A Comparative Study of Analysis Methods in Quantitative Label-free Proteomics

Thesis submitted in accordance with the requirements of the

University of Liverpool

for the degree of Doctor in Philosophy

by

Katherine Isabella Mackay

November 2015

Thesis Abstract – A Comparative Study of Analysis Methods in Quantitative Label-free Proteomics

Katherine Isabella Mackay

The large amounts of data generated by modern proteomics experiments necessitates the use of software pipelines to conduct the bulk of the post-processing. While many software packages (both commercial and open-source) are available to perform some or all of the necessary post-processing steps, it is usual for each research group to use only the instrumentation and software packages with which they are most familiar and/or which are available to analyse their unknown data.

The intention of the studies presented within this thesis was to assess the correlation between the experimental results obtained when;

- a single result dataset is obtained and post-processed in parallel using four separate software pipelines
- a single sample is analysed on two different mass spectrometers and post-processed in parallel

and;

- when different identification thresholds are applied to a dataset prior to parallel quantitation of the resultant data sets

Correlation between different mass spectrometry instruments was assessed and found to yield high r values, especially at the protein level, and was also found to improve following the application of abundance thresholds, however the result of applying score thresholds was unpredictable.

The use of manual FDR thresholds prior to importing data into Progenesis LC-MS yielded interesting results, which suggest that a threshold of 1% peptide FDR and 1 or 2% protein FDR is most effective in terms of yielding accurate ratios while maintaining acceptable sensitivity.

In addition, a consensus method is suggested to utilise the results from multiple software pipelines in order to increase sensitivity and reduce the FDR, through the use of the QPROT tool[1, 2] and manual post-processing.

Table of Contents

1 – Introduction

1.1 – Aims and Objectives

The studies described in this thesis aim to explore the assumptions of reliability and comparability applied to label-free proteomics data, and to suggest possible strategies of evaluating and improving on the accuracy of these assumptions. While there is much innovation in the development of new experimental and post processing techniques, there are few published studies investigating these validation questions. Investigations of this nature are however highly relevant to the field of proteomics and the wider field of molecular biology, especially as both the popularity of label-free proteomic methods and the complexity of those biological systems they are applied to increases.

In the process of the studies presented here, as well as investigating the correlation of results obtained using different post-processing methods on the same data (Chapter Two, page 36), and using different instrument platforms with analogous post processing (Chapter Three, page 70), it was also intended to identify and implement relatively simple and practical strategies that increase the confidence of protein identifications, and hence quantitative values, through the use of multiple post processing pipelines (Chapter Two, page 36) in a manner analogous to the improvement observed when using multiple search engines to add confidence to protein identifications by reducing the false discovery rate (FDR)[3].

Throughout the remainder of this introduction I will present the experimental basis of proteomics and MS, the computation techniques available for the identification and quantification of proteins, and the software and instrumentation available.

1.2 – Proteomics

Proteomics is the study of the proteome, which can be defined as 'the entire protein compliment in a given cell, tissue or organism'[11]. At its simplest proteomics describes an attempt to catalogue all the proteins present in, for example, cells from certain tissues such as muscle or foreskin, or to identify all the proteins expressed by unicellular organisms. This type of identification proteomics has become widely used[12-14], however the questions

now asked by proteomics are much more complex than simple identification (such as asking which proteins are differentially expressed between healthy and diseased cells, or which proteins are modified in given cellular conditions – e.g. following infection of cells with a parasite). Also, it is now often desired to obtain quantitative information from most proteomic experiments. The increased availability of complex instrumentation for use in proteomics workflows for additional information on the instrumentation available) has led to more research groups using proteomics techniques in their studies, and therefore an increased demand for bioinformatics software to process the resultant data[15].

One of the important issues encountered when asking these more complex biological questions is that the abundance of any given protein within the cell is constantly changing and thus any single experiment will provide only a 'snapshot' of the proteins present within that cell at the time of sample collection. While this information is eminently useful for the identification of those proteins present in a given condition, it is necessary to study how protein expression is changing over time in order to obtain a full understanding of the function of proteins within cells and tissues. While time course experiments have been conducted to assess the turnover of proteins within cells[16], this type of study remains expensive and challenging and therefore has not become routine.

It is possible to split the methods used to conduct quantitative proteomic experiments by two main criteria; whether they use labelled or label-free biological samples, and whether they yield relative or absolute quantitation data[15]. Relative quantitation aims to study the abundance of proteins in a sample with respect to each other, or to observe fold changes across different experimental conditions, while absolute quantitation aims to determine the true abundance of the studied proteins in the sample. Both relative and absolute quantitation can be achieved using labelled[17] or label-free methods, achieving relative quantitation stand alone or absolute quantitation using internal standard[18] based strategies.

1.3 – Proteins

Proteins are biological macromolecules possessed of a multidimensional structure as a function of the chemical composition of their constituent polypeptide regions. Each of these polypeptide regions comprises a polymeric chain of amino acids joined by amide bonds, with terminal carboxylic acid and amine moieties, which are joined via amide bonds to form the peptide backbone. A generalised summary of peptide properties and moieties is shown in Figure 1.

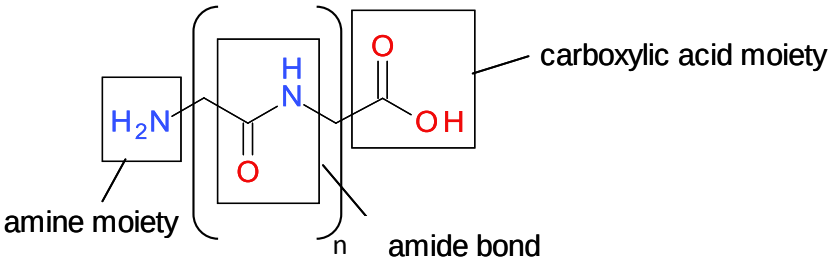

Figure 1: Summary of generic peptide properties.

There are 20 main amino acids that combine in various combinations to create proteins, as shown in Table 1. The order of amino acids in the polypeptide chains, and hence in the final protein chain, is denoted as the primary structure of the protein. The functionality (i.e. the side chains present on the peptide backbone) of each amino acid in the chain contributes to the conformation of the protein chain in space via multiple interactions such as Vann der Waals forces and hydrogen bonding. Hydrogen bonding in particular can lead to a very ordered sub-structure in suitable polypeptide regions of the protein chain, with this structure most commonly taking the form of an α–helix or β–sheet (the properties of each of these types of structure are shown in Table 2). This level of order is denoted as the secondary structure of the protein. Both the α–helix and the β–sheet maximise the pairing of available lone pairs of electrons within hydrogen bonds while minimising steric hindrance between the polypeptide chains. A further layer of organisation, denoted the tertiary structure, refers to the way in which the different polypeptide regions are folded in space. This folding is caused mainly by hydrogen bonding, disulphide bonding or hydrophobic interactions, and is solvent and temperature dependent i.e. the structure may breakdown (or denature) when the protein is heated or dissolved in a new solvent for example. The specific folding of the polypeptide chain allows the protein to become biologically active in specific conditions e.g. within the cell, in order to perform its function within the body[23].

Name		Chemical Formula	
Full	1-letter code	(Neutral Molecule)	Monoisotopic Mass
Alanine	A	$C_3H_7NO_2$	71.0372
Arginine	R	$C_6H_{14}N_4O_2$	156.1011
Asparagine	N	$C_4H_8N_2O_3$	114.0429
Aspartic Acid	D	$C_4H_8NO_4$	115.0269
Cysteine	C	$C_3H_7NO_2S$	103.0092
Glutamic Acid	E	$C_5H_9NO_4$	129.0426
Glutamine	Q	$C_5H_{10}N_2O_3$	128.0586
Glycine	G	$C_2H_5NO_2$	57.0215
Histidine	H	$C_6H_9N_3O_2$	137.0589
Isoleucine	I	$C_6H_{13}NO_2$	113.0841
Leucine	L	$C_6H_{13}NO_2$	113.0841
Lysine	K	$C_6H_{14}N_2O_2$	128.0949
Methionine	M	$C_5H_{11}NO_2S$	131.0405
Phenylalanine	F	$C_9H_{11}NO_2$	147.0684
Proline	P	$C_5H_9NO_2$	97.0528
Serine	S	$C_3H_7NO_3$	87.0320
Threonine	T	$C_4H_9NO_3$	101.0477
Tryptophan	W	$C_{11}H_{12}N_2O_2$	186.0793
Tyrosine	Y	$C_9H_{11}NO_3$	163.0633
Valine	V	$C_5H_{11}NO_2$	99.0684

Table 1: Table of the standard amino acids showing molecular formulae and monoisotopic mass[24].

Secondary Structure	Conformation	Conducive amino acids
α–helix	Polypeptide chains in a right handed helix (3.6aa/turn, length 0.56nm)	Alanine, methionine, leucine, glutamate
β–sheet	Sheet of polypeptide chains of 5-10aa, extended and aligned for hydrogen bonding between NH and C=O groups on adjacent chains	Tyrosine, tryptophan, phenylalanine, valine, threonine

Table 2: Summary of α–helix/β–sheet properties and the amino acids most likely to adopt each structure.

1.4 - Mass Spectrometry

Mass spectrometry essentially involves the measurement of the mass to charge (m/z) ratio of ions within a vacuum. With the exception of time-of-flight instruments and magnetic sector instruments (which are now rarely used), this is achieved via the application of RF and dc voltages to create an electric field in which there is a stable path for only those ions of a specific mass and charge. The mass spectrometry instrument scans sequentially through each mass in a user specified mass range (e.g. 300-3000amu) to identify those analyte ions which are present in the sample.

The original mass spectrometry instruments merely ionised the sample and 'weighed' the ions produced, however it is now more usual to see 'hybrid instruments' which involve a fragmentation stage and thus provide more information about the sub-structure of the ions of interest through the study of (fragment) daughter ions. Though there are several methods of ionisation and mass analysis, the general schematic of the hybrid mass spectrometer remains the same, and is shown Figure 2.

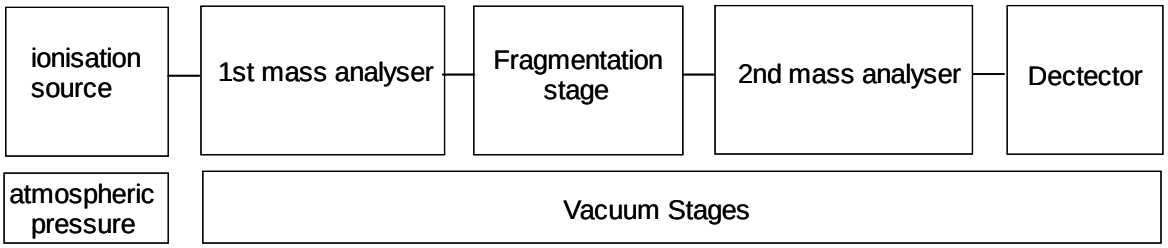

Figure 2: General schematic of the hybrid mass spectrometer.

1.4.1 - Ionisation source:- The ionisation source of choice for the quantitative study of proteins is the nano-electrospray source[25], based upon the standard electrospray source but able to deal with the smaller sample volumes available. Electrospray (ESI) is referred to as a 'soft' ionisation method, with ionisation occurring directly from the sample solution at atmospheric pressure. The ions are then passed through a transition section de-pressurised by a roughing pump and from there into the high vacuum stages which contain the mass analyser and detector. 'Soft ionisation' refers to any ionisation method which produces primarily $[M+H]^+$ ions with minimal fragmentation; this is especially useful when studying unknown analytes as it prevents confusion between parent and daughter ions at the full scan stage ("full scan" refers to a scan which records all ions present with a m/z ratio within the user specified scan range). Manufacturers provide many subtle variations on the basic electrospray ionisation source, however the general principles of operation remain the same. Within the ionisation source the sample solution (a polar solvent in which the sample is soluble) is nebulised from a capillary tube into the source housing via a needle which is held at high potential. The potential difference between the needle and a counter electrode causes charge separation within the liquid and subsequent deformation of the meniscus at the needle end to form a cone (the 'Taylor cone')[26, 27] as shown in Figure 3. At the apex of this cone the charge density is very high and a fine jet of charged liquid is ejected towards the counter electrode. This jet cannot remain stable and breaks up further into charged droplets, which are driven apart by Coulombic repulsion to produce a mist of smaller charged droplets containing the ions which were pre-formed in the polar sample solution.

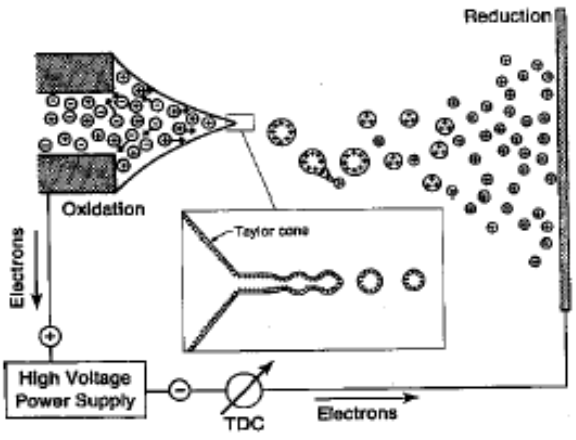

Figure 3: Taylor cone formation at the end of the ESI capillary, and the disintegration of the resulting jet to a mist of small charged droplets[26].

Excess charge density moves to the surface of these droplets as a conducting medium, with the surface charge density increasing as solvent ions are evaporated to give progressively smaller droplets[28]. At some point the surface charge density has increased such that the repulsive Coulombic forces between analyte ions become stronger than the force of the surface tension which binds the droplet together (this point is known as the Rayleigh stability limit). When this point is reached it causes the droplet to 'explode' in a process called Coulombic Fission, which occurs multiple times to produce steadily smaller droplets. Some theories suggested that this Coulombic Fission simply continues until repulsive forces cause the droplet to break up, giving both free and cluster (adduct) ions in the form of charged 'droplets' which inherit the charge of the parent droplet (the Charged-Residue model). More recent studies have suggested that the droplets do not in fact 'explode', but instead eject smaller droplets in a process known as 'droplet jet fission'[29]. This occurs from the elongated end of the microdroplets that are deformed by their flight under the influence of the electric field. The charge density is significantly increased in this area of elongation and a jet of smaller daughter droplets is produced in a process analogous to the formation of the original jet emitted from the Taylor cone at the mouth of the capillary. The daughter droplets formed account for approximately 1-2% of the mass, but 10-18% of the charge of the parent droplet, and are therefore subject to an increased surface charge density[26].

It has also been suggested, by J. Iribarne *et al*, that when a certain radius is reached, ion evaporation becomes favourable over Coulombic Fission and free analyte ions are evaporated from the surface of the droplets – their study calculated this point to equate to a surface charge density of approximately $10^8 Vcm^{-3}$ (this is termed the Ion Evaporation model)[30]. Though there is a continuing debate on the exact method of formation of ions in ESI, it is generally assumed that the CRM is more appropriate for large molecules and the IEM for small molecules, though in actuality the method of formation is likely to be a hybrid of these two models.

For efficient analysis, it is necessary for the charge on the droplet to be the same as that on the analyte ions to allow those ions to move more easily to the surface of the droplet from whence they can evaporate/undergo fission. As peptide molecules are readily protonated, it is usual to conduct mass spectrometry for proteomics in positive ion mode (positive charge on source needle, negative charge on the counter electrode to draw positive ions into the vacuum stages). As the ESI source requires a constant stream of liquid it is usual to couple

the mass spectrometer to liquid chromatography (LC), with the added advantage of increased sample separation and purification of the individual components of the sample into separate chromatographic peaks. Thus in addition to ionization mode, the choice of solvent used in the LC stage should also be carefully considered in terms of ionisation suitability, as well as chromatographic efficiency, as the more 'hydrophobic' an analyte is in a given solvent, the easier it will be to overcome the solvation energy and the ions will be more easily desorbed into the gas phase[28, 31, 32]. However this must be traded with the ability of analyte molecules to form ions or at least strong dipoles in the solvent solution, which enhances ion abundance[29]. Taking both considerations into account, it can be seen that choosing the correct solvent can therefore have a significant effect on the efficiency of ionisation, and the ideal parameters will vary greatly for different analytes. Therefore all solvent characteristics should be optimised for specific analytes – including the pH of the chromatographic mobile phases (as positive ions are most readily formed in acidic solutions and vice versa).

Once desorbed within the electrospray source, the free analyte ions (and any analyte adduct ions) are drawn into the mass spectrometer via the ion sweep cone and passed into the 1st mass analyser.

1.4.2 - MS analysis (1st mass analyser):- The 1st mass analyser conducts full scan MS analysis, recording the analyte ions according to their m/z ratio. In almost all hybrid instruments this first stage will involve a quadrupole mass analyser.

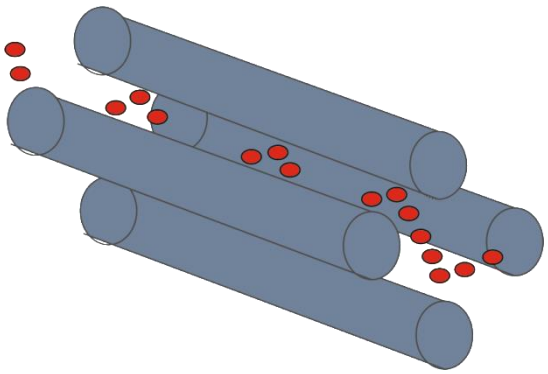

Figure 4: Stable ion trajectories within the quadrupole mass analyser.

The four rods in the quadrupole are divided electrically into two pairs, with the voltages applied to each rod pair being equal in amplitude but opposite in sign. The quadrupoles repeatedly 'scan' through the specified mass spectrum (e.g. 300-3000 atomic mass units) by altering the RF and dc voltages applied to each of the quadrupole rod pairs. Each RF/dc voltage combination will create a stable path through the quadrupole for ions of a certain m/z only (such a stable path is shown in Figure 4), with ions of all other m/z values following a collisionary path and being neutralised (by collision with the rods) or ejected from the cell (by passing out between the rods).

In hybrid instruments with multiple mass analyser stages, the first stage of mass analysis within the quadrupole mass analyser can be used for mass scanning where parent masses are measured without collecting fragmentation data; however in the time-of-flight or orbitrap instruments often used for proteomics it is generally desired to exploit the higher mass resolution of these mass analysers to gain accurate mass data for the parent ions (and hence greater confidence in their identification). In this situation the quadrupole functions as an ion transfer device in the first stage of mass analysis, creating a stable path for all ions within the specified mass range and thus passing them forward to the TOF or orbitrap mass analyser for full scan mass analysis of the parent ions.

1.4.3 - Fragmentation stage:- Fragmentation of the parent ions allows more information to be gained about the structure of the parent molecule via analysis of the fragmentation pattern i.e. the m/z ratios of the fragments formed. The most common method of ion fragmentation used in proteomics is 'collision induced dissociation' (CID), occurring within a 'collision cell' containing (most usually) a second quadrupole functioning as an ion transmission device, with the RF voltage on the rods such that there is a stable path for all ions within the selected mass range of m/z values. The collision cell is pressurised with a non-reactive gas (most usually argon), which is the agent of collision-induced dissociation, i.e. fragmentation. This fragmentation occurs when the analyte ions collide with neutral atoms of the collision gas, thus transferring some translational kinetic energy to internal energy and placing the ion in an excited state. If the transfer of energy is sufficient the ion will dissociate into characteristic fragments, and the degree of fragmentation can be optimised by altering the collision energy applied. The term "collision energy" refers to the application of a dc bias voltage to the gate electrodes preceding the second quadrupole, which creates a potential difference between the source and the collision cell and thus

increases the translational kinetic energy of the ions as they enter the collision cell. As the translational kinetic energy of the ions increases, so does the amount of energy transferred in any given collision and consequently the degree of fragmentation will increase with increased collision energy. Optimisation of the collision energy aims to maximise the production of diagnostic fragments – it becomes difficult to resolve the structure of large fragments as the possible ion combinations to produce a given m/z ratio increases, and small fragments eventually become too common to use in structure elucidation. To ensure there is no crossover of fragments between parent ion masses it is usual for the collision cell to be purged of ions between scans by applying a large voltage of the opposite polarity to the rod pairs.

1.4.4 - MS-MS analysis (2nd mass analyser):- MS-MS analysis provides information about the structural fragments of the analyte molecule, and this information can be used to elucidate the structure of the parent molecule. In protein research and with standard settings fragmentation generally occurs first at the amide bonds of the peptide molecules, though there is an ongoing area of research to identify fragmentation rules for peptide ions under different conditions both during analysis and in the system of interest [33-36]. Within the second mass analyser a full scan analysis is performed on the fragment ions passed from the collision cell. In some instruments this geometry is less intuitive in practise, most notably in the case of some instruments which interface iontrap and orbitrap mass analysers, where ions are passed first to the orbitrap (via the iontrap set to allow a stable path for all m/z ratios within the specified mass range) and then back through the collision cell with the resultant fragment ions passing into the iontrap for analysis. This allows the instrument to achieve the greatest mass accuracy for the parent ions with the lower resolution iontrap being used for fragmentation analysis.

1.4.5 - Mass Analyser Types

1.4.5.1 - Time-of-flight (TOF) mass analyser:- The TOF mass analyser measures m/z as a function of the 'drift time' of a given ion within a flight tube of specific length, when accelerated towards a detector under vacuum conditions.

Before entering the flight tube, ions are accumulated in the collision cell via the application of a voltage of the same polarity as the analyte ions to the end gate of the collision cell, which repells the ions from the gate. These accumulated fragment ions are then allowed into the flight tube to coincide with the next TOF-pulse, by swapping the polarity on the end gate to allow the ions to move out of the collision cell.

The TOF-pulse itself is created by the application of a strong voltage of the same polarity as the analyte ions to orthogonal accelerator (pulsar) plates - this propels the ions out of the end gate towards the detector. In modern instruments the ions often reach the detector not by a linear flight path but via a reflector, or in some cases several reflectors, which serve to normalise the energy difference between ions of identical mass but differing kinetic energy, and also allows the flight tube section of the instrument to remain physically the same size despite increasing the length of the flight path for ions[37]. The reflector itself is a set of plates charged with the same polarity as the analyte ions, which 'reflects' the ions towards the detector by electrical repulsion. Normalisation is achieved as those ions with higher kinetic energy will penetrate further towards the plates and lose their 'excess' energy through repulsion. Thus all ions of the same m/z ratio are caused to reach the detector after the same drift time within the flight tube. A schematic of a TOF-analyser containing a single reflector is shown in Figure 5.

Once ions are detected, the m/z measurement itself is calculated from the drift time (the time between being accelerated by the TOF-pulse and reaching the detector), the acceleration voltage applied and the length of the flight tube.

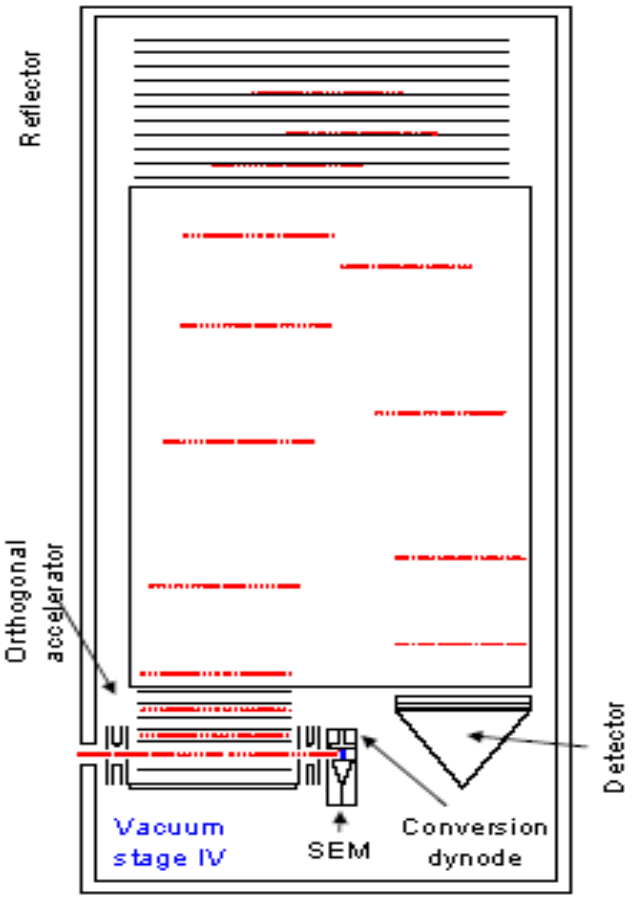

Figure 5: Simple schematic of a TOF tube including a single reflector (as found in the Bruker Microtof Q[38]).

1.4.5.2 - Iontrap mass analyser:- The term 'ion-trap' describes a group of mass analysers including linear and three-dimensional ion-traps. A general schematic for each of these is shown in Figure 6.

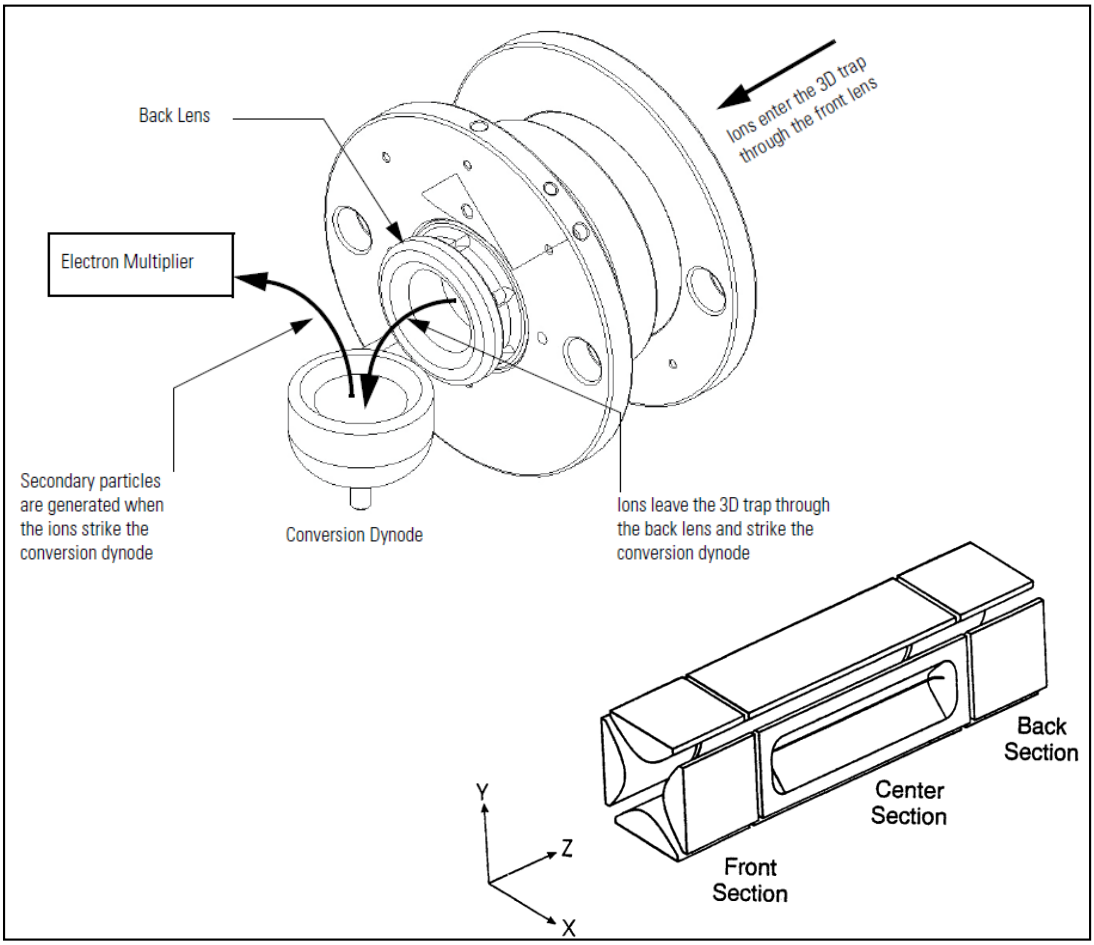

Figure 6: Geometry of the three-dimensional and linear ion-trap mass analyser[39].

The general principle of both iontrap analyser types remains the same, in that ions are trapped within the electrodes as a function of their mass and charge. By altering the applied RF and dc voltages to create potential wells of stability for ions with a selected m/z ratio only, the full mass spectrum may be scanned by sequentially ejecting ions of particular m/z ratios to the detector. Alternatively, as a collision gas is present, ions of a particular m/z ratio may be trapped, accumulated (with the ejection of all other ions), and fragmented with the resultant fragment ions then being scanned out to the detector. The fact that fragmentation and scanning can occur in the same mass analyser allows for multiple levels of analysis, i.e. it is possible to excite and fragment the first, second, etc level fragment ions (this capability is described as MS^n) as well as the original parent ions. This may be done by selecting the m/z ratio of the ions to be trapped and fragmented prior to beginning the experiment, or by

allowing the instrument to select for fragmentation those ions that are observed at the highest ion intensity.

1.4.5.3 - Orbitrap mass analyser:- The orbitrap mass analyser was introduced by Makarov[40] and is a modification of the quadrupole iontrap mass analyser. It is often interfaced with a linear iontrap, which performs the fragmentation stage (as fragmentation within the orbitrap is slow) and mass analysis of the fragment ions, with the orbitrap being used due to its ability to provide high mass resolution for the parent ions. The functional principles are similar to those which describe the iontrap mass analyser, however an electrostatic field (applied dc voltage) only is used to create a stable path for the ions of interest, without the use of RF voltages. The analyser is composed of a spindle electrode and a pair of bell shaped outer electrodes, with the pair being separated by a ceramic insulation ring[41]. A schematic showing the spindle and bell electrodes, as well as the trajectory of trapped ions within them, is shown in Figure 7.

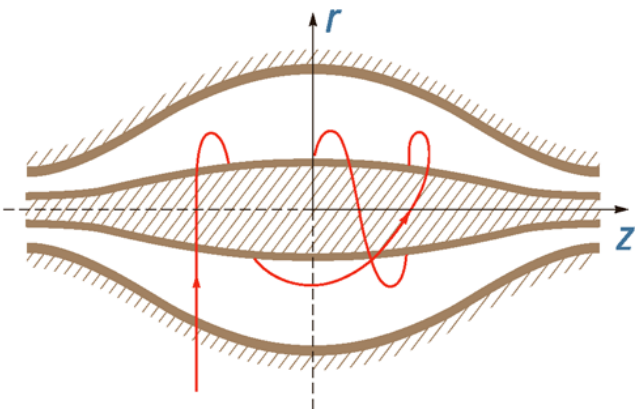

Figure 7: Trajectory path of trapped ions within the orbi-trap mass analyser[42].

Ions which have a velocity vector perpendicular to the spindle electrode, and which obtain a velocity of a selected magnitude (applied by the instrument as a function of the m/z ratio being recorded in a given scan), are trapped into a stable orbit around the central electrode. The axial component of the ion oscillation is detected as an image current on the two halves of the bell electrodes. A fourier transform is then applied to determine the m/z ratio of the ions that are trapped. Once trapped, ions of sequential m/z ratio are released to the detector by altering the dc voltage applied to the spindle electrode (in practice a potential is also applied to the end-cap or 'gate' electrode to regulate the release of ions towards the detector).

1.4.6 LC-MS Output Data

As the sample passes first through a chromatography column before reaching the mass spectrometer, there is chromatographic separation of the individual molecules in the sample prior to ionisation. As each peak i.e. peptide (or group of peptides, when several peptides have the same or very similar physico-chemical properties) elutes from the chromatography column it passes into the ion source of the mass spectrometer. Therefore each elution peak from the chromatography system contains a "packet" of molecules, which are ionised and subsequently recorded at the detector as a peak in the ion abundance. This is reflected in the LC-MS data output, which is known as the total ion chromatogram (TIC)[26]. As the mass spectrometer performs scans, it produces one mass spectrum per scan that shows the m/z ratios of the ions observed in that scan. The total ion abundance recorded in each scan is shown in the TIC, as a function of the retention time[26].

From the TIC, it is possible to computationally construct an extracted ion chromatogram (XIC or EIC) that shows the elution profile only for those sample ions with a user selected m/z ratio. The XIC can be useful to identify related sample molecules, for example one that has been deuterated versus one that has not, that have different m/z ratios but the same elution profile. Using this simple derivatisation strategy it is possible to assess the relative amounts of an analyte in different conditions by performing deuterium exchange on only one of the samples. This will give two peaks in the mass spectrum which are separated by a known mass difference (deuterium is the 'heavy' isotope of hydrogen, possessing a neutron in the nucleus in addition to the proton and electron found in hydrogen and therefore having an atomic mass of 2 – this gives rise to a mass shift of 1 mass unit for each deuterium atom present when studying singly charged molecules).

The mass spectra recorded throughout the TIC can be viewed and analysed to determine which ions are present in each scan[26, 43]. However, rather than a single peak for each ion present, the mass spectrum will show an isotope pattern of several peaks. This is due to the presence of less common isotopes present within the molecule at natural ratios, for example C^{13} and to a lesser extent N^{15}. This isotope pattern includes the monoisoptopic peak, which corresponds to the presence of the most abundant isotopes (eg C^{12} and N^{14}), and less intense peaks corresponding to molecules containing less common isotopes (with the relative

abundance of the peaks mirroring the naturally occurring distribution of those isotopes present). An example isotope pattern is shown in Figure 8.

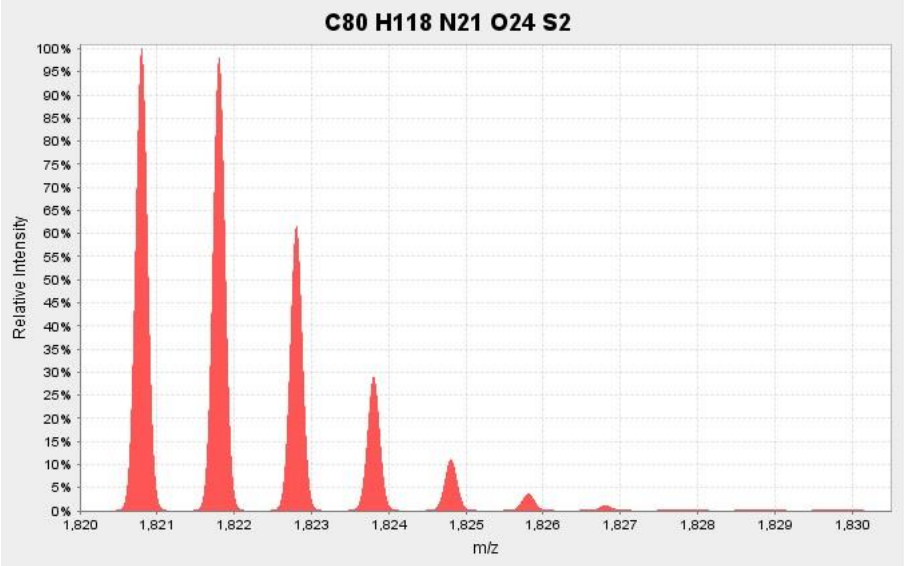

Figure 8: An example isotope pattern, for the given protein (generated using the IPC (Isotope Pattern Calculator) tool[44].

Unfortunately, an experimental mass spectrum is unlikely to display a perfect isotope pattern for a variety of reasons. One possible reason for this is that the low abundance peaks in the isotope pattern may be obscured by noise peaks of similar abundance, and another is that co-eluting analytes may have overlapping isotope patterns that can be hard to resolve from each other. Thus both obtaining chromatography parameters that separate individual analytes as much as is possible and ensuring that any computational algorithms are able to elucidate isotope patterns reliably is essential to meaningful analysis of experimental data.

1.5 - Protein Mass Spectrometry

1.5.1 – Analysis of Whole Proteins

The mass spectrometric analysis of whole proteins may be conducted using ESI, however more usually a iso (Matrix-Assisted Laser Desorption Ionisation) ionisation source is used.

Laser desorption ionisation (LDI) was first introduced in the late 1960s[26], and was readily applicable to the analysis of organic salts (low-mass) and light absorbing compounds. However there was not utility to apply this method to protein analysis until the late 1980s, as it was non-trivial to obtain useful spectra for biomolecules, especially those with mass exceeding 2000 atomic mass units. Two methods were put forward that allowed analysis of biomolecules via LDI; the mixture of the analyte (in glycerine) with ultrafine cobalt power, and the co-crystalisation of the analyte with an organic matrix (MALDI). This second method gained more use as it gave greater sensitivity and versatility as a technique, though both methods are capable of producing useful spectra for molecules with molecular weight up to 100000 atomic mass units.

A standard sample preparation for MALDI sees the analyte dissolved at approximately 0.1mg/ml and the matrix dissolved to saturation, or at a concentration above 10mg/ml. The admixture of these two solutions puts the analyte:matrix ratio within the range of 1000:1 to 100000:1, which is optimal for the production of good MALDI mass spectra. 0.5-2µl of sample is deposited on a plate (or MALDI target) to give a thin layer of sample over the matrix. Two methods to achieve this are the evaporation of the matrix solution followed by application and evaporation of the sample solution without re-dissolving the matrix, or the analyte may be introduced to a pre-prepared matrix using nano-ESI to ionise the sample.

Ionisation in the MALDI technique utilises laser desorption of sample ions from the surface of the matrix. A laser beam (typically ultraviolet light but infrared radiation may also be used) is focussed on a small area of approximately 0.05-0.2mm diameter and the absorption of this laser light by the sample layer causes evaporation and then ionisation of the sample atoms. The method of ionisation is thought to be a combination of the evaporation of pre-formed ions from the matrix and the gas phase photoionisation of evaporated atoms once they are within the plasma plume. The ion source is operated at room temperature and the laser

attenuation may be optimised for each measurement. A schematic of the MALDI ionisation source is shown in Figure 9.

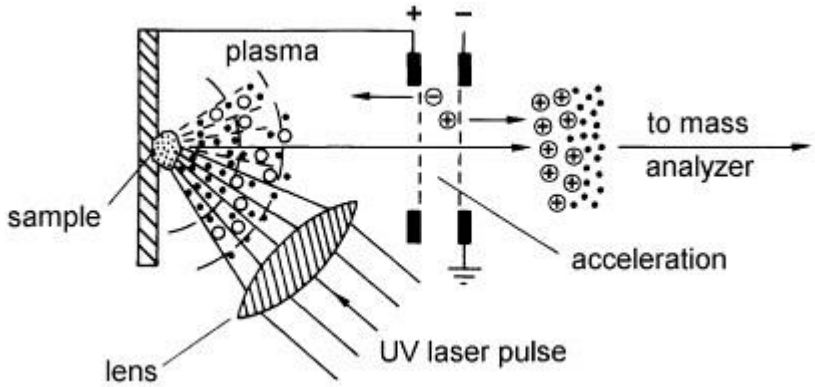

Figure 9: Schematic of the MALDI ionisation source[26].

The mass analyser used with the MALDI technique is most usually a TOF analyser (or an FT-ICR (Fourier Transform Inductively Coupled Plasma instrument, if the analysis of very high mass analytes is required). When an analysis is conducted under specified experimental conditions using a MALDI ion source, a so-called 'peptide mass fingerprint' (PMF)[45] mass spectrum of high mass quasi-molecular ions[46] is produced, which is unique to a given protein. This can then be used diagnostically by comparison with library spectra (in the same way that library spectra have historically been used to identify small molecules analysed by GC-MS with electron impact ionisation (EI)[47]) to rapidly identify a single protein within a gel spot[48], and quantification may be achieved through the use of internal standards of known concentration. The MALDI technique cannot, however, facilitate the identification of unknown proteins within a sample mixture such as is required for the analysis of modern proteomic data (e.g. to study proteins for which there are no library spectra available such as to identify previously unknown proteins within a tissue sample).

1.5.2 – Analysis of Peptides as a Method to Identify Unknown Proteins

Previously unidentified proteins in complex mixtures can be detected and sequenced using a complex strategy, which involves tandem mass spectrometry utilising the MS-MS functionality of hybrid mass spectrometers with a nano-ESI ion source. Prior to analysis the sample protein mixture is digested with a protease enzyme - most usually trypsin, which according to the Kiel rule "cuts" the protein at the C-terminal end of arginine and lysine peptide residues, but not before proline residues[49]. However cleavages before proline have been described experimentally and the suggestion made that excluding the resultant peptides from the pool of theoretical peptides may prevent the identification of peptides which are in fact present in the sample[49]. In addition, variation in the peptides produced following trypsin digestion may be introduced when the protein sample is not completely digested leading to "missed cleavages", where a potential cleavage site is missed giving one larger peptide where two smaller peptides would be expected. Therefore the algorithms used must be capable of considering the possible variations to "perfect" digestion. However, it has also been observed that there is no significant difference in precision, accuracy, specificity or sensitivity on the inclusion or exclusion of peptides resulting from missed cleavages[50].

Due to the protease pre-digestion step the mass spectrometric analysis is actually performed not on the whole proteins within the original sample, but on the peptides resulting from sample digestion and further on their daughter fragments created within the mass spectrometer (e.g. from collision induced dissociation within a collision cell), bringing the molecules being studied below 3000 atomic mass units. The abundance of the constituent peptides in the sample is therefore used as a proxy for the protein abundance present in the original sample, making it extremely important that identified peptides are assigned to the correct parent proteins. As peptides are smaller in size there are less possible combinations of atoms that account for the recorded mass of these peptides and their fragments than there would be for larger molecules, and in addition there are less possible charge state combinations. This makes the task of structure elucidation less computationally demanding, and in addition allows further structural information still to be obtained from the fragmentation data.

Both proteins and peptides are predominantly protonated on the amino groups during ESI/nano-ESI, i.e. at the N-terminus and on arginyl, histidyl and lysyl residues[51]. Most usually singly charged peptides are observed, though 2^+ and 3^+ charge states are also common[51], and whole proteins have been shown to gain approximately one charge per kDa mass[51]. As a general rule for ESI, the signal seen by the mass spectrometer is proportional to the concentration of analyte in the sample[26, 51] and this relationship is flow-rate independent. However this relationship may be affected by the ionization efficiency of the individual peptides, which can give a more confused picture of protein abundance as the intensity of signal from individual peptides may differ even when they arise from the same parent protein molecule present at a given true abundance.

During the second stage of analysis within the collision cell the peptides fragment predominantly at the amide bonds, producing smaller amino acid chain daughter fragments. The main ion types which are produced as a result of fragmentation at standard collision energies have been classified as b- and y-ions. The y-ions arise sequentially from the C-terminus of the parent peptide ion, with y_1 describing the amino acid lost if the first amide bond is cleaved and so on. Correspondingly the b-ions arise from the N-terminus of the peptide ion, again with b_1 describing the amino acid lost if the first amide bond is cleaved. The sites of fragmentation at the amide bond that create both b- and y-ions are shown in Figure 10.

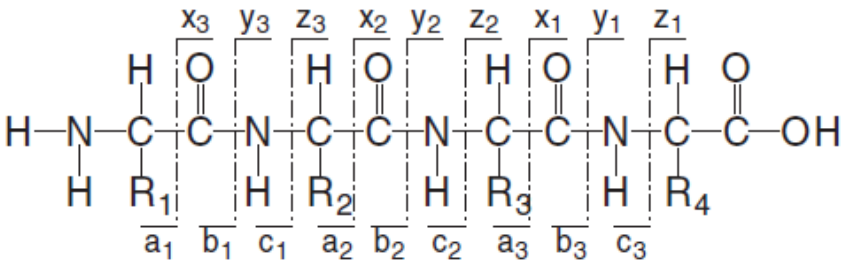

Figure 10: Peptide fragmentation at the amide bond to produce b and y ions[52].

There are several reasons making it unlikely that all of these fragment ions will be observed in a given mass spectrum, including the differing stabilities of both parent and fragment peptide ions within the mass spectrometer, relative bond strengths within the peptides and the applied collision energy. However, those which are observed may be used diagnostically to deduce the sequence of amino acids in the parent peptide and hence identify the peptides (peptide-level data) and ultimately their parent proteins (protein-level data). For the current

study, the terms feature and peptide-level data can be considered synonyms. However, there are packages that can treat features and peptides differently, for example in the case of two different charge states for the same peptide – where these would be treated as two features, but aggregate values used to give a single peptide.

1.6 - Identification of Proteins from Peptide MS-MS data

Though this originally involved *de novo* sequencing from manual analysis of the data, as is done for the analysis of small molecules[24, 53], it is now more usual to use database searching software for protein inference. This increases the automation of proteomic experiments and greatly decreases the time required for this step, therefore bringing the analysis of hundreds of unknown proteins into the realms of practicability. A summary of the steps involved in this automated process is shown in Figure 11.

Following the practical MS analysis the result files are input to protein inference software or scripts for further post processing and analysis. The software receives as input the m/z of parent and fragment ions, the intensity of the fragment ions, plus a text file containing the peptide sequences (in one letter notation) for all the known and hypothetical proteins expressed by the species of interest – a 'fasta' file (compiled from the gene sequence data). According to user specified parameters, most essentially the protease enzyme used, theoretical peptide spectra are then created within the software for the constituent peptides of each protein listed in the .fasta file. These theoretical spectra are then compared to the observed experimental spectra and ranked in terms of how well the two match when overlaid. This sounds simple, but is complicated by several factors including the presence of a, c, z and x ions as well as the more abundantly seen b and y ions, and the presence of noise peaks that can confuse the isotope pattern of features (by appearing to be one of the peaks in the isotope pattern, or by obscuring low intensity "true" peaks) or appear to be peptide features themselves. Therefore the experimental spectrum will always be an imperfect match for the theoretical one and there may be several possible matches for any given observed spectrum, with the highest ranking (by best match) theoretical spectrum usually being assigned as the identity of the peptides present.

As the peptides that are identified as present in the data represent a sub-sequence of their parent proteins, the presence of those proteins within the sample may be inferred from the presence of their constituent peptides. Each protein in the sample may be represented by single or multiple peptides, and there may be ambiguity when a peptide sequence could be assigned to multiple parent proteins. Frequently, a threshold is applied which requires that each protein has been inferred from at least two or three quantified peptide features, to avoid reliance on single peptide features for identification or quantitation of proteins.

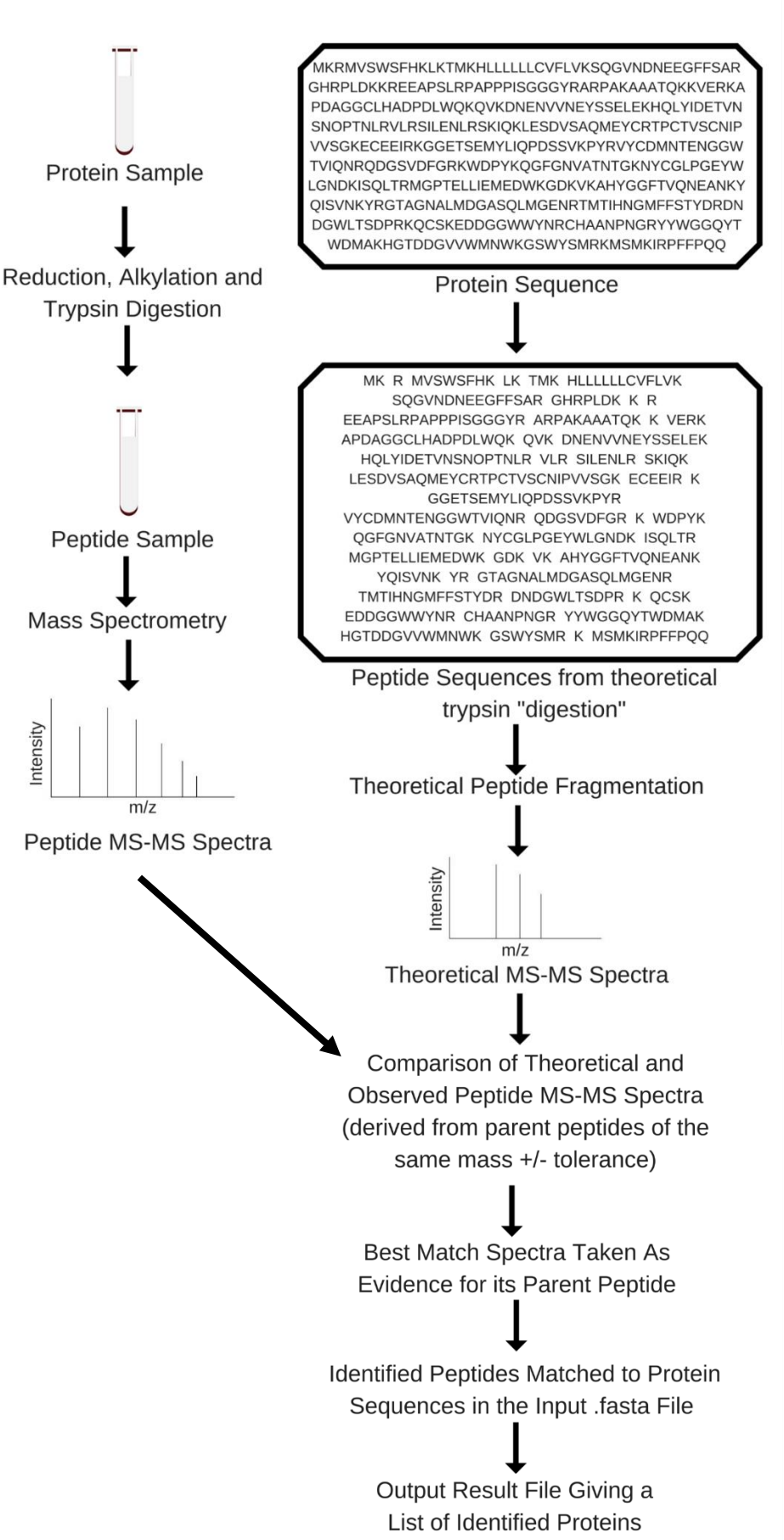

Figure 11: Schematic showing the main stages present in a typical proteomics experiment.

1.7 - Concatenated target-decoy Database Searching

While it is possible to assign protein probabilities calculated from the peptide probabilities of the assigned constituent peptides[54], as implemented in the Trans Proteomic Pipeline web based software, it is increasingly common for a concatenated target-decoy database to be used so that it is possible to calculate a false discovery rate (FDR)[55] as a measure of confidence.

The concatenated target-decoy database is created from original .fasta files containing all the proteins predicted from gene models of the organisms being studied (if more than one organism is present, e.g. when studying the infection of human cells with a parasite, the .fasta files for all the species present are concatenated into one larger .fasta file). The decoy section of the concatenated target-decoy database is most usually created by appending either a reverse set of all the proteins in the .fasta file (generated by reversing the peptide sequences of the predicted proteins present in the original .fasta file), or a set of artificial protein accessions that are assigned to random peptide sequences. The decoy sequences are clearly denoted as such in the target-decoy database .fasta file, most usually by including a "Decoy" or "REV" suffix concatenated with the accession number of the real sequence, or for randomised concatenated target-decoy databases a "random" prefix with an arbitrary unique identifier (such as a letter or a number) may be used. This method of including equal numbers of target and decoy protein sequences in a single file means that there is equal competition between the real and the decoy sequences, with no bias towards either group.

The concatenated target-decoy database is searched in the usual manner in order to find the best fit theoretical spectra in terms of matching to the experimental spectra. Many search engines now also include an option to set an FDR threshold as a measure of confidence in the results returned by that search engine. The assumption is that if a given number of decoy proteins are incorrectly identified to be present in the sample, it can be assumed that the same number of "real" predicted proteins will also have been incorrectly identified as present by the software. Where this is not an option built into a software package it is possible to calculate a manual FDR by looking for the presence of decoy proteins in the result list of protein accessions, using the formula FDR = FP / FP + TP (where FP is the number of

observed false positives i.e. concatenated target-decoy proteins and TP is the number of observed true positives i.e. real protein accessions). This method can be useful to determine the peptide or protein score that corresponds to the desired FDR, when it is applied after the protein list has been sorted by the score of interest, and it is then possible to disregard those protein identifications that do not meet this threshold.

1.8 - Labelling Techniques in Proteomics

Though this thesis is concerned with the bioinformatics methods used in the context of label-free proteomics, it is useful to include an explanation of common labelling techniques within this introduction to provide context for the term "label-free". "Labelling" is the term used to describe the derivatisation of the sample molecules in order to differentiate between samples from different conditions, e.g. before and after drug administration, and may be achieved through *in vivo* metabolic labelling or *in vitro* chemical labelling[15]. Each of these has its own advantages and disadvantages; *in vivo* labelling is more efficient and yet prohibitively expensive for some studies (e.g. when looking at a whole complex organism such as a chicken[56]), while *in vitro* labelling can theoretically be applied to any sample[15].

In addition there are several internal standard based labelling strategies that involve the use of synthetic or proteotypic (the term "proteotypic" describes those peptides which are unique to a given protein, as opposed to those peptides that could be assigned to multiple proteins predicted in the sample) peptides as the internal standard. From the known abundance of this internal standard the abundance of experimental proteins can be calculated using the relative intensities of the internal standard and the experimental protein peaks in the mass spectrum. Typically the internal standard peptides are synthesised or derivatised[19] to give a predictable mass offset allowing for ease of differentiation between the internal standard and the peptides of interest. These internal standard techniques include AQUA[20] (internal standard peptides derivatised to give a known mass difference in the result data), QconCAT[21] (derivatised peptides concatenated into a synthetic protein via bacterial culture) and SISCAPA[22] (derivatised peptides isolated with experimental peptide analogues using immobilised anti-peptide antibodies). As with relative quantitation methods, the protein abundance is calculated from the peptide abundance observed for those peptides assigned to that protein. Thus the same challenges of peptide/protein inference and abundance calculation are present, and there are a great variety of tools designed to address these issues. However there is no one simple tool that can be applied as a standard technique, and therefore this is an ongoing area of research.

1.8.1 - Stable Isotope Labelling with Amino Acids in Cell Culture (SILAC)

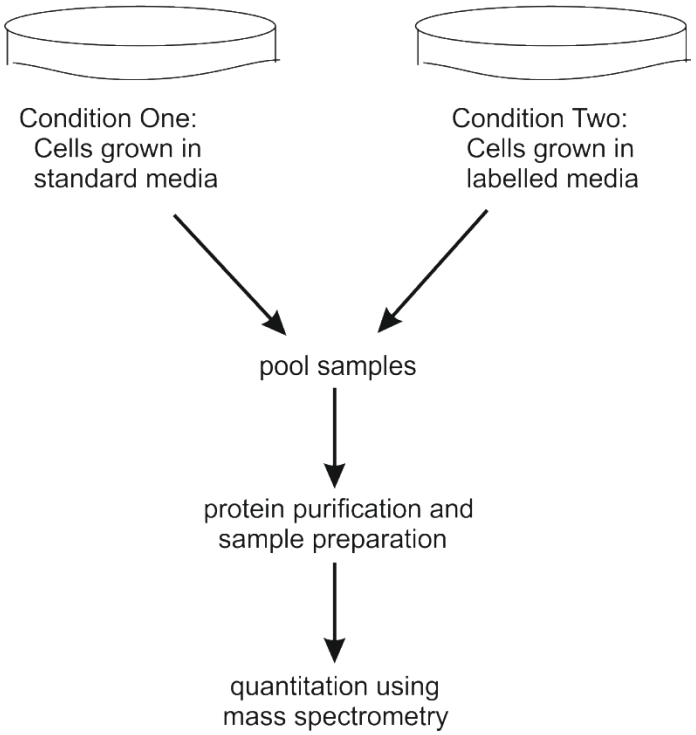

Figure 12: Example SILAC work flow.

This type of labelling experiment is particularly useful when studying cells in different conditions, e.g. when looking at the effect of drug treatment or the changes following infection with a parasite, and a summary of an example workflow is shown in Figure 12. The cells in condition one are grown in standard media and those in condition two are grown in labelled media, i.e. media containing the heavy isotopes of one or more element within the media such as deuterium or ^{18}O. The samples from each condition are then pooled together for the protein purification, reduction, alkylation and digestion steps. Running this pooled sample through a mass spectrometry pipeline produces a final mass spectrum with paired peaks, separated by the mass difference between the labelled and unlabelled peptides. This means that differential expression between the two conditions can then be inferred from the relative intensities of the labelled and unlabelled peaks.

While the least expensive labelling strategy is to use deuterated amino acids, several other labelling atoms are routinely used with the most common strategies using ^{13}C and ^{15}N labelled amino acids which give rise to a larger mass shift. This larger mass shift between paired peaks makes it easier to separate out those paired peaks from a complex mass spectrum. As the labelled and unlabelled amino acids experience the same chromatographic and ionisation conditions these fold change values are highly reliable and it is this accuracy which means that SILAC is one of the most commonly used proteomic techniques.

1.8.2 - Isobaric Tags for Relative and Absolute Quantification (iTRAQ)

This technique concerns the use of commercial iTRAQ reagents, which bind to peptides following protease digestion, and like SILAC it utilises the pooling of samples from different conditions of interest[57]. The iTRAQ reagents are synthetic molecules which contain an N-hydroxysuccinimide (NHS) ester derivative which binds to primary amino groups within the peptide molecules via an amide bond (this derivatisation at the peptide stage increases the complexity of the sample by one to two orders of magnitude and therefore iTRAQ is not best suited to highly complex samples). The NHS reagent also contains a mass balance group of carbonyl moeity and a reporter group based on N-methylpiperazine. The mass balance group compensates for the different reporter moieties, ensuring that labelled peptides from different conditions appear at the same m/z in the parent mass spectrum. During MS-MS analysis the mass balance group is released as a neutral fragment and the reporter group forms an ion which can be observed in the low mass region of the MS-MS spectrum (114-117m/z[15]). This use of reporter ions removes the need to extract mass pairs from potentially complex spectra, with the fold changes between conditions instead being calculated from the relative intensities of the reporter ions. Originally a set of four iTRAQ reagents was available (allowing five experimental conditions to be studied i.e. four labelled conditions and one unlabelled condition) but this has now been increased to eight, allowing differential expression to be assessed across nine experimental conditions (e.g. nine time points in a time course experiment).

1.8.3 - Isotope-coded Affinity Tags (ICATs)

The ICAT[58] method also utilises synthetic molecules to derivatise the proteins in the sample. The ICAT reagents have three specific chemical regions: a reactive section, an isotopically coded linker section, and an affinity tag. Once again the technique is designed to find differential expression between two conditions, but this time both samples are derivatised. The sample from one condition is labelled with the heavy form of the reagent and the sample from the other condition is labelled with the light form of the reagent. In both cases the reagent binds to the side chains of cysteinyl residues in the reduced protein samples. The two samples are then pooled and digested to peptides, and the tagged peptides are then isolated by avidin affinity chromatography. The isolated peptides are then analysed by mass spectrometry, with the identification completed as above from the MS-MS spectra and the differential expression inferred from the intensity ratios between the heavy and light peak pairs in the MS spectra. While effective, this technique by its nature cannot quantify cysteine-free proteins and as the cysteine content of proteins is often low this lowers its utility in practice with real samples.

1.9 - Label-free Proteomics

There are two main limitations to the labelling techniques used in proteomics; the first is the high cost of labelled media and the second is the large amount of time taken to prepare fully labelled biological samples.

Label-free work flows do not carry the increased time requirement and complexity of sample preparation that is associated with labelling techniques, however there is an increased requirement for instrument time as parallel LC-MS-MS runs of each separate sample are required (rather than a single run of pooled samples), and in addition more technical replicates are required to manage the greater variability of label-free samples (with respect to labelled samples where all the experimental steps are common after samples have been pooled rather than processed in parallel). While this may be a concern if instrument time is at a premium (e.g. when many groups share a single instrument), as the LC-MS-MS runs can be automated there is time for the analyst to be preparing further samples while awaiting the results.

Label-free quantitation is achieved via direct comparison of the LC-MS-MS data from several runs, and this means that in order to obtain meaningful results it is necessary that the instruments used are capable of delivering extremely high mass accuracy and reproducible chromatography retention times to allow the separate samples to be accurately compared across runs.

1.9.1 - Quantitation Methods

Quantitation for both labelled and label-free methods is typically carried out either via the analysis of extracted ion chromatograms (XICs) to give intensity values (as peak height or area – these are known as intensity based methods)[59] or by using the number of fragmentation spectra matched to peptides as a proxy for the protein abundance in the sample (these are known as spectral counting methods)[60].

1.9.1.1 – Spectral Counting Methods

Spectral counting methods are based on the assumption that those peptides present at high abundance will be observed as a relatively large peak in the elution profile at the LC stage, and are therefore be both more likely to be detected in relatively large number of scans and to be selected from a given scan for fragmentation analysis (for example as one of the top 3 most abundant ions present), and that therefore a count of the observed MS-MS spectra for that ion will be representative of the peptide abundance in the original sample (and that this will be reproducible in repeat experiments). The estimated parent protein abundance can then be calculated from these peptide abundance values (by taking the sum of all MS-MS spectra recorded from peptides which have been assigned to the parent protein)[61-63]. There are potential flaws in this correlation of spectral counts with protein abundance as the differing chemical composition of a proteins constituent peptides causes differing ionisation efficiencies within the ion source of the mass spectrometer, potentially making the observed "abundance" an average value which may include a high spectral count for peptides which ionise well and a low spectral count for those which don't. In addition, where peptides of similar mass elute together it is non-trivial to determine which peptide a given MS-MS spectra should be assigned to. There have been several attempts to compensate for this potential problem by weighting the calculations according to the expected ionisation

efficiencies based on the peptide properties (hydrophobicity etc) and using machine learning techniques, such as in the APEX Quantitative Proteomics Tool[64, 65] (which does both).

1.9.1.2 – Intensity Based Methods

Intensity based quantitation is achieved via calculation of the peak height or area for each peptide taken from the MS spectra for that parent mass once the peptide has been identified from the MS-MS spectra. The intensity recorded for each peptide will be a summation of the intensities for all ions present for that peptide, e.g. 1^+ and 2^+ ions. A ratio is calculated for each peptide based on the intensities found in each condition, and protein fold change obtained from combining the peptide ratios for all peptides assigned to that protein[61]. This technique requires accurate parent mass values, in order to minimise the issue of peptides of similar mass and eluting together (either in the same peak or in overlapping peaks) being quantified as one peptide.

Recently, there has been a move towards quantification of label-free data via the alignment of accurate mass and retention time across multiple runs to form an aggregate spectrum[66], and a typical work flow for this alignment process is described below;

1. Signal pre-processing

The types of pre-processing required will vary for different datasets, however some of the more common steps are noise removal (noise may arise from contaminants in the sample or mobile phase, or from electronic noise within the mass spectrometer), baseline correction (the baseline is the signal intensity recorded between peaks when no sample peaks are being detected, and this intensity can drift up or down over time) and file conversion (though noise removal and baseline correction are increasingly redundant as instrumentation improves and these steps are incorporated into software packages). The need for file conversion is dependent on the compatibility of the output format from the instrument being used with the chosen analysis software, and therefore may not be required.

2. Feature detection and quantitation

Peaks present in the TIC are converted to two-dimensional centroid data, and those groups of centroid "peaks" which resemble an isotope pattern are selected for analysis[67]. Each isotope pattern will be present in a number of consecutive scans, for the duration of the elution profile of the chromatographic peak containing the peptide. Once detected, features are quantified by modelling the peak as a Gaussian distribution and calculating the area under the curve. This is broadly correct, however chromatographic peaks will show some tailing effects and there are some more complex algorithms which model this behaviour in order to achieve greater accuracy.

3. Retention time alignment

This alignment step aims to remove variation introduced by differences in retention time between runs, which could be caused by slight temperature or pressure differences within the LC instrument. The alignment step brings those features with the same m/z ratio to a common retention time. This can be done using either signal or identity based methods[68, 69].

4. Collection of peptides across runs

This refers to the assignment of peptides detected in different runs to the same feature. Different approaches have been applied to this task, e.g. Progenesis LC-MS prepares a consensus map or aggregate spectrum which combines all runs, while OpenMS collects together all post-alignment features that occur within a given tolerance window (e.g. 50s) of a selected reference feature[69]. The advantage of an aggregate spectrum is that it can be saved into a single file that will be searched against a protein sequence .fasta file in the usual manner. As well as lower time and cost requirements, this label-free method is advantageous because the presence of an identified protein in multiple replicates can add confidence to the identification of that protein, and in addition if a feature is confidently identified in one condition its presence can be inferred in the other conditions (even if MS-MS spectra have only been collected from one condition). This inference of peptide identity allows differential expression analysis to be performed on the data even when peptides are at very low abundance in some conditions and therefore have not been selected for fragmentation analysis in those runs.

5. Peptide identification mapping

At this stage the identified features are tied to theoretical peptides generated as constituents of those proteins which may be present in the sample, via the accurate mass values assigned to the features (see pages 21-23).

6. Peptide to protein quantitation inference

For label-free methods protein quantitation is usually inferred from the summation of the intensity values assigned to the constituent peptides of that protein, though a mean value may be used. For SILAC (see pages 26-27 for a detailed discussion of the SILAC method) a median ratio is normally used. When peptides cannot be resolved to a single protein it is usual to assign and quantify a protein group instead, to avoid discarding data.

7. Intensity normalisation

A normalisation step is necessary as the total ion count will vary from run to run, and therefore the signal intensities reported from each run (being a proportion of the total ion count) will also vary. Normalisation steps aim to minimise this variation so that the separate runs are comparable, using either internal standards or statistical methods. One possible statistical method (which is used later in this thesis – see page 36) is median absolute deviation (MAD) normalisation[70]. The intensity normalisation step has the potential to greatly affect the intensity values calculated for peptides, and hence proteins, and therefore optimisation of this step is extremely important. Since the work presented in this thesis was conducted, there have been several studies which have looked at optimising the normalisation step[71-73].

1.9.2 – Available Quantitation Software

As previously mentioned there are many software packages available to post process and analyse proteomic LC-MS-MS data, including vendor software (e.g. ProteinPilot™ from ABI, ProteinScape™ from Bruker, BioWorks™ and Sieve™ from Thermo, and the ProteinLynx Global Server™ from Waters), commercial software (e.g. Progenesis LC-MS) and software which is open source (e.g. MaxQuant)[74-82]. Though the open source software is freely available, it is generally also specific to a type of instrument platform, a particular experimental procedure[15], or to a specific combination of both. The open source software provision is important for those analysts who do not have access to any vendor software, or

who wish to use in house or third party search engines. For those who wish use the same software across multiple instrument platforms it is important that there are software packages which are not instrument or input format specific, and packages such as Progenesis LC-MS are becoming more universal as they work to provide this. One issue arising from the large range of software packages available is the lack of standardised data formats, meaning that not all software packages can be used to analyse data from all platforms and therefore this must be considered when designing an experimental pipeline. However conversion scripts/software is emerging to allow simpler transition between data formats, allowing more freedom for analysts to choose a preferred search engine and processing software with which to analyse their data and making comparison between results from different platforms possible[15] and indeed practicable.

In addition to post-processing software the high complexity of samples where the ground truth is unknown means that benchmarking studies using sufficiently complex and truly representative standard datasets are increasingly important in order to add confidence to the methods used and hence to the biological conclusions made from the experimental results. As the creation of such standard datasets is extremely non-trivial, it will be highly beneficial to the field if these datasets once created are made freely available from online repositories, such as the NCCR Yeast Resource Centre Public Data Repository[4], PRIDE[5], the Global Proteome Machine Database (GMPdb)[6], PeptideAtlas[7], MassIVE (UCSD)[8], or Tranche[9], as the availability of good datasets will allow a greater number of meaningful benchmarking studies to be conducted by bioinformaticians and experimentalists who may not have the expertise, time or equipment required to design and prepare the complex standard samples themselves. This ready availability of proteomics datasets in standard open-source formats is the aim of the ProteomeXchange (PX) consortium[10], which currently includes PRIDE[5], PeptideAtlas[7] and MassIVE (UCSD)[8]. Another consideration of great importance is the need for long term stability of proteome repositories to avoid the loss of datasets, for example many researchers deposited raw data into Tranche, which then lost financial support resulting in many data sets becoming corrupted or being lost altogether. Therefore an additional goal of the PX consortium is to ensure that such a situation cannot happen again, so that researchers can be confident that datasets will be robustly stored for the foreseeable future.

2 - Quantitative Proteomics Software used in Combination to Reduce False Discovery Rate

2.1 - Aims

This software comparison study aims to determine how much variability is introduced into final results through the use of different software pipelines to analyse the same data set, in terms of the number of proteins reported as differentially expressed, which proteins are identified/quantified, and the false discovery rate for each pipeline. As well as comparing and contrasting results from individual pipelines, the pipelines used allow comparison of intensity-based label free pipelines versus spectral counting pipelines, and in addition this study examines the potential of creating a consensus method which takes data from all pipelines to produce a more robust result than any single pipeline alone.

2.2 - Introduction

The move from identification to quantitation proteomics has become increasingly widespread as mass spectrometry instrumentation has become more capable of delivering accurate mass information. This is particularly important as peptides often overlap in LC-MS, and low resolution instruments cannot resolve the data to individual peptides, leading to poor quality quantitation. As a result of this increased use of proteomic pipelines and desire for quantitative results, the associated software techniques have also become more sophisticated – delivering more accurate data for large numbers of proteins. This study compares and combines the results obtained from the parallel analysis of a single dataset using four separate software pipelines for protein identification and quantification. This is a question of relevance to the field as little work has been reported on this type of comparison, yet the assumption of comparability for results obtained by different groups using different pipelines is implicit in the comparison of the biological conclusions presented by analysts. It is intended that this study will provide information for analysts as they design their experiments, and give an insight into the challenges that are encountered when deciding on the instrument platform and post-processing methods which are appropriate to their aims.

The quantitation methods considered in this study are; emPAI[83] values generated within and reported by Mascot (Matrix Science), spectral count and intensity values reported from Progenesis LCMS (Non-Linear Dynamics), absolute protein expression values from the APEX quantitative proteomics tool[64, 84] (following processing using the Trans-Proteomic Pipeline[85] web interface), and intensity values from MaxQuant[86] (open source). Very little work has been done where this type of comparison has been made between different spectral counting and intensity based software pipelines[87], with discussions of this type largely restricted to the presentation or comparison of novel in-house pipelines with those which are more widely used (both intensity based[88-92] and spectral counting based[93, 94] workflows) rather than an assessment of the software pipelines that are available to the standard user. Despite this, an investigation of the reliability of those software pipelines used for the analysis of proteomic data is an extremely important research question, and increasingly so as these methods become more widely used, and the biological questions asked by these analyses becomes more complex.

Each workflow pipeline was used with all parameters set at the default settings (i.e. those settings recommended in the software guidelines), apart from those parameters that were experiment or instrument specific. These inherent parameters (eg mass tolerance, allowed modifications, the number of missed cleavages allowed and the .fasta database searched for identifications) were set identically across all software in order that the results were comparable. The results from all pipelines were then compared to evaluate the conformity, reliability and reproducibility of the data obtained from and between different processing and experimental workflows when using a standard dataset for all analyses. In order to complete the comparison an assessment was made of any gains obtained by combining the results of multiple pipelines, in terms of improving the ratio of sensitivity to false positives for differentially expressed proteins (analogous to the gains observed when using multiple search engines for protein identification[3]).

2.2.1 - Datasets studied

2.2.1.1 - ABRF iPRG2009

This dataset was created for an iPRG (Proteome Informatics Research Group) study conducted by the ABRF (Association of Biomolecular Resource Facilities) in 2009, the primary goal of which was to "Evaluate protein differentiation tools for MSMS", with the additional secondary goal to produce a "benchmark reference" of 'true' proteins which could then be used as a basis for software development"[95]. Thus the study was not intended to be quantitative, and therefore the dataset was not constructed with complete quantitative analysis in mind. Thirty different labs took part in the 2009 study, and between them they returned 37 submissions of results. Each submission detailed which proteins had been assigned as significantly changed between the two conditions studied, plus a score showing the confidence placed on each reported protein assignment.

In order to create the "standard known dataset" to be used in the study an *E.coli* digest sample was split and run in parallel using 1D SDS-PAGE on two parallel lanes of the same gel. Sections of these gels were then excised to create artificial up and down regulation between the two conditions, and these sample conditions were then arbitrarily labelled as 'Red' and 'Yellow'. All gel sections were reduced and alkylated, followed by an in-gel digestion with trypsin. The resulting final samples were then run in 5 parallel runs on a low resolution Thermo Finnigan LTQ mass spectrometer and the output MS and MS-MS spectra distributed via the Tranche tool[9] in .raw, .mzXML, .mzML, .mgf and .DTA data formats. The tranche tool was an interface to search and access the resources hosted on the Tranche server, and the data from this study was presented on the server as a downloadable folder containing all the .raw data and a .fasta database file, which was compiled by concatenating .fasta files containing BSA (Bovine Serum Albumin), common contaminants and reverse sequences to the SwissProt *E.coli* database .fasta file (as accessed on the 3/6/2008). In order to generate an unbiased Answer Key (to be used to evaluate the results returned from the groups and laboratories taking part in the ABRF study) the ABRF group performed three parallel runs of the excised gel sections (which contain the proteins that will be "missing" from the 'Yellow' and 'Red' samples), with these artificial conditions labelled as 'Blue' and 'Green'. The results were then analysed in parallel by 10 ABRF research group members. 92% of the Answer Key is made up of proteins which were agreed upon by at least 5 group members, with the

remaining 8% made up from proteins which were agreed upon by a minimum of 2 group members and fell within the expected mass range.

2.2.1.2 – CPTAC Study 6

This standard dataset was created for CPTAC (Clinical Proteomic Tumour Analysis Consortium) study 6[96] (published October 2009), which was a benchmarking study seeking to provide a basis which would allow groups to self-assess the performance of their instrumentation and post-processing procedures on a standard dataset where the ground truth is known. The dataset presented consisted of five samples which were the result of splitting a single yeast culture, with the intention to obtain a truly homologous sample as the basis for the prepared standard samples, so that the only difference between the final samples would be the level of spike-in proteins added. Creation of this single culture for use in the study was outsourced by the study group, and was conducted as follows; "*S. cerevisiae* strain BY4741 (MATa, leu2_0, met15_0, ura3_0, his3_1) was grown in a 10-liter batch of rich (yeast extract peptone dextrose) medium at 30 °C in a fermentor to an A_{600} of 0.93"[96]. The sample was then passed back to the study group as a lyophilised yeast lysate, which was reconstituted, reduced and alkylated prior to trypsin digestion. Following digestion the sample (dried digest re-suspended to a concentration corresponding to ~60ng/µl yeast protein before digestion in 0.1% aqueous formic acid) was split into five smaller samples, each of which was spiked with different levels of a standard protein mix (which had also been pre-digested with trypsin). The mix used was the Universal Proteomics Standard 1 (UPS1) standard mix from Sigma-Aldrich[97] (an equimolar protein mix with all proteins present at 5pmols), which was spiked in at a level corresponding to concentrations of 0.25, 0.74, 2.2, 6.7 and 20fmol/µl total protein pre-digestion. This methodology is summarised below in Figure 13.

The resulting five spiked samples were designated A-E (with A containing the lowest and E the highest concentration of spike-in proteins) and sent out to a number of expert labs where the samples were run in triplicate on various ion-trap based MS platforms (LTQ, LTQ-VELOS, LTQ-VELOS Orbitrap and LTQ-Orbitrap) following a standard operating procedure (SOP) which was distributed with the samples. The full LC-MS parameters contained in the SOP can be found in the supplementary material for the CPTAC Study 6[96]. The results files generated from the various MS runs in the external laboratories were then returned to the study group for analysis. The resulting paper[96] reported that fold changes could be

quantified accurately for the UPS proteins, but that these proteins could not be detected in the A sample (hence only samples B-E are used in this analysis), and that the yeast proteins were not seen to be significantly changing between the 5 samples.

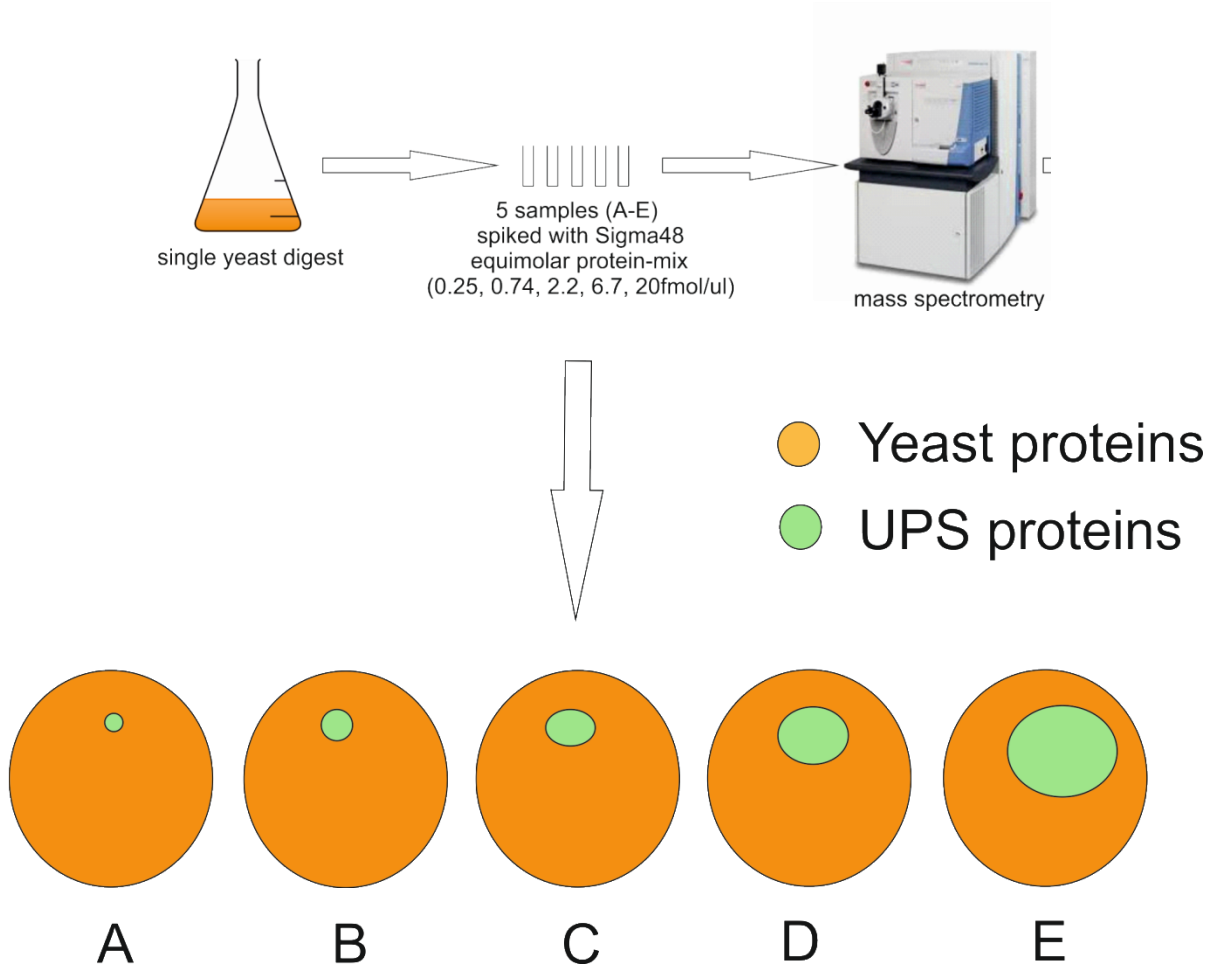

Figure 13: CPTAC samples were created by splitting a single yeast digest into five and spiking each with protein mix (0.25, 0.74, 2.2, 6.7 and 20fmol/ul respectively for samples A-E).

The study group made the .raw spectra from the analyses by all groups available as Thermo .raw files via the Tranche tool[98]. However, for my study only one of these datasets was used – this was the dataset resulting from the analysis using a Thermo LTQ Orbitrap (the dataset designated "LTQ-OrbitrapO@65", in which the 65 represents the laboratory which completed the analysis) - as this was seen to yield the highest number of peptide spectrum matches for yeast proteins.

2.3 - Methods

For ease of reference, a workflow diagram has been prepared which summarises the work that has been carried out in the course of this study and this is presented below in Figure 14.

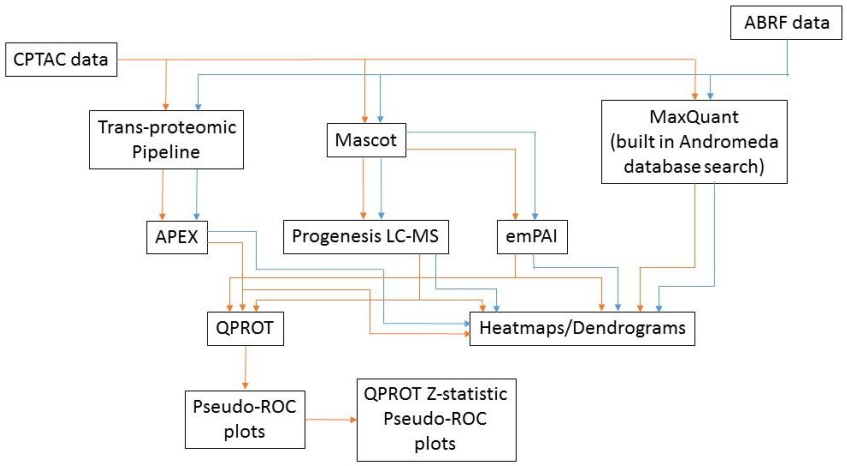

Figure 14: Workflow diagram showing the different stages of analysis for the ABRF and CPTAC data

2.3.1 - FASTA files used

2.3.1.1 – ABRF iPRG2009 data

The official .fasta file distributed via Tranche was used for the analysis of this dataset, with the same .fasta file being input to all software analysis pipelines.

2.3.1.2 - CPTAC Study 6 data

The .fasta file for this dataset was created by concatenating two FASTA files – the current yeast open reading frames transcript downloaded from the Saccharomyces Genome Database (downloaded early 2010) and an official database containing the sequences of the 48 UPS1 proteins.

2.3.2 - Software packages used

2.3.2.1 - emPAI calculation

.fasta files were searched using an in-house Mascot server with the following settings: allow 1 missed cleavage, fixed modifications: Carbamidomethyl (C), variable modifications: Oxidation (M)*, Peptide Tolerance: +/-10ppm, MS-MS tol. +/-0.6Da (the fixed and variable modifications are set as a function of the experimental preparation). The emPAI values used were those exported directly from Mascot. The emPAI value is a modification of simple spectral counting, the basic assumption of which is that if a protein is abundantly present all the constituent peptides of that protein will be observed during mass spectrometry. In order to determine the emPAI value, the PAI (protein abundance index) is first calculated by dividing the "observed" value by the "observable" value for each protein in a sample, where the observed value is the number of non-redundant peptides identified (allowing multiple charge states for each peptide) as experimentally associated with that protein and the observable value is the number of peptides per protein predicted from a theoretical digest which are within the defined scan range of the instrument (in the Mascot implementation this range is not simply that set by the user, but is dependent on the highest and lowest m/z values detected experimentally i.e. any theoretical peptides above or below these will not be included in the observable value used in the calculation). The emPAI value is then calculated from the PAI as

$$emPAI = 10^{\frac{observed}{observable}} - 1.$$

*The fixed modifications set are a function of the experimental procedure conducted in the sample preparation, namely reduction and alkylation. The variable modification of oxidation on methionine is set as this is a common post-translational modification and it is desired to record any proteins whether they have this modification or not.

2.3.2.2 – APEX Quantitative Proteomics Tool

The APEX tool is also a spectral counting based method of scoring protein abundance similar to emPAI. In this quantitative method peptides are considered more or less likely to be observed by mass spectrometry based on their physico-chemical properties and their resulting predicted behaviour within the mass spectrometer. The input for the APEX tool are .protXML files output from the TransProteomic Pipeline (TPP), therefore the files used in this study were pre-processed within the TPP web interface prior to analysis with the APEX tool.

In the TPP web interface the files were converted to .mzXML and the in-house Mascot server used for the searches (with the same parameters as above). The Mascot result files were then converted to .pepXML files and these were run through the next stage of the TPP, PeptideProphet, with the option selected to run ProteinProphet afterwards (using pI, hydrophobicity and RT information, and including results with pepProphet probability <0.05 and minimum peptide length = 7). The final .protXML files were then analysed using the APEX tool. One .protXML file was selected to be used for training (using the default random forest machine learning algorithm[99], which is considered to be the most appropriate training mechanism for APEX), with reference to the relevant .fasta database file, using the default settings (Consecutive Misses: 0, Use Minimum Peptide Length: 3, Use Minimum Peptide Mass: 250, Use Maximum Peptide Mass 7500, P(i)=1.00, APEX normalisation factor = 0.00) and with all the available peptide properties used to assess the ionisation efficiency of theoretical peptides. This creates an .ARFF file of training data, which is then used with the relevant FASTA file to create an .oi file. This .oi file stores the signal intensity and physico-chemical properties of the peptides found in the training file, and is used as the reference for the analysis of other sample files. The APEX scores are internally calculated using the following formula.

ni = observed spectral counts

pi = protein identification probability

Oi = predicted count for one molecule of protein (sum of peptide detection probabilities)

$$APEX = \frac{p_i\left(\frac{n_i}{O_i}\right)}{\sum_{k=1}^{k=N} p_k\left(\frac{n_k}{O_k}\right)}$$

The APEX score, p_i, n_i and o_i are reported in the APEX output .csv file, which can easily be copied into Excel for post-processing.

2.3.2.3 - Progenesis LC-MS (Nonlinear Dynamics Ltd)

Thermo .raw files were loaded into the Progenesis LC-MS software (version 3.0.3840.17781) and all samples aligned to the E1 sample using automatic alignment. The resulting aggregate spectrum was then filtered to include +1, +2 and +3 charge state features only. In the experimental design tab within Progenesis LC-MS the samples were grouped according to the experimental conditions (B-E), and features without MS2 spectra were deleted from the analysis (as MS2 information is necessary for peptide/protein identification). An .mgf file representing the post-alignment aggregate spectrum was exported and searched using

Mascot (with all settings as detailed above). The resulting .xml file was then re-imported to Progenesis LC-MS to assign peptides to features. Two different thresholds were used while importing these identifications, reflecting changes to the guidelines offered by the software manufacturers. First, a threshold of Mascot score≥40, hits>2 was applied, as recommended in a tutorial from NonLinear Dynamics. The analysis was also completed using a Mascot score threshold ≥ 17, as this is the value given by Mascot above which a given protein is confidently identified within this data set. The "Protein Raw Abundance" data type was then used for post-processing.

2.3.2.4 - MaxQuant

The Thermo .raw files were analysed within the Quant.exe (version 1.1.1.36) program of the MaxQuant software, using the appropriate .fasta files and the following parameters: Parent tolerance +/- 10ppm, Variable Modifications: Oxidation (M), Fixed Modifications: Carbamidomethyl (C), CID MS-MS tolerance: 0.6Da (using 6 top peaks per 100Da, Higher charge and allowing water and ammonia losses), Time correlation: 0.7s, Peaks per 100Da: 20, SIL weight: 4, ISO weight: 2, Low mass cutoff: 0, Peptide/Protein FDR: 0.01, Using unmodified, Oxidation(M) and Carbamidomethyl(C) peptides for quantification and "Using Unique Peptides", Keep low-scoring versions of identified peptides, Match between runs: Time window 2 mins. The resultant .txt files were saved in the .csv format for post-processing in Excel. The "LFQ Intensity" data type was used for post-processing as it has been observed to be a more accurate measurement of the true ratios present[100].

2.3.3 - Median Absolute Deviation Normalisation (Intensity based methods)

All calculations involved in this normalisation procedure were conducted manually in Microsoft Excel, in a manner based upon the method used for Progenesis LC-MS as described by Non-Linear Dynamics[101]. In brief, log ratios were calculated for each replicate relative to the E1 sample (chosen since condition E contains the highest concentration of spiked UPS proteins and it is therefore assumed that all UPS proteins that can be detected should be identified in this replicate). The median log ratio for each replicate (ie for each column in the Excel spreadsheet) was calculated. Next, deviations from the median log ratio in each replicate were calculated for every protein and these values used to calculate the median absolute deviation (MAD). This is the median of the calculated deviations from the median log ratio, and is calculated for each replicate separately (i.e. there is a MAD value for each

column). Log ratios which were not within one MAD of the median value for a given replicate were removed as outliers and the remaining values used to calculate a scaling factor for each replicate, which is calculated as the inverse of the mean log ratio per column (calculated following outlier removal).

2.3.5 - Total Count Normalisation (Spectral Count based methods)

Median absolute deviation normalisation is not appropriate for spectral count derived abundance values, since many low abundance proteins produce identical emPAI or APEX values in different conditions (giving a log ratio of 0), and this leads to skewed normalisation factors when MAD normalisation is used. Therefore for spectral count based methods it was decided to use the simpler Total Count Normalisation method, with this calculation also conducted manually in Microsoft Excel. The total summed protein abundance value was calculated for each replicate (ie each column in the spreadsheet) and all replicates normalised against the replicate which gave the highest total abundance value (for the APEX data this was C3, and for the emPAI data E1). The scaling factor for normalisation was calculated as $S_f = \frac{total_h}{total_i}$ where $total_h$ is the highest column total value and $total_i$ is the total abundance for a given column. Scale factors were then used to calculate the normalised abundance values for each column as above.

2.3.6 - Thresholds applied for inclusion of data for analysis

Prior to the application of a statistical test to identify differentially expressed proteins, it is important to remove low quality data from the analysis to avoid the skewing of the results by incorrectly assigned features. Thresholds were set appropriately for each type of data and post-processing method.

2.3.6.1 – Heatmap Data Thresholds

Less stringent thresholds were applied to the protein lists to allow more proteins to be included in the generation of heatmaps, these thresholds required only that each protein had been quantified by at least two software packages and in both the conditions being considered. This was considered reasonable as the generation of heatmaps was intended as a means to investigate the agreement between packages in terms of which proteins were

identified as differentially expressed, rather than an a measurement of quantitation accuracy.

2.3.6.2 – Thresholds for inclusion in pseudo-ROC plots

For the intensity-based data the threshold was set to require that the protein had been quantified using 2 or more unique peptides (the unique peptides metric is reported by both MaxQuant and Progenesis). This threshold is considered sensible as there is risk in using a single peptide to represent protein-level quantitation because any errors in feature detection, map alignment or identification assignment, etc, will be directly reflected in the results. Through the use of more than one peptide, such inaccuracies are mitigated.

However there is no information about unique peptides included in the results data from the spectral counting methods and therefore the threshold of two unique peptides cannot be used in this case. Thus for the spectral count data, since an emPAI or APEX abundance value calculated from a small number of peptide counts would be low quality for the reasons given above, the data matrix was ranked by percentile and a threshold set which excluded the lower 50% of data. This threshold is somewhat arbitrary but was set to avoid protein ratios derived from small numbers of spectral counts being included in the final results. It was observed that this was the lowest proportion of the data which could be removed in order to exclude all of those proteins which were detected in only one of the fifteen condition replicates. The rationale for this threshold is that those proteins that are present at a significant abundance in the original sample should produce peptides which are present at detectable abundance in at least two or more conditions and/or replicates.

Both these threshold protocols were considered sensible for the two data types and though they are not strictly directly comparable it was considered that a “real” comparison of data with appropriate thresholds applied was more appropriate to the questions asked by this study (aiming as it does to look at the comparability of data obtained from different analysis pipelines used in different labs, where the data from each pipeline will have been processed optimally for each individual study).

2.3.7 - Tests for differential expression

Three different "scores" were used to identify proteins as differentially expressed; p-values, absolute fold change and QPROT-FDR. The p-values (obtained from two-tailed Student's t-tests conducted on the data in Excel) and absolute fold changes were calculated from the protein abundance values reported for each replicate. Each score was calculated as a pairwise comparison between the conditions (E/B, E/C and E/D), for each of the four software packages. In addition, the data was processed through the QPROT tool [1, 2], which has been designed specifically to calculate accurate statistics, such as FDR or Z-statistic values, for differentially expressed proteins in label-free spectral counting and intensity data. The QPROT tool[1] is a Linux based tool, which is a further development of the QSPEC tool[102, 103] that uses Bayesian statistics to model the likelihood of true differential expression in label-free data. The QPROT-FDR from the QPROT output was used as the chosen score to order proteins for the pseudo-ROC plot comparison using this "score".

2.3.8 - Calculating FDR and sensitivity for pseudo-ROC plot generation

To calculate an FDR each table of results was ordered by the chosen "score" (pvalue, fold change or QPROT-FDR) of interest, and the FDR was calculated from: FDR = FP / FP + TP, where TP is the count of UPS proteins with the given score or better and FP is the count of all non-UPS proteins with the given score or better. Sensitivity was calculated as sensitivity = TP / 48 (as there were 48 UPS proteins in the UPS1 protein mix spiked into the CPTAC samples) and q-values were calculated by finding the lowest FDR that could achieve the same sensitivity or better for the score of interest. Therefore, for every row of data there is a q-value and an associated sensitivity, and these were plotted as "pseudo-ROC" plots (shown below – these are referred to as pseudo-ROC because a ROC (Receiver Operating Characteristic) plot typically displays true positive rate versus false positive rate so these plots are a variation on the norm). These plots provide a measure of test accuracy, with a plot that follows the left and top border of the space closely indicating high accuracy, and show the effect on sensitivity when different q-value thresholds are applied.

2.3.9 - Consensus across different packages

A straightforward method was sought to compare the results from the different software packages in order to assess whether there was a consensus in the data. A consensus between packages would make it possible to test the hypothesis that the consensus results will give a better measure of differential expression than any single package alone. This hypothesis follows the logic behind improving sensitivity in peptide/protein identification through the use of multiple search engines[3, 104], and extends it to quantitative analyses. The methods chosen to investigate this hypothesis were heatmaps of the data, and pseudo-ROC plots generated using the QPROT Z-statistic. Both the heatmaps and the the pseudo-ROC plots were generated using the R statistical package.

2.3.9.1 - Heatmap generation

Log_{10} ratios were calculated between the E/B, E/C and E/D conditions and the maps constructed using the heat map.2 package in Bioconductor, which is a supplementary statistical package for R. Parameters were set to use the hierarchical clustering function (hclust) as the distance/linkage algorithm to generate the heatmaps. Null and infinity values cause problems for calculating dendrograms, with a null value occuring when the protein is undetected in both conditions being compared, and an infinity value being reported when the protein is detected in only one condition. In both cases these were set to zero, and those rows containing only zeros were excluded as no conclusions could be drawn from the data for these proteins. The scale was fixed so that a zero mean gives black, with up-regulation giving green and down-regulation giving red areas in the heatmap. No optimisation was performed on the parameters used within R to create the dendrograms and no additional thresholds were applied to exclude unreliable data, since we wished to test how well this method would perform in the absence of optimisation which would be difficult for the standard user to perform in a real situation where the ground truth would obviously not be known. The generated heatmaps can be used both as a method to compare the results between pipelines in terms of presence and magnitude of up/down regulation, but can also represent a consensus method giving greater confidence to those proteins that are assigned as similarly differentially expressed by multiple pipelines.

2.3.9.2 - QPROT Z-statistic

This method of comparison used the more stringent thresholds detailed above in 2.3.6.2. The Z-statistic itself is a measure of the distance (in standard deviations) from the mean in normally distributed data. The QPROT tool analyses the global distribution of the data from each software package and thus multiple Z-statistic values are broadly comparable across the different packages. If a given protein had not been measured by a particular package, the Z-statistic for that package was scored as zero to denote that there was no evidence for differential expression.

As the Z-statistic values are comparable across packages, the absolute value of the mean Z-statistic was taken in order to combine the results from the four software packages, bringing all data onto the same scale as the absolute value is an approximation of the global strength of differential expression (either up or down) as calculated by each of the different packages. These values were also plotted on the pseudo-ROC plots to assess this combination method as a way to improve sensitivity with respect to the plots from each individual package.

2.4 - Results

Both datasets were analysed using four different software pipelines in order to test how well each software package could detect the known underlying ratios present within the data, to assess the agreement between pipelines, and to estimate the FDR associated with different methods of determining which proteins were differentially expressed between conditions. Heathmaps and pseudo-ROC plots were used to compare the results obtained from the different software packages and assess any agreement on the direction and magnitude of differential expression between conditions. An attempt was also made to combine the results from all software packages to determine if a consensus method would improve the sensitivity while reducing the number of associated false positives.

2.4.1 – ABRF data

We analysed the ABRF data set (yellow/red comparison) to see how well the different software packages agreed. First we looked at the proteins that were scored as significantly changing ($p<0.05$) by each of the software pipelines and were contained within the answer key. Across the different pipelines, over 500 proteins were scored as significantly changing, with only 22 proteins being common across all pipelines, thus showing the general agreement between pipelines was poor. The same analysis was repeated for the fold change data, and this gave a slight increase in the number of common proteins, but still showed no general agreement in the overall picture. A heat map of the five pipelines (at this stage we were also considering the spectral count values reported from Progenesis, however this metric was dropped from further analysis) for the log ratios of yellow over red is shown in Figure 15. A clear trend emerges from the heat map that there are two clusters of proteins, one showing clear up-regulation (green – 14 proteins) and one showing down-regulation (red – 20 proteins). As the published "answer key" does not include differential expression information the ground truth for the ABRF dataset is unknown. However, the results from the heat map are likely to be relatively robust with a lower FDR with respect to using a single software package, since each software package is likely to introduce a unique set of errors, and any bias introduced by using a single package should be removed by considering the consensus results. On considering the heatmap it can be seen that there are a number of proteins where different packages give opposite results – and therefore there is little confidence in the identification/quantitation of these proteins. It is possible that one package is "correct" and the other "incorrect" which could be deduced by careful manual analysis, but creating heatmaps of the results as a first pass allows the lab scientist to focus first on those proteins where there is a high confidence that differential expression is truly present.

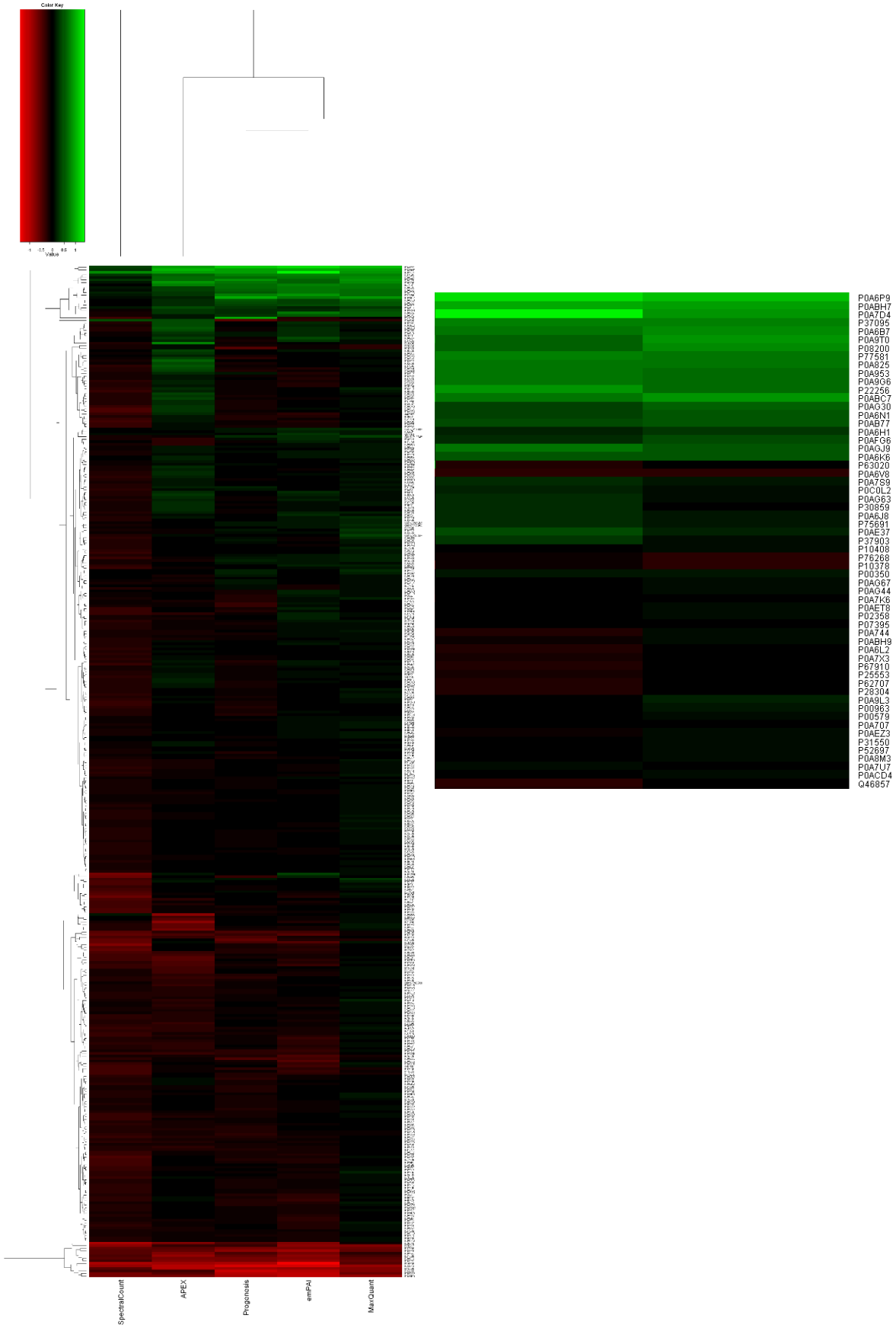

Figure 15: Heatmap showing differential expression results for the ABRF dataset, including zoomed section showing protein accessions (Columns are “Spectral Count” (from Progenesis), “APEX”, “Progenesis”, “emPAI” and “MaxQuant”)

Following the analysis of those heatmaps generated from the ABRF dataset it was concluded that the setup was too artificial, and as the true real answer is unknown it is impossible to draw any conclusions about software quality based on the analysis of this dataset with any confidence. As such further stages of analysis were not completed for this dataset.

2.4.2 - CPTAC ratio calculation

Ratios were calculated for the spiked in UPS proteins in the CPTAC dataset to compare the experimental ratios with the known ratios between the spike-in amounts for the E/B (27 fold change), E/C (9 fold change) and E/D (3 fold change) comparisons, using the four software pipelines: emPAI, APEX, Progenesis LC-MS and MaxQuant. It was observed (see Figure 16) that the two intensity based pipelines gave greater accuracy as measured by the median protein ratios. Progenesis LC-MS gives the closest median ratio values to those expected from this dataset, 26.8, 10.3 and 4.1 for raw values and 20.8, 8.5 and 3.0 after normalisation with the expected values being 27,9 and 3. MaxQuant also produced relatively accurate ratios in the raw data, 34.2, 8.2 and 2.7, but this time normalisation appears to increase the ratios, to 49.6, 11.7 and 3.2. Both intensity-based packages produce similar values for the interquartile range of the data. This indicates that the relative spike-in has been done correctly across the different samples; however there are inherent differences in the abundances of proteins present in the yeast background. Differences in the yeast background can also be observed in box plots of the yeast log ratios before and after normalisation for each package.

The two spectral counting methods underestimate the ratios in all three comparisons. Given that Progenesis LC-MS and MaxQuant (raw abundance) are able to broadly measure the correct ratios without normalisation steps, we can assume that there is a feature of the underlying spectral count method that is inaccurate, at least for the analysis of this dataset. However, as neither emPAI nor APEX were designed for accurate calculation of ratios across multiple conditions, this result is not entirely unexpected. Both emPAI and APEX involve putting the spectral count values onto an exponential scale, and it may be that this relationship does not hold for this dataset, or for a certain subset of the proteins within. For emPAI, the calculation uses observed (peptides)/observable and therefore to get a high dynamic range you would need to observe the majority of the peptides within a given

protein. For many proteins it is likely that a subset of their constituent peptides will have physicochemical properties that lead to poor ioisation, and thus high emPAI values cannot be seen for those proteins.

It was observed for this dataset (see Figure 17) that while the raw values for UPS proteins are broadly accurate in each MS run, there are differences in the global abundance of yeast proteins (which are expected to be in 1:1 ratios across all conditions).

Applying normalisation to the data creates a more uniform background of relative abundance for the yeast proteins in the sample but it also alters the reported abundance of the UPS proteins, which is likely to lead to artificial differences between replicates. As such, a specialised normalisation scheme was designed to correct independently for global differences in UPS and yeast protein abundance within replicates.

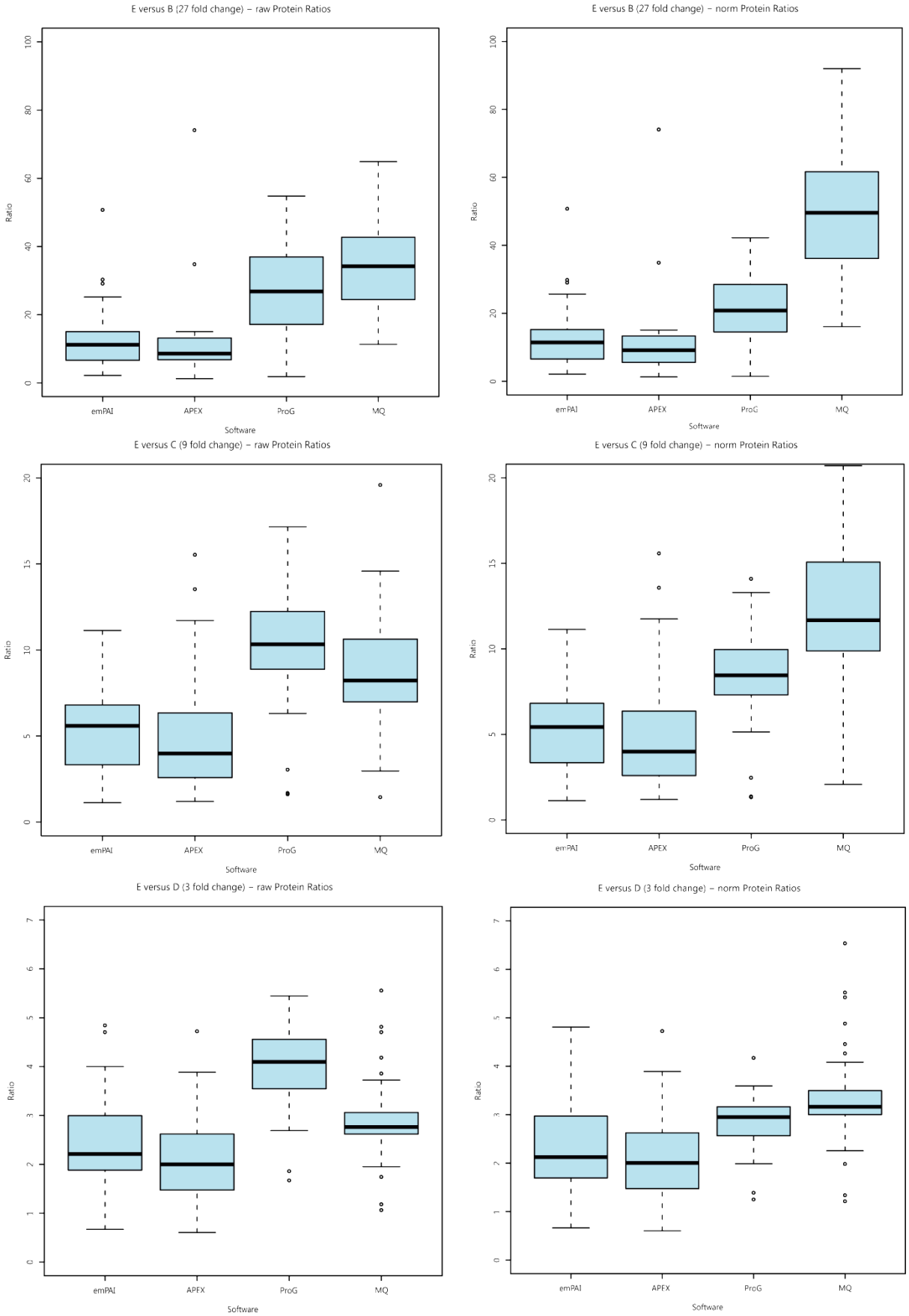

Figure 16: Box plots of log ratios for the UPS1 proteins identified in each comparison by all pipelines, for both raw and normalised data.

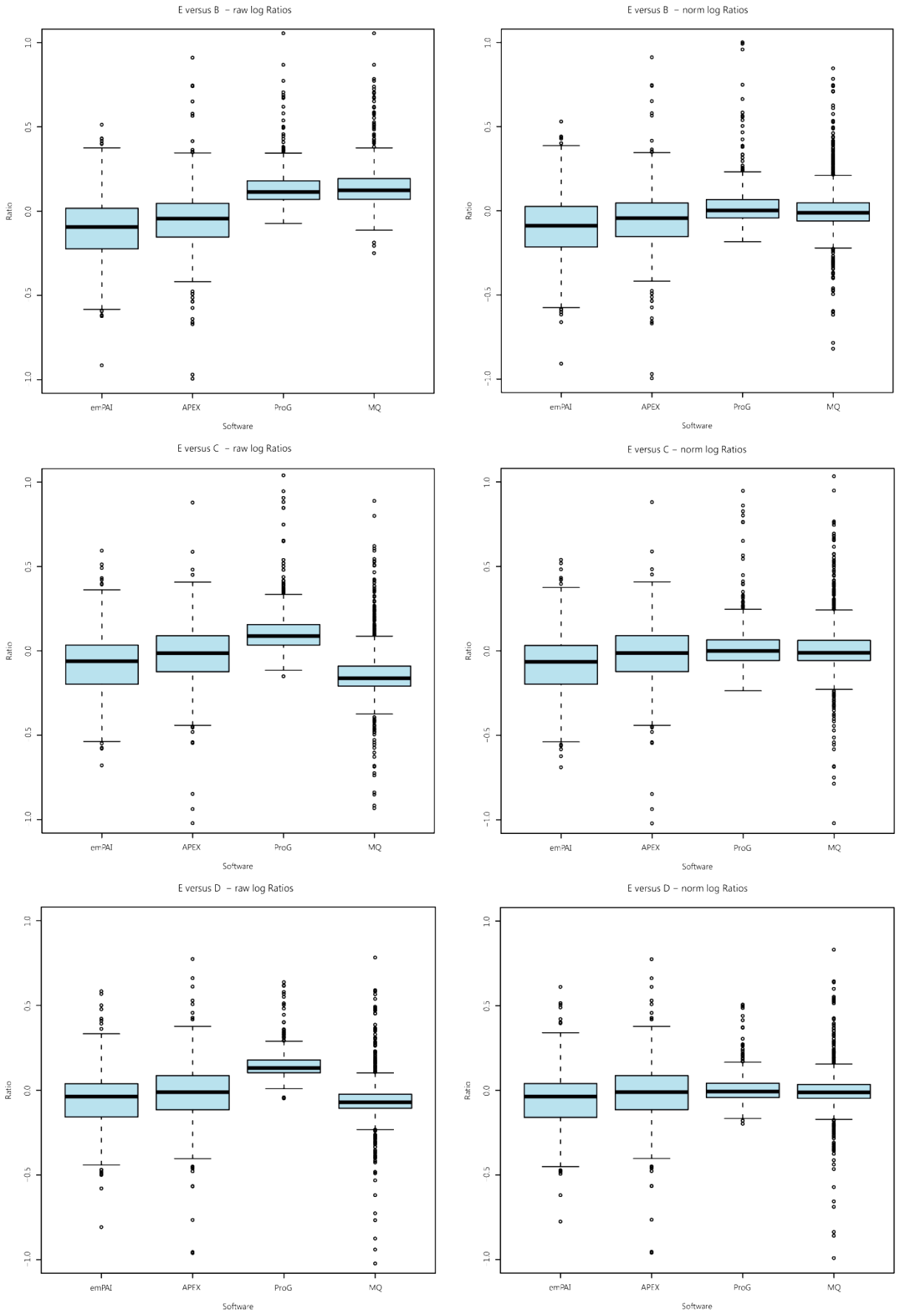

Figure 17: Box plots of log ratios for the yeast background proteins identified in each comparison by all pipelines, for both raw and normalised data.

2.4.3 - Consensus across different tools

Next an assessment was made to evaluate the sensitivity of the different software packages for evaluating differential expression of proteins and the FDR associated with a given sensitivity value, and to study the effect of using different pipelines in a combined approach as a strategy to improve sensitivity while simultaneously lowering the associated FDR.

2.4.3.1 - Measurement of differential expression by Student's t-test

Two-tailed t-tests were performed to calculate p-values from the complete set of CPTAC results for the three comparisons being studied (E/B, E/C and E/D), as analysed using each of the four software packages (two versions of the results obtained from Progenesis LC-MS are included to highlight the effect of changing the identification threshold). The null hypothesis is that a given protein is not changed in abundance, with the alternative hypothesis being that there is differential expression between different conditions.

2.4.3.2 - Pseudo-ROC Plots

A p-value was calculated for every protein and the results table ordered by p-value for each software package and for each comparison. It was possible to estimate an FDR value associated with every p-value, via the count of differentially expressed UPS proteins (true positives) versus differentially expressed yeast proteins (false positives). The FDR values were then converted to q-values and line graphs plotted of the sensitivity (proportion of the 48 proteins scored as differentially expressed for each q-value) to create pseudo-ROC plots. The ROC plots show (see Figure 18) that for all pipelines the use of p-values assigns many yeast proteins as significantly differentially expressed, and therefore a high FDR is necessary to achieve a reasonable sensitivity.

The E/B comparison should be the comparison in which it is easiest for the packages to achieve high sensitivity and low FDR, yet values of ~30-90% FDR are observed in order to achieve 50% sensitivity. For the E/C comparison values of ~40-90% FDR are observed at 50% sensitivity, and for E/D, the most difficult comparison for packages to find differentially expressed proteins in correctly, FDR values of 25% to 80% are observed at 50% sensitivity. In all cases, in order to achieve high sensitivity (say 80%) - FDR values of ~80-90% occur, which would clearly be problematic in any real biological study where the ground truth is unknown. Figure 18 is also annotated with the value of sensitivity and q-value that would

be obtained when a (typical) threshold of $p<0.05$ is set. In all cases the performance is unacceptably bad – generally leading to FDR > 80%. However, the intensity based methods far out-perform the spectral count methods using this statistical test. Within Progenesis LC-MS two import thresholds were used when loading identifications from Mascot, a lower, arbitrary threshold and a higher threshold based on the confidence value given by Mascot. It is seen that the maximum possible sensitivity is decreased when using the higher threshold for loading identifications (see Figure 18, upper three panels – the yellow line represents the lower threshold and the green line represents the higher threshold) - this is an example of the importance of setting this threshold correctly as it shows that setting too high a threshold can mean that true positives are missed. The effect of changing the import thresholds will be considered in more depth below (see Chapter 4, page 98). It is clear that for this study at least, using a t-test to compute p-values is not appropriate. In each study condition, there are only three replicates and it appears that there is too much background variability between replicates to achieve accurate p-values using this method.

A parallel analysis was completed with the protein results ordered by fold change, with the fold change being calculated for every protein and the results table re-ordered by these fold change values for each software package and for each comparison. Again, an FDR was estimated for every fold change value, via the count of differentially expressed UPS proteins (true positives) versus differentially expressed yeast proteins (false positives). The FDR values were then converted to q-values and line graphs plotted of the sensitivity (proportion of the 48 proteins scored as differentially expressed for each q-value) to create pseudo-ROC plots. In the E/B comparison, ordering by fold change leads to an improvement in the sensitivity to FDR relationship for all packages. This is not unexpected since the UPS proteins on average are present in a 27:1 ratio, compared to the yeast proteins that are expected to be present at a 1:1 ratio. In the E/C and E/D comparisons, the performance of all packages gets progressively worse as the differentially expressed UPS proteins become harder to distinguish from the yeast background. However, Progenesis LC-MS and MaxQuant appear roughly comparable when the lower identification threshold (i.e. fewer weak identifications are allowed) is used to import identification results into Progenesis LC-MS.

Annotations have been added to the figures to show the FDR and sensitivity values that would result if a “2 fold threshold” were applied to each pipeline, as this is a threshold that has commonly been used in published studies. There is clearly great variability in all

comparisons and in no pipeline does the use of a 2 fold threshold accurately control FDR and achieve a reasonable level of sensitivity. As such, in order to achieve good sensitivity and low FDR for this data set, an optimised fold-change threshold would need to be applied individually for each comparison. Clearly in a real data set, where the true ratios of the proteins are unknown, the level of these thresholds would be challenging to assess.

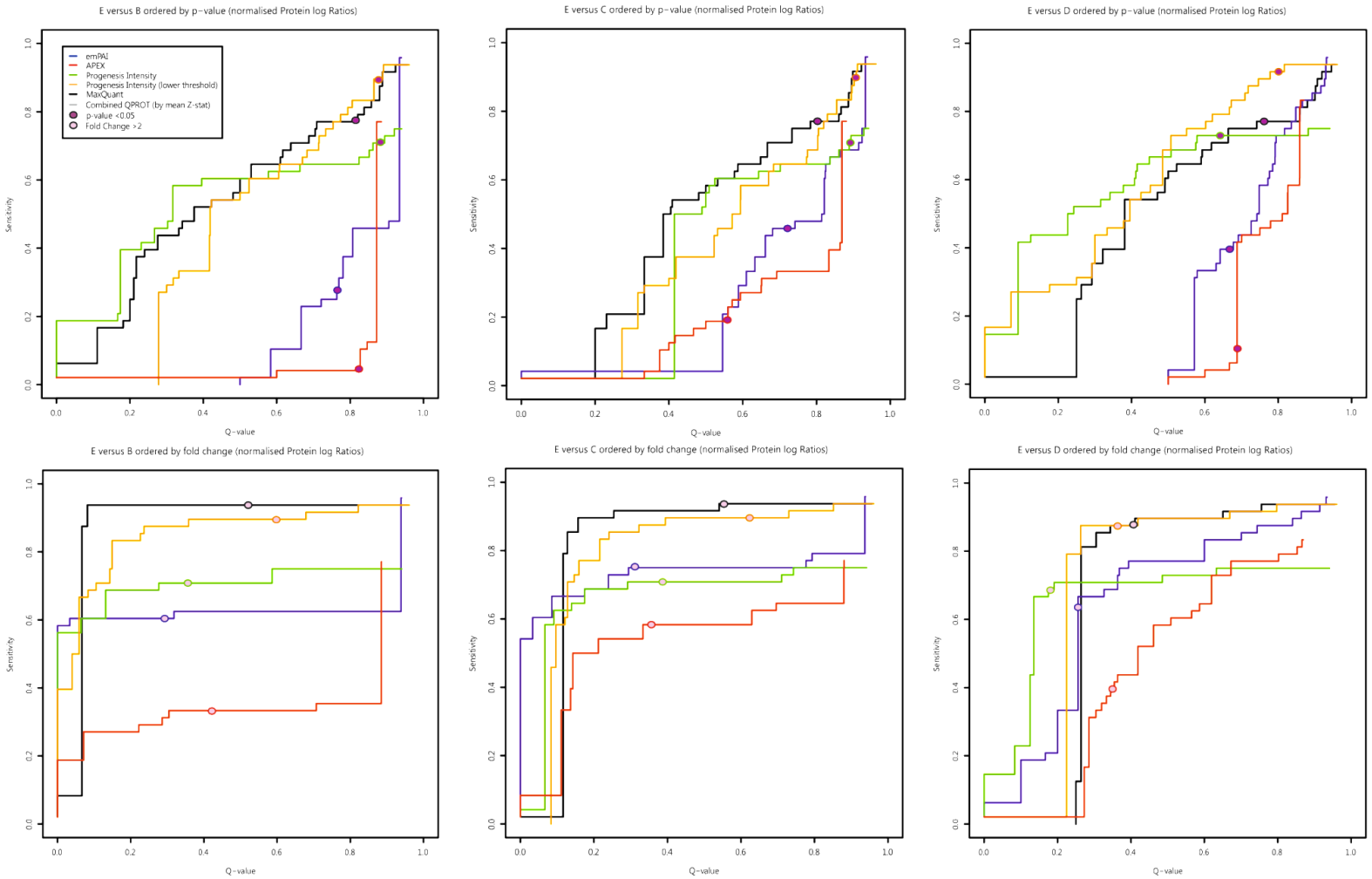

Figure 18: Pseudo-ROC plots for all comparisons ordered by p-value and fold change, annotated with p-value = 0.05 and fold change = 2. Key: Blue line = emPAI, Red line = APEX, Green line = Progenesis LC-MS intensity values (higher threshold), Yellow line = Progenesis LC-MS intensity values (lower threshold), Black line = MaxQuant.

2.4.3.3 – CPTAC Heatmaps

A way to visually assess the consensus across pipelines is to create a heatmap for each of the three comparisons, with dendrograms calculated both for the protein axis and the software axis based on distance. The results from the E/B comparison are shown in Figure 19. Only one cluster of proteins show any clear pattern to the results from all software pipelines - at the top of the heat map. This cluster contains: 42 ups proteins and 29 yeast proteins, giving an FDR of 41% and a sensitivity of 88%. No UPS1 proteins are identified outside of this cluster, giving the heatmap method a significant performance gain over the best individual package in terms of identifying differential expression. For the other comparisons; the heat map for the E/C comparison has a clear differentiated cluster with an FDR of 24% and a sensitivity of 92% and that for the E/D heatmap cluster had an FDR of 23% and a sensitivity of 90%.

These all show significant improvements over the use of any single package and have the advantage that they do not rely on arbitrary p-value or fold change thresholds. Clearly, there is still a substantial FDR with all methods, however there are genuine differences in some yeast proteins in some replicates i.e. it is likely that many of the 29 yeast proteins that cluster on the E/B heatmap are genuinely differentially expressed. In each of the four columns on the heatmap, each individual package finds many other proteins as changing in abundance (any strong green or red signal), but almost none of these are UPS1 proteins, and with almost no agreement with other software packages, implying that the difference is due to the way in which the software has processed the data rather than any underlying feature of the data set itself.

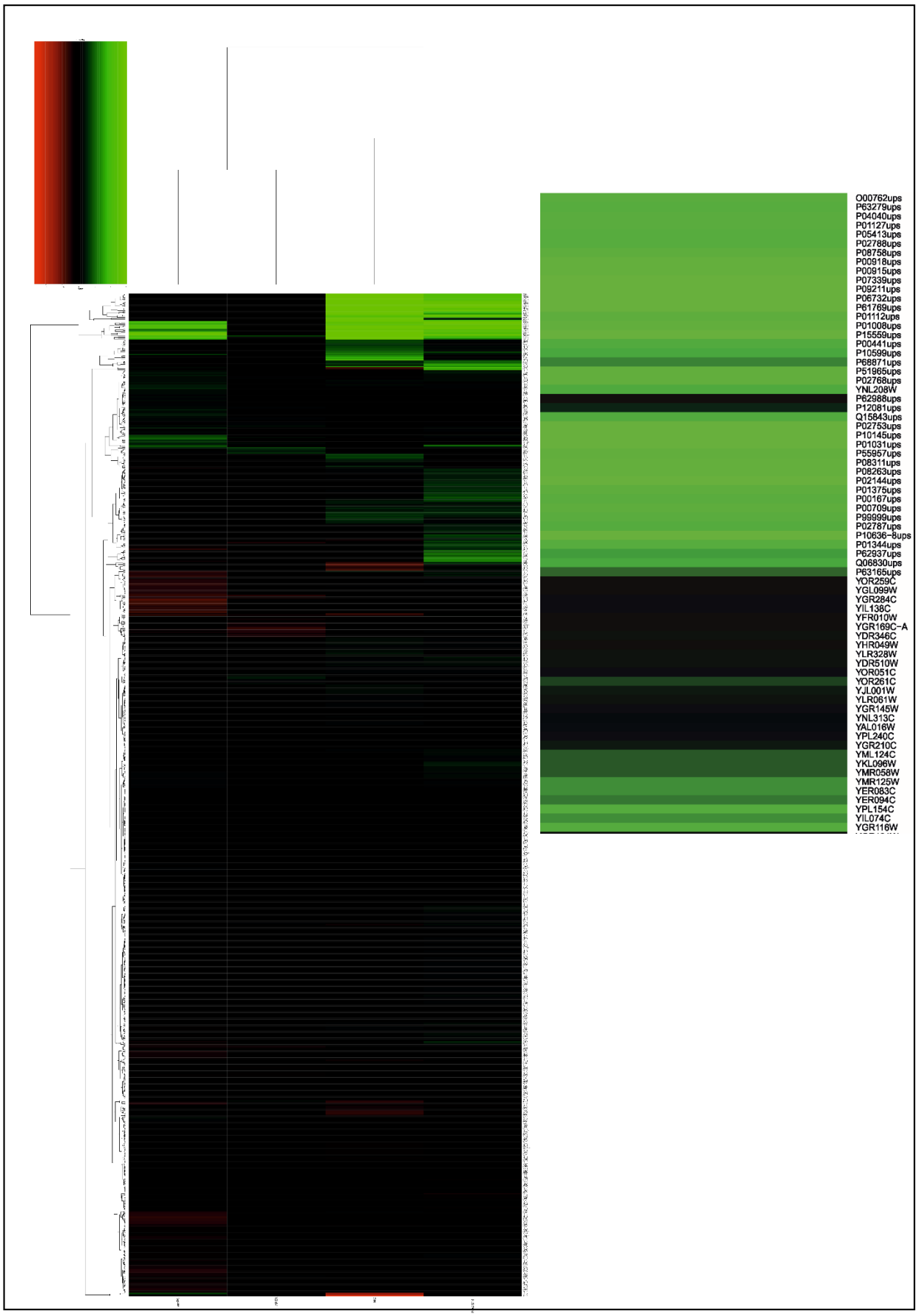

Figure 19: Heatmap (E/B comparison) comparing data from all pipelines, with zoom section of the top cluster of upregulated proteins including the UPS proteins

2.4.3.4 - QPROT post-processing

In an attempt to determine differential expression more accurately, the QPROT tool was used to create an additional set of ROC plots (see Figure 20). In general, QPROT is able to determine differential expression for all packages more accurately than either pair-wise t-tests or fold change measurements. In the E/B and E/C comparisons, MaxQuant performance appears to be best and there is a marked improvement in the emPAI curve, rendering it comparable to the intensity based methods for the E/B and E/C comparisons. When the ratios are smaller (in the E/D comparison) the plot shows that Progenesis LC-MS considerably outperforms the other pipelines when used with the lower identification threshold. In the E/D comparison, all packages perform better with QPROT post-processing than by fold change or t-test, but no package is able to detect more than 80% of the UPS proteins without incurring a substantial FDR (25-80%). Across all conditions the APEX tool appears to perform less well than emPAI. However, it should be noted that the emPAI method is considerably simpler to apply to spectral count data. In contrast, obtaining APEX values requires a number of processing steps through different tools, which may not have been optimised in this analysis. Given that the underlying basis of APEX and emPAI are broadly similar, it would be expected that APEX performance should be comparable to emPAI and though it is likely that each of the processing steps could be optimised to bring the experimental values closer to the ground truth the optimisation required would be unique to each dataset. Clearly then, it would not be possible to ascertain the optimal parameters when the ground truth was not know, i.e. with a real experimental dataset. An interesting question for further work may be to assess the success of analysis of carefully prepared standard known datasets following optimisation with a similar standard dataset. There are issues however with whether such a method would be transferable to the experimental laboratory, as the design of a standard dataset that is suitably similar to experimental samples would be highly challenging.

As illustrated in the previous section, there is currently no universal statistical test and threshold that can control FDR in the analysis of this data set. Our hypothesis is that if different software pipelines agree on a result i.e. that a protein is changing in abundance; this gives greater confidence that this interpretation is correct, since different pipelines are likely to make different kinds of errors. Figure 20 displays the result from a new method which can be used to find consensus between different packages, in terms of those proteins which are called as differentially expressed. In this method, the absolute mean Z-statistic is

calculated from the Z-statistic values given by the QPROT tool for each of the four pipelines studied. In comparisons E/B and E/C, excellent performance is observed, far outperforming any individual software package, even when that packages results have been post-processed with the QPROT tool, achieving around 90% sensitivity before any false positives are incurred. The E/D comparison also shows a considerable performance gain over any individual package, for example, achieving more than 80% sensitivity with less than 10% FDR compared with the best individual package having approximately 25% FDR for the same sensitivity. As an example of why the consensus method is successful, the yeast protein YDR334W is measured as strongly differentially expressed by Progenesis LC-MS in the output from the QPROT tool (second ranking protein overall), with a Z-statistic of 3.49 in the E/D comparison. The other three packages have no value for this protein, meaning it has presumably been misidentified or is present at low abundance, leading to a mean z-statistic of 0.87. Several other yeast proteins share this profile, indicating that the consensus method is acting as a filter for low-quality or low abundance data. A possible explanation for this misidentification within Progenesis LC-MS is that once a protein has been assigned from one spectra it will be inferred to be present in other spectra even if there is no discernible peak and no identification data is present i.e. when no MS-MS spectra have been recorded. A different type of example is the yeast protein YPL154C, scored as strongly differentially expressed by Progenesis LC-MS in the E/D comparison (Z: 2.66), but with Z values from other packages as 0.14 (MaxQuant), 0.51 (emPAI) and 0 (APEX), leading to abs. mean Z = 0.83. In the E/D comparison, there is agreement between Progenesis LC-MS and MaxQuant on two yeast proteins that are strongly differentially expressed - YNL208W and YMR058W, but again with much reduced Z values from the spectral count methods - YNL208W 2.01, 2.94, 0.76 and 0; YMR058W 2.13, 2.74, 0 and 0 (from Progenesis LC-MS, Maxquant, emPAI and APEX respectively).

A further aspect to consider for the consensus method is the setting of appropriate thresholds. On Figure 20 the point corresponding with abs. mean $z > 1.96$ is also annotated, which would be approximately analogous to $p < 0.05$ in a two-tailed t-test. For the E/D and E/C comparisons, this appears to be an appropriate threshold for separating the false positives from true positives. In the E/D comparison however, this threshold is too conservative, reporting FDR = 0 but a sensitivity value of only 0.25.

It is clearly a labour intensive task to run four packages on the same data set. As such, the use of different combinations of software packages was studied to assess whether the same performance gains can be achieved using fewer software packages. The combinations assessed were: MaxQuant, Progenesis and emPAI; MaxQuant and Progenesis; MaxQuant and emPAI; and Progenesis and emPAI. The pseudo-ROC plot profiles (shown in Figure 21) for the E/B and E/C comparisons are relatively similar for all combinations, showing that using an intensity-based package in combination with at least one other package (intensity or spectral counting based) is sufficient to accurately detect differential expression. In the E/D comparison, the performance of most combinations is similar, with a slight drop in performance for MaxQuant and emPAI, presumably since Progenesis LC-MS was the best performing individual package in this comparison and for this dataset.

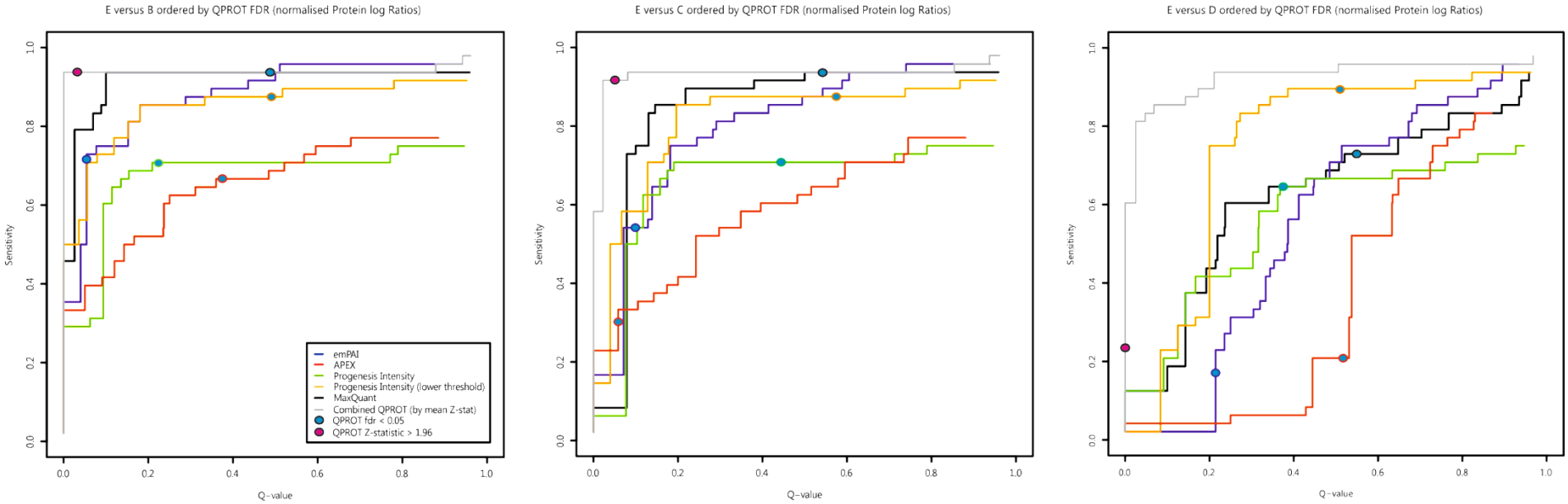

Figure 20: Pseudo-ROC plots for all comparisons ordered by QPROT FDR, annotated with QPROT FDR < 0.05 and QPROT Z-stat > 1.9. Key: Blue line = emPAI, Red line = APEX, Green line = Progenesis LC-MS intensity values (higher threshold), Yellow line = Progenesis LC-MS intensity values (lower threshold), Black line = MaxQuant, Grey line = Combined QPROT (by mean Z-stat).

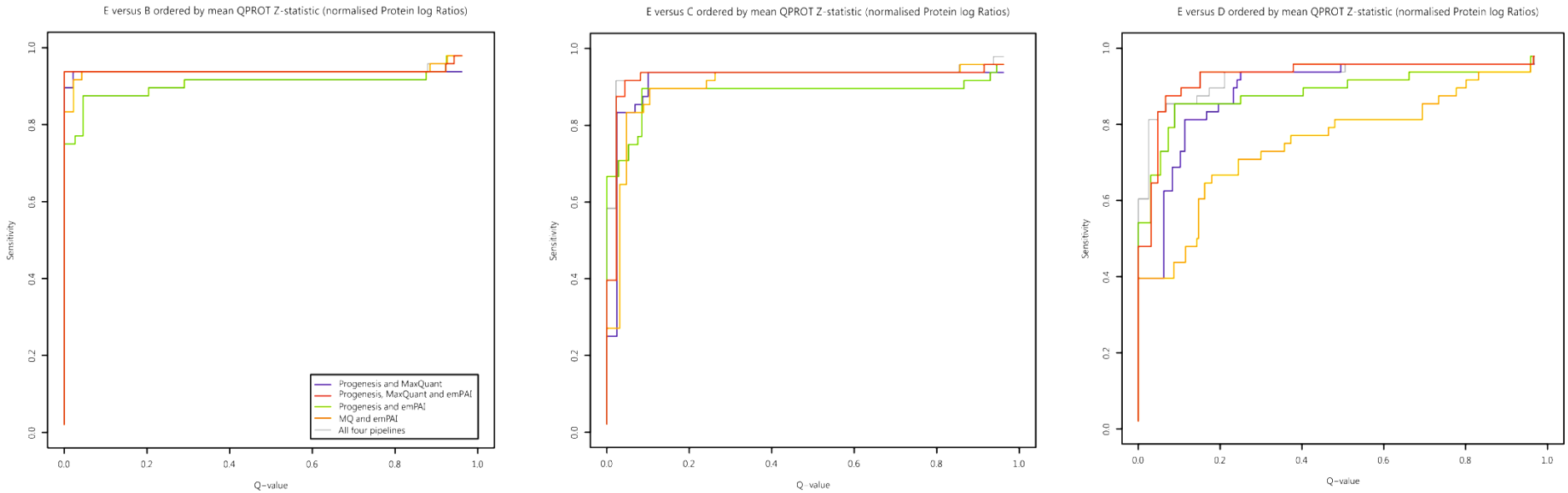

Figure 21: Pseudo-ROC plots for all comparisons ordered by QPROT Z-statistic using multiple packages combined in a consensus method. Key: Blue line = Progesis and MaxQuant, Red line = Progenesis, MaxQuant and emPAI, Green line = Progenesis and emPAI, Grey line = All four pipelines (Progenesis, MaxQuant, emPAI and APEX).

2.5 - Discussion and Conclusions

Many different software packages are available to post-process label-free proteomics data and obtain quantitative values, and there are also several reported methodologies that can also be used independently or alongside these software packages. In this study a single dataset was analysed using four different software packages and the results of these analyses were then compared and post-processing options explored to assess the value of using a combination of software packages to analyse quantitative label-free proteomics data.

The dataset used in this study was created for use in CPTAC Study 6, which required a standard sample of known protein concentrations against a complex background. To achieve this a standard protein mix (UPS1) was spiked into a yeast background sample at several concentrations. In the study presented above ratios were calculated between conditions and compared to the known ratios which should have been present. The two intensity based methods (MaxQuant and Progenesis LC-MS) most closely report the correct UPS1 protein ratios when the raw protein abundance values are used, indicating that the spike-in had been done correctly. However it was observed that applying global normalisation to the data skewed the UPS1 protein ratios – suppressing the ratios in Progenesis LC-MS and increasing the ratios in MaxQuant. This is concluded to be because though the majority of the proteins in the sample are expected to be present at a 1:1 ratio, when one replicate is chosen as the master some replicates show global differences in the abundance of yeast proteins relative to this master. Global normalisation brings these differences into line, however it also changes the values for the UPS proteins. This means that the presence of multiple distributions should be considered when there are multiple populations present in a sample and it would be ideal to apply specifically tailored normalisation schemes to those datasets i.e. performing normalisation to each distribution separately. Evidently this type of strategy requires an in depth knowledge of the dataset that is being studied, which is potentially challenging when studying real biological data. Another consideration which arose from the study of this dataset is the necessity to ensure that the background proteins present in a standard protein mix are truly homogenous over all samples, as in this dataset there were a lot of yeast proteins which were identified to be changing between conditions/replicates (as the ratios for the UPS proteins were broadly correct, this lends confidence to the conclusion that there is true variation in the majority of the yeast proteins reported as changing in abundance).

It was seen that p-value and fold change thresholds were ineffective as a method to increase sensitivity and reduce FDR, with a high FDR being necessary to achieve 80% sensitivity. This is more pronounced in the spectral counting based data where an 80-90% FDR is present, while the intensity based methods achieve both increased sensitivity and decreased FDR. While it could be possible to optimise thresholds for a known dataset, this is not possible in a real dataset where the ground truth is unknown. The use of the QPROT tool renders all the techniques more effective in terms of increasing sensitivity while reducing FDR, and indeed this post processing provides enough improvement to the emPAI data in particular to render it comparable to the intensity based methods. It was observed that p-value calculations are particularly unsuited to label-free data where there is high variability between replicates, particularly when the number of replicates is low. Fold change and p-value should be weakly correlated, however as p-value calculations take the variance between replicates into account and fold change calculations do not, the level of correlation will be dependent upon the present variation across replicates in the given dataset.

Two methods were used to assess the benefit of using multiple packages in conjunction to achieve a consensus result. The first method concerned the generation of heatmaps, which are a useful technique to get a visual representation of a consensus on differential expression in the data. Despite this there is still a large FDR associated with a reasonable level of sensitivity. The second consensus method uses the results output by the QPROT tool, specifically using a mean absolute Z-statistic calculated for each protein from the Z-statistic values output by QPROT for each software pipeline. This value was used to produce pseudo-ROC plots with a greatly improved FDR to sensitivity ratio – giving approximately 90% sensitivity before false positives appear in the results list. The main advantage of this method is that it is unlikely that a false positive protein will be identified with a high Z-statistic value by all software pipelines and therefore the effect is that the false positive proteins are filtered out from the results.

In conclusion there is a definite benefit to using multiple software packages to achieve a consensus result. Even when only two software packages are used in conjunction with the QPROT tool there is a significant improvement in the sensitivity to FDR ratio. This is a realistic way to conduct studies, which could be utilised by study groups working on real biological data, especially as emPAI numbers are reported directly from Mascot, which is very widely

used. The emPAI can then be used in conjunction with the intensity based package of choice to give a consensus on which proteins are truly differentially expressed.

As an extension to this study it would be extremely useful to obtain or design samples such that they contain known amounts of several proteins or groups of proteins, in order that the ground truth is known. Such a dataset was designed by Stefan Tenzer, and presented at various conferences, which contains weighted populations of mouse, yeast and *E-coli* proteins to ensure a uniform total protein amount. Work on this or other similar datasets could also further consider the merits of splitting the experimental data into its constituent data distributions prior to normalisation steps.

In addition to the considerations outlined above, the work for this study was completed in 2009-2011, and it is possible that the results could be improved were the analysis repeated using modern search engines (such as Byonic[105, 106], PEAKS[107, 108] or ProteinPilot (AB Sciex)) for the protein identification step.

3 - Pairwise comparisons of the results obtained when using different mass spectrometry instrument platforms for label-free data

3.1 - Introduction

The recent increased popularity of label-free proteomics has been mirrored by an increase in the number of mass spectrometry instruments which are suitable for that purpose. These instruments are often from different instrument vendors, and have a variety of different mass analysers, ion optics, or combinations of mass analyser types.

Each laboratory performs biological experiments using the instrumentation that they have available, or which they have chosen (from a small or wide range of options) as most appropriate for a given experiment, and the reported results are considered comparable by the proteomics community. This study was designed to test this assumption of comparability between instrument platforms by assessing the correlation between the results obtained when the same biological samples were run on two different mass spectrometers, and post processed analogously (using Non-linear Dynamics' Progenesis LC-MS and self-written Java code). An ideal dataset would come from an experiment in which identical chromatographic conditions had been used on both instruments, however this dataset was obtained using the standard chromatographic conditions for both instruments (the chromatographic conditions used are shown below in Table 3). While this is not ideal, it may represent a more accurate representation of the typical situation where samples are sent to be analysed on multiple instruments or where the mass spectrometric analysis is conducted by a non-expert user with the "standard" instrument parameters.

The two mass spectrometers used in this study were a Thermo Scientific Orbitrap VELOS and a Waters Synapt G2. These are both popular instruments from well-respected vendors, but they possess very different internal ion optics, as detailed below and shown in Figure 22.

Thermo Scientific Orbirtrap VELOS: This is an orbitrap instrument in which ions pass through a linear ion trap to a curved linear ion trap (C-trap) and can then be sent to the orbitrap or onwards to the collision cell. There is capability within the ion optics to analyse both parent and fragment ions using the orbitrap. Initial mass analysis and ion selection is generally performed in the orbitrap, with the ions then being passed through the C-trap and collision cell to a linear ion trap for fragmentation, or passed straight to the scanning electron microscope (SEM) detector. In this experiment the instrument was set up to perform data dependent analysis on the twenty most intense ions in each full MS scan, which are passed sequentially into the collision cell for fragmentation after ion selection and accumulation.

Waters Synapt G2: This is a time-of-flight instrument with an electrical W-shaped flight path within the flight tube, achieved by two ion mirrors. This allows increased resolution in the mass measurements via increasing the length of the ion flight path without an analogous increase in the physical length of the flight tube itself. The ions pass through an initial quadrupole stage and then into a helium cell where, uniquely for this instrument, all ions are fragmented as the instrument switches at high speed between high and low collision energies throughout the experiment. The advantage of this is that there is MS-MS information for every ion observed in the full scan spectrum, whereas data dependent analysis only provides information about the most intense ions. This unique method of analysis is marketed by Waters as MS^E and is more generally referred to as data independent acquisition (DIA).

	Orbitrap	SYNAPT
Column	nanoACQUITY UPLC™ BEH130 C18 15cm x 75µm, 1.7µm capillary column	nanoACQUITY UPLC HSS T3 C18; 15cm× 75µm, 1.7µm capillary column
Gradient	3-40% acetonitrile in 0.1% formic acid for 90min then 40-85% acetonitrile in 0.1% formic acid for 3 min	3-40% acetonitrile in 0.1% formic acid for 120 min
Flow Rate (nl/min)	300	300

Table 3: Chromatographic conditions for the mass spectrometric analysis on each instrument

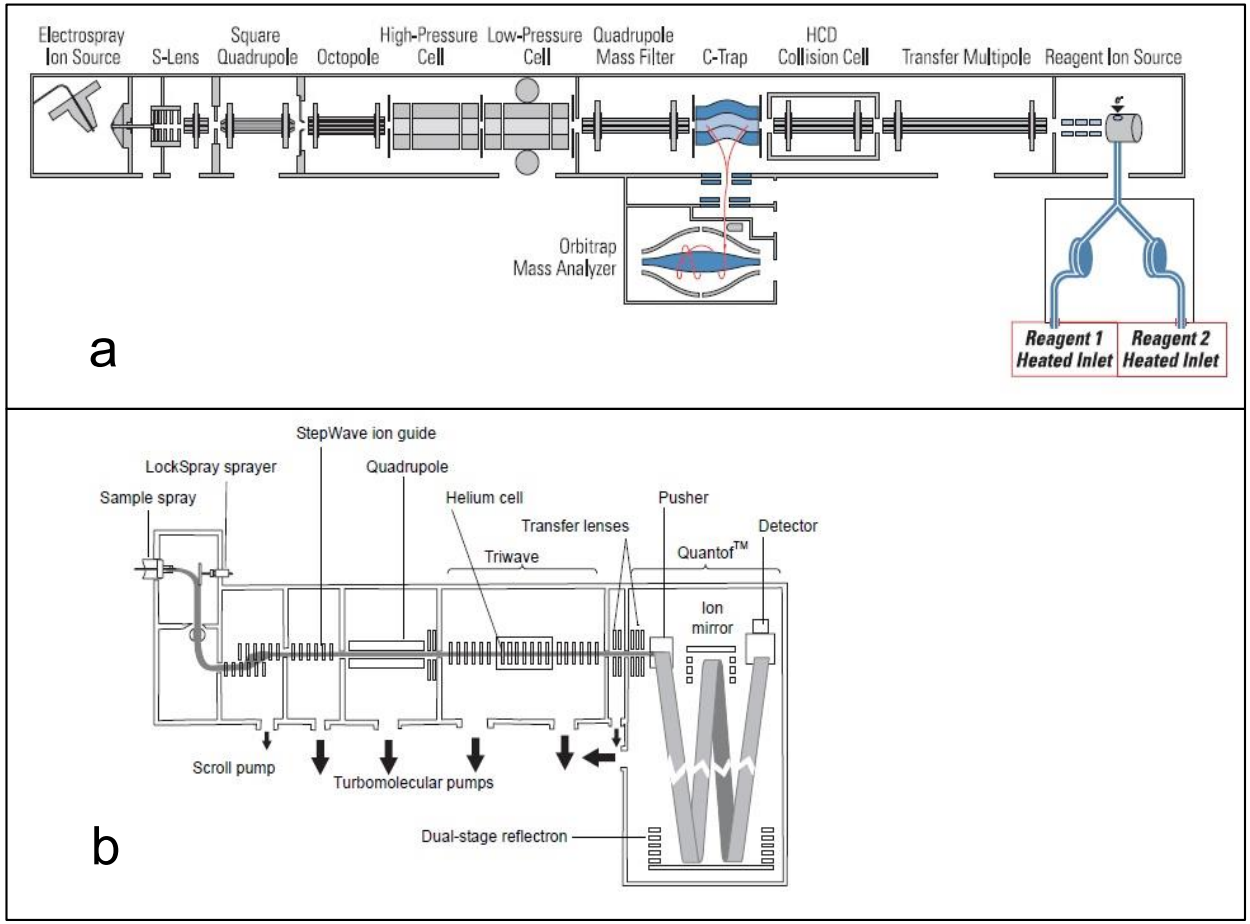

Figure 22: Schematic diagrams showing the internal optics of both the Thermo Scientific Orbitrap VELOS and the Waters Synapt G2, taken from the SYNAPT G2-S HDMS Operator's Overview and Maintenance Guide and the Thermo Scientific LTQ Orbitrap Velos Product Specifications respectively

3.2 - Methods

Time-course data was collected from a culture of human foreskin fibroblast cells infected with *Toxoplasma gondii* VEG (data collected at 2, 4, 8 and 16 hours post infection, data from 2, 4 and 8 hours post infection used for this study). Three biological samples were collected for each timepoint and the sample preparation was completed analogously. Each sample was then run on the two instruments using their respective routinely used optimised chromatographic conditions and comparable analytical columns (C18 15cm x 75µm, 1.7µm capillary column, 300nl/min flow rate, 35^{o}C, gradient: 90mins 3-40%B, 3mins 40-95%B and 120mins 3-40%B for the Thermo Orbitrap VELOS and the Waters Synapt G2 respectively). The wet-lab work was completed by a collaborator (Dr Dong Xia), and the raw data output from the mass spectrometry analysis was obtained and used for this study. Post-processing was completed using Progenesis LC-MS, as comparatively as possible, with only a slight divergence in the methods used. This was due to the fact that only the specifically designed Waters software (ProteinLynx Global Server or PLGS) is capable of processing the MS^{E} data from the Waters Synapt G2 instrument. PLGS was therefore used to perform database searching of the MS^{E} data, using a reverse database generated within PLGS itself. The results of this analysis were then exported using the Progenesis PLGS plug-in and the output imported into Progenesis LC-MS for further analysis. PLGS is vendor software from Waters and therefore cannot be used to process data from instruments sold by other vendors i.e. the raw data obtained for this study from the Thermo Scientific Orbitrap VELOS. Therefore this data was aligned in Progenesis LC-MS to create a merge file representing the aggregate spectrum, which was then searched against the reverse database generated by PLGS using Mascot (Matrix Science) on the in-house Mascot server. The Mascot result file (as an XML) was then imported back into Progenesis LC-MS to complete the parallel analyses.

Once the analysis in Progenesis LC-MS was complete the feature level data was exported as .csv files for further analysis. The exported feature list files were input into self-written Java code which finds the common peptide sequences and reads out the data from both instruments into a single file (peptide sequence, feature ID, modifications, peptide score and abundance values are included in this file), referred to as the *common features file*.

The common features file and the feature list files from both instruments were then saved as Excel workbooks for manual/semi-automated analysis. For the feature level analysis score (10%, 25% and 50% percentiles, i.e. retaining the top 10, 25 or 50% of the data respectively) and abundance (25% and 50% percentiles) thresholds were applied to the common features file and the resultant data saved as separate .csv files containing the abundance data for each time point studied. These were used as input for self-written R code which produces a correlation plot of the data.

For the protein level analysis the common features file was saved as an Excel workbook and that data copied into several duplicate sheets, to which a threshold was applied by either score (again using 10%, 25% and 50% percentiles, i.e. retaining the top 10, 25 or 50% of the data respectively) or abundance values (again using 25% and 50% percentiles) using the PERCENTILE function built into Excel. Those features which passed the relevant thresholds were then moved to separate sheets to be used for reference. A VLOOKUP function was used to search between the Progenesis LC-MS feature list workbooks and the reference sheets in the common features workbook, this imported the data if it passed a given threshold and otherwise reports a blank row. This gave sheets which were in the same format as the original Progenesis LC-MS output but contained only the features which passed the relevant thresholds (the formula reported non-passing features as blank rows, which were then deleted by filtering for blanks and deleting the matching rows). An additional step was required because there are duplicate features assigned to some of the peptide sequences (usually when different charge states had been assigned to the same peptide) in both datasets. To remove the duplicates the data was sorted alphabetically by peptide sequence and then by the abundance of the relevant replicate (or by the score for the score thresholded data). The second level of the search was to ensure that the most abundant (or highest scored) peptide was taken forward through the analysis when the remove duplicates function was used (this function retains the first instance of a duplicate and removes subsequent instances). Before further processing the file was then sorted by number to get the list back into the order as it was reported in the original Progenesis LC-MS output file.

The thresholded feature list files were also input into the Progenesis Post Processor (PPP) software which was written in-house by Dr Da Qi[109]. This package uses the feature data to produce hi3 protein data (from the top three features assigned to each protein i.e. the three assigned features which have the highest ion abundance) which is analogous to that

produced by PLGS for the Waters MS^E data, and also produces total abundance protein data. It was these protein results from the PPP that were used in this study for further protein level analysis.

The protein results from the PPP were edited manually in Excel to create a file with the same headings as a Progenesis LC-MS output file. This could then be input into the self-written Java code to find which proteins were common between the two instruments and read these out into a new file called the common proteins file. This file was used to generate .csv files containing the data for each time point studied, and these were used as input to R code written to produce correlation plots of the data.

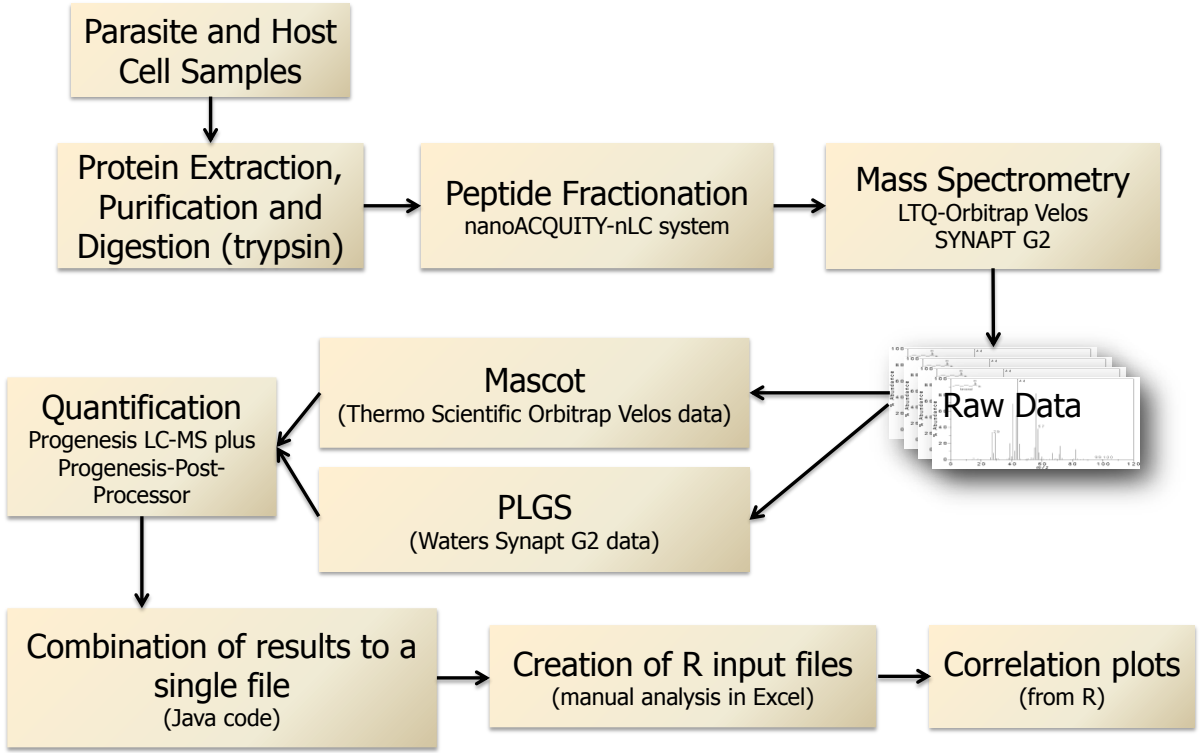

Figure 23: Flowchart describing the process from .raw data to the final correlation plots

3.3 - Results

3.3.1 - Common and unique proteins

The first and most basic exploration of the data was a simple comparison of the number of proteins identified following analysis of the data from each instrument. The numbers for the unique proteins were taken from the original Progenesis LC-MS output files, while the number of common proteins was taken from the common proteins report file generated by my Java code with these output files as the input. It was decided to use these files for this part of the analysis to remove any bias introduced by forcing the use of common features only for protein inference (as is done in the later analysis when only common features/proteins are considered).

Figure 24 shows that approximately one fifth of the proteins (339 out of 1453) identified in the Thermo Orbitrap VELOS data were common to the result sets generated by both instruments, compared to just under half of the proteins identified in the Waters Synapt G2 data (339 out of 381). This translates to a quarter and two thirds respectively when a three peptide threshold is applied.

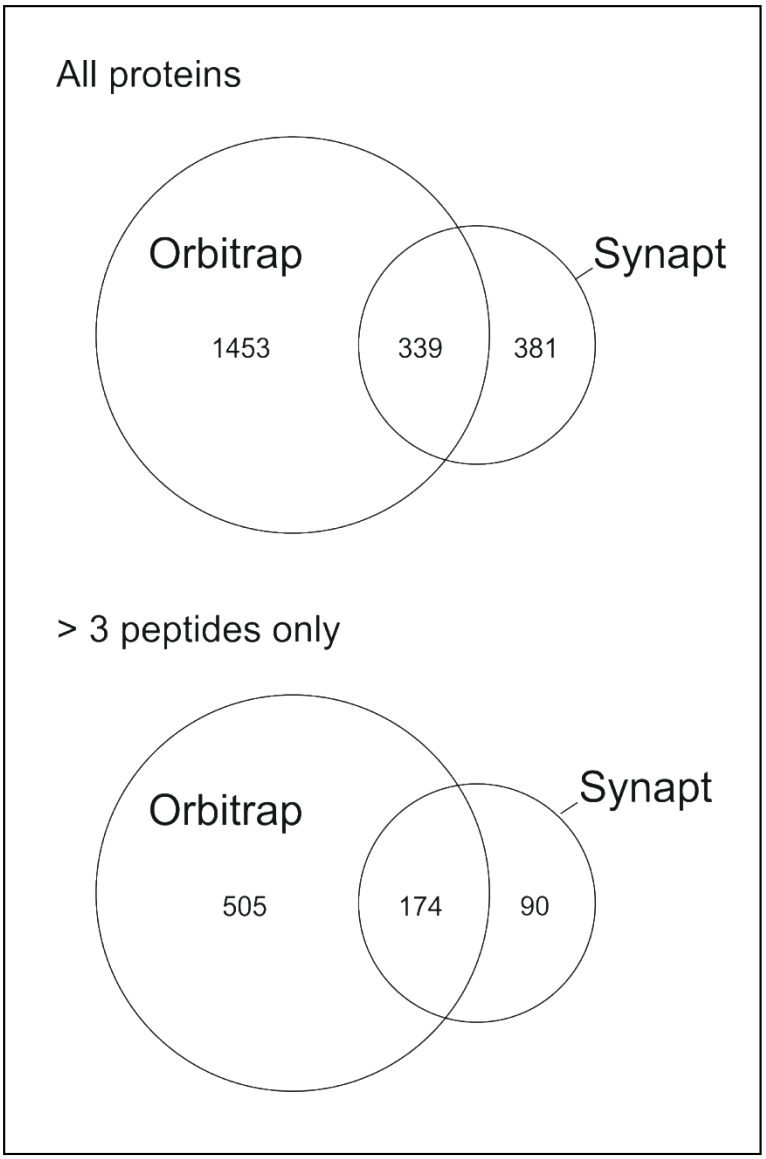

Figure 24: Venn diagrams showing the number of proteins reported by both instruments, with the number of common proteins also shown. Top shows all the reported proteins, while the lower Venn shows the resultant data when a three peptide threshold is applied.

3.3.2 - Pearson correlation

It is essential to accept the assumption that the results obtained from running a single sample on different instruments should give correlated results to allow the comparison of conclusions drawn by different groups about the differential expression of a protein of interest. By producing Pearson correlation plots of the data obtained in this study the assumption of correlation was tested for both hi3 and total abundance data (from PPP output), and following the application of score and abundance thresholds to both.

3.3.2.1 - Feature data correlation at different thresholds

The feature data from both instruments is observed to correlate reasonably well – with Pearson values between 0.5 and 0.9 (see Table 4). The application of score thresholds (i.e. retaining the top 10, 25 or 50% of the data when ordered by peptide score) yields unpredictable results in the feature data. A higher score threshold would be expected to increase the correlation between the two instruments, however an increased score threshold is observed to cause both increased and decreased correlation depending on the time point being studied (see Table 4, left hand panels).

The effect of abundance thresholds is more as expected, i.e. the data shows a positive trend in Pearson value as the threshold is increased. At all time points there is a large jump in the Pearson value following the application of a 25% abundance threshold (with respect to no threshold being applied), and a smaller increase when a 50% threshold is applied (see Table 4, right hand panels). It can also be seen from the correlation plots that it tends to be low abundance features which are identified by only one of the two instruments.

Sample 1.1		
Threshold	Score Correlation	Abundance Correlation
None	0.5598	0.5598
10%	0.5736	n/a
25%	0.4492	0.7492
50%	0.3941	0.8367
Sample 2.1		
Threshold	Score Correlation	Abundance Correlation
None	0.5934	0.5934
10%	0.6068	n/a
25%	0.606	0.8061
50%	0.707	0.8957
Sample 3.1		
Threshold	Score Correlation	Abundance Correlation
None	0.578	0.578
10%	0.5867	n/a
25%	0.5683	0.7689
50%	0.6791	0.8491

Table 4: Pearson value tables for Orbitrap vs Synapt (feature level) data at all timepoints. "Score Correlation" refers to the Pearson value observed when thresholds are applied at 10, 25 and 50% percentiles respectively, to the data when ordered by peptide score. "Abundance Correlation" refers to the Pearson value observed when thresholds are applied at 10, 25 and 50% percentiles respectively, to the data when ordered by peptide abundance.

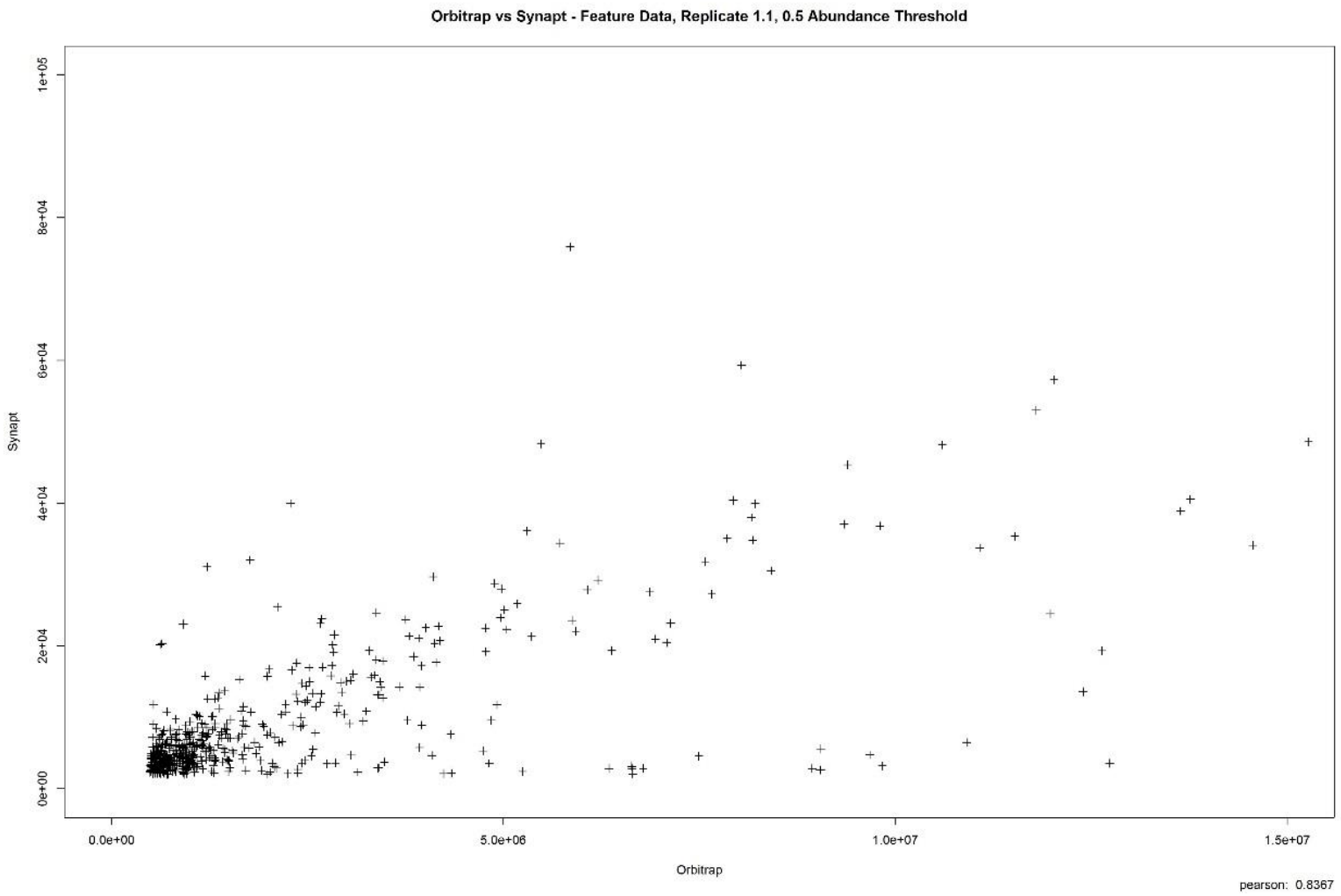

Figure 25: Correlation graph for Orbitrap vs Synapt (feature level) data at 50% abundance threshold for Sample 1.1

3.3.2.2 - Protein data correlation at different thresholds

In general the protein data shows improved correlation between instruments with respect to the feature data (where the Pearson values ranged from 0.5-0.9, see Table 4), as all but one Pearson value being above 0.9 and that one value still being high at 0.8573 (see Table 5). This is considered to be as expected because each feature assigned to a protein contributes to the final protein abundance, and thus the effect of those features which have low abundance and high variability will be somewhat masked at the protein level.

The application of peptide score thresholds is again observed to have no reliable influence on the correlation of the two sets of data, with the Pearson value being above 0.9 at all thresholds (see Table 5, left hand panels). However the application of score thresholds does markedly affect the number of proteins identified – a 50% score threshold reduces the total number of commonly identified proteins from 96 to 16 (14 for hi3 data where NaN values were removed – these arise where the protein has been identified and reported within the total abundance data from Progenesis LC-MS but does not have three or more peptides and therefore does has no reported hi3 abundance values). This large reduction in the number of proteins identified is a trade-off that few would be willing to accept, especially there is no reliable improvement seen in the Pearson values (see Table 5, left hand panels).

The application of abundance thresholds shows more of an effect to the correlation between the two datasets, with the greatest improvement being seen in the total abundance data where an improvement of 0.02 has been observed in the reported Pearson value when a 50% abundance threshold has been applied.

As a general observation it is seen that there is greater correlation in the hi3 data than in the total abundance data (see Tables 5-7 below). One possible explanation for this is that the hi3 process excludes those lower abundance peptides which are more likely to be recorded unpredictably due to differences in ionisation efficiency and ion transfer through the ion optics of the different instruments. Those peptides which are more readily ionised will be present at greater abundance within the mass spectrometer and therefore transfer differences between instruments will have a proportionally lower effect on high abundance ions and thus greater correlation is to be expected for these ions.

The application of abundance thresholds has a less extreme effect on the number of identified proteins than the application of score thresholds, however the reduction in number is still quite large, with a 50% threshold reducing the protein count to 38 from 98.

Protein Correlation – Sample 1.1						
Score Values				Abundance Values		
hi3				hi3		
Threshold	Pearson	No. of Proteins		Threshold	Pearson	No. of Proteins
None	0.9696	96		None	0.9635	98
10%	0.9713	82		25%	0.9714	69
25%	0.9223	59		50%	0.9736	38
50%	0.9755	14				
Total Abundance				Total Abundance		
Threshold	Pearson	No. of Proteins		Threshold	Pearson	No. of Proteins
None	0.9554	96		None	0.9516	98
10%	0.9329	82		25%	0.9587	69
25%	0.9182	59		50%	0.9654	38
50%	0.8573	16				

Table 5: Protein correlation values for hi3 and total abundance data, for the first time point with score and abundance thresholds applied. "Score Values" refers to the Pearson value observed when thresholds are applied at 10, 25 and 50% percentiles respectively, to the data when ordered by peptide score. "Abundance Values" refers to the Pearson value observed when thresholds are applied at 10, 25 and 50% percentiles respectively, to the data when ordered by peptide abundance. The upper panels show values for the hi3 data, while the lower panels show the values for total abundance data.

Protein Correlation – Sample 2.1						
Score Values				Abundance Values		
hi3				hi3		
Threshold	Pearson	No. of Proteins		Threshold	Pearson	No. of Proteins
None	0.7943	96		None	0.7981	99
10%	0.7983	82		25%	0.9594	61
25%	0.9261	59		50%	0.9797	35
50%	0.9563	14				
Total Abundance				Total Abundance		
Threshold	Pearson	No. of Proteins		Threshold	Pearson	No. of Proteins
None	0.9244	96		None	0.927	99
10%	0.9273	82		25%	0.9474	61
25%	0.8788	59		50%	0.9717	35
50%	0.9548	16				

Table 6: Protein correlation values for hi3 and total abundance data, for the second time point with score and abundance thresholds applied. "Score Values" refers to the Pearson value observed when thresholds are applied at 10, 25 and 50% percentiles respectively, to the data when ordered by peptide score. "Abundance Values" refers to the Pearson value observed when thresholds are applied at 10, 25 and 50% percentiles respectively, to the data when ordered by peptide abundance. The upper panels show values for the hi3 data, while the lower panels show the values for total abundance data.

Protein Correlation – Sample 3.1						
Score Values				Abundance Values		
hi3				hi3		
Threshold	Pearson	No. of Proteins		Threshold	Pearson	No. of Proteins
None	0.8233	96		None	0.816	99
10%	0.8472	82		25%	0.9778	71
25%	0.9309	59		50%	0.9737	41
50%	0.9391	14				
Total Abundance				Total Abundance		
Threshold	Pearson	No. of Proteins		Threshold	Pearson	No. of Proteins
None	0.9165	96		None	0.9232	99
10%	0.9145	82		25%	0.9658	71
25%	0.9096	59		50%	0.9587	41
50%	0.8891	16				

Table 7: Protein correlation values for hi3 and total abundance data, for the third time point with score and abundance thresholds applied. "Score Values" refers to the Pearson value observed when thresholds are applied at 10, 25 and 50% percentiles respectively, to the data when ordered by peptide score. "Abundance Values" refers to the Pearson value observed when thresholds are applied at 10, 25 and 50% percentiles respectively, to the data when ordered by peptide abundance. The upper panels show values for the hi3 data, while the lower panels show the values for total abundance data.

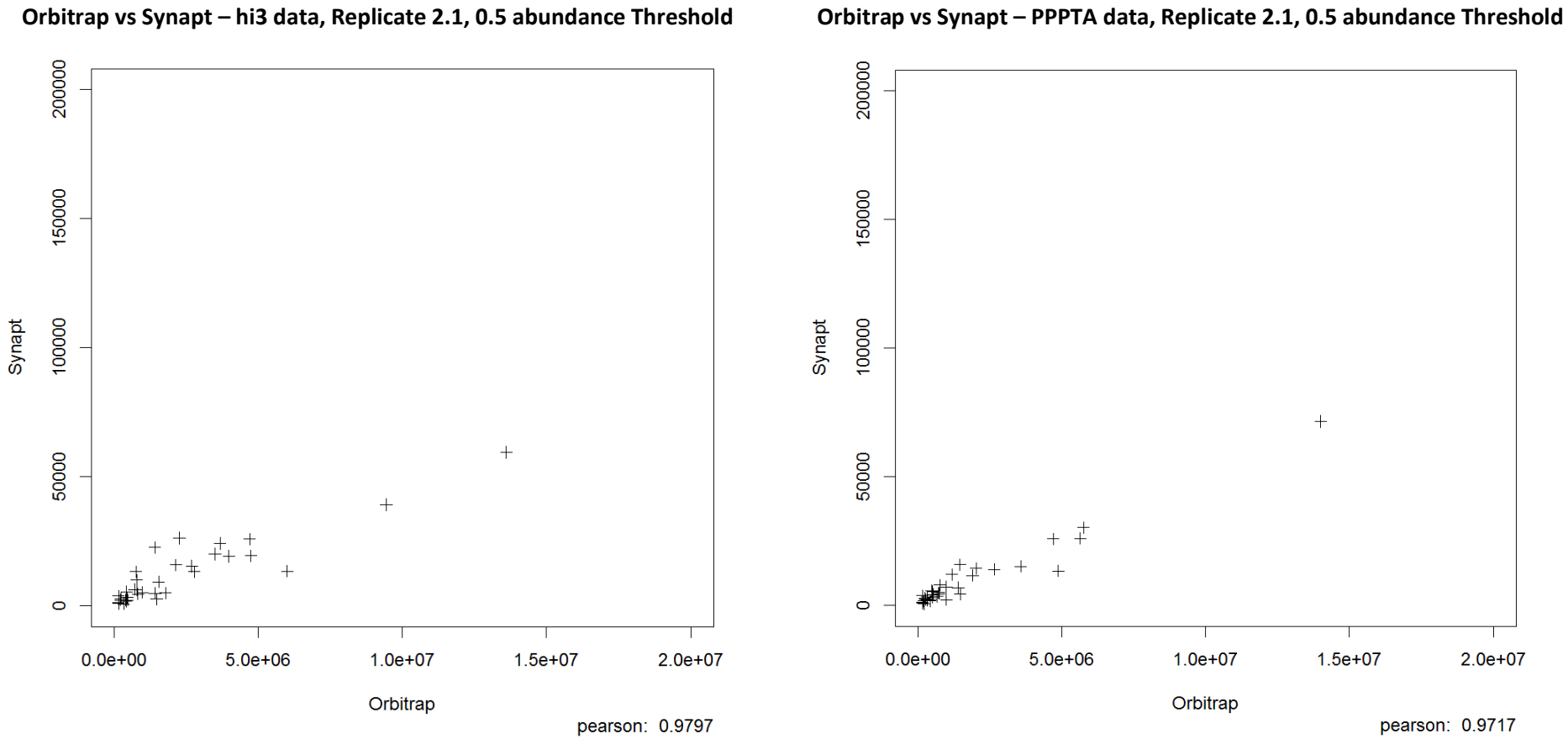

Figure 26: Correlation graphs for Orbitrap vs Synapt (protein level) data at the 50% abundance threshold for Sample 2.1. The left hand panel shows hi3 data and the right hand panel shows total abundance data (both hi3 and total abundance (PPPTA) data taken from the output of the Progenesis Post Processor).

3.3.3 - Coefficient of Variance

3.3.3.1 - Feature data

Another way to examine the data and to compare the performance of the two instruments is to look at the coefficient of variance across the three biological replicates studied at each time point. Though these are biological rather than technical replicates any variation between them should be constant as for each numbered sample a single pellet was made up into two samples which were run in parallel on both instruments. Therefore it is considered that any remaining variation should be instrument dependent.

The coefficient of variance was calculated for three sets of data (see Table 8) – raw values from Progenesis, normalised values from Progenesis, and MAD normalised data (as used for the CPTAC dataset in Chapter 2 above). In all cases the coefficient of variance calculated for the Waters Synapt G2 was greater than that for the Thermo Orbitrap VELOS, with the highest CoV ratio being over two.

It is possible that this is purely a function of Progenesis LC-MS being better optimised for the analysis of Thermo data (for which it was originally designed), but the increased variance could also arise from the unique sampling method employed within the Waters Syanpt G2 instrument. It is likely that the analysis of Waters Synapt data will soon be fully integrated into Progenesis LC-MS, and this would allow a fuller analysis with identical post processing, to fully identify the source of this variation.

Progenesis LC-MS Raw Data						
Sample Name	Average coefficient of variation			Median coefficient of variance		
	Orbitrap	Synapt	S/O	Orbitrap	Synapt	S/O
Inf1	0.326985	0.514743	1.5742075	0.2622783	0.3111407	1.1862999
Inf2	0.314077	0.38989978	1.2414147	0.2792001	0.2878652	1.0310355
Inf3	0.247974	0.42565086	1.7165174	0.1972386	0.2680221	1.3588722
Progenesis LC-MS Normalised Data						
Sample Name	Average coefficient of variation			Median coefficient of variance		
	Orbitrap	Synapt	S/O	Orbitrap	Synapt	S/O
Inf1	0.272694	0.51660522	1.8944527	0.2016511	0.310943	1.541985
Inf2	0.245177	0.329698	1.3447361	0.1909607	0.2118725	1.1095083
Inf3	0.252664	0.41805392	1.6545863	0.2054736	0.25448	1.2385045
Median Normalised Data						
Sample Name	Average coefficient of variation			Median coefficient of variance		
	Orbitrap	Synapt	S/O	Orbitrap	Synapt	S/O
Inf1	0.246791	0.49869671	2.0207273	0.1673684	0.2858771	1.7080708
Inf2	0.228554	0.32818808	1.43593	0.1731137	0.2074545	1.1983712
Inf3	0.248414	0.41800993	1.6827134	0.1997049	0.2513774	1.2587442

Table 8: Coefficient of variance tables for median normalised feature data from both instruments

Box plots were created for the coefficient of variance at all time points (median normalised data shown in Figure 27), and show that there is greater spread in the Waters Synapt G2 data, though the median values from both instruments are similar (there are several outliers present outside the whiskers of the box plots, this is considered to be an artefact of there being thousands of data points within the dataset and is not unusual).

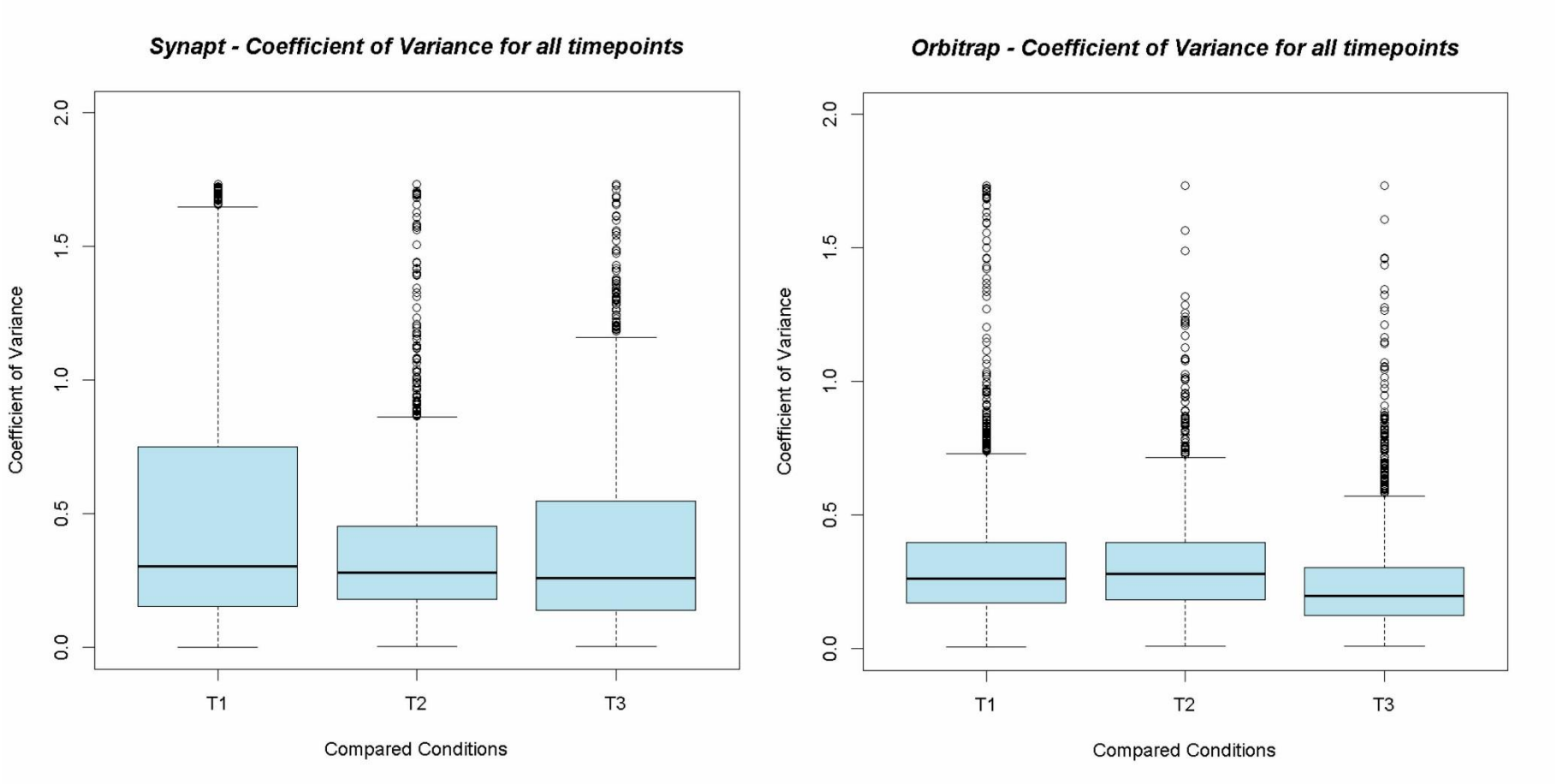

Figure 27: Box plots showing the coefficient of variance for feature data at all time points as analysed on both instruments (median normalised data)

The increased spread in the Waters Synapt G2 data implies that there may be some bad data points present, which skew the mean values. It may be possible to pick out these "bad" data points by manually examining the extracted ion chromatograms and excluding any features that look unreliable. However, based on this data, it is not possible to exclude data processing as the source of the increased variation shown in the Waters Synapt G2 data. It is anticipated that a truly parallel analysis of the data from different instruments will soon be possible using Progenesis LC-MS, and this would allow a more in depth investigation of the sources of variation.

3.3.3.2 - Protein data

The coefficient of variance analysis was repeated with the protein data, again looking at both the raw and normalised data reported from Progenesis LC-MS and the median normalised data (calculated manually in Excel from the raw data). For this part of the analysis the protein data calculated from all the features found by each instrument was used instead of that calculated from the common features only, in order to avoid any bias that could be introduced by reducing the number of available features from which protein abundance is calculated.

In the protein data the coefficient of variance is highly similar between the two instruments (see Table 9), with normalisation of the data both within Progenesis LC-MS and manually by median absolute deviation within Excel causing a decrease in the average coefficient of variance for both instruments. However this affect is more pronounced in the Thermo Orbitrap VELOS data than the Waters Synapt G2 data, meaning that again the coefficient of variance is higher for the Waters Synapt G2 – with the highest ratio (Synapt CoV over Orbitrap CoV) being 1.5 for the first time point.

Progenesis LC-MS Raw Data						
Sample	Average coefficient of variance			Median coefficient of variance		
	Orbitrap	Synapt	S/O	Orbitrap	Synapt	S/O
T1	0.208066	0.224417	1.078585	0.192651	0.164244	0.852546
T2	0.224619	0.231215	1.029368	0.225126	0.219493	0.974978
T3	0.148038	0.196253	1.325693	0.12992	0.155992	1.200674
Progenesis LC-MS Normalised Data						
Sample	Average coefficient of variance			Median coefficient of variance		
	Orbitrap	Synapt	S/O	Orbitrap	Synapt	S/O
T1	0.156016	0.238345	1.5276923	0.136791	0.206549	1.5099616
T2	0.148264	0.175233	1.1818932	0.135648	0.144262	1.0635025
T3	0.146347	0.195426	1.3353595	0.128381	0.155845	1.2139279
Median Normalisation Data						
Sample	Average coefficient of variance			Median coefficient of variance		
	Orbitrap	Synapt	S/O	Orbitrap	Synapt	S/O
T1	0.136987	0.202409	1.477577	0.110148	0.143632	1.303986
T2	0.140895	0.16998	1.206428	0.125881	0.134142	1.065628
T3	0.145043	0.193383	1.333282	0.127887	0.155739	1.217781

Table 9: Coefficient of variance tables for median normalised protein data from both instruments. The "S/O" column shows the ratio of Synapt CoV/Orbitrap CoV.

The same is clear when the normalised data is represented as box plots for both instruments (see Figure 28). Again the median values for coefficient of variance are similar in the data from both instruments, but the spread is greater in the data from the Waters Synapt G2.

On creating box plots of the raw and normalised protein data from both instruments it is seen that though the median coefficient of variance is more comparable between instruments, it is clear that the normalisation has been effective for this data.

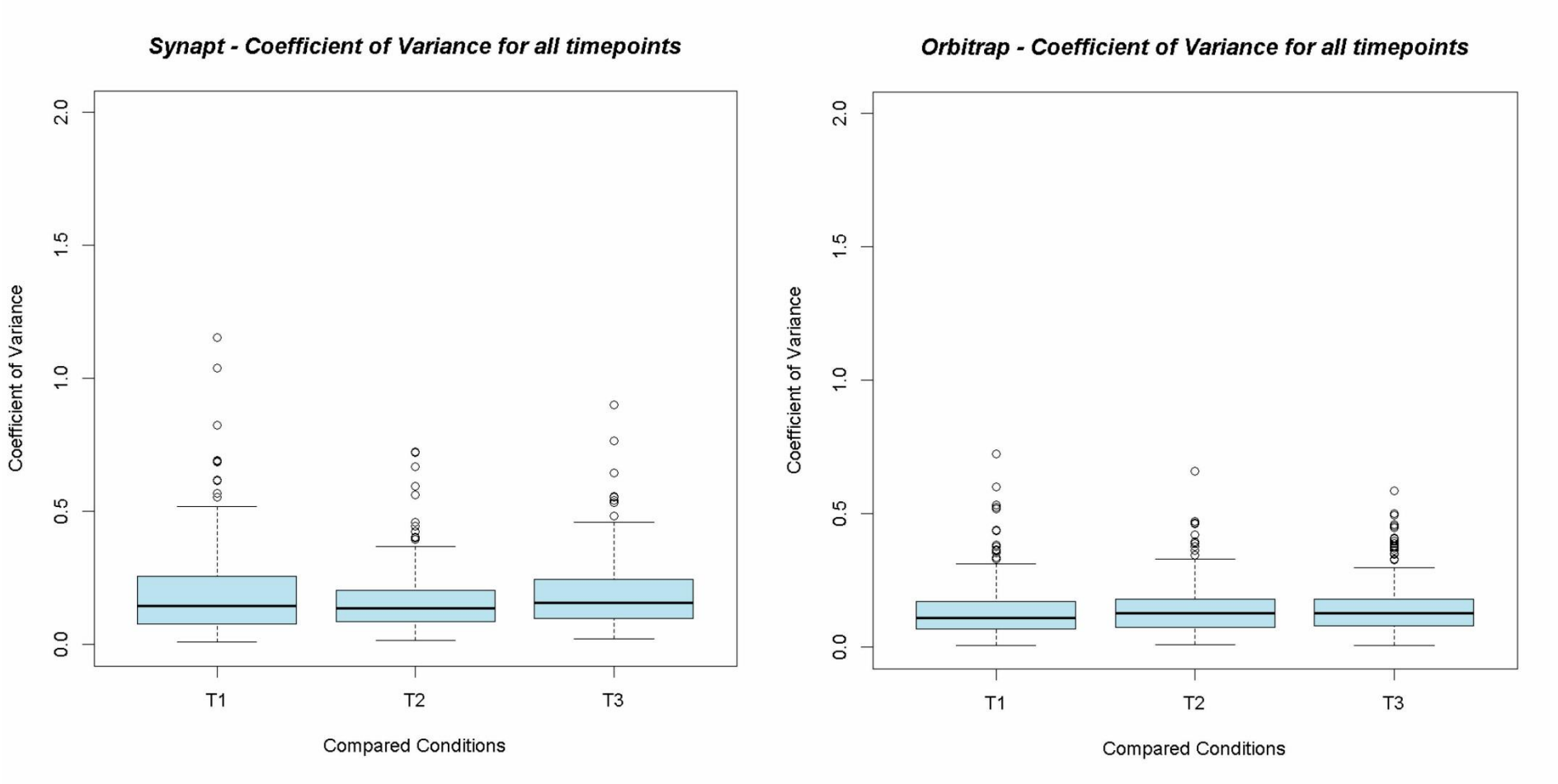

Figure 28: Box plots of coefficient of variance for median absolute deviation normalised protein data at all time points

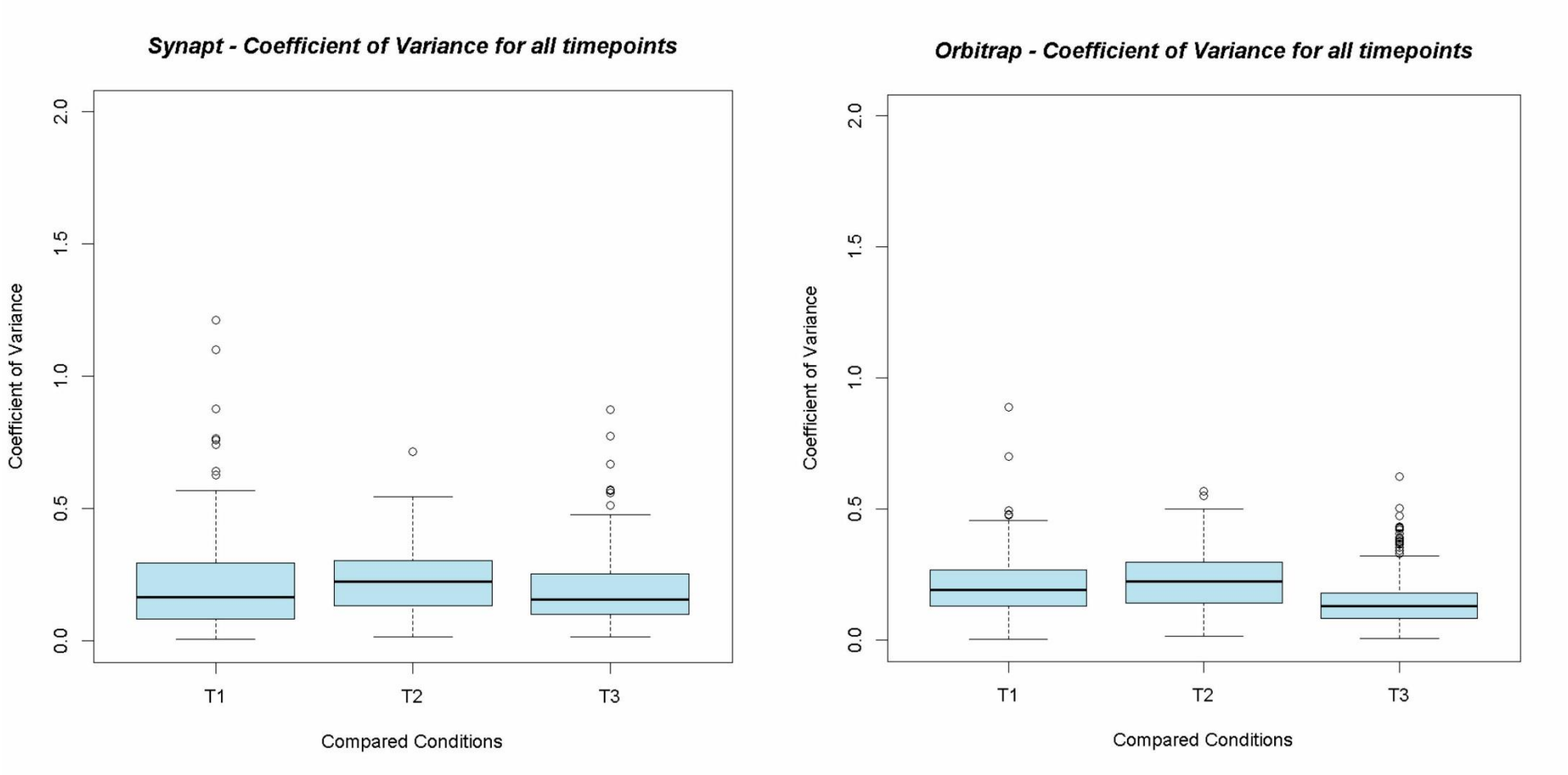

Figure 29: Box plots of coefficient of variance for raw protein data at all time points

3.4 - Discussion and Conclusions

This study was designed to investigate the correlation of the data obtained from the analysis of identical biological samples on two different mass spectrometry instruments. Data from different instruments is often compared in the literature, if not directly, then in the form of the conclusions made based on that data. However there are few studies which look at the direct comparison of results when biologically identical samples are run on different instruments. Though the two instruments used in this study are both considered to be high accuracy, there is potential for variation due to instrumental differences. The Waters Synapt G2 may be expected to give higher accuracy quantitation data as quantitation in time-of-flight instruments is achieved via direct ion counting, rather than through a Fourier transform as in the Thermo Scientific Orbitrap VELOS. However, quantitation accuracy is also improved through increasing the number of peptides observed, and an orbitrap gives increased sensitivity due to the ion trapping and accumulation stage prior to detection.

These results show that, for these two instruments at least, the correlation between the results is high at the protein level, while it is somewhat lower at the feature level. The application of score thresholds (i.e. retaining the top 10, 25 or 50% of the data when ordered by peptide score) is not observed to have a predictable effect on the correlation of the two datasets; however the application of abundance thresholds does show a favourable relationship between increased abundance thresholds and a higher Pearson value. While the feature level correlation can be improved considerably by applying abundance thresholds, this is mitigated at the protein level due to the combination of several peptide features into each protein abundance measurement. Therefore it is considered that based on what is seen in this dataset there is little benefit to setting thresholds at the peptide level, while it could be detrimental in terms of losing proteins at the next level because they no longer possess sufficient quantitation peptides to be retained following the application of peptide thresholds.

The number of proteins identified by each instrument was also investigated, and it was seen that the Waters Synapt G2 identifies fewer proteins than the Thermo Orbitrap VELOS – 1792 vs 720 proteins (2.49 times the number) and that the majority of these proteins are common to the two instruments. These numbers are reduced for both instruments when a three peptide threshold is applied, to 679 vs 264 proteins for the Thermo Orbitrap VELOS and the

Waters Synapt G2 data respectively, but the ratio between the two instruments remains approximately the same (the number of peptides identified by the Thermo Orbitrap VELOS being 2.57 times greater than that for the Waters Synapt G2).

The assessment of the coefficient of variance for both instruments at all time points shows a greater variance in the Waters Synapt G2 data, though the median values for the average coefficient of variance are comparable between the two instruments. It was observed from the creation of box plots that normalisation has a favourable effect on the coefficient of variance observed for both instruments, and has a greater effect on the Thermo Orbitrap VELOS data.

It is concluded that for this data at least, there is little benefit to the application of score thresholds at either the feature or the protein level due to the unpredictable effect of their application on the correlation between the data from the two instruments. The most beneficial threshold to apply for this data seems to be a 25% abundance threshold at the protein level, which gives a positive effect on the correlation between instruments and a tolerable reduction in the number of identified proteins. Median absolute deviation normalisation is seen to reduce the instrument dependent variation seen between the separate biological replicates studied for each time point, and is therefore recommended. Hi3 data is seen to yield a higher Pearson value than total abundance data suggesting that the hi3 method is preferable, at least for this data set.

As future work it would be beneficial to design or acquire a dataset composed of more samples, which had been run on a greater variety of instruments with identical chromatographic conditions and with at least three technical replicates. It would also be beneficial to use multiple spike ins at various protein concentrations (possibly as a protein mix spike in, or by using a weighted sample containing multiple protein populations such as that described earlier in this thesis), in order to make an assessment of accuracy and reproducibility between instruments where the ground truth is fully known. Ideally this analysis would be completed using the most up-to-date instruments, for example using a Thermo Scientific Fusion (quadrupole-iontrap-orbitrap tribrid instrument) versus Waters SYNAPT G2si (TOF). It would also be useful to conduct an analysis with identical chromatographic conditions, as it is likely that one set of the conditions used in this work is more suited to this particular sample and it is not possible to dismiss the use of different

chromatographic gradients as the source of some of the variation between instruments that was observed in this study. However, though identical chromatographic conditions are ideal to create fully comparable results, it can be argued that the comparison as presented is more relevant to the field as biological conclusions compared in the literature will arise from the optimised conditions used in each individual laboratory. While a difference in chromatographic conditions may impact peptide-level identifications, this effect should be mitigated when peptide data is taken up to the protein-level. For those peptides which are identified, peptide quantitation should be broadly unaffected by any difference in the chromatographic conditions used.

4 - The effect of the identification thresholds used on the results obtained from label-free data

4.1 - Introduction

In the software comparison study reported above (Chapter Two, Page 36), it was seen that the use of different score thresholds (e.g. the protein score reported by Mascot) when importing *identification* results had a large effect on the final *quantitative* output from Progenesis LC-MS. The study presented below is intended to extend this observation further and investigate the use of multiple peptide-level FDR thresholds applied to Mascot-derived-search results. It is also intended to briefly test the hypothesis that the use of multiple database search engines will improve the results of the analysis, through the use of Mascot, OMSSA and X!Tandem data post processed with Scaffold. Scaffold is a protein inference viewing and validation software (ProteomeSoftware)[110] capable of accepting identification data from multiple search engines, and allowing multiple peptide and protein level thresholds to be applied within the user interface.

4.2 - Methods

This study was conducted using the CPTAC data files that have been analysed using Progenesis LC-MS as described above (Chapter Two, page 36), with the Mascot result file from that work being used for the Mascot section of the study presented herein.

Firstly the effect of altering the protein identification thresholds used was assessed, with all thresholds set within the Mascot output .csv file (processed manually in Excel). To do this, the data was sorted by Mascot score and a manual peptide FDR calculated based on the number of concatenated target-decoy accessions assigned (counting each reverse accession as a false positive and calculating the FDR from FDR = FP / FP + TP). It was then possible to pick out the Mascot score that corresponded to the desired FDR threshold (0, 1, 5, 10 and 20% FDR thresholds were used), and these Mascot score thresholds were then applied at the import stage to Progenesis LC-MS when importing the Mascot .xml results file. This import stage and post processing was then repeated for each Mascot score threshold in order to

obtain output protein list files (containing quantitative protein abundance values) from Progenesis LC-MS which correspond to each of the above FDR thresholds.

For the second part of this study result files were obtained that had been generated by searching the data with OMSSA and X!Tandem, using analogous search parameters to the Mascot search (allow 1 missed cleavage, fixed modifications: Carbamidomethyl (C), variable modifications: Oxidation (M), Peptide Tolerance: +/-10ppm, MS-MS tol. +/-0.6Da). The Mascot, OMSSA and X!Tandem data files were then imported into Scaffold (Version 4, Proteome Software)[110] as three biosamples, and with the auto-parse option selected to import the .fasta database file. Within Scaffold a low (5%) protein probability threshold was set, with the intention that any observed variation could then be assumed to arise from the changing the peptide FDR threshold. This peptide FDR threshold was set at 0.1, 0.5, 1.0, 2.0 and 5.0, and a result file exported for each threshold. These spectrum report result files were then imported into Progenesis LC-MS. A further analysis was also conducted using a 1% peptide FDR threshold and varying the protein FDR as it was observed that this stringent peptide threshold is required to remove unreliable data from the results.

No manual filters were applied while importing Scaffold result files to Progenesis LC-MS, as thresholds had already been applied to the search result files. Protein list files were then exported for further analysis in Excel. Protein abundance ratios were calculated for each comparison (E/B, E/C and E/D) and the lists ordered to filter the UPS1 proteins to the top of the sheet. The ratio lists were then copied into new .csv files as input to in-house R code in order to generate box plots of the data.

4.3 - Results

4.3.1 - Manual peptide FDR thresholds applied to data searched using Mascot

Box plots showing abundance ratios for the UPS1 proteins in all comparisons (E/B, E/C and E/D) at five peptide FDR thresholds (0, 1, 5, 10 and 20% peptide FDR) are shown below in Figure 28. The results from this analysis are interesting, as it appears that the most stringent peptide FDR threshold does not yield the most accurate median protein abundance ratios, at least for the UPS1 proteins present in this dataset. In fact the "best" results (in terms of the accuracy of median protein abundance ratios and minimal spread of the data) were found when a threshold of 1% peptide FDR was applied, giving median values of 28.81, 10.74 and 3.97 compared to the expected ratios of 27, 9 and 3 respectively. Generally, the E/C and E/D ratios are overestimated while the E/B ratio is underestimated (the two exceptions are the E/B ratio at 1% FDR that is overestimated and the E/C ratio at 20% FDR that is slightly underestimated). The suppression of the ratios at higher peptide FDR thresholds may be explained by the inclusion of lower abundance peptides in the estimation of protein abundance – either low abundance peptides which are incorrectly assigned to that protein, or which are correctly assigned but which are not fully ionised and/or detected within the mass spectrometer.

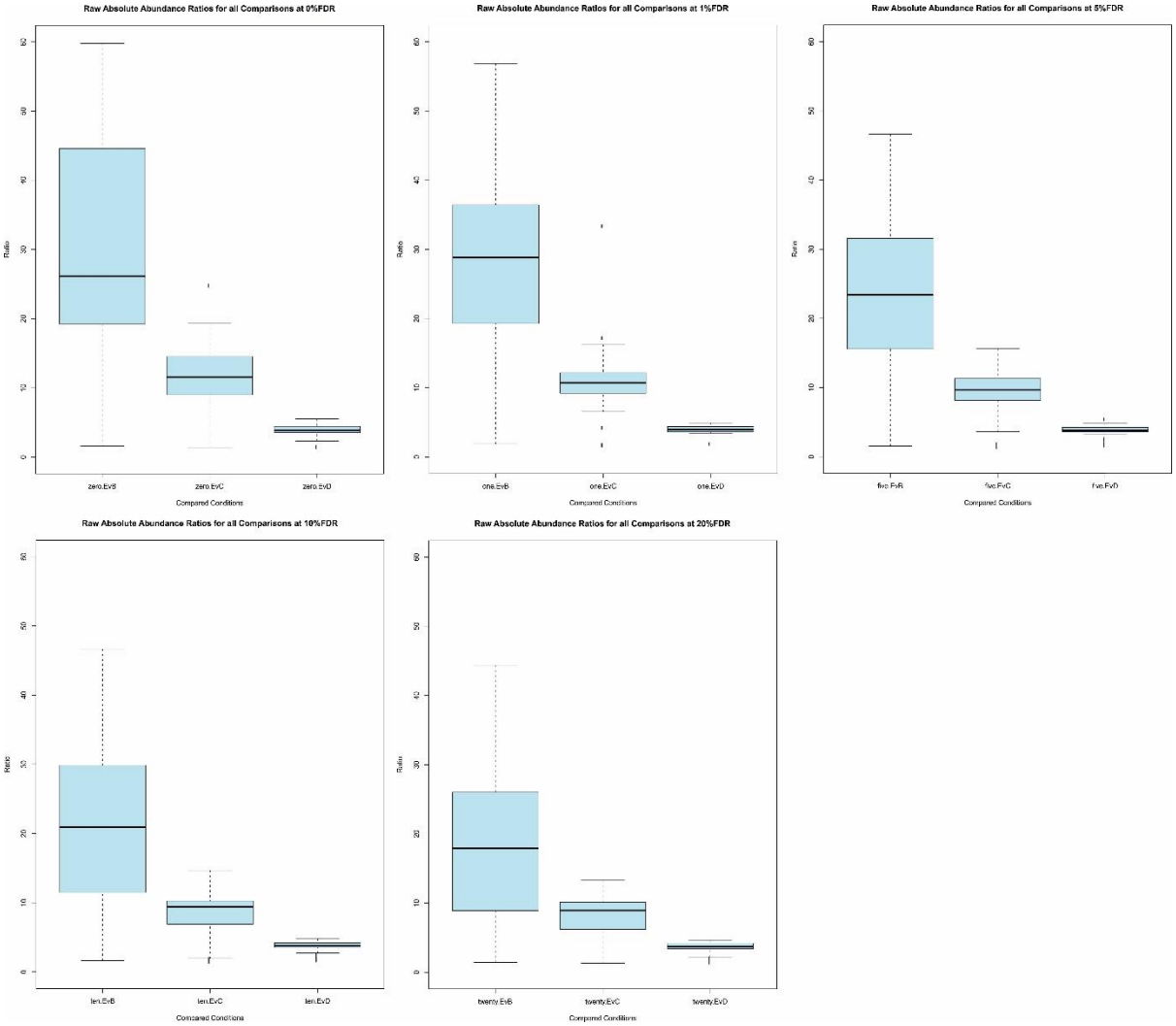

Figure 30: Box plots showing abundance ratios for all comparisons analysed using multiple peptide FDR thresholds with Mascot data in Progenesis LC-MS.

Mascot Data only - UPS1 protein abundance ratios										
	0.0FDR		1.0FDR		5.0FDR		10.0FDR		20.0FDR	
	median	IQ range	median	IQ range	median	IQ range	median	IQ range	median	IQ range
EvB	26.11	25.3	28.81	17.16	23.43	15.99	20.89	18.38	17.91	17.11
EvC	11.53	5.51	10.74	2.96	9.63	3.11	9.43	3.34	8.92	3.91
EvD	3.91	0.92	3.97	0.78	3.84	0.65	3.81	0.68	3.73	0.82

Table 10: Table of median and interquartile-range values for all ratio comparisons at 0.0, 1.0, 5.0, 10.0 and 20.0% peptide FDR thresholds.

The trade off in terms of the number of identified UPS1 proteins was also considered (see Table 11, middle column), and it was seen that the highest proportion of the UPS1 proteins present was identified at the highest peptide FDR threshold. This is as would be expected, and in fact all of the UPS1 proteins were identified at both the 10 and 20% peptide FDR thresholds. However only 45 and 46 proteins were identified respectively when a threshold requiring two or more peptides used for quantitation of the protein was applied. At 1% peptide FDR 44 of the UPS1 proteins were identified, with 41 proteins passing the two peptide threshold, giving sensitivity values of 0.92 and 0.85 respectively. It is considered that this is a reasonable trade off. The total number of yeast proteins reported was also studied (see Table 11, right hand column), and the expected correspondence of increased peptide FDR threshold with an increased number of identified proteins was observed (however it was observed that a high peptide threshold is required to remove unreliable data as hundreds of concatenated target-decoy proteins are reported in the result sheet, and setting different protein probability thresholds has very little affect in terms of reducing the number of decoy identifications, therefore it should be noted that using these thresholds there are unsuitably high numbers of false positives for use in biological study).

FDR threshold (%)	no. of UPS1 proteins (no. passing 2 quant peptide threshold)	no. of yeast proteins (no. passing 2 quant peptide threshold)
0	33 (22)	702 (386)
1	44 (41)	1308 (747)
5	47 (45)	2418 (1126)
10	48 (45)	2686 (1826)
20	48 (46)	2735 (2432)

Table 11: Table showing the number of UPS1 and yeast proteins identified at all FDR thresholds

4.3.2 - Peptide FDR thresholds applied within Scaffold using Mascot, OMSSA and X!Tandem search sesults

On considering the results following multiple search engine analysis and post processing with Scaffold using a 5.0% protein probability and 0.1, 0.5, 1.0, 2.0 and 5.0% peptide FDR thresholds, it is seen that all 48 UPS1 proteins are identified at all peptide FDR thresholds, with 46 proteins passing the two peptide threshold (see Figure 31 and Table 12).

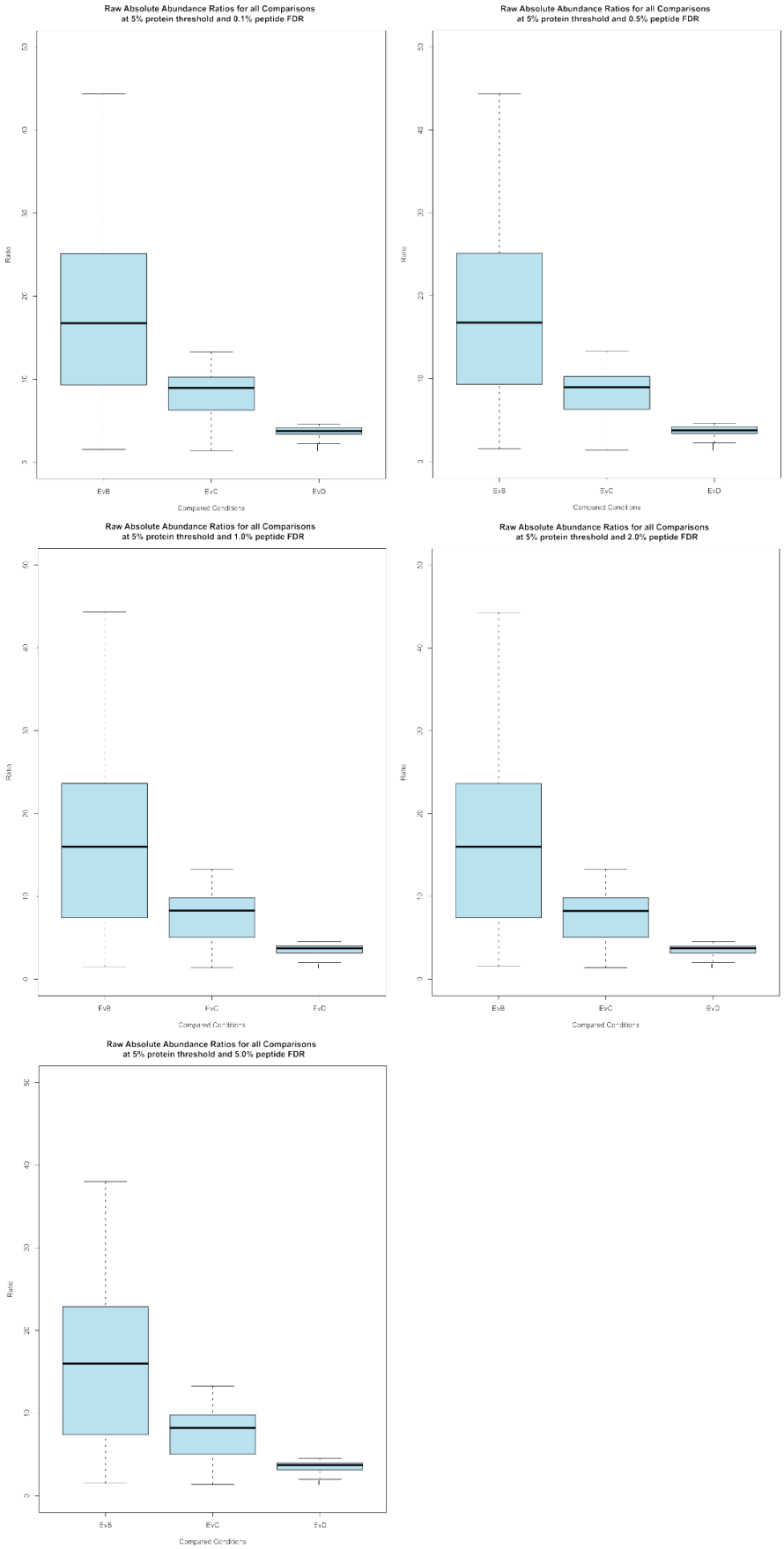

Figure 31: Box plots showing abundance ratios for all comparisons analysed using multiple peptide FDR thresholds in Scaffold with Mascot, OMSSA and X!Tandem data in Progenesis LC-MS

Scaffold Data at 5% Protein Probability - peptide FDR thresholds					
	0.1%FDR	0.5%FDR	1.0%FDR	2.0%FDR	5.0%FDR
EvB	16.75	16.75	16.1	16.1	16.1
EvC	8.95	8.95	8.28	8.24	8.22
EvD	3.75	3.75	3.74	3.73	3.73

Table 12: Table showing median values for all ratio comparisons between conditions (EvB, EvC and EvD) at 5% protein probability and 0.1, 0.5, 1.0, 2.0 and 5.0% peptide FDR thresholds.

Once a stringent peptide FDR threshold is applied (1% peptide FDR, see Table 13) the median ratios reported are far closer to the ground truth (27, 9 and 3 respectively for the EvB, EvC and EvD comparisons), with the results from setting both 1, 2 and 5% protein FDR giving median ratios of 29.43, 10.40 and 4.03. Even at 10% protein FDR the ratios are still close to the ground truth for the EvC and EvD comparisons (9.4 and 3.93), but the ratio reported for the EvB comparison is depressed (10.27). At high protein FDR (1 and 2%) a very small number of decoy hits was observed (4 decoy hits), with only a slight increase in the number of decoy hits at 5% protein FDR (27 decoy hits). The number of decoy hits then increased to over 100 when a 10% protein FDR threshold was used.

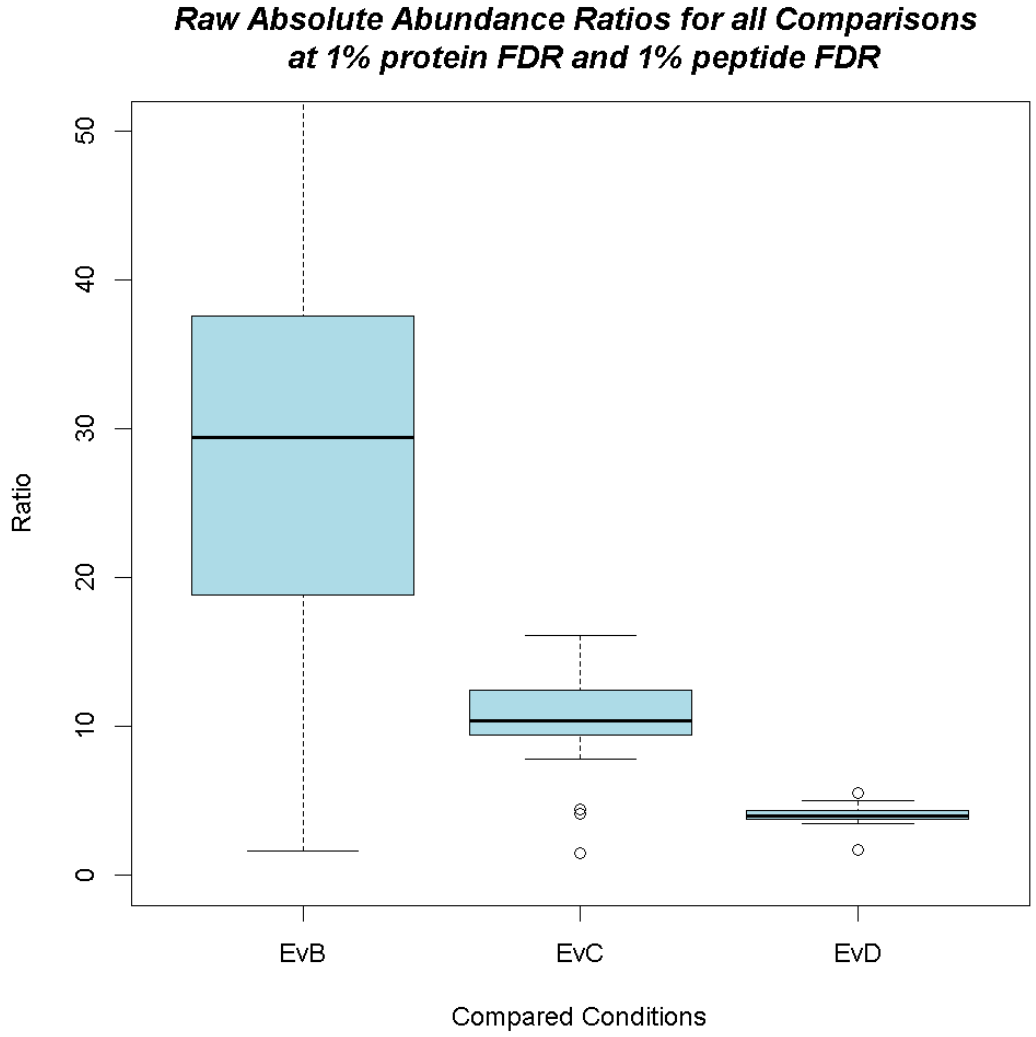

Figure 32: Box plot showing abundance ratios for all ratio comparisons analysed using 1% peptide FDR and 1% protein FDR thresholds in Scaffold with Mascot, OMSSA and X!Tandem data and post processed in Progenesis LC-MS

Scaffold Data - UPS1 protein abundance ratios								
	1.0%FDR		2.0%FDR		5.0%FDR		10.0%FDR	
	median	IQ range	median	IQ range	median	IQ range	median	IQ range
EvB	29.43	18.79	29.43	18.79	29.43	19.19	19.27	16.95
EvC	10.4	2.97	10.4	2.97	10.4	2.93	9.4	3.08
EvD	4.03	0.61	4.03	0.61	4.03	0.59	3.93	0.57

Table 13: Table showing median and inter-quartile range values for all ratio comparisons between conditions (EvB, EvC and EvD) at 1% peptide FDR with 1, 2, 5 and 10% protein FDR thresholds.

4.4 - Discussion and Conclusions

This was a very brief study due to time constraints, however the results from the Mascot analysis demonstrate that the relationship between the reliable identification of proteins and their reliable quantitation is extremely non-trivial and that further study in this area would be beneficial to the field. For this data at least, the suggestion is that a 1% peptide FDR threshold (applied to the data when ordered by the reported Mascot score) is most beneficial in terms of obtaining accurate protein abundance ratios. It would be useful to repeat the analysis using a sample where a greater proportion of the proteins present were of known abundance, or possibly with a simple sample containing only a small number of proteins.

The analysis using Scaffold illustrated the importance of selecting the correct thresholds, with a seemingly sensible strategy giving highly erroneous results in terms of the median ratio values reported. Using a low protein probability threshold gives rise to high FDR and many decoy identifications, and altering the peptide FDR has little effect in terms of reducing the FDR. It is observed that the most sensible strategy appears to be the use of a 1% peptide FDR threshold and a protein FDR threshold of 1-5%, as this gives median ratios closest to the ground truth for this dataset, and with very low numbers of decoy hits.

5– Discussion

5.1– Project Overview

5.1.1 – Achievements

The scope of this project was to consider the methods of analysis that are available to researchers in the field of quantitative label-free proteomics and possible methods to increase the confidence of experimental results through the parallel use of multiple post processing software pipelines. To this end comparisons were made between popular software pipelines, between mass spectrometry instruments, and between different identification thresholds. Out of these comparisons arose suggestions of methods that can be used to increase the confidence in both the identification of proteins and the differential expression of those proteins between different experimental conditions.

5.1.2 – Key points highlighted by this project

5.1.2.1 - There is a great need for representative standard datasets where the ground truth is known, and for these datasets to be made available to the bioinformatics community via data repositories, so as to allow an increase in the number of benchmarking studies looking at methods to increase to confidence and reliability of data obtained from various software pipelines and experimental or post processing methods. In particular, there is a need to ensure that the "background proteins" in standard samples where a spike in has been applied for comparison are truly homologous across all replicates of the sample, as variation in this background causes uncertainty in the assignment of differential expression profiles (ie assignment of differential expression for "background proteins" cannot be inferred as a false positive when there are genuine changes in their abundance between samples).

5.1.2.2 - Global normalisation methods do not appear to be suitable for the analysis of datasets where there is more than one distribution in the data, for example when both host and parasite cells are present, or when proteins are spiked into a background sample. The skewing of differential expression ratios resulting from the application of global normalisation to datasets containing multiple distributions becomes an important consideration for those scientists studying such host-parasite or similar systems, and a

possible method to manage this is suggested within this thesis – namely to apply normalisation to each set of proteins individually (ie to host proteins and parasite proteins) – and this has been used successfully by colleagues on real data.

5.1.2.3 - While statistical post-processing, for example using the QPROT tool, can be very successful in terms of improved sensitivity and reduced FDR, the choice of identification threshold prior to quantitative analysis can also have a great impact on the sensitivity and FDR of the quantitative results. Therefore if possible these input thresholds should be optimised for the dataset in question, and from the studies presented in this thesis an input threshold of 1% peptide level FDR appears to be the most effective in terms of outputting the correct ratios between conditions and maintaining sensitivity without reducing the number of identified proteins past what is acceptable.

5.1.2.4 - The parallel analysis of a single biological sample on different instrument platforms yields intensity values that are well correlated at the protein level, with slightly lower correlation at the feature level, and this correlation is improved (at least for the dataset studied) by using hi3 data as opposed to using total abundance data. This may be due to a combination of the most abundant peptides being most likely to be present at the same abundance as the parent protein, and the possibility of low abundance peptides skewing or confusing the assigned abundance of the parent protein. Also, while the application of quartile thresholds by abundance is efficacious in improving the correlation between instruments (particularly at the feature level as the effect is somewhat masked at the protein level), the effect of score thresholds is less predictable, with an increased score threshold actually reducing the correlation at the feature level in some cases. Presumably this is due to the application of thresholds above an (unknown) optimum point removing true positives from the result list.

5.1.3 – Limitations

The intention throughout this project was to use standard procedures for all software packages, and therefore it is possible that the results obtained could be improved if the methods used were further optimised for the data being considered. However, the use of standard procedures was considered more representative of the situation in which a lab scientist is working with experimental data where the ground truth is unknown and therefore the opportunity for optimisation is highly limited.

In the assessment of correlation between instruments there are some differences in the standard chromatographic conditions for two instruments, and in addition it was necessary

to perform initial analysis and protein inference for the two datasets in separate software environments due to vendor software being necessary to analyse the MS^E data type. Therefore it is not possible to exclude these differences as a potential cause of the variation observed between the data output from the two instruments.

5.1.4 - Suggestions for future work

For both the software and instrument comparison studies it would be beneficial to repeat the studies with a dataset that has a robust background that is truly homologous across all conditions, and which contains multiple spike-ins or protein populations of known abundance and/or at known ratios between conditions.

For the instrument study it would also be useful to repeat the study using two or more instruments that produce data formats that allow truly parallel post-processing, and using identical chromatographic parameters. It would also be useful to perform the analysis using different samples to ensure that the conclusions drawn are not specific to the single biological system studied.

5.2 – Relevance to the field

While there are some limitations in the results presented here, important considerations are raised in relation to data normalisation and threshold selection. Suggestions are made for the use of a simple spectral counting technique (such as emPAI) in tandem with an intensity based technique and additional statistical post-processing to improve sensitivity and reduce FDR in quantitative results. There is also confirmation for the core assumption that abundance data obtained from different instrument platforms is broadly correlated is correct, at least for the instruments used and for the dataset studied as part of this thesis. The work presented here also clearly demonstrates the need for freely available and truly representative standard datasets, as it is necessary for there to be confidence in the ground truth of the dataset to allow reliable conclusions to be made regarding the effect on sensitivity and FDR when using different thresholds and post processing methods.

6 - Acknowledgements

I would like to thank my supervisor Dr Andy Jones for all his help and encouragement, and all my colleagues for their assistance and support with this project (particularly those who advised me through my programming efforts).

7 - Bibliography

1. Choi, H., et al., *QPROT: statistical method for testing differential expression using protein-level intensity data in label-free quantitative proteomics.* Journal of Proteomics, 2015.
2. *QPROT tool: Decription and Download Page.* http://sourceforge.net/projects/qprot/.
3. Jones, A.R., et al., *Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines.* PROTEOMICS, 2009. **9**(5): p. 1220-1229.
4. Riffle, M., L. Malmström, and T.N. Davis, *The Yeast Resource Center Public Data Repository.* Nucleic Acids Research, 2005. **33**(Database issue): p. D378-D382.
5. Jones, P., et al., *PRIDE: a public repository of protein and peptide identifications for the proteomics community.* Nucl. Acids Res., 2006. **34**(suppl_1): p. D659-663.
6. Fenyö, D., J. Eriksson, and R. Beavis, *Mass Spectrometric Protein Identification Using the Global Proteome Machine.* Methods in molecular biology (Clifton, N.J.), 2010. **673**: p. 189-202.
7. Desiere, F., et al., *The PeptideAtlas project.* Nucl. Acids Res., 2006. **34**(suppl_1): p. D655-658.
8. *massIVE*. [cited 2015 10.08]; Available from: http://massive.ucsd.edu/ProteoSAFe/datasets.jsp.
9. Falkner, J.A., J.A. Hill, and P.C. Andrews, *Proteomics FASTA archive and reference resource.* Proteomics, 2008. **8**(9): p. 1756-1757.
10. *ProteomeXchange*. Available from: http://www.proteomexchange.org/.
11. Tyers, M. and M. Mann, *From genomics to proteomics.* Nature, 2003. **422**(6928): p. 193-197.
12. Fröhlich, T. and G.J. Arnold, *Proteome research based on modern liquid chromatography--tandem mass spectrometry: separation, identification and quantification.* Journal of Neural Transmission, 2006. **113**(8): p. 973-994.
13. Mawuenyega, K.G., et al., *Large-scale identification of Caenorhabditis elegans proteins by multidimensional liquid chromatography-tandem mass spectrometry.* Journal of proteome research, 2003. **2**(1): p. 23-35.
14. Peng, J., et al., *Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC–MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome.* Journal of Proteome Research, 2002. **2**(1): p. 43-50.
15. Lau, K.W., et al., *Capture and Analysis of Quantitative Proteomic Data.* Proteomics, 2007. **7**(16): p. 2787-2799.
16. Beynon, R.J., *The dynamics of the proteome: Strategies for measuring protein turnover on a proteome-wide scale.* Briefings in Functional Genomics & Proteomics, 2005. **3**(4): p. 382-390.

17. Hernandez-Castellano, L.E., et al., *Colostrum protein uptake in neonatal lambs examined by descriptive and quantitative liquid chromatography-tandem mass spectrometry.* Journal of Dairy Science, 2015. **98**(1): p. 135-147.
18. Pailleux, F. and F. Beaudry, *Internal standard strategies for relative and absolute quantitation of peptides in biological matrices by liquid chromatography tandem mass spectrometry.* Biomedical Chromatography, 2012. **26**(8): p. 881-891.
19. Barr, J.R., et al., *Isotope dilution--mass spectrometric quantification of specific proteins: model application with apolipoprotein A-I.* Clinical Chemistry, 1996. **42**(10): p. 1676-1682.
20. Gerber, S.A., et al., *Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS.* Proceedings of the National Academy of Sciences, 2003. **100**(12): p. 6940-6945.
21. Pratt, J.M., et al., *Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes.* Nat. Protocols, 2006. **1**(2): p. 1029-1043.
22. Anderson, N.L., et al., *Mass Spectrometric Quantitation of Peptides and Proteins Using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA).* Journal of Proteome Research, 2004. **3**(2): p. 235-244.
23. *Molecular Biology*. 2015.
24. Medzihradszky, K.F. and R.J. Chalkley, *Lessons in de novo peptide sequencing by tandem mass spectrometry.* Mass Spectrometry Reviews, 2015. **34**(1): p. 43-63.
25. Juraschek, R., T. Dülcks, and M. Karas, *Nanoelectrospray—more than just a minimized-flow electrospray ionization source.* Journal of the American Society for Mass Spectrometry, 1999. **10**(4): p. 300-308.
26. Gross, J.H., *Mass Spectrometry: A Textbook*. 1st ed. 2002: Springer.
27. Fernández de la Mora, J., *The Fluid Dynamics of Taylor Cones.* Annual Review of Fluid Mechanics, 2007. **39**(1): p. 217-243.
28. *Applied Biosystems API 4000 LC/MS/MS System Hardware Manual.* Doc. No. D1000013652C, 2002.
29. Kebarle, P. and U.H. Verkerk, *Electrospray: From ions in solution to ions in the gas phase, what we know now.* Mass Spectrometry Reviews, 2009. **28**(6): p. 898-917.
30. Thomson, J.V.I.a.B.A., *On the evaporation of small ions from charged droplets.* J. Chem. Phys., 1976. **64**(6).
31. *Finnigan™ H-ESI™ Probe Operator's Manual.* 2005. **97055-97045**.
32. *TSQ™ Quantum Access™ Hardware Manual.* 2006. **70111-97133 Revision B**.
33. Kelstrup CD, H.O., Francavilla C, Olsen JV, *Pinpointing phosphorylation sites: Quantitative filtering and a novel site-specific x-ion fragmen.* J Proteome Res, 2011. **10**: p. 2937–2948.
34. Simón-Manso Y, N.P., Yang X, Stein SE, *Loss of 45 Da from a2 ions and preferential loss of 48 Da from a2 ions containing methionine in peptide ion tandem mass spectra.* J Am Soc Mass Spectrom, 2011. **22**: p. 280–289.
35. Kilpatrick LE, N.P., Yang X, Simón-Manso Y, Liang Y, Stein SE, *Formation of y + 10 and y + 11 ions in the collision-induced dissociation of peptide ions.* J Am Soc Mass Spectrom, 2012. **23**: p. 655–663.
36. Medzihradszky KF, T.J., *Unusual fragmentation of Pro-Ser/Thr-containing peptides detected in collision-induced dissociation spectra.* J Am Soc Mass Spectrom, 2012. **23**: p. 602–607.
37. *Waters Synapt Mass Spectrometry System Operators Guide.* 2008(71500153502 Revision A).
38. *2002 Bruker Daltonics MicrOTOF Q User Manual* 2007. **Version 1.1**.
39. *Thermo LTQ Hardware Manual.* 2005. **97055-97013 Revision B**.

40. Hardman, M. and A.A. Makarov, *Interfacing the orbitrap mass analyzer to an electrospray ion source.* Analytical Chemistry, 2003. **75**(7): p. 1699-1705.
41. Scigelova, M. and A. Makarov, *Orbitrap Mass Analyzer – Overview and Applications in Proteomics.* PROTEOMICS, 2006. **6**(S2): p. 16-21.
42. *Thermo LTQ Orbitrap XL Hardware Manual.* 2008. **1225830 Revision B**.
43. Johnstone, C.G.H.a.R.A.W., *Mass Spectrometry Basics*. 2002.
44. Laboratory, P.N.N. *Isotope Pattern Calculator*. [cited 2015 11.08]; Available from: OMICS.PNL.GOV.
45. Thiede, B., et al., *Peptide mass fingerprinting.* Methods, 2005. **35**(3): p. 237 - 247.
46. Beavis, R.C., B.T. Chait, and K.G. Standing, *Factors affecting the ultraviolet laser desorption of proteins.* Rapid Communications in Mass Spectrometry, 1989. **3**(7): p. 233-237.
47. McLafferty, F., et al., *An Enlarged Data Base of Electron-Ionization Mass Spectra.* Journal of the American Society for Mass Spectrometry, 1991. **2**(5): p. 432-437.
48. Pappin, D.J.C., *Rapid Identification of proteins by peptide-mass fingerprinting.* Current Biology, 1993. **3**: p. 327-332.
49. Rodriguez, J., et al., *Does Trypsin Cut Before Proline?* Journal of Proteome Research, 2007. **7**(1): p. 300-305.
50. Chiva, C., M. Ortega, and E. Sabidó, *Influence of the digestion technique, protease, and missed cleavage peptides in protein quantitation.* Journal of proteome research, 2014. **13**(9): p. 3979-3986.
51. Mann, M. and M. Wilm, *Electrospray mass spectrometry for protein characterization.* Trends in Biochemical Sciences, 1995. **20**(6): p. 219-224.
52. Yu, C., et al., *Classifying b and y ions in peptide tandem mass spectra*, in *Proceedings of the 6th international conference on Fuzzy systems and knowledge discovery - Volume 5*. 2009, IEEE Press: Tianjin, China. p. 37-41.
53. Krebs, I., et al., *A software solution automatically assigns formulae for construction of fragmentation pathways accelerating drug elucidation with ESI-TOF.* LC GC Europe, 2008: p. 31-33.
54. Nesvizhskii, A.I., et al., *A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry.* Anal. Chem., 2003. **75**(17): p. 4646-4658.
55. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nat Meth, 2007. **4**(3): p. 207-214.
56. Doherty, M.K., et al., *Proteome dynamics in complex organisms: Using stable isotopes to monitor individual protein turnover rates.* Proteomics, 2005. **5**(2): p. 522-533.
57. Wiese, S., et al., *Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research.* PROTEOMICS, 2007. **7**(3): p. 340-350.
58. Gygi, S.P., et al., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.* Nat Biotech, 1999. **17**(10): p. 994-999.
59. Zhu, W., *Mass Spectrometry-Based Label-Free Quantitative Proteomics.* Journal of Biomedicine and Biotechnology, 2010. **2010**.
60. America, A.H.P. and J.H.G. Cordewener, *Comparative LC-MS: A landscape of peaks and valleys.* PROTEOMICS, 2008. **8**(4): p. 731-749.
61. Old, W.M., et al., *Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics.* Molecular & Cellular Proteomics, 2005. **4**(10): p. 1487-1502.
62. Liu, H., R.G. Sadygov, and J.R. Yates, *A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics.* Analytical Chemistry, 2004. **76**(14): p. 4193-4201.

63. Gao, J., et al., *Guidelines for the Routine Application of the Peptide Hits Technique.* Journal of the American Society for Mass Spectrometry, 2005. **16**(8): p. 1231-1238.
64. *APEX Quantitative Proteomics Tool Manual.* 2010. **Version 1.1.0**.
65. Lu, P., et al., *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation.* Nat Biotech, 2007. **25**(1): p. 117-124.
66. Xia, D., et al., *The proteome of Toxoplasma gondii: integration with the genome provides novel insights into gene expression and annotation.* Genome Biology, 2008. **9**(7): p. R116.
67. Hubbard, S.J. and A.R. Jones, *Proteome Bioinformatics*. Methods in Molecular Biology. Vol. 604. 2009: Springer.
68. *MapAlignerPoseClustering*. [cited 2015 10.08.2015]; Available from: http://ftp.mi.fu-berlin.de/pub/OpenMS/documentation/html/TOPP_MapAlignerPoseClustering.html.
69. Weisser, H., et al., *An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics.* Journal of Proteome Research, 2013. **12**(4): p. 1628-1644.
70. Hampel, F.R., *The Influence Curve and its Role in Robust Estimation.* Journal of the American Statistical Association, 1974. **69**(346): p. 383-393.
71. Jow, H., R.J. Boys, and D.J. Wilkinson, *Bayesian identification of protein differential expression in multi-group isobaric labelled mass spectrometry data.* Statistical Applications in Genetics & Molecular Biology, 2014. **13**(5): p. 531-551.
72. Mann, J.C., et al., *Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ.* 2014.
73. Stein, P.A.R., et al., *Improved Normalization of Systematic Biases Affecting Ion Current Measurements in Label-free Proteomics Data.* 2014.
74. Shadforth, I., et al., *Confident protein identification using the average peptide score method coupled with search-specific, <I>ab initio</I> thresholds.* Rapid Communications in Mass Spectrometry, 2005. **19**(22): p. 3363-3368.
75. Shadforth, I., et al., *i-Tracker: For quantitative proteomics using iTRAQTM.* BMC Genomics, 2005. **6**(1): p. 145.
76. Shadforth, I., et al., *GAPP: A Fully Automated Software for the Confident Identification of Human Peptides from Tandem Mass Spectra.* J. Proteome Res., 2006. **5**(10): p. 2849-2852.
77. Leptos, K.C., et al., *MapQuant: Open-source software for large-scale protein quantification.* PROTEOMICS, 2006. **6**(6): p. 1770-1782.
78. Lindell, D., et al., *Photosynthesis genes in marine viruses yield proteins during host infection.* Nature, 2005. **438**(7064): p. 86-89.
79. Schulze, W.X. and M. Mann, *A Novel Proteomic Screen for Peptide-Protein Interactions.* Journal of Biological Chemistry, 2004. **279**(11): p. 10756-10764.
80. Ong, S.-E., et al., *Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics.* Mol Cell Proteomics, 2002. **1**(5): p. 376-386.
81. Ong, S.-E., I. Kratchmarova, and M. Mann, *Properties of 13C-Substituted Arginine in Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC).* Journal of Proteome Research, 2002. **2**(2): p. 173-181.
82. Pan, C., et al., *ProRata: A Quantitative Proteomics Program for Accurate Protein Abundance Ratio Estimation with Confidence Interval Evaluation.* Analytical Chemistry, 2006. **78**(20): p. 7121-7131.

83. Ishihama, Y., et al., *Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein.* Molecular & Cellular Proteomics, 2005. **4**(9): p. 1265-1272.
84. Lu, P., et al., *Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation.* Nature Biotechnology, 2007. **25**(1): p. 117-124.
85. Keller, A., et al., *A uniform proteomics MS/MS analysis platform utilizing open XML file formats.* Mol Syst Biol, 2005. **1**: p. 1744 - 4292.
86. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotech, 2008. **26**(12): p. 1367-1372.
87. Dicker, L., X. Lin, and A.R. Ivanov, *Increased Power for the Analysis of Label-free LC-MS/MS Proteomics Data by Combining Spectral Counts and Peptide Peak Attributes.* Molecular & Cellular Proteomics, 2010. **9**(12): p. 2704-2718.
88. Trudgian, D.C., et al., *Comparative evaluation of label-free SINQ normalized spectral index quantitation in the central proteomics facilities pipeline.* PROTEOMICS, 2011. **11**(14): p. 2790-2797.
89. Finney, G.L., et al., *Label-Free Comparative Analysis of Proteomics Mixtures Using Chromatographic Alignment of High-Resolution µLC–MS Data.* Analytical Chemistry, 2008. **80**(4): p. 961-971.
90. Bellew, M., et al., *A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS.* Bioinformatics, 2006. **22**(15): p. 1902-1909.
91. Radulovic, D., et al., *Informatics Platform for Global Proteomic Profiling and Biomarker Discovery Using Liquid Chromatography-Tandem Mass Spectrometry.* Molecular & Cellular Proteomics, 2004. **3**(10): p. 984-997.
92. May, D., et al., *A Platform for Accurate Mass and Time Analyses of Mass Spectrometry Data.* Journal of Proteome Research, 2007. **6**(7): p. 2685-2694.
93. Chen, Y.-Y., et al., *Refining comparative proteomics by spectral counting to account for shared peptides and multiple search engines.* Analytical and Bioanalytical Chemistry, 2012. **404**(4): p. 1115-1125.
94. Zybailov, B., et al., *Correlation of Relative Abundance Ratios Derived from Peptide Ion Chromatograms and Spectrum Counting for Quantitative Proteomic Analysis Using Stable Isotope Labeling.* Analytical Chemistry, 2005. **77**(19): p. 6218-6224.
95. *ABRF 2009 study presentation slides.* http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm.
96. Paulovich, A.G., et al., *A CPTAC inter-laboratory study characterizing a yeast performance standard for benchmarking LC-MS Platform performance.* Molecular & Cellular Proteomics, 2009.
97. *Sigma-Aldrich UPS1 Protein Standard.* http://www.sigmaaldrich.com/life-science/proteomics/mass-spectrometry/universal-proteomics-standard.html.
98. Mead, J.A., L. Bianco, and C. Bessant, *Recent developments in public proteomic MS repositories and pipelines.* PROTEOMICS, 2009. **9**(4): p. 861-881.
99. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning - Data Mining, Inference, | Trevor Hastie | Springer*. 2 ed. Springer Series in Statistics. 2009: Springer-Veriag New York.
100. Cox, J., et al., *Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ.* Molecular & Cellular Proteomics, 2014. **13**(9): p. 2513-2526.

101. *Progenesis LC-MS: How Normalisation Works.* http://www.nonlinear.com/support/progenesis/lc-ms/faq/how-normalisation-works.aspx.
102. Gregori, J., et al., *An effect size filter improves the reproducibility in spectral counting-based comparative proteomics.* Journal of proteomics, 2013. **95**: p. 55-65.
103. Choi, H., D. Fermin, and A.I. Nesvizhskii, *Significance Analysis of Spectral Count Data in Label-free Shotgun Proteomics.* Molecular & Cellular Proteomics, 2008. **7**(12): p. 2373-2385.
104. Searle, B.C., M. Turner, and A.I. Nesvizhskii, *Improving Sensitivity by Probabilistically Combining Results from Multiple MS/MS Search Methodologies.* J. Proteome Res., 2008. **7**(1): p. 245-253.
105. Bern, M., Y. Cai, and D. Goldberg, *Lookup Peaks: A Hybrid of de Novo Sequencing and Database Search for Protein Identification by Tandem Mass Spectrometry.* Analytical Chemistry, 2007. **79**(4): p. 1393-1400.
106. Bern, M.W. and Y.J. Kil, *Two-Dimensional Target Decoy Strategy for Shotgun Proteomics.* Journal of Proteome Research, 2011. **10**(12): p. 5296-5301.
107. Ma, B., et al., *PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry.* Rapid Communications in Mass Spectrometry, 2003. **17**(20): p. 2337-2342.
108. Zhang, J., et al., *PEAKS DB: De Novo sequencing assisted database search for sensitive and accurate peptide identification.* Molecular & Cellular Proteomics, 2011.
109. Qi, D., et al., *A software toolkit and interface for performing stable isotope labelling and top3 quantification using Progenesis LC-MS.* OMICS: A Journal of Integrative Biology, 2012. **16**(9): p. 489-495.
110. Searle, B.C., *Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies.* PROTEOMICS, 2010. **10**(6): p. 1265-1269.