APPLICATION OF LABEL-FREE MASS SPECTROMETRY-BASED PROTEOMICS TO BIOMARKER DISCOVERY

Susan Elizabeth Slade B. Sc. (Hons)

A thesis submitted for the degree of Doctor of Philosophy

University of Warwick
School of Life Sciences

April 2013

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am indebted to my supervisor Professor Jim Scrivens for offering me the opportunity to complete my PhD. study part-time and for his continued encouragement and support in my professional career.

I would like to acknowledge my collaborators Prof. Steve Thornton and Dr. Eamonn Breslin, for their clinical input to the project. In particular, Eamonn for his assistance with sample preparation, at best a tedious task.

Jim Langridge, Hans Vissers, Joanne Connolly, Chris Hughes and in particular Jonathan Fox at Waters Corporation, Manchester for their continued support and assistance with the project.

This work was financially supported by the Prof. Kypros Nicolaides and the Fetal Medicine Foundation, London, UK.

I feel very lucky to have worked with the Scrivens' group, many of whom joined me after completing their second year undergraduate studies looking to take part in the Warwick intercalated year training program. Vib, Charlie, Sarah, Nisha, Elle and Matt, your developing careers confirm the potential I saw in you all those years ago. Kostas, Gill, Jonathon, Fran, George, Baharak and Krisztina have all made working with the Scrivens' group a pleasure with their continued support.

I would like to thank my family, in particular my mother for her love and support over the years having been widowed in her thirties and left with two young children to bring up alone. I apologise for not coming for Sunday lunch for so long due to writing this thesis and thank you for understanding.

To Simon and Emily, for supporting me when so much of my personal life has been devoted to my studies for the last five years but in particular the last 12 months.

Declaration

I hereby declare that this thesis, submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy and entitled "Application of label-free mass spectrometry-based proteomics to biomarker discovery", represents my own work and has not been previously submitted to this or any other institution for any degree, diploma or other qualification. Work undertaken by my colleagues is explicitly stated where appropriate.

Susan E. Slade
April 2013

Summary

Mass spectrometry is an analytical technique which is used extensively in the fields of chemistry and physics. Developments in the field over the last two decades have permitted the analysis of a wide variety of biological molecules from a range of sources. The term proteomics relates to the study of the protein complement of a cell or organism with particular interest in the identification and quantification of these analytes.

A biomarker is a characteristic that can be measured and evaluated to give an indication of normal, biological processes, or pharmacological responses to a therapeutic intervention. Bodily fluids are a rich source of potential biomarkers as they can be obtained in reasonable quantity and their extraction is generally minimally invasive. The plasma proteome, which contains many thousands of analytes spanning a dynamic range greater than 10 orders of magnitude, reflects the status of the many tissues and organs in the body serving as an ideal medium for potential biomarker discovery.

The analytical challenges posed by the plasma proteome are significant. Depletion of the highly abundant proteins is usually a prerequisite of any biomarker study and no technique has the dynamic range to study all of the proteins present. Comprehensive characterisation of the plasma proteome requires significant experimental effort and cost. Use of pooled samples in biomarker studies is widespread and the majority of biomarkers, which have been identified in the discovery phase, have not passed clinical validation.

A data independent, label-free quantitative approach has been evaluated for the study of depleted maternal plasma proteomes taken in the first trimester. Plasma was characterised from individual and groups of patients from three obstetric conditions using single and multi-dimensional chromatography. Potential biomarkers from each source were identified and evaluated.

Multi-dimensional chromatography was used to simplify the complexity of the analytes introduced to the mass spectrometer and the benefits and limitations of the approach in terms of biomarker discovery have been demonstrated.

Abbreviations

1D	One-dimensional
1D RP-LC	One-dimensional reversed phase nano ultra pressure liquid chromatography
2D	Two-dimensional
2D RP-RP-LC	High pH reversed phase – low pH reversed phase nano ultra pressure liquid chromatography
3D	Three-dimensional

A

ACN	Acetonitrile
ADH	Alcohol dehydrogenase
AFP	α-Fetoprotein
AMRT	Accurate mass retention time
APEX	Absolute protein expression
AQUA	Absolute quantitation (peptides for)
AUC	Area under the curve

B

BEH	Bridged ethyl hybrid
BLAST	Basic Local Alignment Search Tool
BMI	Body mass index
BPI	Base peak intensity
BSA	Bovine serum albumin
β-hCG	Human choriogonadotropin subunit beta

C

CID	Collisional-induced decomposition
CV	Coefficient of variation
CVF	Cervical vaginal fluid

D

Da	Dalton
DDA	Data dependent acquisition
DIA	Data independent acquisition
DiGE	Differential in-gel electrophoresis
DNA	Deoxyribonucleic acid
DS	Down’s syndrome

E

E. coli	*Escherichia coli*
ECD	Electron capture dissociation
ELISA	Enzyme-linked immunosorbent assay
EM	Electron multiplier
emPAI	Exponentially modified protein abundance index
ENO	Enolase
ESI	Electrospray ionisation
ETD	Electron transfer dissociation
eV	Electron volt

F

FAB	Fast atom bombardment
FDR	False discovery rate
FFR	Field free region
fmol	Femtomole
FPR	False positive rate
FT	Fourier transform
FT-ICR	Fourier transform ion cyclotron resonance
FWHM	Full width at half maximum

G

g	Centrifugal acceleration
Gb	Gigabyte
GELFrEE	Gel-eluted liquid fraction entrapment electrophoresis
GFP	Glu^{1}-Fibrinopeptide B peptide

H

HILIC	Hydrophilic interaction liquid chromatography
HPLC	High performance liquid chromatography
HPPP	Human Plasma Proteome Project
HSS	High strength silica
HUPO	Human Proteome Organisation

I

ICAT	Isotope-coded affinity tagging
IEF	Isoelectric focussing
IMMS	Ion mobility mass spectrometry
IRMPD	Infrared multi-photon dissociation
iTRAQ	Isobaric tags for relative and absolute quantification
IU	International unit

K

kDa	Kilodalton
kV	Kilovolt

L

L	Litre
LC	Liquid chromatography
LIT	Linear ion trap
LOD	Limit of detection
LOQ	Limit of quantification

M

M	Molar
m/*z*	Mass-to-charge ratio
MALDI	Matrix assisted laser desorption ionisation
MCP	Microchannel plate detector
mg	Milligram
min	Minute
mL	Millilitre
mM	Millimolar
MoM	Multiples of the median
mRNA	Messenger RNA
MRM	Multiple reaction monitoring
MS	Mass spectrometry
msec	Millisecond
MS/MS or MS^n	Tandem mass spectrometry
MS^E	Alternating low/elevated collision energy data independent acquisition
MUDPiT	Multidimensional protein identification technology
µg	Microgram
µL	Microlitre
µm	Micrometre
µM	Micromolar
µsec	Microsecond

N

NaI	Sodium iodide
NanoES	Nanoelectrospray
ng	Nanogram
nm	Nanometre
nM	Nanomolar
ns	Nanosecond
NT	Nuchal Translucency

O

oa	Orthogonal acceleration
OD	Optical density
ORF	Open reading frame

P

PAGE	Polyacrylamide gel electrophoresis
PAI	Protein abundance index
PAPP-A	Pappalysin-1
PhosB	Glycogen phosphorylase B
PLGS	ProteinLynx Global Server
PMF	Peptide mass fingerprinting
ppm	Parts per million
PSG	Pregnancy specific glycoproteins
PTM	Post-translational modification

Q

Q-ToF	Quadrupole time-of-flight
QC	Quality control

R

RF	Radio frequency
RPLC	Reversed phase liquid chromatography
RT	Retention time

S

sec	Second
SCX	Strong cation exchange
SDS	Sodium dodecyl sulphate
SEC	Size exclusion chromatography
SELDI	Surface enhanced laser desorption/ionisation time of flight
SID	Surface-induced dissociation
SILAC	Stable isotope labelling in cell culture
SURUSS	Serum urine and ultrasound screening study

T

TDC	Time-to-digital converter
ToF	Time-of-flight
TPR	True positive rate
T21	Trisomy 21

U

UE_3	Unconjugated oestriol
UPLC	Ultra performance liquid chromatography
UV	Ultra-violet

V

V	Volt

X

XDIA	Extended data-independent acquisition
XIC	Extracted ion current

Chapter One: Introduction

1.1 Mass Spectrometry

In 1906, Joseph John Thomson was awarded the Nobel Prize for Physics "in recognition of the great merits of his theoretical and experimental investigations on the conduction of electricity by gases". His pioneering work on the particles he termed electrons (Thomson 1899) led to the development of the first mass spectrometer, a parabola spectrograph which was used to visualise the isotopes of neon (Thomson 1911). Further developments by Thomson's student Francis W. Aston over the next few years led to an instrument with improved resolving power capable of the study of other non-radioactive isotopes of elements (Aston 1919) providing Aston with the Nobel Prize for Chemistry in 1922. Developments in instrumentation including ionisation sources, increased sensitivity and resolving power continue to take mass spectrometry from a technique used exclusively by chemists and physicists into the field of biology where it has become an essential tool for the study and characterisation of individual biomolecules through to complex systems making it a truly interdisciplinary research tool.

1.1.1 What is a mass spectrometer?

Mass spectrometry is the measurement of the mass-to-charge ratio (*m/z*) of ions in the gas phase. A mass spectrometer consists of a sample inlet which typically may involve a type of chromatography to simplify the mixture of analytes entering the instrument, a source where the gas phase ions are generated, a mass analyser and detector. A computer system is used to control the mass spectrometer and to convert the abundance of the mass/charge separated ions reaching the detector into an accessible and readable format such as a mass spectrum. The sample inlet may be at atmospheric pressure (760 torr) whilst the detector is typically at 10^{-6} torr and so a series of vacuum pumps are required to generate and maintain a vacuum gradient through the instrument.

1.1.2 Ionisation methods

Prior to the 1980's, ionisation techniques in mass spectrometry were limited to electron and chemical impact used on low molecular weight, volatile, thermally stable compounds. With the advent of a fast atom bombardment (FAB) source (Barber, Bordoli et al. 1981), smaller biological molecules were able to be analysed using mass spectrometry without the need for prior chemical derivatisation. It was towards the end of 1980's when two ionisation sources were developed that enabled the field of biological mass spectrometry to develop and still continue to grow today.

In 2002, John Fenn and Koichi Tanaka were awarded the Nobel Prize for Chemistry "for their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules". Tanaka's work on soft laser desorption (Tanaka, Waki et al. 1988) was filed as a patent application in 1985 but the development of matrix assisted laser desorption ionisation (MALDI) by Michael Karas and Franz Hillenkamp (Karas and Hillenkamp 1988) has had a much greater commercial impact.

Electrospray ionisation (ESI) unlike conventional MALDI takes place at atmospheric pressure with the analytes of interest solubilised in a volatile buffer or solvent system (Yamashita and Fenn 1984; Fenn, Mann et al. 1989). Like MALDI, ESI is a soft ionisation technique with its potential impact recognised for biological mass spectrometric applications "also inviting is the prospect of extending the applicability of mass spectrometric analysis to large organic molecules that are too complex, too fragile, or too nonvolatile for ionization by more conventional methods" (Yamashita and Fenn 1984). With the increased sensitivity afforded by the micro-ESI source (Andren, Emmett et al. 1994; Emmett and Caprioli 1994) combined with development of low-µL flow rate liquid chromatography (LC) and high pressure systems (MacNair, Lewis et al. 1997; MacNair, Opiteck et al. 1997; MacNair, Patel et al. 1999) ESI with in-line LC is integral to many proteomics configurations in use today.

The nanoelectrospray ion source (nanoES) was developed by Wilm and Mann and utilised a 1-2 µm spraying orifice achieving flow rates of approximately 20 nL min^{-1},

permitting the analysis of samples for long periods of time (Wilm and Mann 1996). The nanoES source is particularly suited to the analysis of proteins as it is more tolerant of the salts required to maintain non-covalent interactions, data averaging over a period of time can be performed permitting accurate mass measurements and with its stable flow of sample ions optimisation of the interface conditions for the study of non-covalent interactions can be achieved. In nanoES the diameter of the droplets formed are of the order of 200 nm compared to 1-2 μm in ESI, equivalent to 2-3 orders less in volume, and at concentrations of 1 μM each droplet contains on average one analyte molecule (Wilm and Mann 1996). NanoES was used exclusively in this work.

The analyte is passed through a conductive capillary (metal or gold coated borosilicate glass) at high potential typically 1 - 4 kV. Modifiers are typically added to the buffer system to aid the ionisation process and in proteomics applications this is usually formic acid, Figure 1. 1**.**

Under the influence of an electric field, the charged analyte solution forms a Taylor cone and droplets are released. Two main mechanisms have been proposed for the ESI formation of charged gas phase ions. In the ion evaporation model: the solvent evaporates and the size of the droplets reduces but the charge on the droplet remains static. At a given point the Coulombic (charge) repulsion overcomes the surface tension and Coulombic fission results with the cycle repeating until the droplet reaches a critical diameter (<10 nm) when the ions are desorbed directly into the gas phase i.e. ion evaporation replaces Coulombic fission. For molecular species with a low number of charges and *m/z* this mechanism is believed to predominate.

For multiply charged species with increased *m/z* it is accepted that the charge residual model is dominant whereas in the ion evaporation model, repeated cycles of evaporation and Coulombic fission occur. In the charge residual model evaporation continues until all solvent is lost and the charged analyte species enters the gas phase (Kebarle 2000). Evidence suggesting that the charge residual model applies for multiply charged macromolecules was shown by de la Mora and co-workers. Their results indicated that most proteins (6.5 kDa – 1.4 MDa) had charge states where the charge was close to the Rayleigh limit, when the radius of the droplet was similar to

the radius of the protein (Fernandez de la Mora 2000). Other groups have obtained evidence to further validate the charge residual model for native proteins analysed by ESI (Felitsyn, Peschke et al. 2002; Peschke, Blades et al. 2002; Heck and Van Den Heuvel 2004; Peschke, Verkerk et al. 2004). Felitsyn *et al.* concluded that "the formation of charged proteins in the gas phase via charge residual model is very gentle and is expected to lead to relatively small changes of structure in the transition to the gas phase" unlike the ion evaporation model which would disrupt non-covalent interactions that can be observed in ESI.

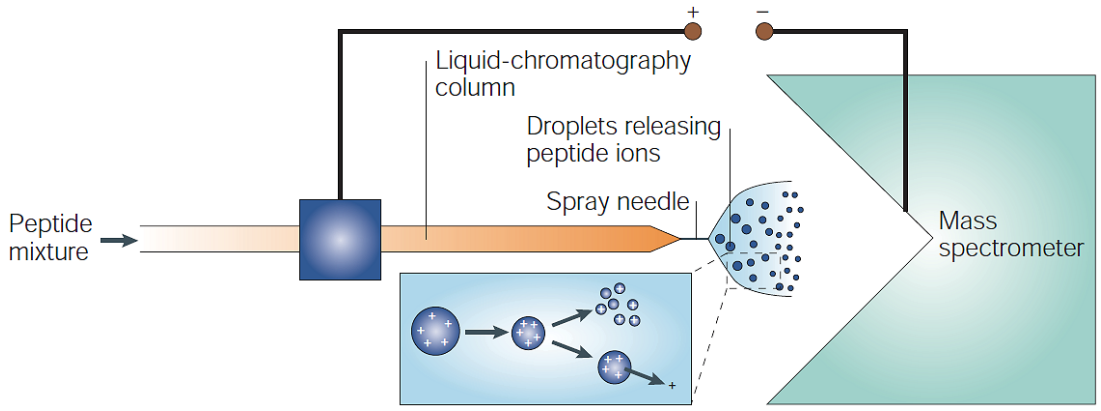


Figure 1. 1 : Electrospray ionisation.
Taken from (Steen and Mann 2004).

1.1.3 Mass analysers

Ions within a mass spectrometer are separated by a mass analyser according to their *m/z* ratio. Mass analysers vary in their ability to separate ions of similar or near-identical *m/z* (resolving power), the range of ions which are transmissible and the efficiency of that transmission. Many types of mass analysers are used in modern proteomics environments, since each have their own strengths and weaknesses for different applications.

1.1.3.1 Ion trap

The quadrupole ion trap (IT) mass analyser was conceived in the early 1950's (Paul and Steinwedel 1953) and patented by Wolfgang Paul in 1960. He was awarded the Nobel Prize for Physics in 1989 for his work on ion traps. Ions are pulsed into the trapping region and stored by the use of radio frequency (RF) and direct current (DC) voltages. Originally used as a mass-selective detection or storage device in which the RF/DC ratio is maintained to allow single *m/*z ions to be stored in the trap, these instruments were used by only a small number of groups (Cooks, Glish et al. 1991) until the Finnegan MAT Corporation (Stafford Jr, Kelley et al. 1984) developed a mass-selective instability mode in which all the ions are stored and then sequentially scanned from the trap using ramped voltages.

The 2D or linear IT is composed of four linear rods (quadrupole) in which the ions are trapped within the potential well between the rods and subsequently ejected rapidly either radially or axially. A 3D trap is composed of three hyperbolic electrodes (one ring and two endcap) and uses helium at 1 mTorr as a buffer or damping gas within the trap to improve the resolution of the mass analysis. All IT mass analysers can suffer from space-charge effects if too many ions are stored within the trap at any given point in time, resulting in poor mass resolution. Modern instruments restrict the trapping time by utilising a pre-scan which calculates the maximum fill time allowed based on the number of ions present. Space-charge effects can result in a restriction of dynamic range of IT mass analysers to typically 2 orders. Resolution is typically nominal mass unless a narrow *m/z* range is selected for analysis.

1.1.3.2 Quadrupole

A quadrupole mass analyser (filter) consists of four parallel round rods, connected as opposite pairs, to which either RF or RF and DC are applied (Paul and Steinwedel 1953). Under a fixed electric field, only certain ions within a narrow window of *m/z* will have a stable trajectory through the quadrupole and reach the detector, whilst all other ions with unstable paths are lost to the rods and walls of the mass analyser.

A quadrupole mass analyser can be operated in multiple modes. In the RF-only mode, a wide range of *m/z* ions traverse the quadrupole (ion guide). Operated in static electric field a quadrupole only allows a narrow window of ions through. By varying the voltages applied to the quadrupole, a wide range of *m/z* ions can be detected over a short period of time.

Quadrupole mass analysers typically have unit mass resolution and low sensitivity when in MS scan mode but high sensitivity and wider dynamic range when used in a targeted analysis for specific compounds.

1.1.3.3 Time-of-Flight

First described by Stephens, as a mass spectrometer under construction (Stephens 1946), that would be "well suited for gas composition control, rapid analysis, and portable use", the Time-of-Flight (ToF) mass analyser determines the *m/z* of an ion based on the time it takes to traverse a field-free region (FFR) within a linear vacuum tube. If a set of ions of different mass/charge are first accelerated and given fixed kinetic energy prior to entering a flight tube, they will each achieve a different velocity and thus travel the fixed distance in a time inversely proportional to the square root of their mass (de Hoffmann and Stroobant 2007).

Theoretically, there is no upper mass limit for a ToF analyser although resolution is dependent on flight time and kinetic energy spread. The latter issue has been addressed by the development of time lag focussing or delayed pulse extraction (Wiley and McLaren 1955; Vestal, Juhasz et al. 1995) which aims to reduce peak broadening observed as a result of ions of identical *m/z* within the extraction region. A time delay is applied between ion formation and the extraction voltage being applied. An ion with an initial higher velocity will feel the extraction pulse less due to its increased distance from the plate. An ion of identical *m/z* but possessing a lower velocity will feel a greater acceleration in comparison, and both ions will reach the detector at the same time. The optimum delay is mass-dependent which requires that the time required is optimised for the mass range of interest.

Significant improvements in resolution on ToF mass analysers have been achieved by the addition of a reflectron or ion mirror within the flight tube. Developed by the Russian physicist Mamyrin, the reflectron is a hollow tube that uses an electric field to first reduce the velocity of the ions to zero and then reverse the direction of the ions back to the FFR of the flight tube (Karataev, Mamyrin et al. 1972; Mamyrin, Karataev et al. 1973; Mamyrin 1994) and on to the detector. Ions of identical *m/z* but varying kinetic energies will travel different distances within the reflectron, which is composed of a set of ring electrodes, with the ions of greater kinetic energy travelling furthest before being accelerated out of the ion mirror. As a result ions of identical *m/z* reach the detector at the same time reducing peak broadening and improving resolution by having effectively doubled the length of the flight tube. A dual stage reflectron employs two electric fields or regions with ions entering the highest field first (Mamyrin, Karataev et al. 1973). It is able to compensate for a much wider range of kinetic energies than a single stage reflectron.

Orthogonal acceleration (oa) compensates for the spatial spread of ions through the use of a large voltage pulse (up to 4 kV) which pushes the ions in a direction orthogonal to ion generation (G.J.O'Halloran, R.A.Fluegge et al. 1964; Dawson and Guilhaus 1989; Coles and Guilhaus 1993). Mass analysis is performed off-axis making this technique well suited for continuous ionisation sources in combination with the use of a dual stage reflectron. The incorporation of LC-MS/MS with an oa configuration on Q-ToF instrumentation was first reported with fragment ion mass accuracy of 0.1 Da and the ability to differentiate singly, doubly and triply charged peptide ions due to mass resolution of 3,000 (Bateman, Green et al. 1995; Morris, Paxton et al. 1996). Factors limiting the resolving power of oa-ToF instruments include velocity and space distribution of ions in the pusher region, misalignment of ions in the accumulation region and acceleration voltage drift/instability (Dodonov, Kozlovski et al. 2000).

With biologically-relevant sensitivity resulting from a high duty cycle, higher mass resolution and sub-second scan rates, a ToF instrument is well suited for capillary and nanoflow LC systems in use in proteomics applications today. The instruments used in this study combined a quadrupole with oa-ToF configuration. These were a

Q-ToF Ultima Global and Synapt HDMS (Waters Corporation, Manchester, U.K.) both fitted with a microchannel plate (MCP) detector. An MCP detector is composed of an array of parallel miniature electron multipliers, typically over 100/plate where each channel is at a slight angle to the plate surface (Wiza 1979). MCP detectors are used for proteomics applications since they are capable of detecting large numbers of ions simultaneously. Where accurate mass and quantitative measurements are being made, it is important to ensure that the detector is not saturated with ions at any given point during the analysis. Each channel requires a recovery time and any ions hitting the channel during this period will not be measured.

1.1.3.4 Orbitrap

Building on the design of the Kingdon ion trap (Kingdon 1923), in 2000 Alexander Makarov published details on a new type of mass analyser (Makarov 2000) which became commercialised as the Orbitrap. Ions orbit around a central axial electrode in harmonic oscillations (frequency proportional to $m/z^{-1/2}$) which can be detected using ion image currents and converted to mass spectra using Fourier Transform (FT) techniques.

The resolution of the Orbitrap is scan time and *m/z* dependent. For the latter, the relationship is inversely proportional to the square root of *m/z*, thus a resolution of 60,000 at *m/z* 400 on an Orbitrap Velos reduces to 30,000 at *m/z* 1600 with the same scan rate. In order to obtain higher resolving powers the early Orbitrap instrument required multi-second scan rates which were incompatible with the ultra fast chromatography systems on the market, but more recent developments allow sub-second scan rates whilst maintaining high resolution measurements (Michalski, Damoc et al. 2011) with high sensitivity and wide dynamic range. In order to achieve a resolving power of 60,000 on an Orbitrap Velos at *m/z* 400, a scan rate of 1Hz is required. On an Orbitrap Elite this scan rate can be achieved at 4 Hz. The maximum resolving power at this *m/z* for the Orbitrap Velos and Elite models are 100, 000 and 240,000 respectively.

For some mass analysers, such as TOF there is minimal change in resolution across the mass range shown in Figure 1. 2, but both the Orbitrap and FT-ICR show a decrease in resolution at increased *m/z*. The Q-ToF mass resolution indicated in Figure 1. 2 is based on a Synapt G2 mass spectrometer and the Orbitrap values are for a Q-Exactive.

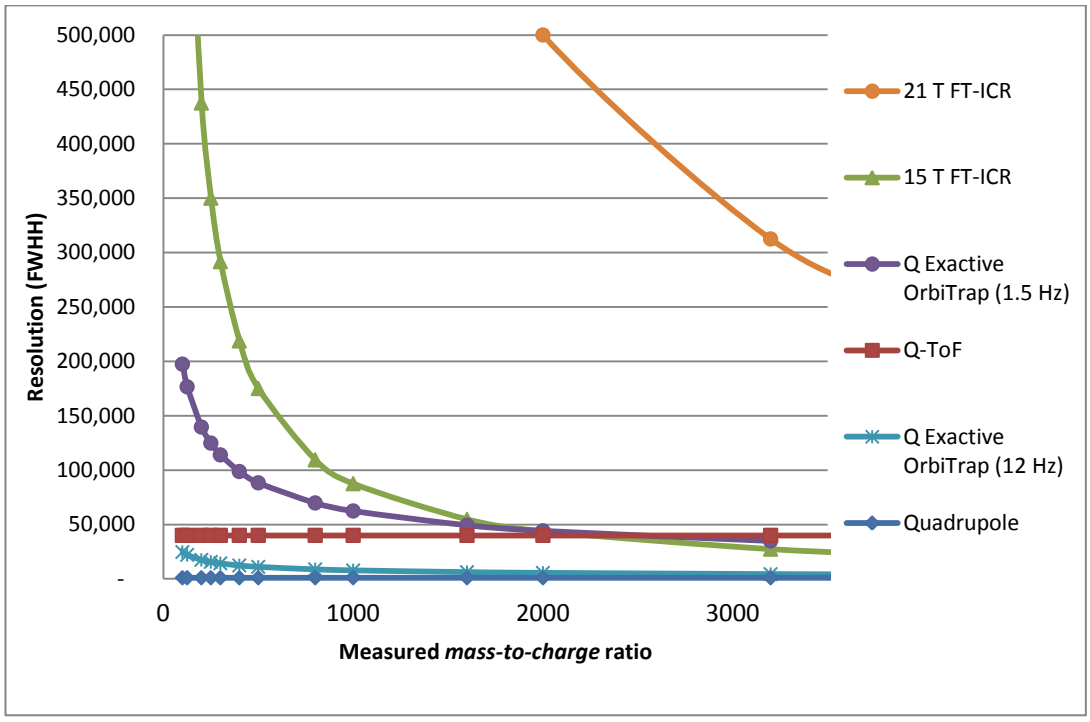


Figure 1. 2 : Effect of measured *mass/charge* on mass analyser resolution.

1.1.3.5 Fourier Transform Ion Cyclotron Resonance

Like the Orbitrap, mass measurement in a Fourier Transform Ion Cyclotron Resonance (FT-ICR) analyser is based on the oscillations of ions, in this case in a strong magnetic field (Comisarow and Marshall 1974). The ions are accelerated or excited into a cyclotron radius, the accelerating field is removed and the movement of the ions induces an image current on a pair of plates which is subsequently detected. The mass spectrum is generated from a Fourier transform of the data. Resolution is dependent on the time for detection and the strength of the magnetic field and is inversely proportional to *m/z*.

A 14.5 Tesla FT-ICR (200,000 resolving power at *m/z* 400) fitted with a nanoflow LC system, using a domain acquisition period of <800 msec, has been configured for the characterisation of algal polar lipids (He, Rodgers et al. 2011). This application is an example of the use of ultra high resolution mass analysis to provide unequivocal elemental composition of the lipids with sufficiently fast acquisition rates necessitated by the nanoflow LC peak widths.

1.2 Tandem mass spectrometry

Tandem mass spectrometry or MS/MS involves two stages of mass selection or mass analysis. For the MS/MS experiments conducted as part of this study on the Q-ToF instruments the first mass analyser (quadrupole) was used to select a precursor peptide for MS/MS. After precursor selection the peptide undergoes a controlled collisional-induced decomposition and the second mass analyser (ToF) is used for mass analysis of the fragment or product ions generated. The product ions can then be used to characterise the species under investigation, in this case to identify the peptide based on its amino acid sequence.

On instruments where the two stages of mass analysis are spatially distinct, this is termed in-space tandem mass spectrometry and is performed on instruments such as tandem or triple quadrupole, Q-ToF and ToF/ToF. Where the ions are trapped and then subsequently allowed to undergo dissociation e.g. ion traps and FT-ICR mass spectrometers, mass selection and analysis is performed using the same analyser, this is defined as in-space tandem mass spectrometry. In the latter example, multiple rounds of MS/MS can be performed since a fragment ion from the MS/MS experiment can then be mass selected, dissociated and the fragments subsequently subjected to further mass analysis (MS^3) and so on (MS^n). The sensitivity of the experiment falls significantly with each additional round of MS/MS.

A number of scan modes of tandem mass spectrometry are available and are suitable for different applications:

Product ion scan – the first mass analyser is static, selecting a specific ion *m/z* for dissociation whilst the second analyser scans the fragment ions.

Precursor ion scan – the second mass analyser sits static on a specific product ion *m/z* whilst the first is scanning for possible precursors that generate that fragment ion.

Neutral loss scan – both mass analysers are scanning but at a mass offset characteristic of the neutral loss under investigation.

Selected reaction monitoring – both mass analysers are static, the first selecting the precursor ion *m/z* and the second a specific fragment ion characteristic of the compound of interest.

A number of fragmentation methods are utilised in mass spectrometry and of most relevance to this study, which used Q-ToF instruments, is collisional-induced decomposition (CID) (McLafferty and Bryce 1967; Haddon and McLafferty 1968; Jennings 1968). The selected precursor ion is subjected to collisions within a chamber containing inert gas molecules e.g. argon converting some of the translational kinetic energy into an increase in internal vibrational energy which induces fragmentation of the ion at the lowest energy backbone cleavage sites. In these low energy CID experiments the fragmentation occurs primarily at amide bonds forming *b* and *y* ions.

Electron capture dissociation (ECD) causes fragmentation primarily of the peptide amine bond (nitrogen atom and the α-carbon) producing *c* and *z* ions (Zubarev, Kelleher et al. 1998). Dissociation is caused by the interaction of an electron and a multiply charged peptide or protein ion, an odd-electron, free-radical driven fragmentation. Found primarily on FT-ICR instruments, ECD has been used to characterise labile post-translation modifications (PTM) such as phosphorylation (Stensballe, Jensen et al. 2000) and glycosylation (Mirgorodskaya, Roepstorff et al. 1999).

Electron transfer dissociation (ETD) is analogous to ECD. It was developed initially for ion trap instruments (Syka, Coon et al. 2004) and uses a radical anion e.g. anthracene or nitrosobenzene to rapidly transfer an electron to multiply charged peptides or proteins. The fragmentation pathways involved are similar to those in ECD. ECD and ETD, unlike CID, are nonergodic processes i.e. do not involve intramolecular vibrational energy redistribution. The capability to perform ETD

experiments on a hybrid ion mobility-enabled Q-ToF has been demonstrated (Campuzano, Brown et al. 2010). ETD is incompatible with MS^E data acquisition as an ETD experiment does not utilise collision energy to generate fragment ions, therefore DDA is performed. This work utilised MS^E acquisition with CID employing trypsin as the enzyme for proteolytic cleavage. Alternative proteases have been employed to generate higher molecular weight peptides (higher charge states) for ETD analysis, as these have been shown to generate improved sequence information from MS/MS spectra (Pitteri, Chrisman et al. 2005), including Lys-C, Glu-C (Molina, Horn et al. 2007)and Lys-N (Taouatas, Drugan et al. 2008). The use of ETD would have increased the experimental time required to perform the experiments on another instrument but could be used to characterise the post-translational modifications on plasma proteins, in particular glycosylation (Mikesh, Ueberheide et al. 2006).

The nomenclature for the product ions of peptide fragmentation was initially proposed by Roepstorff and Fohlman (Roepstorff and Fohlman 1984) and subsequently modified (Johnson, Martin et al. 1987). Product ions will only be detected if they carry one or more charges; if the charge is retained on the N-terminal fragment the ion is termed *a*, *b* or *c*. Conversely, C-terminally charged ions are referred to as *x*, *y* or *z*. The subscript employed indicates the number of amino acid residues in the fragment, Figure 1. 3

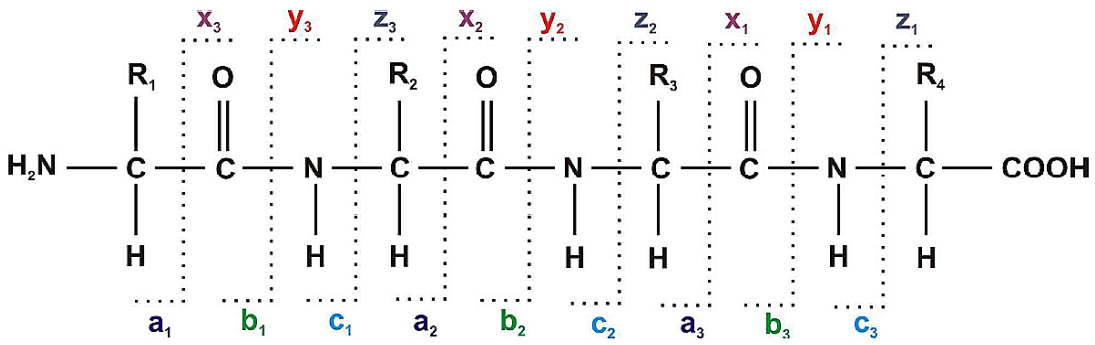


Figure 1. 3 : Product ion nomenclature for a peptide

Internal fragments can be produced from a double cleavage of the peptide backbone (*b* and *y* cleavage) and are termed amino-acylium ions. Immonium ions which can be formed are also the result of internal fragmentation, containing a single side chain and result from combined *a* and *y* cleavage having the general structure $RCH = NH_2^+$ where R is the amino acid side chain. Immonium ions can be useful for the interpretation of MS/MS spectra providing information indicating the possible presence of residues in the peptide (Falick, Hines et al. 1993).

MS/MS spectra can be used to infer a peptide sequence, although frequently the product ion spectrum does not provide sufficient information to fully and unambiguously characterise the peptide. The procedure can be very labour intensive if carried out manually. With the advent of software algorithms to perform this role and the development of protein databases for data interrogation, the requirement for the manual interpretation of MS/MS spectra *de novo* has reduced significantly in recent years.

1.3 Mass spectrometry-based proteomics

In 1994 the term PROTEOME was first used and defined as the PROTEin complement of the genOME by Marc Wilkins and co-workers at Macquarie University, Australia (Wasinger, Cordwell et al. 1995). Two years later the first use of PROTEOMICS appeared referring to the study of the PROTEOME (Wilkins, Pasquali et al. 1996).

Proteomics can be described as the characterisation of the protein products expressed by a genome under a defined environmental condition. Whereas the genome is a static entity, the proteome is constantly evolving in response to environmental stimuli, resulting in changes in protein expression or post-translational modification events. The study of the proteome is therefore, essential to understanding biological systems.

Due to an exponential increase in transcriptomic information available, some correlation between mRNA prediction and respective protein level can now be

predicted from the genome. Gygi reported stable mRNA levels with corresponding protein levels varying by 20-fold and conversely examples of stable protein levels whilst mRNA transcripts varied 30-fold, this work was based on 150 proteins from *Saccharomyces cerevisiae* (Gygi, Rochon et al. 1999). The group reported that "simple deduction from mRNA transcript analysis is insufficient".

Schwanhausser attempted to correlate not only protein and mRNA levels but also to assess their respective turnover rates based on over 5000 genes from mammalian mouse fibroblasts using pulse labelling experiments (Schwanhausser, Busse et al. 2011). They reported no correlation between protein and mRNA half-lives R^2=0.02 (log–log scale). mRNAs were on average five times less stable than proteins (median 9 hr compared to 46 hr) whilst proteins spanned a wider concentration range. On average, protein levels were significantly higher (900-fold) than their respective mRNA and they reported a higher correlation between the two levels than previously observed. The study showed that for non-synchronised, exponentially growing mouse fibroblasts 40% of the observed protein levels could be accounted for by mRNA levels with different translation efficiencies contributing to the higher dynamic range of proteins observed, with abundant proteins translated 100-fold more efficiently. In general housekeeping genes tend to have stable mRNAs and proteins whilst gene products required for cellular responses tended to have unstable mRNAs and proteins, e.g. transcription factors, cell signalling and cell cycle functions. Secreted proteins tended to have stable mRNAs unlike the proteins themselves.

Vogel and co-workers experimentally measured absolute protein and matching mRNA levels for >1000 genes in the Daoy medulloblastoma human cell line, using a combination of shotgun proteomics and microarrays (Vogel, de Sousa Abreu et al. 2010). The study identified sequence characteristics, with dominant functions in the regulation of translation and protein degradation. A model including mRNA and sequence features was proposed to explain 67% of the variation of protein abundance in the mammalian system. The contribution of translation and protein degradation was shown to be as important as that of mRNA transcription/stability to the protein abundance. The authors demonstrated that protein and matching mRNA levels correlated significantly, with variation in mRNA expression explaining ~25–30% of the variation in protein abundance. Another 30–40% of the variation could be

accounted for by sequences characteristics including sequence length, amino-acid frequencies and also nucleotide frequencies.

With the development of soft ionisation techniques that allowed the study of proteins and peptides by mass spectrometry, not only could proteins be rapidly identified but also quantified and the presence, absence or stoichiometry of PTMs be achieved on a large scale (Aebersold and Mann 2003; Domon and Aebersold 2006; Han, Aslanian et al. 2008; Yates, Ruse et al. 2009; Walther and Mann 2010).

1.3.1 Challenges of proteomics

The challenges faced by proteomic researchers should not be underestimated and some are detailed below.

In a typical prokaryotic cellular system the range of concentration between most and least abundant protein would typically be $10^4 – 10^5$. No single experiment has yet been able to identify *and* quantify the entire proteome over this dynamic range. Experimental approaches involving sample prefractionation prior to MS are necessary to probe deep into the low abundance proteome. In plasma, the dynamic range is even greater approaching 10^{12} (Anderson and Anderson 2002). A single protein, albumin, dominates approximately 50% of the proteome and the 12 most abundant account for 94% of the proteome. Proteins present at less than 100 copies/cell are usually below detectable limits (Schwanhausser, Busse et al. 2011).

A single methodology has previously been insufficient to fully probe the proteome but a study by Nagaraj utilised long LC gradients and an Orbitrap-based instrument (Q-Exactive) to characterise the yeast proteome, identifying almost 4,000 proteins in each 1D separation, close to the number of proteins that would be expected to be expressed (Nagaraj, Kulak et al. 2012). In most proteomic experiments, orthogonal approaches are often required, increasing the sample requirement and experimental time significantly.

Post-translational modifications significantly increase the complexity of the proteome. The RESID Database aims to be a comprehensive resource of

information on both naturally occurring PTMs and chemically induced modifications containing 572 entries in release 68.0 on 31st December, 2011 (http://www.ebi.ac.uk/RESID) (Garavelli 2004). PTMs such as phosphorylation and glycosylation are commonly occurring but cannot entirely be predicted from the genome sequence. Although potential sites of occupancy may exist, this does not mean that the site is always occupied or that there is homogeneity at that site within the proteome.

Enrichment of the PTM-specific proteome may be required for detailed study, either at the protein or peptide level (McLachlin and Chait 2001; Macek, Mann et al. 2009). Phosphopeptides in particular may be present at low abundance and their ionisation efficiency can be significantly lower than the native peptide. For phosphoserine and phosphothreonine-containing peptides the reduced ionisation efficiency observed is between two and five-fold and for phosphotyrosine 10-fold at equimolar stoichiometry, compared to the non-phosphorylated counterpart (author's own observations, data not shown). Typically a fractionation step (strong cation exchange or hydrophilic chromatography) followed by immobilised metal affinity (iron, gallium, TiO_2 or ZnO_2) is used. No technique is suitable for all phosphopeptides and a recent comparison of methodologies revealed that although there was some overlap, each technique enriched a subset of modified peptides (Bodenmiller, Mueller et al. 2007; Fila and Honys 2011).

In eukaryotes, variability in protein products that can be produced from a single gene is achieved by alternative pre-mRNA splicing resulting in protein isoforms (Black 2003). Identification of the correct isoform of a protein from proteomics data sets can be particularly difficult and may require subsequent further experiments to confirm initial observation (Hatakeyama, Ohshima et al. 2011; Moskaleva, Zgoda et al. 2011; Wu, Tolic et al. 2011). Ribosome profiling based on deep sequencing of ribosome-protected mRNA fragments has yielded information on the potential for alternative translation products. Although recognition of the correct translation initiation site for many proteins is essential to ensure its correct localisation and biological functionality, the analysis of the N-termini of 706 *Saccharomyces cerevisiae* proteins identified up to 89 potential alternate translation initiation sites (Helsens, Van Damme et al. 2011; Menschaert, Van Criekinge et al. 2013).

Improvements in analytical techniques used for the quantitation of the proteome have demonstrated that biological replicates may not be always be biologically identical. Proteomic measurements are based on the average abundance of each protein in a heterogeneous sample. Care must be taken to ensure that, where possible, cells are harvested, stored and processed at the same time points in an identical manner. If the samples have been treated, consideration needs to be given to the effect on cell growth and cell cycle. With cell lines, the cells need to be synchronised at the point of treatment and collection. If samples are combined from numerous sources, care needs to be taken that the pooled sample is representative of all the individual protein levels from each source. Good experimental design is crucial before any proteomic study can be undertaken (Wilkins and Hunt 2007; Song, Bandow et al. 2008; Caffrey 2010) .

1.4 Proteomic approaches

Experimental approaches taken in proteomic studies can be categorised as profiling, comparative (relative) or absolute quantitation. In profiling experiments the objective is to identify as many proteins as possible from the proteome, whereas in comparative studies the identified proteins are quantified *relative* to those measured in each proteome under investigation. For proteins that have been identified as changing in expression, levels are reported in terms of fold-change. These approaches have been described as the discovery stage of a proteomic study. In absolute quantitation the level of an identified specific sub-set of proteins is quantified (usually those identified from the discovery phase). This is termed targeted proteomics as it requires the use of internal standards which mean that the selected proteins in each assay can be reported in terms of their concentration in molarity (femtomole or attomole), mass (pg) or copies/cell. Approaches for targeted studies will be discussed further in Section 1.7.

Proteomic studies that characterise the proteome using *intact* proteins are referred to as top-down whilst the analysis of digested proteins at the peptide level is termed a bottom-up approach (Chait 2006).

1.4.1 Top-down proteomics

Top-down proteomics experiments are typically performed using ESI on a high resolution FT-ICR mass spectrometer. Purified or semi-purified proteins are mass measured, precursors of interest are selected and then dissociated, typically by ECD, although the use of sustained off-resonance irradiation (SORI), infrared multi-photon dissociation using a CO_2 laser (IRMPD), blackbody radiation, CID and surface-induced dissociation (SID) have been reported (Bogdanov and Smith 2005). Excellent sequence coverage can be achieved by top-down approaches and the technique is often used in the study of labile PTMs (Breuker, Jin et al. 2008) and protein interactions (Breuker and McLafferty 2003).

The disadvantages of a top-down approach include the costs involved in the purchasing, maintenance and running of the instrument and the need for highly skilled operators (Marshall 2000). The sample introduced into the mass spectrometer needs to be a relatively pure protein available in reasonable quantities, the MS/MS spectrum requires significant interpretation and the intact molecular mass of the protein needs to be ideally relatively small (<50 kDa) although some success has been reported on larger proteins using a combination of top- and middle-down (limited proteolysis) approaches (Ge, Rybakova et al. 2009). The use of ESI additives, heated vaporisation, independent non-covalent and covalent bond dissociation (Han, Jin et al. 2006) or variable thermal and collisional activation (McLafferty, Breuker et al. 2007) has proven useful for proteins >200 kDa.

FT-ICR mass spectrometry of proteins typically required off-line purification of the protein of interest prior to analysis by direct infusion. Thus the application of top-down proteomics to complex mixtures of proteins was not feasible. Liquid-based protein separation strategies such as gel-eluted liquid fraction entrapment electrophoresis (GELFrEE) or a combination of solution-based IEF with GELFrEE followed by LC-MS/MS has permitted the characterisation of human cell line proteins up to 25 kDa and 72 kDa molecular weight respectively (Lee, Kellie et al. 2009; Tipton, Tran et al. 2012). This GELFrEE separation LC-MS/MS methodology has been used to study whole membrane proteins with molecular weight below 60

kDa containing up to 8 transmembrane helices by a top-down approach (Catherman, Li et al. 2013).

1.4.2 Bottom-up proteomics

In a bottom-up proteomic approach as shown schematically in
Figure 1. 4**,** the protein-containing sample is digested enzymatically, chemically or both and the resulting peptides analysed using mass spectrometry. In large-scale studies, trypsin is invariably used as the protease of choice for enzymatic digestion since it has a high specificity for hydrolysing peptide bonds at the carboxylic sides of arginine or lysine residues. In commercial preparations of trypsin, the lysine residues have been reductively methylated, to produce a stable enzyme resistant to autolysis and treated with L-1-tosylamide-2-phenylethyl chloromethyl ketone to remove chymotryptic activity. Trypsin is stable under a wide set of conditions and will tolerate the presence of sodium dodecyl sulphate (SDS - 0.1% w/v) and chaotrophs (2 M urea or guanidine). The tryptic peptides generated are ideally sized for MS/MS experiments and span a range of hydrophobicity making them amenable to a reversed phase chromatographic separation. The generation of a peptide with a basic C-terminal residue typically generates a *y*-ion series in the product ion spectrum that can be interpreted fully or partially aiding in the identification of the peptide.

Where ESI is used in a bottom-up approach, an in-line LC separation of the peptides will typically be used to reduce the complexity of the analytes entering the instrument at any given point in time. The use of LC in proteomics is discussed further in Section 1.4.2.5.

In MALDI experiments the digested samples can be spotted directly onto a target for analysis, or for more complex samples off-line spotting of the LC eluate mixed with matrix can be used to pre-fractionate and simplify the analytes ionised.

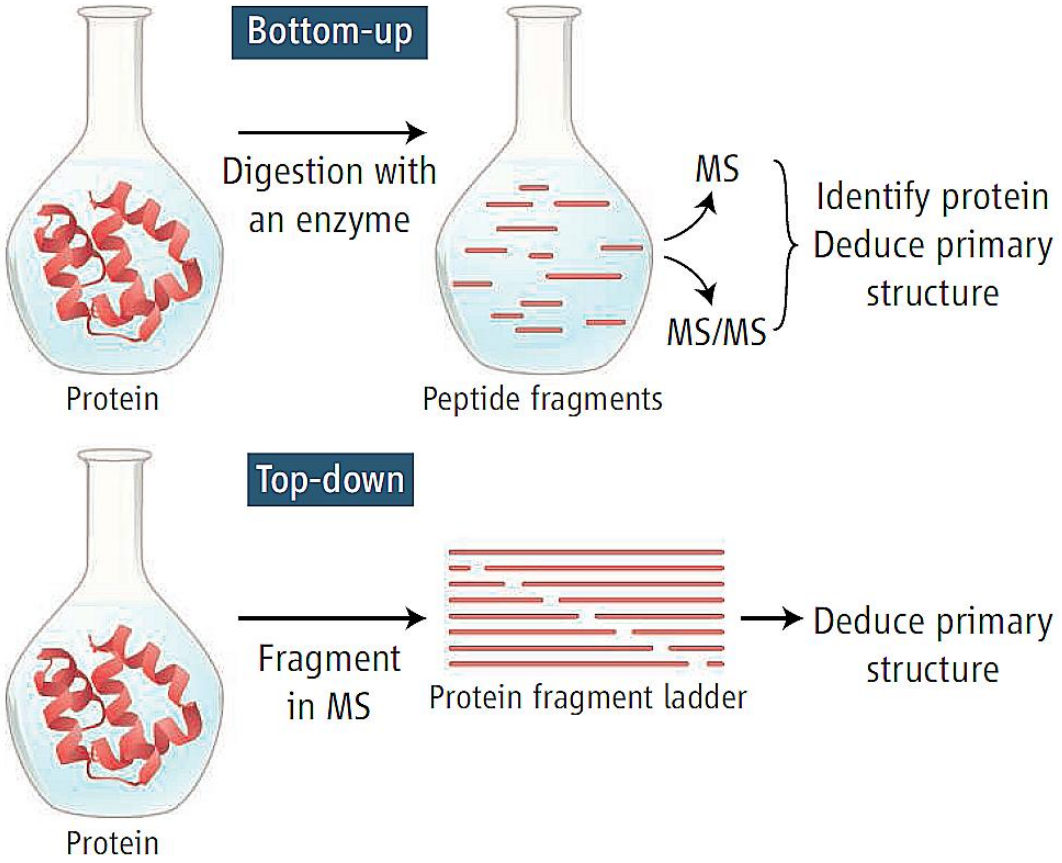


Figure 1. 4 : Bottom-up and top-down proteomics approaches.
Taken from (Chait 2006).

1.4.2.1 Peptide mass fingerprinting

The use of peptide mass fingerprinting (PMF) in combination with the use of software algorithms was first described for the identification of proteins in 1993 by a number of groups (Henzel, Billeci et al. 1993; James, Quadroni et al. 1993; Mann, Hojrup et al. 1993; Pappin, Hojrup et al. 1993). The approach utilised peptide masses, in the absence of tandem MS information, to identify a protein from a database containing many thousands of sequences. This approach was rapid and was most effective when the sample under investigation was relatively simple, either purified or pre-fractionated. For more complex organisms, such as eukaryotes, many protein sequences in the database could give rise to the PMF observed by ESI or MALDI, potentially leading to an incorrect identification.

1.4.2.2 Protein identification incorporating MS and tandem MS information

In 1994, the Yates group published an approach that utilised both peptide mass and tandem MS data, using a computer algorithm to identify proteins from *Escherichia coli* and *Saccharomyces cerevisiae* cell lysates and peptides released from major histocompatibility complex class II cell lines (Eng, McCormack et al. 1994). The algorithm used the protein sequences in the Genpept database to predict the fragmentation pattern that would be expected to be generated from each tryptic peptide. The algorithm identified all the potential peptides from the database based on the observed experimental mass (within a ± 0.05% or 1 mass unit window). For each peptide the observed fragmentation (± 1 mass unit) was compared with the predicted pattern and given a score. The highest scoring peptide sequences were then reported. In these experiments the MS and tandem MS experiments were performed independent of one another as an automated system of collecting both sets of data had yet to be developed.

1.4.2.3 Data dependent acquisition

In 1996, the technique subsequently referred to as data dependent (McCormack, Schieltz et al. 1997) or data directed acquisition was first published (Stahl, Swiderek et al. 1996). This allowed MS and tandem MS data collection to occur within the same analysis minimising the amount of sample required. The data system was programmed to select suitable peptides for MS/MS analysis based on a number of criteria including signal-to-noise and quality of MS/MS data in *real time*. Typically 4-8 MS/MS precursor spectra were collected from each precursor and common contaminant ions from in-gel digestions were excluded from the selection process. The number of precursors that can be selected for MS/MS has been increasing with instrument developments and up to 20 can now be selected.

A data dependent acquisition is, by the nature of the criteria that control precursor peptide selection, a concentration-dependent technique. A small number of peptides are selected from a single MS scan (typically 3-8) and then for a period of time a product ion scan is collected from each of the precursors and the process repeated for a specified time or number of scans, Figure 1. 5. During this time all information on

co-eluting peptides entering the instrument but not selected for MS/MS is lost. In order to obtain the highest quality MS/MS data, the software will select the most abundant peptides at a given point and the less abundant (or less ionising) peptides will not be selected. This can result in peptides from a highly abundant protein being sampled more frequently than those from a low abundance protein which may be observed by only a single peptide (Gygi, Rist et al. 1999b; Link, Eng et al. 1999).

The Yates group assessed the level of random sampling in DDA experiments on a yeast proteome in order to identify low abundance proteins (Liu, Sadygov et al. 2004). From 9 identical analyses, incorporating two-dimensional LC to simplify the mixture, a total of 1751 proteins were identified. Of those, 35.4% were found in every replicate and 24% were identified just once. For each additional experiment over the 9 performed, the model predicted only a small increase in the number of proteins observed. Thus for effective coverage of low abundance proteins analysed using DDA, more than 3 replicates are required.

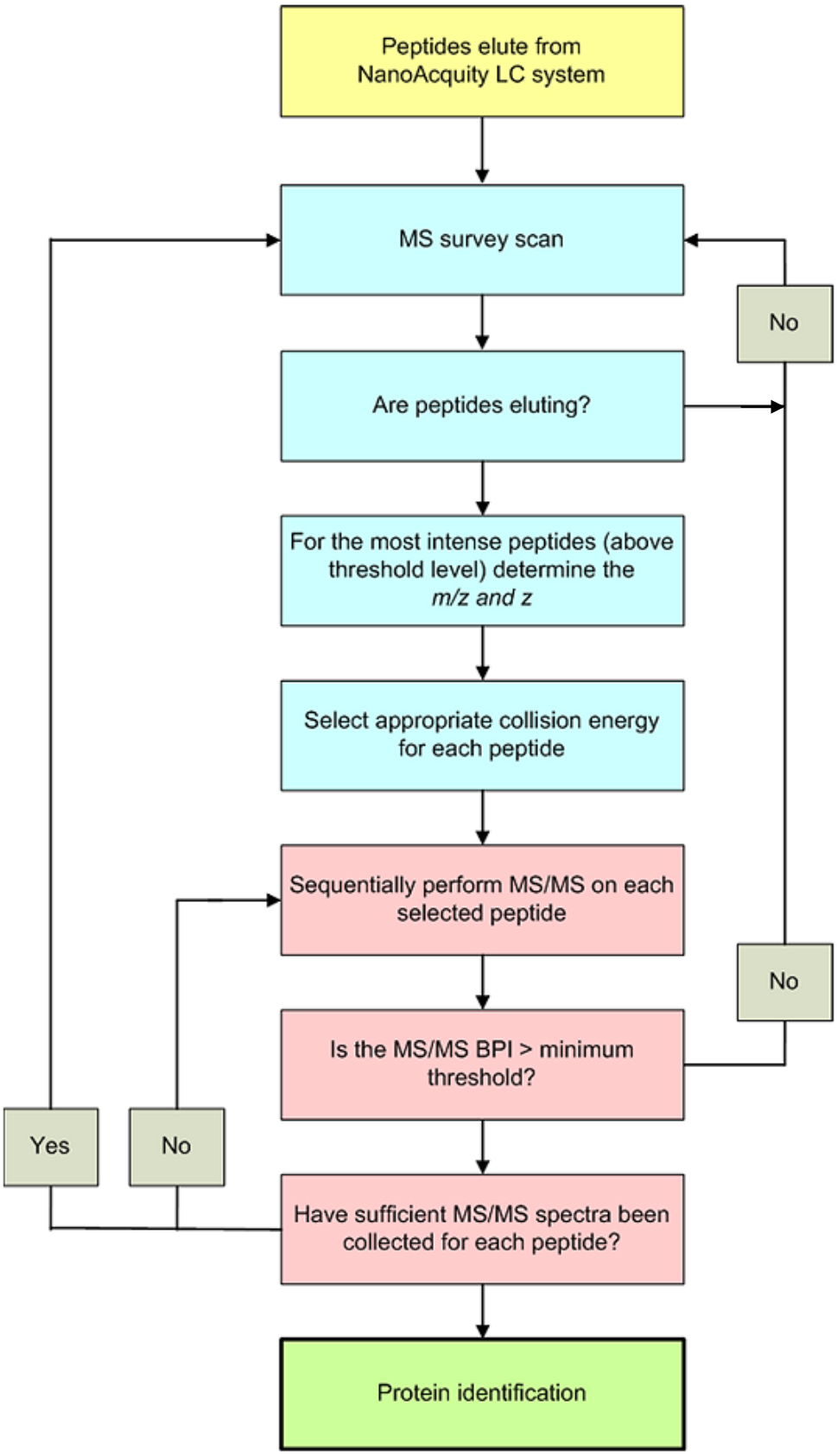


Figure 1. 5 : Schematic of a data dependent acquisition.

1.4.2.4 Data independent acquisition

A variation on DDA uses relatively large precursor windows (*m/z* 10 mass units) within an ion trap for isolation and subsequent fragmentation of peptides across the mass range under study (Venable, Dong et al. 2004). Approximately 20% more spectra were collected using this approach, but this did not significantly increase the number of peptides or proteins identified.

In 2003, a methodology that involved a data independent acquisition (DIA) was published. This approach utilised two different nozzle-skimmer voltages (V_{NS}) on an ion trap. This was referred to as shotgun CID and compared with DDA (Purvine, Eppel et al. 2003). For shotgun CID two experiments were performed separately, in the first the V_{NS} was set low to minimise fragmentation in the source region and then a second V_{NS} was used that would result in product ion formation. The results from a database search of the DDA and DIA data indicated results were comparable between the techniques with only one peptide that was identified by DIA alone.

The authors reported the ambiguity of determining for each product ion which precursor it arose from, but suggested this could be addressed by comparison of the chromatographic profiles for each precursor to that of the potential product ions, in essence time-alignment of the precursor-product ion lineage. The authors pointed out that this technique could be performed not only in-source on MS instrumentation but also in a collision cell such as that employed on a Q-ToF instrument and that the limiting factor was the software needed to process the data.

The collection of combined low/high collision energy mass spectra was reported by Bateman *et al.* on a Q-ToF instrument to identify phosphorylated peptides by the observation of either a neutral loss of 98 Da (H_3PO_4) or a phosphorylated immonium ion at *m/z* 216 under high energy MS acquisitions (Bateman, Carruthers et al. 2002). Waters Corporation commercialised data independent acquisition based on this same premise of using alternating high/low collision energy to obtain precursor and product ion (pseudo MS/MS fragmentation) spectra. It was termed MS^E, where ^{E} refers to the use of elevated collision energy during the acquisition process.

In 2005 Silva *et al.* reported the use of accurate mass retention time pairs collected by LC-MS^E on a Q-ToF instrument for relative comparison of peptide intensities (Silva, Denny et al. 2005). The data were collected in low energy MS using a collision energy (CE) of 10 V and the elevated MS was performed using a ramped CE of 28-35 eV. They reported the need for accurate mass measurements, such as those obtained on Q-ToF instruments, in conjunction with reproducible chromatography, to allow the data obtained across multiple experiments and potentially many LC column changes to be compared. The peptides that were found to be changing in relative intensity were selected and submitted for PMF identification with a mass accuracy of typically ± 5 ppm.

The use of time-alignment to correlate product and precursor ions obtained by LC-MS^E was published the following year (Silva, Denny et al. 2006). The software algorithm assumes that any fragment ions have an elution apex within one scan of the precursor, in this case 1.8 sec. The mass accuracy of both precursor and fragment ions (± 5 ppm) could be used to correlate or reject particular fragment ions. A probabilistic peptide fragmentation model generated from empirical data from well characterised samples (Skilling, Denny et al. 2004) was incorporated into the software ProteinLynx Global Server (PLGS - Waters Corporation, MA, USA) was used to provide additional validation of product-precursor ion lineage. The precursor and product ion data from LC-MS^E were deisotoped and charge state reduced generating a set of mass-corrected, monoisotopic ions for database interrogation providing qualitative and quantitative information on the proteins identified.

The software algorithm used in PLGS rel. 2.3 and 2.4 will be discussed in Section 1.4.3.6 whilst the quantitative aspects of LC-MS^E will be detailed in Sections 1.6.4 and 1.7.1.

The incorporation of ion mobility into an MS^E experiment has been shown to increase the proteome coverage up to 60% providing higher confidence peptide and protein identifications (Shliaha, Bond et al. 2013). Additional separation is achieved within the instrument, using travelling wave ion mobility separation of the peptide precursors with MS^E data acquisition. Precursors and their respective fragment ions are aligned based on their mobility properties (drift time alignment). Quantitation

can be achieved using mobility separated precursor ion area. Ion mobility has been shown to increase the dynamic range of the proteomics experiment, through enhanced deconvolution of both low abundance (background noise and contaminating components) and high abundance species (co-eluting peptides of similar *m/z*) (Valentine, Counterman et al. 1998). The use of travelling wave ion mobility separation causes a peptide specific decrease in sensitivity and dynamic range resulting from detector saturation, but the effects are reproducible and do not compromise experimental precision (Shliaha, Bond et al. 2013).

1.4.2.5 Liquid chromatographic separations in proteomics

For many decades mass spectrometry has been used in conjunction with chromatographic separation upstream of the instrument. Initially the technique used was gas chromatography, but in the case of proteomics high pressure liquid chromatography (HPLC), or as in this study, ultra high pressure liquid chromatography (UPLC) (Karpievitch, Polpitiya et al. 2010; Donato, Cacciola et al. 2011) tends to be employed.

The mixture of peptides generated by an enzymatic digest can vary in complexity from tens (single protein) to many thousands (proteome). An LC separation of the peptides is required to reduce the complexity of the analyte entering the instrument at any given point in time. If all the peptides were infused directly into the instrument, low abundance peptides would not be observed, peptides of similar *m/z* would be difficult to distinguish as separate entities, even with a high resolving power instrument. The precursor window for MS/MS would not be selective enough to allow a single peptide through for CID, generating chimaeric product ion spectra. Additionally the use of a chromatographic step will significantly increase the sensitivity of the analysis to biologically relevant concentrations.

Reversed phase (RP) chromatography is generally the preferred chromatographic separation approach if only a single dimension of chromatography is employed. The mobile phase is generated from an aqueous solution and organic solvent. Trifluoroacetic acid, which suppresses the ESI signal, is not generally used unless essential for a particular separation to be achieved. Here, the mobile phase A

consisted of aqueous 0.1% formic acid (FA) with a mobile phase B, acetonitrile containing 0.1% FA.

A C18 chemistry analytical column is usually chosen in capillary (µL min^{-1}) or nanoflow (nL min^{-1}) format. The peak capacity of a chromatographic system (maximum number of peaks that can be theoretically separated within a given gradient time) is dependent on particle size, chemistry, column length, mobile phases, slope of the gradient, linear velocity and temperature. As the stationary phase particle size decreases, the column peak capacity increases but so also does the pressure. By doubling the length of the column, peak capacity increases by about 40% but run times are extended and pressures increase (Gilar, Daly et al. 2004). In the UPLC systems used here, the backpressure generated by the analytical column was typically 5,500 – 7,000 psi (380 – 485 bar) using a 75 µm internal diameter (i.d.) column of 25 cm length with 1.7 µm particle size under aqueous conditions at 40 ^{o}C. Proteomic samples may frequently contain a number of buffer components that may interfere with the ESI process, suppressing the signal from the eluting peptides. An off-line desalting step may be employed or a trapping column installed in the LC configuration prior to the analytical column. The trapping column can contain similar chemistry to the analytical column, but typically will be shorter with a larger diameter and particle size to allow a relatively large volume of aqueous buffer to pass through at relatively low pressures. A trapping column also enables focusing of the sample onto the analytical column, improving peptide resolution.

The tryptic extract is loaded onto the trapping column equilibrated in mobile phase A containing a low % of mobile phase B and a divert valve opened allowing the pre-analytical mobile phase to flush directly to waste taking with it contaminating buffer components. Depending on the buffer composition this step can take many minutes before the divert valve is closed and a linear gradient elution over tens of minutes to hours elutes the peptides into the MS instrument. Peptides are eluted from a C18 chemistry column in order of hydrophobicity, Figure 1. 6**.** During long trapping steps some highly hydrophilic peptides may be lost from the trapping column to waste. Reproducibility of retention times is improved with the use of a column heater to stabilise the column temperature.

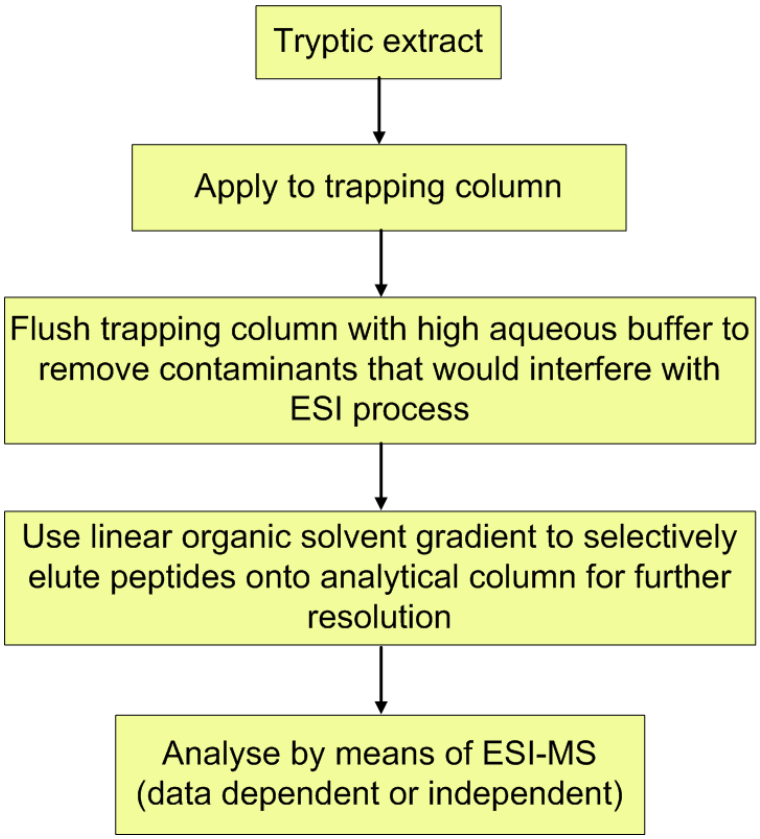


Figure 1. 6 : Schematic of a one dimensional chromatographic separation.

The peak capacity required for highly complex samples may not be achievable in a single one-dimensional (1D) chromatographic step. Multi-dimensional chromatography has been evaluated by a number of groups using orthogonal separations in each dimension to improve the LC separation, reviewed in (Zhang, Fang et al. 2010). Combinations of ion exchange, strong cation or anion exchange (SCX and SAX respectively) with RP, RP-RP **(**Figure 1. 7**)** or HILIC-RP have all been utilised. Here, RP-RP was employed to more extensively probe the proteome of depleted human maternal plasma, refer to Chapter 5 for a review of the RP-RP approach.

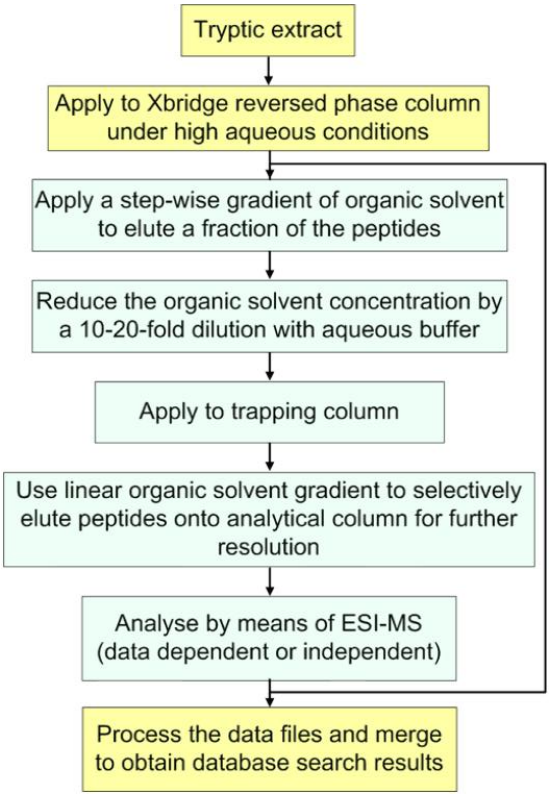


Figure 1. 7: Schematic of an on-line two dimensional reversed phase-reversed phase chromatographic separation.

1.4.3 Computational proteomics

Computational proteomics refers to the computational methods, software algorithms, protein or genomic databases and methodologies used to process, analyse, interpret and manage the data produced in proteomics experiments (Cannataro 2008).

1.4.3.1 *De novo* sequencing

De novo sequencing of a peptide from an MS/MS spectrum is laborious and can be very time consuming. Commercial software has been available for a number of years to assist in this task including SEQUIT! (Proteome Factory, Berlin, Germany), MassSeq (Waters Corporation), Peaks Studio (Bioinformatics Solutions, Ontario, Canada) and Mascot Distiller (Matrix Science, London, UK). *De novo* sequencing is typically used when a peptide cannot be identified from an existing database e.g.

no genome is available such as in a meta-proteomic study or in identification/confirmation of a PTM.

1.4.3.2 Uninterpreted database interrogations

Although software programs for *de novo* sequencing have improved significantly over the years, uninterpreted database interrogations are still the method of choice for proteomic studies, Section 1.4.2.2. Commercial software for the identification of peptides and proteins from uninterpreted data include Mascot (Matrix Science, UK) which is not vendor-specific, PLGS (Waters Corporation), SEQUEST (Thermo Fisher Scientific Inc., MA, USA) and ProteinPilot (Applied Biosystems/MDS Sciex).

A number of academic groups also provide access to in-house software such as ProteinProspector, Open Mass Spectrometry Search Algorithm (OMSSA) and X! Tandem.

One approach to improving the confidence in the protein assignments is to validate the MS/MS results obtained from one or more sources. MSQuant http://msquant.alwaysdata.net/) works in conjunction with Mascot using processed result files, and ProteinProphet (http://proteinprophet.sourceforge.net/index.html) with SEQUEST. Scaffold (Proteome Software, Inc., OR, USA) will process Mascot, Sequest, Phenyx and PLGS output and comes with X! Tandem. It uses statistical algorithms to calculate a probability for each protein identified including quantitative outputs. Peaks Studio (Bioinformatics Solutions, Canada) will also import results from Mascot, Sequest, X!Tandem and OMSSA to generate consensus reports in one browser.

1.4.3.3 Protein databases

It is essential that an appropriate and comprehensive database for the system under study is used. The database should contain all protein entries including trypsin (if used), internal standards and any proteins used in the sample preparation such as lysozyme for cell lysis. If a host-pathogen interaction is studied then the two independent sets of protein sequences should be interrogated, usually as a

concatenated database. The common Repository of Adventitious Proteins (cRAP) is a useful collection of protein sequences that can frequently be identified in proteomic experiments that may not be derived from the system under study. These include common laboratory proteins, proteins from dust or physical contact (keratins) and proteins used as molecular weight or mass spectrometry quantitation standards (http://www.thegpm.org/crap/index.html).

Although the number of protein databases available on-line is growing there are still many that are inaccessible or confidential. Where access is granted to the content, some may require modification to enable the software to successfully read the sequences and output the results in a sensible format.

Publicly available databases include UniProtKB (http://www.uniprot.org/), National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/protein/) and Integr8 from European Bioinformatics Institute, which has not been updated since 2011. UniProtKB/TrEMBL contains protein sequences from computationally generated annotation and large-scale functional characterisation. UniProtKB/Swiss-Prot (reviewed) is a manually annotated, non-redundant protein sequence database whilst the NCBInr database is the largest and least annotated. Species-specific or taxonomic divisions can be downloaded from these sites and used for interrogation with MS data.

1.4.3.4 Protein isoforms

An issue that arises with higher eukaryotic organisms is the handling by software algorithms of protein isoform sequences in databases. Frequently, identified peptides may be present in multiple protein entries in the database and shared across isoforms (Ahmed, Oliva et al. 2005; Misek, Kuick et al. 2005). In a bottom-up approach it may not be possible to confirm which of the protein forms is present in the sample and so additional validation experiments may be required e.g. a top-down or targeted analysis for unique proteotypic peptides from one of the isoforms (She, Rosu-Myles et al. 2012).

1.4.3.5 Calculation of false discovery rate

With the generation of large proteomic data sets there is a need for a determination of the accuracy of identification and/or quantitation. This is usually expressed as the false discovery rate (FDR) and is required for publication purposes at both the protein and peptide level (http://www.mcponline.org/site/misc/PhialdelphiaGuidelinesFINALDRAFT.pdf). The FDR can be calculated using a decoy database of *in silico* generated protein sequences (Eddes, Kapp et al. 2002; Cargile, Bundy et al. 2004; Elias, Haas et al. 2005; Huttlin, Hegeman et al. 2007). Commercial software packages provide the option to search a decoy database and report the FDR for the dataset.

In order to calculate the FDR for large data sets, a decoy database in which the original sequences have been reversed or randomised is created and either interrogated separately from the protein database or concatenated to it. Identical search parameters are used for both interrogations and the FDR is calculated and expressed as a percentage, where FP is the number of proteins identified from the decoy database and TP is the number for the protein database, Equation 1.2.

$$FDR = \frac{FP}{(FP + TP)}$$

Equation 1.2 Calculation of False Discovery rate for large-scale proteomic data

For the purposes of this study calculation of FDR was based on the automatic generation of a decoy database by the PLGS software that contained one random protein entry for each original, concatenated to the original database. Each random sequence had the same amino acid composition as the original and would generate the same number of tryptic peptides but the amino acid sequence was randomised. Each random sequence was given the header ">RANDOM*x* Random Sequence *x*" where *x* denotes the entry number in the decoy database.

1.4.3.6 Protein identification from MS^E data

All the label-free MS^E data generated in this study were processed and proteins quantified using the Identify and Expression algorithms in PLGS releases 2.4 and later. As these algorithms are unique to the handling of MS^E data, more detailed review of the software for these later releases is discussed below (Li, Vissers et al. 2009). The collision energy ramps employed were either optimised on the instrument as part of this work or suggested by the instrument manufacturer as appropriate for MS^E acquisition. Protein quantification from LC-MS^E data is achieved by integration of the three dimensional volume of each peptide (time, *m/z* and intensity from all charge states) derived from the LC separation in the low energy scans.

The algorithms within PLGS utilise both the properties of peptides in the gas phase and liquid phase during the LC-MS^E acquisition to build a model of their behaviour. The algorithms refer to these properties multiple times during the database search to continually re-score and re-rank identified peptides and proteins based on how well they conform to the model. If a sufficient number of peptides are identified exceeding a minimum score, the in-built retention time and fragmentation models are refined, based on empirical data, and the peptides identified are given a score. All identified peptides are assigned back to their respective proteins which are then also given a score. The peptide precursors and product ions from the highest scoring protein are depleted from the dataset so that they are not incorrectly assigned in later protein identifications. This procedure is repeated until the pre-specified FDR is exceeded or the minimum protein score has not been met.

The data from an LC-MS^E analysis is collected in separate data functions. Function 1 contains low energy MS, function 2 elevated (ramped) collision energy MS and data from the reference compound for accurate mass correction. Initially **ion detection** is performed, with the first two functions processed (Silva, Denny et al. 2005) into an inventory containing the monoisotopic accurate mass-peak apex retention time information as well as the summed peak area (all isotopes from all observed charge states), average charge state, LC peak and RT start and end points.

The precursor and product ions are **time-aligned** using their peak apex (± 10% chromatographic peak width calculated from function 1). If multiple precursors elute from the LC, then the product ions observed will be associated to *all* potential precursors until depleted later in the processing as a result of being specifically attached to a protein.

The time-aligned precursor and product ions are **filtered** to remove low mass ions i.e. those resulting from low molecular weight tryptic peptides, in order to maintain the high specificity of the initial protein identifications. Ions below 750 Da under low energy and 350 Da in elevated energy are sidelined and recovered later in the processing stage. Additional filtering removes product ions higher in mass and intensity than the precursor.

A **randomised database** (one random entry/original) is created by PLGS if one has not been created by the user and a **presearch** is conducted using the same parameters as for the main search. The algorithm will then refine its model of properties of peptides in the gas/liquid phase using only the unambiguously identified peptides. If the number of peptides exceeds 250 with a minimum score, the algorithm will create a new model of behaviour for this set of peptides. This includes a retention time model (in real-time), product ion mass distribution and fragmentation model. A comparison is made between sequence length, charge state and precursor intensity to the summed number of product ions from that precusor. The summed product ion intensity is then related to the precursor ion intensity along with the total of identified consecutive and complementary *b* and *y* ions. Finally for the observed charge state(s) for each peptide, both the composition and sequence order of the peptide is compared to the sum of the ratio of *y*/*b* ion intensity.

Any product ions initially identified by **database search pass 1** as belonging to a precursor are compared against the results of the model described above. This predicts for a specific peptide of given length, charge state, intensity and the number of product ions that should be produced. At this stage all fixed modifications are considered based on the change in precursor and fragment ion *m/z*. Any tryptic peptides processed must have a low energy precursor mass error of < 10 ppm and at least 3 fragment ions with a < 20 ppm mass error. These are assigned a score before

comparison with the model, which then refines the score further. An additional adjustment of the score is made after a comparison with the predicted fragmentation model. This results in an increased score if a consecutive series of product ions are identified with a greater weighting if those ions are greater in *m/z* than the precursor.

The peptide scores may be increased in the event of further fitting to the models generated. These properties include retention time, product ions resulting from neutral loss of water and ammonia from certain amino acids, correlation between sequence and observed charge state, ratio of *y*/*b* product ion intensity resulting from presence of certain N-terminal amino acids and complementary N- and C-terminal product ions. The peptide with the highest score is given an arbitrary value of 1 and the others 0.

All peptides ranked with a value of 1 are assigned to their prospective protein(s) which are given a score based on the summed product ion intensity. The protein with the highest score is given an arbitrary value of 1 and the others 0. The selected protein is then assigned a score based on the matching tryptic peptides (with a rank of 1) and normalised to peptide length and summed intensity of the three most intense peptides. This score is refined based on the model behaviour of peptides from the presearch, which includes a comparison of assigned product ions compared to what would be expected for a peptide of the same length and precursor ion intensity.

After an examination of the observed ionisation efficiency distribution compared to that expected for a protein of similar molecular weight and concentration, the total number of continuous and complementary *y*/*b* ions with product ions observed at the preferred fragment sites, sequence coverage and the ratio of the total product ion intensity to precursor intensity the protein with the highest score is assigned and the precursor and product ions identified depleted from any subsequent database interrogations.

The process continues with proteins ranked, scored using the physicochemical model with the associated ions depleted until a protein score falls below the minimum specified or the FDR has been exceeded. The number of identified proteins allowed

from the decoy database is calculated by multiplying the specified FDR by the number of proteins identified from the database *before* the decoy protein was identified. Each time a decoy protein is identified the calculation is repeated until both numbers agree. At this point database search pass 1 has been completed.

Database search pass 2 generates a subset database containing only the proteins identified in search pass 1. Using accurate mass-retention time information, it tentatively identifies non-specific cleavage events, fragment ions generated in-source e.g. loss of water/ammonia or a PTM, to unassigned peptides that could have originated from the subset of proteins.

In the first iteration, in-source fragments and their respective product ions are assigned if they are time-aligned with their respective precursor. A second iteration identifies time-aligned precursor loss of H_2O and NH_3 and the third iteration identifies specified variable modifications including any missed cleavages. In the event of a peptide being tentatively assigned to more than one possible variant, the one scoring the highest number of matching fragments ions is additionally weighted if the losses are indicative of the modification. Where product ion information is insufficient to validate the potential variant, the modification is not reported.

In **database search pass 3** all remaining precursor information is used to search the full database, but now the total product ion intensity is allowed to exceed that of the precursor to account for potential in-source fragmentation of highly labile peptides. Multiple modifications on a single peptide are tentatively assigned, scored, ranked and assigned to their protein. The protein table is then re-scored, re-ranked and the newly assigned ions depleted from the dataset until the minimum protein score and/or FDR criteria are met.

1.4.3.7 Protein Quantification

Since most proteomic data sets contain some element of quantitation, software packages are available to incorporate identification of proteins with relative quantitation.

For gel-based studies, Progenesis SameSpots (Nonlinear Dynamics, Newcastle upon Tyne, UK), DeCyder (Uppsala, Sweden), PDQuest (Bio-Rad Laboratories, CA, USA) and MELANIE (Swiss Institute of Bioinformatics, Lausanne, Switzerland) software suites are commercially available.

Major software packages including Mascot, PLGS, ProteinPilot and Proteome Discoverer (Thermo Fisher Scientific Inc., MA, USA) include within their software suites the ability to process and quantify proteins from samples containing metabolically or chemically labelled peptides, detailed in Sections 1.6.1 and 1.6.2.

Commercial products for the processing of label-free quantitative data include Progenesis LC-MS (Nonlinear Dynamics, UK), Mascot Distiller (Matrix Science, UK), and ProteoIQ (NuSep, GA, USA). There are also in-house programs available for analysing spectral count data which will be reviewed in Section 1.6.3. The handling of quantitative data by PLGS will be reviewed in Section 1.6.4.

1.5 Gel-based and profiling strategies

Qualitative profiling of proteomic samples can be achieved from gel or solution-based samples. The aim of these studies is to confidently identify as many proteins as possible from the proteome. For comprehensive proteome coverage it is essential that sample preparation solubilises as many proteins as possible from the source material. A discussion of preparation strategies for a number of sample types including cytosolic extracts, membrane-associated, tissue and biopsy materials is presented below.

1.5.1 Sample preparation strategies

Proteomic samples can originate from a wide range of sources including bacterial, plant, viral, fungal, animal and human. The proteins may already be in-solution or require solubilising prior to separation and analysis. Where a solubilisation step is involved, the use of chaotrophs and/or detergents can significantly improve recovery

but consideration must be given to any potential interference that may occur in downstream analysis.

Prokaryotic cells may be lysed enzymatically, by osmotic shock, sonication or through passage through a French pressure cell. To generate a cytosolic extract the solution needs to be centrifuged at 100,000 *g* for at least one hour and the supernatant collected.

Differential centrifugation can be employed to fractionate cell compartments e.g. membrane, nucleus, cell wall and then, after washing to remove the soluble components, the pelleted material is solubilised. If detergents are used, one that can be removed by dialysis or the use of a protein precipitation step that leaves the detergent in-solution is preferable.

For tissue sections, the author has successfully modified a protocol for formalin-fixed, paraffin-embedded tissues (Nirmalan, Hughes et al. 2011) which uses Rapigest (Waters Corporation, USA) (Chen, Cociorva et al. 2007) an acid-labile surfactant designed to improve digestion efficiency, without causing modification to the proteins and compatible with MS analysis.

Homogenisation and cell disruption is required for mammalian cell lines in association with chaotrophs and/or detergents. The filter-aided sample preparation (FASP) protocol involves completely solubilising the proteome in sodium dodecyl sulphate (SDS), this is then exchanged for urea on a filtration device with a low molecular weight cut-off (10 kDa). Peptides are eluted after digestion through the filter and analysed using MS (Wisniewski, Zougman et al. 2009). This methodology has been used by the author to identify and quantify over 150 proteins from a rat liver biopsy using LC-MS^{E}.

Extracellular/secreted proteins can be selectively precipitated, typically using acetone or three-phase partitioning (Pacheco, Slade et al. 2011) and then re-solubilised prior to tryptic digestion and MS analysis.

An acetone precipitation step has been used in sample preparation strategies to remove interfering compounds prior to tryptic digestion and MS analysis. A selective modification can be shown to occur on proteins where residual acetone has been left in the sample post-precipitation (Simpson and Beynon 2010). Other protein precipitation strategies may be equally efficient at removing unwanted contaminants.

1.5.2 One-dimensional gel electrophoresis

Denaturing SDS polyacrylamide gel electrophoresis (PAGE) is frequently used in laboratories to separate proteins according to molecular weight, typically with a ± 10% accuracy. The proteins are first solubilised in a buffer containing SDS, β-mercaptoethanol, glycerol and bromophenol blue at 100 ^{o}C. The negatively charged SDS binds to the proteins (ratio 1.4 g SDS : 1 g protein) (Reynolds and Tanford 1970) with a net charge proportional to the length. Under the influence of an electric field the proteins will migrate through an acrylamide gel with the lower molecular weight species travelling furthest. A set of standard proteins undergoes separation in the same gel and an estimation of molecular mass can be inferred from the calibration curve generated based on the migration of the standards. The presence of SDS in the buffers allows a wide variety of proteins to be solubilised and then resolved. After visualisation of the proteins, bands may be excised and processed for tryptic digestion and MS analysis, Section 1.5.4. Unless the initial sample was relatively pure, most bands from a 1D gel would contain multiple proteins.

In native electrophoresis (non-denaturing) SDS is omitted from the buffers and proteins migrate through the gels based on surface charge and conformation, this is buffer pH-dependent. Any modification that affects the conformation may result in a change in exposed charge and thus differing migration in the gel. This approach is often used to resolve intact protein complexes, the components of which can then be identified by MS analysis.

1.5.3 Two-dimensional gel electrophoresis

Until relatively recently, it was accepted that a study of the proteome would require a fractionation or resolution step at the protein level prior to analysis. Typically this was achieved through the use of two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) (O'Farrell 1975). The proteome is solubilised, denatured and reduced prior to isoelectric focusing resolving the proteins according to their isoelectric point (pI). A second orthogonal separation by SDS-PAGE further resolves the proteins according to molecular mass. Some proteins may be fully resolved using this approach, although typically proteins of similar pI and molecular weight will occupy near-identical positions on the gel. The proteins can be visualised using dyes (e.g. Coomassie, silver, Sypro, fluorescent dyes) and the density of stain used to quantify the protein levels in each sample using software algorithms, Section 1.4.3.7. Proteins identified as changing in expression between proteomes may be excised for protein identification using MS analysis.

The preparation and running of 2D gels is a highly time consuming process, the variability between gels run simultaneously can be significant and the analysis of the gels themselves can take many days or weeks in order to obtain quantitative data. Inter-gel variation has been partly addressed by the use of fluorescent dyes with narrow excitation and emission bands, in a methodology described as difference in-gel electrophoresis (DIGE). Two proteomes and a control may be labelled with CyDye which attaches covalently to the epsilon amino group of lysine via an amide linkage (Unlu, Morgan et al. 1997). The samples are combined and electrophoresed together, scanned using the appropriate excitation/emission wavelength prior to analysis using appropriate software, Section 1.4.7.

Limitations of 2D gels include problems of solubility, particularly with membrane-associated proteins, species at the extremes of molecular weight and pI not being resolved and extraction efficiency of peptides from the gel can be low (typically around 20%). The focus has turned to the use of mass spectrometry-based approaches not only for the identification of proteins but also to determine quantitative changes in expression.

1.5.4 In-gel tryptic digestion

The identification of excised proteins within gel bands may be achieved by subjecting each band to an appropriate destaining procedure. This includes a reduction of disulphide bonds with dithiothreitol followed by chemical modification of each cysteine residue (alkylation with iodoacetamide) and subsequent tryptic digest to generate peptides terminating with an arginine or lysine residue. The peptides are extracted from the gel and typically analysed using liquid chromatography electrospray ionisation tandem mass spectrometry (LC-ESI-MS/MS) using a data dependent acquisition (DDA), Figure 1.8.

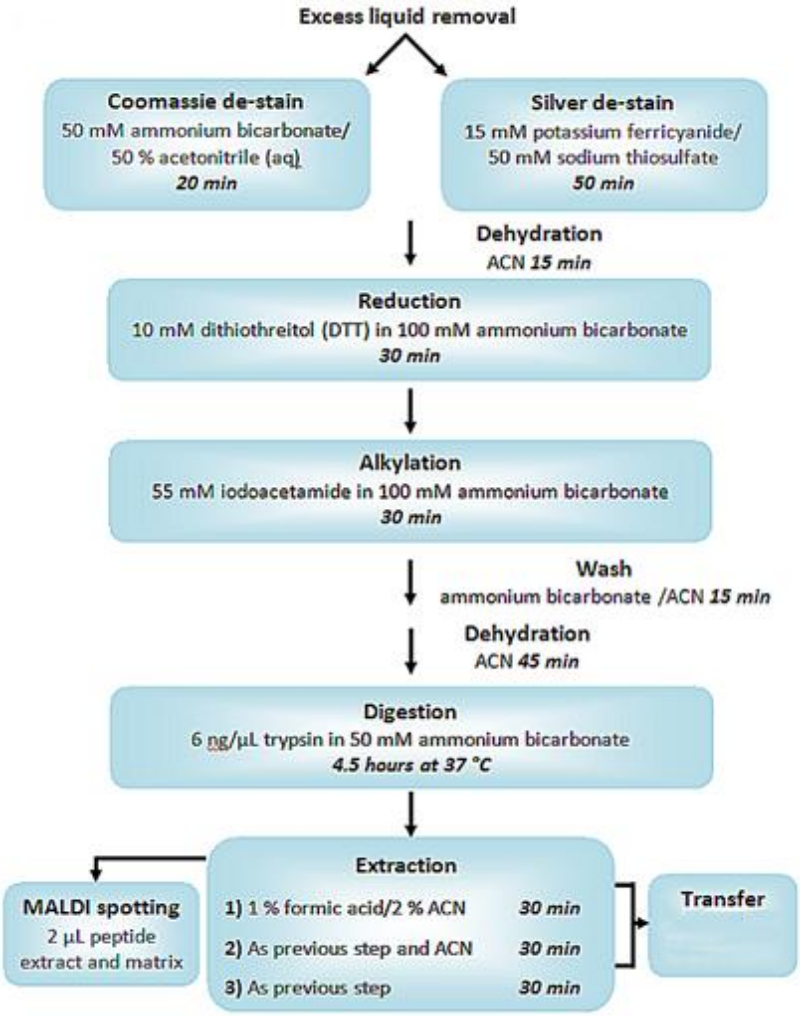


Figure 1. 8 : Automated procedure for an in-gel tryptic digestion
Key : ACN (acetonitrile)
The extracted peptides are transferred to a cooled tray, typically a 96-well microtitre plate.

1.6 Non-directed quantitative proteomic strategies

Quantitative proteomic strategies fall into one of two categories, relative or absolute. In absolute quantitation the amount of a protein in a sample is defined by a measure of its concentration in femtomoles or picograms per amount of cell material or copies/cell. Experimental approaches taken to determine absolute levels of a protein are described in Section 1.7.

Relative protein expression is defined in terms of up- or down-regulation and is expressed as fold-change. A number of approaches have been developed to assess the relative quantitation between proteomic samples these include metabolic and chemical labelling of proteins or peptides, label-free quantitation based on peak intensity in LC-MS chromatograms or spectral count. Some of the more commonly used approaches are detailed in the following sections.

1.6.1 Metabolic labelling in culture

In a proteomics study, metabolic labelling involves the incorporation of one or more stable, isotopically labelled elements during the growth phase of an organism. A study by the Chait group reported whole cell incorporation of ^{15}N into *Saccharomyces cerevisiae* in a study of protein abundance levels in wild type and mutant populations (Oda, Huang et al. 1999). Other organisms have been successfully metabolically labelled. These include *Caenorhabditis elegan*s and *Drosophila melanogaster*, feeding with ^{15}N-labeled *Escherichia coli* or yeast respectively (Krijgsveld, Ketting et al. 2003), rat (Wu, MacCoss et al. 2004) and *Arabidopsis thaliana* (Nelson, Huttlin et al. 2007). The cost implications of ^{15}N labelling are high and, due to the varying number of nitrogen atoms in a protein sequence, predicting the associated mass shifts may be challenging. The Mascot search engine (Matrix Science) now supports metabolic labelling by ^{15}N only or ^{15}N with ^{13}C.

Stable isotope labelling of amino acids in culture (SILAC) relies on the incorporation of isotopically labelled *amino acids* into proteins. The labelled amino acids,

typically lysine or arginine, are added to the growth medium for a number of cell turnover cycles to ensure a high level of incorporation of the label. The two cultures to be compared are grown under identical growth conditions but one uses the labelled growth medium. The proteomic samples are extracted from each culture, combined in equal quantities and analysed using mass spectrometry. The first use of SILAC (Ong, Blagoev et al. 2002) utilised deuterated leucine (d3-Leu) which caused a 3 Da mass increase for each leucine residue in each tryptic peptide.

SILAC can be used to compare more than two conditions, if appropriately labelled amino acids are available. As the number of labels increases, so does the mass spectral complexity. Not all cell types are suitable for SILAC since the essential amino acid must not be synthesised by the organism and, if arginine is used, a consideration of its conversion to proline must be incorporated into the experimental design.

1.6.2 Chemical labelling approaches

In 1999 isotope-coded affinity tags (ICAT) were described (Gygi, Rist et al. 1999a). A chemical label composed of three regions, a thiol-specific active group, a linker region with either 0 or 8 deuterium atoms and biotin was used. Two samples could be compared by labelling their cysteine residues. This was performed prior to combination of the samples and proteolytic digestion. The tryptic peptides were then avidin affinity purified, reducing the complexity of the sample for analysis. Quantitation, as in SILAC experiments, was determined in MS-mode.

An acid-cleavable biotin version of ICAT was developed. This was termed ciCAT (Hansen, Schmitt-Ulms et al. 2003; Li, Steen et al. 2003) and resolved the issues regarding differences in LC retention times due to the presence of deuterium by using nine ^{13}C atoms in the heavy label. This allowed removal of biotin prior to MS analysis, improving the quality of the MS/MS spectra obtained.

Cysteine is the least frequently observed amino acid in proteins. In the UniProtKB/Swiss-Prot protein knowledgebase release 2012_03, it composed only 1.36% of the 189,901,164 amino acid residues of the 535,248 sequence entries

whereas the random mathematical occurrence for it would be 3.28%. Cysteine has been demonstrated to be underrepresented in all organisms with its appearance correlating with organism complexity (Miseta and Csutora 2000). This lack of frequency can result in quantitation being performed on a small number of peptides per protein, and in approximately 10% of the cases quantitation cannot be determined due to an absence of Cys residues in the protein.

Amine-reactive isobaric tagging reagents (Ross, Huang et al. 2004) commercialised by Applied BioSystems and marketed as iTRAQ, allowed the multiplexing of 4 or 8 samples with all the peptides in a sample labelled via lysine side chains and their peptide N-terminus. The isobaric tags are composed of a reporter group (114 - 117 or 113 – 121 Da for the 4- or 8-plex reagent respectively), a mass balance region (increasing the mass of the tag to 145 or 305 Da in each case respectively) and an amine-reactive group. The proteomes for analysis are tryptically digested prior to labelling, after which they are combined in equal quantity prior to MS analysis. Quantitation is performed in MS/MS mode based on the measured reporter ion abundances.

iTRAQ labelling is expensive, time consuming and laborious and issues with reporter ion compression have been reported by a number of groups (DeSouza, Grigull et al. 2007; DeSouza, Romaschin et al. 2009; Patel, Thalassinos et al. 2009) limiting its effective dynamic range. Ow and co-workers suggested that background noise hinders quantification of low intensity reporter ions resulting in a decline in accuracy and precision and the suppression of dynamic range (Ow, Salim et al. 2009). Similarly mixed MS/MS from more than one precursor was shown to dampen large intensity differences in complex samples. Considerations suggested by the authors to improve the accuracy of quantitation included applying isotopic correction and use of high resolution instruments to prevent contamination of the 121 *m/z* reporter ion from phenylalanine immonium ion contribution.

Although chemical labelling has been widely used in quantitative proteomic studies, issues of incomplete labelling (chemical or metabolic), cost, increased sample requirements, time and complexity of sample preparation with the use of specific software for quantitation has resulted in developing interest in label-free approaches.

These approaches have in theory, no limit in terms of the number of samples that can be compared.

1.6.3 Label-free relative quantitation by spectral counting

Spectral counting approaches are based on the assumption that the number of identified MS/MS spectra collected from a tryptic digest using multidimensional protein identification technology (MudPIT) will increase proportionately to the protein concentration in a sample. As the abundance of a protein and the respective peptides increase so will the occurrence of MS/MS events, the observation of unique peptides and increased sequence coverage and this can be measured and compared over multiple data sets (Washburn, Wolters et al. 2001). Spectral count (total number of identified MS/MS spectra) has been shown to have a linear correlation over two orders of magnitude with relative protein abundance (Liu, Sadygov et al. 2004).

As spectral count data is collected in a data dependent acquisition it may be subject to bias towards more abundant peptides from abundant proteins, which may mask peptides which are present at low concentrations. At higher concentrations, detector saturation may occur further limiting the linear dynamic range. A linear response is assumed for each protein present and this does not account for variation between experimental runs.

A normalised spectral abundance factor (NSAF) has been developed to address some of these issues (Florens, Carozza et al. 2006; Zybailov, Mosley et al. 2006; Dong, Venable et al. 2007) and takes into consideration the fact that larger proteins tend to generate more peptides/spectra than smaller proteins. The NSAF for a given protein may be calculated from the number of spectral counts identifying that protein, divided by its length, divided by the sum of *all* spectral counts/length for *all* the proteins in the experiment.

1.6.4 Label-free relative quantitation by LC-MS and LC-MS^E using area under the curve

In a LC-MS analysis, ions having specific *m/z*, charge, retention time and intensity are recorded. As sample ion concentration increases, so would recorded signal intensity, within the linear dynamic range of the detector (Voyksner and Lee 1999). This approach has been tested for a tryptic digest generated from horse myoglobin (Chelius and Bondarenko 2002). LC-MS was performed on the digest (containing 10fmol – 100 pmol) and the average peak area under the curve (AUC) for five peptides calculated from three technical replicates. A linear correlation was observed (r^2=0.991), Figure 1. 9**.**

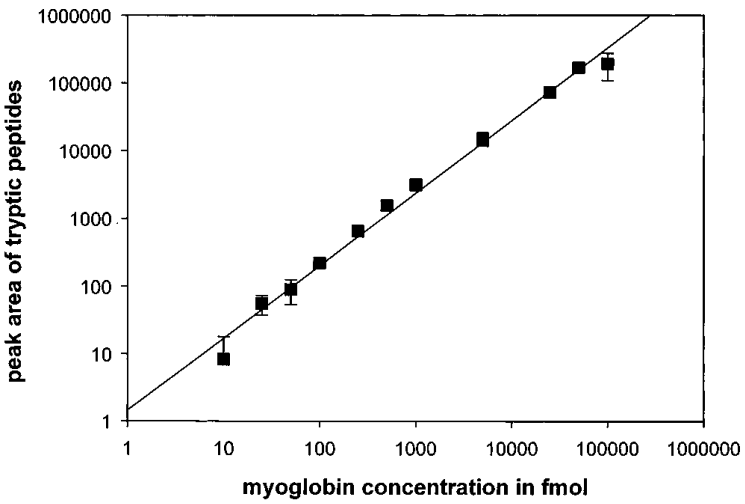


Figure 1. 9 : Correlation between average tryptic peptide LC-MS peak area and protein concentration.
The average peak area measurements were based on 5 peptides and 3 technical replicates (Chelius and Bondarenko 2002).

The correlation was also observed with a protein spiked at varying concentrations into complex mixtures (Chelius and Bondarenko 2002) and further improved (r^2=0.9978) by the use of normalised peak areas (Bondarenko, Chelius et al. 2002).

Relative quantitation using AUC is dependent on a number of factors including high resolution accurate mass measurement, reproducible LC retention times/peak shape, and assumes that sufficient data points have been collected across the curve to calculate the peak area. This can be an issue with DDA data since the MS scans

used for AUC measurement would not be collected during periods when MS/MS scans, used for protein identification, were being collected. Experimental variation between injections and samples can be reduced by data normalisation, using spiked internal standards or abundant housekeeping proteins in the sample, or total ion area.

In contrast to DDA, MS^E acquisitions collect alternating low energy MS scans which are also used for quantitation, with elevated energy scans for peptide identification. Assuming a sufficiently fast scan rate, LC-MS^E data can be used in a similar way to LC-MS for relative quantitation studies (Silva, Gorenstein et al. 2006). An average coefficient of variance (CV) has been reported of <15% for a data set (Silva, Denny et al. 2005). Levin *et al.* evaluated the LC-MS^E approach in terms of the quantitative performance which could be obtained using a four-protein mixture doped into low, medium and complex samples (Levin, Hradetzky et al. 2011b). The authors reported observed ratios to be correct to within 26.3% (±12.6% SD) with a limit of quantification of 61 amol/uL for simple mixtures, rising to 488 amol/uL for the more complex samples of depleted plasma and rat frontal cortex tissue extract. The majority of endogenous peptides were correctly reported as having no change in abundance between samples. Some compression of the protein ratios was observed and was assumed to be the result of ion supression.

1.7 Absolute protein quantitation

1.7.1 Label-free absolute quantitation

For some studies it may be necessary to calculate the *absolute* amount of a protein in a sample, which would typically be expressed as fmol or ng/unit of sample.

The protein abundance index (PAI), later improved to the empirically modified PAI (emPAI) was based on number of identified peptides divided by the number of theoretically observable peptides for each protein ((Rappsilber, Ryder et al. 2002; Ishihama, Oda et al. 2005) respectively).

Absolute protein expression (APEX) calculates the protein abundance per cell based on the proportionality between the observed peptides in a MUDPiT experiment and the abundance. Correction factors are used to estimate each protein abundance from the fraction of peptide mass spectra identified to it, corrected by the prior expectation of observation (Lu, Vogel et al. 2007).

The Hi3 absolute quantitation approach is based on the observation that the average MS signal from the three most intense peptides from a protein per mole (counts/fmol) was constant (CV ±10%). The relationship between protein concentration and the average Hi3 signal response was plotted with a linear curve fit with an r^2 value of 0.9939. By incorporating an internal standard, doped at a known concentration into a sample, it could be used to calculate a universal response factor (counts/mol) (Silva, Gorenstein et al. 2006). From this value, the concentration of all identified proteins could be calculated using the Hi3 approach. A six-protein standard was added to a complex sample containing human serum using alcohol dehydrogenase as the internal standard and the absolute quantification results calculated for each standard protein, Table 1. 1.

Levin *et al.* reported a linear dynamic range for quantitation using the LC-MS^E quantitative approach in complex samples of 2-2.5 orders approaching 3 orders for simpler mixtures (Levin, Hradetzky et al. 2011a).

Protein	**Theoretical concentration (pmol)**	**Calculated concentration (pmol)**	**Error**
Enolase	15.0	13.6	-9.3%
Serum albumin	12.5	12.9	3.3%
Alcohol dehydrogenase	**10.0**	**10.0**	**0.0%**
Phosphorylase B	6.0	6.5	8.7%
Hemoglobin (β)	5.0	4.7	-5.3%
Hemoglobin (α)	5.0	4.3	-13.3%

Table 1. 1 : Absolute quantitation calculated using LC-MS^E and Hi3 approach on proteins spiked in human serum.
Alcohol dehydrogenase was used as the internal reference (Silva, Gorenstein et al. 2006) in the six-protein standard mix.

Hi3 absolute and relative expression quantitation from LC-MS^E data have been used in this research. Single and multidimensional chromatography have been utilised for peptide separation prior to MS analysis.

1.7.2 Targeted absolute quantitation

An accepted gold standard of absolute quantitation would require the use of an internal standard for each analyte of interest incorporating a stable isotope label, usually ^{2}H, ^{13}C or ^{15}N.

The use of multiple reaction monitoring (MRM) for the quantitation of proteins was developed initially by the Gygi group (Gerber, Rush et al. 2003) based on the quantitation of tryptic peptides, using isotopically labelled analogues (AQUA peptides) which were added to each sample at known amounts. By quantifying at the peptide level, the protein concentration could then be estimated. The labelled AQUA peptide co-elutes (or near co-elutes) with the endogenous tryptic peptide and is thus subject to any matrix interference present in the subsequent MS analysis. In the mass spectrometer the peptides behave near-identically in all other respects other than the induced mass shift in the MS and MS/MS spectra resulting from the incorporation of ^{13}C and ^{15}N labels, typically resulting in a 6-10 Da mass difference.

Isotopically labelled AQUA peptides are marketed commercially by both Sigma Aldrich and Thermo Scientific and as SpikeTides™ by JPT Peptide Technologies. The peptides are chemically synthesised, purified and quantified which can make the cost of purchasing large numbers for a study prohibitive. An alternative approach termed QconCAT, involves expression of a recombinant protein comprising a large number of labelled concatenated peptide sequences which are then released by tryptic digestion in equimolar concentrations (Pratt, Simpson et al. 2006; Rivers, Simpson et al. 2007). Reports of incomplete cleavage at tryptic sites suggested further optimisation of the process may be required and the order of the peptide sequences was found to affect *in vitro* translation success rates (Mirzaei, McBee et al. 2008). The authors reported that there was not one superior approach when considering QconCAT and chemically synthesised peptides and that any method of quantification should be assessed on a case-by-case basis.

MS quantification of labelled and endogenous peptides can be achieved by extracting ion chromatograms for each peptide pair from an LC-MS experiment and using the peak ratios to calculate the peptide (and thus protein) concentration, Figure 1. 10**.**

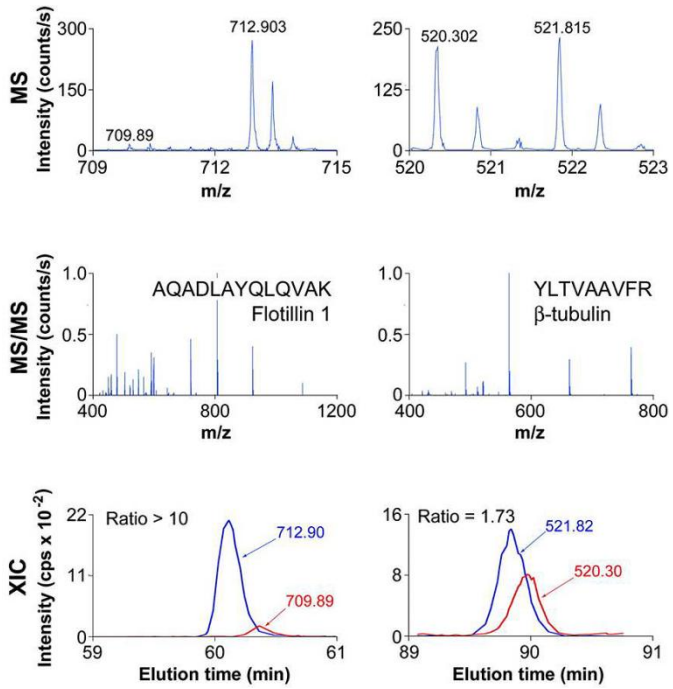


Figure 1. 10 : Representative MS, MS/MS, and extracted ion chromatograms for peptides from two proteins flotillin 1 and β-tubulin, quantified using isotopically labelled standards. Ion chromatograms (intensity vs. elution time) of the endogenous (Leu-), shown in blue and labelled (LeuD3) peptides, in red, were extracted from the MS scans and integrated (Foster, de Hoog et al. 2003).

An alternative approach using an MRM assay for each of the peptides of interest would combine the use of a triple or tandem quadrupole instrument with chromatographic separation of the peptides. The first quadrupole (Q1) performs mass selection on one of the peptides of interest, which would then be subsequently fragmented in the collision cell (Q2 on a triple quadrupole) with Q3 being used to select one or more of the diagnostic fragment ions (transitions) for the peptide. This improves selectivity allowing quantitation only from the peptide rather than any other species of similar *m/z* and retention time, Figure 1. 11**.** The process is repeated

for a list of peptides (endogenous and isotopically labelled) with their transitions for a given retention time window. Quantitation is determined from the relative response of the labelled analogue to the peptide of interest. An increase in sensitivity of up to 100-fold (attmol) can be achieved over conventional MS/MS in an MRM assay.

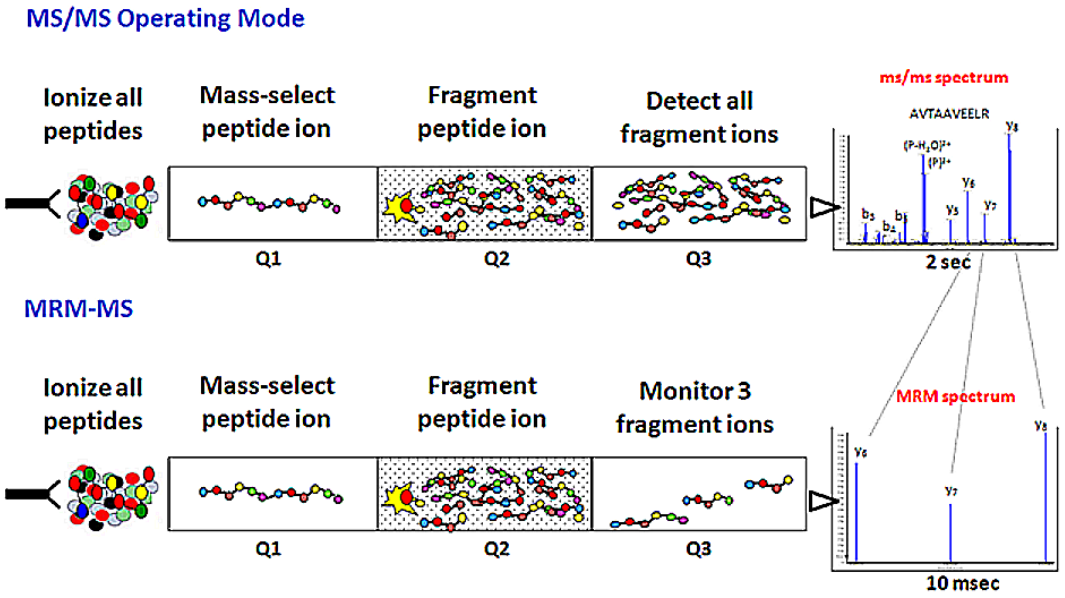


Figure 1. 11 : Comparison of MS/MS and MRM mode of analysis on a triple quadrupole instrument.
In an MS/MS experiment the selected peptide precursor is fragmented and all the fragment ions detected, whereas in an MRM experiment specific diagnostic fragment ions (transitions) are detected. The Q3 signal provides the area for quantification. Where more than one transition is selected for a given precursor ion, the accumulative counts are used for quantitation.
(http://www.broadinstitute.org/scientific-community/science/platforms/proteomics/mrm-multiple-reaction-monitoring).

A study by Anderson *et al.* using MRM assays for human (MARS-6 depleted) plasma proteins quantified L-selectin at 200 amol on-column (Anderson and Hunter 2006) demonstrating the increased sensitivity afforded by the MRM approach. The authors reported a dynamic range of 4.5 orders with typical CVs in the 2-7% range for abundant proteins rising to 11-22% for proteins near the limit of quantitation.

1.7.3 MRM assay design

Careful design of an MRM-based assay is required to ensure that the peptide measurements performed accurately reflect the *protein* concentration in the sample. From the discovery stage of a proteomics study, candidate proteins are identified and peptides chosen which are unique to the protein of interest and exhibit sufficient signal response in the mass spectrometer to obtain good sensitivity in the experiment. At least two peptides per protein should be selected, if possible not subject to post-translational or chemical modification. Suitable transitions are selected for each peptide to ensure specificity in the assay (Lange, Picotti et al. 2008; Kiyonami and Domon 2010; Holman, Sims et al. 2012).

A repository of observed peptides in proteomics experiments, generated from the literature, can be accessed at the Global Proteome Machine (http://www.thegpm.org) and interrogated to identify suitable peptides for MRM-based assays. Other publically available resources can be found at National Institute of Standards (http://www.nist.gov/index.html) and the Institute for Systems Biology ((https://www.systemsbiology.org/). Instrument vendors provide software to assist in the development of MRM assays including MRMPilot™ Software (AB Sciex) and Verify^{E} from Waters Corporation. Computational tools are available to predict suitable peptides (Mallick, Schirle et al. 2007; Fusaro, Mani et al. 2009) e.g. Skyline software from MacCoss group (Prakash, Tomazela et al. 2009) (https://skyline.gs.washington.edu/labkey/project/home/software/Skyline/begin.view).

1.8 Plasma Proteomics

The study of human body fluids has long been conducted by the health profession for the detection and quantification of biomolecules to aid in disease diagnosis. In particular urine, serum and plasma have been studied since they can be obtained in relatively large quantities. The basic premise that underlies their use is that the composition of a particular biofluid will be a reflection of tissues present in the body, whether healthy or diseased. In essence, a biofluid contains a snapshot of the condition of the human body at that point in time and thus is a potential mine of

information concerning the patient under study. It was therefore not too surprising that the field of proteomics research has been employed to identify potential biomarkers in human bodily fluids.

There are three functions or categories of biomarker: diagnostic, prognostic and predictive. A diagnostic biomarker can detect disease and may determine the extent/relapse of the disease state. A prognostic marker can be used to predict the disease history and can be used to forecast patient prognosis and disease recurrence. Predictive markers can be used to optimise treatment and therapies on a personal rather than population level (Winter, Yeo et al. 2013).

A typical protein biomarker study would consist of a series of defined elements, each with its own logistical and analytical requirements. The five phases that a biomarker needs to pass through to produce a useful tool for population screening have been detailed in (upper).

The first (discovery) stage of the biomarker pipeline would begin with a relatively small number of samples for comprehensive analysis to identify potential protein candidates, (lower). Taking a number of months, this would be followed by a verification and validation stage carried out on a number of selected *target* proteins (identified at the discovery stage) using a larger number of samples. Many of the candidates would be rejected at this stage, which can take many months to over a year to complete. The remaining biomarker/s would then have antibodies raised against them for use in further assays, usually Enzyme-Linked ImmunoSorbent Assay-based *(*ELISA). This would typically involve an orthogonal assay to those used in the verification and validation stage and would be taken forward in collaboration with commercial partners for clinical validation, taking many years and many thousands of samples to complete. Finally, approval from the regulatory bodies would then be sought for the assay to be rolled out.

Phases:	Phase 1 Preclinical Exploratory	Phase 2 Clinical Characterization & Assay Validation	Phase 3 Clinical Association: Retrospective Repository studies	Phase 4 Clinical Association: Prospective Screening studies	Phase 5 Disease control
Objective	Target Biomarker Identification, Feasibility	Study assay in people with & without disease	Case-control studies using repository specimens	Longitudinal studies to predict disease	Clinical use
Site	Biomarker Development Lab	Biomarker Validation Lab	Clinical Epidemiologic Centers	Cohort Studies	Community
Design	Cross-sectional	Cross-sectional	Case-control	Prospective	RCT
Sample Size	Small	Small	Modest	Medium	Large
Validity	Content & construct validity	Criterion validity	Predictive validity	Efficacy of strategy	Effectiveness
Result	Assay precision reliability, sensitivity	Reference limits, intra-individual variation	Screening characteristics, true & false+ rates	ROC analyses	No.-needed-to screen/treat

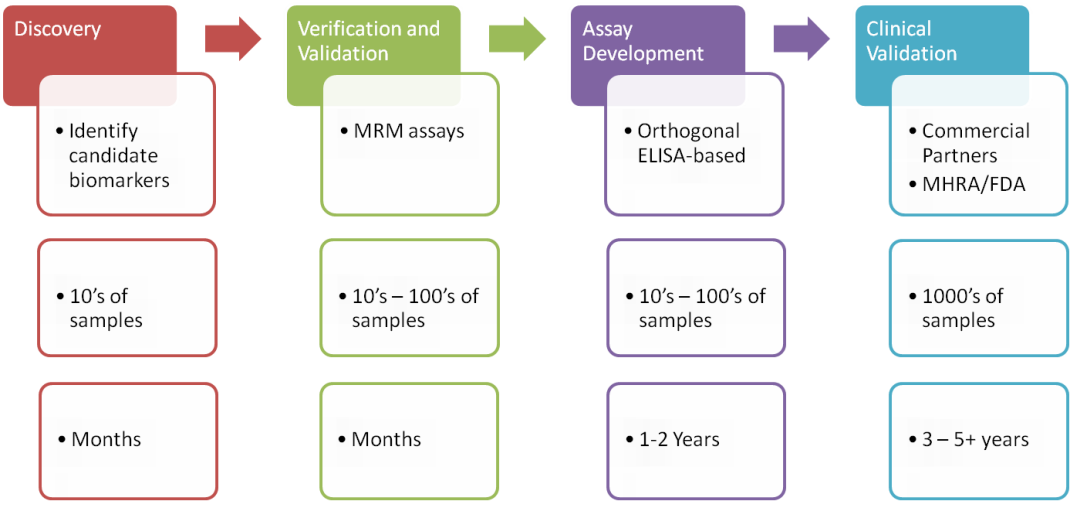


Figure 1. 12 : Biomarker pipelines indicating the stages from candidate discovery to clinical application.
The number of candidate proteins decreases from left to right as the number of patient samples increases, adapted from http://www.vicc.org/jimayersinstitute/devt/ and (Pepe, Etzioni et al. 2001).

1.8.1 Challenges of plasma proteomics

The dynamic range of proteins present in plasma exceeds 10 orders of magnitude (Anderson and Anderson 2002) whilst being composed of potentially many thousands of protein species. In the 2002 Anderson study, 289 plasma proteins were quantified using a variety of techniques, including immunoassays and since then the number of proteins identified has increased. By 2005 this number had increased to 960 proteins identified from over 6,900 peptides analysed using a variety of sample preparation and proteomics methods (Deutsch, Eng et al. 2005) and in 2008, 697 non-immunoglobulin plasma proteins were stringently identified and validated (Schenk, Schoenhals et al. 2008). In 2011, the Aebersold group reported identifying 20,433 distinct peptides from which they inferred a non-redundant set of 1929 proteins from 91 LC-MS/MS data sets (Farrah, Deutsch et al. 2011). As the numbers of proteins identified continues to grow, it is clear that full characterisation of the plasma proteome still eludes the proteomics community. The studies described utilised a variety of techniques to identify large numbers of proteins and involved many tens to hundreds of hours of experimental and instrumental effort. Whilst this is invaluable in terms of deep proteome coverage, this intensity of effort would not be practical on the large number of plasma samples required to undertake a biomarker discovery experiment.

A reduction in the complexity of the plasma proteome can be achieved by either depleting high abundance proteins or utilising an enrichment step for specific classes of biological molecules. More than 50% of human proteins are thought to be glycosylated (Van den Steen, Rudd et al. 1998) with glycoproteins playing an important role biomarker discovery (Durand and Seta 2000; Freeze 2001; Spiro 2002). A number of approaches have been used for glycoprotein enrichment from plasma samples. Lectin affinity chromatography (Yang, Harris et al. 2006) or solid phase extraction using hydrazide chemistry capture are effective methods targeting N-linked glycoproteins (Pan, Wang et al. 2006). Glycocapture methodologies using hydrazide residues chemically linked to beads can be used to enrich at the peptide (after digestion) or protein level. The use of magnetic beads has been reported with detection of glycoproteins in depleted plasma at levels of 10-100 pmol mL^{-1}

demonstrating improved efficiency over macroporous beads (Berven, Ahmad et al. 2010).

1.8.2 Depletion of abundant proteins from plasma

Recent improvements in mass spectrometry design have increased instrumental dynamic range on a discovery proteomics-based platform to around 10^4. We are faced with a complex fluid that, if analysed in its native state, produces identification of a very small percentage of the total protein complement. Early studies of 2D-PAGE-resolved plasma typically identified <100 proteins although many more spots were visualised on the gels. Use of immunoaffinity chromatography to deplete plasma of a number of high abundance proteins was first reported by Pieper *et al.*, using a technique they termed multicomponent immunoaffinity subtraction chromatography using purified, immobilised polyclonal antibodies to albumin, immunoglobulin G, immunoglobulin A, transferrin, haptoglobin, α-1-antitrypsin, hemopexin, transthyretin, α-2-HS glycoprotein, α-1-acid glycoprotein, α-2-macroglobulin and fibrinogen (Pieper, Su et al. 2003). After depletion, a reported increase of between 350 and 400 Coomassie Blue G-250 proteins spots was observed.

Commercially available systems (spin or liquid chromatography-based) offer the depletion of single or multiple proteins or alternatively selective protein enrichment from bodily fluids such as plasma, cerebrospinal fluid and urine, Table 1. 2**.** The Multiple Affinity Removal System (MARS™, Agilent Technologies) offers depletion of 6, 7 or 14 abundant plasma proteins, Figure 1. 13**.** Enrichment of the non-depleted proteins using these approaches, as determined by MS/MS spectral counting was reported to be approximately 4-fold for both the MARS-7 and MARS-14 fractions (Tu, Rudnick et al. 2010) with a 25% increase in the number of proteins identified (220) compared to raw plasma. Twenty three low abundance proteins, concentration in plasma <10 ng mL^{-1}, were identified but they constituted less than 6% of the total depleted sample including an isoform D of proteoglycan-4, tenascin-X and thyroglobulin (1.0, 3.8 and 5.1 ng mL^{-1} in raw plasma respectively).

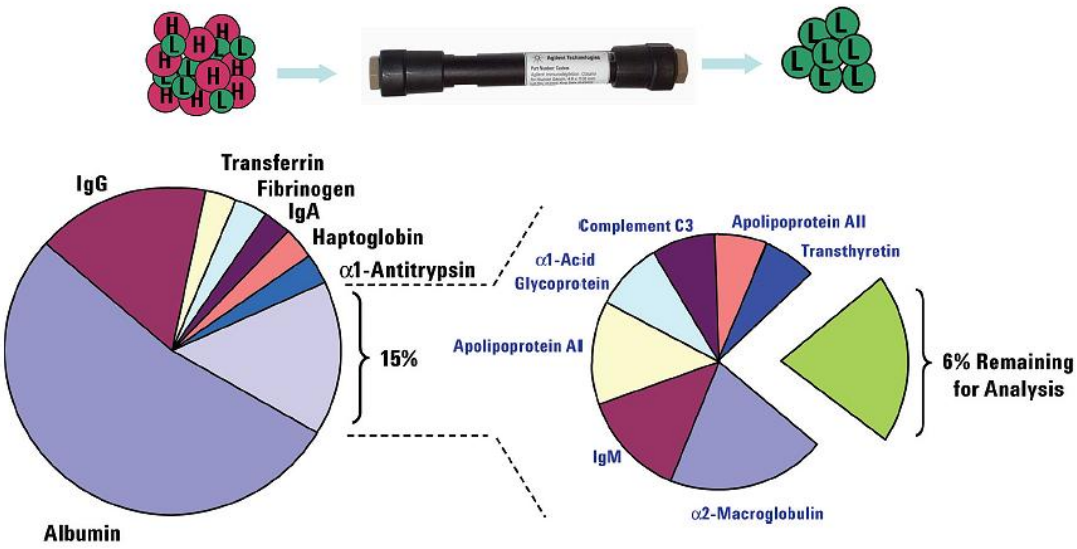


Figure 1. 13 : Proteins immunodepleted by the Multiple Affinity Removal System.
On the left, the Human 7 depleted proteins are indicated and (right) the additional proteins depleted using the Human 14 system, adapted from (Mrozinski, Zolotarjova et al. 2008)**.**

Depletion System	Protein Affinity Ligands	Manufacturer/Distributor
ProteoPrep® Blue Albumin & IgG Depletion Kit	Albumin and IgG. Medium is a mixture of two media, a blue dye conjugated to an agarose base matrix and Protein G agarose	Sigma-Aldrich
ProteoPrep® Immunoaffinity Albumin & IgG Depletion Kit	Mixture of two beaded immunoaffinity mediums containing recombinantly expressed, small single-chain antibody ligands	Sigma-Aldrich
MARS™ Hu-6	Albumin, IgG, antitrypsin, IgA, transferrin and haptoglobin	Agilent Technologies Inc.
MARS™ Hu-7	As Hu-6 plus fibrinogen	
MARS™ Hu-14	As Hu-7 plus α2-macroglobulin, α1-acid glycoprotein, IgM, apolipoprotein AI, apolipoprotein AII, complement C3 and transthyretin	
Proteominer	Combinatorial bead-based library of hexapeptide ligands for protein enrichment	Bio-Rad Laboratories
ProteomeLab™ IgY-12	Refer to Section 3.1.	Beckman Coulter Inc./GenWay Biotech Inc.
Seppro® IgY-14	Albumin IgG, α1-Antitrypsin IgA, IgM, Transferrin, Haptoglobin, α2-Macroglobulin, Fibrinogen, Complement C3, α1-Acid, Glycoprotein (Orosomucoid), HDL (Apolipoproteins A-I and A-II) and LDL (mainly Apolipoprotein B)	Sigma-Aldrich
Seppro® Supermix	Ligands developed by immunising chickens with flow-through fraction from IgY12/14 column	Sigma-Aldrich
ProteoPrep® 20	Combination of small recombinant immunoaffinity ligands and antibodies remove 97-98% of the total protein from plasma	Sigma-Aldrich

Table 1. 2 : Plasma depletion or fractioning systems available commercially.

The MARS-6™ (Agilent Technologies) and Seppro MIXED 12 (GenWay Biotech Inc.) depletion systems were compared and the flow-through and bound fractions were tryptically digested and analysed using 2D chromatography (SCX-RP) MS/MS (Gong, Li et al. 2006). A total of 529 plasma proteins were identified across all the fractions with a FDR of 1%. Over 300 proteins were identified from the MARS and Seppro immunodepleted fractions (307 and 329 respectively of which 238 were common to both fractions) with 194 and 217 proteins in the respective bound fractions with 150 proteins in common. In common with a number of previous studies (Mehta, Ross et al. 2003; Huang, Stasyk et al. 2005), the authors reported a population of low molecular weight peptides/proteins associated with the high abundance high molecular weight species (albumin, transferrin and α1-acid glycoprotein), acting as molecular carriers or sponges. To test this theory a co-immunoprecipitation with antibody against human albumin was performed. There was a 65% overlap between the proteins identified in the co-immunoprecipitant and the two bound fractions from the depletion systems providing supporting evidence for interactions in plasma between albumin and circulating low molecular weight species.

Many of the technical challenges of identifying disease biomarkers from plasma remain unresolved but the importance of a reproducible and robust depletion strategy to partly address the dynamic range issue has long been acknowledged (Jacobs, Adkins et al. 2005; Zolotarjova, Martosella et al. 2005; Brand, Haslberger et al. 2006; Gong, Li et al. 2006; Tu, Rudnick et al. 2010).

1.8.3 Individual patient or pooled plasma for biomarker studies?

The prohibitive cost of sample preparation (typically depletion) and multi-dimensional chromatography with MS/MS analysis has resulted in the use of pooled plasma samples from clinical groups rather than individual patient samples in many studies published to-date. This is not surprising given that the cost of Seppro IgY-14 LC-based depletion for one sample can cost over £40 with instrument time running into many hours to obtain comprehensive quantitative coverage of the proteome. Additional time would then be required for technical and/or biological replicates to be obtained driving the experimental time required to a day or more.

Few studies have compared the effect of pooling plasma or serum samples on the observed protein profiles. A study using surface enhanced laser desorption/ionisation time of flight (SELDI-ToF) mass spectrometry to analyse individual and pooled sera from invasive aspergillosis cases reported the loss (below peak detection limit) of 14 out of 35 discriminating biomarker peak clusters, identified in individual samples that were absent from the pooled samples. Thirteen new peak clusters were only observed in the pooled sample (Sadiq and Agranoff 2008). A typical example of the change in intensity of peak clusters is shown in Figure 1. 14.

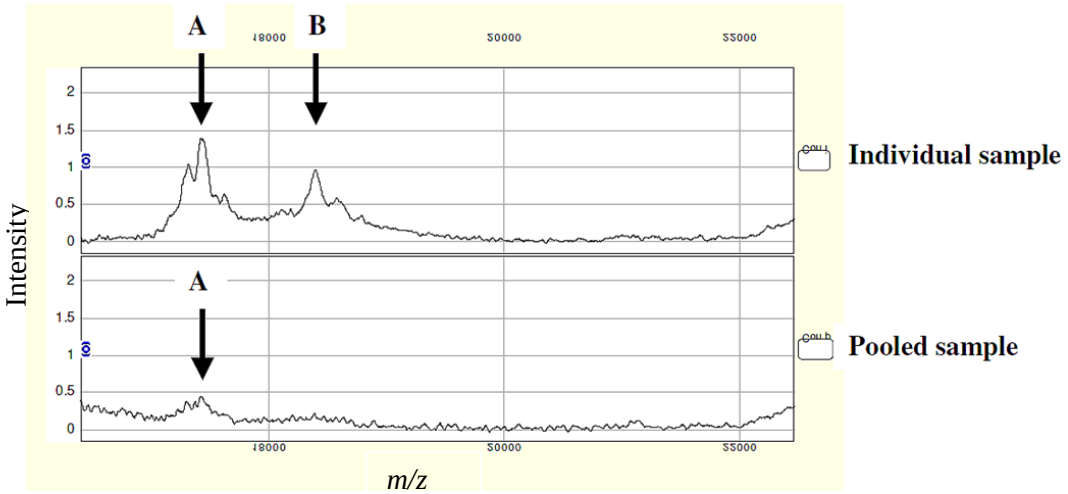


Figure 1. 14 : Comparison of individual and pooled sera analysed using SELDI-ToF showing the effect on measured peak intensity.
Peak A was observed in the pooled sample at a lower intensity and peak B below the peak detection threshold, extracted from (Sadiq and Agranoff 2008).

1.8.4 Requirements for a good biomarker study.

A primary requirement for any biomarker study to be successful would be a source of well characterised and defined clinical samples representative of the target population (Surinova, Schiess et al. 2011). This would include information on patient history, disease status at the time of sampling, presence of comorbidities (other diseases), medication status and lifestyle choices (alcohol and smoking). Age, gender, ethnicity and body mass index (BMI) should also be considered when selecting suitable samples for the study (Pepe, Etzioni et al. 2001).

In 2005, Rai *et al.* published guidance as part of the HUPO Plasma Proteome Project for specimen collection and handling (Rai, Gelfand et al. 2005). The authors described a number of pre- and post-analytical variables that could occur, described the importance of suitable quality control and reference materials and suggested standardised methods based on empirical observations to reduce variability in the proteomic analyses (Rai, Gelfand et al. 2005). The analytical techniques employed in the biomarker study should be robust, reproducible and the technical variance in the measurements should be significantly less than the expected biological variance within the target population.

The importance of collaborating with leading clinicians is vital to a project, to prevent a biomarker study becoming a list of up- and down-regulated proteins with no biological insight. A comparison of control versus disease samples will always yield a set of differentially controlled proteins but the majority of these are neither clinically relevant nor sufficiently selective for a diagnostic test. A review of 2D gel-based studies published in the journal *Proteomics* (volumes 4-6, 2004-2006) by Petrak and co-workers produced a list of repeatedly identified proteins from human, rat and mouse sources regardless of the disease under study (Petrak, Ivanek et al. 2008). From a list of 4,700 regulated proteins, a list of the TOP 15 was produced of which 70% of the articles studied reported at least two of the proteins from the list. Approximately 25% of the articles listed at least 5 of the TOP 15 proteins with the most frequently identified regulated protein being enolase 1 (alpha-enolase), a glycolytic enzyme. The list contained predominantly heat shock, stress and cytoskeletal (keratins) components. The authors concluded that these proteins/protein families could be representative of cellular sensors regulated as a response to stimuli.

A similar study by Wang *et al.*, based on proteins predominantly identified by 2D gel-MALDI techniques from 66 publications (2000-2007) encompassed 5 organisms including fruit fly and worm. The study concluded that 44 proteins were repeatedly observed which accounted for over 23% of the identified species. Twenty eight (37.6%) of frequently detected protein families were from heat shock, actin, ribonuclear and proteasome families (Wang, Bouwman et al. 2009).

Clinical insight into the condition under study can help to rapidly reduce the original number of candidate biomarkers based on lack of selectivity and specificity. The inclusion of a third condition i.e. control vs. condition A vs. condition B can also assist in the deselection of non-clinical candidates.

1.8.5 Gestational diseases of pregnancy

During pregnancy the maternal immune response must be modified to allow the foetus to develop in the uterus, from initial implantation through the gestational period until successful delivery. Up to 10% of pregnancies can be affected significantly by maternal or perinatal morbidity and mortality (or both) caused by dysfunctions and abnormalities (Shankar, Cullinane et al. 2004).

Pre-eclampsia is one of a number of gestational diseases which can lead to severe complications and death. It is most common amongst women in their first pregnancy and risk factors include diabetes, kidney, autoimmune diseases and chronic hypertension (Mabie, Pernoll et al. 1986; Hernandez and Cunningham 1990; Qiu, Williams et al. 2003). The risk factor for pre-eclampsia ranges from 2-8% according to the World Health Organization International Collaborative Study of Hypertensive Disorders of Pregnancy (1988) rising to 15-25% in those with gestational hypertension (Saudan, Brown et al. 1998). Women at risk of pre-eclampsia would be assessed at the first antenatal visit based on clinical history and examination. Routine screening for pre-eclampsia is based on measurement of blood pressure and proteinuria (excess levels of protein in urine).

Currently no effective single test exists for the prediction of onset of pre-eclampsia in early pregnancy. Where women are identified at risk of pre-eclampsia, close monitoring and pharmacological intervention with aspirin, calcium supplementation where necessary and lifestyle changes such as rest, exercise and diet are recommended (Thangaratinam, Langenveld et al. 2011).

Trisomy 21 (T21) is a condition in which a person has 47 chromosomes instead of 46, resulting from an additional chromosome 21. T21 is the cause of 90% of Down's syndrome cases and is the most common cause of birth defect, affecting the

brain and body development for which there is no cure. There are currently no known environmental risk factors but with increasing maternal age the risk factor increases significantly from 1 in 1500 at age 20 to 1 in 100 at age 40. Of approximately 700,000 births in the UK in 2009, 765 were born with Down's syndrome (http://www.nhs.uk/Conditions/pregnancy-and-baby/pages/screening-amniocentesis-downs-syndrome.aspx).

Early screening in the UK can involve either a blood test or a combined blood test with ultrasound scan (nuchal translucency – NT) to indicate if the pregnancy is above or below the national cut-off for risk. The median values obtained from an NT scan in an unaffected pregnancy would be 1.5 mm compared to 2.7 mm in Down's syndrome, Figure 1. 15 (Bestwick, Huttly et al. 2010). Confirmatory genetic tests would be offered in cases where the risk was indicated above the national cut-off, such as chorionic villus sampling (1%-2% risk of miscarriage) or amniocentesis (0.5-1%).

Figure 1. 15 : Nuchal translucency measurement measuring the fluid under the skin at the back of the neck.
Down's syndrome is associated with decreased muscle tone (hypotonia, or floppy infant) which is reflected in the increased fluid detected early in pregnancy.

The blood test typically screens for the levels of the hormone free beta-hCG (human chorionic gonadotrophin) and placenta associated plasma protein A (PAPP-A) which can show increased and decreased levels, respectively, in the cases of Down syndrome compared to normal pregnancies. Using these two proteins with the NT scan, 90% of Down's cases can be identified with less than a 5% false positive rate. The inclusion of other proteins in the assay such as alpha-fetoprotein, unconjugated estriol (decreased levels) and inhibin A (increased) increases the specificity and sensitivity, reducing the number of pregnancies requiring genetic testing. The screening tests will be discussed in further detail in Section 4.1.1.

1.8.6 Proteomics in reproductive medicine

The wide range of physicochemical and biological changes taking place during pregnancy and the combination of foetal and maternal proteins circulating in the blood means that plasma could serve as an ideal material from which the identification of biomarkers for diseases of pregnancy could be achieved. Modern proteomic platforms with their high sensitivity and selectivity could allow the comprehensive characterisation of the gestational proteome and these have been exploited in a number of studies relating to reproductive medicine.

Cervical vaginal fluid (CVF), which is present in the entire reproductive tract, not only contains low molecular weight species but also proteins, enzymes and cells. From the fluid, analysed using 2D-PAGE over 400 protein spots were identified from 5 healthy, pregnant women at term, of which 157 were common to all the gels (Di Quinzio, Oliva et al. 2007). Tang and co-authors reported that almost half of the proteins they identified, using 2D gel MALDI-TOF/TOF analysis from CVF, were plasma components and they were unable to differentiate symptomatic from asymptomatic sufferers of vulvovaginal candidiasis (Tang, De Seta et al. 2007). The profiles from CVF taken from women at 16-22 weeks gestation from normal pregnancies were shown to have a number of proteins common to both plasma and amniotic fluid when analysed using 1D gels and MudPIT approaches (Dasari, Pereira et al. 2007).

A comprehensive analysis of amniotic fluid utilising three different fractionation techniques was performed by Cho *et al.* They identified 842 non-redundant proteins which included most of the currently used biomarkers for pregnancy-associated disorders such as preterm delivery, intra-amniotic infection, and chromosomal anomalies of the foetus (Cho, Shan et al. 2007)

Plasma would be a preferred matrix for a diagnostic test, over the fluids described above, since blood samples are routinely taken during the term of a pregnancy and can be collected locally rather than requiring a hospital visit. They pose little risk to the mother or foetus and are minimally invasive. Protein biomarkers identified could be added to any current ELISA-based tests in use, reducing the overall cost of the diagnosis. A biomarker could also be used on its own to indicate a risk factor or used in conjunction with existing markers to improve sensitivity and/or selectivity.

This study assessed the suitability of current proteomics approaches for the identification of plasma biomarkers for two diseases of pregnancy.

1.9 Project aims

This study aims to bring together a number of newly developed mass spectrometric and liquid chromatographic methodologies to characterise the plasma proteome. The use and effectiveness of depletion of highly abundant proteins has been explored and the applicability of the techniques in the discovery stage of a biomarker study for two gestational diseases of pregnancy, trisomy 21 and early onset pre-eclampsia, evaluated.

NanoLC with nanoES ionisation mass spectrometry was selected for this work in preference to gel-based approaches due to its high resolving power at the peptide level, both chromatographic and mass spectrometric, providing increased confidence in both peptide and protein identifications with improved quantification reliability. An alternative approach would have been to use LC-MALDI which would have required the use of chemical tags, such as iTRAQ for quantitation. Poor resolution, sensitivity and mass accuracy of the fragment ions in MALDI MS/MS experiments combined with a wide precursor mass selection window are factors contributing to

the lower confidence in protein identification and quantification results obtained from these types of experiments.

The hypothesis that underpins this work is that one or more proteins are expressed at altered levels in the plasma of women during the first trimester of pregnancy that can be used diagnostically for the obstetric conditions pre-eclampsia or trisomy 21.

The main aims of the research were to:-

- Develop data independent acquisition (MS^E) on a Q-ToF Ultima Global instrument and evaluate performance against a commercial offering, the Synapt HDMS system.

- Utilise a data independent acquisition approach to characterise maternal plasma samples, from individual patients, depleted of 12 highly abundant proteins.

- Develop an automated LC-based methodology for the depletion of 14 highly abundant proteins and quantitatively characterise the fractionated plasma from control, pre-eclampsia and trisomy 21 gestations from individual patient samples.

- Evaluate the use of online 2D chromatography for the profiling and quantitative analysis of pooled, depleted maternal plasma from control and trisomy 21 gestations.

1.10 Research papers

Patel, Nisha A., Crombie, Andrew, **Slade, Susan E.**, Thalassinos, Konstantinos, Hughes, Chris, Connolly, Joanne B., Langridge, James, Murrell, J. Colin and Scrivens, James H.. (2012) *Comparison of one- and two-dimensional liquid chromatography approaches in the label-free quantitative analysis of methylocella silvestris.* Journal of Proteome Research, Vol.11 (No.9). pp. 4755-4763. ISSN 1535-3893

Dawkar, Vishal V., Chikate, Yojana R., Gupta, Vidya S., **Slade, Susan E.** and Giri, Ashok P.. (2011) *Assimilatory potential of Helicoverpa armigera reared on host (Chickpea) and nonhost (Cassia tora) diets.* Journal of Proteome Research, Vol.10 (No.11). pp. 5128-5138. ISSN 1535-3893

Rodgers, U. R., **Slade, Susan E.**, Scrivens, James H. and Clark, I. M. (2011) *Identifying substrates for MMP-28 using a label-free proteomics approach.* In: Autumn Meeting of the British-Society-for-Matrix-Biology, Univ East Anglia (UEA), Norwich, England, 6-7 Sep 2010. Published in: International Journal of Experimental Pathology, Vol.92 (No.3). A29-A29.

Ehsan, S., **Slade, Susan E.**, Balls, G., Jones, D. J. L., Butt, H. Z., London, N. J. M., Sayers, R. D. and Bown, M. J.. (2011) *Discovery of plausible "candidate biomarkers of AAA" at protein level : a pilot study.* British Journal of Surgery, Vol.98 (Suppl. 2). p. 18. ISSN 0007-1323

Pacheco, Luis G. C., **Slade, Susan E.**, Seyffert, Núbia, Santos, Anderson R., Castro, Thiago L. P., Silva, Wanderson M., Santos, Agenor V., Santos, Simone G., Farias, Luiz M., Carvalho, Maria A. R., Pimenta, Adriano M. C., Meyer, Roberto, Silva, Artur, Scrivens, James H., Oliveira, Sérgio C., Miyoshi, Anderson, Dowson, Christopher G. and Azevedo, Vasco. (2011) *A combined approach for comparative exoproteome analysis of Corynebacterium pseudotuberculosis.* BMC Microbiology, Vol.11 . article no. 12. ISSN 1471-2180

Nicolaides, K., **Slade, Susan E.**, Breslin, Eamonn, Scrivens, James H. and Thornton, S. (2011) *Identification of predictive biomarkers for pre-eclampsia by plasma proteomic profiling.* In: 30th Annual Meeting of the Society-for-Maternal-Fetal-Medicine, Chicago, Ilinois, 01-06 Feb 2010. Published in: American Journal of Obstetrics and Gynecology, Vol.204 (Suppl.1). S294.

Grabenauer, Megan, Wyttenbach, Thomas, Sanghera, Narinder, **Slade, Susan E.**, Pinheiro, Teresa J. T., Scrivens, James H. and Bowers, Michael T.. (2010) *Conformational stability of Syrian hamster prion protein PrP(90-231).* Journal of the American Chemical Society, Vol.132 (No.26). pp. 8816-8818. ISSN 0002-7863

Hilton, Gillian R., Thalassinos, Konstantinos, Grabenauer, Megan, Sanghera, Narinder, **Slade, Susan E.**, Wyttenbach, Thomas, Robinson, Philip J., Pinheiro, Teresa J. T., Bowers, Michael T. and Scrivens, James H. (2010) *Structural analysis of prion proteins using drift cell and traveling wave ion mobility mass spectrometry.* In: Asilomar Conference on Ion Spectroscopy, Pacific Grove, CA, October 16-20, 2009. Published in: Journal of The American Society for Mass Spectrometry, Vol.21 (No.5). pp. 845-854.

Patel, Vibhuti J., Thalassinos, Konstantinos, **Slade, Susan E.**, Connolly, Joanne B., Crombie, Andrew, Murrell, J. C. (J. Colin) and Scrivens, James H.. (2009) *A comparison of labeling and label-free mass spectrometry-based proteomics approaches.* Journal of Proteome Research, Vol.8 (No.7). pp. 3752-3759. ISSN 1535-3893

Thalassinos, Konstantinos, Grabenauer, Megan, **Slade, Susan E.**, Hilton, Gillian R., Bowers, Michael T. and Scrivens, James H.. (2009) *Characterization of phosphorylated peptides using traveling wave-based and drift cell ion mobility mass spectrometry.* Analytical Chemistry, Vol.81 (No.1). pp. 248-254. ISSN 0003-2700

Jackson, Anthony T., **Slade, Susan E.**, Thalassinos, Konstantinos and Scrivens, James H.. (2008) *End-group characterisation of poly(propylene glycol)s using electrospray ionisation-tandem mass spectrometry (ESI-MS/MS).* Analytical and Bioanalytical Chemistry, Vol.392 (No.4). pp. 643-650. ISSN 1618-2642

Clokie, Martha R. J., Thalassinos, Konstantinos, Boulanger, Pascale, **Slade, Susan E.**, Stoilova-McPhie, Svetla, Cane, Matt, Scrivens, James H. and Mann, Nicholas H.. (2008) *A proteomic approach to the identification of the major virion structural proteins of the marine cyanomyovirus S-PM2.* Microbiology, Vol.154 (Part 6). pp. 1775-1782. ISSN 1350-0872

Invited Oral Presentations

2010 University of East Anglia

2009 Waters Corporation Users meeting and Waters Corporation LC-MS Meeting

2009 East Midlands Proteomics Workshop

2008 Third Annual Congress of the Italian Proteomic Association, Brindisi, Italy

Taught Courses

Practical Proteomics Course held at Warwick University over 3 days.
http://www2.warwick.ac.uk/fac/sci/lifesci/study/shortcourses/proteomics/

Warwick Practical Proteomics Course held in Bangalore, India over 4 days and sponsored by Waters Corporation.
(http://www2.warwick.ac.uk/fac/sci/lifesci/research/facilities/proteomics/proteomicsbangalore).

Chapter Two: Development of data independent MS^E acquisition on a Q-ToF Ultima Global instrument

2.1 Introduction

Common approaches to protein identification using mass spectrometry utilise data dependent approaches, described in Section 1.4.2.3. Throughout the chromatographic separation a number of peptides are mass selected for MS/MS based on the MS survey scan. This approach is effective for identifying a small number of proteins from gel-resolved samples, but has a number of potential limitations particularly for the comprehensive characterisation of complex mixtures, such as tryptic digests derived from proteomes.

Selection of ions for MS/MS in a data dependent acquisition is biased towards the most ionisable peptides eluting from the LC column at that point in time. Selection of abundant precursor ions, if fragmentation is favourable, should generate high quality MS/MS spectra each with a high signal: noise ratio ensuring peptide identification from the database interrogation. Selection of low abundance ions generates fragment ions that may be difficult to differentiate from the background and provide poor or misleading identifications.

There is a limitation on the number of ions that can be mass selected for MS/MS from a single survey scan. On the Q-ToF Ultima Global (Waters Corporation, Milford, MA, USA) used in this study the maximum number is eight whilst the LTQ OrbiTrap Elite (Thermo Fisher Scientific, Waltham, MA, USA) is capable of obtaining one MS survey scan (240,000 resolution @ *m/z* 400, 768 msec transient) with 20 CID scans all within a 2.7 second cycle (Michalski, Damoc et al. 2012). The absolute number of MS/MS events/second is not the limiting factor, as instrument scan rate and LC resolution must also be taken into consideration.

High resolution ultra high pressure LC systems, such as the NanoAcquity UPLC (Waters Corporation, Milford, MA, USA) used in this study use 75 μm i.d. columns incorporating a 1.8 μm particle size stationary phase and typically produce chromatographic peak widths of 6 sec at FWHH. Selecting 8 peptide ions for MS/MS with a scan rate of 1 sec may only generate good quality MS/MS spectra from the first few precursors, assuming that the MS survey scan was performed prior to the chromatographic peak apex. Reducing the scan rate to sub-second scan rates,

will ensure that more precursors are selected but only at the expense of quality of MS/MS signal: noise. There will always be a trade off between LC resolution, scan rate and quality of MS/MS data in any data dependent acquisition.

For complex mixtures where >100,000 peptide species may be generated from a tryptic digest of a proteome, many tens of peptides may be co-eluting from the LC column at a given point in time. Some of those peptides will have very similar *m/z* and may be selected for MS/MS within the precursor ion selection window, generating a chimaeric MS/MS spectrum that is composed of the fragmentation of both peptides, affecting peptide and protein identifications rates. A proteomic analysis of HeLa cervix carcinoma cells labelled using SILAC, observed 101,726 isotope clusters (charge state >1) of which 16,924 were targeted for MS/MS during the LC separation using a top-10 approach on an LTQ Orbitrap Velos (Thermo Fisher Scientific, Waltham, MA, USA). Of these only 9,797 peptides were identified (58% of those targeted with a 1% FDR) using a 4 Th isolation window (Michalski, Cox et al. 2011),

Figure 2. 1**.** The incorporation of an additional chromatographic separation would reduce the complexity of the mixture entering the MS instrument for data dependent acquisition, but at the expense of additional experimental and instrument time.

An alternative to data dependent acquisition was developed by Waters Corporation for their Q-ToF Premier and later instrumentation and is described in Section 1.4.2.4. No mass selection of peptide precursors for MS/MS in a data independent acquisition occurs and all scans are collected in MS mode (Silva, Denny et al. 2005). This approach is termed MS^E.

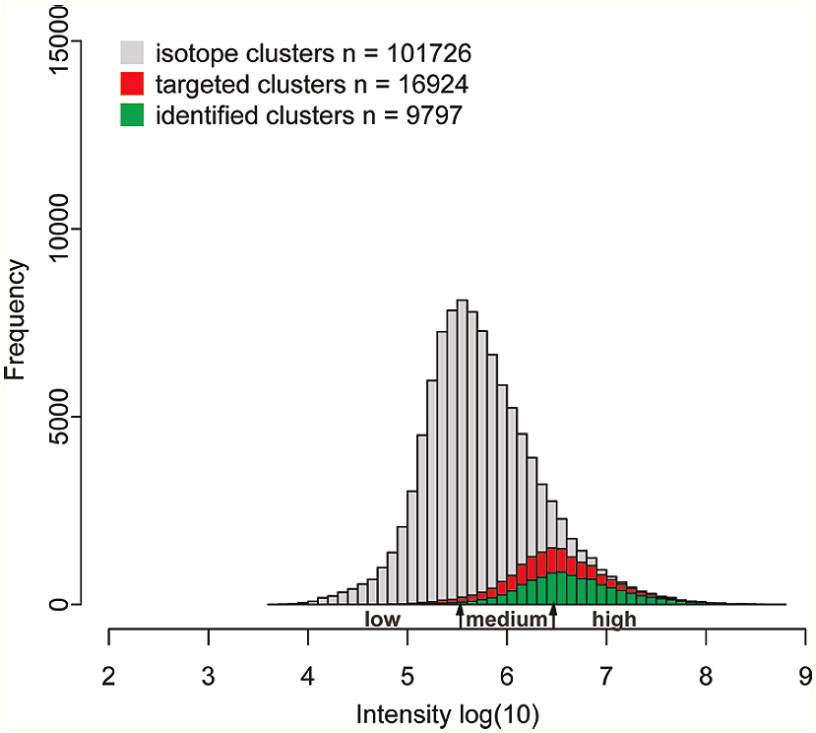


Figure 2. 1 : Histogram of intensity of detected peptides from an LC-MS/MS analysis of SILAC-labelled HeLa cervix carcinoma cells.
Bars in green indicate MS/MS selected and identified peptides, red bars show targeted peptides and grey bars all peptides (charge state >1)**,** taken from (Michalski, Cox et al. 2011).

In MS^E the instrument alternates between two collision energy states. For the first MS scan, a low collision energy is employed analogous to the survey scan used in a DDA. This is immediately followed by a scan in which the collision energy is ramped from low to elevated during the scan period,

Figure 2. 2. In the low energy scan, information is collected on the *m/*z of the eluting intact tryptic peptides and their respective retention times. During the elevated energy scan, information on intact precursors, together with all the fragment ions produced, regardless of precursor and their respective retention times is obtained. As no mass selection takes place, bias towards the identification of high abundance peptides is reduced and information on the eluting peptides will always be collected (Patel, Thalassinos et al. 2009). The correlation of fragment ions to their respective precursor and peptide/protein identification has been described in depth in Section 1.4.3.6.

2.1.1 Instrumental considerations for MS^E data independent acquisition

A number of instrumental parameters need to be optimised in order to collect MS^E data suitable for processing and database interrogation by the software ProteinLynx Global Server (PLGS - Waters Corporation, Milford, MA. USA). Each of these experimental parameters will be reviewed in detail.

2.1.2 Low collision energy instrumental considerations

In the low collision energy scan, information on intact peptide precursors needs to be collected. Selection of a collision energy that minimises fragmentation within the collision cell is essential. Fragmentation of the peptides can also result from conditions within the source region, such as source pressure, collision gas pressure and cone voltage. Transmission of low *m/z* background ions needs to be reduced in order to reduce the processing time taken by PLGS to differentiate between peptide ions and chemical/background noise.

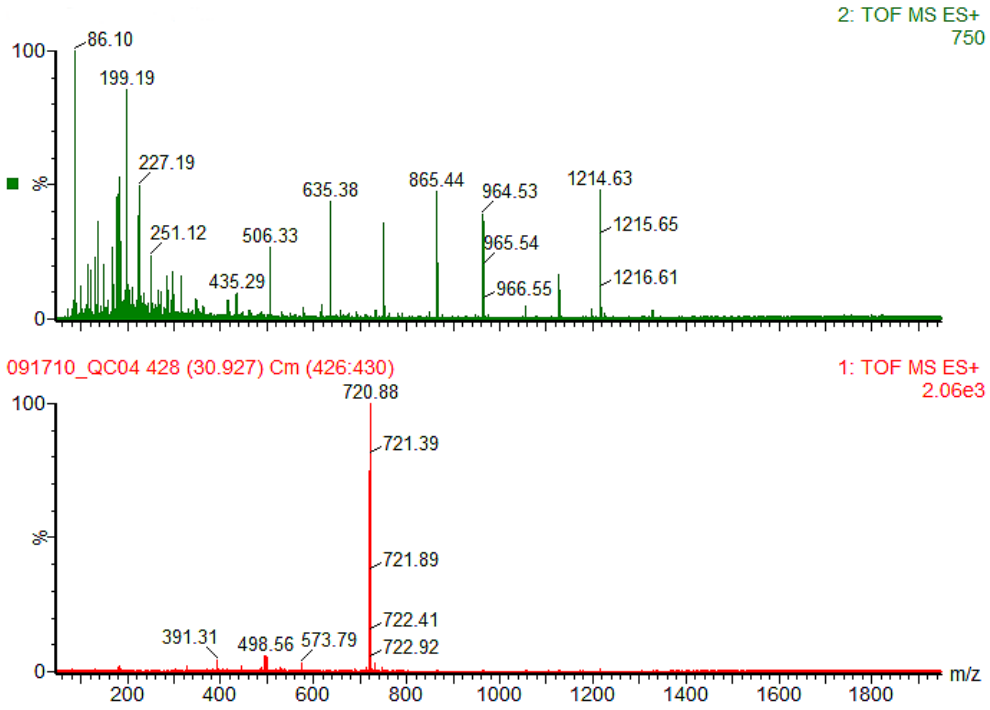


Figure 2. 2 : Low and elevated collision energy spectra obtained from a peptide analysed using MS^E acquisition.
Elevated collision energy spectrum (6 – 30 V) displayed in upper panel with low energy MS (6 V) shown below.

2.1.3 Elevated collision energy instrumental considerations

In a DDA using a Q-ToF instrument, the selection of collision energy for a given peptide precursor is based on the *m/z* and *charge* of each ion as determined by the instrument control software. A set of look-up tables for each charge state are subdivided by *m/*z and are then used to determine the appropriate collision energy for each peptide.

In an MS^E acquisition, multiple peptide precursors may be co-eluting from the LC spanning a wide range of *m/z* and *charge*. An optimal collision energy ramp should therefore fragment both low and high *m/z* ions from singly charged through to multiply charged species (5^+ and 6^+). Any collision energy ramp will not be optimised for all of the eluting peptides, but it should be possible to empirically optimise the ramp using experimentally derived data using the protein identification results from PLGS to determine the efficiency of each collision energy ramp used.

Optimised transmission of fragment ions from the collision cell would improve peptide identification rates which would in turn increase the sensitivity of the MS^E experiment. A schematic of a Q-ToF Ultima Global (Figure 2. 3) is shown illustrating the instrumental regions requiring optimisation for MS^E acquisition.

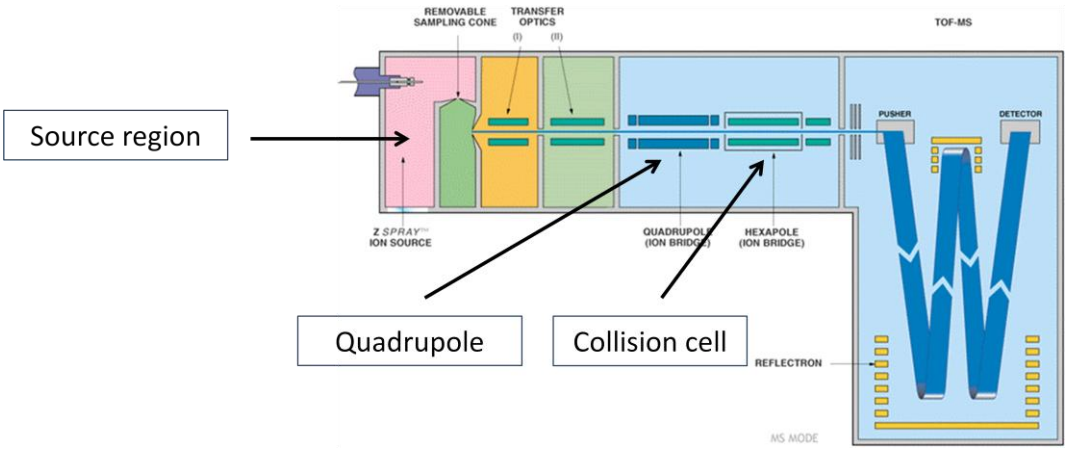


Figure 2. 3 : Schematic of a Q-ToF Ultima Global with arrows depicting regions of instrument requiring optimisation for a MS^E data independent acquisition.

2.1.4 Data processing considerations

In an MS^E acquisition, the MassLynx software (Waters Corporation, Milford, MA. USA) creates three functions within the raw data file. Function 1 contains the low collision energy information, function 2 contains elevated energy data and function 3 reference compound MS data used for lock mass correction of the raw data during processing, Figure 2. 4. Instrument control is defined by the MS Method Editor by selecting MS^E acquisition.

Incorrect instrumental conditions result in data files that the software PLGS cannot process which require extensive time to process (hours per run) which is not feasible in a high throughput proteomics environment.

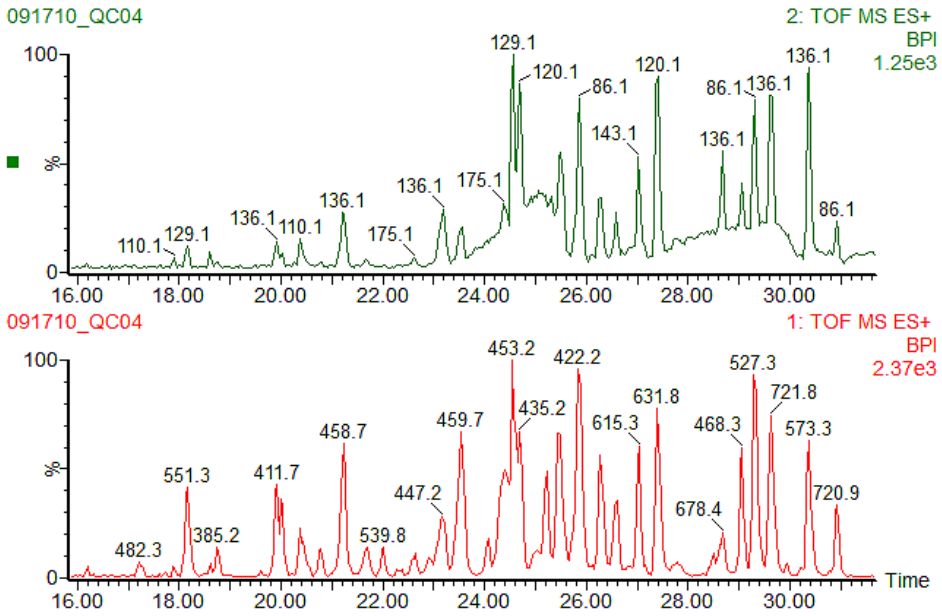


Figure 2. 4 : BPI chromatograms indicating low and elevated collision energy data collected by MS^E acquisition.
The low energy BPI chromatogram (function 1) displayed in lower panel and elevated energy (function 2) upper. Each peak has been annotated with the base peak *m/z*.

In this study, a method was developed to allow a Q-ToF Ultima Global instrument to collect MS^E data that could be used for proteomic analyses. On these older Q-ToF instruments an option for MS^E acquisition within the MS Method Editor does not exist, and therefore a modified method was created within the Editor. A modified neutral loss acquisition was utilised for this purpose and data from the optimised

experimental conditions evaluated against a data dependent acquisition and a commercial instrument designed for MS^E acquisition.

The main aims of this work were to:-

- Implement non-labelled, data independent acquisition (MS^E) experiments on an older Q-ToF instrument not commercially set up for such analyses.
- Compare the information content of data independent and data dependent experiments on a complex tryptic digest.

2.2 Materials and methods

2.2.1 Material suppliers

All tryptically digested MassPREP™ *Escherichia coli* and protein standard 1 were obtained from Waters Corporation (Milford, MA, USA). Glu^1-Fibrinopeptide B peptide (human) was purchased from Sigma Aldrich (Gillingham, UK). Mass spectrometry solvents were supplied by MallinckrodtBaker Inc. (Phillipsburg, NJ, USA). Sample vials (LCMS Certified) were purchased from Waters Corporation (Milford, MA, USA) fitted with pre-slit PTFE/silicone septa in the caps.

2.2.2 Sample preparation

Lyophilised MassPREP™ digestion standard mix 1 contains four tryptically digested proteins in approximately equimolar ratios. These are glycogen phosphorylase (rabbit - PhosB), alcohol dehydrogenase (yeast - ADH), enolase (yeast) and serum albumin (bovine - BSA). One vial was rehydrated in 1 mL of 10% v/v aqueous acetonitrile and then an aliquot was further diluted in 0.1% v/v aqueous formic acid, giving a final concentration of approximately 50 fmol μL^{-1} of each digested protein in solution. The single protein digest of MassPREP™ glycogen phosphorylase (rabbit) was prepared to a final concentration of 50 fmol μL^{-1} in 0.1% v/v aqueous formic acid.

Lyophilised MassPREP™ *E. coli* digestion standard was reconstituted to a final concentration of 500 ng μL^{-1} in 0.1% v/v aqueous formic acid.

Glu^1-Fibrinopeptide B peptide (GFP) was reconstituted in 50% v/v aqueous acetonitrile containing 0.1% formic acid to a final concentration of 500 fmol μL^{-1}.

2.2.3 Liquid chromatography configuration

All nanoscale liquid chromatographic separations were performed using a directly-coupled NanoAcquity UPLC system and a nanoelectrospray source (Waters

Corporation, Milford, MA, USA). The system was composed of a binary solvent, auxiliary solvent and sample manager fitted with a heating and trapping module.

LC separations were performed using a Symmetry C18 trapping column (180 µm x 20 mm 5 µm) and a HSS T3 analytical column (75 µm x 150 mm 1.8 µm). The composition of solvent A was 0.1% v/v aqueous formic acid and solvent B 0.1% v/v formic acid in acetonitrile. An aliquot of each sample (50 fmol of glycogen phosphorylase, 100 fmol of digestion standard 1 or 500 ng of *E. coli* digest) was applied to the trapping column and flushed with 0.1% solvent B for 1 minute at a flow rate of 20 µL min^{-1}. Sample elution was performed using a flow rate of 300 nL min^{-1} by increasing the organic solvent concentration from 5 to 40% B over 20 min. For the *E. coli* DDA analyses, the LC gradient was extended to 60 and 120 min. All analyses were conducted as either single injections or in technical triplicate.

The reference compound, GFP, used for lockmass correction was infused at a constant rate of 500 nL min^{-1} at 500 fmol μL^{-1}.

2.2.4 Mass spectrometry configurations

The precursor ion accurate masses and associated fragment ion spectra of the tryptic peptides were mass measured with a Q-ToF Ultima Global mass spectrometer (Waters Corporation, Milford, MA, USA) operated in electrospray data independent MS^E and data dependent mode controlled by MassLynx v4.0. In-source decomposition was reduced by decreasing the cone voltage (RF Lens 1) to 25 V and using a source temperature of 70 ^oC. The mass profile on the quadrupole was set to 400, 500 and 600 *m/z* with ramp and dwell times of 25%, 25%, 25% and 25% respectively.

The time-of-flight analyser of the mass spectrometer was externally calibrated using a MS/MS spectrum obtained from the doubly charged precursor of the GFP peptide from *m/z* 50 to 1300. The calibration was manually validated with an average ppm error across the mass range of <10 ppm being obtained.

All subsequent data were post-acquisition lockmass-corrected using the monoisotopic ion of the doubly charged precursor of GFP (*m/z* 785.8426). The GFP

was delivered to the mass spectrometer via a NanoLockSpray interface and sampled every 60 seconds.

The mass spectrometer was fitted with a universal nanoflow sprayer (Waters Corporation, Milford, MA, USA) with an applied capillary voltage of 3.5 kV. The spectral acquisition scan rate was 0.9 sec with a 0.1 s interscan delay for both MS^E and DDA.

In data dependent acquisition mode over a *m/z* range of 50-2000, CID experiments were performed on the four most intense, multiply charged peptides as they eluted from the column at any given time. Once data had been collected the next four most intense peptides were selected and the process repeated.

MS^E accurate mass data were collected over the *m/z* range 50 – 1950 using a modified neutral loss acquisition created through the MS Method Editor. Function 1 was equivalent to a MS survey scan using a range of low collision energies from 6 eV. During function 2, a large number of collision energy ramps were utilised starting as low as 6 V through to a maximum of 85 eV. No neutral loss mass was specified in the file but in the event that an MS/MS acquisition was triggered, a minimum intensity threshold of 1×10^7 was used to ensure an immediate return to low and elevated collision energy switching. The use of a time delay during the ramp in function 2 was evaluated, commencing 250 msec to 500 msec into the scan.

MS^E data acquisition was performed using a Synapt HDMS instrument (Waters Corporation, Milford, MA, USA), configured for MS^E through the MS Method Editor controlled by MassLynx v4.1. The instrument was calibrated as described for the Q-ToF Ultima Global. In low energy MS mode, data were collected at a constant trap collision energy of 6 eV. In elevated energy MS mode, the trap collision energy was ramped from 15 V to 30 V whilst the transfer collision energy was held at 3 V and 10 V for low and elevated conditions respectively. All subsequent data were post-acquisition lockmass-corrected using the monoisotopic ion of the doubly charged precursor of GFP (*m/z* 785.8426).

2.2.5 Processing of data dependent and MS^E acquired data

The uninterpreted MS/MS data from the data dependent acquired analyses were processed using ProteinLynx Global Server (PLGS) v2.5.1 (Waters Corporation, Milford, MA, USA). Experimental data were smoothed, background subtracted, centred and deisotoped. All data were lockspray calibrated against GFP using data collected from the reference line during acquisition.

The MS^E data were processed using PLGS v2.5.1. The ion detection, clustering and protein identification steps have been explained in detail in Section 1.4.3.6. In brief, lockmass-corrected spectra are centroided, deisotoped and charge-state-reduced to produce a single accurately mass measured monoisotopic mass for each peptide and its associated fragment ions. Correlation of precursor and potential fragment ions is achieved using time alignment (Geromanos, Vissers et al. 2009). Data processing parameters specified 200, 30 and 750 for the low, elevated and intensity threshold values respectively in PLGS.

2.2.6 Database interrogation

All processed data were used to interrogate a database containing the proteins of interest for the MassPREP™ protein standards using PLGS v2.5.1. Protein sequences were downloaded and compiled from the UniProtKB database (http://www.uniprot.org/).

A database was downloaded from UniProtKB (http://www.uniprot.org/) of reviewed and unreviewed proteins for *E.coli* strain BL21 release 2011-05, for use with the MassPREP™ *E.coli* standard.

For DDA data a fixed modification of carbamidomethyl cysteine was specified with oxidation of methionine as a variable modification. One missed trypsin cleavage site was permissible. Search parameters specified a 50 ppm mass error tolerance against the database and a minimum of one matched peptide.

For the MS^E data, database search parameters were as described above also included variable modifications of acetyl N-terminus and deamidation of asparagine and glutamine.

Precursor and fragment ion tolerances were determined automatically by PLGS. Protein identification criteria included the detection of at least three fragment ions per peptide, seven fragment ions per protein and at least one peptide per protein

2.3 Results and discussion

2.3.1 Protein identifications from data dependent acquisition

Single injections of 50 fmol of a tryptic digest of PhosB and 100 fmol of MassPREP™ digestion standard 1 were analysed using DDA on a Q-ToF Ultima Global instrument. A total of 23 peptides were identified from PhosB with sequence coverage of 25% being obtained.

For the digest standard containing four proteins including PhosB analysed at 2-fold concentration, peptides were identified from all four proteins with 11 from BSA, 8 from PhosB, 5 from ADH and 3 from enolase. The observed sequence coverage for each of the proteins was 19%, 10%, 16% and 6% respectively.

For *E. coli* cytosolic extract analysed using 30, 60 and 120 min LC separations in triplicate using DDA, the average number of proteins identified was 5.3, 19.3 and 34.3 as the LC gradient was extended.

These results exemplify the bias observed in a DDA experiment in a sample containing only a small number of proteins (4). With more complex samples, a greater bias would be expected towards dominant proteins.

2.3.2 Benchmarking MS^E data acquisition

In order to establish a benchmark for MS^E data quality, a 50 fmol injection of PhosB was analysed using LC-MS^E on a Synapt HDMS instrument. Over three replicate injections the sequence coverage ranged from 47.5% to 49.8%. A peptide coverage map is shown in Figure 2. 5**.**

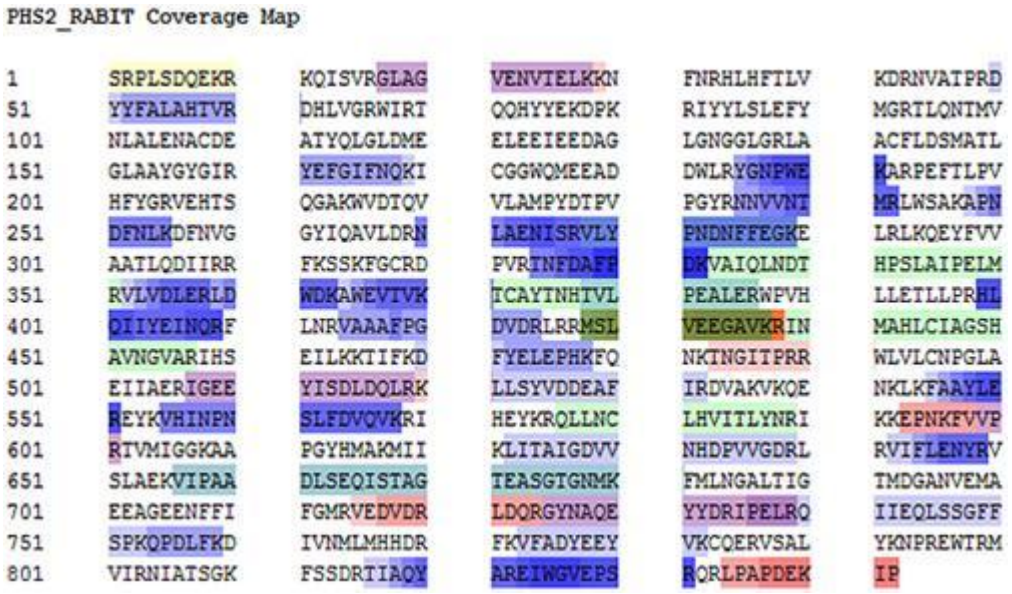


Figure 2. 5 : Peptide coverage map obtained from an MS^E acquisition of 50 fmol glycogen phosphorylase tryptic digest on a Synapt HDMS instrument.
The average sequence coverage was 48.6% with the coloured sections indicating the type of peptide identified, e.g. modified or missed cleavage.

2.3.3 Optimising collision energy ramp on Q-ToF Ultima Global

A wide range of start and end voltages were assessed for their suitability for MS^E data acquisition on the Q-ToF Ultima Global including the use of a ramp delay. The use of a neutral loss acquisition, with high thresholds for MS/MS transition was found to be successful in obtaining MS^E-like data. The creation of an additional raw data function, where MS/MS data would have been collected (function 3), did not affect the processing of the data as MS^E in PLGS. The collection of data from the reference line for lock mass correction as function 4 (instead of 3 as on a Synapt instrument) also did not affect PLGS processing of the collected data.

Optimisation of collision energy ramps was performed on the Q-ToF Ultima Global using tryptic digests from a single protein, a four protein mixture and an *E. coli* cytosolic extract in triplicate. The use of a delayed ramp did not improve the quality of the MS^E data obtained (data not shown). For each of the sample types, the results obtained from collision energy ramps 6 – 20 V to 6 – 50 V is presented.

For the four protein mixture, four proteins were identified in all ramp conditions except 6 – 50 V, Figure 2. 6**.**

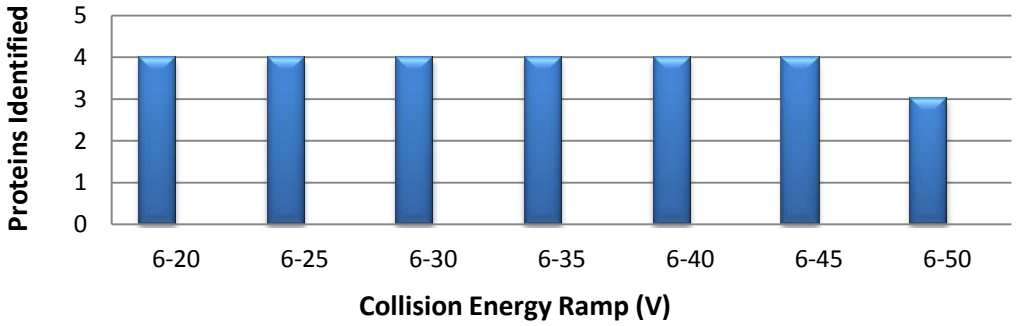


Figure 2. 6 : Number of proteins identified by a range of MS^E collision energy ramps on a Q-ToF Ultima Global

For the three proteins, ADH, enolase and BSA the optimal ramp conditions were not identical. For BSA, an average sequence coverage of 66.5% was observed with the 6 – 20 V ramp whereas for enolase the 6 – 25 and 6 – 30 V ramps gave an average 56.5% and 55% coverage respectively. The optimal ramp for ADH was observed using 6 – 25 V with an average of 54.2% coverage, Figure 2. 7**.**

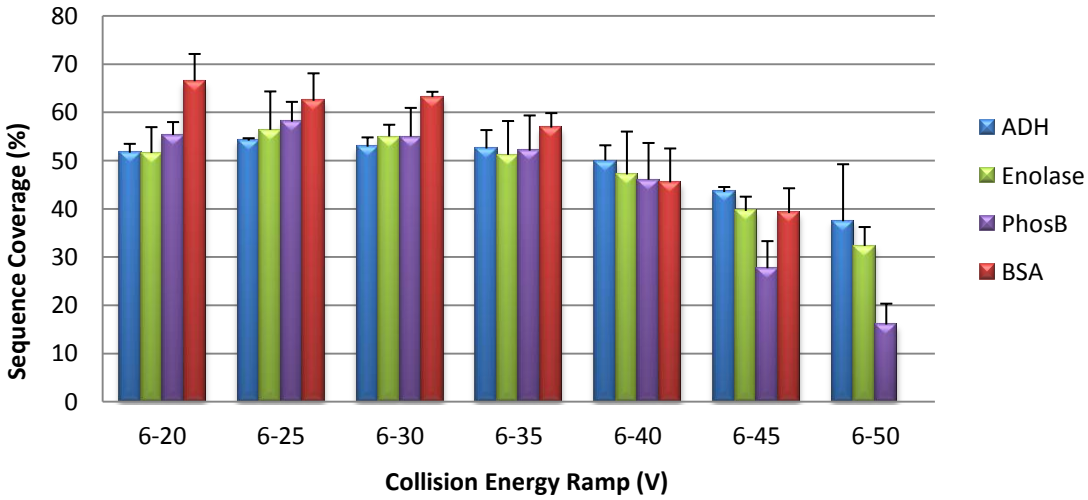


Figure 2. 7 : Average sequence coverage observed from proteins utilising a range of collision energy ramps.
Key: alcohol dehydrogenase (ADH), enolase, glycogen phosphorylase (PhosB) and bovine serum albumin (BSA). Results based on three technical replicates.

A comparison of results obtained from the four protein tryptic standard analysed using the 6 – 30 V MS^E collision gradient and the DDA analysis revealed a significant increase in the number of peptides identified by MS^E. The number of observed peptides from PhosB, BSA, ADH and enolase by MS^E was 63, 39, 35 and 25 respectively, which equates to a 4-fold increase for BSA and almost 10-fold for enolase, Figure 2. 8**.**

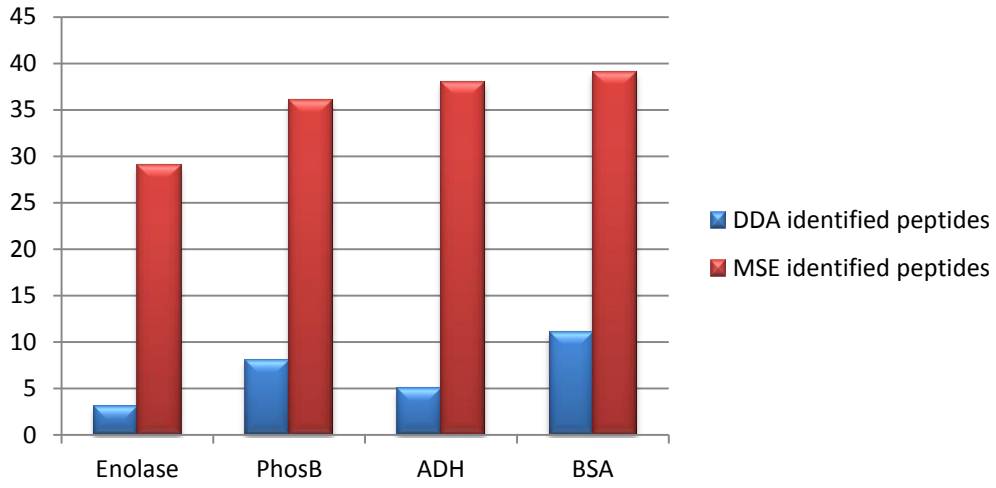


Figure 2. 8 : Number of peptides identified from a four protein digest standard analysed using DDA and MS^E.
Key: Enolase, glycogen phosphorylase (PhosB), alcohol dehydrogenase (ADH) and bovine serum albumin (BSA). Data collected using a 6 – 30 V collision gradient.

The four proteins used in the standard digest span a range of molecular weight from 35 kDa to 100 kDa, so any one given ramp may be not be optimal for every protein in a complex sample. In order to assess ramp conditions on a more complex sample, a 500 ng *E. coli* digest was utilised. The optimal ramp condition was determined as 6 – 30 V identifying just under 200 proteins each time using a 60 min LC gradient. The number of proteins identified from each ramp, in technical triplicate, were normalised to the number obtained from the 6 – 30 V ramp and plotted against the results from the DDA analyses of the same sample using a 30, 60 and 120 min LC gradient as comparison, the data is shown in Figure 2. 9**.**

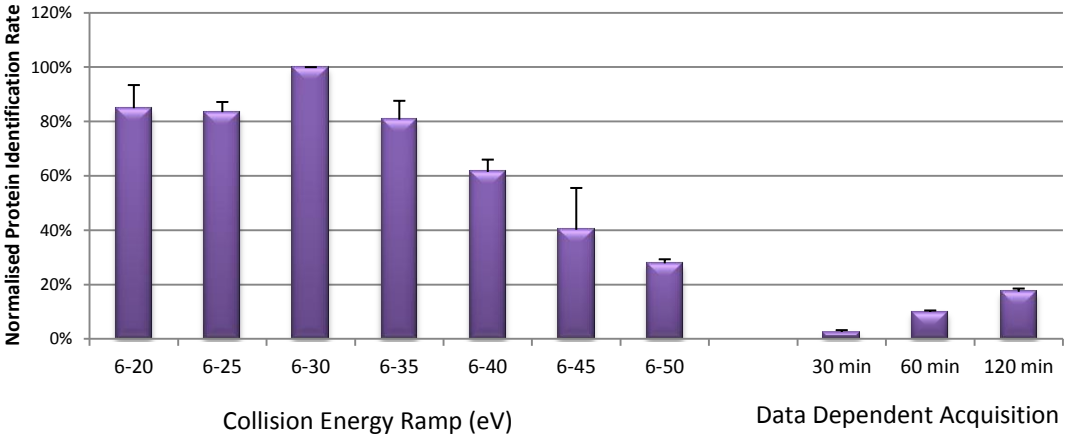


Figure 2. 9 : Normalised number of *E. coli* proteins identified from MS^E collision energy ramps and data dependent acquisition on a Q-ToF Ultima Global.
MS^E data were collected using a 60 minute LC separation.

Based on the optimised collision energy ramp conditions for MS^E from the *E. coli* digest, 50 fmol of Phos B was analysed on the Q-ToF Ultima Global using a 6 – 30 V ramp for comparison against the benchmark Synapt HDMS data. Three technical replicates provided 54.8%, 54.8% and 48.1% sequence coverage compared to an average 48.6% from the Synapt HDMS, Figure 2. 10. A peptide coverage map is shown in Figure 2. 11.

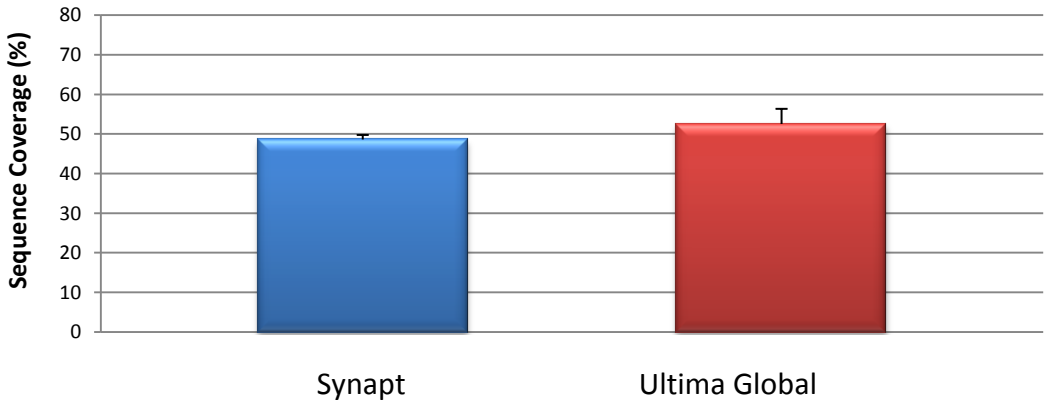


Figure 2. 10 : Comparison of sequence coverage obtained from a 50 fmol tryptic digest of glycogen phosphorylase analysed using MS^E on a Synapt HDMS and Q-ToF Ultima Global.
A 6 – 30 V collision energy ramp was utilised on the Q-ToF Ultima Global.

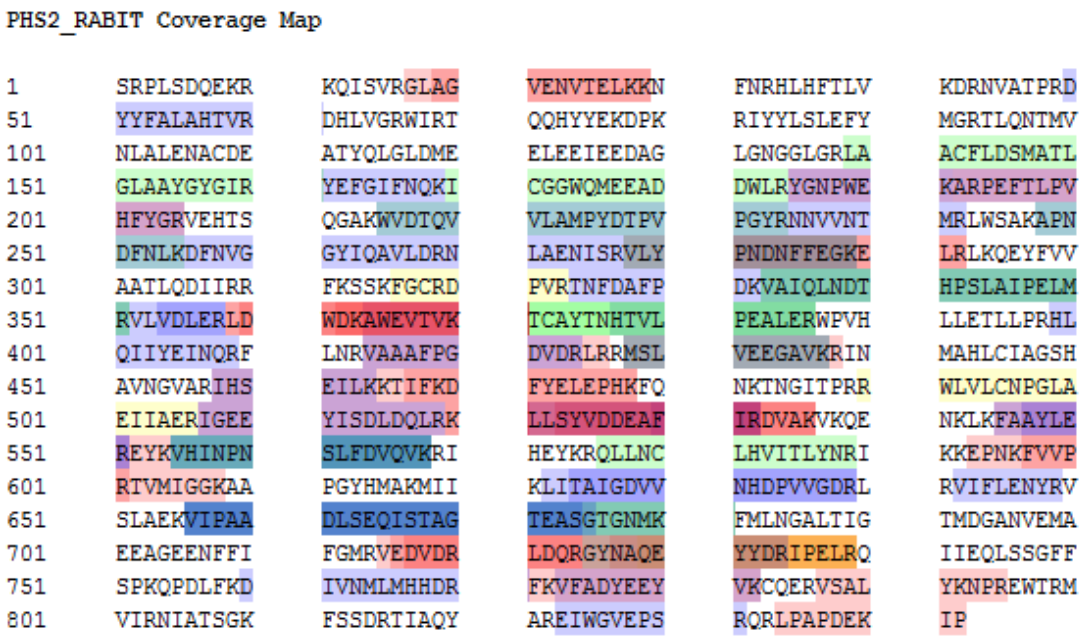


Figure 2. 11 : Peptide coverage map obtained from an MS^E acquisition from 50 fmol glycogen phosphorylase tryptic digest on a Q-ToF Ultima Global.
The average sequence coverage was 52.5% using an optimised collision energy gradient.

2.3.4 Comparison of *E. coli* MS^E acquired data with two-dimensional gel electrophoresis analysis

Protein identifications from the *E. coli* digest analysed using LC-MS^E using an optimised collision energy ramp on the Q-ToF Ultima Global provided information regarding both molecular weight and isoelectric point (pI), as predicted by sequence. This allowed the proteins identified to be plotted on a virtual 2D gel which could then be compared to experimental 2D gels. Information on the strain and growth conditions of the *E. coli* standard was not available, so two typical 2D gels from strain K12 W110 are shown for comparison.

In Figure 2. 12 (left) Blankenhorn *et al.* visualised 300 proteins from strain K12 W110 stained by Coomassie blue using gels optimised for the molecular weight range up to 80 kDa spanning the pI range 4.5 – 6.5 (Blankenhorn, Phillips et al. 1999). Figure 2. 12 (right) shows a 2D reference map for *Escherichia coli* taken

from the SWISS-2DPAGE Two-dimensional Polyacrylamide Gel Electrophoresis Database (http://world-2dpage.expasy.org/swiss-2dpage/viewer).

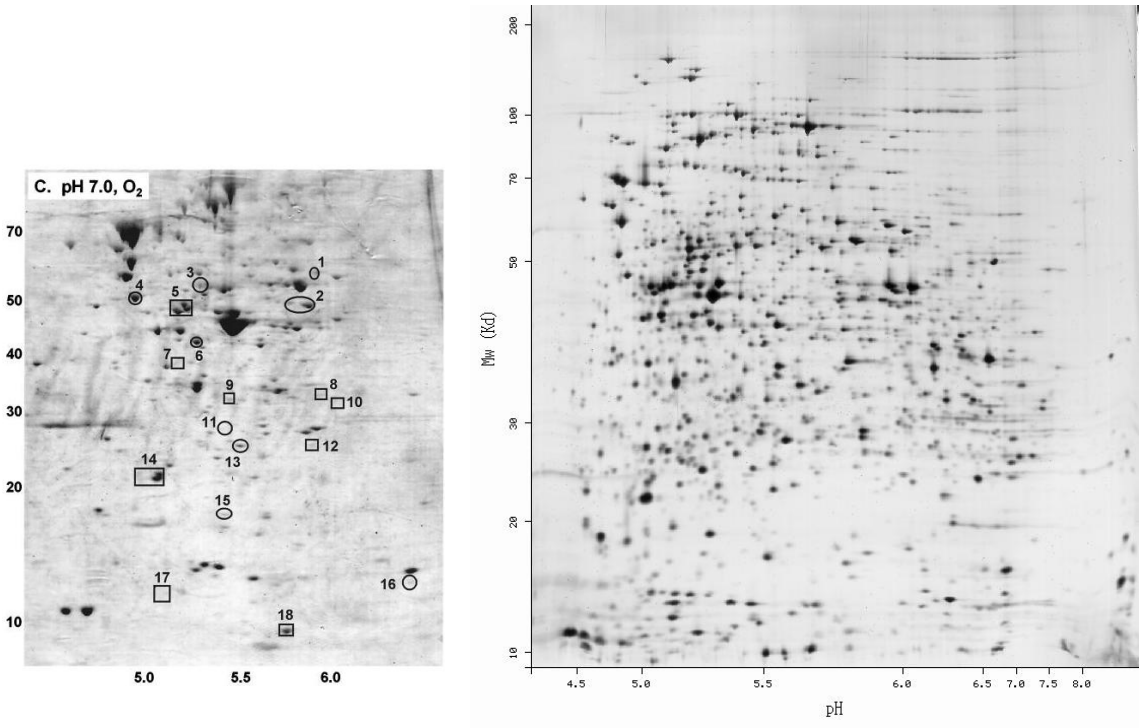


Figure 2. 12 : 2D gel images from *Escherichia coli*.
Strain K12 W110 aerobically grown at neutral pH (left) from (Blankenhorn, Phillips et al. 1999) and reference map taken from SWISS-2DPAGE Two-dimensional Polyacrylamide Gel Electrophoresis Database (right).

As can be seen in the reference gel shown in Figure 2.12 (right), the proteome of *E. coli* spans a wide molecular weight and pI range although most experimentalists target small regions of the proteome typified by Figure 2.12 (left) in order to optimise the resolution of the protein spots. On the reference gel there appears to be bunching of proteins in the pI>8.0 region with poor resolution of spots and the resolution of low molecular weight species is poor below 10 kDa. The experimental time to obtain good quality 2D gel images can be significant.

In contrast, MS^E data obtained from a single analysis of the *E. coli* standard required no chromatographic optimisation, was performed in 60 min and required as little as 500 ng to identify approximately 200 proteins. Unlike 2D gels where spots need to be removed for later identification, in an LC-MS^E experiment all proteins above the detection limit may be identified without additional experimental time. The virtual

2D gel generated from the MS^E data is shown in Figure 2. 13. The pI region 4 – 8 is well represented in the MS^E analysis compared to the reference and experimental 2D gel. A significant number of proteins in this region also fall within the molecular weight range 20 – 80 kDa. For 2D gels these would require further resolution by narrow range isoelectric focusing strips to ensure that only one protein was present in each spot ensuring accurate quantitation. In MS^E analyses, proteins can have near identical molecular weight and/or pI and still be identified.

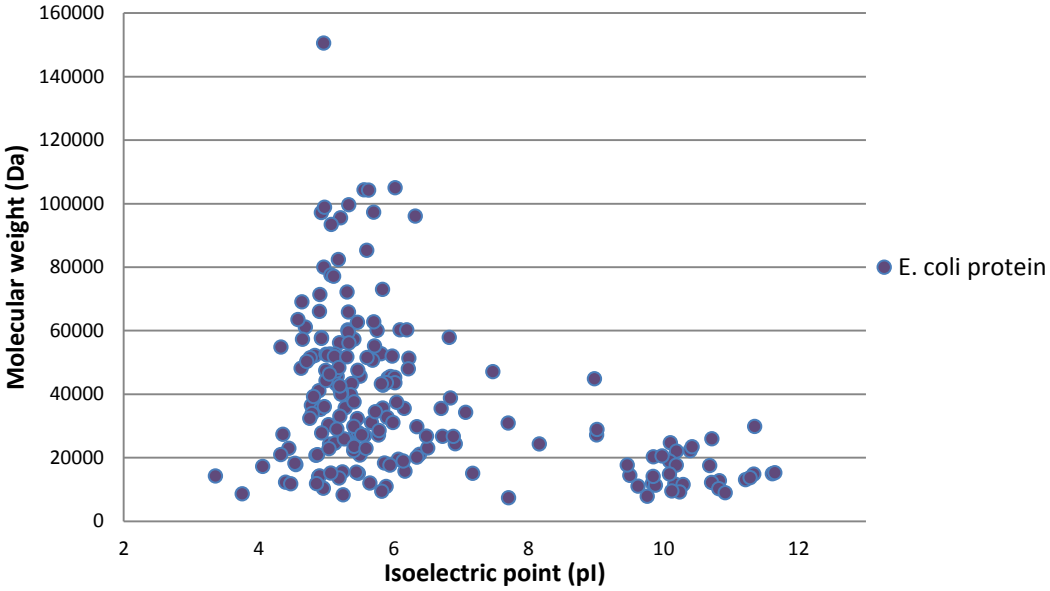


Figure 2. 13 : Virtual 2D gel plotted from E. coli proteins identified by LC-MS^E on Q-ToF Ultima Global.

Twenty three proteins were observed in the MS^E analysis with a pI>8.0 whilst 8 proteins had a molecular weight <10 kDa. The average number of peptides per protein identified from the single MS^E data set obtained on the Q-ToF Ultima Global was 13.4 and the average sequence coverage 37.2%. Of all the proteins identified, 36.9% were identified with >40% sequence coverage. No proteins were identified by a single peptide and only 5% of the proteins were identified by <4 peptides.

2.4 Conclusions

Data independent MS^E acquisition has been shown to dramatically increase the number of proteins identified from a complex mixture without the need for

additional chromatographic steps, when compared to a conventional data dependent approach. This work has demonstrated that using an optimised, modified neutral loss acquisition on a Q-ToF Ultima Global MS^E type data can be obtained, processed and used in database interrogations as if it were MS^E data.

This work has also shown that the data quality from an optimised MS^E acquisition using a Q-ToF Ultima Global, established in terms of protein identification and sequence coverage, was very similar to that obtained using a commercial MS^E-capable instrument, the Synapt HDMS. Recent data collected which compared MS^E using a Synapt G2 HDMS and Q-ToF Ultima Global (with identical sample loading), revealed a very similar number of proteins identified from a liver homogenate. A quantitative proteomic analysis of exoproteome variation in the prokaryotic pathogen *Corynebacterium pseudotuberculosis* obtained by MS^E on the Q-ToF Ultima Global has recently been published (Pacheco, Slade et al. 2011).

MS^E utilising a single dimension chromatographic separation has been demonstrated to identify highly and moderately abundant proteins in complex samples. This makes it a suitable analytical method for the analysis of human plasma which is described in the following chapters.

Chapter Three: Proteomic analysis of IgY-12 fractionated maternal plasma

3.1 Introduction

Rifai and co-workers described a biomarker as "a measurable indicator of a specific biological state, particularly one relevant to the risk of contraction, the presence or the stage of disease" (Rifai, Gillette et al. 2006). The National Institutes of Health Biomarker Working Group defined a biomarker as any *characteristic* that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Atkinson, Colburn et al. 2001). Many medical tests for disease involve direct analysis of biological material such as plasma, urine, cerebrospinal fluid or tissue biopsies and include measurements such as enzyme or hormone concentration, but other techniques such as imaging or gene phenotype distribution can provide biomarker information.

Routine screening of pregnant women at 11-13 weeks gestation in the UK involves a blood sample being taken for the identification of risk factors for trisomy 21, thus providing a suitable material to identify protein biomarkers. This is an ideal sample as the population of women tested would span a wide range of obstetric conditions, ethnicity, age, body mass index (BMI) and parity and any biomarker information could be used to improve existing diagnostic tests (trisomy 21) or identify markers of disease allowing earlier medical intervention (pre-eclampsia).

Biomarker discovery in plasma is challenging due to the complexity of the matrix under investigation but also holds the most promise as the circulatory system is thought to be representative of the physiological and pathogical processes underway in any individual (Anderson and Anderson 2002). Over the last 15 years, the average introduction of new protein analytes has remained static at an average of 1.5 per year and of the 109 unique FDA-approved serum or plasma protein tests, 80% were introduced before 1993 (Anderson 2010) despite the enormous numbers of publications citing biomarker discovery.

Most biomarker studies involving plasma utilise a depletion step to remove the highly abundant proteins, refer to Section 1.8.2. The depletion step needs to be effective and reproducible across many tens of samples, to avoid incorrect

identification of candidate biomarkers and to reduce the cost of analysis. Due to the cost and experimental time required, many studies have utilised pooled plasma samples without validating intra-patient variation within clinical groups. The analytical approach taken should identify and quantify accurately and precisely, the protein levels in plasma including those depleted, to ensure consistency across the dataset.

The ProteomeLab™ IgY-12 partitioning system used in this study was based on avian antibody (IgY) – antigen affinity interactions. IgY antibodies were generated by immunising laying hens with purified human antigens (highly abundant plasma proteins) which are listed in Table 3. 1 (Huang, Harvie et al. 2005). The advantage of using IgY rather than IgG antibodies was a reduction in non-specific cross-reactive binding of other plasma proteins such as Fc receptor, complement and rheumatoid factors, Proteins A, G and IgM bind to the Fc region of the antibodies whilst maintaining high avidity. After immunisation the antibodies were secreted by hens into egg yolks generating approximately 100 mg of total IgY/egg which equates to 5 g of antibody produced from each hen (40-60 eggs). The antibodies were covalently coupled to a microbead support which can be used in a LC column or, as utilised in this study, spin column format. The diluted plasma can be applied to the column and the partitioned plasma proteome, enriched in the moderately and low abundance proteins eluted from the column. The bound highly abundant proteins can subsequently be eluted prior to the next plasma application. The ProteomeLab™ IgY-12 partitioning system is guaranteed for 100 partitioning procedures (50 samples per spin column) and was chosen for this study for its robustness.

In this work, individual maternal plasma samples were depleted of 12 highly abundant proteins and analysed utilising one dimensional nanoscale LC-MS^E analysis on a Q-ToF Ultima Global instrument. All samples were taken from age, BMI and ethnicity matched women with no obstetric complications reported.

Target Protein	Typical Depletion Efficiency (%)*	Target Protein	Typical Depletion Efficiency (%)
Albumin	99.6	IgM	99.0
IgG total	99.1	α 1-Antitrypsin	99.7
Transferrin	99.1	Haptoglobulin	99.3
Fibrinogen	96.9	α1-acid Glycoprotein	99.1
IgA	99.0	Apolipoprotein A-I	99.2
α 2-Macroglobulin	94.4	Apolipoprotein A-II	99.2

Table 3. 1 : Proteins partitioned by the ProteomeLab™ IgY-12 human plasma kit. ProteomeLab™ IgY Protein Partitioning Solutions from Phenomenx Inc. (Torrence, CA, USA), 2007.

3.2 Materials and methods

3.2.1 Material suppliers

A ProteomeLab IgY-12 spin column kit was purchased from Beckman Coulter (Fullerton, USA). This included 2 spin columns, dilution, stripping and neutralisation buffers. Double centrifuged maternal plasma samples were supplied by Prof. Kypros Nicolaides of King's College Hospital, London, UK with full ethical approval.

Eleven plasma samples were selected for this part of the study, from Caucasian women of a narrow age, body mass index (BMI) and gestational range. The average maternal age was 31.9 years, BMI 23.5 and gestational age 87.5 days. Full sample details are included in Appendix A.

Rapigest surfactant was obtained from Waters Corporation (Milford, MA, USA). Dithiothreitol was supplied by Melford Labs. (Ipswich, UK). Glu^1-Fibrinopeptide B peptide (human), sodium azide, iodoacetamide and ammonium bicarbonate were purchased from Sigma Aldrich (Gillingham, UK), sequencing grade trypsin from

Promega (Madison, WI, USA) and mass spectrometry solvents were supplied by MallinckrodtBaker Inc. (Phillipsburg, NJ, USA). Spin-X cellulose acetate centrifuge tube filters were supplied by Costar (Corning Inc., Tewksbury MA, USA) and 5 kDa nominal molecular weight cut-off (NMWCO) spin columns by Millipore (Billerica, MA, USA). Sample vials (LCMS Certified) were purchased from Waters Corporation (Milford, MA, USA) fitted with pre-slit PTFE/silicone septa in the caps.

3.2.2 Sample preparation

Each individual maternal plasma sample was thawed from -70 ^{o}C at room temperature, inverted a number of times to ensure homogeneity prior to a 20 μL aliquot being removed. Each aliquot was diluted to 500 μL with 1 x dilution buffer and centrifuged using a 0.22 μm Spin-X cellulose acetate centrifuge tube filter prior to depletion

3.2.2.1 Fractionation of human plasma using an IgY-12 spin column

Filtered, diluted maternal plasma dilution buffer was transferred to a spin column and sealed. The unit was incubated at room temperature for 15 min on a rotary shaker with multiple inversions.

The end cap was removed and the unit centrifuged at 2,000 g for 30 sec with the flow-through (depleted plasma) collected in a new 2 mL collection tube. A 500 μL aliquot of 1 x dilution buffer was added to the spin column and the process repeated, with the flow through collected and combined with the initial eluate. Depleted plasma was then ready for processing for tryptic digestion, yielding approximately 160 μg.

Bound, highly abundant plasma proteins on IgY-12 spin column beads were eluted using the following procedure. Three 500 μL aliquots of 1 x dilution buffer were added to the spin column, inverted multiple times and centrifuged at 2,000 g for 30 sec with the flow-through collected. A stripping buffer (500 μL, 1 x concentration) was added to the spin column in four steps, incubated with shaking and inversion for 3 min at room temperature each time, prior to centrifugation at 2,000 g for 30 sec

with the flow-through (bound plasma proteins) collected into a new 2 mL collection tube. The combined eluate from the stripping procedure (2 mL) was immediately neutralised with 200 µL of 10 x neutralisation buffer, as supplied.

The spin column stripping procedure was completed within 15 min and the column neutralised by the addition of 600 µL of 1 x neutralisation buffer (diluted from the 10 x stock supplied), incubated at room temperature for 5 min with shaking and inversion and centrifuged at 2,000 g for 30 sec with the flow-through collected into a new 2 mL collection tube. The spin column beads were completely resuspended in 500 µL of 1 x dilution buffer containing 0.02% sodium azide prior to storage at 4 ^{o}C until next use. A flow diagram of the IgY-12 depletion procedure is shown in

Figure 3. 1**.**

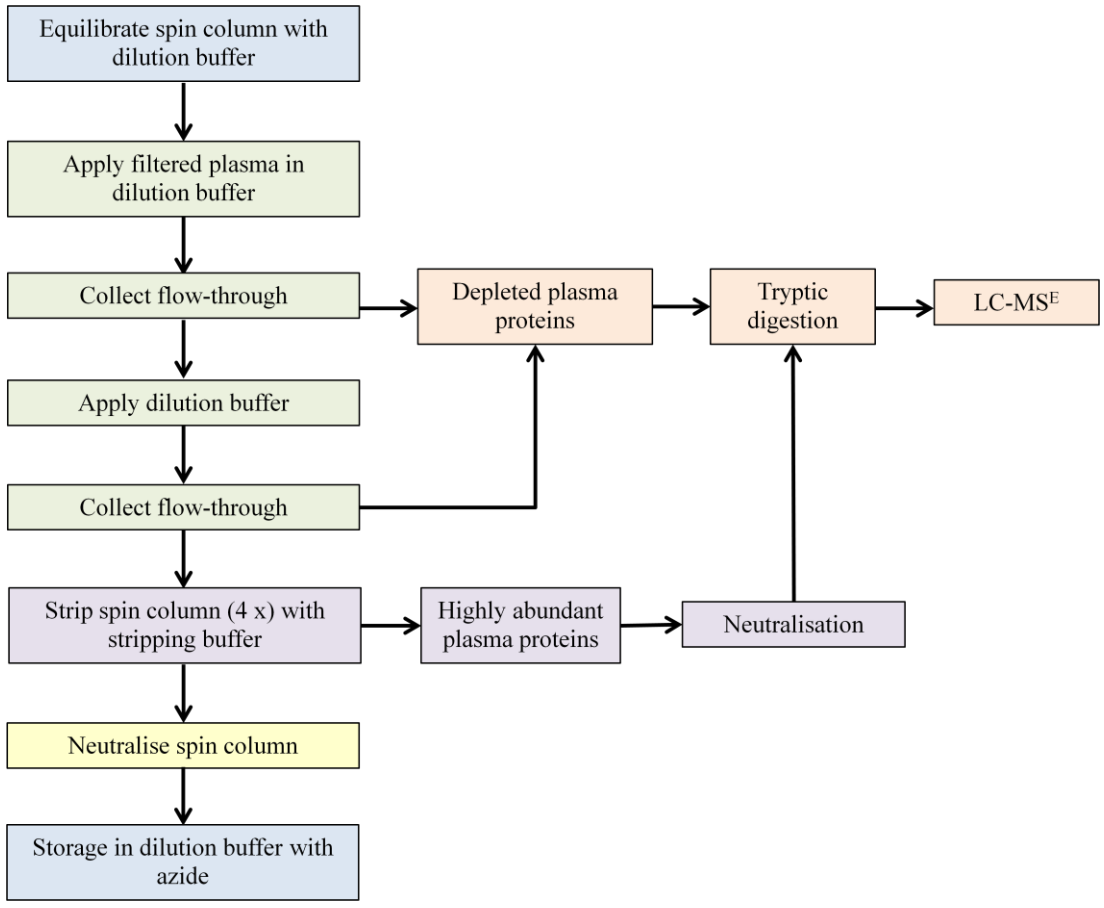


Figure 3. 1 : Flow diagram of the depletion procedure for plasma using an IgY-12 spin column.

3.2.2.2 Tryptic digestion of IgY-12 fractionated plasma

The IgY-12 depleted plasma fraction (total volume approximately 1 mL) was transferred to a vial containing lyophilised Rapigest surfactant, with a final concentration of 0.1% w/v and gently agitated until fully dissolved. The contents were transferred to an 5 kDa NMWCO spin column and centrifuged at 14,000 g, 4 ^{o}C until the volume was approximately 50 µL.

The contents were transferred to a 0.5 mL microfuge tube (Fisher Scientific Ltd, Loughborough, UK) and incubated in a water bath at 80 ^{o}C for 15 min. A 5 µL aliquot of 100 mM dithiothreitol in 100 mM ammonium bicarbonate (NH_4HCO_3) was added to the plasma and thoroughly agitated prior to incubation at 60 ^{o}C for 15 min, followed by the addition of 5 µL of 200 mM iodoacetamide in 100 mM NH_4HCO_3 and incubated at room temperature, in the dark for 30 min.

A vial containing 20 µg of trypsin was fully resolubilised in 20 µL of 100 mM NH_4HCO_3 with 2 µL (2 µg) transferred to the plasma sample and thoroughly agitated. The sample was incubated overnight at 37 ^{o}C.

The following day, 2 µL of concentrated formic acid were added to the sample and incubated at 37 ^{o}C for 15 min, prior to filtration through a 0.22 µm Spin-X cellulose acetate centrifuge tube filter. An aliquot of the tryptically digested sample was removed for analysis and the remainder stored at -20 ^{o}C until required. Typically 45 - 50 µL of tryptic digest was obtained for each depleted plasma sample giving a final concentration of approximately 3 µg μL^{-1} depleted plasma.

An aliquot containing 50 µL of raw plasma obtained from a complication-free pregnancy (N86757) was filtered through a 0.22 µm Spin-X cellulose acetate centrifuge tube filter, resuspended in 950 µL of 100 mM NH_4HCO_3 and transferred to a vial containing lyophilised Rapigest. The contents were then processed identically to the depleted plasma fraction from the IgY-12 spin column. Based on an average plasma protein concentration of 60 – 85 g L^{-1} the total protein content of the undepleted tryptic digest would be in the region of approximately 4 mg.

For the IgY-12 depleted plasma sample an equal volume of tryptic digest was combined with a solution containing 100 fmol μL^{-1} MassPREP™ glycogen phosphorylase (PhosB) tryptic digestion standard in 0.1% v/v aqueous formic acid, giving a final concentration of 50 fmol μL^{-1} PhosB. Endogenous glycogen phosphorylase was not observed in any of the analyses in this work, and so was suitable for use as an internal standard for the estimation of protein concentrations using the Hi3 approach (Silva, Gorenstein et al. 2006). An overview of the tryptic digestion process is shown in Figure 3. 2**.**

The undepleted plasma protein tryptic digest was diluted to an approximate concentration of 0.4 μg μL^{-1} and combined with a solution containing 1000 fmol μL^{-1} MassPREP™ glycogen phosphorylase (PhosB) tryptic digestion standard in 0.1% v/v aqueous formic acid, giving a final concentration of 500 fmol μL^{-1} PhosB.

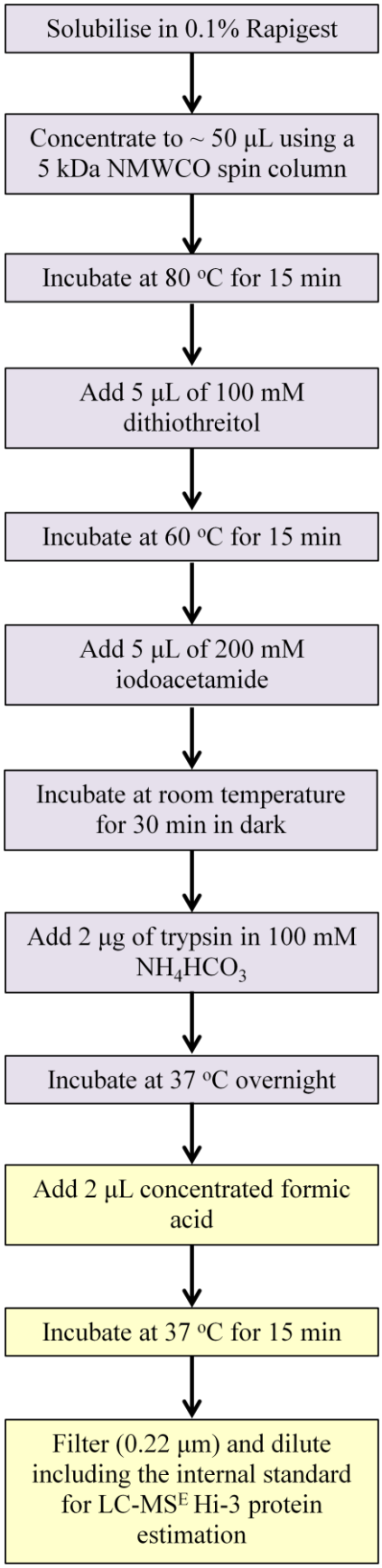


Figure 3. 2 : Flow diagram of the process for tryptic digestion of plasma samples.

3.2.3 LC-MS^E configuration

All nanoscale liquid chromatographic separations were performed using a directly-coupled NanoAcquity UPLC system and a nanoelectrospray source (Waters Corporation, Milford, MA, USA) fitted to a Q-Tof Ultima Global instrument (Waters Corporation, Milford, MA, USA). The system was composed of a binary solvent, auxiliary solvent and sample manager fitted with a heating and trapping module.

LC separations were performed using a Symmetry C18 trapping column (180 μm x 20 mm 5 μm) and a BEH C18 analytical column (75 μm x 250 mm 1.7 μm). The composition of solvent A was 0.1% v/v aqueous formic acid and solvent B 0.1% v/v formic acid in acetonitrile.

An aliquot of each sample was applied to the trapping column and flushed with 0.1% solvent B for 2 min at a flow rate of 15 μL min^{-1}. Sample elution was performed at a flow rate of 250 nL min^{-1} by increasing the organic solvent concentration from 3 to 40% B over 90 min, with a total run time of 115 min. The mass spectrometer was fitted with a universal nanoflow sprayer (Waters Corporation, Milford, MA, USA) and an applied capillary voltage of 3.5 kV was used. All analyses were conducted in technical triplicate for the undepleted plasma and quadruplicate for the IgY-12 fractionated samples.

Prior to and after each set of technical replicates, a quality control (QC) injection of 50 fmol PhosB was analysed using LC-MS^E and the data processed by PLGS. Where the peptide sequence coverage fell below 35% for PhosB, no further sample data were collected and the cause of the loss of peptide identification investigated and resolved. A minimum of four QCs were collected every 24 hr during sample data collection.

The precursor ion accurate masses and associated fragment ion spectra of the tryptic peptides were measured using a Q-ToF Ultima Global mass spectrometer (Waters Corporation, Milford, MA, USA) operated in electrospray data independent MS^E mode controlled by MassLynx v4.0 over the *m/z* range 50 - 1950. In-source decomposition was minimised by decreasing the cone voltage (RF Lens 1) to 25 V and using a source temperature of 80 ^{o}C. The mass profile on the quadrupole was set

to 400, 500 and 600 *m/z* with ramp and dwell times of 25%, 25%, 25% and 25% respectively. A collision energy ramp from 6 – 35 V was used for the elevated collision energy conditions. The spectral acquisition scan rate was 0.9 sec with a 0.1 s interscan delay

The time-of-flight analyser of the mass spectrometer was externally calibrated using the MS/MS spectrum obtained from the doubly charged precursor of the GFP peptide over a range of *m/z* 50 to 1300. The calibration was manually validated with an average ppm error across the mass range <10 ppm being obtained. GFP was used for lockmass correction (*m/z* 785.8426) and was infused via a NanoLockSpray interface at a constant rate of 500 nL min^{-1} at 500 fmol μL^{-1} and sampled every 60 seconds.

3.2.4 Processing of MS^E acquired data

The MS^E data were processed using PLGS v2.3 (IgY-12 depleted) or v2.4 (undepleted) and lockspray calibrated against GFP using data collected from the reference line during acquisition. The ion detection, clustering and protein identification have been explained in detail in Section 1.4.3.6. In brief, lockmass-corrected spectra are centroided, deisotoped and charge-state-reduced to produce a single accurately mass measured monoisotopic mass for each peptide and its associated fragment ions. Initial correlation of a precursor and potential fragment ions is achieved using time alignment (Geromanos, Vissers et al. 2009). Data processing parameters of 250, 100 and 1500 for the low, elevated and intensity threshold values respectively were specified in PLGS software.

3.2.5 Database interrogation using MS^E data

All PLGS processed data were used to interrogate an appropriate database using the same release of PLGS as that used for processing. Within the PLGS software, each database was randomised to form a concatenated database of genuine and random entries.

The IgY-12 depletion data were used to interrogate the IPI human database rel. 3.51 downloaded from (http://www.ebi.ac.uk/IPI/IPIhuman.html), appended with the sequences for porcine trypsin and rabbit glycogen phosphorylase (P00761 and P00489 - http://www.uniprot.org/). The undepleted plasma data were interrogated against IPI human database rel. 3.69.

For the MS^E data, database search parameters included a fixed modification of carbamidomethyl cysteine, one missed trypsin cleavage site with variable modifications of acetyl N-terminus, oxidation of methionine and deamidation of asparagine and glutamine.

The precursor and fragment ion tolerances were determined automatically by PLGS. The protein identification criteria included the detection of at least three fragment ions per peptide, seven fragment ions per protein and at least one peptide per protein using a 4% false discovery rate (FDR). PhosB was specified as the internal standard and the concentration specified (in fmol) in the PLGS workflow template to allow the Hi3 estimation of protein concentration.

The protein identifications obtained from each of the tryptic digests, analysed in triplicate were exported from the PLGS browser into Microsoft Office Excel. Within Excel, the protein identifications from the technical replicates were then filtered for replication ≥ 2 to generate a list of proteins identified from each individual maternal plasma sample.

3.3 Results and discussion

3.3.1 LC-MS^E analysis of undepleted plasma

Raw plasma (undepleted) from a complication-free gestation was tryptically digested and analysed using LC-MS^E. The inclusion of an internal standard of PhosB allowed estimation of the sample loading for each injection and the individual levels of each of the proteins identified. For each technical replicate, 74, 84 and 85 proteins were identified including the internal standard and any random database entries. The

average loading, calculated from summing the protein levels reported by PLGS within each replicate was 691 ng ± 66ng.

In total across all 3 technical replicates, 153 proteins were identified with an average sequence coverage of 24.8% and 14.6 peptides per protein. These values include the random (decoy) protein entries of which there were 40, providing a false discovery rate of 26.1%. The abundance of each protein was estimated in each replicate based on the reported level (ng) from PLGS and calculated as a percentage of the summed loading within the replicate, excluding the internal standard (ng).

Using the pivot table function in Excel, the protein information was filtered for replication with proteins observed in only one replicate being rejected. After filtering, 54 proteins were retained including the internal standard with a calculated FDR of 0.0% i.e. no random proteins were identified >1. The abundance levels for each protein were not recalculated after replication filtering but were calculated as an average across the technical replicates.

The average replication rate for the undepleted plasma proteins identified (replication ≥2) was 2.65 with 19.3 peptides per protein and 28.5% sequence coverage, Figure 3. 3**.**

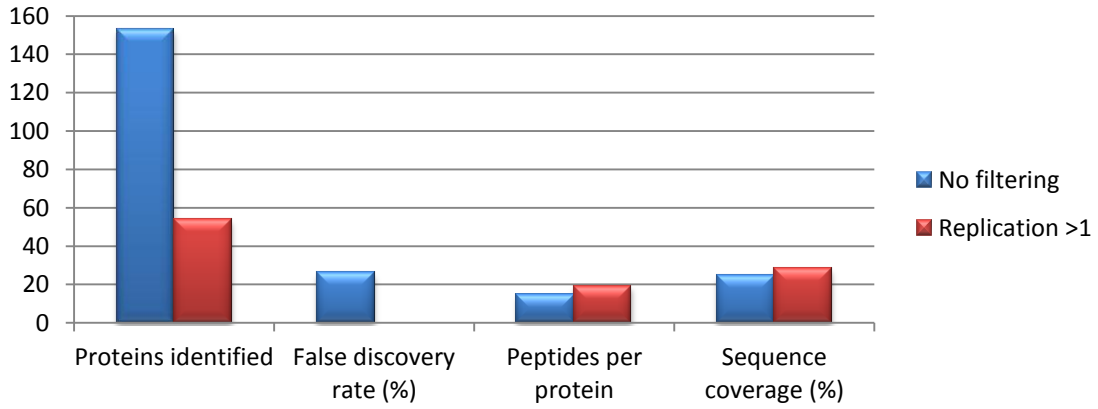


Figure 3. 3 : Effect of filtering protein identification results (replication ⩾2) from normal undepleted plasma analysed using LC-MS^{E} across three technical replicates.

Average abundance across all three technical replicates was estimated for each of the 54 proteins (replication ≥2) from the undepleted plasma. The most abundant protein

identified, as expected, was serum albumin at 24.6% with an average of 86 peptides and 61.9% sequence coverage, Figure 3. 4**.** All of the proteins depleted by an IgY-12 affinity column were identified in the analysis and 11 were among the 30 most abundant.

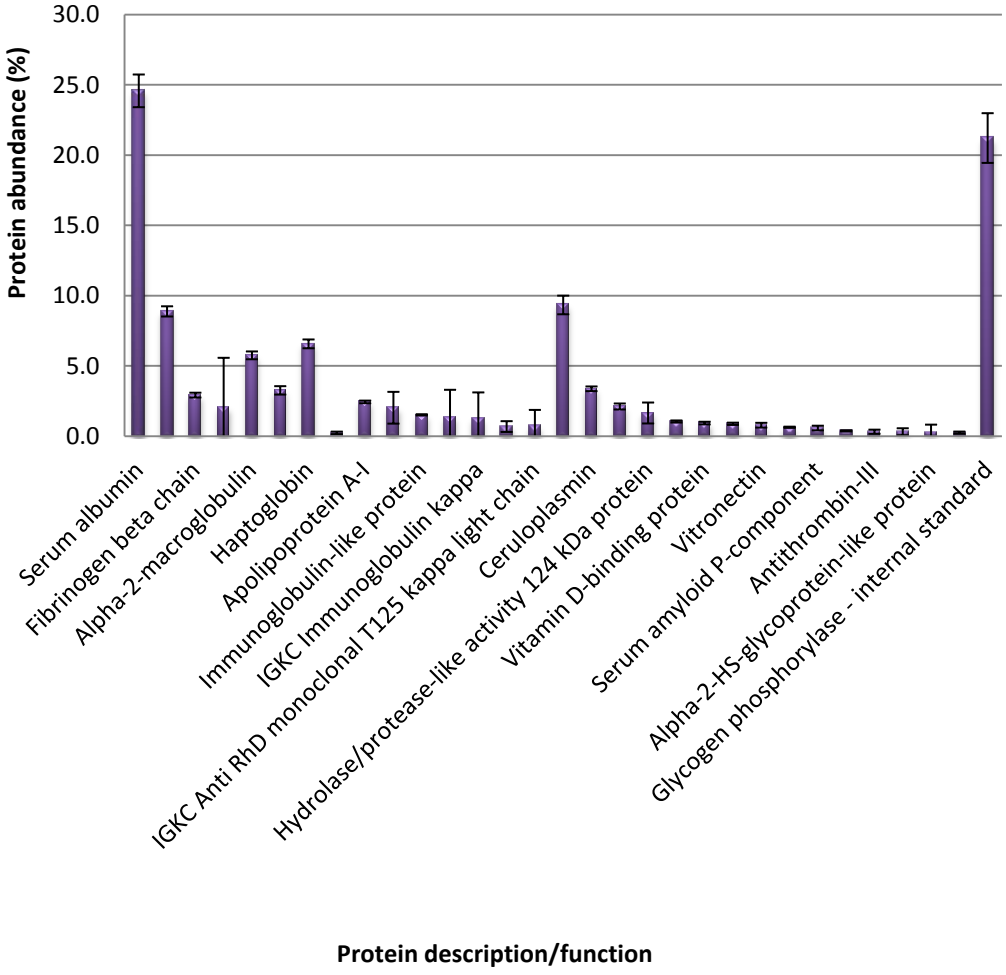


Figure 3. 4 : Protein abundance levels in undepleted normal maternal plasma estimated across three technical replicates analysed using LC-MS^{E} (replication $\geqslant$2).
The proteins on the left of the chart can be depleted with the use of an IgY-12 column.

In Figure 3. 5 a typical 2D gel is shown for undepleted plasma, demonstrating the dominance of the highly abundant proteins including albumin, indicated by an arrow.

Figure 3. 5 : 2D gel image of undepleted plasma.
The highly abundant albumin protein is identified by an arrow (taken from ProteomeLab IGY-12 High Capacity Partitioning Kit Brochure DS-10049A, issued 2006).

None of the proteins identified in the undepleted plasma are currently used in diagnostic tests for trisomy 21, confirming that depletion of the plasma would be required in order to target the biologically and clinically relevant part of the plasma proteome.

3.3.1 Protein identification from IgY-12 depleted individual maternal plasma

Eleven maternal plasma samples from normal, unaffected gestations were individually depleted of 12 highly abundant proteins using an IgY-12 spin column and analysed using LC-MS^E in technical quadruplicate.

In total, across all 44 experiments, 6,318 proteins were identified by PLGS with an average of 143.6 per replicate, without filtering of the data. This relates to 1,164 unique proteins of which 745 were only observed in one of the 44 LC-MS^E runs (64%) and an average 105.8 proteins identified per sample. The false discovery rate (FDR) prior to data filtering was 380 random identifications from 1,164, which equates to 32.7%, an average of 8.64 random entries per replicate analysis.

Using a replication filter retaining proteins observed in ≥2 LC-MS^E analyses out of the 44 collected, the number identified was 419 including the internal standard, of which 13 were random entries (FDR 3.1% for the data set). For the random entries identified, 7 were single peptide hits, 5 were observed with 2 peptides and 1 with 3 peptides. The average number of peptides identified and the sequence coverage was

6 per protein and 22.2% respectively. For the 20 proteins that were observed most frequently in the LC-MS^{E} runs, the average sequence coverage and number of peptides identified have been shown in Figure 3. 6**.**

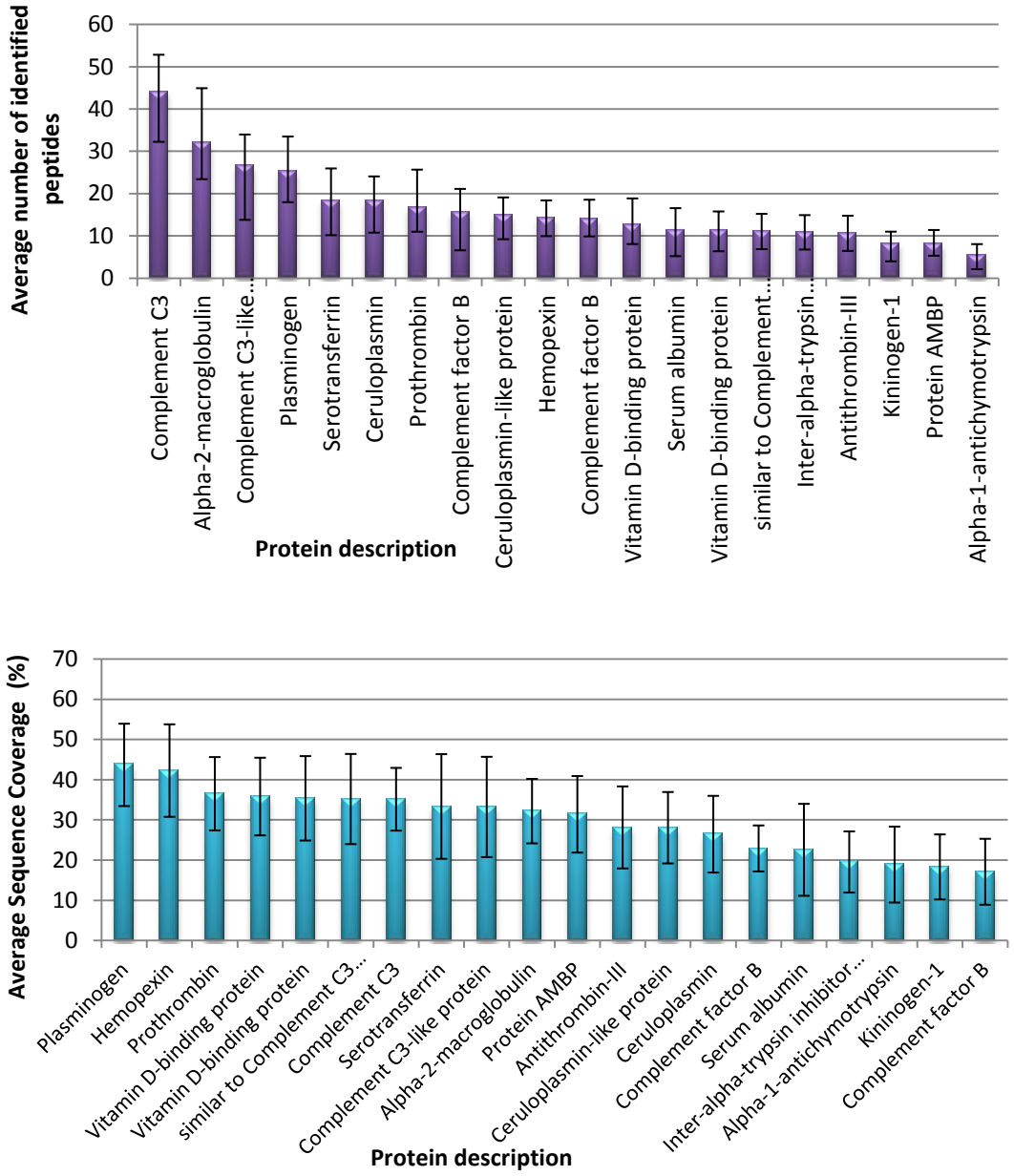


Figure 3. 6 : Bar charts depicting average number of peptides and sequence coverage analysed using LC-MS^{E} for the 20 most identified proteins from IgY-12 depleted plasma. The average number of peptides displayed in upper panel and sequence coverage in lower panel. Values are based on four technical replicates and include standard deviation error measurements after replication filtering ≥2.

The number of proteins identified in this study using the LC-MS^E approach on individual plasma depleted of 12 highly abundant proteins compares very favourably with conventional 2D-PAGE with LC-MS/MS methodology (Huang, Harvie et al. 2005). For our LC-MS^E approach an average 691 ng of tryptic digest was loaded in each technical replicate and after filtering ≥2 observations, 405 plasma proteins were identified compared to 47 proteins identified from 200 µg used in the gel-based experiment described by Huang. Identified peptide numbers from each protein in both MS approaches were relatively similar considering that the sample loading was almost 290-fold higher on the gel.

Three of the proteins that were depleted during the IgY-12 fractionation process were still evident in the 20 most frequently observed proteins. These were albumin, serotransferrin and alpha-2-macroglobulin. The suggested average depletion efficiencies detailed by the supplier of the column were 99.6%, 99.1% and 94.4% respectively. Typically as the sample concentration decreases, the number of identified peptides and sequence coverage reduces until the peptides are below the detection limit for the experiment. For albumin, the average sequence coverage was reduced from 61.9% with 86 identified peptides to 22.6% and 11.5 identified peptides. A similar reduction was observed for serotransferrin, indicating that the levels of these proteins had been extensively reduced. In the case of alpha-2-macroglobulin, although the average number of peptides dropped by 31.4%, the sequence coverage exhibited only a 6.5% decrease, Table 3. 2**.** This suggested that either the spin column used was not efficiently removing the alpha-2-macroglobulin or that the level of this protein in the plasma prior to depletion exceeded the maximum binding capacity.

Protein	Average Sequence Coverage		Average number of identified peptides	
	Undepleted	IgY-12 depleted	Undepleted	IgY-12 depleted
Albumin	61.9	22.6	86	11.5
Serotransferrin	58.6	33.4	60	18.4
α-2-macroglobulin	34.4	32.2	46.7	32.0

Table 3. 2 : Effect of IgY-12 depletion on average sequence coverage and peptide identification.
The concentration of three proteins, albumin, serotransferrin and α-2-macroglobulin, were compared to their respective levels in undepleted plasma**.**

A virtual 2D gel is depicted in Figure 3. 7 which compares the proteins identified from raw and IgY-12 depleted plasma, replication ≥2. The image clearly shows a 7.7-fold increase in the number of proteins identified from IgY-12 depleted plasma (purple) compared to undepleted (yellow). The depletion strategy has increased the range of both molecular weight and pI over which proteins are identified, increasing the depth of the plasma proteome coverage, by decreasing the levels of the highly abundant proteins. The undepleted plasma protein markers in yellow are shown in order to visualise those proteins identified in both sets of analyses i.e. with a yellow marker and purple outline.

Two proteins appear to be absent in the IgY-12 LC-MS^E analyses when compared to undepleted plasma, namely an uncharacterized protein (pI 6.35/mol. wt. 123,937 Da – UniProtKB E7ETN3) and fibronectin splice variant E, isoform 1 (pI 5.32/mol. wt. 262,439 Da – UniProtKB P02751-1). For the latter protein a number of other fibronectin splice variant species were identified in the IgY-12 depleted plasma, including P02751-2, -4, -5, -6, -10, -12, -13 and -14.

On closer investigation of the PLGS results for P02751-1 in undepleted plasma it was apparent that although P02751-1 isoform 1 was listed in the software browser, in fact 12 additional isoforms provided a possible identity of this protein with the same 13 tryptic peptides identified. Sequence coverage ranged from 7.7% to 9.7% with isoform 1 identified as most probable with 8.0% sequence coverage. A further 3 isoforms were identified with 12 peptides, Figure 3. 8**.** A number of these isoforms have molecular weights very similar to fibronectin isoform 1 and can be visualised on the virtual 2D gel of IgY-12 depleted plasma proteins as a cluster of purple markers just below the 250 kDa position.

In earlier versions of PLGS including 2.3 and 2.4 these types of protein identification issues were quite apparent. It was often not possible to ascertain why PLGS had assigned a protein identity to one species over another; in this case it has selected one of 13 proteins in the mid-range of the sequence coverage. This type of problem occurs particularly when identifying eukaryotic proteins using a database containing isoform species. In PLGS 2.5 and later, the browser reports the protein with the highest sequence coverage, which means that the same protein should be reported each time, if identical peptides are identified. For plasma using the IPI databases used herein, on average for each protein reported in the browser, there were 4.7 isoforms, fragments or splice variants in the database with some or all matching peptides. For prokaryotic protein species this is less of a problem.

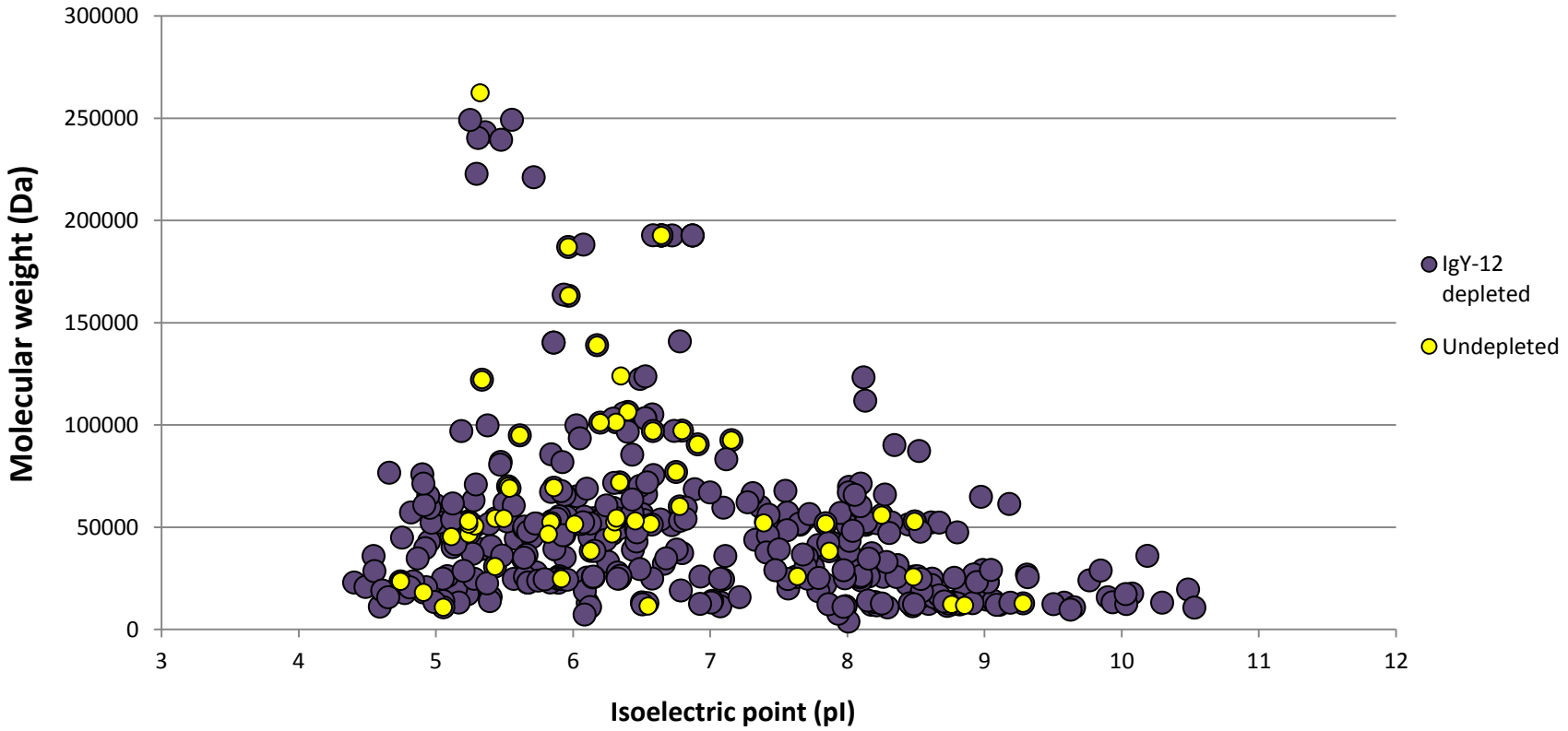


Figure 3. 7 : Virtual 2D gel comparing proteins identified in undepleted and IgY-12 depleted plasma.
The undepleted plasma proteins are depicted using smaller markers, such that a protein identified in both undepleted (yellow) and depleted (purple) LC-MS^{E} analyses are depicted by a yellow marker with a purple outline. Proteins displayed have been filtered for replication ≥2.

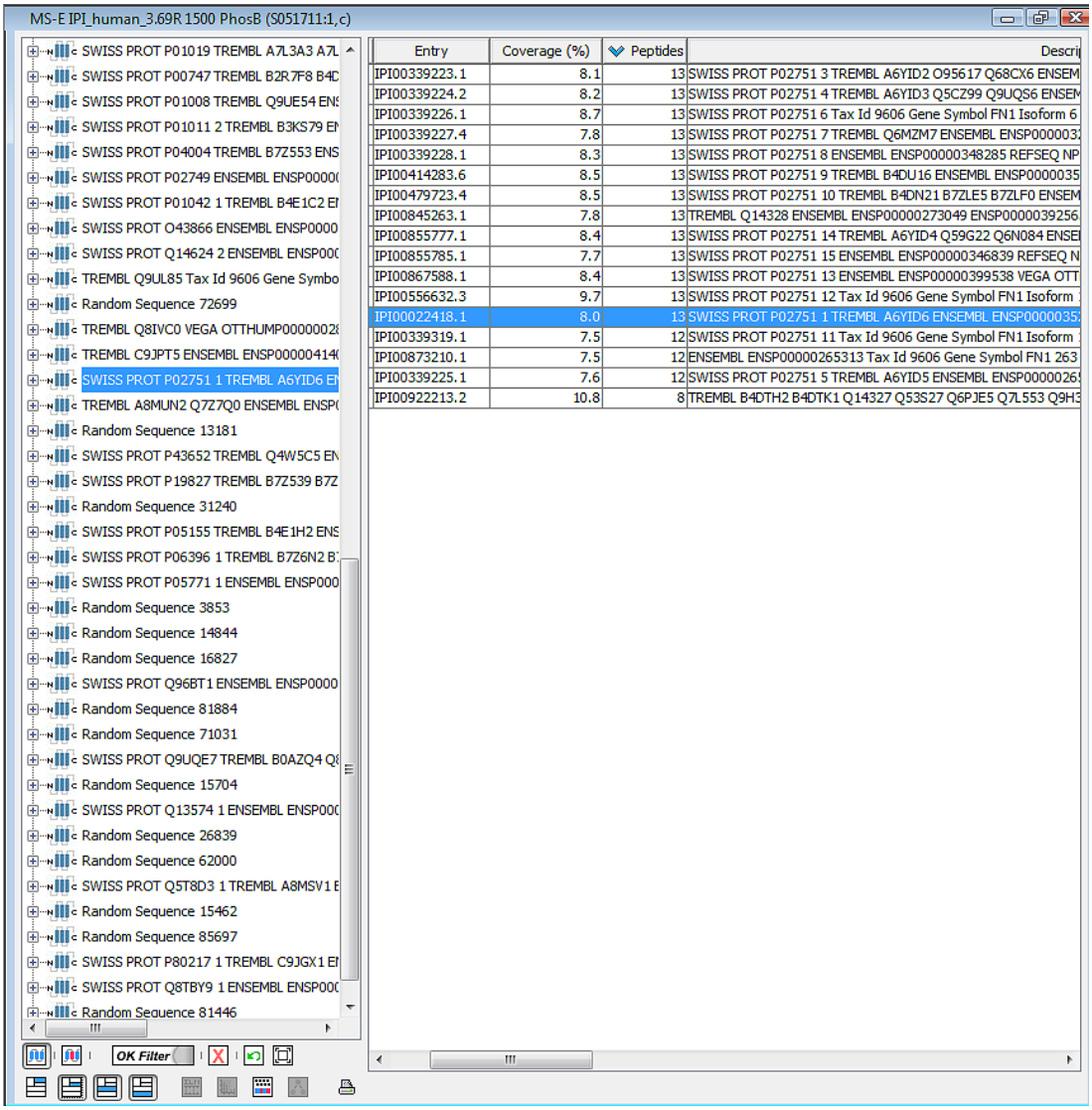


Figure 3. 8 : PLGS v2.4 report identifying fibronectin splice variant E, isoform 1 as the most probable identity for a protein in undepleted plasma.
A closer examination of this protein (right panel, highlighted in blue) indicated that 13 possible protein species contain the same 13 identified tryptic peptides.

This raised the question as to which of the many and varied fibronectin isoform/s were present in each sample, when comparing multiple data sets. In many cases where these ambiguities exist, a bottom-up proteomics experiment may not be able to answer that question. One can avoid some of the issues by excluding isoforms from the database under interrogation, and then if a protein is of particular interest (regulated expression), go back retrospectively into the data and explore the possible isoforms or splice variants and identify the most likely candidate/s. In some cases, it

may be necessary to purify and fully characterise the protein of interest in order to confirm its biochemical role or to ascertain if a mixture of isoforms were present.

Alternative splicing of pre-mRNAs in eukaryotes has been estimated to take place in up to 60% of all genes (Modrek and Lee 2002; Modrek and Lee 2003) and approximately 95% of multi-exon genes (Pan, Shai et al. 2008). Databases such as IPI used in this study, TrEMBL and Ensembl, incorporate putative sequences generated from genomic sequencing and annotation, indicating likely protein coding regions. Evidence of translation of gene products may not exist for some or many of the protein sequences in the database. Blakeley and co-workers estimated that in the human Ensembl 48 database, there were an average 1.8 protein isoforms/gene and, in genes encoding more than one non-redundant isoform, 41% encoded >1 isoform with an average of 3 per gene (Blakeley, Siepen et al. 2010). In the study, ioforms of a protein TGFβ were compared between the Ensembl 48 and SwissProt databases and <2% of the peptides were found to be unique to one of the annotations, indicating the difficulty in unambiguously identifying protein isoforms from a single experiment. The authors recommended a targeted strategy, to identify unique N-terminal or exon spanning peptides and using a combination of proteomics databases to ensure that all novel isoform sequences are considered in database interrogations. The use of alternative proteases such as Lys-C or Arg-C were considered since they in general generate longer peptides, potentially spanning unique regions of proteins isoforms (exon junctions) whilst maintaining the quality of the MS/MS spectra obtained due to the C-terminal basic residue.

The second protein absent from the IgY-12 dataset but observed in undepleted plasma (uncharacterized protein E7ETN3) has sequence identity with UniProtKB P00751 Complement factor B. P00751 was listed in the PLGS software as having 14 identical tryptic peptides to E7ETN3 which was identified with 18 peptides. Confusingly, there were two proteins in the list that were identified with 2 additional tryptic peptides to that found for E7ETN3 (20) but were not listed in the browser as the most probable identity.

3.3.2 Stringent filtering of IgY-12 depleted plasma protein identifications

The protein identifications generated by PLGS for the IgY-12 depleted plasma samples were more stringently filtered. Previously a replication filter of ≥2 was used however, to stringently filter the results the protein identifications were retained only if they were identified in ≥2 replicates *per plasma sample*. This reduced the number of proteins identified to 345 across all the samples excluding the internal standard. There were no random entries within this data set and so the FDR was determined to be 0.0%.

For the 345 stringently filtered proteins, the frequency of identification was determined and plotted in
Figure 3. 9 with 83 proteins identified in only one sample. The stringent filtering of the data had reduced the number of protein identifications by approximately 18% but this did not account for the variation in frequency of identification across the data set, although one explanation for this could be inter-sample variation. It was feasible that a number of protein levels across the sample set could be above the detection limit in some cases (this to some extent is dependent on the stringency of filtering) and below in others thus resulting in the apparent presence or absence of proteins between samples.

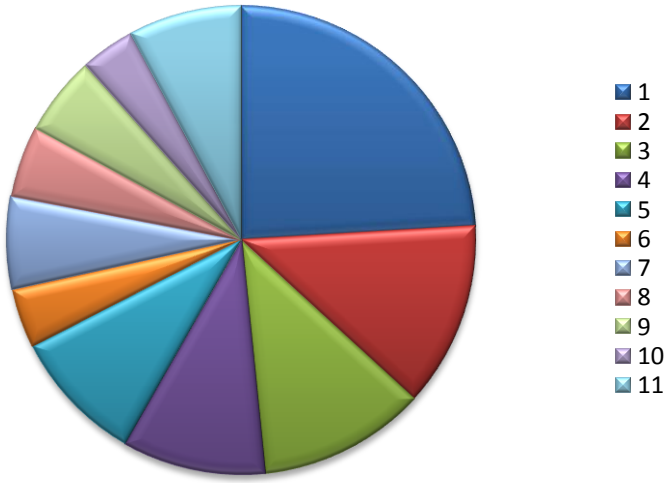


Figure 3. 9 : Pie chart indicating the frequency of protein observation in 11 plasma samples using stringently filtered identifications in LC-MS^E experiments.
The total number of proteins identified was 345 (excluding internal standard) with 24% observed in one sample (replication ≥2 per sample).

3.3.3 Semi-quantitative measurements of IgY-12 depleted individual maternal plasma protein levels

It was assumed that since human plasma protein concentration is relatively stable (60 – 85 g L^{-1}) then if all samples were treated identically there should be similar levels of the proteins within the samples after depletion and tryptic digestion. Thus the results described here may be described as semi-quantitative.

An equal volume of each IgY-12 tryptically digested maternal plasma sample was analysed using LC-MS^E in the presence of the internal standard PhosB. Using the Hi3 protein estimation approach, approximate levels of each of the identified proteins were determined.

Of the 44 replicate LC-MS^E analyses, in 30 (10 samples) the internal standard was identified and thus the protein level in each replicate experiment could be estimated. The absence of the internal standard from the experiment indicated that the optimal level of sample loading had been exceeded. Where proteins were common to all samples, it would be expected that those levels would also be similar, since all samples originated from uncomplicated pregnancies in women of a narrow restricted age, BMI and ethnicity range. In contrast, for many proteins the levels determined varied widely. As shown previously in Figure 3.4, the intra-variation in quantitative measurements between replicate analyses of the same sample is very low, and so the variation that was being observed across the 10 samples was due to variation in protein levels within the group, rather than experimental error, Figure 3. 10**.**

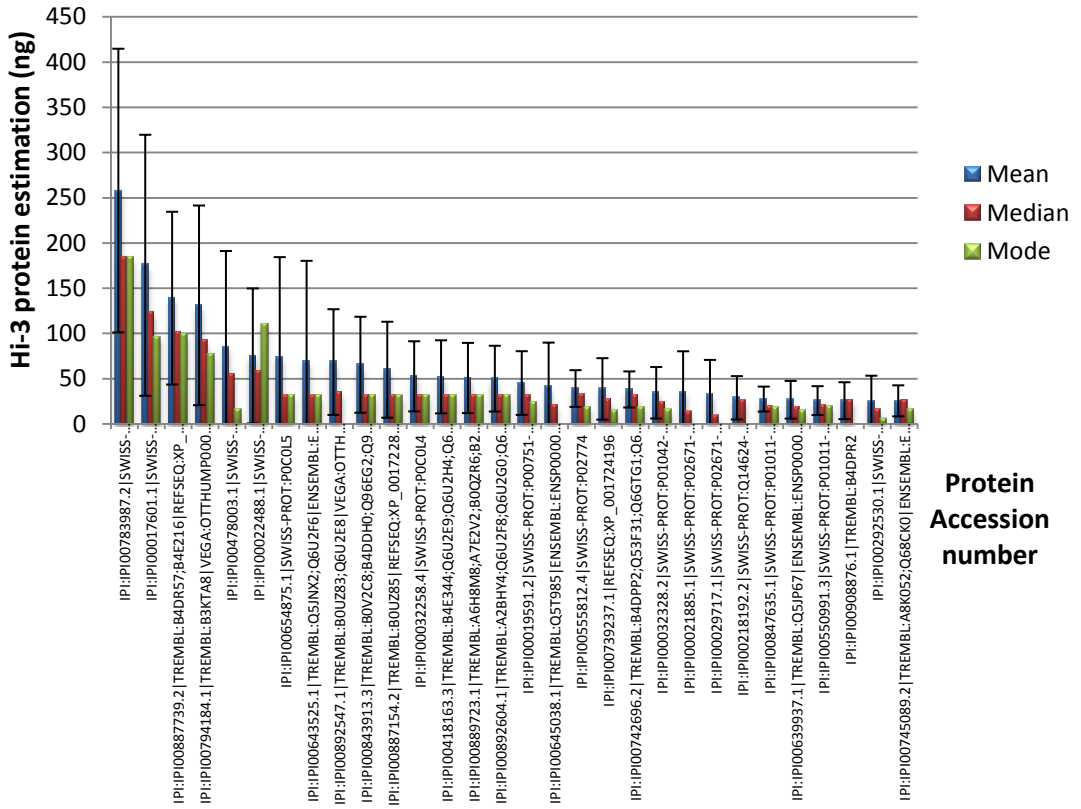


Figure 3. 10 : Semi-quantitative protein levels based on ten IgY-12 depleted normal plasma samples estimated using the Hi3 approach.
The mean (with standard deviation), median and modal values (ng) based on three technical replicates measurements/sample are expressed, reflecting inter-sample variation in plasma protein concentration.

Plasma components from normal healthy individuals, when measured biochemically fall within concentration windows termed reference ranges. Reference ranges have been published for a large number of plasma constituents e.g. hormones, metabolites, sugars and proteins (Tietz and Amerson 1990) and are routinely used by NHS pathology laboratories. An example of plasma reference ranges obtained from normal pregnancies across three trimesters is shown in Table 3. 3. The plasma component concentrations can vary considerably between patients in the same trimester, e.g alkaline phosphatase 82 – 274 U/L (36 weeks), but also across the gestational period. Thus any successful proteomics approach to identifying candidate biomarkers *must* be able to accurately measure the concentration range within which each protein falls in both normal and diseased conditions, which further emphasises the requirement to have low technical variance in the quantitation measurements to allow the biological variance to be correctly assessed.

	12 WEEKS	24 WEEKS	36 WEEKS
Albumin g/L	35 – 45	30 –38	22 – 37
Alkaline phosphatase U/L	27 – 90	32 – 108	82 - 274
Cholesterol mmol/L	3.3- 7.3	4.2 – 9.3	4.9 – 10.8
Creatinine μmol/L	48 – 78	41 – 78	47 – 87
Glucose mmol/L	2.9 – 5.9	2.7 – 5.3	2.7- 5.5
HDL cholesterol mmol/L	1.3 – 3.1	1.4 – 3.4	1.4 – 3.3
Triglycerides mmol/L	1.1 – 3.7	1.7 – 4.1	2.8 – 7.1
Urate mmol/L	0.10 – 0.27	0.12 – 0.31	0.16 – 0.42
Urea mmol/L	1.9 – 6.2	1.8 – 5.6	1.6 – 5.0

Table 3. 3 : Reference ranges for plasma components in normal pregnancy.
Adapted from (Lockitch and Wadsworth 1993)

3.4 Conclusions

A raw plasma sample was tryptically digested and analysed using LC-MS^E and the proteins identified compared with those from eleven individual plasma samples depleted of 12 highly abundant proteins. The depletion approach employed was effective at significantly reducing the levels of the 12 highly abundant proteins. Although generation of the depleted (flow-through) fraction was relatively quick, the stripping, neutralisation and equilibration of the spin columns was laborious. The depleted plasma samples provided deeper plasma proteome characterisation with a 6-fold increase the number of proteins identified, confirming the need for depletion in biomarker discovery studies. LC-MS^E methodology identified more proteins than gel-based LC-MS/MS experiments using less than 1/200th the amount of depleted plasma.

Quantification of gel-based proteins can be problematic as this can be a time consuming task to ensure that the spot volume has been correctly determined by the software. Frequently, the identification of multiple proteins from one 2D gel spot invalidates any measurement of the protein level as all of the proteins present contribute to the density of staining. In contrast, in LC-MS^E experiments the proteins are identified and quantified in the complex mixture.

Use of technical replicates in the LC-MS^E experiments allowed for filtering of the protein identification results. Two types of filtering were employed, based on the

number of observations across *all* analyses (regardless of sample) or the more restrictive approach that required a protein to be observed in >1 technical replicate *per sample*. The latter resulted in the least number of proteins being identified to a sample with a false discovery rate of 0%, indicating that perhaps this type of filtering may be too stringent and that many proteins that were identified, but then discarded from each sample have been rejected unnecessarily.

Samples for IgY-12 depletion were selected from normal gestations and fell within a confined range of BMI, age and gestational age. The LC-MS^E analyses performed were semi-quantitative based on an assumption that all plasma samples contained similar levels of total protein. Within this narrow sample group a wide variation in terms of the proteins identified and the levels of each was observed confirming inter-patient variation. This naturally occurring biological variation would be lost if plasma samples from a number of patients were pooled together as only single point measurements would be collected.

A number of issues were identified in the protein identification results from this dataset regarding the potential presence of homologous isoform sequences in the database. Different isoforms of the same protein were identified between samples and between technical replicates further complicating the analysis of the data. On occasion it was not possible to determine why the search engine had preferred one isoform as the most probable identification over another.

Based on the requirement for individual samples to be depleted of highly abundant proteins, the presence of a number of high abundance proteins *after* IgY-12 depletion and the time consuming nature of the spin column approach, a decision was taken to increase the number of proteins depleted to 14 and to use an LC-based approach which could be fully automated, to increase sample throughput. In addition, the depleted samples would be analysed using LC-MS^E using near identical sample loading to increase the confidence in the quantitative results obtained.

Chapter Four: Automated IgY-14 fractionation of maternal plasma analysed using LC-MS^{E}

4.1 Introduction

Manual depletion of 12 highly abundant proteins from maternal plasma, as described in the previous chapter, was found to be highly labour intensive and not suitable for the partitioning of a large number of samples that would need to be quantitatively analysed in the next phase of the study. An automated depletion step had the potential to improve throughput and through the use of an IgY-14 column could be used to deplete a further subset of plasma proteins.

The IgY-14 partitioning system depletes the 12 highly abundant proteins previously described in Table 3. 1 as well as complement C3 and apolipoprotein B. Complement C3 was observed in all the replicate analyses of IgY-12 plasma and the concentration on column was estimated at an average of 258 ng, the most abundant protein observed.

An automated depletion system would be essential for a validation trial using MRM-based assays if the candidate biomarkers proteins were below the limit of quantitation or detection in raw plasma.

4.1.1 Screening tests for Down's syndrome in the UK

The Serum, Urine and Ultrasound Screening Study (SURUSS) was a large collaborative review of antenatal screening for Down's syndrome in the UK involving 25 centres, funded by the Health Technology Assessment Programme to determine best practice (Wald, Rodeck et al. 2003). At the time of publication it was the largest study on women assessed in both the first and second trimester. Since 1988, a combination of serum markers with maternal age has become the screening method of choice. Centres adopted either the double or triple test, α-fetoprotein (AFP) with choriogonadotropin protein (hCG) or AFP, hCG with unconjugated oestriol (uE_3) respectively. The SURUSS project showed that based on efficacy, safety and cost the integrated test yielded an 85% detection rate (DR) with 1.2% false positive rate (FPR). In the first stage of the integrated test, two measurements are taken, the concentration of Pregnancy Associated Plasma Protein-A (PAPP-A)

and Nuchal Translucency (NT) using an ultrasound scan at 10 completed weeks. In the second trimester (14-20 completed weeks) AFP, uE_3, free ß-hCG and inhibin A levels were determined. The measurements of the six markers in combination with information on the woman's age were used to estimate the risk of having a pregnancy with Down's syndrome. Where no NT measurement was available, for a DR of 85% the FPR was 2.7%. If the risk is determined to be ≥1 in 150 this would be classed as screen-positive for Down's syndrome and a further invasive diagnostic test is advised. About 1 in 65 women screened will be classed as screen-positive, of which 1 in 6 women will have an affected pregnancy.

In the event that a woman first presents in the second trimester, the SURUSS report suggested that a quadruple test involving AFP, uE_3, inhibin A and ß-hCG would be the preferred screening approach, giving a FPR of 6.2% with a DR of 85%.

For serum markers and NT measurements there is a distribution of values for both Down's syndrome and unaffected pregnancies, with a region of overlapping values from both conditions, Figure 4. 1. Results from screening tests are reported as multiples of the median (MoMs) rather than absolute levels (Cuckle, Wald et al. 1987) at a given gestational age. Thresholds are applied e.g <0.5 MoM or >2.5 MoM in order to assess risk. In Figure 4. 2 the calculation of the MoM for a hypothetical biomarker at a range of completed weeks of gestation is shown. In the example, the median level at 14 weeks would be 100 IU mL^{-1} for unaffected pregnancies. A woman with a level of 50 IU mL^{-1} would be classed as 0.5 MoM (50/100=0.5). The MoMs for all the serum predictive markers are collected and the combined risk assessed. A number of factors affect risk assessment and so adjustments for the following need to be taken into account:

- Serum concentrations tend to be decreased in heavier women and increased in lighter women
- Ethnicity – AFP, free ß-hCG and PAPP-A levels tend to be higher in Afro-Caribbean women
- *In vitro* fertilisation can affect the levels of free ß-hCG and uE_3
- AFP and uE_3 levels tend to be reduced in cases of insulin-dependent diabetes mellitus

- Smoking can affect the levels of free ß-hCG, PAPP-A and inhibin A in serum.

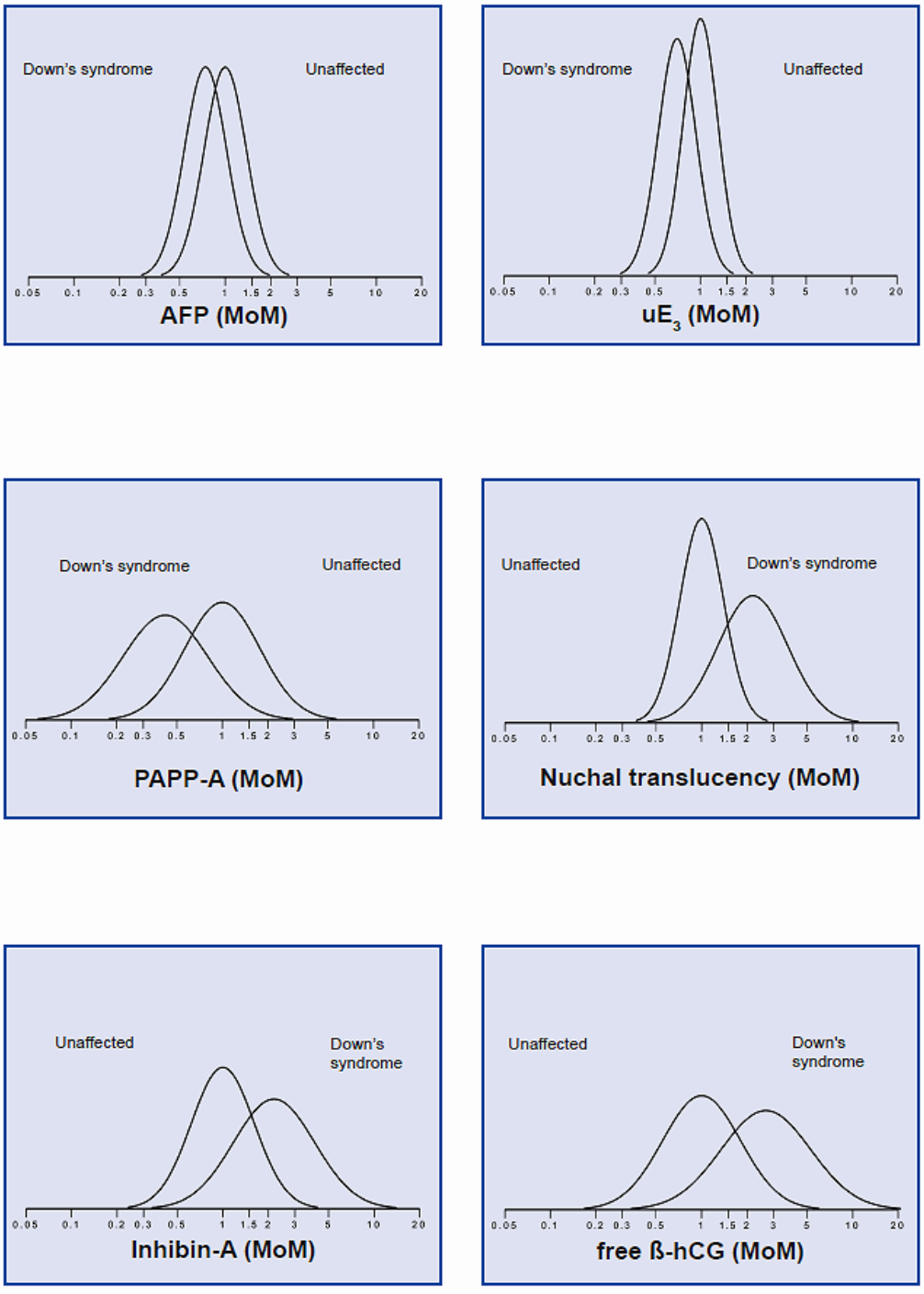


Figure 4. 1: Relative frequency distributions of the markers in Down's and unaffected pregnancies.
The risk of Down's syndrome is the same as the background risk in the general population at the intersection points.

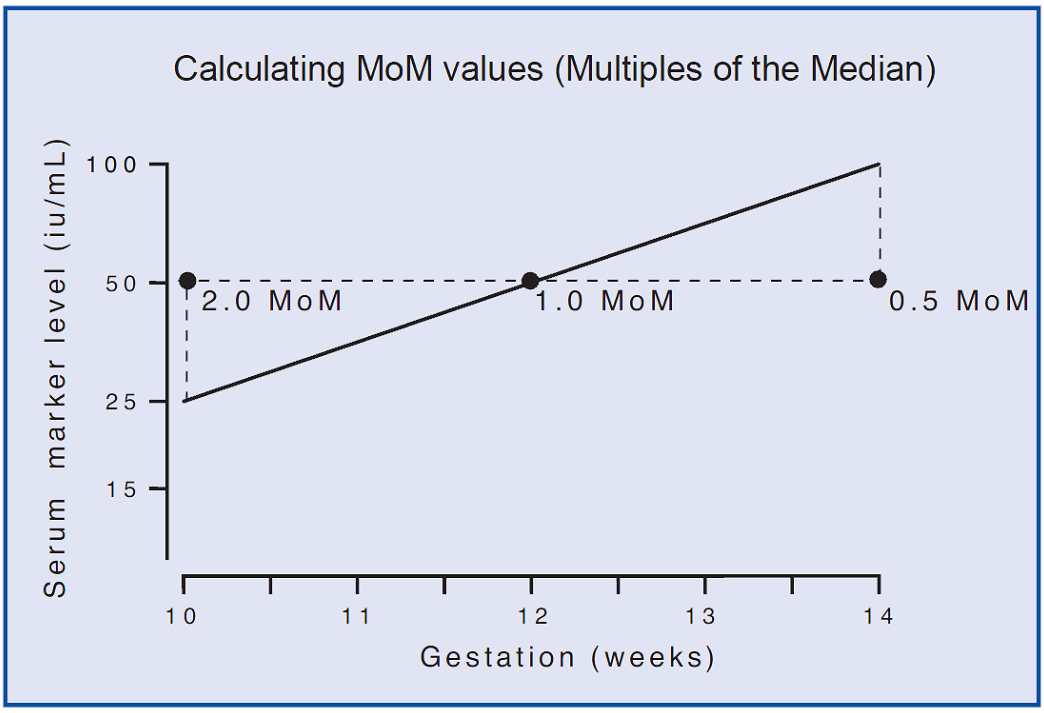


Figure 4. 2: Example of multiples of the median determination.
A hypothetical serum marker for trisomy 21 has a median level (1.0 MoM) of 25, 50 and 100 IU mL^{-1} at 10, 12 and 14 weeks gestation. If a woman presents at 10 completed weeks with a marker level of 50 IU mL^{-1} this would be converted to 2.0 MoM (50/25=2.0) i.e. twice the median value at that gestational age.

Methodologies for screening serum predictive markers of Down's syndrome are based on enzyme-linked immunosorbent assays (ELISA),

Figure 4. 3. Analyte (serum or standard) is applied to a microplate pre-coated with immobilised capture antibodies to the biomarker of interest. After an incubation step any unbound material is washed off and an enzyme-linked second antibody to the captured biomarker is applied and incubated. After a second wash step the substrate for the enzyme is added, incubated and a colour develops dependent on the level of biomarker present.

Variations on the ELISA principle include the DELFIA® system (dissociation-enhanced lanthanide fluorescent immunoassay) from PerkinElmer Inc., USA which uses time-resolved fluorometric assay technology providing a wider dynamic range than traditional ELISA since the reporter is not enzyme-based (4-5 orders). The KRYPTOR system from B·R·A·H·M·S* GmbH, Germany utilises TRACE

technology (time resolved amplified europium cryptate emission) involving non-radioactive energy transfer from a donor molecule to an acceptor (XL665, a phycobilliprotein pigment purified from red algae) following a successful immune reaction. The fluorescence is proportional to the biomarker concentration obtained through a double selection of spectral (wavelength) and temporal (time resolved) measurements. This principle is based on the 1987 Nobel Prize winning work of French chemist Jean-Marie Lehn.

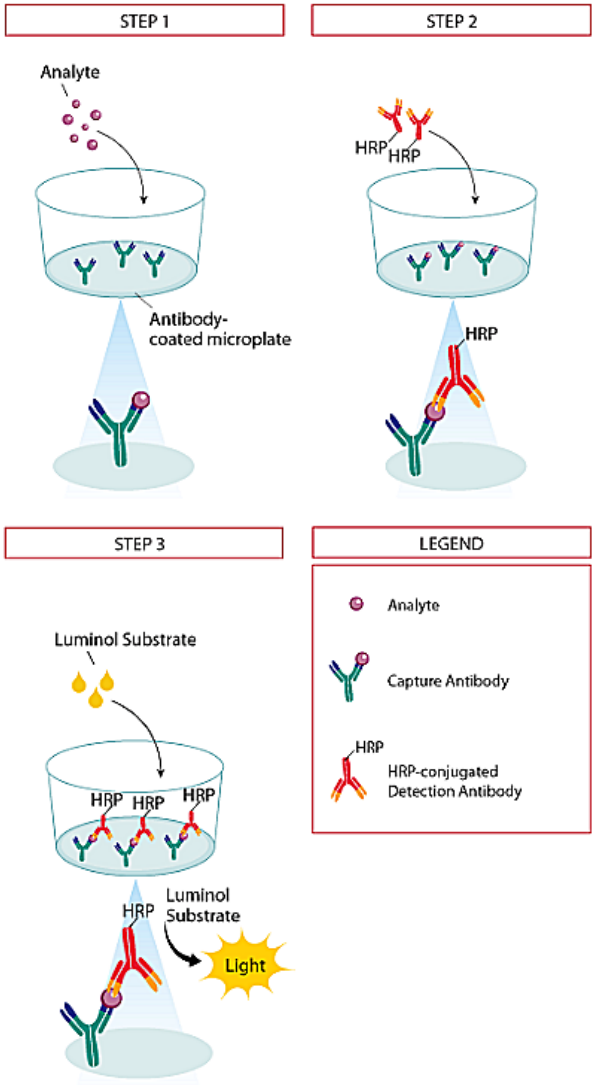


Figure 4. 3: Assay principle for enzyme-linked immunosorbent assay
The serum or standard is applied to a microplate containing capture antibodies. A second enzyme-linked antibody (in this case horseradish peroxidase) is applied and when the substrate for the enzyme, tetramethylbenzidine, is added a blue colour develops of which the absorbance at 450 nm is proportional to the amount of bound biomarker.
Taken from ELISA Reference Guide & Catalog from R&D Systems Inc., USA.
(http://www.rndsystems.com/resources/images/6836.pdf).

Table 4. 1 highlights a selection of commercial ELISA-based assays systems indicating sample volume requirements and limits of detection. Total assay times vary from <10 minutes to in excess of an hour, depending on the biomarker and technology in use.

Target Protein	**Limit of Detection**	**Sample Volume (µL)**	**Manufacturer/Distributor**
PAPP-A	0.19 µg mL^{-1}	10	Bio-Line, Belgium
	1.8 ng mL^{-1}	50	B·R·A·H·M·S* GmbH, Germany
Free β-hCG	0.16 IU L^{-1}	26	B·R·A·H·M·S* GmbH, Germany
	0.25 IU L^{-1}	10	GenWay Biotech Inc., USA
Inhibin A	5 pg mL^{-1}	50	DEMEDITEC Diagnostics GmbH, Germany
	1 pg mL^{-1}	100	Antibodies-online Inc., USA
AFP	0.23 ng mL^{-1}	14	B·R·A·H·M·S* GmbH, Germany
PIGF	7 pg mL^{-1}	100	Quantikine® R&D Systems Inc., USA
	<0.5 pg mL^{-1}		DELFIA®, PerkinElmer Inc, USA

Table 4. 1: Limits of detection for commercial ELISA-based assays.
Target proteins are predictive markers for trisomy 21 and/or pre-eclampsia
Key: PAPP-A – pregnancy-associated plasma protein A
β-hCG – Free choriogonadotropin subunit beta protein
AFP – alpha-fetoprotein
PIGF – placenta growth factor

The aims for this work were:

- To develop an automated depletion strategy to partition maternal plasma.
- Comprehensively characterise and quantify the depleted plasma from three obstetric conditions using an LC-MS^E approach.
- Compare the proteomics approach with commercial predictive assays used in Down's syndrome screening.
- Compare biomarkers suggested by the literature for trisomy 21 and pre-eclampsia with the results from this study and evaluate their potential benefits in a screening process.

4.2 Materials and methods

4.2.1 Material suppliers

A Seppro® IgY14 LC2 column kit (Sigma Aldrich, Gillingham, UK) was obtained including an LC column (6.4 x 63.0 mm bed volume 2 mL), dilution, stripping and neutralisation buffers.

Double centrifuged maternal plasma samples were supplied by Prof. Kypros Nicolaides of King's College Hospital, London, UK with full ethical approval.

Thirty maternal plasma samples were selected for this part of the study, spanning a range of age, BMI, ethnicity with equal numbers of samples identified as normal, trisomy 21 or severe onset pre-eclampsia pregnancies. The average maternal age was 33.6 years, BMI 26.2 and gestational age 89.3 days. Full sample details have been included in Appendix B.

Rapigest surfactant was obtained from Waters Corporation (Milford, MA, USA). Dithiothreitol was supplied by Melford Labs. (Ipswich, UK). Glu^1-Fibrinopeptide B peptide (human), sodium azide, iodoacetamide, ammonium bicarbonate, LC-grade water and acetonitrile were purchased from Sigma Aldrich (Gillingham, UK) and sequencing grade trypsin from Promega (Madison, WI, USA). Mass spectrometry solvents were supplied by MallinckrodtBaker Inc. (Phillipsburg, NJ, USA). Spin-X cellulose acetate centrifuge tube filters were supplied by Costar (Corning Inc.,

Tewksbury MA, USA), 2 mL square 96-well trays by Beckman Coulter (Fullerton, USA) and 5 kDa nominal molecular weight cut-off (NMWCO) spin columns and 0.45 µm nitrocellulose filter discs by Millipore (Billerica, MA, USA). Sample vials (LCMS Certified) were purchased from Waters Corporation (Milford, MA, USA) fitted with pre-slit PTFE/silicone septa in the caps.

4.2.2 Sample preparation

Each individual maternal plasma sample was thawed from -70 ^{o}C at room temperature and inverted a number of times to ensure homogeneity prior to a 200 µL aliquot being removed. Each aliquot was diluted to 1 mL with 1 x dilution buffer and centrifuged using a 0.22 µm Spin-X cellulose acetate centrifuge tube filter prior to depletion.

4.2.2.1 Fractionation of human plasma using an IgY-14 LC2 column

An automated method was developed to deplete maternal plasma using an IgY-14 LC2 format chromatography column, utilising the HPCF module of a ProteomeLab™ PF2D system (Beckman Coulter, Fullerton, USA). The HPCF module comprised of a single quaternary pump, sample loop injector with UV detector and fraction collection (FC/I module) using a 2 mL capacity square 96-well plate. All aqueous solutions used in the depletion protocol were filtered through a 0.45 µm nitrocellulose filter disc and degassed under vacuum for at least 15 minutes.

The entire system was flushed with water to remove any traces of residual organic solvent in the flow path, using a connector in the place of the column. The four buffers required for the depletion step were configured as follows:

- Line A1 – 1x Dilution Buffer
- Line A2 – 1x Stripping Buffer
- Line A3 – 1x Neutralisation Buffer
- Line A4 – 1x Dilution Buffer containing 0.02% sodium azide

A method was developed to load the diluted plasma onto the IgY-14 column whilst sequentially using dilution, stripping, neutralisation and dilution buffer over a 50 min period. For the last depletion protocol of each day, a modified method was utilised to equilibrate the column into dilution buffer containing sodium azide. Fractions were collected from the column during the chromatographic steps and detection of protein elution was achieved using absorbance at 280 nm with a 5 Hz sampling rate. The timetable for the chromatographic separation is shown in Table 4. 2.

Time	**Buffer Line**	**Buffer**	**Flow rate (ml min^{-1})**	**Fraction Collection Rate (min^{-1})**
0.00	A1	Dilution	0.2	1
17.01	A1	Dilution	1.5	1
22.01	A2	Stripping	1.5	1
36.01	A3	Equilibration	1.5	1
42.01	A1 or A4*	Dilution or dilution with sodium azide	1.5	1
50.00			0	0

Table 4. 2: Timetable for chromatographic steps in an IgY-14 LC2 depletion methodology. * Two methods were developed differing in the buffer used to equilibrate the column during the final chromatographic step. For the final separation of each day, the column was equilibrated into dilution buffer containing sodium azide.

With the connector remaining in place of the column, the system was manually flushed with each of the four buffers in turn to prevent contamination of the column once fitted.

The column was connected to the system and a full loop 250 µL injection of 1x dilution buffer was used to perform a blank chromatographic run using the timetable shown in Table 4. 2**,** before any plasma sample was applied. A blank run was completed between different plasma samples being applied to the column.

Two depletion steps, each containing the equivalent of 50 μL of plasma, were performed for each of the thirty maternal plasma samples to ensure that sufficient sample was generated for LC-MS^E analysis. The (depleted) fractions that eluted between 5 and 18 min from both chromatographic separations from each sample, were combined and concentrated using 5 kDa NMWCO spin columns, to generate a single sample for tryptic digestion.

4.2.2.2 Tryptic digestion of IgY-14 fractionated plasma

IgY-14 depleted plasma fraction (total volume approximately 1 mL) was transferred to a vial containing lyophilised Rapigest surfactant, final concentration 0.1% w/v and gently agitated until fully dissolved. The contents were transferred to an 5 kDa NMWCO spin column and centrifuged at 14,000 g, 4 ^oC until the volume was approximately 50 μL.

The contents were transferred to a 0.5 mL microfuge tube (Fisher Scientific Ltd, Loughborough, UK) and incubated in a water bath at 80 ^oC for 15 min. A 5 μL aliquot of 100 mM dithiothreitol in 100 mM ammonium bicarbonate (NH_4HCO_3) was added to the plasma and thoroughly agitated prior to incubation at 60 ^oC for 15 min, followed by the addition of 5 μL of 200 mM iodoacetamide in 100 mM NH_4HCO_3 and incubated at room temperature, in the dark for 30 min.

A vial containing 20 μg of trypsin was fully resolubilised in 20 μL of 100 mM NH_4HCO_3 with 2 μL (2 μg) transferred to the plasma sample and thoroughly agitated. The sample was incubated overnight at 37 ^oC.

The following day, 2 μL of concentrated formic acid were added to the sample and incubated at 37 ^oC for 15 min, prior to filtration through a 0.22 μm Spin-X cellulose acetate centrifuge tube filter. An aliquot of the tryptically digested sample was removed for analysis and the remainder stored at -20 ^oC until required. Typically 45 - 50 μL of tryptic digest was obtained for each depleted plasma sample giving a final concentration of approximately 4.8 μg μL^{-1} (from two combined IgY-14 depletions).

Sample loading was optimised based on protein identification rate. A range of sample loading was investigated from 26 ng to 3.2 μg with the maximum number of

proteins identified in the range 480 ng – 780 ng of tryptic digest on column, Appendix C.

An aliquot of each tryptic digest was diluted 8-fold and combined with a solution containing 200 fmol μL^{-1} MassPREP™ glycogen phosphorylase (PhosB) tryptic digestion standard in 0.1% v/v aqueous formic acid. This produced a sample containing approximately 300 ng μL^{-1} IgY-14 depleted plasma and 100 fmol μL^{-1} PhosB. Endogenous glycogen phosphorylase was not observed in any of the analyses in this work, and was used as an internal standard for the estimation of observed protein concentrations using the Hi3 approach (Silva, Gorenstein et al. 2006).

4.2.3.1 LC-MS^E configuration

All nanoscale liquid chromatographic separations were performed using a directly-coupled NanoAcquity UPLC system and a nanoelectrospray source (Waters Corporation, Milford, MA, USA). The system was composed of a binary solvent, auxiliary solvent and sample manager fitted with a heating and trapping module.

LC separations were performed using a Symmetry C18 trapping column (180 μm x 20 mm 5 μm) and a BEH C18 analytical column (75 μm x 250 mm 1.7 μm). The composition of solvent A was 0.1% v/v aqueous formic acid and solvent B 0.1% v/v formic acid in acetonitrile.

An aliquot of each sample containing internal standard was applied to the trapping column and flushed with 0.1% solvent B for 2 min at a flow rate of 15 μL min^{-1}. Sample elution was performed at a flow rate of 250 nL min^{-1} by increasing the organic solvent concentration from 3 to 40% B over 90 min, with a total run time of 115 min. The mass spectrometer was fitted with a universal nanoflow sprayer (Waters Corporation, Milford, MA, USA) using an applied capillary voltage of 3.5 kV. All analyses were conducted in technical triplicate.

Prior to and after each set of technical replicates, a quality control (QC) injection of 50 fmol PhosB was analysed using LC-MS^E and the data processed by PLGS. Where the peptide sequence coverage fell below 35% for PhosB, no further sample data were collected and the cause of the loss of peptide identification investigated

and resolved. A minimum of four QCs were collected every 24 hr during sample data collection.

MS^E data acquisition was performed on a Synapt HDMS instrument (Waters Corporation, Milford, MA, USA), configured for MS^E through the MS Method Editor controlled by MassLynx v4.1. The time-of-flight analyser of the mass spectrometer was externally calibrated using the MS/MS spectrum obtained from the doubly charged precursor of the GFP peptide over the range *m/z* 50 to 1300. The calibration was manually validated with an average ppm error across the mass range <10 ppm being obtained. GFP was used for lockmass correction (*m/*z 785.8426) infused via a NanoLockSpray interface at a constant rate of 500 nL min^{-1} at 500 fmol μL^{-1} and sampled every 60 seconds.

In low energy MS mode, data were collected at constant trap collision energy of 6 eV. In elevated energy MS mode, the trap collision energy was ramped from 15 V to 30 V whilst the transfer collision energy was held at 3 V and 10 V for low and elevated conditions respectively. The trap vacuum was held at a constant 8.2 mbar. All subsequent data were post-acquisition lockmass-corrected using the monoisotopic ion of the doubly charged precursor of GFP (*m/z* 785.8426).

4.2.4 Processing of MS^E acquired data

MS^E data were processed using PLGS v2.4 and lockspray calibrated against GFP using data collected from the reference line during acquisition. The ion detection, clustering and protein identification has been explained in detail in Section 1.4.3.6. In brief, lockmass-corrected spectra are centroided, deisotoped and charge-state-reduced to produce a single accurately mass measured monoisotopic mass for each peptide and its associated fragment ions. Initial correlation of a precursor and potential fragment ions is achieved using time alignment (Geromanos, Vissers et al. 2009). Data processing parameters specified 250, 100 and 1500 for the low, elevated and intensity threshold values respectively in PLGS. These values were recommended by the manufacturer as appropriate for this release of PLGS.

4.2.5 Database interrogation using MS^E data

The PLGS processed data were used to interrogate the IPI human database rel. 3.69 downloaded from (http://www.ebi.ac.uk/IPI/IPIhuman.html), randomised once to form a concatenated database of genuine and random entries and appended with the sequences for porcine trypsin and rabbit glycogen phosphorylase (P00761 and P00489 - http://www.uniprot.org/).

For MS^E data, database search parameters included a fixed modification of carbamidomethyl cysteine, one missed trypsin cleavage site with variable modifications of acetyl N-terminus, oxidation of methionine and deamidation of asparagine and glutamine.

Precursor and fragment ion tolerances were determined automatically by PLGS. Protein identification criteria included the detection of at least three fragment ions per peptide, seven fragment ions per protein and at least one peptide per protein with a 4% false discovery rate (FDR). PhosB was specified as the internal standard and the concentration specified (in fmol) in the PLGS workflow template to allow the Hi3 estimation of protein concentration.

Sample loading onto the LC-MS^E system was optimised at approximately 600 ng on-column for each injection to ensure that quantitative measurements obtained were within the dynamic range of both the chromatography column and the Synapt HDMS detector.

Protein identifications obtained from each of the tryptic digests, analysed in triplicate were exported from the PLGS browser into Microsoft Office Excel. Within Excel, the protein identification results were then further filtered for replication. A simple filter would require a protein to be observed in more than one analysis out of the 87 collected in this work. More stringent filtering would require the protein to be observed in ≥2 replicates *per sample*. The latter would generate a list of proteins identified from each individual maternal plasma sample with a significantly higher degree of confidence but would potentially reject identification of proteins that may have been observed more confidently in other samples.

4.3 Results and discussion

4.3.1 Automated IgY-14 depletion of maternal plasma

Thirty maternal plasma samples, from three obstetric conditions, were depleted of 14 highly abundant proteins using the Seppro IgY-14 LC2 format chromatography column. Two aliquots of each plasma sample were depleted, combined and then concentrated into a single sample for tryptic digestion and subsequent LC-MS^E analysis.

A typical chromatogram obtained from the separation (absorbance at 280 nm) is shown by the black trace in Figure 4. 4. The depleted proteins elute from the column prior to the highly abundant proteins which elute from 24 min onwards.

4.3.2 LC-MS^E analysis and protein identification of IgY-14 depleted plasma

For each of the thirty IgY-14 LC2 depleted plasma samples from three obstetric conditions, an initial aliquot was analysed using LC-MS^E on a Synapt HDMS instrument. Each aliquot contained a known amount of the internal standard (PhosB typically at 200 fmol on-column) allowing the total identified protein concentration to be estimated using the Hi3 approach. Further samples were prepared and analysed to ensure that a similar protein loading and number of identifications were obtained for the entire data set. Quantitative data were not obtained from one of the T21 samples due to a shortage of material and so in total 87 data files were collected for comparison in PLGS software.

A total of 9,806 protein observations from the IPI database were reported by PLGS from 87 data files, equating to 1,265 non-redundant protein identifications. Of these, 894 (70.7%) were reported only once by PLGS and 74 (5.8%) in just two data files. Proteins that were identified in the majority of the data files (≥80) totalled 45 (3.6%). For identifications observed in >1 data file, the frequency of observation has been plotted and is shown in Figure 4. 5**.**

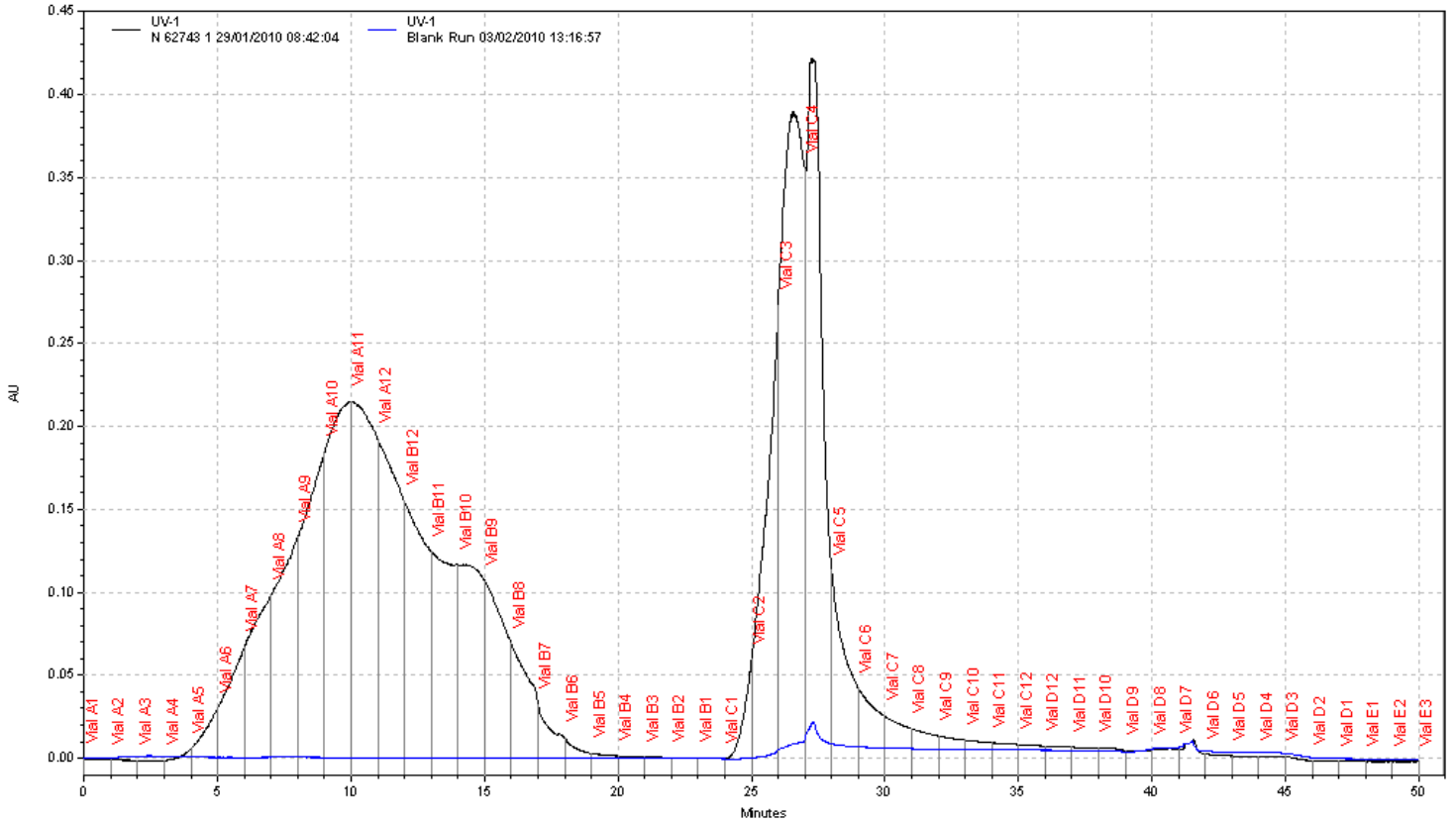


Figure 4. 4: Chromatograms obtained from automated IgY-14 LC2 chromatographic depletion.
The black trace indicates absorbance at 280 nm from a maternal plasma sample, whereas the blue trace was obtained from a blank injection of dilution buffer. Fractions were collected every minute and their positions in the collection tray are indicated in red. Fractions collected between 5 and 18 minutes were combined and tryptically digested from each sample.

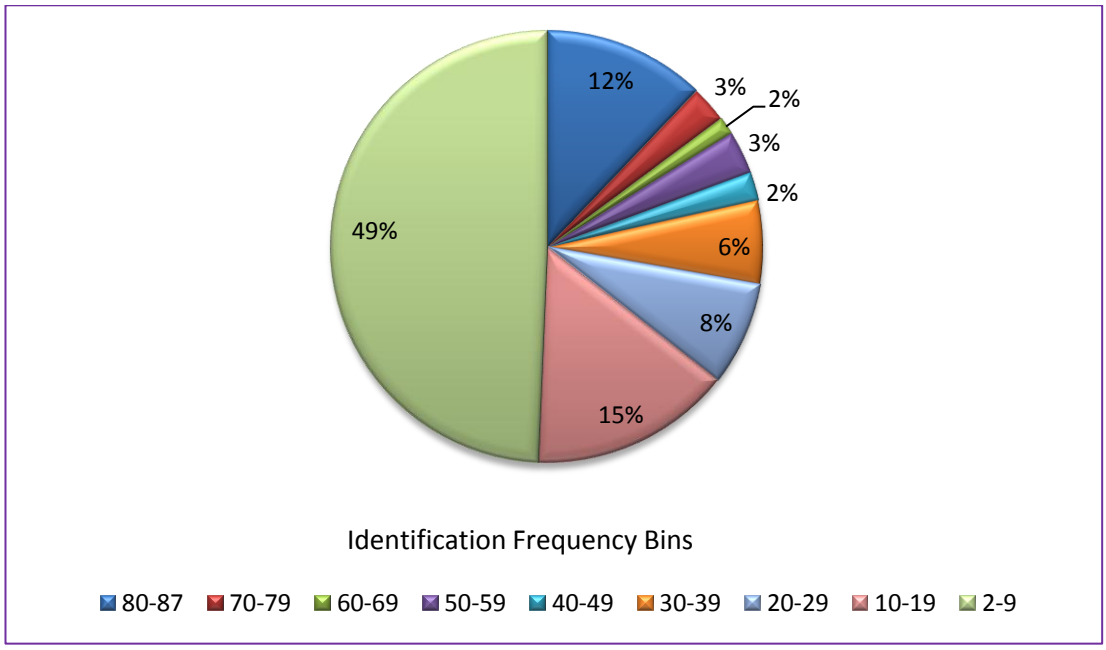


Figure 4. 5: Frequency of protein identification.
The frequency of identification of proteins from IgY-14 LC2 depleted plasma (replication >1) based on 87 data files, 29 samples analysed in technical triplicate.

If no additional filtering of the PLGS results was performed, an average of 9,422 peptides were identified in the 87 LC-MS^{E} analyses with an average 7.5 peptides/protein. As the stringency of filtering was increased to restrict the identifications reported to those observed in at least 1, 2 , 3 or more *analyses* the average number of peptides identified per analysis decreased to 4,360, 3,759 and 3,607 respectively. These results are shown in Figure 4. 6. The average number of peptides/protein increased to 11.8, 12.7 and 13.0 respectively, shown in Figure 4. 7. Average sequence coverage increased from 21.3% with no filtering to 28.3%, 29.8% and 30.1% with increased replication filtering, shown in Figure 4. 8.

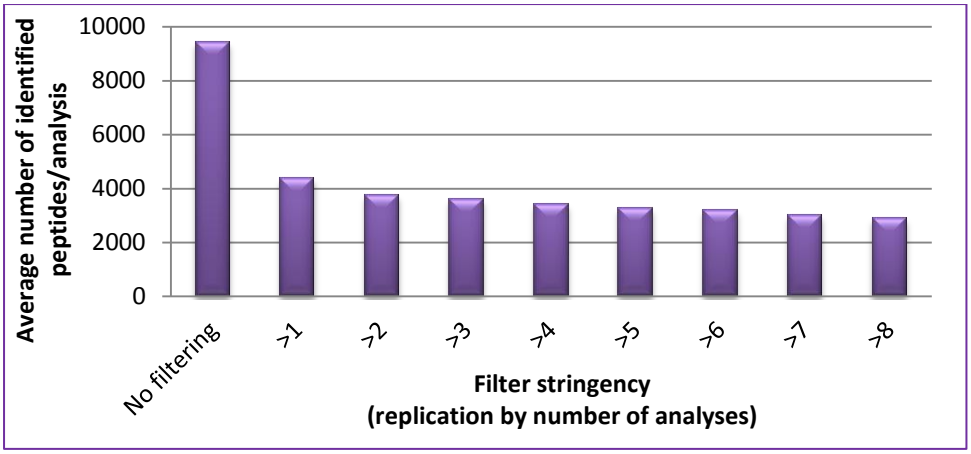


Figure 4. 6: Effect of filtering on the average number of peptides identified in each LC-MS^{E} analysis.

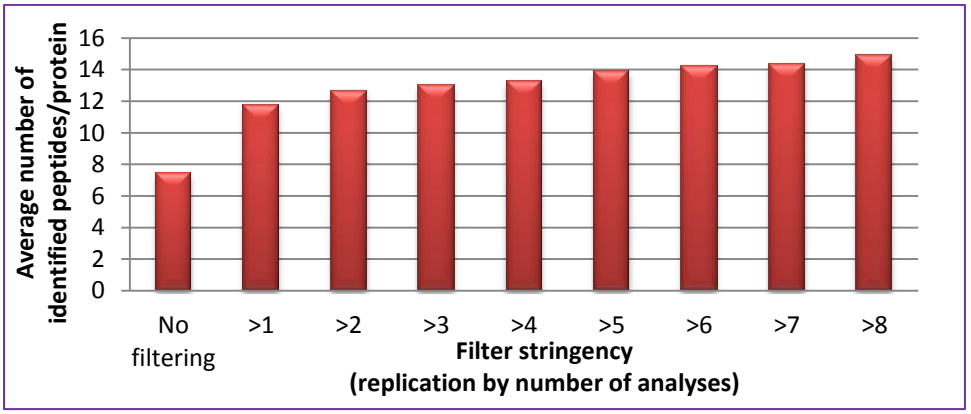


Figure 4. 7: Effect of filtering on the average number of peptides identified from each protein in each LC-MS^{E} analysis.

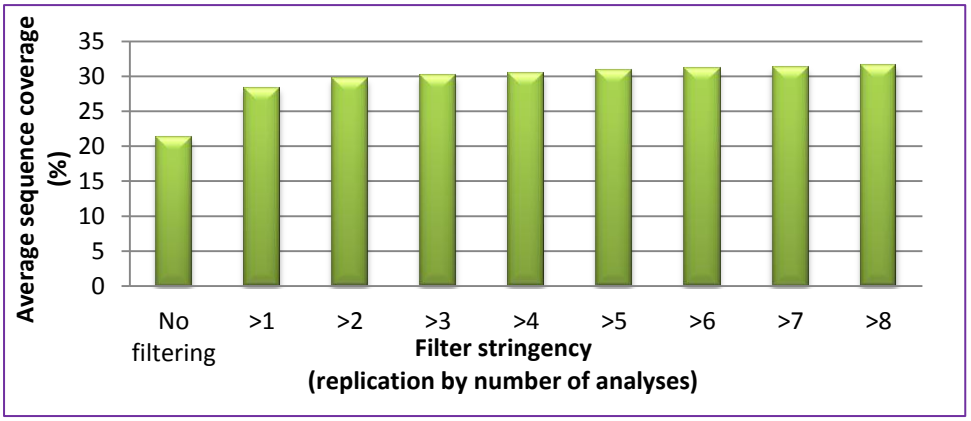


Figure 4. 8: Effect of filtering on the average sequence coverage from each LC-MS^{E} analysis.

For quantitative results obtained from the IgY-14 LC2 depleted plasma, stringent filtering rules were applied to the data i.e. a protein had to be identified in >1 technical *replicate* per sample to be included, even though many of the proteins rejected by this filtering may be valid identifications. Some may have been present at or near the limit of detection and/or alternate isoforms of the protein may have been identified in different replicate analyses. The average number of proteins identified from each IgY-14 LC2 depleted maternal plasma was 113 with an average loading of 617.4 ng on-column. For the three obstetric conditions, normal, PET and T21 the average number of proteins and loading were 112, 115, 112 and 630.8 ng, 645.3 ng and 571.5 ng respectively**,**

Figure 4. 9**.**

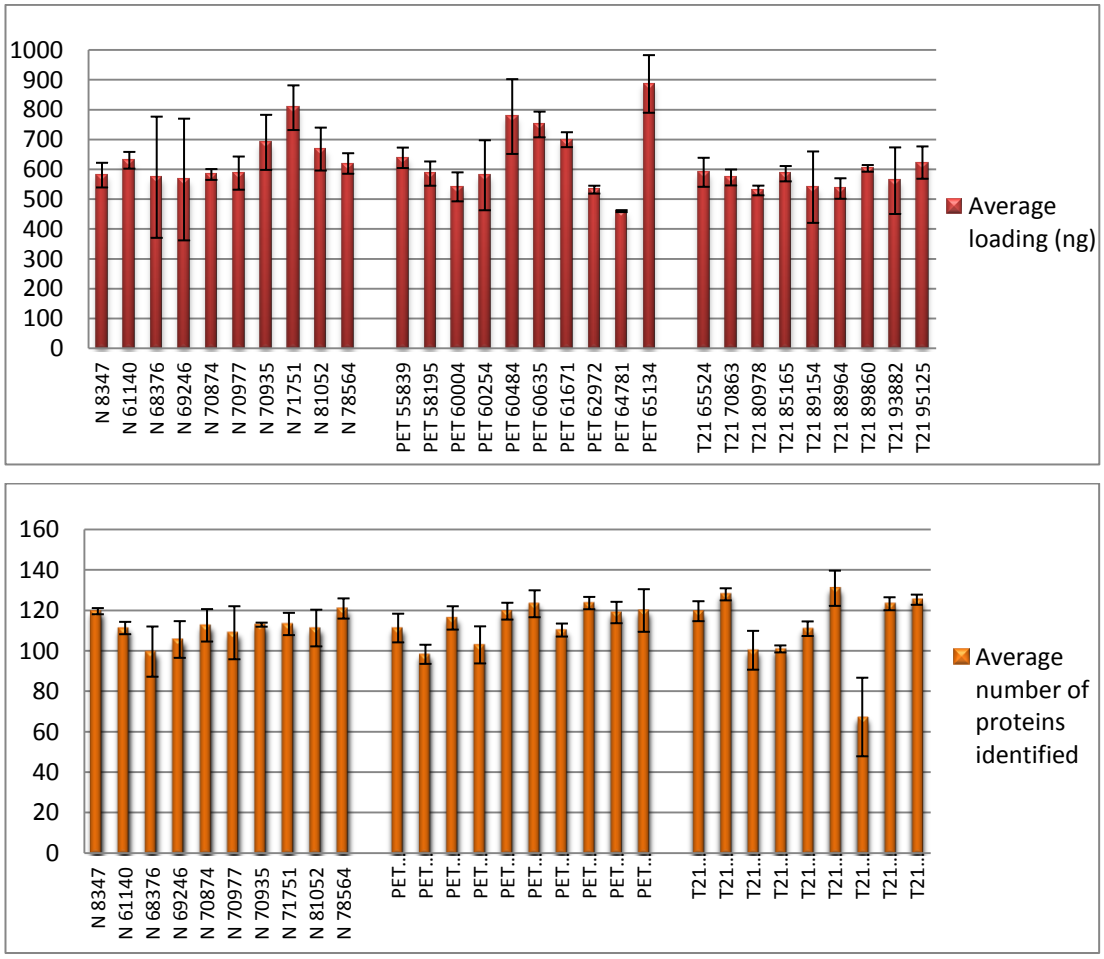


Figure 4. 9: Average sample loading and number of proteins identified from tryptically digested IgY-14 LC2 depleted plasma for quantitative analysis.

Based on 3 technical replicates per sample using the Hi3 approach for protein concentration, error bars indicate standard deviation.

4.3.3 Quantification of proteins identified from IgY-14 depleted plasma

PLGS output included an estimate of the amount of each identified protein in both fmol and ng levels using the Hi3 approach. The abundance of these proteins was converted to a percentage of the total identified protein (in ng) in Excel and then filtered for replication >1 per sample, to improve the confidence in the quantitative results and the average abundance for each sample and obstetric condition determined.

For a number of proteins, such as prothrombin and vitamin D binding protein, the average abundance levels observed between samples were relatively consistent, Figure 4. 10 and Figure 4. 11**.** For prothrombin protein, the standard deviation in the measurements was 0.27, 0.3 and 0.27 for normal, PET and T21 obstetric groups. Similarly for vitamin D binding protein, the standard deviation was 0.6, 0.5 and 0.7 respectively.

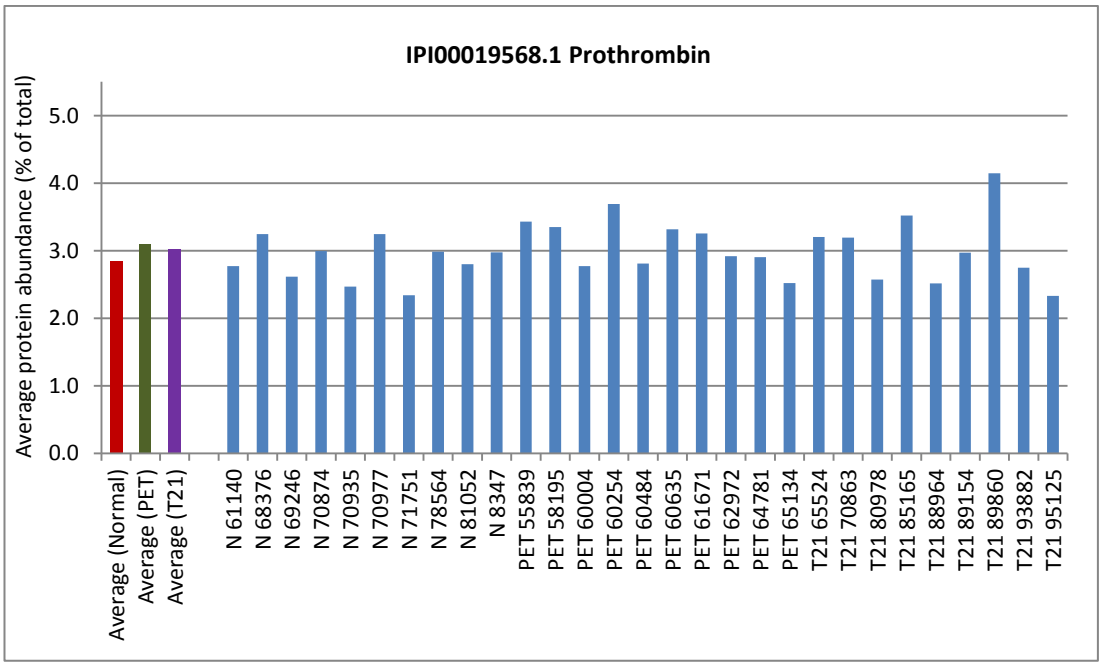


Figure 4. 10: Abundance of prothrombin in IgY-14 LC2 depleted plasma.
Abundance calculated as a percentage of the total identified protein content based on ng levels reported by the Hi3 approach in PLGS.

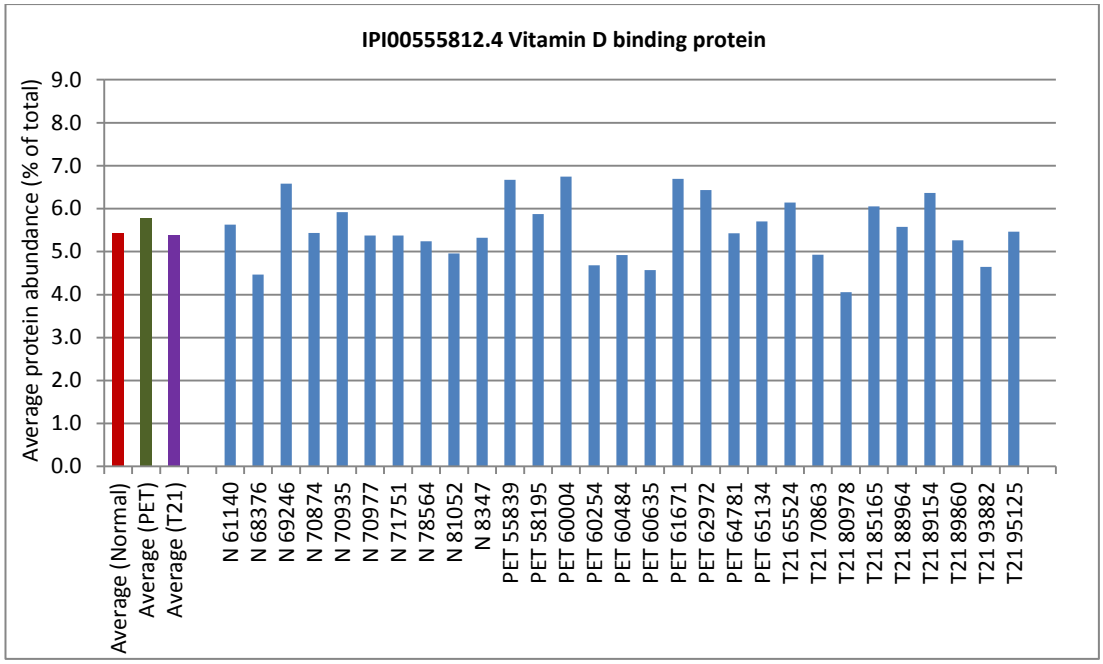


Figure 4. 11: Abundance of vitamin D binding protein in IgY-14 LC2 depleted plasma. Abundance calculated as a percentage of the total identified protein content based on ng levels reported by the Hi3 approach in PLGS.

For other proteins, such as ceruloplasmin, the inter-sample variation observed was more pronounced but still independent of obstetric condition. Average protein abundance varied from 1.56% to 10.33% between the samples but average levels for each obstetric group were very similar at 6.71%, 7.37% and 6.92% for normal, pre-eclampsia and trisomy 21 respectively, Figure 4. 12.

Average abundance of Complement C3, depleted during sample processing, varied greatly between samples (from 0.29% to 12.85%). Four of the pre-eclampsia samples exhibited high levels of complement C3 at an average of 9.18% whereas the other 6 were observed at an average of 0.57%. This is reflected in the average abundance for the protein for the PET obstetric condition at 4.01% compared to 2.05% and 1.85% for the normal and T21 samples respectively**,** Figure 4. 13. Had the samples from the three obstetric conditions been pooled, this protein could have been identified as a biomarker for PET with a 1.96-fold increase in concentration compared to the normal pooled sample and 2.17-fold with respect to T21. Complement C3 was also identified in one normal (12.85%) and one T21 sample

(9.3%) at elevated levels, suggesting that this protein is not specific to the PET condition.

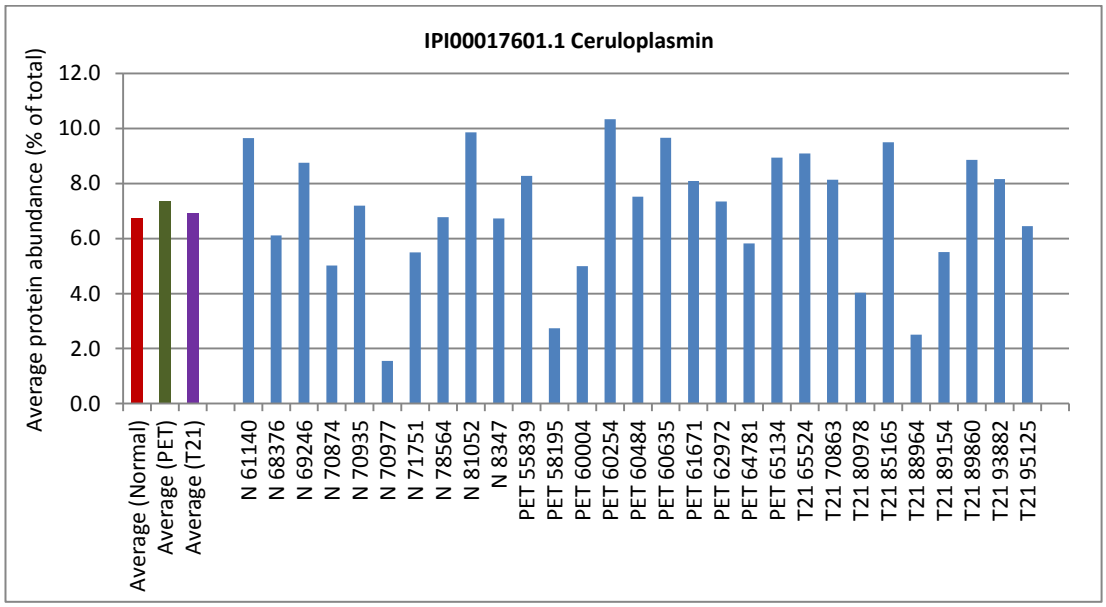


Figure 4. 12: Abundance of ceruloplasmin in IgY-14 LC2 depleted plasma.
Abundance calculated as a percentage of the total identified protein content based on ng levels reported by the Hi3 approach in PLGS.

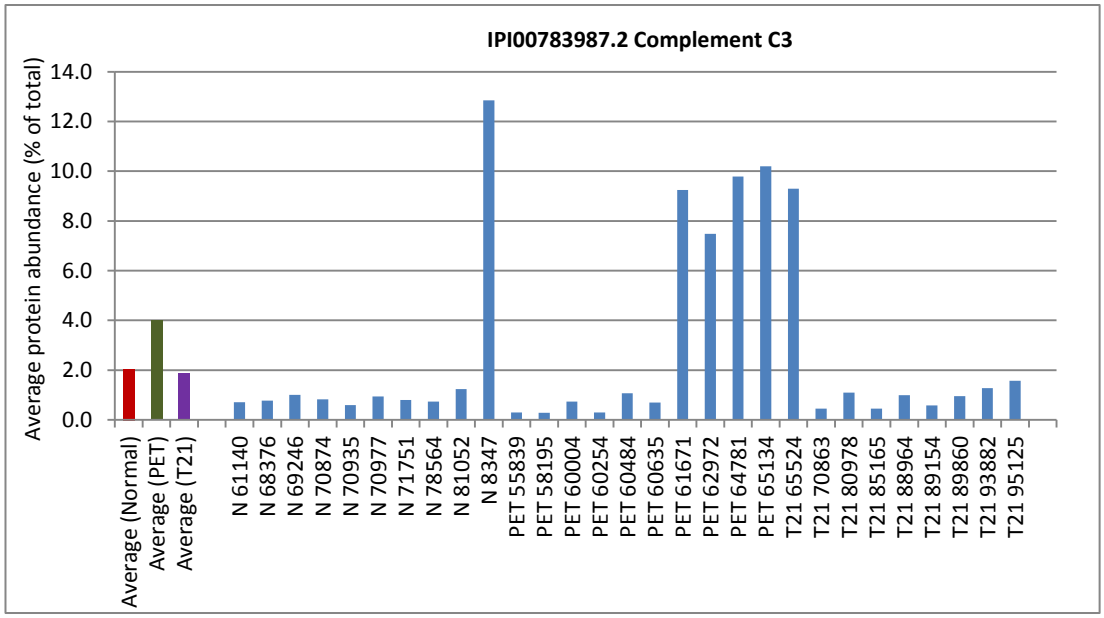


Figure 4. 13: Abundance of complement C3 in IgY-14 LC2 depleted plasma.
Abundance calculated as a percentage of the total identified protein content based on ng levels reported by the Hi3 approach in PLGS.

4.3.4 Quantification of depleted proteins from IgY-14 LC2 partitioned plasma

An advantage of the LC-MS^E approach for protein identification and quantification of depleted or fractionated plasma proteins was the ability to monitor the efficiency of the depletion process during the lifetime of the study. The plasma samples where possible, were depleted in a rotating cycle of obstetric groups, Appendix D.

The average abundance of each of the proteins depleted by the IgY-14 approach from each sample was determined, Figure 4. 14**.** The level of albumin in the partitioned samples fell significantly to an average abundance of 3.48%, 2.08% and 3.39% for the three obstetric groups, normal, PET and T21. Serotransferrin was reduced to an average abundance of 1.05%, 1.44% and 1.00% for normal, PET and T21 samples compared to 8.89% in normal undepleted plasma analysed using LC-MS^E. Haptoglobulin fell from 6.58% (undepleted) to 0.3%, 0.43% and 0.22% respectively.

Complement C3 was described in the previous section 4.3.4, as a protein which was depleted by the IgY-14 system but identified at varying levels in some of the samples. This could be due to inefficient depletion as a result of column aging or higher levels of the protein being present in the raw plasma prior to depletion thus exceeding the maximum loading capacity for the column. No correlation was observed between the levels of complement C3 in depleted plasma and the age of the column. No correlation was observed between the depletion efficiency for the highly abundant proteins (immunoglobulins excluded) and column age during this study, which involved >200 rounds of partitioning (personal communication Edmond Wilkes). The hypothesis that some of the abundant proteins were present at high levels in raw plasma could be investigated using the LC-MS^E approach, by analysing the residual original plasma supplied, possibly indicating that the maximum loading for the specific protein exceeded the binding capacity.

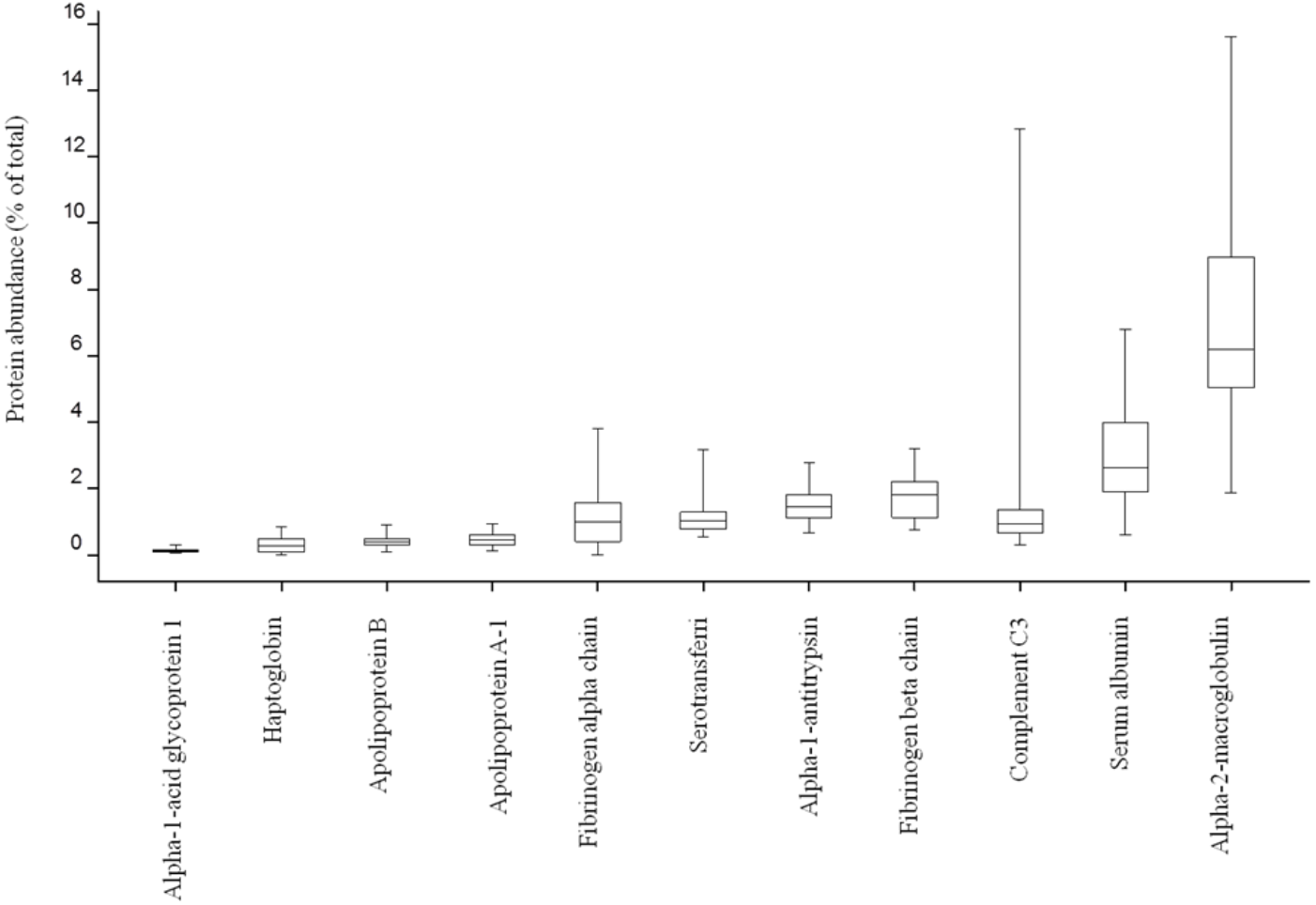


Figure 4. 14: Depleted protein abundance following IgY-14 LC2 chromatography.
The bottom and top of each box represents the 25th and the 75th percentile respectively; the line represents the median value. Whiskers extend to the most extreme data points.
Abundance calculated as a percentage of the total identified protein content based on ng levels reported by the Hi3 approach in PLGS.

4.3.5 Identification of trisomy 21 predictive proteins

Proteins characterised as part of the antenatal screening for Down's syndrome include pappalysin-1 (Pregnancy-associated plasma protein A, PAPP-A), human choriogonadotropin subunit beta (β-hCG), inhibin A, alpha fetoprotein (AFP) as well as the hormone unconjugated oestriol (uE_3). Protein identification results from IgY-14 depleted plasma encompassing the three obstetric conditions were reviewed for the identification of these proteins.

4.3.5.1 Pappalysin-1

PAPP-A was not identified in any of the 87 LC-MS^E data files generated from the 29 samples. The median level of serum PAPP-A increases during the first and second trimester, from approximately 0.7 IU L^{-1} at 9 weeks to 10 IU L^{-1} at 21 weeks in normal pregnancies, Figure 4. 15 In comparison, the level of PAPP-A in males and non-pregnant females would be 14 mIU L^{-1}.

In trisomy 21 pregnancies the median PAPP-A level is 0.5 multiple of the median (MoM) i.e. 50% of the level expected in an unaffected normal pregnancy. A reduced level of PAPP-A has also been suggested for use as a marker for PET especially in combination with other tests such as uterine artery Doppler ultrasound imaging (Yu, Khouri et al. 2008; Poon, Maiz et al. 2009), Figure 4. 16**.** A screening method using a combination of maternal history, mean arterial pressure, uterine artery pulsatility index, serum PAPP-A and PlGF levels at 11 to 13 weeks gestation was able to identify 93.1% of cases of early onset PET with a FDR of 5% (Poon, Kametas et al. 2009). The absence of identification of PAPP-A in PET and T21 conditions was not unsurprising given that it was not observed in the normal samples.

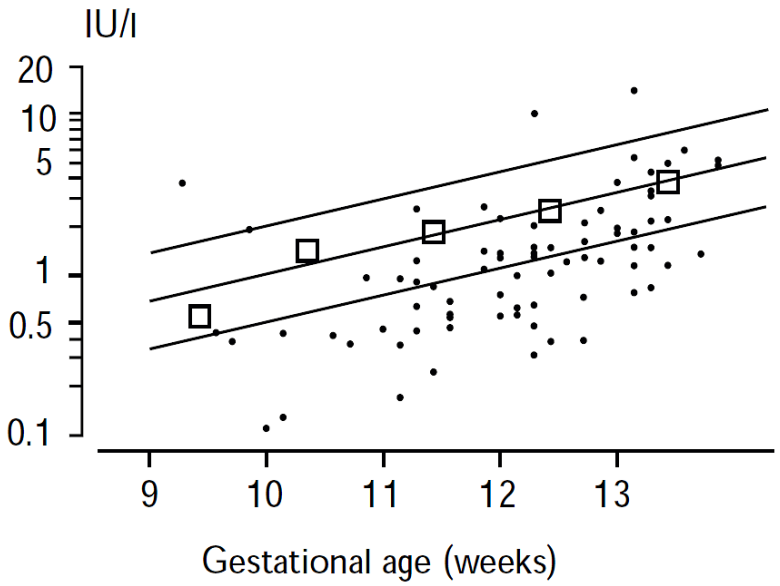


Figure 4. 15: Relationship between gestational age and plasma PAPP-A level.
Open squares are median controls, with the three centiles corresponding to 2.0, 1.0 and 0.5 MoM respectively for unaffected pregnancies, taken from (Wald, Rodeck et al. 2003).

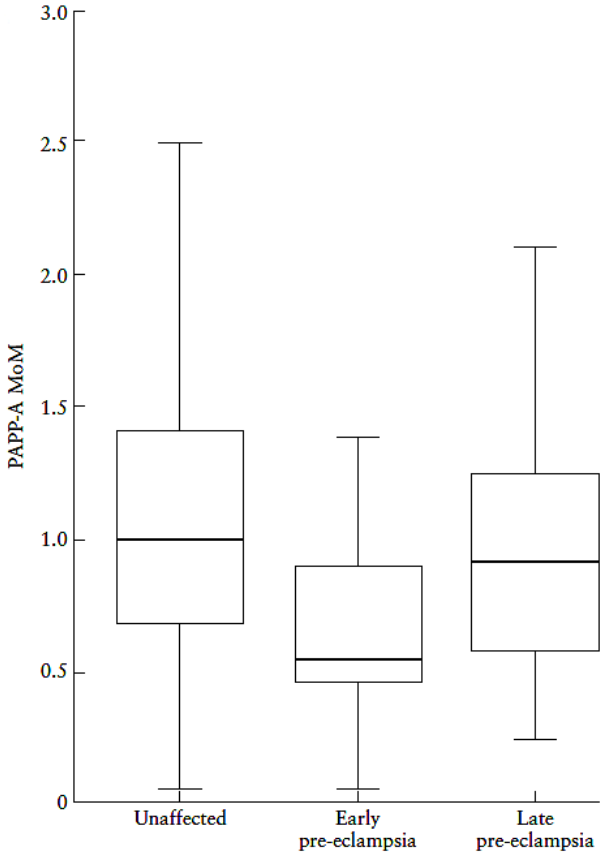


Figure 4. 16: Pregnancy-associated plasma protein-A levels.
MoM multiple of the median; median, interquartile range and range are shown with levels determined at 11 + 0 to 13 + 6 weeks, taken from (Poon, Maiz et al. 2009)

The concentration of PAPP-A increases up to 150-fold during pregnancy and for the samples taken for this study (median 88 days gestational age) the median protein concentration in an unaffected pregnancy would be approximately 3.2 IU L^{-1}. Using the conversion factor of 1 IU L^{-1} = 4.5 mg L^{-1} and an estimated molecular weight of 700 kDa, this would equate to approximately 41 fmol μL^{-1} in the tryptic digest of PAPP-A or 5 fmol on column in each LC-MS^E analysis of IgY-14 depleted plasma from a normal sample, approaching the limit of detection for these types of experiments. This assumes that there were no losses during sample preparation. Nagalla and co-workers adopted a number of proteomic strategies to identify novel biomarkers for Down's syndrome from first and second trimester MARS-6 depleted plasma including 2D DIGE, 2D-LC-chromatofocusing (CF), MudPIT (SCX-RP-MS/MS) and MALDI-MS (Nagalla, Canick et al. 2007). Their study identified proteins from three major functional groups: protease inhibitors, acute-phase response proteins and serum carrier proteins. None of the four current proteins used in antenatal screening were identified in any of their approaches including the 2D-LC-CF approach which utilised 5-7 mg of depleted plasma, an excess of 8300-fold compared to this study.

4.3.5.2 Choriogonadotropin subunit beta

Free choriogonadotropin subunit beta protein (β-hCG) levels in plasma peak at around 9 weeks into pregnancy and reduce over subsequent weeks. In first trimester trisomy 21 cases the measured median β-hCG level is approximately 2.0 MoM compared to unaffected pregnancies. β-hCG levels in urine form the basis of commercially available home pregnancy test kits.

A number of isoforms of β-hCG were present in the database interrogated and two of these variants of β-hCG were identified from the 87 data files in this study. P01233-1 (IPI00000870.1) is a 165 amino acid-containing protein including a 20 residue signal peptide whereas Q6NT52-1 (IPI00746788.8) is a 195 residue protein including a 50 residue signal peptide. Closer examination of the mature protein sequences indicated that the two had 100% identity and so are treated here as the same protein, Figure 4. 17.

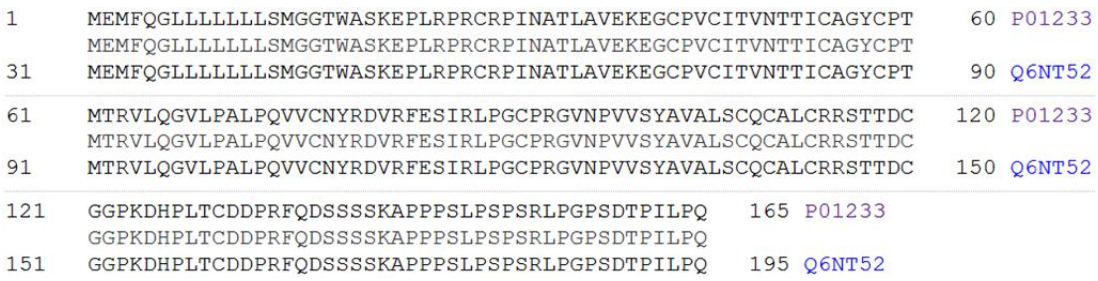


Figure 4. 17: BLAST alignment of β-hCG variants
Choriogonadotropin subunit beta variant 2 (Q6NT52-1) has 100% identity with choriogonadotropin subunit beta (P01233-1) from residue 31. The mature processed product is 145 residues in length.

P01233-1 was identified in 39 replicates with an average 2.5 peptides and Q6NT52-1 in 5 replicates with 4.4 peptides (50.6% of the total data set). The average sequence coverage was 25.6% and 27.7% respectively. Presumably due to the low numbers of peptides identified as belonging to β-hCG, no measure of protein abundance was available from these experiments indicating that the protein concentration was near to the limit of detection. For the samples taken for this study (median 88 days gestational age) the median protein concentration in an unaffected pregnancy would expected to be approximately 34 IU L^{-1} or 34 ng mL^{-1} (using the conversion factor 1 IU L^{-1} = 1 ng mL^{-1}), Table 4. 3. Assuming a molecular weight of 15.5 kDa for processed β-hCG, this equates to 68 pg in the tryptic digest (4.5 fmol μL^{-1}) or <0.6 fmol on column for each analysis of an unaffected pregnancy sample.

With elevated β-hCG levels in trisomy 21 cases, it would be expected that it would be identified (if not quantified) from more replicate injections from T21 than normal samples, assuming the concentration was on the limit of MS identification. This was observed, β-hCG was identified in 9 out of 30 analyses from normal cases (30%) and 14 out of 27 in T21 (52%).

Gestational Age (completed weeks)	Median β-hCG (IU L^{-1}) unaffected pregnancies	Median β-hCG (IU L^{-1}) trisomy 21 pregnancies
11	44.83	
12	39.33	
13	30.32	
14	22.29	71.4
15	15.0	37.4
16	12.2	31.3
17	9.1	25.4
18	8.4	14.0

Table 4. 3: β-hCG levels in first and second trimester serum
Values based on a continuous study of 5,000 pregnant women by the Fetal Medicine Foundation (UK) confirmed by 35,000 additional samples using the B·R·A·H·M·S* KRYPTOR free β-hCG assay.

The highest frequency of β-hCG observation was from the PET samples with the protein identified in 70% of the analyses (21 out of 30). Di Lorenzo and co-workers identified β-hCG as strongly associated with early onset PET and suggested its use in combination with other markers (placental growth factor and chronic hypertension) in first trimester screening (Di Lorenzo, Ceccarello et al. 2012).

4.3.5.3 Inhibin A

Inhibin A is a dimeric glycoprotein consisting of an alpha (P05111, IPI00007080.1) and beta-A (P08476, IPI00028670.1) chain. Serum concentrations of inhibin A increase in normal pregnancies to a level of about 550 pg mL^{-1} at 8-9 weeks of gestation, which decreases to a plateau of 180 pg mL^{-1} at 15 weeks. In Down's syndrome, elevated serum levels of inhibin A with a median value of 2.06 MoM compared to normal pregnancies have been observed (Aitken, Wallace et al. 1996).

In this study inhibin A was not identified in any of the analyses. This is not surprising since based on a 550 pg mL^{-1} concentration of the protein in serum and a

molecular weight of the dimeric form of the protein of ~86 kDa, the estimated level of inhibin A in each tryptic digest would have been 55 pg equating to a concentration of 12.8 amol μL^{-1} or (<2 amol on column).

4.3.5.4 Alpha-fetoprotein

Alpha-fetoprotein (AFP - P02771, IPI00022443.1) is a 68 kDa single chain glycoprotein of the albuminoid superfamily which includes albumin and vitamin D binding protein. Altered levels of AFP have been associated with a number of cancers, particularly of the liver.

Maternal serum levels of AFP continue to increase during pregnancy until the third trimester when levels begin to decrease (Mizejewski 2003). In trisomy 21 pregnancies the serum levels of AFP were found to be reduced compared to unaffected cases with a median 0.67 MoM (Johnson, Cowchock et al. 1991), Figure 4. 18.

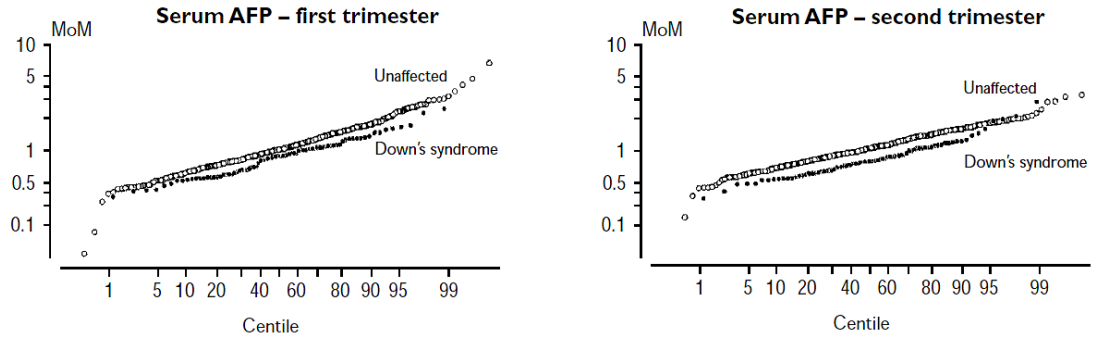


Figure 4. 18: Probability plots for serum α-fetoprotein in first and second trimester. Extracted from (Wald, Rodeck et al. 2003)

Typical median levels of AFP in unaffected pregnancies at 88 days gestational age would be expected to be 20.5 ng mL^{-1} (~300 amol μL^{-1} in plasma) compared to approximately 14 ng mL^{-1} in T21, equating to 75 amol and 51 amol on-column in this study for normal and T21 outcomes. AFP was not identified from any of the 87 analyses of depleted plasma from the three obstetric conditions.

Elevated levels of AFP have been associated with an increase of the likelihood of pre-eclampsia and in particular early onset preeclampsia (Olsen, Woelkers et al. 2012), based on a study of 7,110 pregnant women, Figure 4. 19. The study concluded similar results for both β-hCG and inhibin A, as well as AFP where >2 MoM indicated a higher risk.

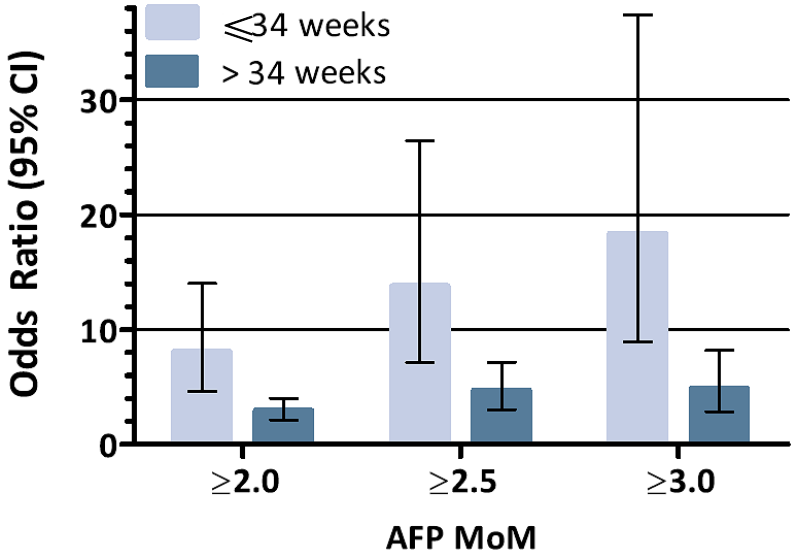


Figure 4. 19: Odds ratio for early and late onset pre-eclampsia in women with elevated α-fetoprotein.
Extracted from (Olsen, Woelkers et al. 2012), the higher the MoM the greater the likelihood of pre-eclampsia particularly early onset. Light blue bars indicate delivery before 34 weeks, dark blue after 34 weeks.
Key: AFP - α-fetoprotein, MoM – multiple of the median

4.3.6 Identification of literature-based obstetric biomarker proteins from IgY-14 LC2 partitioned plasma

Nagalla and co-workers adopted a number of proteomics methodologies for the comprehensive analysis of maternal plasma depleted of 6 highly abundant proteins in their endeavour to identify biomarkers for Down's syndrome (Nagalla, Canick et al. 2007). 2D DIGE, 2D-LC-chromatofocusing (CF), MudPIT (SCX-RP-MS/MS) and MALDI-MS were employed to analyse first and second trimester depleted plasma.

2D DIGE analysis visualised 1,816 and 1,842 proteins spots from the first and second trimester gel sets. Comparative analysis indicated 28 and 26 spots respectively with differential abundance >1.5-fold change ($p<0.05$) from which 19 and 16 proteins were identified by MS/MS. In some of the spots, multiple proteins were identified and peptide counts were used to determine the significant protein.

2D-CF identified 95 and 80 absorbance bands where differential intensities ranged from 0.004 to 0.638 AU, which translated into 25 and 6 proteins characterised by MS in first and second trimester samples respectively, generating complementary results to those obtained from 2D DIGE. The SCX-RP analysis identified 9 proteins using spectral counting (from one hundred selected based on high-confidence identification, peptides ≥2) with trends similar to that observed in 2D DIGE and 2D-CF experiments.

A number of proteins identified by Nagalla were highly abundant in plasma such as apolipoproteins involved in lipid metabolism, complement factors and α-1-acid glycoprotein which are depleted by an IgY-14 strategy. Table 4. 4 summarises the observed fold-change reported in their study by the various approaches in both trimesters.

Abundance levels obtained from this study were plotted for each of proteins listed in Table 4. 4 (all replicates from three obstetric conditions, based on percentage of total identified protein in ng).

Two proteins listed in Table 4. 4 were not detected in this study (serum amyloid A and transthyretin) and two were detected but no abundance levels were generated (clusterin and pregnancy-specific β-1 glycoprotein 1).

Afamin was quantified by 2D DIGE as almost 7-fold differentially abundant in first trimester Down's syndrome plasma but the results from this study are not in agreement (Nagalla, Canick et al. 2007). The median abundance (% total in ng) for the three conditions normal, pre-eclampsia and trisomy 21 were 0.65, 0.70 and 0.68 and average abundance values were 0.74, 0.69 and 0.72 respectively, Figure 4. 20. These equate to median MoMs of 1.08 (PET/normal and 1.04 (T21/normal). It is

possible that, in the 2D DIGE analysis, multiple proteins contributed to the spot quantified using the software (Phoretix 2D Evolution, NonLinear Dynamics, UK) as increased in abundance but not identified using MS/MS.

		Fold-change first trimester (DS/control)		Fold-change second trimester (DS/control)		
Protein	**Accession**	**2D DIGE**	**2D-CF**	**2D DIGE**	**2D-CF**	**MudPIT spectral counting**
Afamin	P43652	6.84		1.89	↑	
Ceruloplasmin	P00450	1.89		2.13		
Clusterin	P10909	2.51	↑		↑	
Fibronectin	P02751					1.8
Ficolin 3	O75636	-3.67				
Gelsolin	P06396			1.53		
Histidine-rich glycoprotein	P04196		↓		↓	
Kininogen I	P01042		↑		↑	
Pigment epithelium-derived factor	P36955	4.52		1.75		
Plasminogen	P00747		↑			
Pregnancy-specific β-1 glycoprotein 1	P11464				↑	
Serum amyloid A	P02735			2.71		11.2
Sex hormone binding globulin	P04278	2.69			↑	2.1
Tetranectin	P05452		↑	4.91	↑	
Transthyretin	P02766	2.00	↑	1.83	↑	2.2

Table 4. 4: Differential abundance of first and second trimester plasma proteins in unaffected and Down's syndrome cases.
The table excludes the highly abundant proteins depleted using an IgY-14 partition, from (Nagalla, Canick et al. 2007).
Key: DS – Down's syndrome

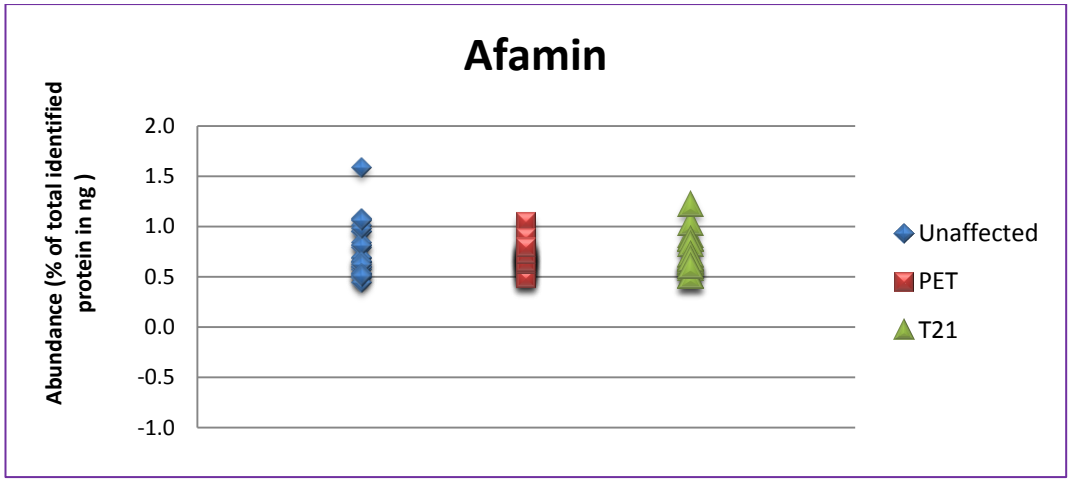


Figure 4. 20: Abundance of afamin in IgY-14 depleted maternal plasma.
Abundance calculated as a percentage of the total identified protein estimated in ng. The study by Nagalla indicated afamin was increased in abundance 6.84-fold by 2D DIGE.

Ceruloplasmin and sex hormone binding globulin were quantified at elevated levels in Down's cases by 2D DIGE at 1.9-fold and 2.69-fold, but here these observations were not replicated, Figure 4. 21 and Figure 4. 22. A greater biological variation within each obstetric group was observed for both proteins in this study and it is conceivable that this was reflected in the 2D DIGE analyses.

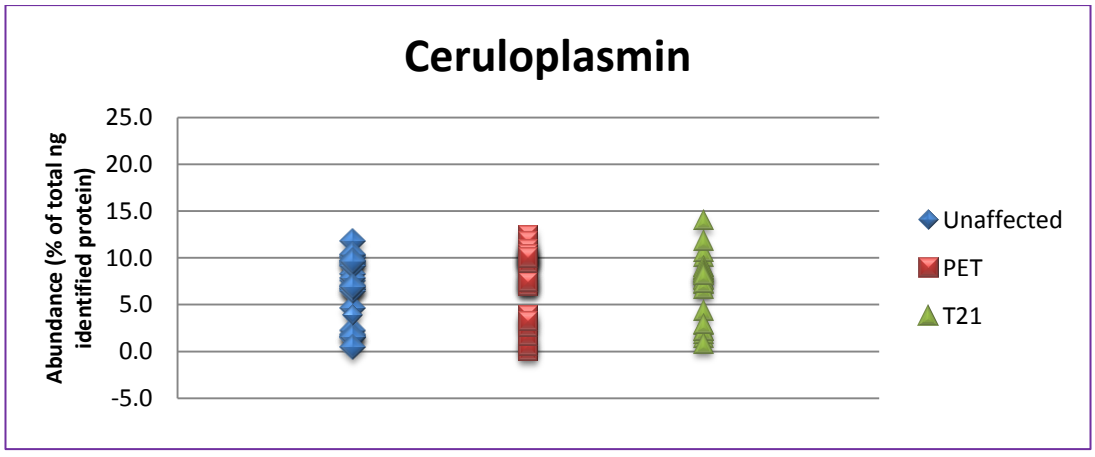


Figure 4. 21: Abundance of ceruloplasmin in IgY-14 depleted maternal plasma.
Abundance calculated as a percentage of the total identified protein estimated in ng. The study by Nagalla indicated ceruloplasmin was increased in abundance 1.9-fold by 2D DIGE.

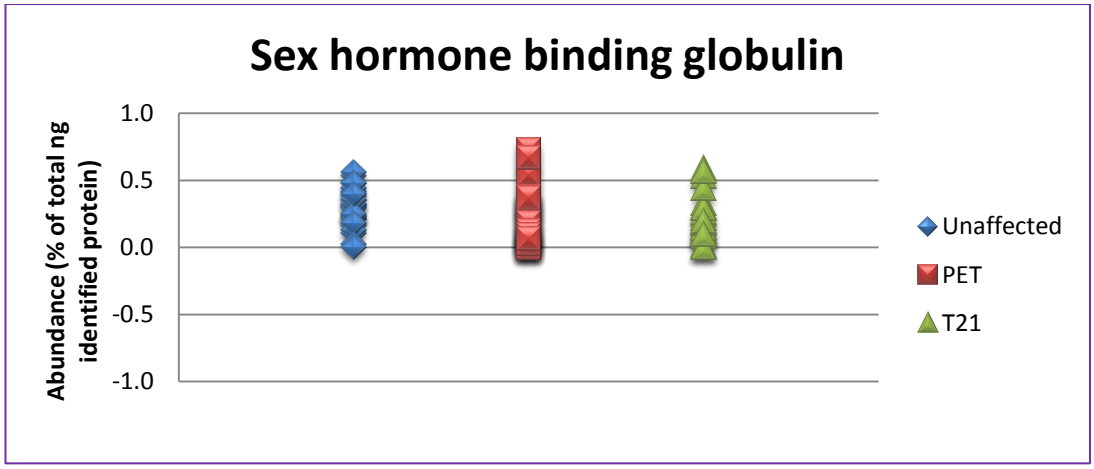


Figure 4. 22: Abundance of sex hormone binding globulin in IgY-14 depleted maternal plasma.
Abundance calculated as a percentage of the total identified protein estimated in ng. The study by Nagalla indicated sex hormone binding globulin was increased in abundance 2.69-fold by 2D DIGE.

Sex hormone binding globulin was suggested as a predictive biomarker for pre-eclampsia in first and second trimester (Wolf, Sandler et al. 2002; Carlsen, Romundstad et al. 2005; Spencer, Yu et al. 2005a), but this has subsequently been discounted (Spencer, Yu et al. 2005b; Valdés R, Lattes A et al. 2012). The results of this study also indicate that sex hormone binding globulin would not be a suitable biomarker for PET.

Histidine-rich glycoprotein was identified by Nagalla *et al.* as present at reduced levels in Down's cases based on 2D-LC-CF using UV absorbance to determine quantitative differences. In this study, the protein was observed in T21 cases with increased biological variation compared to the other two conditions. The median abundance level for the unaffected and PET cases were 0.38% and 0.39% respectively and 0.55% for T21, equivalent to median MoMs of 1.0 (PET/normal) and 1.5 (T21/ normal). The average abundance levels were 0.39%, 0.44% and 0.62% for normal, PET and T21 respectively, Figure 4. 23.

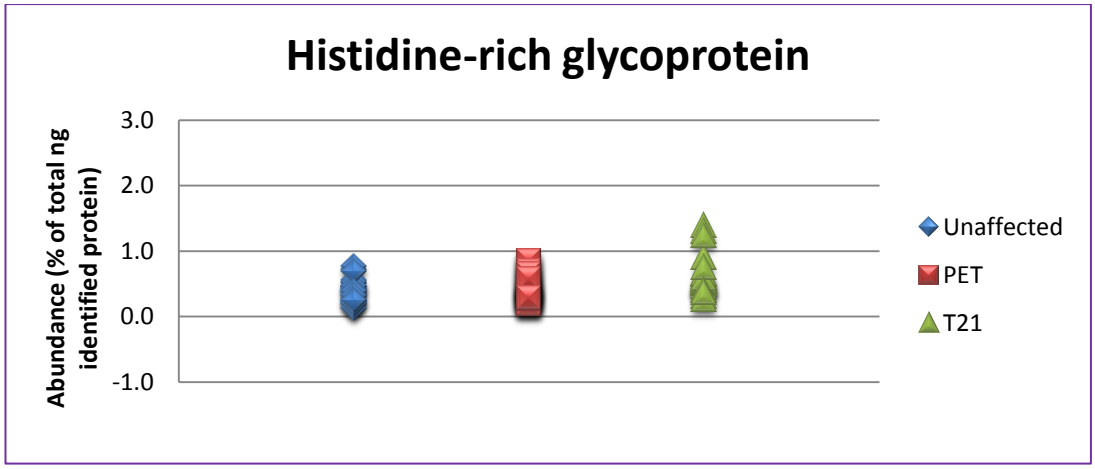


Figure 4. 23: Abundance of histidine-rich glycoprotein in IgY-14 depleted maternal plasma. Abundance calculated as a percentage of the total identified protein estimated in ng. The study by Nagalla indicated histidine-rich glycoprotein was decreased in abundance by 2D-LC- CF.

Proteins that have been detected at increased levels in pre-eclampsia and implicated in its pathogenesis include soluble FLT-1 (vascular endothelial growth factor receptor 1) and soluble endoglin. Placental protein-13 has been observed at low levels in the first trimester of PET cases as have proteins involved in the development of new blood vessels (angiogenesis) from endothelium for placental development such as vascular endothelial growth factor and placental growth factor (PIGF) reviewed by (Carty, Delles et al. 2008), Figure 4. 24. This study did not identify these proteins in IgY-14 depleted plasma. PlGF levels at 13 to 16 weeks gestation were determined to be 142 pg mL^{-1} in unaffected cases and 90 pg mL^{-1} in pregnancies where pre-eclampsia develops (Levine, Maynard et al. 2004). This is equivalent to a PIGF concentration of 6.3 amol $\mu L^{-1}$and 4.0 amol μL^{-1} in the tryptic digest for analysis. Adiponectin which has been implicated in PET pathology (Conde-Agudelo, Villar et al. 2004), was observed in only two of the analyses in this study and so no inference into its applicability as a biomarker for PET could be achieved, Figure 4. 24.

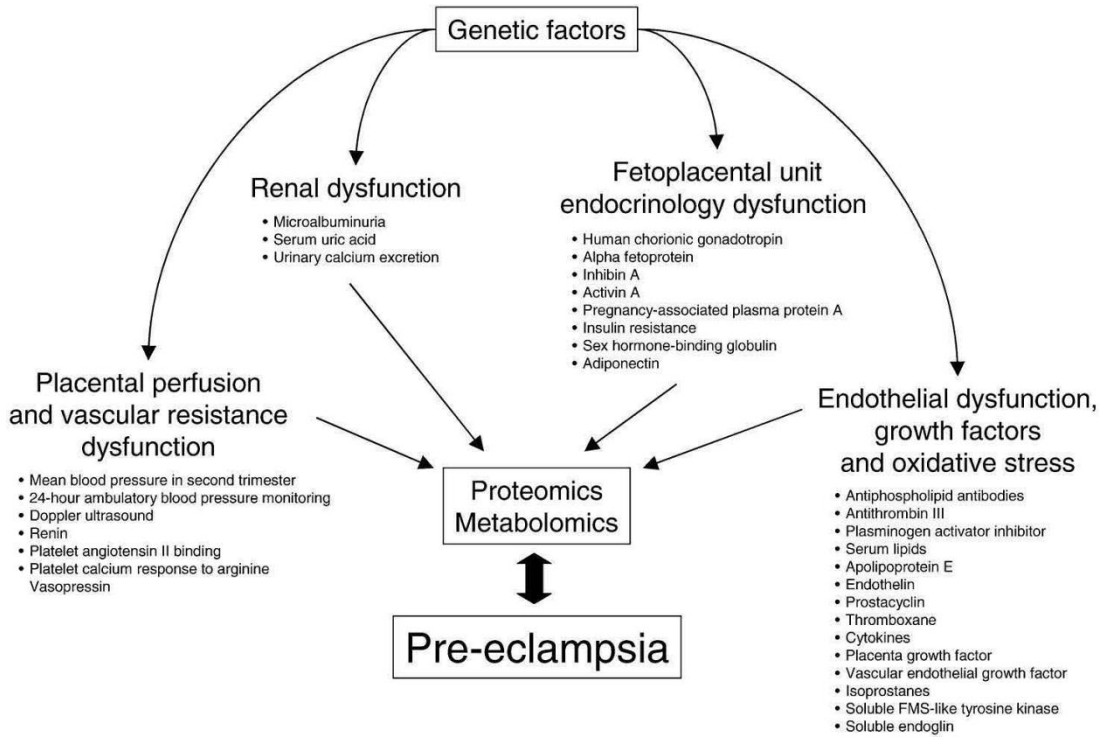


Figure 4. 24: Biomarkers of pre-eclampsia.

Taken from (Carty, Delles et al. 2008), biomarkers for PET are grouped into four categories (Conde-Agudelo, Villar et al. 2004).

4.4 Conclusions

An automated liquid chromatography-based method for the depletion of 14 highly abundant proteins from plasma has been developed. Although initial commissioning of the LC system and column was more time consuming than the spin column approach, the ability to deplete many samples each day was found to be beneficial for a larger scale study.

Here two aliquots of each plasma sample were depleted, combined, concentrated and tryptically digested to ensure sufficient material was available for analysis. In most cases a single depletion would have yielded more sample than was required to obtain quantitative LC-MS^E data, further increasing the sample throughput approximately

2-fold compared to spin columns. In addition, the LC-based methodology released the user to perform other tasks during the depletion of each sample, such as tryptic digestion or data processing. Blank runs were performed between samples but no evidence was obtained to suggest that carryover was a significant problem.

No evidence was obtained from this study to indicate that the IgY-14 column was exhibiting decreased depletion efficiency. The LC-MS^E approach allowed the efficiency to be monitored in every run thus highlighting any potential problems with column aging as they became apparent.

A total of 29 IgY-14 depleted, tryptic digest samples from three obstetric conditions (unaffected pregnancies, trisomy 21 and pre-eclampsia) were analysed using near identical sample loading within the dynamic range of the detector and columns in technical triplicate. Proteins were identified and quantified using the Hi3 approach and the effect of filtering assessed on the number of peptides and proteins identified as well as the average number of peptides/protein and sequence coverage. In the following chapter the effects of filtering the PLGS results will be addressed comparing the proteins identified by the 1D to those using multi-dimensional chromatography approaches.

Using a single chromatographic separation only one of the current predictive markers for trisomy 21, β-hCG, was identified. This was present at elevated level in trisomy 21 pregnancies in the first trimester and was observed in a higher proportion of the T21 samples than those from unaffected pregnancies. β-hCG was most frequently identified in the PET group (70% of analyses) in agreement with studies implicating its elevated level in association with early onset pre-eclampsia.

Two of the predictive markers for Down's syndrome would have been present in the analysed samples below the limit of detection for the discovery proteomics experiment limited to a single chromatographic separation used in this study. The concentration of AFP and inhibin A on column would have been approximately 75 and 1.6 amol respectively assuming no losses, beyond the limits for MRM-based assays even with their extended dynamic range. PAPP-A would have been present at ~5 fmol on column in normal unaffected pregnancy samples but at a reduced level in

the T21 group and was not identified in this study. The ELISA-based assays have a limit of detection for the predictive markers in the very low ng mL^{-1} or even pg mL^{-1} levels and do not require a depletion step, indicating that proteomics approaches have to mine significantly deeper into plasma to even compete with current screening techniques. Elisa-based assays can be completed in minutes to a few hours compared to days for the depletion, tryptic digestion and LC-MS^E analysis to be conducted.

Proteins were identified and quantified that showed little variation between samples or obstetric conditions and could potentially be selected for use as internal standards in an MRM-based assay to compensate for differences in sample loading. Some proteins were observed at levels that varied greatly between samples but were not specific to one obstetric condition.

Potential biomarkers for Down's syndrome and pre-eclampsia were identified from the literature and considered in the context of this study. In almost all of the examples, no evidence was obtained to validate the published biomarker for the suggested obstetric condition identified.

In some cases, levels of a protein varied up to 10-fold between samples which distorted the average level of that protein for the obstetric group. The advantage of analysing individual samples meant that this observation could be identified as patient-specific and so conclusions were not drawn about the protein's suitability as a biomarker. In pooled samples this natural variability would not be observed and incorrect assumptions may be drawn.

The use of pooled samples giving an average level of each protein in plasma contrasts with the use of median values for predictive screening for Down's syndrome. LaBaer suggested that comparing average values from biomarker studies can "mislead the investigator" and that high confidence statistical differences do not necessarily indicate a good biomarker (LaBaer 2005). The range of values obtained from a *population* should be used to determine if suitable cutoff values for selectivity and specificity can be employed. This would require a much larger number of

samples from each population of obstetric conditions (than used in this study) to be analysed comprehensively.

Filtering the data affected the number of proteins and peptides identified in each analysis. The most stringent filtering was reserved for the quantitative data where each protein reported was identified in >1 replicate per sample. This significantly reduced the number of measurements reported from the entire data set: though a protein may have been identified in many other samples, it would be rejected in others suggesting perhaps, that in a large data set such as this alternative filtering methods could be effectively employed. Results were filtered for observation in a minimum number of *analyses* rather than sample replicates. Although initially a large decrease in the average number of proteins and peptides identified was observed when any filter was employed, the average value for sequence coverage, peptides/protein and peptides/analysis varied little with increased stringency of filtering >2 analyses. The data suggests that a less stringent approach to assessing LC-MS^E results may be more appropriate, particularly in the case of large sets of analytical data. The effects of filtering results will be evaluated further in the following chapter where single and multiple chromatographic separations have been employed.

Chapter Five: Comparative analysis of maternal plasma by utilising single and multi-dimensional chromatography

5.1 Introduction

In a proteomics experiment a complex mixture of proteins is typically enzymatically digested to generate a sample containing many 10's of thousands of peptides for identification which would require an MS/MS scan rate of 25 s^{-1} for selection of all the eluting features from the LC separation (Michalski, Cox et al. 2011). Reversed phase (RP) liquid chromatography has been extensively used for the separation of peptides prior to mass spectrometric analysis but in itself can be insufficient to resolve all the peptide species. Thus, more peptides enter the instrument than can be identified.

In RP chromatography hydrophobic molecules are adsorbed onto a hydrophobic immobilised stationary phase in the presence of a polar, aqueous solvent. Separation is achieved through a partitioning mechanism in which equilibrium is established between the analytes and the mobile and stationary phases. Decreasing the polarity of the mobile phase by increasing the organic solvent content results in desorption of the analytes from the stationary phase. More hydrophobic analytes require higher concentrations of organic solvent to promote desorption. Under low pH conditions RP separations are achieved by the addition of an ion pairing agent such as trifluoroacetic acid or, for proteomic applications, formic acid.

Multi-dimensional chromatography approaches for proteomics applications usually employ strong cation exchange (SCX) followed by in-line or off-line RP. Gilar *et al.* evaluated a number of 2D approaches including size exclusion-RP, hydrophobic interaction-RP, SCX-RP and RP-RP (Gilar, Olivova et al. 2005a). Peak capacity can be viewed as the separation efficiency of the LC system and may be defined as the number of components that could be theoretically separated on a column within a given gradient time (Ghrist, Stadalius et al. 1987; Stadalius, Ghrist et al. 1987; Neue, Carmody et al. 2001). For multi-dimensional systems the peak capacity can be calculated as the multiplication of the individual separations from both dimensions. This would only hold true if the two separation strategies were orthogonal and in reality this may not be the case.

A combination of two types of chromatography has been shown not to be sufficient in some cases to significantly improve the peak capacity. In the Gilar study, two types of RP column were utilised, Atlantis dC18 and Phenyl using a five protein digest mix with six non-tryptic peptides added. They reported that although the two columns were suitable for some applications, such as the separation of critical peptide pairs, in general the separation had limited orthogonality.

In contrast the use of RP-RP with mobile phases at different pH has demonstrated similar levels of orthogonal separation to other 2D separations (Gilar, Olivova et al. 2005a). The study reported that acidic peptides were retained more at low pH and basic peptides at high pH. As both column stationary phases were C_{18}-based the effects were predominantly due to pH. The authors reported that in SCX separations selectivity was based on peptide charge, as expected, but since tryptic peptides are predominantly doubly and triply charged, separation space was restricted to a narrow retention window. Longer peptides which are retained by RP separations (more hydrophobic) exhibited reduced retention times compared to shorter peptides. HILIC-RP was identified as having a greater degree of orthogonal separation but solubility of peptides in high organic solvent (70% acetonitrile) for the first dimension chromatography was expressed as a concern.

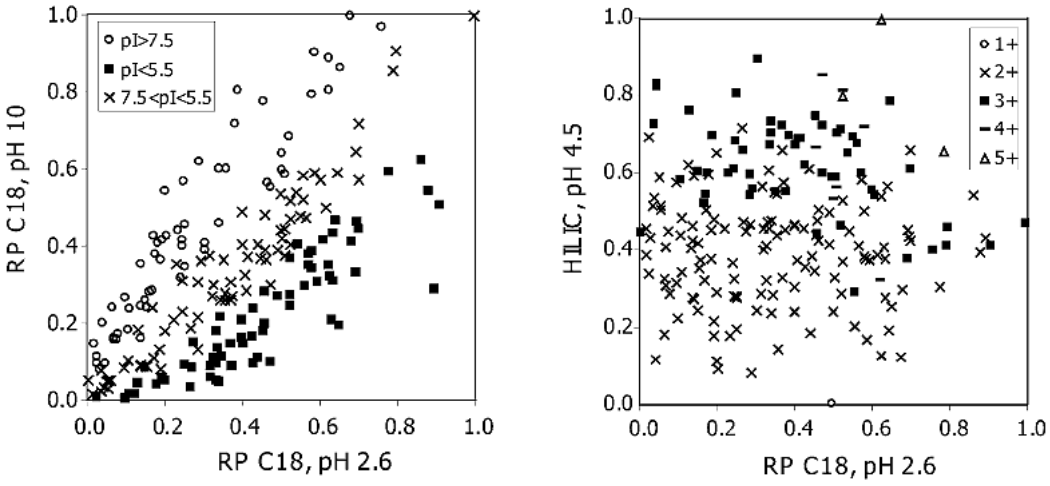


Figure 5. 1: Normalised retention time plots for two-dimensional chromatographic separations.
In the left panel, high/low pH RP-RP separation of peptides and right hydrophilic interaction-RP was utilised, taken from (Gilar, Olivova et al. 2005a)

In another study by Gilar and co-workers they investigated the use of different types of RP sorbents, effect of pH and ion pairing agents (Gilar, Olivova et al. 2005b) and found that the use of solvents at pH 10 and 2.6 for the first and second dimension respectively provided substantial orthogonality, even when using the same C_{18} column material. They identified a reduced level of peptide loss compared to SCX-RP and better recovery of both long and short peptides in RP-RP without the complication of salt buffers. The authors also suggested that the predictable retention times of peptides eluting from RP separations both at low and high pH could be used as a data filter improving peptide identification. In contrast, SCX retention time provides information on peptide charge but there can be significant overlap between the RT windows occupied by each of the multiply charged species.

Toll *et al.* reported the separation of peptides at elevated temperatures with low and high pH buffers analysed in both positive and negative ion ESI-MS with similar peak capacities and column efficiency (Toll, Oberacher et al. 2005). The authors suggested that although the low and high pH separations were not fully orthogonal, they showed a considerable degree of complementarity making this 2D approach applicable for the analysis of complex peptide mixtures. Lower information content was reported for the deprotonated MS/MS spectra, due to gaps and truncations in the fragment ion series resulting in uninterpretable spectra, combined with the reduced sensitivity expected in negative ion mode.

Zhou and co-workers compared SCX-RP and RP-RP for the analysis of peptides from a commercial *E. coli* digest, a HeLa S3 cell lysate and multi-component protein complexes isolated using tandem affinity purification, (Zhou, Cardoza et al. 2010). The authors observed that the RP-RP approach identified more peptides and proteins than SCX-RP, using identical MS conditions, with more peptides/proteins observed from 200 ng of an *E. coli* digest. No statistically significant bias was observed in terms of the peptides identified by RP-RP. Similar observations were reported from the analysis of a tandem affinity purified Ku protein complex. The authors concluded that RP-RP outperforms SCX-RP with respect to the total number of identified proteins and dynamic range as a result of the additional peak capacity afforded by the first dimension.

The development of robust multi-dimensional approaches for the separation of complex tryptic digests combined with MS^E data acquisition has been shown to provide significant potential for the analysis of complex biological mixtures. In this study, a 2D RP-RP chromatographic separation approach was employed using an on-line 2D NanoAcquity system (Waters Corporation, Milford, MA, USA). Samples are loaded onto the first dimension column at high pH, using the flow path configuration shown in Figure 5. 2. This allows for on-line dilution of the eluate from the first dimension column with aqueous buffer from the second binary solvent manager (shown in blue) prior to the trapping step. Analytical separation in the second dimension is achieved using an increasing linear gradient of organic solvent with the flow path shown in Figure 5. 3. Subsequent fractionation of the sample is achieved using a pulse at lower pH (increased acetonitrile content) generated by the first binary solvent manager (shown in red), prior to on-line dilution and trapping, as described. During analytical separation, the first dimension column is held at high pH.

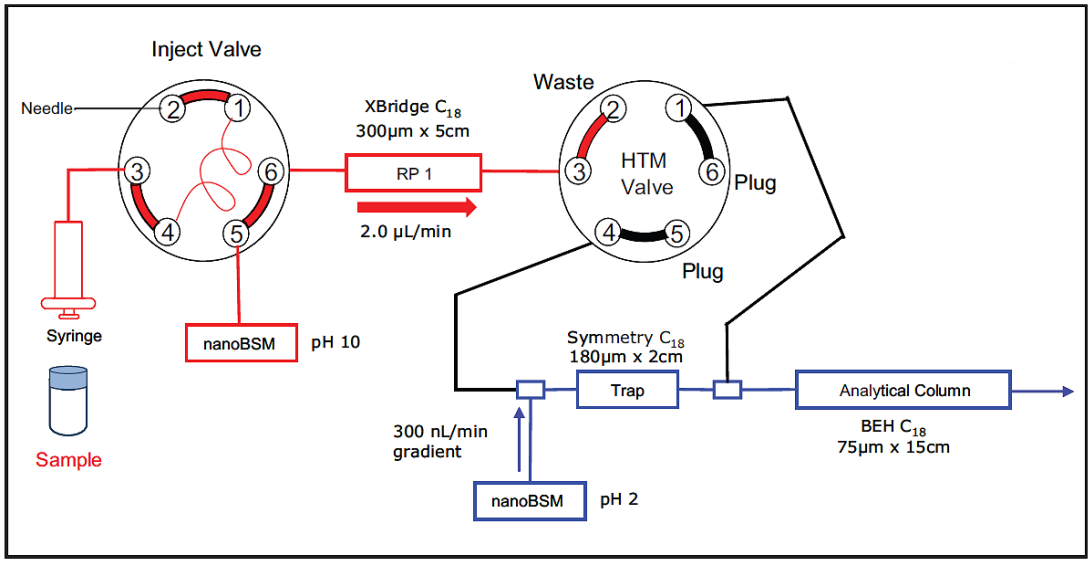


Figure 5. 2: Fluidic flowpath for a 2D RP-RP NanoAcquity system with on-line dilution during sample loading, fractionation and trapping.
Extracted from Waters Application Note 20003174EN LB-PDF (July 2009).

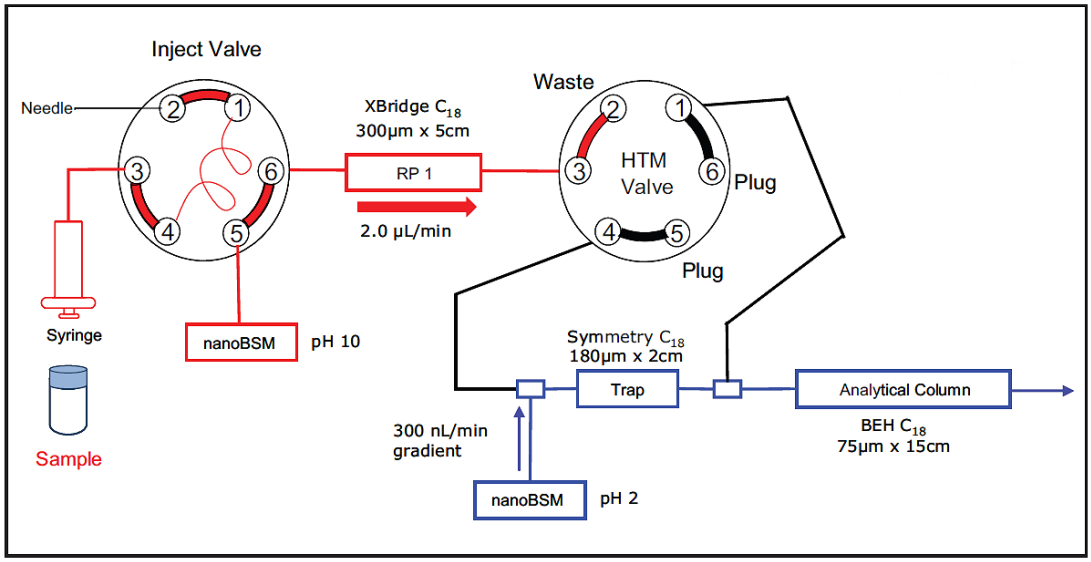


Figure 5. 3: Fluidic flowpath for a 2D RP-RP NanoAcquity system with on-line dilution during analytical separation.
Extracted from Waters Application Note 20003174EN LB-PDF (July 2009).

The aims for this work were:

- Deplete two plasma samples generated from normal and trisomy 21 maternal plasma not used in previous experiments.
- To utilise a two-dimensional (RP-RP) chromatographic separation with MS^E data acquisition for the analysis of pooled, depleted maternal plasma.
- Compare single and multi-dimensional approaches in terms of protein and peptide identification rates.
- Assess suitable filter thresholds for MS^E data based on empirical evidence obtained in this study.
- Determine comparative protein levels assessed from single and multi-dimensional separations and different patient groups.
- Assess putative biomarker proteins identified using each approach.

.

5.2 Materials and methods

5.2.1 Material suppliers

A Seppro® IgY14 LC2 column kit (Sigma Aldrich, Gillingham, UK) was obtained which included an LC column (6.4 x 63.0 mm bed volume 2 mL), dilution, stripping and neutralisation buffers.

Double centrifuged maternal plasma samples were supplied by Prof. Kypros Nicolaides of King's College Hospital, London, UK with full ethical approval.

Forty maternal plasma samples were selected for this part of the study, spanning a range of age, BMI, ethnicity with equal numbers of samples identified as normal and trisomy 21 gestations. The average maternal age was 33.5 years, BMI 23.9 and gestational age 89.9 days. Full sample details are included in Appendices D and E.

Rapigest surfactant and MassPREP™ ADH and phosphorylase B digestion standards were obtained from Waters Corporation (Milford, MA, USA). Dithiothreitol was supplied by Melford Labs. (Ipswich, UK). Glu^1-Fibrinopeptide B peptide (human), sodium azide, iodoacetamide, ammonium bicarbonate, LC-grade water and acetonitrile were purchased from Sigma Aldrich (Gillingham, UK) and sequencing grade trypsin from Promega (Madison, WI, USA). Mass spectrometry solvents were supplied by MallinckrodtBaker Inc. (Phillipsburg, NJ, USA). Spin-X cellulose acetate centrifuge tube filters were supplied by Costar (Corning Inc., Tewksbury MA, USA), 2 mL square 96-well trays by Beckman Coulter (Fullerton, USA) and 5 kDa nominal molecular weight cut-off (NMWCO) spin columns and 0.45 µm nitrocellulose filter discs by Millipore (Billerica, MA, USA). Sample vials (LCMS Certified) were purchased from Waters Corporation (Milford, MA, USA) fitted with pre-slit PTFE/silicone septa in the caps.

5.2.2 Sample preparation

Each individual maternal plasma sample was thawed from -70 ^{o}C at room temperature and inverted a number of times to ensure homogeneity prior to a 20 µL

aliquot being removed. Each aliquot from 20 normal, unaffected pregnancies was combined, diluted to 2 mL with 1 x dilution buffer and centrifuged using a 0.22 µm Spin-X cellulose acetate centrifuge tube filter prior to depletion. Each aliquot from 20 trisomy 21 maternal plasma samples was combined, diluted and filtered as described above.

5.2.2.1 Fractionation of pooled human plasma using an IgY-14 LC2 column

The automated method developed to deplete maternal plasma using an IgY-14 LC2 format chromatography column was utilised to partition the two pooled plasma samples. This is described in detail in Section 4.2.2.1. Fractions were collected from the column using a 2 mL capacity square 96-well plate during the chromatographic steps and detection of protein elution was achieved using absorbance at 280 nm with a 5 Hz sampling rate. All aqueous solutions used in the depletion protocol were filtered through a 0.45 µm nitrocellulose filter disc and degassed under vacuum for at least 15 minutes. The timetable for the chromatographic separation is described in Table 4. 2.

After priming the system fully, the column was connected to the system and a full loop 250 µL injection of 1x dilution buffer was used to perform a blank chromatographic run using the timetable shown in Table 4. 2. A blank run was undertaken between the two plasma samples.

Four depletion steps, each containing the equivalent of 50 µL of plasma, were performed for each of the two pooled maternal plasma samples to ensure that sufficient sample was generated for 2D-LC-MS^E analysis. The (depleted) fractions that eluted between 5 and 18 min from both chromatographic separations from each sample, were combined and concentrated using 5 kDa NMWCO spin columns, to generate a single sample for tryptic digestion.

5.2.2.2 Tryptic digestion of IgY-14 fractionated pooled plasma

The IgY-14 depleted plasma fraction (total volume approximately 1 mL, n=20 for each obstetric group) was transferred to a vial containing lyophilised Rapigest surfactant, final concentration 0.1% w/v and gently agitated until fully dissolved. The contents were transferred to an 5 kDa NMWCO spin column and centrifuged at 14,000 g, 4 ^{o}C until the volume was approximately 50 µL.

Contents were transferred to a 0.5 mL microfuge tube (Fisher Scientific Ltd, Loughborough, UK) and incubated in a water bath at 80 ^{o}C for 15 min. A 5 µL aliquot of 100 mM dithiothreitol in 100 mM ammonium bicarbonate (NH_4HCO_3) was added to the plasma and thoroughly agitated prior to incubation at 60 ^{o}C for 15 min, followed by the addition of 5 µL of 200 mM iodoacetamide in 100 mM NH_4HCO_3 and incubated at room temperature, in the dark for 30 min.

A vial containing 20 µg of trypsin was fully resolubilised in 20 µL of 100 mM NH_4HCO_3 with 2 µL (2 µg) transferred to the plasma sample and thoroughly agitated. The sample was incubated overnight at 37 ^{o}C.

The following day, 2 µL of concentrated formic acid were added to the sample and incubated at 37 ^{o}C for 15 min, prior to filtration through a 0.22 µm Spin-X cellulose acetate centrifuge tube filter. An aliquot of the tryptically digested sample was removed for analysis and the remainder stored at -20 ^{o}C until required. Typically 45 - 50 µL of tryptic digest were obtained for each depleted plasma sample giving a final concentration of approximately 4.8 µg μL^{-1} (from two combined IgY-14 depletions).

An aliquot of each tryptic digest was diluted and combined with a solution containing either MassPREP™ glycogen phosphorylase (PhosB) or alcohol dehydrogenase (ADH) tryptic digestion standard in 0.1% v/v aqueous formic acid. This typically produced a sample providing 50 fmol PhosB on column for the LC-MS^E analyses. For 2D-LC-MS^E experiments the concentration of ADH on column was either 48 fmol or 76 fmol and this was used as an internal standard for estimating observed protein concentrations using the Hi3 approach (Silva, Gorenstein et al. 2006).

5.2.3.1 LC-MS^E configuration

All nanoscale liquid chromatographic separations were performed using a directly-coupled NanoAcquity UPLC system and a nanoelectrospray source (Waters Corporation, Milford, MA, USA). The system was composed of a binary solvent, auxiliary solvent and sample manager fitted with a heating and trapping module.

LC separations were performed using a Symmetry C18 trapping column (180 μm x 20 mm 5 μm) and a BEH C18 analytical column (75 μm x 250 mm 1.7 μm). The composition of solvent A was 0.1% v/v aqueous formic acid and solvent B 0.1% v/v formic acid in acetonitrile.

An aliquot of each sample containing internal standard was applied to the trapping column and flushed with 0.1% solvent B for 2 min at a flow rate of 15 μL min^{-1}. Sample elution was performed at a flow rate of 250 nL min^{-1} by increasing the organic solvent concentration from 3 to 40% B over 90 min, with a total run time of 115 min. The mass spectrometer was fitted with a universal nanoflow sprayer (Waters Corporation, Milford, MA, USA) and an applied capillary voltage of 3.5 kV was used. All analyses were conducted in technical triplicate.

Prior to and after each set of technical replicates, a quality control (QC) injection of 50 fmol PhosB was analysed using LC-MS^E and the data processed by PLGS. Where the peptide sequence coverage fell below 35% for PhosB, no further sample data were collected and the cause of the loss of peptide identification investigated and resolved. A minimum of four QCs were collected every 24 hr during sample data collection.

MS^E data acquisition was performed on a Synapt HDMS instrument (Waters Corporation, Milford, MA, USA), configured for MS^E through the MS Method Editor controlled by MassLynx v4.1. The time-of-flight analyser of the mass spectrometer was externally calibrated using the MS/MS spectrum obtained from the doubly charged precursor of the GFP peptide from *m/z* 50 to 1300. The calibration was manually validated with an average ppm error across the mass range <10 ppm. GFP was used for lockmass correction (*m/z* 785.8426) infused via a NanoLockSpray interface at a constant rate of 500 nL min^{-1} at 500 fmol μL^{-1} and sampled every 60 seconds.

In low energy MS mode, data were collected at constant trap collision energy of 6 V. In elevated energy MS mode, the trap collision energy was ramped from 15 V to 30 V whilst the transfer collision energy was held at 3 V and 10 V for low and elevated conditions respectively. The trap vacuum was held at a constant 8.2 mbar. All subsequent data were post-acquisition lockmass-corrected using the monoisotopic ion of the doubly charged precursor of GFP (*m/z* 785.8426).

5.2.3.2 2D-LC-MS^E configuration

All nanoscale liquid chromatographic separations were performed using a directly-coupled 2D NanoAcquity UPLC system with a nanoflow source (Waters Corporation, Milford, MA, USA). The system comprised of two binary and one auxiliary solvent manager with a sample manager fitted with a heating and trapping module.

An aliquot of pooled plasma tryptic digest was loaded onto the first dimension column, Xbridge C18 (300 µm x 5 cm 5 µm) equilibrated in 20 mM ammonium formate pH 10 at 2 µL min^{-1}. A discontinuous 6-step gradient of acetonitrile was used (11.1, 14.5, 17.4, 20.8, 45 and 65%) to elute peptides onto the trapping column, Symmetry C18 (180 µm x 20 mm 5 µm). The concentration of acetonitrile pulses was generated automatically within MassLynx software upon selection of number of fractions for the 2D experiment. The fractions containing organic solvent were diluted ten-fold using aqueous flow from the 2nd dimension binary solvent manager prior to trapping. For the 2nd dimension analytical separations, a BEH C18 column (75 µm x 200 mm 1.7 µm) was employed using a 300 nL min^{-1} flow rate.

5.2.4 Processing of MS^E acquired data

LC-MS^E data were processed using PLGS v2.4 or v2.5.1 and lockspray calibrated against GFP using data collected from the reference line during acquisition. Ion detection, clustering and protein identification have been explained in detail in Section 1.4.3.6. In brief, lockmass-corrected spectra are centroided, deisotoped and charge-state-reduced to produce a single accurately mass measured monoisotopic mass for each peptide and its associated fragment ions. Initial correlation of a

precursor and potential fragment ions is achieved using time alignment (Geromanos, Vissers et al. 2009). Data processing parameters specified 250, 100 and 1500 for the low, elevated and intensity threshold values respectively for PLGS v2.4 or 200, 100 and 750 in PLGS v2.5.1.

Data obtained from each 2D-LC-MS^E fraction was processed independently using PLGS v2.4. Processing parameters specified 200, 75 and 1500 for the low, elevated and intensity threshold values respectively. Each of the 6 processed data files that comprised a single 2D-LC-MS^E analysis were then merged into a single file prior to database interrogation.

5.2.5 Database interrogation using MS^E data

PLGS processed LC-MS^E data were used to interrogate the IPI human databases rel. 3.69 and 3.85 downloaded from (http://www.ebi.ac.uk/IPI/IPIhuman.html) randomised once to form a concatenated database of genuine and random entries and appended with the sequences for porcine trypsin, rabbit glycogen phosphorylase and alcohol dehydrogenase (P00761, P00489 and P00330 - http://www.uniprot.org/).
For MS^E data, database search parameters included a fixed modification of carbamidomethyl cysteine, one missed trypsin cleavage site with variable modifications of acetyl N-terminus, oxidation of methionine and deamidation of asparagine and glutamine.

Precursor and fragment ion tolerances were determined automatically by PLGS. Protein identification criteria included the detection of at least three fragment ions per peptide, seven fragment ions per protein and at least one peptide per protein with a 4% false discovery rate (FDR). PhosB or ADH was specified as the internal standard and the concentration specified (in fmol) in the PLGS workflow template to allow the Hi3 estimation of protein concentration.

Protein identifications obtained from each of the tryptic digests, analysed in triplicate were exported from the PLGS browser into Microsoft Office Excel. Within Excel, the protein identification results could then be further filtered for replication based either on the number of times the protein was observed from each sample or from the entire data set.

In some cases, a protein sequence had been identified in a previous release of the IPI human database but its accession number had been deleted from more recent versions. In these cases the sequence was retrieved from UniParc (http://www.uniprot.org/uniparc/) and a BLAST sequence similarity seach was conducted to identify a homologous alternative.

5.2.6 Relative protein expression

Relative protein expression analysis was performed using the Expression algorithm in PLGS v2.4 and 2.5.1. Results from each of the replicate 2D analyses were grouped according to obstetric conditions (normal or T21) within PLGS and the data were processed at the protein level using auto normalisation.

Protein lists were filtered within PLGS to remove proteins that were only observed in one data file/obstetric condition and had a score <1000. For each reported ratio, a probability value was assigned by PLGS from 0 to 1; those values that are 0.05 or lower and 0.95 and higher, represented regulation likelihood greater than 95% and were assumed to be statistically significant.

5.3 Results and discussion

5.3.1 Comparison of pooled plasma samples analysed using 1D and 2D-LC-MS^E

Two pooled samples representing the obstetric conditions trisomy 21 and normal, unaffected pregnancies were generated from 20 individual plasma samples each. The pooled samples were depleted of 14 highly abundant proteins using an IgY-14 LC2 format column, tryptically digested and analysed using 1D and 6-fraction 2D-LC-MS^E in triplicate. A typical set of chromatograms from each of the six fractions has been shown in Figure 5.4.

Average sample loading for the 1D experiment was 1.1 μg and for the 2D analysis 0.71 μg identifying a total of 287 and 451 proteins respectively, if no filtering of the results was performed. This equated to an increase of 1.6-fold in protein identification rate when the 6-fraction 2D experiments were performed despite a reduced sample loading, Table 5. 1.

The number of peptides identified using 1D analysis ranged from 1,297 to 2,090 with an average of 1,766 for the normal pregnancy sample and 1,628 for trisomy 21 when no filtering of the results was performed. The additional separation capacity of the 2D chromatographic separations was demonstrated by an increase in peptide identification of up to 4.4-fold in the 2D experiments (an average 5,248 for the pooled normal sample and 5,740 for T21 and a 2-fold increase in the average number of peptides identified from each protein from 12.4 to 26.4 for 1D and 2D separations respectively, Table 5. 1. Previous studies have observed a similar increase in peptide and protein identification due to the increased peak capacity. Zhou and co-workers reported 1.7-fold and 1.2-fold increases respectively by the use of RP-RP compared to SCX-RP with identical sample loading in the characterisation of TAP enzyme complexes concluding that the effect was as a result of additional peak capacity in the first dimension RP step (Zhou, Cardoza et al. 2010). The RP-RP chromatographic separation used here with no additional sample loading, outperformed the single dimension analysis as a result of the additional peak

capacity and the degree of orthogonality in the two RP separations (Gilar, Olivova et al. 2005a). It has been suggested that peak capacity for an RP separation on a state-of-the-art column at high or low pH can approach 300 resulting in a larger potential separation space (Gilar, Olivova et al. 2005b).

The similar sample loading used for both 1D and 2D separations here constrained the improvements usually observed with the use of multi-dimensional chromatography. Loading would usually be increased by the number of first dimension fractions e.g. if 500 ng was the optimal loading for RP only, then a 6-fraction RP-RP experiment would utilise a 3 µg level (6 x 500 ng). A sample loading of <800 ng for the 6-fraction 2D experiments described here was utilised. Gilar and co-workers demonstrated the improvement in peptide and protein identification when comparing 1D and 2D (RP-RP)-LC-MS^E analysis of undepleted serum (Gilar, Olivova et al. 2009). They reported 52 proteins identified by 316 unique peptides for the 1D approach increasing to 191 proteins and 1083 peptides in the RP-RP analysis with 95% of the peptides detected in the 1D experiment observed in both approaches, using a 10-fold greater molar load. It would be anticipated that a similar increase in peptide, protein and sequence coverage rates would be observed with additional sample loading utilising the 2D approach.

Patel *et al.* also observed an increased number of protein identifications with an RP-RP-MS^E approach compared to 1D (Patel, Crombie et al. 2012). Employing a 6-fraction first dimension, 603 proteins were identified from soluble cytoplasm of the prokaryote *Methylocella silvestris* compared to 178 from the single dimension. The use of 11 fractions only provided an additional 174 proteins. The authors concluded that an increase in the number of fractions beyond six had diminishing returns in terms of output. In contrast Zhou reported a continual increase in the number of peptide and protein identifications as a function of fraction depth (Zhou, Cardoza et al. 2010) using a DDA approach selecting either the 3 or 5 most abundant peptides with charge state 2^+ to 4^+.

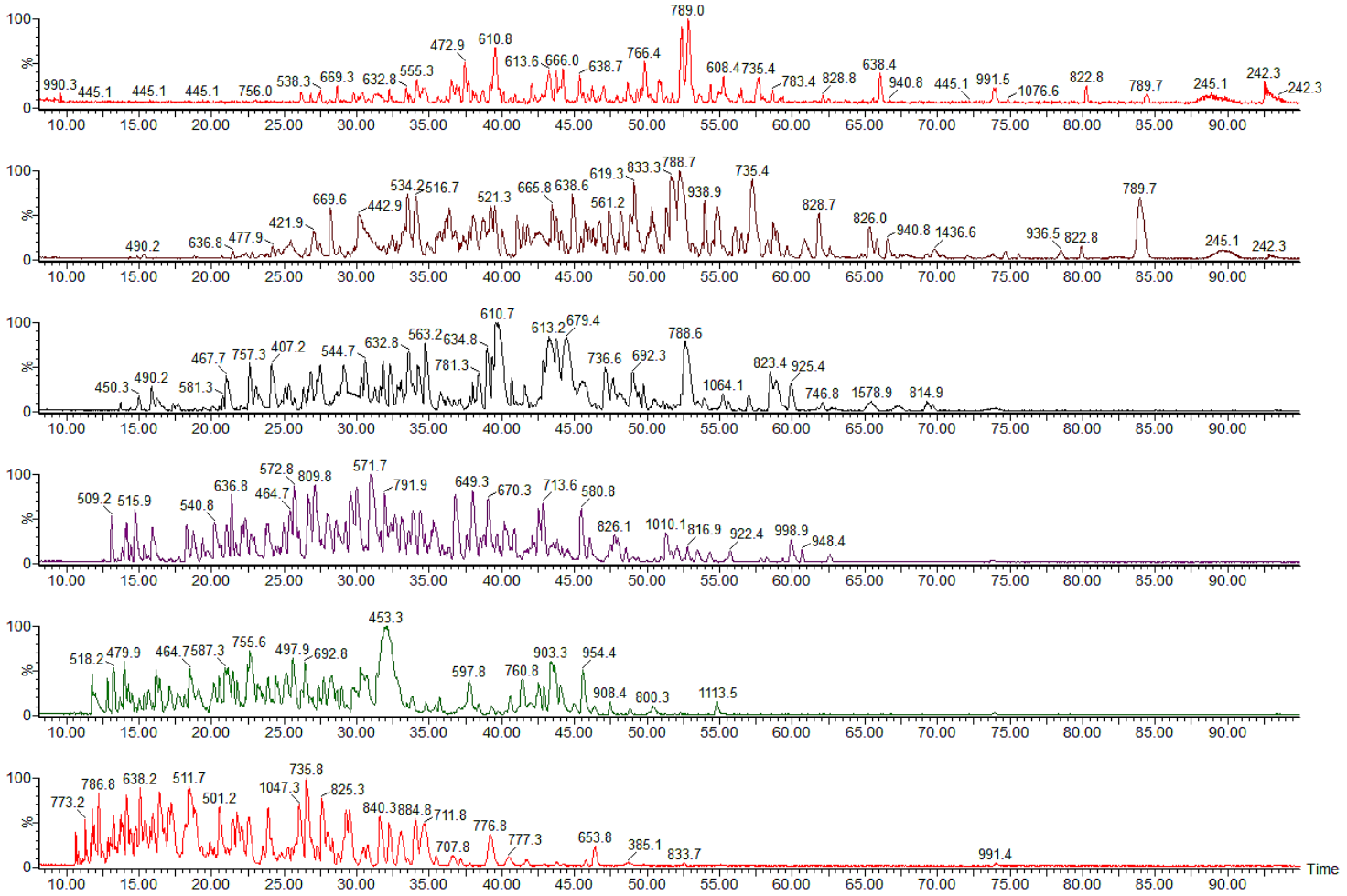


Figure 5. 4: Base peak intensity chromatograms obtained from a six-fraction RP-RP separation of depleted human plasma. The acetonitrile pulses used were 11.1%, 14.5%, 20.8%, 45% and 65%, shown above from lower to upper panels.

	2D Pooled Normal Plasma	2D Pooled T21 Plasma	1D Pooled Normal Plasma	1D Pooled T21 Plasma	Fold-increase (2D/1D) Normal plasma	Fold-increase (2D/1D) T21 plasma
Average loading excl. internal standard (ng)	796.96	628.61	1111.16	1210.22		
Proteins identified - no filter	312	290	199	183	1.57	1.58
Proteins identified >1 file	146	131	85	81	1.72	1.62
Proteins identified >2 files	110	98	62	61	1.77	1.61
Average number of peptides/protein - no filter	26.53	26.36	12.30	12.49	2.16	2.11
Average number of peptides/protein >1 file	39.87	44.57	19.31	17.20	2.06	2.59
Average number of peptides/protein >2 files	45.61	38.14	19.50	19.76	2.34	1.93
Average sequence coverage/protein - no filter	26.38	26.83	24.50	23.45	1.08	1.14
Average sequence coverage/protein >1 file	33.66	36.40	32.69	30.18	1.03	1.21
Average sequence coverage/protein >2 files	35.83	36.08	32.91	32.23	1.09	1.12
Average number of peptides Replicate 1	5400	5757.00	1312.00	1297.00	4.12	4.44
Average number of peptides Replicate 2	5218	5540.00	2090.00	1916.00	2.50	2.89
Average number of peptides Replicate 3	5125	5923.00	1897.00	1672.00	2.70	3.54
Average number of peptides all analyses	5247.67	5740.00	1766.33	1628.33	2.97	3.53
Random proteins identified - no filter	70	66	46	40		
Random proteins identified >1 file	2	1	0	1		
Random proteins identified >2 files	0	0	0	0		
False Discovery Rate (FDR) - no filter	22.44%	22.76%	23.12%	21.86%		
FDR >1 file	1.37%	0.76%	0.00%	1.23%		
FDR >2 files	0.00%	0.00%	0.00%	0.00%		

Table 5. 1: Comparison of results obtained from pooled plasma analysed using 1D and 2D-LC-MS^{E}.

5.3.2 Assessment of the stringency of filtering protein identifications

Protein identification results from a LC-MS^E experiment generated using default PLGS parameters would typically have a false discovery rate of ~4%. With an increasing number of replicates, there was a concomitant increase in the number of random entries observed, however the number of non-random entries identified did not increase at the same rate. This disproportionate representation by false positive protein identifications can result in a high FDR when compiling the results from multiple technical replicates, where no filtering of the identification has been performed. The FDR calculated from three technical replicates analysed using 2D or 1D-LC-MS^E for both normal and T21 pooled samples was 22.6% and 22.5% respectively, Table 5. 1. The Aebersold group have reported that the size of a proteomic data set can have an impact on the protein false discovery rates compared to the peptide level (Reiter, Claassen et al. 2009). In particular, the false positive single peptide occurrences were nearly two orders of magnitude larger than the average FDR for the entire set. The authors developed the MAYU software to establish FDR at the protein level using assemblies of peptide spectrum matches.

Two approaches to filtering the results have been utilised in this study. The most stringent filter required a protein to be observed in >1 technical replicate (Patel, Thalassinos et al. 2009; Patel, Crombie et al. 2012). This typically reduced the FDR to 0.0% but resulted in a significant reduction in the number of identifications from each sample. This filter was used for the quantitative results presented. A less rigid filtering approach was adopted in Chapter 4, due to the number of LC-MS^E analyses performed (87 in total) which required that a protein be observed in >1 analysis across all the experiments as it was observed that many proteins were identified multiple times but not necessarily in the same sample. An assessment of the suitability of this filter (>1 analysis) has been performed.

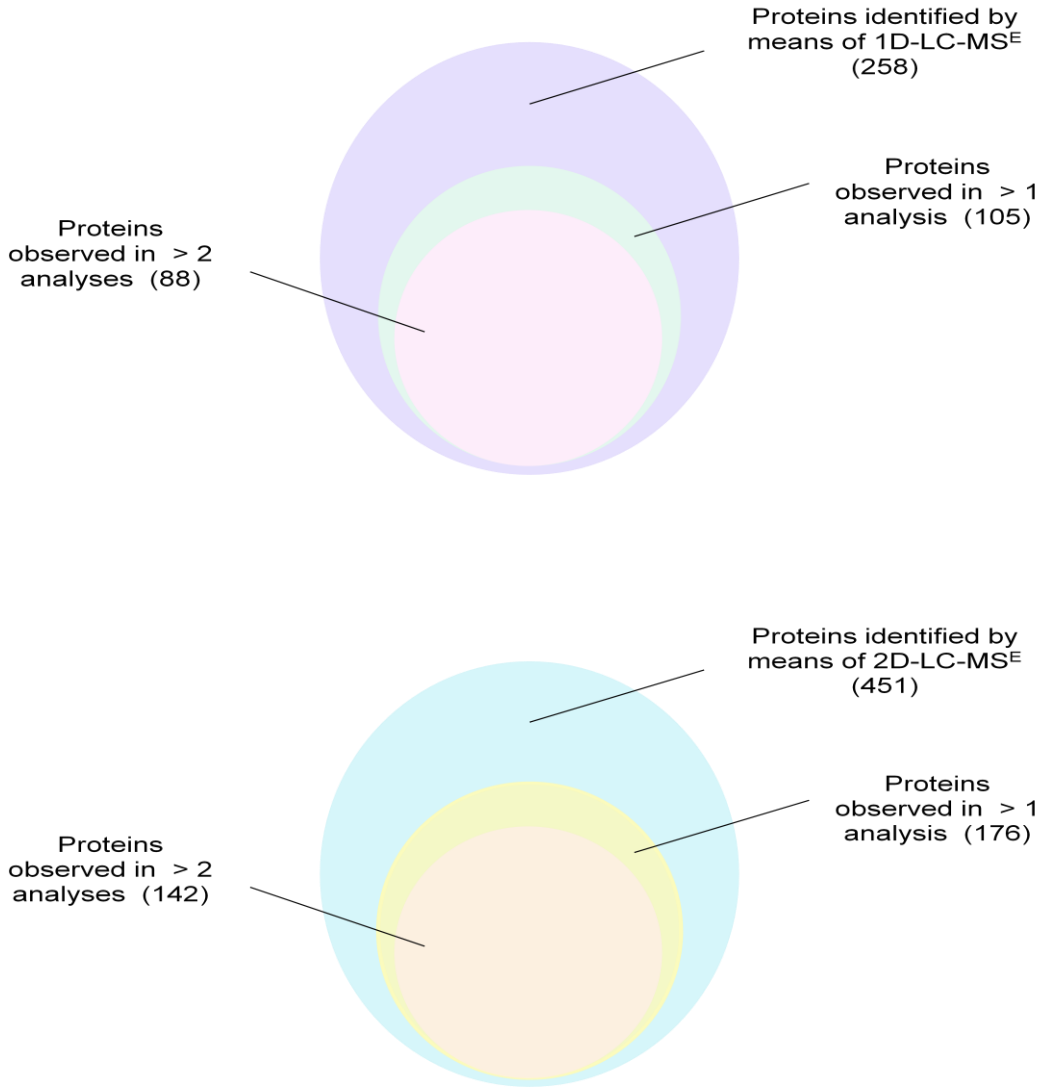


Figure 5. 5: Venn diagram indicating proteins identified using LC-MS^E from pooled depleted plasma.
Proteins identified by 1D-LC-MS^E are shown in the upper panel and 2D-LC-MS^E in the lower panel.

To assess a suitable filter for this and further studies, a comparison of protein identification results from pooled 1D and 2D experiments was conducted, utilising all 6 analyses from each (three from normal and three from T21 depleted plasma). An assumption was made that if a protein was identified in >1 analysis across the 2D experiments then it was classed as a genuine, confident identification of a plasma protein and 176 identifications were used for this purpose out of a total 451. This allowed proteins identified from the 1D pooled analyses to be cross-checked against the list of 176 confident plasma proteins from the 2D experiments. From all six 1D pooled analyses, 287 proteins were observed, of which 105 proteins were identified in >1 1D analysis and 94 of those were included the 2D list. Proteins observed in just one of the 1D analyses totalled 182 and 21 of those were observed in the 2D confident plasma protein list. The 21 1D one hit wonders could be further segregated by their average score assigned by PLGS software with 6 exceeding 1000 and 11 less than 249, Figure 5. 6. A score of 1000 was selected as this is typically used as a cut-off within PLGS software to reduce FDR and reject low scoring proteins in comparative proteomic experiments. The scores below 1000 were then divided into four groups in score increments of 250.

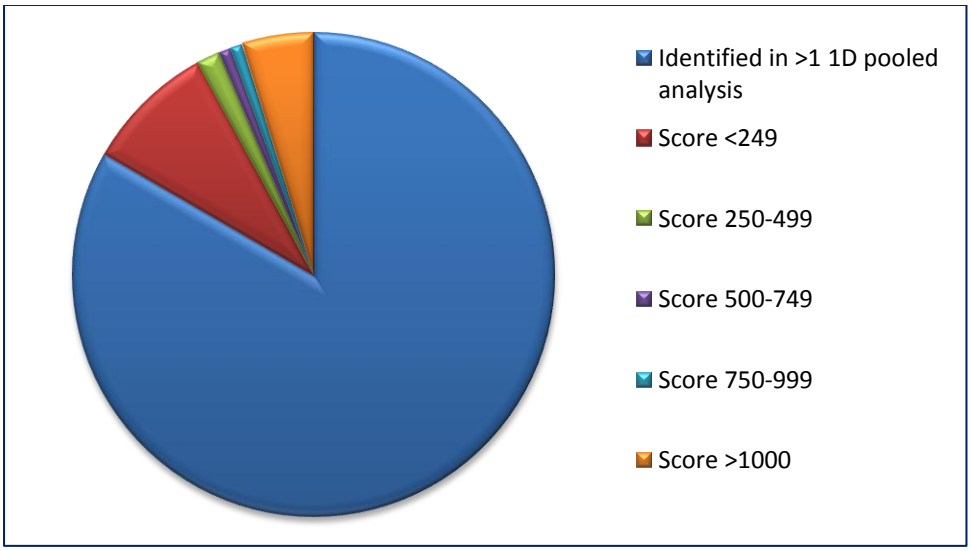


Figure 5. 6: Proteins observed in 1D pooled plasma analyses confidently identified in 2D pooled plasma.
A protein was classed as confidently identified if observed in >1 pooled 2D analysis. A total of 105 proteins from the 1D pooled experiments were included in the confident identification list and the one hit wonders were subdivided further according to their average score.

Results obtained from pooled 1D and 2D experiments were filtered for observation in >1 analysis,

Figure 5. 7. This resulted in an increase in the average sequence coverage to 35% for 2D and 31.5% for the 1D experiment respectively and the average number of peptides/protein identified increased to 42.2 and 18.3, Table 5. 1**.** The average score/protein increased from 6,347 to 11,035 with the incorporation of additional chromatographic separation utilising the >1 analysis filter. The FDR for all six 2D experiments was 1.7% using the >1 analysis filter and for the 1D analyses it was 0.95%, compared to over 29% with no filtering. If the one hit wonder proteins with a score >1000 from the 1D pooled analyses were included to increase the number of proteins identified from this dataset this did not result in additional random proteins being identified.

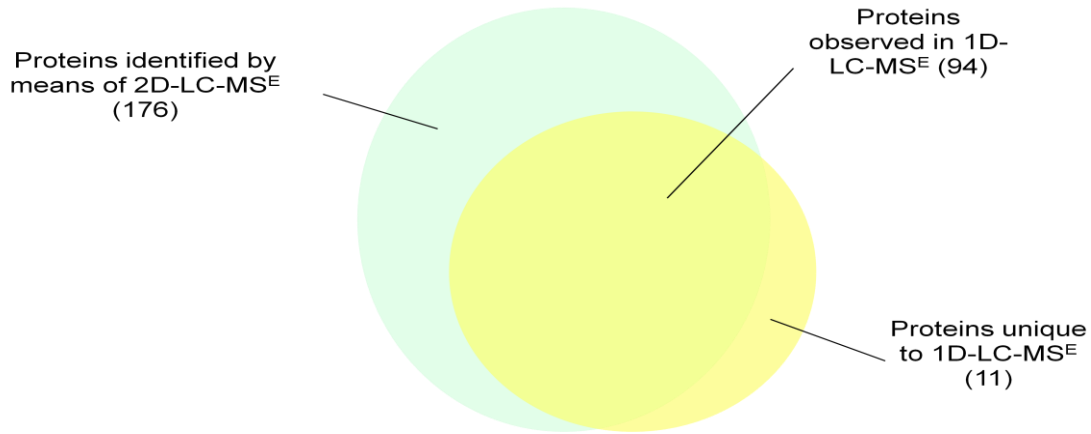


Figure 5. 7: Venn diagram indicating proteins identified from pooled depleted plasma using single and multi-dimensional chromatography.
Protein identifications from each approach were identified in at least 1 analysis from either normal or trisomy 21 plasma.

A protein identified from 1D but absent from the 2D confident protein list does not exclude it from being a plasma protein, but this aided in the assessment of a suitable filter for the data obtained from LC-MS^E experiments. Patel and co-workers identified a subset of peptides and proteins that were not identified in the multi-dimensional approaches and suggested that in 2D separations peptides may elute to waste during the extended trapping times utilised in 2D separations (Patel, Crombie

et al. 2012). These are employed to elute a fraction of peptides from the 1st dimension column, whilst reducing the concentration of acetonitrile in the eluate (through combination with aqueous 0.1% buffer B from the 2nd dimension binary solvent manager) prior to application to the trapping column. Each step may take 10-15 minutes to complete compared to a few minutes for a 1D separation.

In order to fully validate a suitable filter for the data, further 2D-LC-MS^E experiments could be performed utilising a significantly higher sample load than described here. Typically 2.5-3.0 µg would be used for a 6-fraction chromatographic separation or 5 µg for a 10-fraction experiment. This would yield significantly increased numbers of proteins identified confidently that could be compared with the data sets previously collected. The data could also then be used to assess if one hit wonder proteins could be acceptable identifications if a suitable average score was obtained, and the effect this would have on the false discovery rate for the dataset determined.

5.3.3 Identification of trisomy 21 biomarker proteins from pooled depleted plasma

PAPP-A, inhibin A and AFP were not identified in pooled normal or T21 depleted plasma analysed using 1D or 2D-LC-MS^E. β-hCG was identified in both normal and T21 pooled samples separated using 2D chromatography with an average of 8 peptides observed in all three replicates and sequence coverage >30%, although no measure of abundance was reported by PLGS.

5.3.4 Relative plasma protein levels analysed using 2D-LC-MS^E

Relative protein expression analysis was performed to compare the protein levels in normal and T21 depleted pooled plasma. Differences were determined using the PLGS Expression algorithm built into PLGS v2.5.1, with the output filtered for >1 replicate, score >1000 and regulation likelihood greater than 95%.

The filtered PLGS protein expression analysis table indicated that there were 8 proteins determined to be unique to the pooled normal plasma analysed using 2D-

LC-MS^E and 8 proteins unique to T21 pooled plasma. A number of the protein accession numbers have been archived in UniProtKB and in those cases the protein description/function has been assigned based on sequence homology.

A further 60 proteins were identified as occurring at differing levels between the two obstetric conditions, of which 24 were reported as varying by >1.5-fold,
Table 5. 3. A minimum expression cut off of 1.5-fold was selected in this work as the experimental error within the MS^E measurements has been shown to be <1.2-fold (Patel, Thalassinos et al. 2009) and one of the commercial biomarkers for trisomy 21, α-fetoprotein has an MoM value consistent with a 1.5-fold alteration in level. A number of the proteins identified were keratin contaminants, residual depleted proteins such as ceruloplasmin, fibrinogen, α-1-antitrypsin and apolipoprotein or immune response proteins. The internal standard, alcohol deydrogenase was reported as changing 1.68-fold between T21 and normal plasma. The actual difference in the internal standard concentration between the two samples was 1.58-fold (76 fmol on column compared to 48 fmol), providing high levels of confidence in the expression levels reported from the 2D approach.

Protein Description	Unique to pooled plasma
Fibrinogen gamma chain	Normal
Neutrophil defensin 1	Normal
Complement C3	Normal
Vitamin K-dependent protein C	Normal
Homology to Vitamin K-dependent protein S and Protein S	Normal
Clusterin	Normal
Lambda light chain of human immunoglobulin surface antigen-related protein	Normal
Inter-alpha-trypsin inhibitor heavy chain H3	Normal
Ig kappa chain C region	T21
Highly similar to Alpha-2-HS-glycoprotein	T21
Ig alpha-1 chain C region	T21
Inter-alpha-trypsin inhibitor heavy chain H1	T21
Highly similar to Hemopexin	T21
Kininogen-1	T21
Extracellular matrix protein 1	T21
Fibronectin	T21

Table 5. 2: Proteins identified from 2D-LC-MS^{E} analysis of pooled plasma as unique to an obstetric condition.
The protein expression results were filtered for >1 replicate, score >1000 and regulation likelihood greater than 95%.

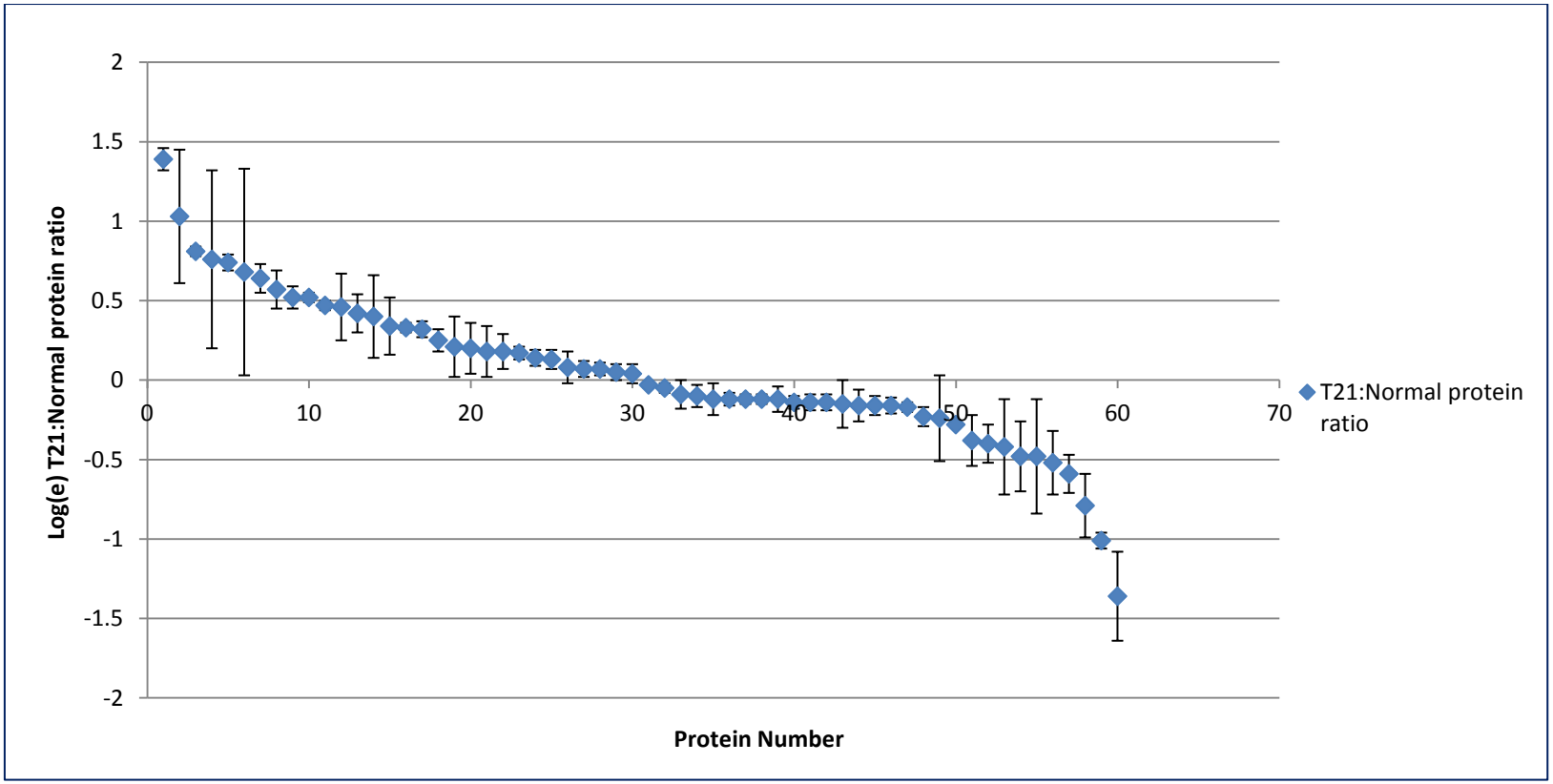


Figure 5. 8: Scatter plot of protein expression ratios derived from pooled trisomy 21 and normal plasma analysed using 2D-LC-MS^{E}.
Error bars indicate standard deviation as log(e) measurements.
The internal standard (protein number 27 on the plot) was calculated to have a 1.68-fold increase in level in the T21 sample using the protein expression analysis, in agreement with the 1.58-fold used.

Protein Description	Score	T21:Normal_Ratio	T21:Normal_Log(e)Ratio
Ceruloplasmin - similar to	104668.4	4.01	1.39
Putative pregnancy-specific beta-1-glycoprotein 7	1862.55	2.80	1.03
Fibrinogen beta chain	23027.28	2.25	0.81
26 kDa protein - Anti-RhD monoclonal T125 kappa light chain	4697.9	2.14	0.76
Fibrinogen alpha chain	22812.82	2.10	0.74
Complement component C6	4570.46	1.97	0.68
Alpha-1-antitrypsin	8914.89	1.90	0.64
Apolipoprotein B-48	1194.18	1.77	0.57
Alcohol dehydrogenase (internal standard)	8581.85	1.68	0.52
Fibronectin isoform	25036.07	1.68	0.52
Fibronectin isoform	25423.13	1.60	0.47
Glutathione peroxidase 3	2276.64	1.58	0.46
Sex hormone-binding globulin	12165.84	1.52	0.42
Complement C1q subcomponent subunit C	1213.32	1.49	0.4
Serum paraoxonase/arylesterase 1	15268.66	0.67	-0.4
Ig mu heavy chain disease protein	2279.62	0.66	-0.42
Keratin, type II cytoskeletal 1	1441.9	0.62	-0.48
PSG3 Pregnancy-specific beta-1-glycoprotein 3 precursor	1792.1	0.62	-0.48
Keratin, type I cytoskeletal 9	1983.91	0.59	-0.52
Inter-alpha-trypsin inhibitor heavy chain H4	30352.56	0.55	-0.59
A-2-macroglobulin	20041.75	0.45	-0.79
Ceruloplasmin	107550.1	0.36	-1.01
Pregnancy zone protein	8793.26	0.26	-1.36

Table 5. 3: Proteins identified at altered levels in trisomy 21 or normal pooled plasma analysed using 2D-LC-MS^{E}.

5.3.5 Relative plasma protein levels based on 1D-LC-MS^E expression analysis

In Chapter Four, the analysis of 29 individual plasma samples was described providing information on variation in protein abundance between patients and across the three obstetric groups of normal, unaffected pregnancies, trisomy 21 and pre-eclampsia. Protein abundance was estimated using the Hi3 approach in PLGS. In this section the results obtained from the individual samples were grouped by obstetric condition and analysed using the Expression algorithm in PLGS v2.5.1 and a comparison of normal and T21 protein levels made.

A number of immunoglobulin-like and immune response proteins have been identified as unique to both the normal and T21 obstetric groups. Vitamin K-dependent protein C was identified as unique to the T21 condition in contrast to the 2D-LC-MS^E expression analysis which suggested that the protein was unique to the normal pregnancy plasma sample.

Pregnancy specific beta 1 glycoproteins (PSGs) are believed to play immunomodulatory roles during pregnancy and are produced in high quantity during gestation. PSG3 was identified at 1.6-fold increased level in normal plasma, whereas PSG7 was suggested as a biomarker for the T21 condition with a 2.8-fold increase in level in T21 plasma from the 2D expression analysis. PSG3 and PSG7 share a significant number of tryptic peptides in common (90% sequence identity) and 6 of these were identified from both isoforms, however it was possible to distinguish between them for relative expression analysis as 8 peptides unique to PSG3 were observed and 9 unique to PSG7.

Two proteins from 1D individual expression analysis exhibited a >1.5-fold change in level between conditions after removal of depleted and immune response species. These were β-hCG and a neuroblastoma breakpoint family member. A 2.15-fold increase in level was reported for T21 plasma compared to normal in excellent agreement with the MoM value of 2.0 for β-hCG. Neuroblastoma breakpoint family members 10 and 14 were reported at elevated levels in T21 plasma but not observed in the 2D analysis.

Two proteins were observed at decreased levels in T21 individual plasma from the expression analysis using 1D, excluding depleted proteins. Both galectin-3-binding protein and retinol binding protein 4 plasma protein were determined to be present at 1.6-fold elevated levels in normal plasma but this observation was not confirmed using the 2D expression analysis.

5.3.6 Comparison of biomarkers for trisomy 21 obtained from different biological groups

Here, two sets of independent LC-MS^E analyses have been performed on different groups of patient plasma. Any useful predictive biomarkers for T21 should be apparent in both sets of data and follow the same trend, either unique to an obstetric condition or up- or down-regulated. Results presented in the previous two sections indicated that although each of the separate analyses identified potential biomarkers for trisomy 21, there was little overlap in the proteins they suggested.

In order to compare the results from each biological group (2D LC-MS^E using pooled plasma and 1D-LC-MS^E with individual patient plasma) those proteins identified at differing levels in T21 or normal plasma from the 2D analysis were selected. Then, for each of those proteins, the ratio of the mean abundance was *calculated* from the 1D-LC-MS^E data (T21 abundance/normal abundance, based on the Hi3 approach) and converted to a log(e) ratio,

Figure 5. 9. For the majority of the proteins identified as potential biomarkers for T21 by the 2D approach, the ratio result obtained from individual 1D analysis was either not observed or there was no correlation. A similar trend in T21/normal plasma ratio was observed for 22 of the proteins, although the correlation was weak in a number of cases.

Proteins including AMBP (α-1-microglobulin), gelsolin, angiotensinogen, afamin, attractin and antithrombin were identified at levels less than 1.3-fold different between obstetric conditions, making their potential use in a diagnostic test limited.

Excluding complement and apolipoproteins, eight predictive biomarkers for the trisomy 21 obstetric condition were identified from both the 2D pooled and 1D individual sample sets exhibiting similar trends in protein level, Table 5.4.

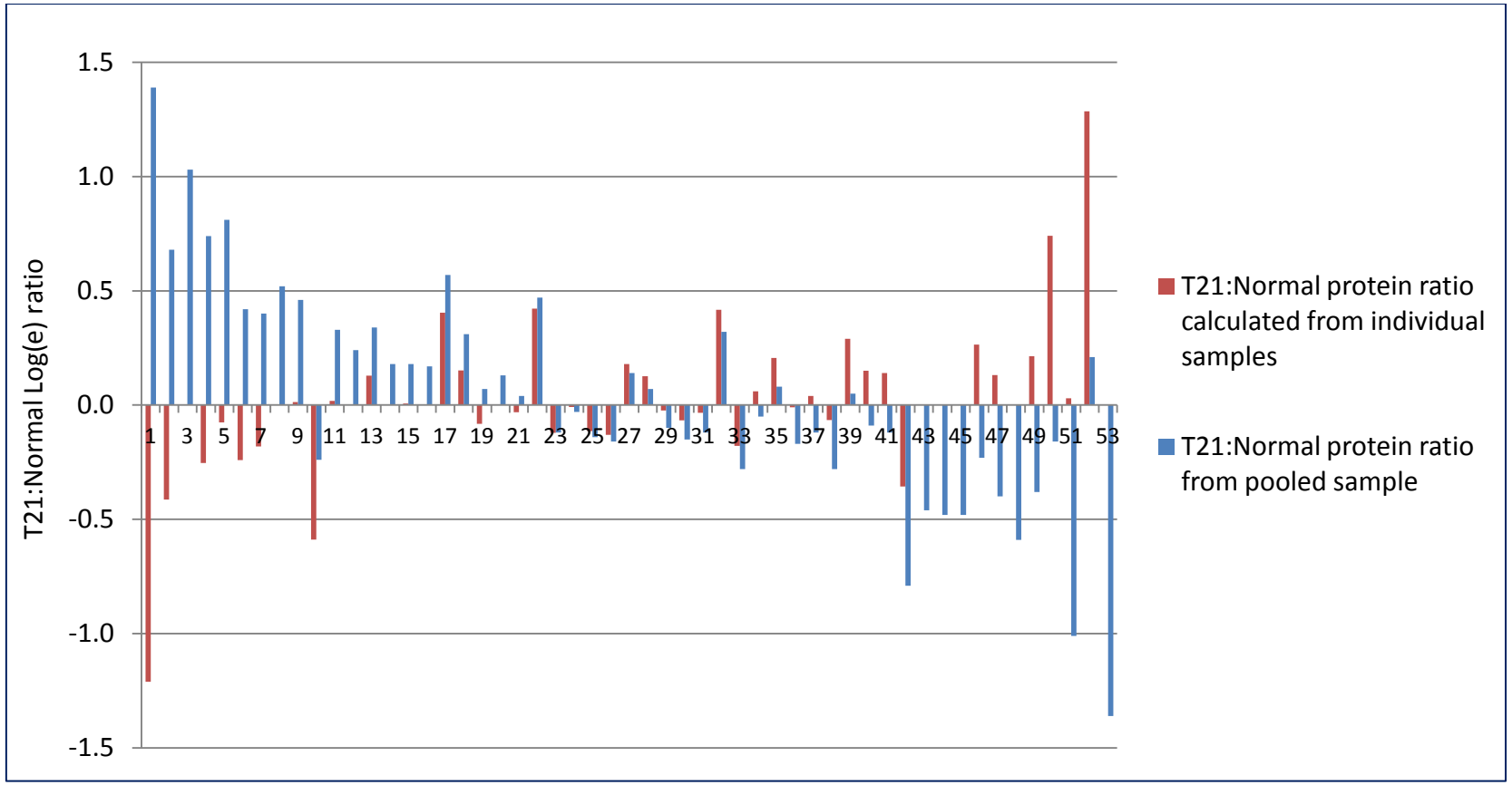


Figure 5. 9: Comparison of protein level ratios calculated from two sets of biological samples.
Red bars indicate the calculated abundance ratio of trisomy 21 to normal plasma protein levels from individual samples from 1D-LC-MS^{E} analysis
Blue bars indicate the protein level ratio determined by the PLGS v.2.4 Expression algorithm based on two pooled samples and 2D-LC-MS^{E} analysis.

Description	Abundance individual 1D T21:Normal_Log(e)Ratio	Pooled 2D T21:Normal_Log(e)Ratio
Fibronectin	0.42	0.47
Complement factor B	-0.12	-0.12
α-1-microglobulin (AMBP)	-0.11	-0.14
Gelsolin	-0.13	-0.16
Angiotensinogen	0.18	0.14
Antithrombin-III	0.13	0.07
Afamin	-0.02	-0.1
Carboxypeptidase N catalytic chain	-0.07	-0.15
Complement C1r subcomponent	-0.03	-0.12
Apolipoprotein A-IV	0.42	0.32
Attractin	-0.18	-0.28

Table 5. 4: Proteins identified at differing levels in trisomy 21 and normal plasma from two independent biological sample sets.
The 1D analysis was performed on individual samples, with the mean abundance ratio calculated for T21:normal based on Hi3 approach.
The 2D protein level ratios were determined by the PLGS v.2.4 Expression algorithm based on pooled obstetric samples and 2D-LC-MS^E analysis.

Fibronectin, a glycoprotein found in amniotic fluid and placental tissue was identified at 1.5-fold and 1.6-fold elevated levels in T21 plasma from 1D and 2D experiments respectively. It was identified as unique to T21 by PLGS v2.5.1. It is expressed at elevated levels both in early and late pregnancy and has been identified in a number of studies as a potential marker for pre-eclampsia, as such potentially lacking the specificity required by a T21 diagnostic test (Pignot and Busine 1989; Paternoster, Stella et al. 1996; Shaarawy and Didy 1996; Bodova, Biringer et al. 2011). The level of vaginal fibronectin has also been suggested as a predictive marker for spontaneous preterm birth particularly during the latter stages of pregnancy (Honest, Bachmann et al. 2002) and a commercial test is available with a high degree of accuracy in predicting birth within 7-10 days.

5.3.7 Limit of protein identification

Using the >1 analysis filter described previously the 2D pooled protein identification results were evaluated to determine the limit of protein identification in the dataset. A total of 22 proteins (12.5%) were identified at an average level of <5 fmol on column and 48 (27.3%) between 5 and 10 fmol on column. For the pooled 1D dataset 3 proteins (2.9%) were identified at <5 fmol on column and 5 (4.8%) between 5 and 10 fmol. In this study the sample loading for the 2D experiments was very similar to that used for 1D, so it may not be expected that significantly lower abundance proteins would be observed, but this study does indicate that the additional separation of peptides using the 2D approach enabled more low abundance proteins to be identified in any given sample.

The study of the plasma proteome over the last forty years has changed dramatically since publication of the first volume of *The Plasma Proteins* (Putnam 1960). The number of proteins identified by proteomic approaches has continued to increase from 289 (Anderson and Anderson 2002) to 1929 confident assignments based on 20,433 peptides with a FDR of 1% and 0.16% at the protein and peptide level respectively (Farrah, Deutsch et al. 2011).

The HUPO Human Plasma Proteome Project (HPPP) was established in 2002 creating a human plasma proteome database based on the results from 55 laboratories across the world. Protocols included combinations of depletion, fractionation, with mass spectrometry and immunoassay. Now in its second phase, the HPPP includes very large data sets compiled using standardised analysis (TransProteomic Pipeline) creating the Human Plasma PeptideAtlas.

In contrast to this study where a number of samples were characterised and therefore extensive fractionation was not feasible, in many cases the objective of the proteomic analysis is to identify as many proteins as possible from a given sample. Combinations of approaches have been employed to reach this goal. Millioni and co-workers utilised a four-dimension approach to the analysis of plasma including combinatorial peptide ligand library enrichment and IEF of the tryptic peptides using IPG strips which were subsequently cut into 8 pieces. SCX fractionation was then

employed (4 fractions each giving a total of 32 fractions from IEF-SCX) with RP-LC-MS/MS analysis (Millioni, Tolin et al. 2012). In total, 417 proteins were identified with at least two peptides by this approach. The authors reported identifying low abundance proteins including angiogenin (10^{-9} g L^{-1}), pigment epithelium growth factor (10^{-8} g L^{-1}) and thrombospondin (10^{-6} g L^{-1}) in plasma. Angiogenin and thrombospondin were not identified here, but pigment epithelium growth factor (PEDF) was observed in the pooled experiments and in 83 of the 87 analyses from the individual samples. Relative expression analysis utilising all curated peptides from the protein in each analysis indicated a 1.2-fold increase in the PET group compared to normal, unaffected gestations. Fold-changes of 1.04 and 0.85 were determined for PEDF in T21 compared to normal and T21 compared to PET groups respectively. PEDF is expressed at high levels early during gestation and down-regulation of PEDF has been suggested to alter the placental vasculature with heightened angiogenesis, contributing to adverse perinatal outcomes, such as stillbirth (Plunkett, Fitchev et al. 2008) and the pathogenesis of ovarian endometriosis (Huang, Chen et al. 2012).

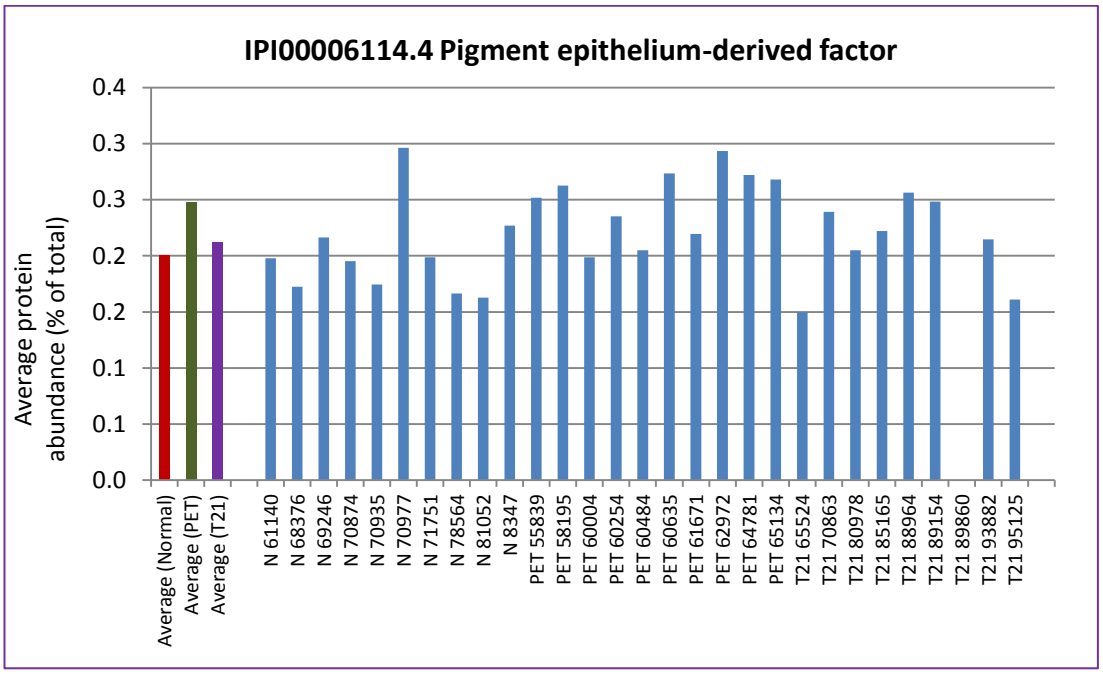


Figure 5. 10: Abundance of pigment epithelium growth factor in IgY-14 LC2 depleted plasma.
Abundance calculated as a percentage of the total identified protein content based on ng levels reported by the Hi3 approach in PLGS.

A selection of 10 proteins identified in >1 2D analysis at <10 fmol on column in this study, were compared with the plasma proteins reported with high confidence in PeptideAtlas (Farrah, Deutsch et al. 2011). All the proteins were recorded in PeptideAtlas spanning the molecular weight and concentration ranges 22 kDa - 92 kDa and 20 ng mL^{-1} – 58 µg mL^{-1} respectively, with eight estimated to have a plasma concentration between 140 and 620 ng mL^{-1} in non-maternal plasma, Table 5. 5.

The estimated concentration (in ng mL^{-1}) of 1929 high confidence plasma proteins reported by Farrah, based on spectral counting is shown in
Figure 5. 11. The results were derived from 91 high quality LC-MS/MS data sets encompassing a variety of sample preparation approaches (depleted and non-depleted), fractionation methods and enrichment techniques and indicate a dynamic range of ~6 orders. Highlighted by a green circle, the most abundant protein in this study, ceruloplasmin (not depleted by the IgY-14 approach) is indicated as well as low abundance identified proteins in the red oval, detailed in Table 5. 5. The 2D approach used here demonstrated a dynamic range in excess of 3 orders and with an increased sample loading a 4-order dynamic range would be achievable, in agreement with the observations of Patel who employed an 11-fraction 2D RP-RP approach to increase the dynamic range of the analyses (Patel, Crombie et al. 2012). The concentrations of the predictive biomarkers for trisomy 21 have been indicated in Figure 5.11 with yellow markers, but were not identified by Farrah and co-workers.

Greater probing of the plasma proteome comes at a price of both experimental and data processing/storage costs. Three technical replicates of a biological sample analysed using a 6-fraction RP-RP-LC-MS^E approach, takes many hours to collect and process the data, even when utilising a graphics processing unit for this purpose and occupies over 50 Gb of storage space. After the results have been generated the output has to be further manipulated in Excel or other suitable programs. Newly developed software such as TransOmics (Waters Corporation, MA, USA) and Progenesis LC-MS, both powered by Nonlinear Dynamics algorithms (Newcastle upon Tyne, UK) aim to integrate and streamline both the identification and relative quantitation of proteomics data into a single package, with statistical analysis built-in.

Protein conflicts can be resolved within the software packages, comparisons at both the peptide and protein level can be achieved and visualised with relative ease, ensuring access to *biologically relevant* information.

5.4 Conclusions

Two pooled samples have been prepared from normal, unaffected and trisomy 21 maternal plasma (n=20) and depleted of 14 highly abundant proteins. The pooled samples were analysed using both a single and multi-dimensional chromatography approach using RP and high/low pH RP-RP separations using similar sample loadings. RP-RP outperformed RP in terms of peptide, protein and sequence coverage due to the additional separation space afforded by the multi-dimensional approach. Although 2D analyses were not performed at an optimal sample level, more proteins were identified at low fmol level on-column.

The confident protein output from the 2D experiments was used to determine a suitable filter for use in further studies. Observation in two replicates significantly reduced the number of proteins identified, so a filter based on observation across ≥2 analyses was assessed, resulting in an increased number of identifications whilst the FDR remained <1.5%. For single protein identifications from the pooled 1D data set, the use of score as a filter was evaluated. In the PLGS software v2.5.1, a score of 1000 was deemed to be a threshold which allowed additional proteins to be identified but not at the expense of an increased FDR. This would need to be further evaluated on a larger data set before implementation.

Of the four proteins used for the predictive assay for trisomy 21, only β-hCG was identified in the pooled 2D experiments. β-hCG, inhibin A, AFP and PAPP-A were not reported in plasma by Millioni (Millioni, Tolin et al. 2012) using a four-dimensional approach to the analysis of human plasma or by Farrah despite the increased depth of their analyses, which in the latter study exceeded 6 orders (Farrah, Deutsch et al. 2011).

UniProtKB Accession	Description	Molecular Weight (kDa)	Estimated plasma concentration * (ng mL^{-1})	Sequence Coverage * (%)	Concentration from 2D experiments (fmol on-column) **	Sequence coverage from 2D experiments (%) **
P23142	Isoform of Fibulin-1	77.2	620	66	0.6	31
O00391	Sulfhydryl oxidase 1	82.6	150	56	0.9	14
P05452	Tetranectin	22.5	58000	92	1.1	46
P43251	Biotinidase	61.1	490	52	4.6	19
P13727	Bone marrow proteoglycan	25.2	20	52	4.9	38
P08571	Monocyte differentiation antigen CD14	40.1	420	71	5.3	11
Q15582	Transforming growth factor-beta-induced protein ig-h3	74.7	140	56	2.0	21
P26927	Hepatocyte growth factor-like protein	80.3	250	49	6.7	26
P80108	Phosphatidylinositol-glycan-specific phospholipase D	92.3	460	61	6.7	15
Q08380	Galectin-3-binding protein	65.3	440	64	9.7	22

Table 5. 5: Low abundance proteins identified from 2D pooled depleted maternal plasma.
Proteins identified in >1 analysis from normal and trisomy 21 depleted plasma analysed using 2D RP-RP LC-MS^{E}.
* Estimated plasma concentration, based on spectral counting and sequence coverage taken from (Farrah, Deutsch et al. 2011).
** Estimated values for plasma concentration, based on Hi3 approach and sequence coverage were based on all observations in normal and T21 plasma.

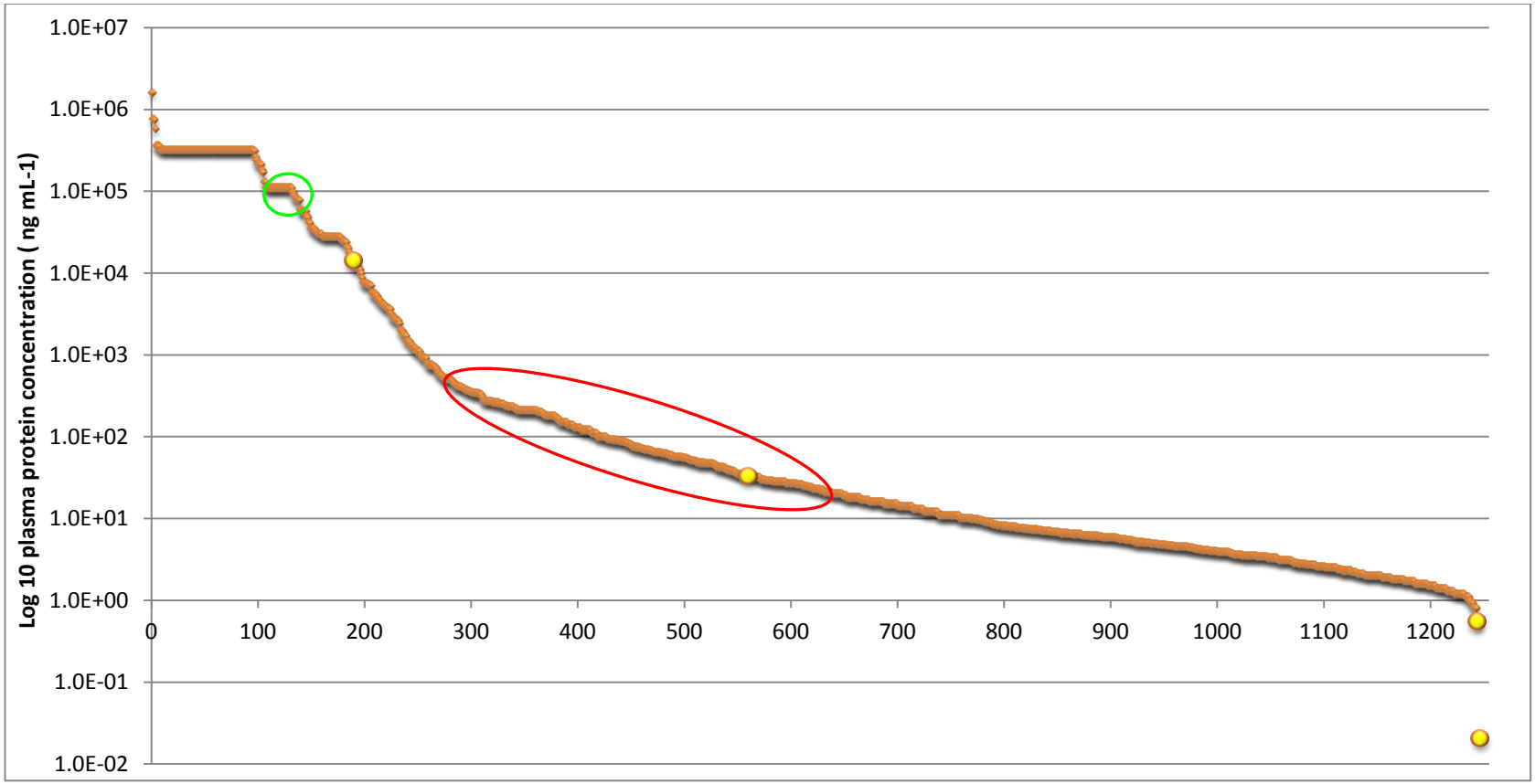


Figure 5. 11: Estimated plasma protein concentration based on spectral counting measurements.
Taken from (Farrah, Deutsch et al. 2011).
Ceruloplasmin, green circle, indicates the most abundant, non-depleted protein identified in the 2D experiments in this study. Red oval indicates the low abundance proteins identified in >1 2D analysis from normal and T21 plasma, demonstrating a dynamic range over 3 orders of magnitude. First trimester concentrations of PAPP-A, β-hCG, inhibin A and AFP (yellow markers) are shown, but were not identified by Farrah *et al.*

The relative quantitative analysis of maternal plasma either as individual samples or pooled, indicated the high degree of confidence in the output. The internal standard in the 2D experiments and the β-hCG level in the individual 1D analyses were correctly identified at their respective levels (CVs 6% and 7.5% respectively).

Relative protein analysis of each set of biological samples resulted in a number of proteins identified as unique by the PLGS software i.e. above the detection limit in one obstetric condition. Additional proteins were identified at differing but statistically significant levels. A comparison of these potential biomarker proteins indicated that the majority did not correlate across the biological groups. A small number of proteins were observed at similar levels between obstetric and biological groups but were below that deemed to be biologically significant.

Fibronectin was identified as a potential biomarker for trisomy 21 at a >1.5-fold increased level. The level of fibronectin in first trimester maternal plasma would need to be fully evaluated as the protein has been implicated in a number of other obstetric conditions, which may affect its use diagnostically. Its use in combination with the existing predictive markers would need to be assessed.

The limit of confident protein identification has been established at low fmol levels on-column evaluated using a combination of the confidently identified proteins from the 2D RP-RP approach to assess suitable thresholds for protein acceptance.

The dynamic range of the experiments conducted exceeded 3 orders and with optimal sample loading a 4 order range could be achievable using multi-dimensional RP-RP-LC-MS^E approaches. Further depletion and/or fractionation would be required to achieve identification over a greater dynamic range.

Chapter Six: Conclusions and future directions

6.1 Conclusions

Successful biomarker studies require the acquisition of high confidence protein identification and quantification data. Data independent acquisition, here exemplified by the MS^E experiment, has been shown to yield higher information content than data dependent approaches. This has been demonstrated by the growing number of publications since its first application in 2006 (Silva, Denny et al. 2006). Other commercial data independent developments include SWATH acquisition (AB Sciex, Foster City, CA, USA) (Gillet, Navarro et al. 2012) and instrument control software for the Q-Exactive (Thermo Fisher Scientific Inc., MA, USA) permitting DIA.

This research has demonstrated the applicability of using an LC-MS^E approach for discovery biomarker studies. Multi-dimensional chromatography, with its increased peak capacity provided more confident protein identifications and permitted the identification of an increased number of low abundance proteins. With optimised sample loading and increased fractionation, the dynamic range of the proteomics experiment can be extended, at the expense of experimental time and effort. With its orthogonal separation capability RP-RP high/low pH chromatography offers significant advantages over other multi-dimensional approaches.

Relaxing the stringency of filtering results from large scale proteomic projects can yield increased numbers of proteins identifications with a minimal increase in the false discovery rate. The identification of alternate isoforms from large eukaryotic databases however continues to be problematic.

Depletion of highly abundant proteins from plasma is essential to obtain the dynamic range required to identify biomarkers of clinical relevance. The partitioning strategy must deplete *a significant number* of abundant proteins in a robust and reproducible manner. The efficiency should be monitored to ensure consistency throughout a study. MS^E acquisition is ideally suited for this purpose as information may be obtained for each of the depleted proteins in each analysis.

Use of pooled samples should be avoided where possible in biomarker discovery studies, since information on the natural variation of plasma protein concentration is lost. This information can be used to calculate risk factors from biomarker assays.

A comparison of two conditions (e.g. diseased and control) will identify proteins that differ in concentration. This research has demonstrated that the majority of the potential biomarkers identified, in this case for two obstetric conditions, could not be validated when compared across different biological groups or individuals. Additional experimental effort may be required at the discovery stage of a biomarker project to ensure that a smaller number of predictive markers are quantified with a higher degree of confidence prior to a clinical validation step.

A potential predictive biomarker for trisomy 21, fibronectin, was identified in first trimester maternal plasma from multiple biological groups but requires further validation in combination with existing protein markers used in commercial tests. The protein has been implicated in other obstetric conditions and so would need to be further assessed in terms of its specificity for the condition.

6.2 Future directions

In excess of 2 Tb of quantitative data have been collected during this research. The results presented here form a small subset of the total biological information contained. Without the need for further data collection, the data could be further interrogated for the following: -

- Variation in the concentration of peptides between conditions, rather than the protein level used here.
- Altered post-translational modification states, particularly glycosylation which has been implicated in a number of disease conditions (Hall, Cawdell et al. 1983; Schrader, Jovanovic-Peterson et al. 1995; Matei 1997).
- Effect of relaxing the data processing parameters and protein identification thresholds.

- Identification of potential biomarkers for pre-eclampsia.

Probing the plasma proteome for proteins present below the ng mL^{-1} level requires extensive depletion and multi-dimensional fractionation based on the dynamic range capacity of modern instruments. The experimental effort required to complete this task on a large number of individual plasma samples is significant and onerous. Creation of data repositories, such as the data independent precursor and product ion relational database (Thalassinos, Vissers et al. 2012), a collection of high quality data which can be queried in a targeted manner with either raw or processed data, allows data information collected from multiple groups to be queried. PRIDE, PRoteomics IDEntifications database is a centralised, public repository for proteomics data, which includes protein and peptide identifications, post-translational modifications with supporting spectral evidence.

An ion mobility enabled, or assisted, data independent LC-MS^E approach shows significant promise for increased proteins and peptide identification from complex samples (Rodriguez-Suarez, Hughes et al. 2013). Protein and peptide identification increases by 160% for 1D and over 260% for 2D separations were reported when combined with ion mobility separation. No additional experimental effort is required to collect the ion mobility enabled data but data processing demands increase significantly. A 7.5-fold increase in protein identification rate from rat exosome was reported for the lower one-third of the dynamic range (low fmol level).

In all omic studies, the challenge lies not only in the collection of significant amounts of high quality data, but also from inferring the biological significance contained within this information. Continued development of bioinformatics software is essential to interrogate and integrate content from omic sources and, in combination with developments in mass spectrometry instrumentation, will be essential to allow proteomics research to continue to evolve over the coming decades.

References

Aebersold, R. and M. Mann (2003). Mass spectrometry-based proteomics. *Nature.* **422**, 198-207.

Ahmed, N., K. T. Oliva, G. Barker, P. Hoffmann, S. Reeve, I. A. Smith, M. A. Quinn and G. E. Rice (2005). Proteomic tracking of serum protein isoforms as screening biomarkers of ovarian cancer. *Proteomics.* **5**, 4625-4636.

Aitken, D. A., E. M. Wallace, J. A. Crossley, I. A. Swanston, Y. van Pareren, M. van Maarle, N. P. Groome, J. N. Macri and J. M. Connor (1996). Dimeric Inhibin A as a Marker for Down's Syndrome in Early Pregnancy. *New England Journal of Medicine.* **334**, 1231-1236.

Anderson, L. and C. L. Hunter (2006). Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol Cell Proteomics.* **5**, 573-88.

Anderson, N. L. (2010). The Clinical Plasma Proteome: A Survey of Clinical Assays for Proteins in Plasma and Serum. *Clinical Chemistry.* **56**, 177-185.

Anderson, N. L. and N. G. Anderson (2002). The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics.* **1**, 845-67.

Andren, P. E., M. R. Emmett and R. M. Caprioli (1994). Micro-electrospray: zeptomole/attomole per microliter sensitivity for peptides. *Journal of the American Society for Mass Spectrometry.* **5**, 867-869.

Aston, F. W. (1919). A positive ray spectrograph. *Philosophical Magazine.* **38**, 707-714.

Atkinson, A. J., W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, R. T. Schooley, B. A. Spilker, J. Woodcock and S. L. Zeger (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework*. *Clin Pharmacol Ther.* **69**, 89-95.

Barber, M., R. S. Bordoli, R. D. Sedgwick and A. N. Tyler (1981). Fast atom bombardment of solids (F.A.B.): a new ion source for mass spectrometry. *Journal of the Chemical Society, Chemical Communications.* 325-327.

Bateman, R. H., R. Carruthers, J. B. Hoyes, C. Jones, J. I. Langridge, A. Millar and J. P. Vissers (2002). A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation. *J Am Soc Mass Spectrom.* **13**, 792-803.

Bateman, R. H., M. R. Green, G. Scott and E. Clayton (1995). A combined magnetic sector–time-of-flight mass spectrometer for structural determination studies by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry.* **9**, 1227-1233.

Berven, F. S., R. Ahmad, K. R. Clauser and S. A. Carr (2010). Optimizing performance of glycopeptide capture for plasma proteomics. *J Proteome Res.* **9**, 1706-15.

Bestwick, J. P., W. J. Huttly and N. J. Wald (2010). Distribution of nuchal translucency in antenatal screening for Down's syndrome. *J Med Screen.* **17**, 8-12.

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry.* **72**, 291-336.

Blakeley, P., J. A. Siepen, C. Lawless and S. J. Hubbard (2010). Investigating protein isoforms via proteomics: a feasibility study. *Proteomics.* **10**, 1127-40.

Blankenhorn, D., J. Phillips and J. L. Slonczewski (1999). Acid- and Base-Induced Proteins during Aerobic and Anaerobic Growth of Escherichia coli Revealed by Two-Dimensional Gel Electrophoresis. *J Bacteriology.* **181**, 2209–2216.

Bodenmiller, B., L. N. Mueller, M. Mueller, B. Domon and R. Aebersold (2007). Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat Meth.* **4**, 231-237.

Bodova, K. B., K. Biringer, K. Dokus, J. Ivankova, J. Stasko and J. Danko (2011). Fibronectin, plasminogen activator inhibitor type 1 (PAI-1) and uterine artery Doppler velocimetry as markers of preeclampsia. *Dis Markers.* **30**, 191-6.

Bogdanov, B. and R. D. Smith (2005). Proteomics by FTICR mass spectrometry: Top down and bottom up. *Mass Spectrometry Reviews.* **24**, 168-200.

Bondarenko, P. V., D. Chelius and T. A. Shaler (2002). Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem.* **74**, 4741-9.

Brand, J., T. Haslberger, W. Zolg, G. Pestlin and S. Palme (2006). Depletion efficiency and recovery of trace markers from a multiparameter immunodepletion column. *Proteomics.* **6**, 3236-42.

Breuker, K., M. Jin, X. Han, H. Jiang and F. W. McLafferty (2008). Top-down identification and characterization of biomolecules by mass spectrometry. *J Am Soc Mass Spectrom.* **19**, 1045-53.

Breuker, K. and F. W. McLafferty (2003). Native electron capture dissociation for the structural characterization of noncovalent interactions in native cytochrome C. *Angew Chem Int Ed Engl.* **42**, 4900-4.

Caffrey, R. (2010). A Review of Experimental Design Best Practices for Proteomics Based Biomarker Discovery: Focus on SELDI-TOF. The Urinary Proteome. A. J. Rai, Humana Press. **641:** 167-183.

Campuzano, I., J. Brown, J. Williams and K. Compson (2010). The Implementation and Characterization of Electron Transfer Dissociation (ETD) on an Ion Mobility Enabled Q-TOF Mass Spectrometer. *J. Biomol. Tech.* **21**, S36.

Cannataro, M. (2008). Computational proteomics: management and analysis of proteomics data. *Brief Bioinform.* **9**, 97-101.

Cargile, B. J., J. L. Bundy and J. L. Stephenson, Jr. (2004). Potential for false positive identifications from large databases through tandem mass spectrometry. *J Proteome Res.* **3**, 1082-5.

Carlsen, S. M., P. Romundstad and G. Jacobsen (2005). Early second-trimester maternal hyperandrogenemia and subsequent preeclampsia: a prospective study. *Acta Obstet Gynecol Scand.* **84**, 117-21.

Carty, D. M., C. Delles and A. F. Dominiczak (2008). Novel biomarkers for predicting preeclampsia. *Trends Cardiovasc Med.* **18**, 186-94.

Catherman, A. D., M. X. Li, J. C. Tran, K. R. Durbin, P. D. Compton, B. P. Early, P. M. Thomas and N. L. Kelleher (2013). Top Down Proteomics of Human Membrane Proteins from Enriched Mitochondrial Fractions. *Analytical Chemistry.* **85**, 1880-1888.

Chait, B. T. (2006). Chemistry. Mass spectrometry: bottom-up or top-down? *Science.* **314**, 65-6.

Chelius, D. and P. V. Bondarenko (2002). Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res.* **1**, 317-23.

Chen, E. I., D. Cociorva, J. L. Norris and J. R. Yates, 3rd (2007). Optimization of mass spectrometry-compatible surfactants for shotgun proteomics. *J Proteome Res.* **6**, 2529-38.

Cho, C. K., S. J. Shan, E. J. Winsor and E. P. Diamandis (2007). Proteomics analysis of human amniotic fluid. *Mol Cell Proteomics.* **6**, 1406-15.

Coles, J. and M. Guilhaus (1993). Orthogonal acceleration — a new direction for time-of-flight mass spectrometry: Fast, sensitive mass analysis for continuous ion sources. *TrAC Trends in Analytical Chemistry.* **12**, 203-213.

Comisarow, M. B. and A. G. Marshall (1974). Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physics Letters.* **25**, 282-283.

Conde-Agudelo, A., J. Villar and M. Lindheimer (2004). World Health Organization systematic review of screening tests for preeclampsia. *Obstet Gynecol.* **104**, 1367-91.

Cooks, R. G., G. L. Glish, S. A. McLuckey and R. E. Kaiser (1991). Ion Trap Mass Spectrometry. *Chemical & Engineering News.* **69**, 26-41.

Cuckle, H. S., N. J. Wald and S. G. Thompson (1987). Estimating a woman's risk of having a pregnancy associated with Down's syndrome using her age and serum alpha-fetoprotein level. *Br J Obstet Gynaecol.* **94**, 387-402.

Dasari, S., L. Pereira, A. P. Reddy, J. E. Michaels, X. Lu, T. Jacob, A. Thomas, M. Rodland, C. T. Roberts, Jr., M. G. Gravett and S. R. Nagalla (2007). Comprehensive proteomic analysis of human cervical-vaginal fluid. *J Proteome Res.* **6**, 1258-68.

Dawson, J. H. J. and M. Guilhaus (1989). Orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Communications in Mass Spectrometry.* **3**, 155-159.

de Hoffmann, E. and V. Stroobant (2007). Mass Spectrometry: Principles and Applications, 3rd Edition, Wiley.

DeSouza, L. V., J. Grigull, S. Ghanny, V. Dube, A. D. Romaschin, T. J. Colgan and K. W. Siu (2007). Endometrial carcinoma biomarker discovery and verification using differentially tagged clinical samples with multidimensional liquid chromatography and tandem mass spectrometry. *Mol Cell Proteomics.* **6**, 1170-82.

DeSouza, L. V., A. D. Romaschin, T. J. Colgan and K. W. Siu (2009). Absolute quantification of potential cancer markers in clinical tissue homogenates using multiple reaction monitoring on a hybrid triple quadrupole/linear ion trap tandem mass spectrometer. *Anal Chem.* **81**, 3462-70.

Deutsch, E. W., J. K. Eng, H. Zhang, N. L. King, A. I. Nesvizhskii, B. Lin, H. Lee, E. C. Yi, R. Ossola and R. Aebersold (2005). Human Plasma PeptideAtlas. *Proteomics.* **5**, 3497-500.

Di Lorenzo, G., M. Ceccarello, V. Cecotti, L. Ronfani, L. Monasta, L. Vecchi Brumatti, M. Montico and G. D'Ottavio (2012). First trimester maternal serum PIGF, free beta-hCG, PAPP-A, PP-13, uterine artery Doppler and maternal history for the prediction of preeclampsia. *Placenta.* **33**, 495-501.

Di Quinzio, M. K., K. Oliva, S. J. Holdsworth, M. Ayhan, S. P. Walker, G. E. Rice, H. M. Georgiou and M. Permezel (2007). Proteomic analysis and characterisation of human cervico-vaginal fluid proteins. *Aust N Z J Obstet Gynaecol.* **47**, 9-15.

Dodonov, A. F., V. I. Kozlovski, I. V. Soulimenkov, V. V. Raznikov, A. V. Loboda, Z. Zhen, T. Horwath and H. Wollnik (2000). High-resolution electrospray ionization orthogonal-injection time-of-flight mass spectrometer. *European Journal of Mass Spectrometry.* **6**, 481-490.

Domon, B. and R. Aebersold (2006). Mass spectrometry and protein analysis. *Science.* **312**, 212-7.

Donato, P., F. Cacciola, L. Mondello and P. Dugo (2011). Comprehensive chromatographic separations in proteomics. *Journal of Chromatography A.* **1218**, 8777-8790.

Dong, M. Q., J. D. Venable, N. Au, T. Xu, S. K. Park, D. Cociorva, J. R. Johnson, A. Dillin and J. R. Yates, 3rd (2007). Quantitative mass spectrometry identifies insulin signaling targets in C. elegans. *Science.* **317**, 660-3.

Durand, G. and N. Seta (2000). Protein glycosylation and diseases: blood and urinary oligosaccharides as markers for diagnosis and therapeutic monitoring. *Clin Chem.* **46**, 795-805.

Eddes, J. S., E. A. Kapp, D. F. Frecklington, L. M. Connolly, M. J. Layton, R. L. Moritz and R. J. Simpson (2002). CHOMPER: a bioinformatic tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies. *Proteomics.* **2**, 1097-103.

Elias, J. E., W. Haas, B. K. Faherty and S. P. Gygi (2005). Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Meth.* **2**, 667-675.

Emmett, M. R. and R. M. Caprioli (1994). Micro-electrospray mass spectrometry: ultra-high-sensitivity analysis of peptides and proteins. *Journal of the American Society for Mass Spectrometry.* **5**, 605-613.

Eng, J. K., A. L. McCormack and J. R. Yates Iii (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry.* **5**, 976-989.

Falick, A., W. Hines, K. Medzihradszky, M. Baldwin and B. Gibson (1993). Low-mass ions produced from peptides by high-energy collision-induced dissociation in tandem mass spectrometry. *Journal of the American Society for Mass Spectrometry.* **4**, 882-893.

Farrah, T., E. W. Deutsch, G. S. Omenn, D. S. Campbell, Z. Sun, J. A. Bletz, P. Mallick, J. E. Katz, J. Malmstrom, R. Ossola, J. D. Watts, B. Lin, H. Zhang, R. L. Moritz and R. Aebersold (2011). A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol Cell Proteomics.* **10**, M110 006353.

Felitsyn, N., M. Peschke and P. Kebarle (2002). Origin and number of charges observed on multiply-protonated native proteins produced by ESI. *International Journal of Mass Spectrometry.* **219**, 39-62.

Fenn, J. B., M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science.* **246**, 64-71.

Fernandez de la Mora, J. (2000). Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. *Analytica Chimica Acta.* **406**, 93-104.

Fila, J. and D. Honys (2011). Enrichment techniques employed in phosphoproteomics. *Amino Acids.*

Florens, L., M. J. Carozza, S. K. Swanson, M. Fournier, M. K. Coleman, J. L. Workman and M. P. Washburn (2006). Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods.* **40**, 303-11.

Foster, L. J., C. L. de Hoog and M. Mann (2003). Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors. *Proceedings of the National Academy of Sciences.* **100**, 5813-5818.

Freeze, H. H. (2001). Update and perspectives on congenital disorders of glycosylation. *Glycobiology.* **11**, 129R-143R.

Fusaro, V. A., D. R. Mani, J. P. Mesirov and S. A. Carr (2009). Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat Biotechnol.* **27**, 190-8.

G.J.O'Halloran, R.A.Fluegge, J. F. Betts and W.J.Everett (1964). Determination of chemical species prevalent in a plasma jet., Bendix Corporation Research Laboratories Division, South®eld Michigan under Contract Nos AF33(616)-8374 and AF33(657)-11018. A.F. Materials Laboratory Research and Technology Division Air Force Systems Command.

Garavelli, J. S. (2004). The RESID Database of Protein Modifications as a resource and annotation tool. *Proteomics.* **4**, 1527-1533.

Ge, Y., I. N. Rybakova, Q. Xu and R. L. Moss (2009). Top-down high-resolution mass spectrometry of cardiac myosin binding protein C revealed that truncation alters protein phosphorylation state. *Proceedings of the National Academy of Sciences.* **106**, 12658-12663.

Gerber, S. A., J. Rush, O. Stemman, M. W. Kirschner and S. P. Gygi (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A.* **100**, 6940-5.

Geromanos, S. J., J. P. C. Vissers, J. C. Silva, C. A. Dorschel, G.-Z. Li, M. V. Gorenstein, R. H. Bateman and J. I. Langridge (2009). The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics.* **9**, 1683-1695.

Ghrist, B. F., M. A. Stadalius and L. R. Snyder (1987). Predicting bandwidth in the high-performance liquid chromatographic separation of large biomolecules. I. Size-exclusion studies and the role of solute stokes diameter versus particle pore diameter. *J Chromatogr.* **387**, 1-19.

Gilar, M., A. E. Daly, M. Kele, U. D. Neue and J. C. Gebler (2004). Implications of column peak capacity on the separation of complex peptide mixtures in single- and two-dimensional high-performance liquid chromatography. *Journal of Chromatography A.* **1061**, 183-192.

Gilar, M., P. Olivova, A. B. Chakraborty, A. Jaworski, S. J. Geromanos and J. C. Gebler (2009). Comparison of 1-D and 2-D LC MS/MS methods for proteomic analysis of human serum. *Electrophoresis.* **30**, 1157-67.

Gilar, M., P. Olivova, A. E. Daly and J. C. Gebler (2005a). Orthogonality of separation in two-dimensional liquid chromatography. *Anal Chem.* **77**, 6426-34.

Gilar, M., P. Olivova, A. E. Daly and J. C. Gebler (2005b). Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. *J Sep Sci.* **28**, 1694-703.

Gillet, L. C., P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter, R. Bonner and R. Aebersold (2012). Targeted data extraction of the MS/MS spectra

generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics.* **11**, O111 016717.

Gong, Y., X. Li, B. Yang, W. Ying, D. Li, Y. Zhang, S. Dai, Y. Cai, J. Wang, F. He and X. Qian (2006). Different immunoaffinity fractionation strategies to characterize the human plasma proteome. *J Proteome Res.* **5**, 1379-87.

Gygi, S. P., B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb and R. Aebersold (1999a). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol.* **17**, 994-9.

Gygi, S. P., B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb and R. Aebersold (1999b). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotech.* **17**, 994-999.

Gygi, S. P., Y. Rochon, B. R. Franza and R. Aebersold (1999). Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol.* **19**, 1720-30.

Haddon, W. F. and F. W. McLafferty (1968). Metastable ion characteristics. VII. Collision-induced metastables. *Journal of the American Chemical Society.* **90**, 4745-4746.

Hall, P. M., G. M. Cawdell, J. G. Cook and B. J. Gould (1983). Measurement of glycosylated haemoglobins and glycosylated plasma proteins in maternal and cord blood using an affinity chromatography method. *Diabetologia.* **25**, 477-81.

Han, X., A. Aslanian and J. R. Yates, 3rd (2008). Mass spectrometry for proteomics. *Curr Opin Chem Biol.* **12**, 483-90.

Han, X., M. Jin, K. Breuker and F. W. McLafferty (2006). Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science.* **314**, 109-12.

Hansen, K. C., G. Schmitt-Ulms, R. J. Chalkley, J. Hirsch, M. A. Baldwin and A. L. Burlingame (2003). Mass spectrometric analysis of protein mixtures at low levels using cleavable 13C-isotope-coded affinity tag and multidimensional chromatography. *Mol Cell Proteomics.* **2**, 299-314.

Hatakeyama, K., K. Ohshima, Y. Fukuda, S. Ogura, M. Terashima, K. Yamaguchi and T. Mochizuki (2011). Identification of a novel protein isoform derived from cancer-related splicing variants using combined analysis of transcriptome and proteome. *Proteomics.* **11**, 2275-82.

He, H., R. P. Rodgers, A. G. Marshall and C. S. Hsu (2011). Algae Polar Lipids Characterized by Online Liquid Chromatography Coupled with Hybrid Linear Quadrupole Ion Trap/Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Energy & Fuels.* **25**, 4770-4775.

Heck, A. J. and R. H. Van Den Heuvel (2004). Investigation of intact protein complexes by mass spectrometry. *Mass Spectrom Rev.* **23**, 368-89.

Helsens, K., P. Van Damme, S. Degroeve, L. Martens, T. Arnesen, J. Vandekerckhove and K. Gevaert (2011). Bioinformatics analysis of a Saccharomyces cerevisiae N-terminal proteome provides evidence of alternative translation initiation and post-translational N-terminal acetylation. *J Proteome Res.* **10**, 3578-89.

Henzel, W. J., T. M. Billeci, J. T. Stults, S. C. Wong, C. Grimley and C. Watanabe (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc Natl Acad Sci U S A.* **90**, 5011-5.

Hernandez, C. and F. G. Cunningham (1990). Eclampsia. *Clin Obstet Gynecol.* **33**, 460-6.

Holman, S. W., P. F. Sims and C. E. Eyers (2012). The use of selected reaction monitoring in quantitative proteomics. *Bioanalysis.* **4**, 1763-86.

Honest, H., L. M. Bachmann, J. K. Gupta, J. Kleijnen and K. S. Khan (2002). Accuracy of cervicovaginal fetal fibronectin test in predicting risk of spontaneous preterm birth: systematic review. *BMJ.* **325**, 301.

Huang, H. L., T. Stasyk, S. Morandell, M. Mogg, M. Schreiber, I. Feuerstein, C. W. Huck, G. Stecher, G. K. Bonn and L. A. Huber (2005). Enrichment of low-abundant serum proteins by albumin/immunoglobulin G immunoaffinity depletion under partly denaturing conditions. *Electrophoresis.* **26**, 2843-9.

Huang, L., G. Harvie, J. S. Feitelson, K. Gramatikoff, D. A. Herold, D. L. Allen, R. Amunngama, R. A. Hagler, M. R. Pisano, W. W. Zhang and X. Fang (2005). Immunoaffinity separation of plasma proteins by IgY microbeads: meeting the needs of proteomic sample preparation and analysis. *Proteomics.* **5**, 3314-28.

Huang, X., L. Chen, G. Fu, H. Xu and X. Zhang (2012). Decreased expression of pigment epithelium-derived factor and increased microvascular density in ovarian endometriotic lesions in women with endometriosis. *Eur J Obstet Gynecol Reprod Biol.* **165**, 104-9.

Huttlin, E. L., A. D. Hegeman, A. C. Harms and M. R. Sussman (2007). Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. *J Proteome Res.* **6**, 392-8.

Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber and M. Mann (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics.* **4**, 1265-72.

Jacobs, J. M., J. N. Adkins, W. J. Qian, T. Liu, Y. Shen, D. G. Camp, 2nd and R. D. Smith (2005). Utilizing human blood plasma for proteomic biomarker discovery. *J Proteome Res.* **4**, 1073-85.

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

Jennings, K. R. (1968). Collision-induced decompositions of aromatic molecular ions. *International Journal of Mass Spectrometry and Ion Physics.* **1**, 227-235.

Johnson, A., F. S. Cowchock, M. Darby, R. Wapner and L. G. Jackson (1991). First-trimester maternal serum alpha-fetoprotein and chorionic gonadotropin in aneuploid pregnancies. *Prenat Diagn.* **11**, 443-50.

Johnson, R. S., S. A. Martin, K. Biemann, J. T. Stults and J. T. Watson (1987). Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Anal Chem.* **59**, 2621-5.

Karas, M. and F. Hillenkamp (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem.* **60**, 2299-301.

Karataev, V. I., B. A. Mamyrin and D. V. Shmikk (1972). New Method for Focusing Ion Bunches in Time-of-Flight Mass Spectrometers. *Soviet Physics Technical Physics.* **16**, 1177-1179.

Karpievitch, Y. V., A. D. Polpitiya, G. A. Anderson, R. D. Smith and A. R. Dabney (2010). Liquid Chromatography Mass Spectrometry-Based

Proteomics: Biological and Technological Aspects. *Ann Appl Stat.* **4**, 1797-1823.

Kebarle, P. (2000). A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *J Mass Spectrom.* **35**, 804-17.

Kingdon, K. H. (1923). A Method for the Neutralization of Electron Space Charge by Positive Ionization at Very Low Gas Pressures. *Physical Review.* **21**, 408-418.

Kiyonami, R. and B. Domon (2010). Selected reaction monitoring applied to quantitative proteomics. *Methods Mol Biol.* **658**, 155-66.

Krijgsveld, J., R. F. Ketting, T. Mahmoudi, J. Johansen, M. Artal-Sanz, C. P. Verrijzer, R. H. Plasterk and A. J. Heck (2003). Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics. *Nat Biotechnol.* **21**, 927-31.

LaBaer, J. (2005). So, you want to look for biomarkers (introduction to the special biomarkers issue). *J Proteome Res.* **4**, 1053-9.

Lange, V., P. Picotti, B. Domon and R. Aebersold (2008). Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol.* **4**, 222.

Lee, J. E., J. F. Kellie, J. C. Tran, J. D. Tipton, A. D. Catherman, H. M. Thomas, D. R. Ahlf, K. R. Durbin, A. Vellaichamy, I. Ntai, A. G. Marshall and N. L. Kelleher (2009). A Robust Two-Dimensional Separation for Top-Down Tandem Mass Spectrometry of the Low-Mass Proteome. *Journal of the American Society for Mass Spectrometry.* **20**, 2183-2191.

Levin, Y., E. Hradetzky and S. Bahn (2011a). Quantification of proteins using data-independent analysis (MSE) in simple and complex samples: a systematic evaluation. *Proteomics.* **11**, 3273-87.

Levin, Y., E. Hradetzky and S. Bahn (2011b). Quantification of proteins using data-independent analysis (MSE) in simple andcomplex samples: a systematic evaluation. *Proteomics.* **11**, 3273-87.

Levine, R. J., S. E. Maynard, C. Qian, K. H. Lim, L. J. England, K. F. Yu, E. F. Schisterman, R. Thadhani, B. P. Sachs, F. H. Epstein, B. M. Sibai, V. P. Sukhatme and S. A. Karumanchi (2004). Circulating angiogenic factors and the risk of preeclampsia. *N Engl J Med.* **350**, 672-83.

Li, G. Z., J. P. Vissers, J. C. Silva, D. Golick, M. V. Gorenstein and S. J. Geromanos (2009). Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *Proteomics.* **9**, 1696-719.

Li, J., H. Steen and S. P. Gygi (2003). Protein profiling with cleavable isotope-coded affinity tag (cICAT) reagents: the yeast salinity stress response. *Mol Cell Proteomics.* **2**, 1198-204.

Link, A. J., J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik and J. R. Yates (1999). Direct analysis of protein complexes using mass spectrometry. *Nat Biotech.* **17**, 676-682.

Liu, H., R. G. Sadygov and J. R. Yates (2004). A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Analytical Chemistry.* **76**, 4193-4201.

Lockitch, E. and L. D. Wadsworth (1993). Handbook of Diagnostic Biochemistry and Hematology in Normal Pregnancy, CRC Press Inc.

Lu, P., C. Vogel, R. Wang, X. Yao and E. M. Marcotte (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol.* **25**, 117-24.

Mabie, W. C., M. L. Pernoll and M. K. Biswas (1986). Chronic hypertension in pregnancy. *Obstet Gynecol.* **67**, 197-205.

Macek, B., M. Mann and J. V. Olsen (2009). Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu Rev Pharmacol Toxicol.* **49**, 199-221.

MacNair, J. E., K. C. Lewis and J. W. Jorgenson (1997). Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns. *Anal Chem.* **69**, 983-9.

MacNair, J. E., G. J. Opiteck, J. W. Jorgenson and M. A. Moseley, 3rd (1997). Rapid separation and characterization of protein and peptide mixtures using 1.5 microns diameter non-porous silica in packed capillary liquid chromatography/mass spectrometry. *Rapid Commun Mass Spectrom.* **11**, 1279-85.

MacNair, J. E., K. D. Patel and J. W. Jorgenson (1999). Ultrahigh-pressure reversed-phase capillary liquid chromatography: isocratic and gradient elution using columns packed with 1.0-micron particles. *Anal Chem.* **71**, 700-8.

Makarov, A. (2000). Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Analytical Chemistry*. **72**, 1156-1162.

Mallick, P., M. Schirle, S. S. Chen, M. R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster and R. Aebersold (2007). Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol.* **25**, 125-31.

Mamyrin, B. A. (1994). Laser assisted reflectron time-of-flight mass spectrometry. *International Journal of Mass Spectrometry and Ion Processes.* **131**, 1-19.

Mamyrin, B. A., V. I. Karataev, D. V. Shmikk and V. A. Zagulin (1973). The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution. Sov. Phys. JETP 37: 45. *Sov. Phys. JETP*. **37**, 45.

Mann, M., P. Hojrup and P. Roepstorff (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol Mass Spectrom.* **22**, 338-45.

Marshall, A. G. (2000). Milestones in fourier transform ion cyclotron resonance mass spectrometry technique development. *International Journal of Mass Spectrometry*. **200**, 331-356.

Matei, L. (1997). Plasma proteins glycosylation and its alteration in disease. *Rom J Intern Med.* **35**, 3-11.

McCormack, A. L., D. M. Schieltz, B. Goode, S. Yang, G. Barnes, D. Drubin and J. R. Yates (1997). Direct Analysis and Identification of Proteins in Mixtures by LC/MS/MS and Database Searching at the Low-Femtomole Level. *Analytical Chemistry*. **69**, 767-776.

McLachlin, D. T. and B. T. Chait (2001). Analysis of phosphorylated proteins and peptides by mass spectrometry. *Curr Opin Chem Biol.* **5**, 591-602.

McLafferty, F. W., K. Breuker, M. Jin, X. Han, G. Infusini, H. Jiang, X. Kong and T. P. Begley (2007). Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics. *FEBS J.* **274**, 6256-68.

McLafferty, F. W. and T. A. Bryce (1967). Metastable-ion characteristics: characterization of isomeric molecules. *Chemical Communications (London).* 1215-1217.

Mehta, A. I., S. Ross, M. S. Lowenthal, V. Fusaro, D. A. Fishman, E. F. Petricoin, 3rd and L. A. Liotta (2003). Biomarker amplification by serum carrier protein binding. *Dis Markers.* **19**, 1-10.

Menschaert, G., W. Van Criekinge, T. Notelaers, A. Koch, J. Crappe, K. Gevaert and P. Van Damme (2013). Deep proteome coverage based on ribosome profiling aids MS-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol Cell Proteomics.*

Michalski, A., J. Cox and M. Mann (2011). More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res.* **10**, 1785-93.

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

Michalski, A., E. Damoc, O. Lange, E. Denisov, D. Nolting, M. Müller, R. Viner, J. Schwartz, P. Remes, M. Belford, J.-J. Dunyach, J. Cox, S. Horning, M. Mann and A. Makarov (2012). Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes. *Molecular & Cellular Proteomics.* **11**.

Mikesh, L. M., B. Ueberheide, A. Chi, J. J. Coon, J. E. P. Syka, J. Shabanowitz and D. F. Hunt (2006). The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics.* **1764**, 1811-1822.

Millioni, R., S. Tolin, G. P. Fadini, M. Falda, B. van Breukelen, P. Tessari and G. Arrigoni (2012). High confidence and sensitivity four-dimensional fractionation for human plasma proteome analysis. *Amino Acids.* **43**, 2199-202.

Mirgorodskaya, E., P. Roepstorff and R. A. Zubarev (1999). Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. *Anal Chem.* **71**, 4431-6.

Mirzaei, H., J. K. McBee, J. Watts and R. Aebersold (2008). Comparative Evaluation of Current Peptide Production Platforms Used in Absolute Quantification in Proteomics. *Molecular & Cellular Proteomics.* **7**, 813-823.

Misek, D. E., R. Kuick, H. Wang, V. Galchev, B. Deng, R. Zhao, J. Tra, M. R. Pisano, R. Amunugama, D. Allen, A. K. Walker, J. R. Strahler, P. Andrews, G. S. Omenn and S. M. Hanash (2005). A wide range of protein isoforms in serum and plasma uncovered by a quantitative intact protein analysis system. *Proteomics.* **5**, 3343-3352.

Miseta, A. and P. Csutora (2000). Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms. *Molecular Biology and Evolution.* **17**, 1232-1239.

Mizejewski, G. J. (2003). Levels of alpha-fetoprotein during pregnancy and early infancy in normal and disease states. *Obstet Gynecol Surv.* **58**, 804-26.

Modrek, B. and C. Lee (2002). A genomic view of alternative splicing. *Nat Genet.* **30**, 13-9.

Modrek, B. and C. J. Lee (2003). Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet.* **34**, 177-80.

Molina, H., D. M. Horn, N. Tang, S. Mathivanan and A. Pandey (2007). Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc Natl Acad Sci U S A.* **104**, 2199-204.

Morris, H. R., T. Paxton, A. Dell, J. Langhorne, M. Berg, R. S. Bordoli, J. Hoyes and R. H. Bateman (1996). High sensitivity collisionally-activated decomposition tandem mass spectrometry on a novel quadrupole/orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom.* **10**, 889-96.

Moskaleva, N. E., V. G. Zgoda and A. I. Archakov (2011). [Mass-spectrometric measurements of P450 isoform specific content and corresponding enzyme activities]. *Bioorg Khim.* **37**, 149-64.

Mrozinski, P., N. Zolotarjova and H. Chen (2008). Human Serum and Plasma Protein Depletion – Novel High-Capacity Affinity Column for the Removal of the "Top 14" Abundant Proteins - Application note. I. Agilent Technologies.

Nagalla, S. R., J. A. Canick, T. Jacob, K. A. Schneider, A. P. Reddy, A. Thomas, S. Dasari, X. Lu, J. A. Lapidus, G. M. Lambert-Messerlian, M. G. Gravett, C. T. Roberts, Jr., D. Luthy, F. D. Malone and M. E. D'Alton (2007). Proteomic analysis of maternal serum in down syndrome: identification of novel protein biomarkers. *J Proteome Res.* **6**, 1245-57.

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

Nelson, C. J., E. L. Huttlin, A. D. Hegeman, A. C. Harms and M. R. Sussman (2007). Implications of 15N-metabolic labeling for automated peptide identification in Arabidopsis thaliana. *Proteomics.* **7**, 1279-92.

Neue, U. D., J. L. Carmody, Y. F. Cheng, Z. Lu, C. H. Phoebe and T. E. Wheat (2001). Design of rapid gradient methods for the analysis of combinatorial chemistry libraries and the preparation of pure compounds. *Adv Chromatogr.* **41**, 93-136.

Nirmalan, N. J., C. Hughes, J. Peng, T. McKenna, J. Langridge, D. A. Cairns, P. Harnden, P. J. Selby and R. E. Banks (2011). Initial development and validation of a novel extraction method for quantitative mining of the formalin-fixed, paraffin-embedded tissue proteome for biomarker investigations. *J Proteome Res.* **10**, 896-906.

O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J Biol Chem.* **250**, 4007-21.

Oda, Y., K. Huang, F. R. Cross, D. Cowburn and B. T. Chait (1999). Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A.* **96**, 6591-6.

Olsen, R. N., D. Woelkers, R. Dunsmoor-Su and D. Y. Lacoursiere (2012). Abnormal second-trimester serum analytes are more predictive of preterm preeclampsia. *Am J Obstet Gynecol.* **207**, 228 e1-7.

Ong, S. E., B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey and M. Mann (2002). Stable isotope labeling by amino acids in cell culture,

SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics.* **1**, 376-86.

Ow, S. Y., M. Salim, J. Noirel, C. Evans, I. Rehman and P. C. Wright (2009). iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J Proteome Res.* **8**, 5347-55.

Pacheco, L. G., S. E. Slade, N. Seyffert, A. R. Santos, T. L. Castro, W. M. Silva, A. V. Santos, S. G. Santos, L. M. Farias, M. A. Carvalho, A. M. Pimenta, R. Meyer, A. Silva, J. H. Scrivens, S. C. Oliveira, A. Miyoshi, C. G. Dowson and V. Azevedo (2011). A combined approach for comparative exoproteome analysis of Corynebacterium pseudotuberculosis. *BMC Microbiol.* **11**, 12.

Pan, Q., O. Shai, L. J. Lee, B. J. Frey and B. J. Blencowe (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet.* **40**, 1413-5.

Pan, S., Y. Wang, J. F. Quinn, E. R. Peskind, D. Waichunas, J. T. Wimberger, J. Jin, J. G. Li, D. Zhu, C. Pan and J. Zhang (2006). Identification of glycoproteins in human cerebrospinal fluid with a complementary proteomic approach. *J Proteome Res.* **5**, 2769-79.

Pappin, D. J., P. Hojrup and A. J. Bleasby (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol.* **3**, 327-32.

Patel, N. A., A. Crombie, S. E. Slade, K. Thalassinos, C. Hughes, J. B. Connolly, J. Langridge, J. C. Murrell and J. H. Scrivens (2012). Comparison of one- and two-dimensional liquid chromatography approaches in the label-free quantitative analysis of Methylocella silvestris. *J Proteome Res.* **11**, 4755-63.

Patel, V. J., K. Thalassinos, S. E. Slade, J. B. Connolly, A. Crombie, J. C. Murrell and J. H. Scrivens (2009). A Comparison of Labeling and Label-Free Mass Spectrometry-Based Proteomics Approaches. *Journal of Proteome Research.* **8**, 3752-3759.

Paternoster, D. M., A. Stella, P. Simioni, A. Girolami and M. Plebani (1996). Fibronectin and antithrombin as markers of pre-eclampsia in pregnancy. *Eur J Obstet Gynecol Reprod Biol.* **70**, 33-9.

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

Pepe, M. S., R. Etzioni, Z. Feng, J. D. Potter, M. L. Thompson, M. Thornquist, M. Winget and Y. Yasui (2001). Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst.* **93**, 1054-61.

Peschke, M., A. Blades and P. Kebarle (2002). Charged states of proteins. Reactions of doubly protonated alkyldiamines with NH(3): solvation or deprotonation. Extension of two proton cases to multiply protonated globular proteins observed in the gas phase. *J Am Chem Soc.* **124**, 11519-30.

Peschke, M., U. H. Verkerk and P. Kebarle (2004). Features of the ESI mechanism that affect the observation of multiply charged noncovalent protein complexes and the determination of the association constant by the titration method. *J Am Soc Mass Spectrom.* **15**, 1424-34.

Petrak, J., R. Ivanek, O. Toman, R. Cmejla, J. Cmejlova, D. Vyoral, J. Zivny and C. D. Vulpe (2008). Deja vu in proteomics. A hit parade of repeatedly identified differentially expressed proteins. *Proteomics.* **8**, 1744-9.

Pieper, R., Q. Su, C. L. Gatlin, S. T. Huang, N. L. Anderson and S. Steiner (2003). Multi-component immunoaffinity subtraction chromatography: an

innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics.* **3**, 422-32.

Pignot, S. and A. Busine (1989). [Fibronectin: an early marker of pre-eclampsia]. *J Gynecol Obstet Biol Reprod (Paris).* **18**, 867-70.

Pitteri, S. J., P. A. Chrisman, J. M. Hogan and S. A. McLuckey (2005). Electron transfer ion/ion reactions in a three-dimensional quadrupole ion trap: reactions of doubly and triply protonated peptides with SO2*. *Anal Chem.* **77**, 1831-9.

Plunkett, B. A., P. Fitchev, J. A. Doll, S. E. Gerber, M. Cornwell, E. P. Greenstein and S. E. Crawford (2008). Decreased expression of pigment epithelium derived factor (PEDF), an inhibitor of angiogenesis, in placentas of unexplained stillbirths. *Reprod Biol.* **8**, 107-20.

Poon, L. C., N. Maiz, C. Valencia, W. Plasencia and K. H. Nicolaides (2009). First-trimester maternal serum pregnancy-associated plasma protein-A and pre-eclampsia. *Ultrasound Obstet Gynecol.* **33**, 23-33.

Poon, L. C. Y., N. A. Kametas, N. Maiz, R. Akolekar and K. H. Nicolaides (2009). First-Trimester Prediction of Hypertensive Disorders in Pregnancy. *Hypertension.* **53, Numb 5**, 812-818.

Prakash, A., D. M. Tomazela, B. Frewen, B. Maclean, G. Merrihew, S. Peterman and M. J. Maccoss (2009). Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. *J Proteome Res.* **8**, 2733-9.

Pratt, J. M., D. M. Simpson, M. K. Doherty, J. Rivers, S. J. Gaskell and R. J. Beynon (2006). Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat. Protocols.* **1**, 1029-1043.

Purvine, S., J. T. Eppel, E. C. Yi and D. R. Goodlett (2003). Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics.* **3**, 847-50.

Putnam, F. W. (1960). The Plasma Proteins, Academic Press.

Qiu, C., M. A. Williams, W. M. Leisenring, T. K. Sorensen, I. O. Frederick, J. C. Dempsey and D. A. Luthy (2003). Family history of hypertension and type 2 diabetes in relation to preeclampsia risk. *Hypertension.* **41**, 408-13.

Rai, A. J., C. A. Gelfand, B. C. Haywood, D. J. Warunek, J. Yi, M. D. Schuchard, R. J. Mehigh, S. L. Cockrill, G. B. Scott, H. Tammen, P. Schulz-Knappe, D. W. Speicher, F. Vitzthum, B. B. Haab, G. Siest and D. W. Chan (2005). HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. *Proteomics.* **5**, 3262-77.

Rappsilber, J., U. Ryder, A. I. Lamond and M. Mann (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Res.* **12**, 1231-45.

Reiter, L., M. Claassen, S. P. Schrimpf, M. Jovanovic, A. Schmidt, J. M. Buhmann, M. O. Hengartner and R. Aebersold (2009). Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics.* **8**, 2405-17.

Reynolds, J. A. and C. Tanford (1970). Binding of dodecyl sulfate to proteins at high binding ratios. Possible implications for the state of proteins in biological membranes. *Proc Natl Acad Sci U S A.* **66**, 1002-7.

Rifai, N., M. A. Gillette and S. A. Carr (2006). Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotech.* **24**, 971-983.

Rivers, J., D. M. Simpson, D. H. L. Robertson, S. J. Gaskell and R. J. Beynon (2007). Absolute Multiplexed Quantitative Analysis of Protein Expression during Muscle Development Using QconCAT. *Molecular & Cellular Proteomics.* **6**, 1416-1427.

Rodriguez-Suarez, E., C. Hughes, L. Gethings, K. Giles, J. Wildgoose, M. Stapels, K. E. Fadgen, S. J. Geromanos, J. P.C. Vissers, F. Elortza and J. I. Langridge (2013). An Ion Mobility Assisted Data Independent LC-MS Strategy for the Analysis of Complex Biological Samples. *Current Analytical Chemistry*. **9**, 199-211.

Roepstorff, P. and J. Fohlman (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom.* **11**, 601.

Ross, P. L., Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson and D. J. Pappin (2004). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics.* **3**, 1154-69.

Sadiq, S. T. and D. Agranoff (2008). Pooling serum samples may lead to loss of potential biomarkers in SELDI-ToF MS proteomic profiling. *Proteome Science.* 16.

Saudan, P., M. A. Brown, M. L. Buddle and M. Jones (1998). Does gestational hypertension become pre-eclampsia? *Br J Obstet Gynaecol.* **105**, 1177-84.

Schenk, S., G. J. Schoenhals, G. de Souza and M. Mann (2008). A high confidence, manually validated human blood plasma protein reference set. *BMC Med Genomics.* **1**, 41.

Schrader, H. M., L. Jovanovic-Peterson, W. C. Bevier and C. M. Peterson (1995). Fasting plasma glucose and glycosylated plasma protein at 24 to 28 weeks of gestation predict macrosomia in the general obstetric population. *Am J Perinatol.* **12**, 247-51.

Schwanhausser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen and M. Selbach (2011). Global quantification of mammalian gene expression control. *Nature.* **473**, 337-42.

Shaarawy, M. and H. E. Didy (1996). Thrombomodulin, plasminogen activator inhibitor type 1 (PAI-1) and fibronectin as biomarkers of endothelial damage in preeclampsia and eclampsia. *Int J Gynaecol Obstet.* **55**, 135-9.

Shankar, R., F. Cullinane, S. P. Brennecke and E. K. Moses (2004). Applications of proteomic methodologies to human pregnancy research: a growing gestation approaching delivery? *Proteomics.* **4**, 1909-17.

She, Y.-M., M. Rosu-Myles, L. Walrond and T. D. Cyr (2012). Quantification of protein isoforms in mesenchymal stem cells by reductive dimethylation of lysines in intact proteins. *Proteomics.* **12**, 369-379.

Shliaha, P. V., N. J. Bond, L. Gatto and K. S. Lilley (2013). Effects of Traveling Wave Ion Mobility Separation on Data Independent Acquisition in Proteomics Studies. *J Proteome Res.*

Silva, J. C., R. Denny, C. Dorschel, M. V. Gorenstein, G. Z. Li, K. Richardson, D. Wall and S. J. Geromanos (2006). Simultaneous qualitative and quantitative analysis of the Escherichia coli proteome: a sweet tale. *Mol Cell Proteomics.* **5**, 589-607.

Silva, J. C., R. Denny, C. A. Dorschel, M. Gorenstein, I. J. Kass, G.-Z. Li, T. McKenna, M. J. Nold, K. Richardson, P. Young and S. Geromanos (2005). Quantitative Proteomic Analysis by Accurate Mass Retention Time Pairs. *Analytical Chemistry.* **77**, 2187-2200.

Silva, J. C., M. V. Gorenstein, G. Z. Li, J. P. Vissers and S. J. Geromanos (2006). Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics.* **5**, 144-56.

Simpson, D. M. and R. J. Beynon (2010). Acetone precipitation of proteins and the modification of peptides. *J Proteome Res.* **9**, 444-50.

Skilling, J., R. Denny, K. Richardson, P. Young, T. McKenna, I. Campuzano and M. Ritchie (2004). ProbSeq--a fragmentation model for interpretation of electrospray tandem mass spectrometry data. *Comp Funct Genomics.* **5**, 61-8.

Song, X., J. Bandow, J. Sherman, J. D. Baker, P. W. Brown, M. T. McDowell and M. P. Molloy (2008). iTRAQ Experimental Design for Plasma Biomarker Discovery. *Journal of Proteome Research.* **7**, 2952-2958.

Spencer, K., C. K. Yu, G. Rembouskos, R. Bindra and K. H. Nicolaides (2005a). First trimester sex hormone-binding globulin and subsequent development of preeclampsia or other adverse pregnancy outcomes. *Hypertens Pregnancy.* **24**, 303-11.

Spencer, K., C. K. H. Yu, G. Rembouskos, R. Bindra and K. H. Nicolaides (2005b). First Trimester Sex Hormone-Binding Globulin and Subsequent Development of Preeclampsia or Other Adverse Pregnancy Outcomes. *Hypertension in Pregnancy.* **24**, 303-311.

Spiro, R. G. (2002). Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds. *Glycobiology.* **12**, 43R-56R.

Stadalius, M. A., B. F. Ghrist and L. R. Snyder (1987). Predicting bandwidth in the high-performance liquid chromatographic separation of large biomolecules. II. A general model for the four common high-performance liquid chromatography methods. *J Chromatogr.* **387**, 21-40.

Stafford Jr, G. C., P. E. Kelley, J. E. P. Syka, W. E. Reynolds and J. F. J. Todd (1984). Recent improvements in and analytical applications of advanced ion trap technology. *International Journal of Mass Spectrometry and Ion Processes.* **60**, 85-98.

Stahl, D. C., K. M. Swiderek, M. T. Davis and T. D. Lee (1996). Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *Journal of the American Society for Mass Spectrometry.* **7**, 532-540.

Steen, H. and M. Mann (2004). The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol.* **5**, 699-711.

Stensballe, A., O. N. Jensen, J. V. Olsen, K. F. Haselmann and R. A. Zubarev (2000). Electron capture dissociation of singly and multiply phosphorylated peptides. *Rapid Commun Mass Spectrom.* **14**, 1793-800.

Stephens, W. E. (1946). A Pulsed Mass Spectrometer with Time Dispersion. Proceedings of the American Physical Society. Cambridge, MA., USA, Physical Review. **69** 691.

Surinova, S., R. Schiess, R. Huttenhain, F. Cerciello, B. Wollscheid and R. Aebersold (2011). On the development of plasma protein biomarkers. *J Proteome Res.* **10**, 5-16.

Syka, J. E., J. J. Coon, M. J. Schroeder, J. Shabanowitz and D. F. Hunt (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A.* **101**, 9528-33.

Tanaka, K., H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida and T. Matsuo (1988). Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry.* **2**, 151-153.

Tang, L. J., F. De Seta, F. Odreman, P. Venge, C. Piva, S. Guaschino and R. C. Garcia (2007). Proteomic analysis of human cervical-vaginal fluids. *J Proteome Res.* **6**, 2874-83.

Taouatas, N., M. M. Drugan, A. J. Heck and S. Mohammed (2008). Straightforward ladder sequencing of peptides using a Lys-N metalloendopeptidase. *Nat Methods.* **5**, 405-7.

Thalassinos, K., J. P. Vissers, S. Tenzer, Y. Levin, J. W. Thompson, D. Daniel, D. Mann, M. R. DeLong, M. A. Moseley, A. H. America, A. K. Ottens, G. S. Cavey, G. Efstathiou, J. H. Scrivens, J. I. Langridge and S. J. Geromanos (2012). Design and application of a data-independent precursor and product ion repository. *J Am Soc Mass Spectrom.* **23**, 1808-20.

Thangaratinam, S., J. Langenveld, B. W. Mol and K. S. Khan (2011). Prediction and primary prevention of pre-eclampsia. *Best Pract Res Clin Obstet Gynaecol.* **25**, 419-33.

Thomson, J. J. (1899). On the masses of the ions in gases at low pressures. *Philosophical Magazine.* **48**, 547-567.

Thomson, J. J. (1911). Rays of positive electricity. *Philosophical Magazine.* **6**, 752-767.

Tietz, N. W. and A. B. Amerson (1990). Clinical guide to laboratory tests, Elsevier Health.

Tipton, J. D., J. C. Tran, A. D. Catherman, D. R. Ahlf, K. R. Durbin, J. E. Lee, J. F. Kellie, N. L. Kelleher, C. L. Hendrickson and A. G. Marshall (2012). Nano-LC FTICR Tandem Mass Spectrometry for Top-Down Proteomics: Routine Baseline Unit Mass Resolution of Whole Cell Lysate Proteins up to 72 kDa. *Analytical Chemistry.* **84**, 2111-2117.

Toll, H., H. Oberacher, R. Swart and C. G. Huber (2005). Separation, detection, and identification of peptides by ion-pair reversed-phase high-performance liquid chromatography-electrospray ionization mass spectrometry at high and low pH. *J Chromatogr A.* **1079**, 274-86.

Tu, C., P. A. Rudnick, M. Y. Martinez, K. L. Cheek, S. E. Stein, R. J. Slebos and D. C. Liebler (2010). Depletion of abundant plasma proteins and limitations of plasma proteomics. *J Proteome Res.* **9**, 4982-91.

Unlu, M., M. E. Morgan and J. S. Minden (1997). Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis.* **18**, 2071-7.

Valdés R, E., K. Lattes A, H. Muñoz S and M. Ángel Cumsille (2012). Evaluación de la globulina transportadora de hormonas esteroidales (SHBG) durante el embarazo como factor predictor de pre-eclampsia y restricción del crecimiento intrauterino. *Revista médica de Chile.* **140**, 589-594.

Valentine, S. J., A. E. Counterman, C. S. Hoaglund, J. P. Reilly and D. E. Clemmer (1998). Gas-phase separations of protease digests. *J Am Soc Mass Spectrom.* **9**, 1213-6.

Van den Steen, P., P. M. Rudd, R. A. Dwek and G. Opdenakker (1998). Concepts and principles of O-linked glycosylation. *Crit Rev Biochem Mol Biol.* **33**, 151-208.

Venable, J. D., M. Q. Dong, J. Wohlschlegel, A. Dillin and J. R. Yates (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods.* **1**, 39-45.

Vestal, M. L., P. Juhasz and S. A. Martin (1995). Delayed extraction matrix-assisted laser desorption time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry.* **9**, 1044-1050.

Vogel, C., R. de Sousa Abreu, D. Ko, S.-Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte and L. O. Penalva (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol.* **6**.

Voyksner, R. D. and H. Lee (1999). Investigating the use of an octupole ion guide for ion storage and high-pass mass filtering to improve the quantitative performance of electrospray ion trap mass spectrometry. *Rapid Commun Mass Spectrom.* **13**, 1427-37.

Wald, N. J., C. Rodeck, A. K. Hackshaw, J. Walters, L. Chitty and A. M. Mackinson (2003). First and second trimester antenatal screening for Down's syndrome: the results of the Serum, Urine and Ultrasound Screening Study (SURUSS). *J Med Screen.* **10**, 56-104.

Walther, T. C. and M. Mann (2010). Mass spectrometry–based proteomics in cell biology. *The Journal of Cell Biology.* **190**, 491-500.

Wang, P., F. G. Bouwman and E. C. Mariman (2009). Generally detected proteins in comparative proteomics--a matter of cellular stress response? *Proteomics.* **9**, 2955-66.

Washburn, M. P., D. Wolters and J. R. Yates, 3rd (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol.* **19**, 242-7.

Wasinger, V. C., S. J. Cordwell, A. Cerpa-Poljak, J. X. Yan, A. A. Gooley, M. R. Wilkins, M. W. Duncan, R. Harris, K. L. Williams and I. Humphery-Smith (1995). Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis.* **16**, 1090-4.

Wiley, W. C. and I. H. McLaren (1955). Time-of-Flight Mass Spectrometer with Improved Resolution. *Review of Scientific Instruments.* **26**, 1150-1157.

Wilkins, M. and S. N. Hunt (2007). Bioinformatics and Experimental Design for Biomarker Discovery. Proteomics of Human Body Fluids. V. Thongboonkerd, Humana Press**:** 147-174.

Wilkins, M. R., C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J.-C. Sanchez, J. X. Yan, A. A. Gooley, G. Hughes, I. Humphery-Smith, K. L. Williams and D. F. Hochstrasser (1996). From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis. *Nat Biotech.* **14**, 61-65.

Wilm, M. and M. Mann (1996). Analytical properties of the nanoelectrospray ion source. *Anal Chem.* **68**, 1-8.

Winter, J. M., C. J. Yeo and J. R. Brody (2013). Diagnostic, prognostic, and predictive biomarkers in pancreatic cancer. *J Surg Oncol.* **107**, 15-22.

Wisniewski, J. R., A. Zougman, N. Nagaraj and M. Mann (2009). Universal sample preparation method for proteome analysis. *Nat Meth.* **6**, 359-362.

Wiza, J. L. (1979). Microchannel plate detectors. *Nuclear Instruments and Methods.* **162**, 587-601.

Wolf, M., L. Sandler, K. Munoz, K. Hsu, J. L. Ecker and R. Thadhani (2002). First trimester insulin resistance and subsequent preeclampsia: a prospective study. *J Clin Endocrinol Metab.* **87**, 1563-8.

Wu, C. C., M. J. MacCoss, K. E. Howell, D. E. Matthews and J. R. Yates, 3rd (2004). Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. *Anal Chem.* **76**, 4951-9.

Wu, S., N. Tolic, Z. Tian, E. W. Robinson and L. Pasa-Tolic (2011). An integrated top-down and bottom-up strategy for characterization of protein isoforms and modifications. *Methods Mol Biol.* **694**, 291-304.

Yamashita, M. and J. B. Fenn (1984). Electrospray ion source. Another variation on the free-jet theme. *The Journal of Physical Chemistry.* **88**, 4451-4459.

Yang, Z., L. E. Harris, D. E. Palmer-Toy and W. S. Hancock (2006). Multilectin affinity chromatography for characterization of multiple glycoprotein biomarker candidates in serum from breast cancer patients. *Clin Chem.* **52**, 1897-905.

Yates, J. R., C. I. Ruse and A. Nakorchevsky (2009). Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. *Annual Review of Biomedical Engineering.* **11**, 49-79.

Yu, C. K., O. Khouri, N. Onwudiwe, Y. Spiliopoulos and K. H. Nicolaides (2008). Prediction of pre-eclampsia by uterine artery Doppler imaging: relationship to gestational age at delivery and small-for-gestational age. *Ultrasound Obstet Gynecol.* **31**, 310-3.

Zhang, X., A. Fang, C. P. Riley, M. Wang, F. E. Regnier and C. Buck (2010). Multi-dimensional liquid chromatography in proteomics—A review. *Analytica Chimica Acta.* **664**, 101-113.

Zhou, F., J. D. Cardoza, S. B. Ficarro, G. O. Adelmant, J. B. Lazaro and J. A. Marto (2010). Online nanoflow RP-RP-MS reveals dynamics of multicomponent Ku complex in response to DNA damage. *J Proteome Res.* **9**, 6242-55.

Zolotarjova, N., J. Martosella, G. Nicol, J. Bailey, B. E. Boyes and W. C. Barrett (2005). Differences among techniques for high-abundant protein depletion. *Proteomics.* **5**, 3304-13.

Zubarev, R. A., N. L. Kelleher and F. W. McLafferty (1998). Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *Journal of the American Chemical Society.* **120**, 3265-3266.

Zybailov, B., A. L. Mosley, M. E. Sardiu, M. K. Coleman, L. Florens and M. P. Washburn (2006). Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J Proteome Res.* **5**, 2339-47.

Appendix A - Sample Information for IgY-12 Partitioned Plasma

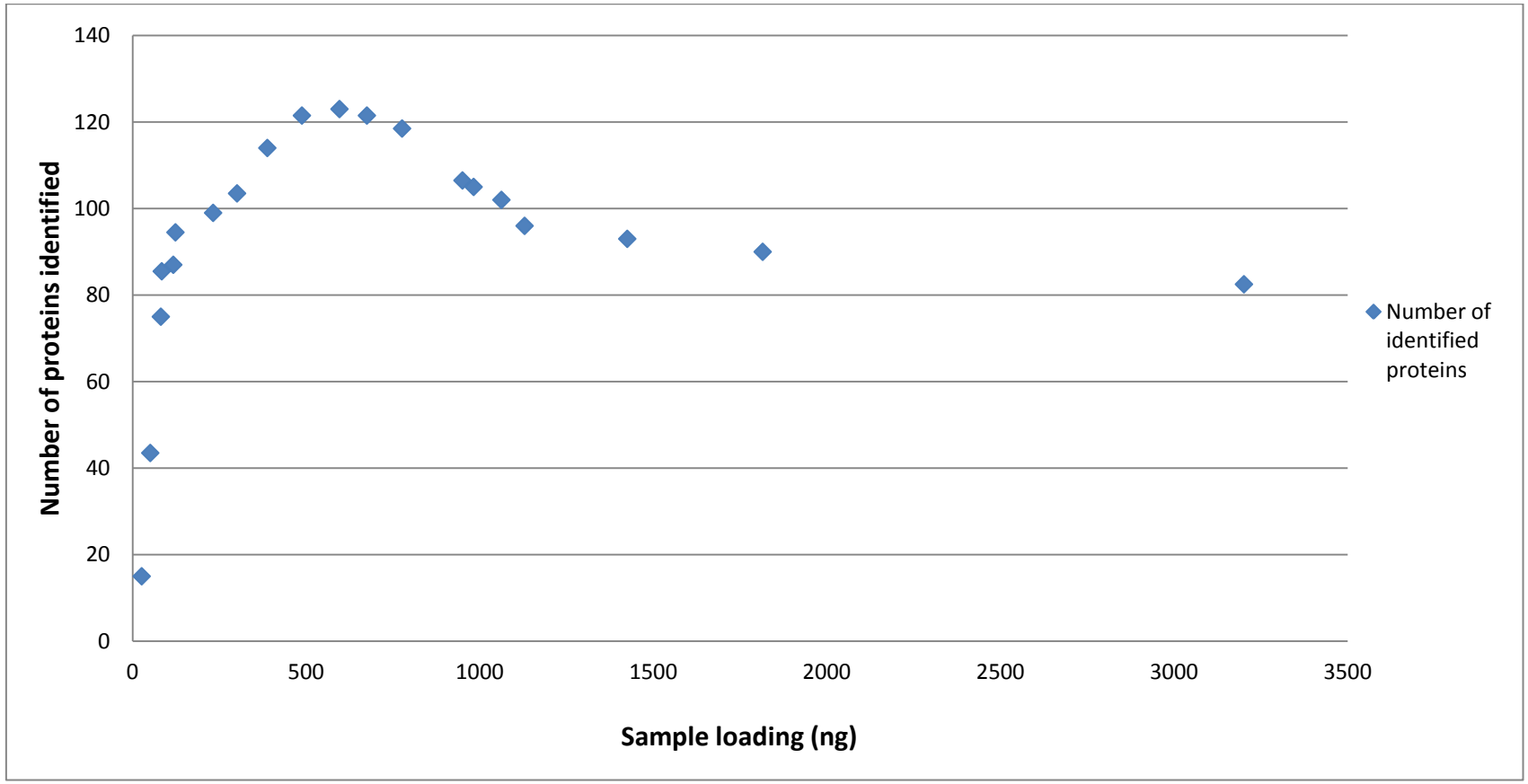
HBRC ID	Lab ID	Outcome	GA (days)	CRL (mm)	Age (yrs)	Wt (kg)	Ht (mt)	BMI	Ethnicity	Parity	Sex
14898	6060	Normal	94	75.9	25.9	51.0	1.61	19.7	Caucasian	5	male
60873	2685	Normal	85	62.6	26.3	69.0	1.62	26.3	Caucasian	0	male
61749	4645	Normal	88	61.8	27.4	69.0	1.78	21.8	Caucasian	1	female
64984	5426	Normal	83	58.3	31.6	55.0	1.53	23.5	Caucasian	0	female
59892	1767	Normal	86	58.3	32.1	66.0	1.63	24.8	Caucasian	0	female
63001	4037	Normal	87	60.5	32.4	67.0	1.65	24.6	Caucasian	0	female
61303	3068	Normal	86	62.1	33.2	72.0	1.64	26.8	Caucasian	0	female
63601	4417	Normal	91	72.0	33.7	66.0	1.73	22.1	Caucasian	0	male
61116	2897	Normal	84	59.3	35.3	63.0	1.61	24.3	Caucasian	0	female
15174	6766	Normal	87	65.0	36.2	66.0	1.67	23.7	Caucasian	1	male
61852	3413	Normal	91	74.4	36.4	61.0	1.73	20.5	Caucasian	0	female
MEDIAN			**87.0**	**62.1**	**32.4**	**66.0**	**1.6**	**23.7**			
AVERAGE			**87.5**	**64.6**	**31.9**	**64.1**	**1.7**	**23.5**			
STANDARD DEVIATION			**3.3**	**6.5**	**3.8**	**6.3**	**0.1**	**2.3**			

Appendix B - Sample Information for IgY-14

Partitioned Plasma

HBRC ID	Lab ID	Outcome	GA (days)	CRL (mm)	Age (yrs)	Wt (kg)	Ht (mt)	BMI	Ethnicity	Parity	Sex
8347	5832	Normal	86	62.5	43.5	70.0	1.62	26.7	White	2	male
61140	18530	Normal	85	60.6	29.4	55.0	1.58	22.2	White	1	male
68376	9463	Normal	85	65.2	24.9	52.0	1.67	18.6	Black	0	male
69246	10147	Normal	86	60.0	41.3	88.0	1.70	30.4	White	2	female
70874	10253	Normal	93	72.0	22.2	89.0	1.67	31.9	White	0	male
70935	11558	Normal	88	70.9	33.9	71.0	1.63	26.9	White	0	male
70977	11398	Normal	85	62.1	25.9	84.0	1.69	29.4	Black	1	female
71751	11197	Normal	95	83.0	20.7	64.0	1.70	22.1	Black	0	female
78564	13871	Normal	86	66.0	22.0	58.0	1.50	25.8	Black	0	male
81052	14749	Normal	86	54.3	24.5	73.0	1.75	23.8	Black	0	female
55839	9030	PET	96	77.2	27.1	69.0	1.64	25.7	Afro-Caribbean	2	female
58195	162	PET	87	67.0	27.5	65.0	1.68	23.0	Afro-Caribbean	1	male
60004	1873	PET	86	54.6	31.7	85.0	1.75	27.9	Caucasian	0	male
60254	2102	PET	87	54.1	30.0	54.0	1.57	21.9	Caucasian	0	Female
60484	2317	PET	86	55.3	32.7	75.0	1.50	33.3	Caucasian	1	male
60635	2464	PET	88	70.6	39.2	63.0	1.66	22.9	Caucasian	1	female
61671	4156	PET	85	58.1	30.9	66.0	1.52	28.6	Caucasian	0	female
62972	4018	PET	94	71.4	42.1	72.0	1.58	29.0	Caucasian	2	female
64781	5233	PET	93	71.0	39.4	74.0	1.62	28.2	Afro-Caribbean	2	female
65134	5958	PET	94	71.4	38.6	86.0	1.60	33.6	Caucasian	3	male
65524	5862	T21	96	73.4	39.2	48.1	1.58	19.4	White	0	female
70863	10246	T21	91	65.0	33.6	59.0	1.55	24.6	White	2	female
80978	14744	T21	86	64.2	43.6	101.7	1.75	33.1	White	2	female
85165	16642	T21	94	70.0	38.0	73.0	1.67	26.2	White	2	female
88964	18423	T21	92	58.9	39.1	81.0	1.75	26.4	White	0	female
89154	18538	T21	85	69.1	39.1	60.0	1.68	21.4	White	0	female
89860	22465	T21	91	69.3	38.7	64.0	1.55	26.6	Oriental	1	male
93882	20717	T21	92	63.0	34.6	91.0	1.83	27.2	White	1	female
95125	21325	T21	91	65.5	41.2	63.0	1.66	22.9	White	1	male
MEDIAN			**88.0**	**65.5**	**33.9**	**70.0**	**1.7**	**26.4**			
AVERAGE			**89.3**	**65.7**	**33.6**	**70.8**	**1.6**	**26.2**			
STANDARD DEVIATION			**3.9**	**7.1**	**7.0**	**13.1**	**0.1**	**4.0**			

Appendix C – Optimisation of sample loading by protein identification rate



A range of sample loading was investigated from 26 ng to 3.2 µg with the maximum number of proteins identified in the range 480 ng – 780 ng of tryptic digest on column.

Appendix D – Dates of Plasma Partition using IgY-14 LC2 Chromatography

Outcome	HBRC ID	DATE DEPLETED
Normal	8347	21.01.10
Normal	61140	22.01.10
Normal	68376	03.02.10
Normal	69246	04.02.10
Normal	70874	04.02.10
Normal	70935	05.02.10
Normal	70977	09.02.10
Normal	71751	10.02.10
Normal	81052	12.02.10
Normal	84672	16.02.10
PET	55839	13.01.10
PET	58195	11.01.10
PET	60004	19.01.10
PET	60254	20.01.10
PET	60484	20.01.10
PET	60635	20.01.10
PET	61671	20.01.10
PET	62972	21.01.10
PET	64781	28.01.10
PET	65134	09.02.10
T21	65524	21.01.10
T21	70863	22.01.10
T21	80978	03.02.10
T21	85165	04.02.10
T21	88964	05.02.10
T21	89154	09.02.10
T21	89860	09.02.10
T21	93882	10.02.10
T21	95125	11.02.10

Appendix E - Sample Information for Pooled IgY-14 Partitioned Plasma (normal, unaffected pregnancy)

Box No	HBRC ID	Lab ID	Outcome	GA (days)	CRL (mm)	Age (yrs)	Wt (kg)	Ht (mt)	BMI	Ethnicity	Parity	Sex
1	62743	3873	Normal	83	58.0	18.0	60.0	1.57	24.3	Mixed	0	female
1	93117	20330	Normal	83	56.1	25.6	58.0	1.68	20.6	White	0	male
1	95167	21336	Normal	83	53.2	31.8	60.0	1.68	21.4	White	0	male
1	68376	9463	Normal	85	65.2	24.9	52.0	1.67	18.6	Black	0	male
1	70977	11398	Normal	85	62.1	25.9	84.0	1.69	29.4	Black	1	female
1	61140	18530	Normal	85	60.6	29.4	55.0	1.58	22.2	White	1	male
1	64765	21327	Normal	85	58.4	32.1	53.1	1.60	20.7	White	1	male
1	8347	5832	Normal	86	62.5	43.5	70.0	1.62	26.7	White	2	male
1	69246	10147	Normal	86	60.0	41.3	88.0	1.70	30.4	White	2	female
1	78564	13871	Normal	86	66.0	22.0	58.0	1.50	25.8	Black	0	male
1	81052	14749	Normal	86	54.3	24.5	73.0	1.75	23.8	Black	0	female
1	96759	22151	Normal	86	58.9	21.5	55.0	1.62	21.0	Black	1	male
1	70935	11558	Normal	88	70.9	33.9	71.0	1.63	26.9	White	0	male
1	94877	21198	Normal	88	63.5	34.1	60.0	1.65	22.0	White	1	male
1	86757	17363	Normal	90	67.7	44.9	62.0	1.53	26.5	White	1	male
1	62454	5265	Normal	93	71.6	39.8	68.5	1.65	25.3	Mixed	0	male
1	70874	10253	Normal	93	72.0	22.2	89.0	1.67	31.9	White	0	male
1	84672	16417	Normal	94	74.7	36.2	50.0	1.56	20.5	Oriental	1	female
1	71751	11197	Normal	95	83.0	20.7	64.0	1.70	22.1	Black	0	female
2	97177	22324	Normal	100	81.1	26.5	59.0	1.56	24.2	White	0	female
MEDIAN				**86.0**	**63.0**	**28.0**	**60.0**	**1.6**	**24.0**			
AVERAGE				**88.0**	**65.0**	**29.9**	**64.5**	**1.6**	**24.2**			
STANDARD DEVIATION				**4.7**	**8.4**	**8.1**	**11.6**	**0.1**	**3.6**			

Appendix E - Sample Information for Pooled IgY-14 Partitioned Plasma (trisomy 21 pregnancy)

Box No	HBRC ID	Lab ID	Outcome	GA (days)	CRL (mm)	Age (yrs)	Wt (kg)	Ht (mt)	BMI	Ethnicity	Parity	Sex
1	89154	18538	T21	85	69.1	39.1	60.0	1.68	21.4	White	0	female
1	80978	14744	T21	86	64.2	43.6	101.7	1.75	33.1	White	2	female
2	96296	22203	T21	86	59.7	32.7	45.0	1.58	18.0	Mixed	0	male
2	97178	22326	T21	87	66.0	40.5	58.0	1.63	21.9	White	2	male
1	78718	13953	T21	90	70.5	32.2	67.0	1.70	23.2	White	0	female
1	70863	10246	T21	91	65.0	33.6	59.0	1.55	24.6	White	2	female
1	95125	21325	T21	91	65.5	41.2	63.0	1.66	22.9	White	1	male
2	89860	22465	T21	91	69.3	38.7	64.0	1.55	26.6	Oriental	1	male
1	88964	18423	T21	92	58.9	39.1	81.0	1.75	26.4	White	0	female
1	93882	20717	T21	92	63.0	34.6	91.0	1.83	27.2	White	1	female
1	96758	22150	T21	92	75.3	38.2	62.0	1.73	20.8	White	1	female
1	78601	13900	T21	94	77.6	39.4	64.0	1.58	25.8	White	0	male
1	85165	16642	T21	94	70.0	38.0	73.0	1.67	26.2	White	2	female
1	95190	21338	T21	95	69.6	32.0	59.0	1.65	21.6	White	1	male
1	65524	5862	T21	96	73.4	39.2	48.1	1.58	19.4	White	0	female
1	72339	22144	T21	98	73.0	40.2	56.0	1.63	21.2	White	2	female
1	101609	24469	T21	98	83.8	41.0	70.8	1.73	23.7	White	1	female
1	102134	24697	T21	98	80.8	31.0	61.0	1.58	24.6	Asian	3	male
1	104169	25609	T21	97	64.7	37.0	58.0	1.63	21.9	White	2	female
1	102697	25404	T21	84	49.3	31.0	63.0	1.68	22.3	White	0	male
MEDIAN				**92.0**	**69.2**	**38.5**	**62.5**	**1.7**	**23.0**			
AVERAGE				**91.9**	**68.4**	**37.1**	**65.2**	**1.7**	**23.6**			
STANDARD DEVIATION				**4.5**	**7.9**	**3.8**	**13.3**	**0.1**	**3.3**			