Quantitative Label-Free Comparative Proteomic Analyses of Eukaryotic Tissues via Mass Spectrometry

by

David Andrew Kramer

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Biochemistry

University of Alberta

Abstract

In recent years, rapid advances in genomic and transcriptomic sequencing technologies have enabled the compilation of vast libraries of protein sequences and thus an explosion in bioinformatics-based fields of research. Among these is proteomics – the study of global protein abundance within a biological system. Proteomics has risen to considerable prominence within the biological and health sciences due to its ability to grasp the subtle complexities of protein biochemistry on an impressive, system-wide scale. In conjunction with the compilation of protein-sequence libraries, advances in liquid chromatography and mass spectrometry have allowed for the reliable and rapid deconvolution, identification – and more recently, quantitation – of proteins within complex mixtures.

Majorly comparative in design, modern proteomics experiments aim to aid in our understanding of how biological systems respond to specific conditions. Because of this, most proteomic quantitation is relative, typically being achieved using stable isotopic labels; proteins originating from two separate experimental conditions are independently labelled with heavy or light stable isotope tags, then mixed together in equivalent proportions. Subsequent analysis of the proteins present in the sample by mass spectrometry allows for the direct comparison of proteins' abundances relative to each other, and inference of causality with respect to the experimental variable. While incredibly elegant in design, such techniques are often impractical, being intensive with respect to cost, time, and sample-handling.

With applications ranging from the study of individual proteins and their biological functions to the complexities of disease pathogenesis, there has recently been a push toward the development

of robust methods for label-free comparative proteomic quantitation. As a result, several techniques for quantitative label-free proteomics have been developed, typically relying on one of two strategies for determining a protein's abundance within a sample; namely spectral counting (counting the number of peptide fragments observed which match a protein's theoretical fragmentation pattern) and extraction of ions' absolute intensities (integration of the total abundance of ions determined to correlate with a protein).

However, several caveats exist for each method and its implementation. Proper methods for data correction and normalization, the treatment of missing values between datasets, and statistical testing/correcting procedures all remain contentious and active areas of research. As such, there exists a lack of consensus on which strategies – and their execution – are best.

Yet, due to the practicality and suitability of label-free proteomic quantitation in the study and characterization of nearly any biological system – including those frequented by diagnostic medicine – I have become a strong proponent of its use. This advocacy has led to our development of a robust, reliable, reproducible, and practical approach to label-free proteomic analyses. Through sample-specific normalization in addition to building upon previously proposed techniques, we provide framework for future label-free proteomic studies with application in any of a myriad of biological systems.

This thesis herein explores the development and application of our mass spectrometry-based label-free semi-quantitative comparative proteomics technique, utilizing the sample-specific normalization of proteins' absolute ion abundances in the characterization of:

1. The proteomic composition of murine hepatic lipid droplets; how they change in response to dietary stress experienced by periods of fasting or fasting followed by re-feeding; and the implications the dynamics of these organellar proteomes may have in their physiological function.

2. The proteomic changes observed *in vivo* for EL4-lymphoma tumours either untreated or treated with an etoposide-cyclophosphamide chemotherapeutic cocktail, and the implications these changes may have towards our understanding of tumour-death.

3. The proteomic differences of luminal-subtype estrogen-receptor positive breast tumours from patients experiencing either disease-free or disease-recurrent survival; the identification of sub-populations of these tumours based on patients' recurrence status, as defined by protein abundance; and the identification of several proteins potentially predictive of a patient's disease-free survival.

Preface

This thesis is an original work by David Andrew Kramer. Animal studies presented in Chapters 2 and 3 were performed in accordance with the guidelines of the Canadian Council on Animal Care (CCAC), with approval from the University of Alberta Animal Welfare Committee (Animal User Protocol #402) and the Cross Cancer Institute Animal Ethics Committee (Protocol #AC10171), respectively. Human clinical tumour samples, with corresponding de-identified patient information as presented in Chapter 4, were obtained with sponsorship from the Alberta Cancer Foundation (ACF), with approval from the University of Alberta Health Research Ethics Board – Biomedical Panel ("Alberta Cancer Proteome Platform", ID# Pro-00049009, 08/26/2014).

For Chapter 2, animal work, RT-qPCR, protein purification, and immunoblot verification was performed by Drs. Ariel Quiroga, Jihong Lian, and Richard Lehner. All mass spectrometry and subsequent data analysis was performed by myself. A version of Chapter 2 of this thesis has been published as:

Kramer, D. A., Quiroga, A. D., Lian, J., Fahlman, R. P., & Lehner, R. (2018). Fasting and refeeding induces changes in the mouse hepatic lipid droplet proteome. *Journal of Proteomics, 181*, 213–224. http://doi.org/10.1016/J.JPROT.2018.04.024

For Chapter 3, the animal work presented was performed by Dr. Melinda Wuest, with immunoblot verification of endogenous caspase abundance performed by Dr. Mohamed Eldeeb. All sample preparation, mass spectrometry, and data analysis was performed by myself. A version of Chapter 3 of this thesis has been published as:

Kramer, D. A., Eldeeb, M. A., Wuest, M., Mercer, J., & Fahlman, R. P. (2017). Proteomic characterization of EL4 lymphoma-derived tumours upon chemotherapy treatment reveals potential roles for lysosomes and caspase-6 during tumour cell death in vivo. *PROTEOMICS, 17(12),* 1700060. http://doi.org/10.1002/pmic.201700060

For Chapter 4, clinical tumour samples and tissue microarrays were selected and provided by Drs. Judith Hugh and Sambasivarao Damaraju, respectively. Partial sample preparation was performed by Ramanaguru Siva Piragasam, and immunoblot validation of PIGR was performed by Yifei Wu. Sample preparation, mass spectrometry, data analysis, and immunohistochemistry was performed by myself, and was visually scored by Dr. Wei-Feng Dong.

Supplemental tables for each chapter can be found with their respective publications, or collectively at: Supplemental Tables

Through art and science in their broadest senses it is possible to make a permanent contribution towards the improvement and enrichment of human life and it is these pursuits that we students are engaged in.

-Frederick Sanger

Acknowledgements

The work presented in this thesis has only been made possible through constant collaboration and support from many mentors, friends, and family members. Because of this, I would first like to thank my supervisor, Dr. Richard P. Fahlman, for providing me with the environment, tools, guidance, and much-needed encouragement in my pursuit of knowledge. Since first taking up a project with him as an undergraduate in 2010, Richard has always made it possible for me to freely pursue my research interests, in addition to constantly providing me with opportunities to build my skillset. My abilities as a scientist owe themselves largely to you. For your continued knowledge, friendship, and mentorship over the past 10 years, I express my deepest gratitude.

The vast majority of my research, as presented in this thesis, has been made possible through collaborative projects with Drs. Richard Lehner, Ariel Quiroga, and Jihong Lian of the Department of Pediatrics; Drs. John Mercer and Melinda Wuest of the Department of Experimental Oncology; and finally, Drs. Judith Hugh and Samasivarao Damaraju of the Department of Laboratory Medicine and Pathology - thank you for your assistance enabling my research and trusting in my abilities. I must also thank Jack Moore, Paul Semchuk, and Audric Moses of the Alberta Proteomics & Mass Spectrometry (APM) Facility and Lipidomics Core, for their knowledge, technical expertise, patience, and friendship; I will miss the conversations, laughs, and coffee breaks. I must also thank Daryl Glubrecht for teaching me how to perform immunohistochemical staining of tissue microarrays, and Dr. Wei-Feng Dong for scoring the results. An immense amount of gratitude must also be expressed to the Canadian Institutes for Health Research (CIHR), Alberta Innovates – Technology Futures (AITF), the Government of Alberta, and the Department of Biochemistry for their financial support.

The members of the Fahlman lab, past and present – in addition to my close friends and those I have made in the Department of Biochemistry and beyond – have been instrumental to the completion of this thesis. Specifically, to Susan Xu, Prarthna Nagar, Guru Siva Piragasam, Luana Leitao, Andrew Locke, Ply Pasarj, Alex Bodman, Connie Le, Philip Krause, Cory Olsson, Cory Shukaliak, and Drs. Steven Chaulk, Roshani Payoe, Angela Fung, Mohamed Eldeeb, Quang Tran, Justin Fedor, Robert Mercer, and John Maringa, among others; thank you for teaching,

supporting, and challenging me. My time in the graduate program has been an incredible experience, and I am happy to have shared it with all of you. For their work and trust in my mentorship I would also like to thank Yifei Wu, Leslie Shewchuk, Anna Stuart, Kim Ho, and Austin Bautista.

Furthermore, I would like to express my gratitude to my supervisory committee members Drs. Ing Swie Goping and John Mackey. They have always been incredibly insightful, supportive, and available to me in my academic pursuits.

My time in the MD-PhD program thus far has been an incredibly rewarding experience. For their acceptance of me into, and for their continued interest and support of me in this program, I am ingratiated to Drs. Ken Butcher and Alan Underhill. Likewise, many thanks to Janis Davis, Jennifer Freund, Kelsey Robertson, and Kimberly Arndt for constantly keeping me apprised of funding, application, and registration deadlines.

I would also like to thank Drs. Michael Ellison and Nicolas Touret for the research opportunities they afforded me during my time as an undergraduate, and for taking part in my candidacy examination. For agreeing to chair and participate in my thesis defense, I would also like to express my appreciation and gratitude to Dr. David Stuart and Drs. Mohan Babu and Olivier Julien, respectively.

Lastly, my deepest gratitude and appreciation must be expressed to my close friends and family. Without their constant and unwavering support of my academic pursuits I would have no doubt lost momentum on this arduous journey long ago. To my fiancée Kristen Lusk, my siblings Renée Kramer and Adam Kramer, my sister-in-law Carla Kramer, my grandmothers Janet Mazur and Barbara Kramer, the Lusk family, Viola and Phil Sunohara, and most importantly, my parents Colleen and Kim Kramer: Thank you. Words cannot express the reverence I possess for the unrelenting love, support, and encouragement you provide me every day in the betterment of myself and my pursuit of knowledge.

Table of Contents

List of Figures

List of Tables

List of Equations

List of Commonly Used Abbreviations

-omic	proteomics, transcriptomics, metabolomics, genomics
2DE	two-dimensional electrophoresis
3D	three-dimensional
$\vec{a}$	acceleration
Å	angstrom
AA	amino acid
Arg/R	arginine
ACN	acetonitrile
Acsl1	acyl-CoA synthetase long chain family member 1
α	critical value, significance threshold, type I error-rate
ANOVA	analysis of variance
APEX	absolute protein expression
ATGL	adipose triglyceride lipase
a_u	ion trapping parameter, DC potential dependent
AUC	area under the curve
β	type-II error rate
β_u	ion stability solution
BH	Benjamini-Hochberg
BP	biological process
C18	silica-bonded octodecacarbon
C57BL/6	C57 black-6 mouse strain
C; ^{12}C;^{13}C	carbon; non-isotopic carbon; heavy carbon
C_α	alpha carbon of amino acid
CASP	caspase
CAT	cathepsin
CC	cellular component
Ces1	carboxylesterase 1
$CI_{95\%}$	95% confidence interval
CID	collision-induced dissociation
Cnx	calnexin
C_o	carbocylic acid of amino acid
CRT	cathode ray tube
C-terminus	free-carboxylic acid terminus of polypeptide chain
C-trap	C-shaped quadrupole ion trap
$CuSO_4$	copper (ii) sulfate
Da	Dalton
DAVID	database for annotation, visualization, and integrated discovery
DC	direct current
°C	degrees Centigrade
Δ	change in
DF	disease-free

DF_{RL}	disease-free recurrent-like
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DR	disease-recurrent
ECM	extracellular matrix
EDTA	ethylenediaminetetraacetic acid
EIC	extracted ion chromatogram
EL4	lymphoma cell line originating from C57BL/6 mice
emPAI	exponentially modified protein abundance index
ER	endoplasmic reticulum
ER+	estrogen receptor positive
ESI	electrospray ionization
EtOH	ethanol
FDP	false-discovery proportion
FDR	false-discovery rate
FWHM	full-width at half-maximum intensity
g	Gram
g	gravity
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GM	global minimum
GO	gene ontology
H	hydrogen (or mass of hydrogen as 1 mass unit)
h	Hours
H_2O_2	hydrogen peroxide
H_3N	amino group
H4	histone H4
H_a	alternative hypothesis
HCA	hierarchical clustering analysis
HCD	high-energy collisional dissociation
HCl	hydrochloride
HER2-	human epidermal growth-factor receptor 2 negative
HLM	hypotonic lysis medium
H_o	null hypothesis
HPLC	high-performance liquid chromatography
HR	hazard ratio
HRP	horseradish peroxidase
iBAQ	intensity based absolute quantification
ICAT	isotope-coded affinity tags
ICD	image current detection
ICPL	isotope-coded protein labels
IEC	ion-exchange chromatography
IEF	isoelectric focusing
IgA	immunoglobulin-A

IgG	immunoglobulin-G
IgM	immunoglobulin-M
Igs	immunoglobulins
IHC	immunohistochemistry
iTRAQ	isobaric tags for relative and absolute quantification
ITT	ion transfer tube
kDa	kiloDalton
KEGG	Kyoto encyclopedia of genes and genomes
kg	kilogram
Ki-67	ki-67 protein; marker of proliferation
kNN	*k*-nearest neighbours imputation
KPNA2	importin subunit alpha 1
LC	liquid chromatography
LCM	laser-capture microdissection
LC-MS/MS	liquid chromatography with tandem mass spectrometry
LD	lipid droplet
lFDR	local false-discovery rate
$LiCO_3$	lithium carbonate
LIT	linear ion trap
LM	local minimum
LOD	limit of detection
Log_{10}	logarithm, base-10
Log_2	logarithm, base-2
LTQ	linear triple quadrupole
Lys/K	Lysine
m	Mass
M	parent molecule mass
m/z	mass-to-charge ratio
MA	mass accuracy
MALDI	matrix-assisted laser-desorption ioniation
MAR	missing at random
MCAR	missing completely at random
MCP	multiple comparisons problem
MF	molecular function
mg	milligram
MICE	multivariate imputation with chained equations
min	minute; minimum
MIP18	mitotic spindle-associated MMXD complex subunit MIP18
mL	millilitre
MLE	minimum least estimate
mM	millimolar
MNAR	missing not at random
MRM	multiple-reaction monitoring

mRNA	messenger ribonucleic acid
MS	mass spectrometry
MS/MS or MS^2	tandem mass spectrometry
MSC	minimum set cover
mu	mass units
MUP/Mup	major urinary protein
MV	missing value
MW	molecular weight
MWU	Mann-Whitney U-test
N; ^{14}N; ^{15}N	nitrogen; non-isotopic nitrogen; heavy nitrogen
nm	nanometer
NPC	normal-phase chromatography
NSAF	normalized spectral abundance factor
NSI	nanospray ionization
NSPs	number of sibling peptides
N-terminus	free-amine terminus of polypeptide chain
OH	hydroxyl group
Ω	radial frequency of V_{RF}
ω	ion secular frequency
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PAI	protein abundance index
PANTHER	protein analysis through evolutionary relationships
PAT	perilipin-1, ADRP/Perilipin-2, TIP47/Perilipin-3
PBS	phosphate-buffered saline
PBST	phosphate-buffered saline – tween 20
PCA	principal component analysis
PCR	polymerase chain reaction
PDCD6	programmed cell death protein 6
Pemt	polytopic membrane protein phosphatidylethanolamine *N*-methyltransfe
pFDR	positive false-discovery rate
PGM3	phosphoacetylglucosamine mutase
PgR+	progesterone receptor positive
pH	*pondus hydrogenii*
PHYHD1	phytanoyl-CoA dioxygenase domain-containing protein 1
pI	isoelectric point
PIA	probabilistic inferrence algorithm
PIGR	polymeric immunoglobulin receptor
π_0	null-distribution
PITC	phenyl isothiocyanate
Plin2	perilipin-2
Plin5	perilipin-5
PMSF	phenylmethylsulfonyl fluoride

PPAR	peroxisome proliferator-activated receptor
ppm	parts-per-million
PRIDE	proteomics identifications database
PSM	peptide-spectrum match
PTH	phenylthiohydantoin
PTM	post-translational modification
p-value	probability value
q	charge of ion
QqQ	triple quadrupole
Q-ToF	quadrupole-time-of-flight
qu	ion trapping parameter, V_{RF} dependent
q-value	Storey-Tibshirani literature FDR adjusted p-value
R	resolution
$R_{1,2}$	electrode radii (orbitrap)
RF	radio-frequency
R_m	characteristic radius of electric field (orbitrap)
RNA	ribonucleic acid
r_0	radius of electric field
ROC	receiver operating characteristic
RP	resolving power
RPC	reverse-phase chromatography
RT-qPCR	reverse-transcription quantitative polymerase chain reaction
SC	secretory component
SCX	strong-cation exchange chromatography
SDS	sodium dodecyl sulfate
SEC	size exclusion chromatography
Σ	sum of
SILAC	stable isotope labelling of amino acids in cell-culture
SINQ	normalized spectral index quantitation
SRM	selected reaction monitoring
ST	Storey-Tibshirani
STRING	search tool for retrieval of interacting genes/proteins
t	time
TAILS	terminal-amine isotopic labeling of substrates
TBS	tris-buffered saline
TBST	tris-buffered saline – tween 20
TCA	tricarboxylic acid
TENN	tenascin-N
TG	triacylglycerol
theta	electric potential
TIC	total ion current
TK-HSD	Tukey-Kramer honest significant difference test
TMA	tissue microarray

TMT	tandem-mass-tags
TMX1	thioredoxin-related transmembrane protein 1
TNBC	triple negative breast cancer
ToF	time of flight
$\Psi(t)$	applied alternating potential
U	DC voltage
µg	microgram
µL	microlitre
µm	micrometer
UniProtKB	universal protein resource knowledgebase
UPLC	ultra-performance liquid chromatography
v/v	millilitres per 100 millilitres
VRF	radio-frequency voltage
w/v	grams per 100 millilitres
z	integer charge number

Chapter 1 : Fundamentals of Proteomics

1.1. Overview of Proteomics

Originally coined in 1994 by Marc Wilkins of Macquarie University, the term 'proteome' – a portmanteau of *protein* and *genome* – refers to the complete composition and quantity of proteins expressed within a biological system[1]. Making up the largest abundance of macromolecules within living organisms [2–4], virtually all biological processes within cells are carried out by, or with the aid of, proteins. Physical and functional associations of proteins with other biomolecules (such as DNA, RNA, other proteins, lipids, small molecule metabolites, or some combination thereof), orchestrated at specific times and locations within the cellular environment form the backbone of biology; whether through the enzymatic catalysis of biochemical reactions, or the provision of organellar and cellular architecture and transport systems, proteins provide a means of order in the otherwise chaotic and disordered chemical soup we have come to call life.

With the advent of modern nucleotide sequencing techniques – DNA and RNA alike – and the completion of several genome sequencing projects[5–7], the fields of modern genomics and transcriptomics have been able to garner an impressive amount of data regarding the sequences of protein-coding genes. This has been used to compile expansive databases of confirmed and putative protein sequences[8–10], in turn providing comprehensive proteome maps for a multitude of organisms. While the substituents of a proteome can be identified by an organism's coding genome, proteomes are highly dynamic in nature. The composition of a system's proteome is not only dependent on the genetic makeup of the cell, but also a myriad of factors and processes including but not limited to spatial[11–13] and temporal[14] location, epigenetic regulation[15], cell-cycle progression[16,17], rate of protein translation[18], protein half-life[19,20], cell-type[21,22], cellular metabolism and energy demands[23,24], and the cellular/organism environment[25–27].

While such a degree of complexity and individual factor-based variability can seem staggering from a research perspective, the changes that occur within proteomes on a per-factor basis are precisely what has enabled the field of proteomics to rise to such prominence over the past several years. By performing comparative analyses of proteomes in response to a specific condition – whether it be, in a broad sense, exposure to a chemical stimulus, environmental condition, disease state, or simply time – a vast amount of information can be elucidated about a biological

system. How proteins function as a collective can be determined; cross-pathway connectivity can be observed, providing a means of understanding not only individual protein functionality, but also how protein networks interact to transform the cellular landscape to best suit the cells' new environment. Information at the whole-proteome level is capable of highlighting proteins, pathways, and networks of importance during processes such as cellular differentiation, disease pathogenesis, metabolism, division, and death, providing a new global understanding of not only how cellular processes are carried out, but also a means to manipulate them.

The comprehensive study of proteomics has historically been incredibly tasking. Prior to the 1980s, proteins were identified on an individual basis after painstaking isolation through techniques most often involving electrophoresis, the most prominent being two-dimensional gel electrophoresis (2DE)[28,29]. Using 2DE, proteins are first separated based on their isoelectric properties via isoelectric focusing (IEF); using an immobilized pH gradient in a polyacrylamide gel and an electric current, proteins migrate to the pH at which they are electrostatically neutral. While quite sensitive, IEF suffers when applied to membrane proteins (low solubility), or proteins' whose pIs are incredibly alkaline (i.e. pI > 11, ~ 3% of the proteome)[30]. In addition, post-translational modifications (PTMs) can alter the pI of a protein species[31], altering migration between runs; while useful if studying a very limited range of proteins, when analyzing large proteomic datasets this can result in cross-contamination of 'spots' on the gel (PTMs can cause one protein to migrate the same as another, resulting in two species residing in the same position on the gel). Following IEF, proteins are then separated based on size via sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)[32]. Like IEF, SDS-PAGE suffers similar problems of protein migration, but with respect to low separation of very large MW proteins; this ultimately led to the development of discontinuous gels to increase resolving power[33].

Following a 2DE separation, gels are visualized and isolated proteins identified by N-terminal sequence degradation ('Edman degradation'), a technique developed by Pehr Edman in 1950[34]. Edman degradation utilizes phenyl isothiocyanate (PITC) to selectively cleave the N-terminal AA of a peptide of length *n*, resulting in the formation of a thiazolinone derivative containing the N-terminal AA of the peptide, and the residual peptide chain of length *n-1*. The thiazolinone

derivative is then extracted and heated under acidic conditions to produce a phenylthiohydantoin (PTH)-AA derivative, which can be analyzed via chromatography to determine the AAs' identity (**Figure 1.1**). By repeating the process, the sequence of a peptide can be determined residue-by-residue, and the resultant peptide sequence can be used to identify the protein in question. Although the process was automated in 1967[35,36], this method has several downfalls; in addition to being laborious and time-intensive, a relatively large amount of starting material with high purity is needed, and the process begins to fail for peptides exceeding 50 residues. Another problem with this technique occurs when studying eukaryotic proteins – most eukaryotic proteins' N-terminal residues are acetylated, preventing nucleophilic attack of the N-terminal amine by PITC. To circumvent these problems, proteins would need to be digested prior to analysis, and the resultant peptides purified, increasing the amount of time needed per protein being identified.

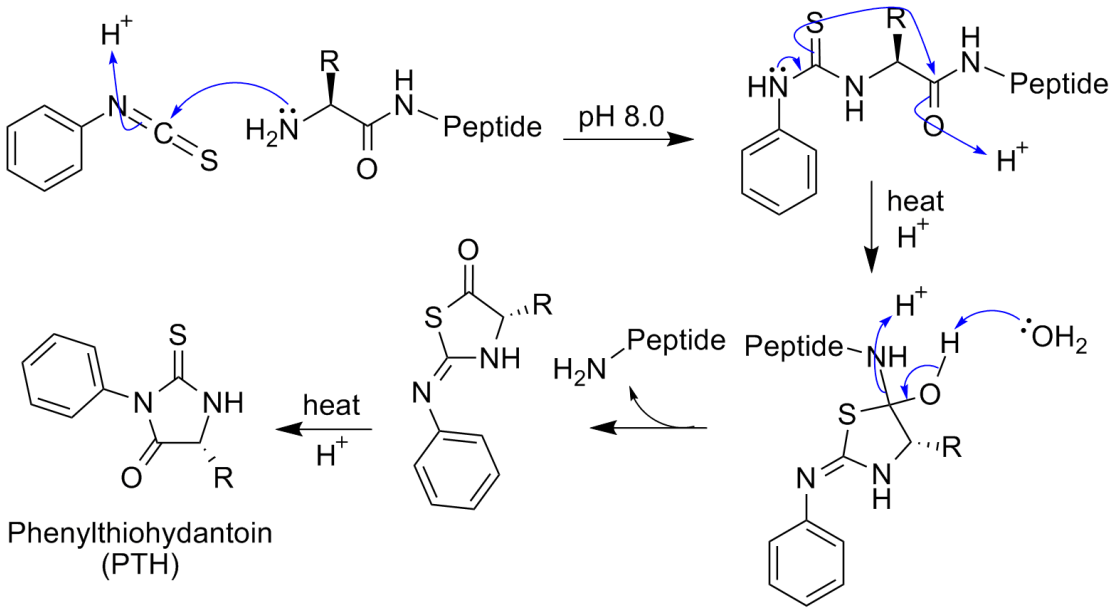


Figure 1.1 Edman degradation.

Under slightly basic conditions, the N-terminal amine's lone pair acts as a nucleophile and reacts with the thiocyanate group of PITC, forming a phenylthiocarbamoyl derivative. Heating this product under acidic conditions results in the formation of a thiazolinone derivative containing the N-terminal residue of the peptide, and release of the remainder of peptide chain. Extraction and acidification of the thiazolinone-AA results in the formation of a PTH-AA conjugate which can be identified through chromatography.

Over the next several decades, computerization began to revolutionize and accelerate the biotechnology industries, outpacing traditional separation methods and chemical sequencing techniques such as Edman degradation. Thus, the field started turning towards faster, more promising methods of molecular separation and identification, that in turn also provided higher sensitivities, signals, and dynamic ranges of detection – in particular, liquid chromatography and mass spectrometry.

1.2. Basic Principles of Mass Spectrometry in Proteomics

1.2.1. The Origins of Mass Spectrometry

Simply put, mass spectrometry is a technique that measures an ionized molecule's mass-to-charge ratio (m/z) via an electric field. Birthed in the late nineteenth-century to study the properties of electricity, early mass spectrometers were relatively crude instruments consisting of a cathode ray tube (CRT) – a vacuum tube with a free cathode (negative electrode) opposite a phosphorescent screen – coupled to an electromagnetic field. In 1886, Eugen Goldstein discovered that filling a CRT with atmospheric gas produced rays that behaved differently from cathode rays (electrons)[37,38]; these rays travelled in the opposite direction to the cathode rays, and as such, were named 'anode rays' or 'canal rays'. In 1898, Wilhelm Wein discovered that canal rays could be deflected in the opposite direction of cathode rays using electric and magnetic fields[39]; these rays were determined to have a positive charge, and a mass much higher than that of the electron (determined one year prior by J.J. Thomson[40]), but rather closer to that of hydrogen atoms. Utilizing Wein's method, J.J. Thomson noticed that the positive rays followed a unique parabolic path when their respective canal rays were deflected through an electric field, and was surprised to see several different species[41,42]. J.J. Thomson determined that if one m/z ratio was determined, the m/z ratios for the other positively-charged species he had observed could be calculated. Over the next two decades J.J. Thomson and Francis William Aston's (Thomson's student) work would lead to the discovery of isotopes[38,42,43], and lay the groundwork for Arthur Jeffery Dempster's mass spectrometer design in 1918[44] and Aston's in 1919[45]; Aston and Dempster's basic theory and principles are still used to this day.

By the mid-1980s, several scientists were trying to utilize mass spectrometry for the analysis of large molecules, such as polymers and proteins. This could possibly allow for the circumvention of chemical peptide sequencing for the identification of proteins. There was a problem however; mass spectrometry requires analytes to be ionized and in the gaseous state. The difficulty then, was being able to reliably and reproducibly ionize chemical species into the gaseous state that are as massive as an entire protein. Dr. John B Fenn, an American chemist who worked in the field

of mass spectrometry was quoted as saying, "the idea of making proteins or polymers "fly"...seemed as improbable as a flying elephant"[46]. However, two breakthroughs helped to pave the way for making mass spectrometry a suitable technique for studying large molecules.

The first came in 1985, with Franz Hillenkamp and Michael Karas' development of matrix-assisted laser-desorption ionization (MALDI)[47]. Using MALDI, samples are first co-crystalized with a photo-ionizable chemical matrix. Irradiation of this matrix with laser-light causes desorption of the matrix-sample mixture and ionization of the sample species, rather than decomposition[48,49]. Hillenkamp and Karas' method initially was only able to ionize peptides up to ~2.8 kDa, using 266nm laser-light and matrices of simple aromatic biomolecules such as tryptophan. In 1988, three years after the development of MALDI, Koichi Tanaka published a refined version of Hillenkamp and Karas' method, by combining 30nm cobalt particles with glycerol and irradiation via a 337nm nitrogen laser[50]; this method enabled Tanaka to be able to ionize molecules as large as 34.5 kDa.

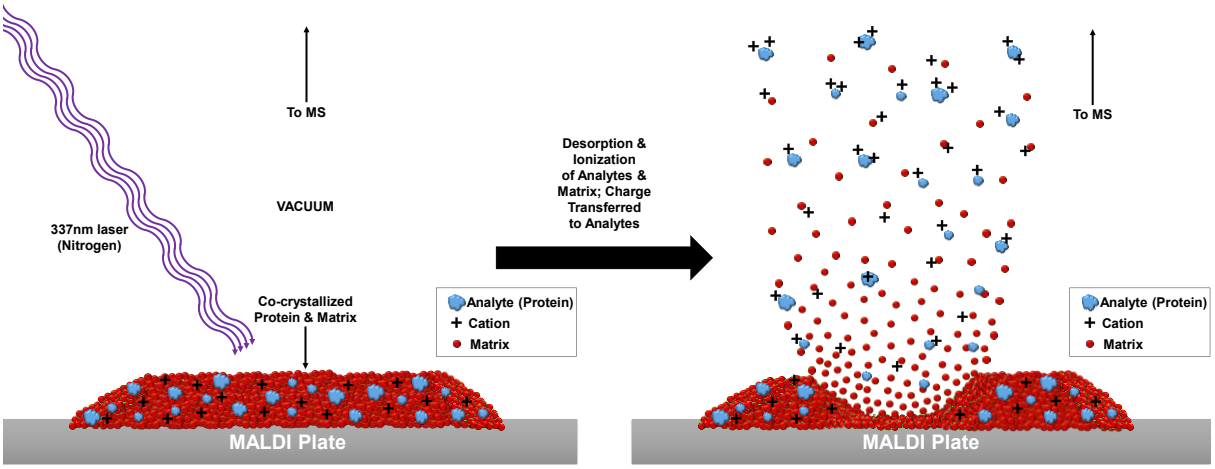


Figure 1.2 Matrix-assisted laser desorption ionization (MALDI).

Proteins co-crystallized with a photo-absorptive matrix are spotted on a metallic plate and irradiated with a 337nm laser in an electric field under vacuum (**left**). This results in rapid desorption and ionization of the matrix and analytes, respectively (**right**).

The second breakthrough was the development of electrospray ionization (ESI) by John B Fenn and Masamichi Yamashita[51]. First described in 1984, ESI makes use of analytes dissolved in an

ion-containing solvent; the solvated sample is nebulized through a fine-tip aperture into a high-voltage electric field. As the aerosol drop travels through the electric field, solvent gradually evaporates from the drop, concentrating the charge. Eventually the electrostatic repulsion exceeds the surface tension of the droplet (the 'Rayleigh limit' – first theorized by Lord Rayleigh in 1882[52]) resulting in a 'Coulomb explosion' and ionization of all solvated species into the gaseous state. By 1989 John Fenn and his colleagues were easily capable of ionizing biomolecules as large as 76 kDa[53,54]. For their work developing methods to ionize large biomolecules, Koichi Tanaka and John Fenn shared part of the 2002 Nobel Prize in Chemistry.

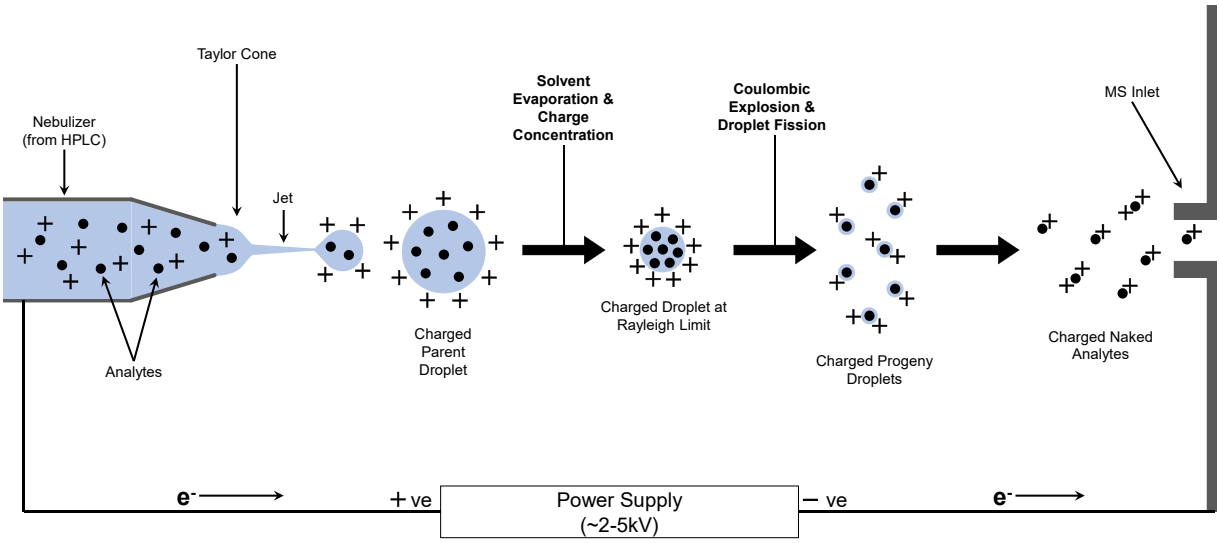


Figure 1.3 Electrospray ionization (ESI).

Analytes eluting from the liquid chromatography (LC) column are subject to electrical polarization via a potential applied between the nebulizer head (LC outlet) and MS inlet. Electrically charged droplets emitted from the nebulizer undergo rapid solvent evaporation and charge concentration, until the charges reach the Rayleigh limit. Droplet fission occurs when the electrostatic repulsion of surface charges exceeds that of the droplet's surface tension, resulting in ionization of analytes. Adapted from [55]

Since the development of reliable methods of protein ionization, mass spectrometers have become instrumental in the field of proteomics. *De novo* sequencing[56,57] – a technique which involves the fragmentation of peptides into their constituent AAs – has made rapid protein identification possible, leading to an explosion in techniques for analyzing protein structure[58],

post-translational modifications[59,60], protein-protein interactions[61,62], protein quantitation[63,64], and more recently whole-proteome quantitation[65–68].

1.2.2. The Mass Spectrum

As previously mentioned, mass spectrometers measure the mass-to-charge ratio (m/z) of an ion[69]. This is an important distinction from an ion's individual mass; a molecular ion that is doubly charged (±2) will have an m/z that is half of the same molecule that is singly charged (±1). As charge-state can often be defined as the loss or gain of a proton (or other ionic species), the mass-to-charge ratio for positive ions (gain of protons) is:

$$(\boldsymbol{m/z}) = \frac{[\mathbf{M} + \boldsymbol{z}\mathbf{H}]}{\boldsymbol{z}}$$

Equation 1.1

And for negative ions (loss of protons) is:

$$(\boldsymbol{m/z}) = \frac{[\mathbf{M} - \boldsymbol{z}\mathbf{H}]}{\boldsymbol{z}}$$

Equation 1.2

where M represents the species' molecular mass in Daltons (Da), H represents the molecular mass of a proton (1 Da), and z is the integer value of the ion's charge-state. When analyzing a singular molecule represented by various m/z ratios, it becomes relatively simple to determine the charge-state of the molecule, as well as its molecular mass. ESIProt[70], a popular mass spectrometry tool, is designed to perform exactly this procedure, albeit with high accuracy. An example of this procedure, for a modified troponin-C molecule is depicted in **Figure 1.4**.

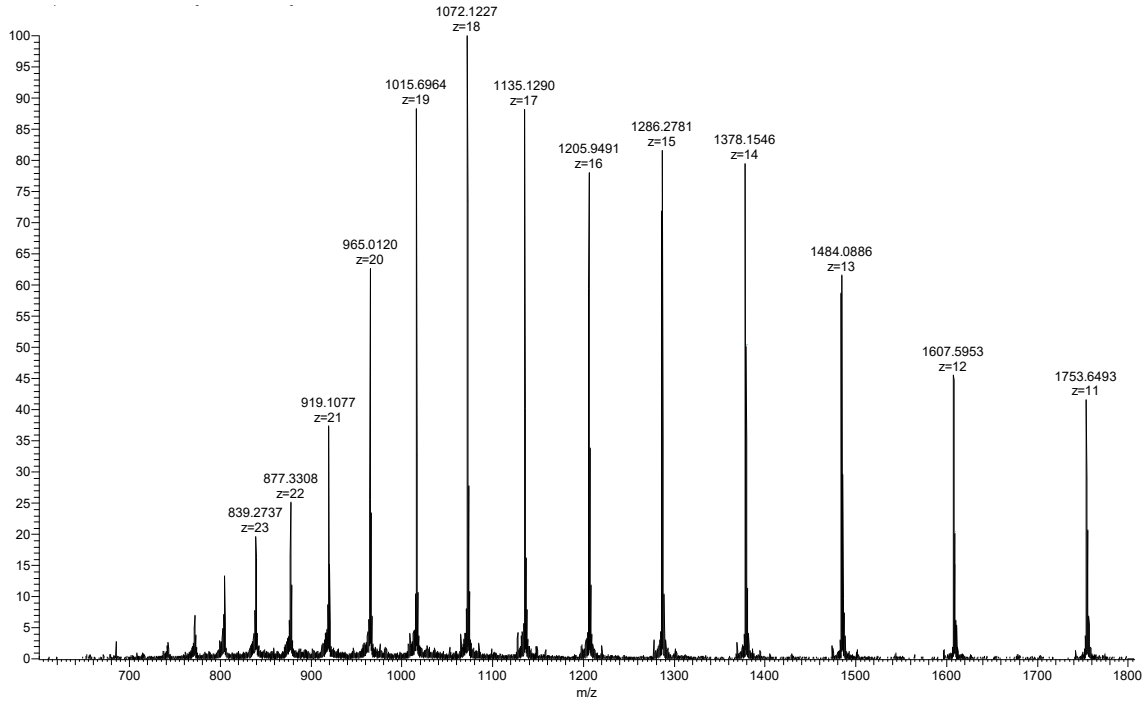


Figure 1.4 Determining mass from multiple charge states.

Mass spectrum resulting from multiple charge states of tagged troponin-C. The peak at 1753.6493 mass units (mu) represents the lowest charge; increasing the $\boldsymbol{z}$-value results in a decrease in m/z. Utilization of ***Equation 1.1*** allows for the determination of both z-states and m. Setting $1753.6493 = [(\boldsymbol{m} + (\mathbf{1}) \times \boldsymbol{z})/\boldsymbol{z}]$, and $1607.5953 = [(\boldsymbol{m} + (\mathbf{1}) \times (\boldsymbol{z} + \mathbf{1}))/(\boldsymbol{z} + \mathbf{1})]$, and continuing this process for subsequent peaks, reveals the $z_{1753.6493} = +11$, $z_{1607.5953} = +12$ and so forth. Utilization of these z-states allows for the mass of the modified troponin-C to be determined to be ~ 19279.5105 with a standard deviation of 0.5260 Daltons.

One of the features that distinguishes mass spectrometry from other analytical techniques is resolution and accuracy. Mass spectrometers have a history of possessing incredibly high resolving power and resolution, capable of separating a molecule's average mass into it's isotopic distribution. In fact, early mass spectrometers led to the discovery of isotopes[38,42,43]. Mass spectrometers' resolution[71–73] can be defined as:

$$\mathbf{R} = \frac{\boldsymbol{m}}{\boldsymbol{\Delta m_{FWHM}}}$$

Equation 1.3

where the resolution (R) of an m/z peak is the average mass (m) divided by the width of the peak at 50% the peak's intensity (full-width at half maximum; Δm_{FWHM}). Resolving power[72–74] represents the instrument's ability to distinguish between two ions (of similar intensity) separated by a small increment, measured at a peak height where the 'valley' between them reaches no more than 10% the maximum height of either peak. This can be calculated as:

$$\mathbf{RP} = \frac{\boldsymbol{m_2}}{(\boldsymbol{m_2} - \boldsymbol{m_1})}$$

Equation 1.4

where resolving power (RP) between two m/z peaks is calculated as the measured mass of the heavier peak (m_2) over the difference between the heavier and lighter (m_1) peaks. If not using the '10% valley' rule, the degree of overlap between the two peaks must be stated.

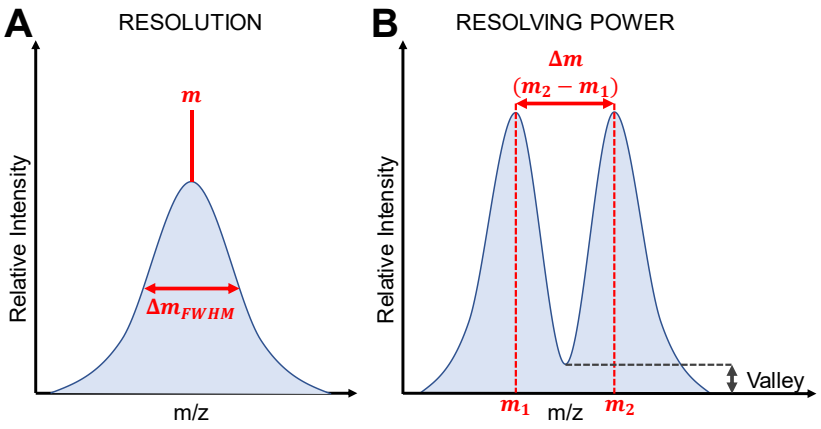


Figure 1.5 Mass resolution *versus* resolving power.

(**A**) Resolution refers to the measured mass divided by the peak's full width at 50% max intensity (FWHM). (**B**) Resolving power refers to the ability to differentiate two partially overlapping peaks, when the valley between them is ~10% the maximum relative intensity.

The accuracy of mass spectrometers, known as 'mass accuracy' or 'mass error', is typically measured in parts per million (ppm). Mass accuracy (MA)[73] is calculated as the difference in the experimentally measured 'exact mass' (m_E) of a monoisotopic species (ion containing no heavy isotopes) from its theoretical 'actual mass' (m_A), divided by the theoretical mass:

$$\mathbf{MA} = \frac{(\boldsymbol{m_E} - \boldsymbol{m_A})}{\boldsymbol{m_A}} \times \mathbf{10^6}$$

Equation 1.5

For mass spectrometers utilized in the field of proteomics, most have dynamic ranges for resolution, which is dependent on the time allotted for performing a scan. For newer instruments, the range of resolution is typically anywhere from 30,000 to ≥100,000[75–77], easily allowing for separation of the various isotopic species of a single ion species. Likewise, the accepted margin of error is typically <±10ppm. Such high resolving power has made the determination of mass and charge relatively simplistic for an isotopic m/z cluster; the gap between peaks within an isotopic cluster will differ by a single Dalton, and this difference can be used to determine the monoisotopic mass and charge state; as all the peaks will differ by a single mass unit, the charge state will be the reciprocal of the measured distance between the peaks (**Figure 1.6**).

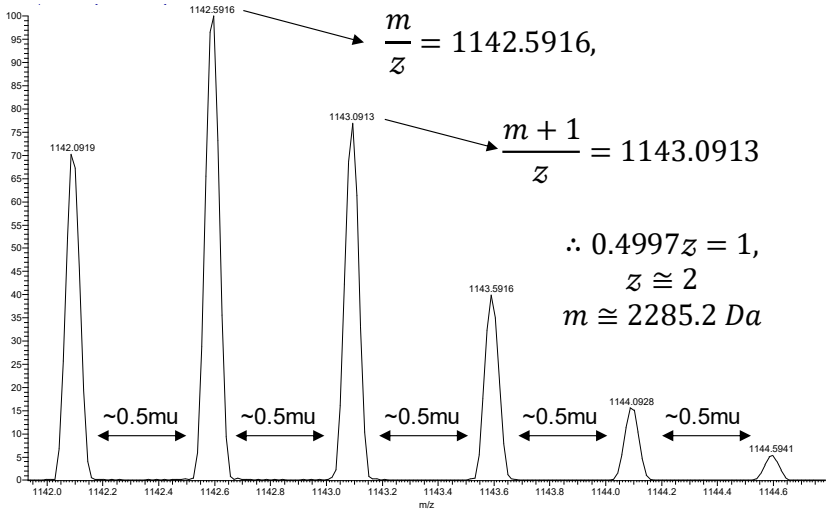


Figure 1.6 Mass determination of an isotopic cluster.

Using the approach outlined above, the first peak in an isotopic series is determined to be the monoisotopic mass. Setting this as the m/z, and the second peak as $(m+1)/z$, and isolating m allows for the determination of z, which can then be used to determine m. As a rule of thumb, because z is constant for an isotopic cluster and mass peaks differ by 1 neutron, the gap between peaks is $1/z$.

1.2.3. Modern Mass Spectrometry in Proteomics

All modern mass spectrometers consist of three components – an ion source, a mass analyzer, and an ion detector[78]. While only a select few types of ion detectors are utilized, there exists a high diversity and modularity of ion sources and mass analyzers, resulting in a plethora of varying types of mass spectrometers. Within the field of proteomics however, a select few types

of instrumentation have become predominant. These include the ion sources: MALDI, ESI, and nanospray ionization (NSI), and the mass analyzers: time-of-flight (ToF), quadrupole-ToF (Q-ToF), triple-quadrupoles (QqQ), linear ion-traps, and orbitraps[78]. Furthermore, the type of ion source and mass analyzer used is dependent on what kind of an approach is being taken; top-down versus bottom-up ("shotgun") proteomics[79,80].

Top-down proteomics is an approach used to analyze intact proteins to reveal information regarding individual proteins' structure, post-translational modifications, interacting partners, and functional proteoforms (all variations of the protein product from a single expressed gene)[81,82]. Using this method, proteins are purified and immediately analyzed via mass spectrometry. Classically this technique utilized MALDI-ToF instruments, but recent advances in linear ion-trap/orbitrap mass resolution has led to the implementation of ESI/NSI ionization with linear ion-traps/orbitraps.

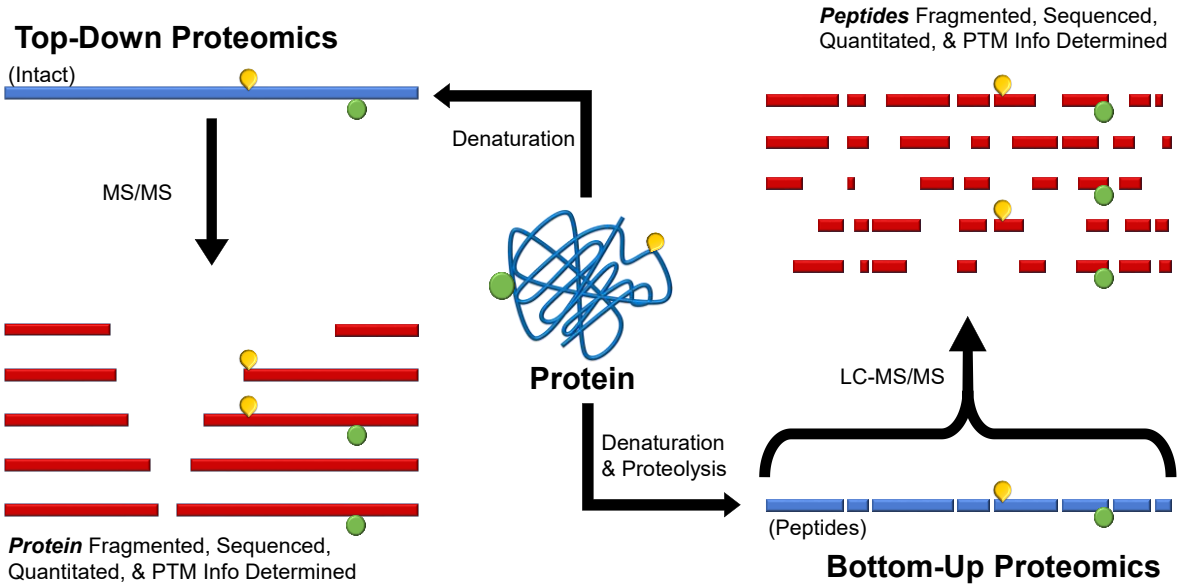


Figure 1.7 'Top-Down' *versus* 'Bottom-Up' proteomics.

Experimental designs for typical top-down (**Left**) and bottom-up (**Right**) proteomics approaches. For top-down, intact proteins are denatured and directly subject to tandem mass spectrometry (MS/MS) analysis, allowing for the determination of structural and PTM characteristics of a protein. For bottom up, proteins are first denatured, then digested into constituent peptide mixtures. These peptide mixtures are separated via LC and subject to MS/MS, allowing the peptides to be sequenced and quantified.

Bottom-up proteomics (colloquially known as a 'shotgun' approach) is a technique where proteins are first proteolytically digested prior to mass spectrometric analysis[83]. Due to the complex mixture of peptides resulting from the digest, bottom-up proteomics requires better front-end separation of analytes; peptide mixtures are simplified through chromatography prior to analysis. This technique greatly increases the coverage of proteins analyzed and allows for the more precise quantification of expressed proteins; by integrating multiple peptides from various proteoforms (post-translationally modified variants) of an individual gene and mapping them to a single peptide map, a comprehensive view of a genes' functionally expressed protein(s) is acquired[83]. Therefore, by sacrificing information regarding the individual proteoforms of a specified gene, total quantification of the gene's protein products is achieved. As mentioned earlier, due to this technique requiring proteolytic digestion and subsequent peptide mixture simplification, ionization is almost always achieved through ESI/NSI operating in tandem with chromatography. Bottom-up proteomics typically utilizes ion-trapping mass analyzers such as linear ion-traps, orbitraps, or a combination of the two[83].

All work described in this thesis was performed using a bottom-up proteomics approach utilizing an ultra-performance or high-performance liquid chromatograph (UPLC or HPLC) in tandem with NSI into an orbitrap or linear ion-trap/orbitrap combination mass spectrometer; as such, only principles and theories applying to the instruments used will be discussed in detail.

1.2.3.1. Proteolytic Digestion of Proteins into Peptides

When using a bottom-up proteomics approach, the first step following protein purification is to perform a proteolytic digest of the sample[83]. This is most often performed with charged-residue sequence-specific proteases such as trypsin, Lys-C, Arg-C, Glu-C, Asp-N, or Lys-N; utilization of these proteases results in the cleavage of polypeptides either following or preceding a charged AA (denoted by 'C' for the C-terminal, or 'N' for the N-terminal side of the named residue, respectively)[84]. By cleaving the peptide backbone with a terminally charged residue, this guarantees at least a ±2 charge on the resulting peptide product (following acidification/alkalinisation), allowing for easier solvation and subsequent ionization. Trypsin is

by far the most commonly used protease for protein digestion[83,84]. Having a sequence specificity for cleavage of the peptide backbone on the C-terminal side of lysine (K) or arginine (R) (except when either residue is followed by a proline), this results in relatively short peptides that, when solubilized in an acidic solution, yields a minimum of a +2 charge per peptide (**Figure 1.8**). For a list of commonly used proteases for protein digestion in mass spectrometry, including their cleavage sites, refer to (**Table 1.1**).

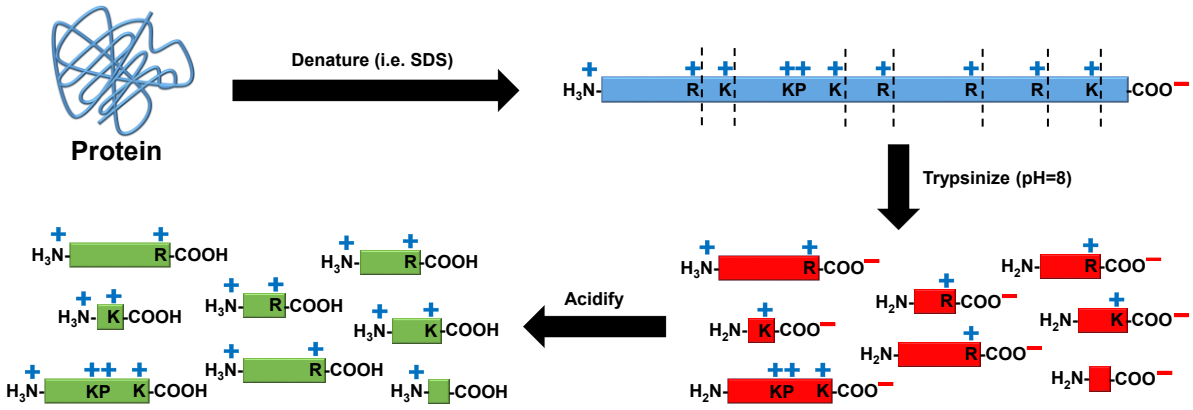


Figure 1.8 Trypsinization of a polypeptide.

Denatured proteins are subject to tryptic digestion, being cleaved on the C-terminal side of Lys/Arg (except when followed by Proline). This results in the formation of short oligopeptides with at least a single positively charged residue. Acidification of these oligopeptides causes any free carboxyl groups to gain a proton and become neutral, while N-terminal amines gain a proton, becoming positively charged. As a result, all tryptic peptides possess at least a +2 charge.

Table 1.1 Commonly used proteases in proteomics. (Adapted from[84])

FAMILY	PROTEASE	CLEAVAGE SITE	ADVANTAGES	DISADVANTAGES
ASPARTIC PROTEASE	Pepsin	C-term of Y/F/W	• Determine disulfide bond sites • Active at low temp; allows for ^{2}H exchange experiments	• Specificity is pH-dependent • Peptides hard to interpret
CYSTEINE PROTEASE	ArgC	C-term of R	• Allows investigation of PTMs	• Long peptides
METALLO-PROTEASES	AspN	N-term of D	• Cleavage site • Active pH-range is broad (4-9)	• Detergents alter site specificity • Long peptides
	LysN	N-term of K	• Resistant to detergents	• Long peptides
	LysargiNase	N-term of R/K	• Mirrors trypsin cleavage	• Expensive
SERINE PROTEASES	GluC	C-term of D	• Allows investigation of PTMs	• Activity highly dependent on pH and buffers • Long peptides
	LysC	C-term of K	• Efficient & specific	• Long peptides
	Chymotrypsin	C-term of F/Y/L/W/M	• Complements trypsin • Good for membrane proteins	• Efficiency varies for different AAs
	Trypsin	C-term of R/K	• Efficient & specific • Inexpensive • Gold Standard	• Short Peptides • C-term peptides hard to see

1.2.3.2. Liquid Chromatographic Separation and Ionization of Peptide Mixtures

The resultant peptide mixture following proteolysis is incredibly complex. To resolve this issue, peptide mixtures are subject to liquid chromatographic separation. Chromatography refers to the physical separation of a mixture of chemical species through the utilization of a mobile phase moving in a singular, continuous direction over a stationary phase (column)[85,86]. The chemical species being separated are placed into the mobile phase and separation is achieved through the species' interaction with the column, impeding elution. As different species have different partitioning coefficients, the degree to which they are impeded by the stationary phase varies. Over time, the characteristics of the mobile phase can be changed to favour highly impeded species to partition back into the mobile phase and elute from the column (gradient elution); while not necessary, gradient elution is widely utilized as it results in 'tightening' of individual species eluting from the chromatograph[87]. Alternatively, with isocratic elution the characteristics of the mobile phase remain constant. This results in species with high partitioning coefficients

for the stationary phase eluting over a much longer timeframe, and ultimately being incredibly dilute[87]. Chemical species can be separated based on various properties – notably their size (size exclusion chromatography; SEC)[88], charge state (ion-exchange/strong-cation exchange chromatography; IEC/SCX)[89,90], hydrophilicity (normal-phase chromatography; NPC)[91], or hydrophobicity (reverse-phase chromatography; RPC)[85–87,92].

While peptide mixtures can be separated using any of the chromatographic separation methods, the most common – RPC – utilizes peptides' hydrophobicity[85–87,92]. In theory, following digestion with a protease such as trypsin (trypsinization), where the digestion proceeds to 100% completion, all peptides should have a minimum charge of +2 (or +4 charge where R/K is followed by P), rendering separation via IEC obsolete. Likewise, SEC's low resolving power is insufficient for separating small peptides, and many peptides are insoluble in the organic solvent starting conditions required for NPC. These factors, in addition to RPC's versatility and superb resolving power when dealing with peptides has resulted in its dominance for both analytical and preparatory techniques[80,86].

When used in tandem with mass spectrometers, the obvious choice for ionization of peptide analytes is through ESI[80,86], and more recently NSI[93,94]. As described earlier, spray ionization utilizes analytes solvated in an ion-rich solution. When dealing with proteolyzed peptides, these tend to be protons, generated through a decrease of the sample solution pH. The physical behaviour of emerging eluent from the chromatograph's capillary outlet (nebulizer) is governed by several forces, including the flow rate, the dimensions of the nebulizer opening, the surface tension of the droplet, and the voltage applied between the capillary outlet and the mass spectrometer inlet[95]. Voltage applied between the nebulizer head and MS inlet results in polarization of solvated ions. Initially, the surface tension of the liquid being emitted from the nebulizer (p_γ) exceeds that of the electrostatic pressure (p_E). As voltage is increased however, $p_E > p_\gamma$, resulting in formation of a Taylor cone, and the jettison of liquid in the form of positive-ion containing droplets from the tip of the cone; these droplets travel towards the negative potential at the MS inlet, during which they experience rapid solvent evaporation. As the solvent evaporates, the ions are concentrated until electrostatic repulsion between ions exceeds the

surface tension of the droplet, resulting in the 'Coulombic explosion' leading to droplet fission and emanation of the positive ions into the gaseous state[95].

Due to the relatively gentle nature of the ionization process ('soft ionization'), large biomolecules stay intact, and thus, pick up more charges from solution; this effectively extends the mass range which mass analyzers can work, easily facilitating the analysis of molecules in the kDa range up into the MDa range. The key differences between ESI and NSI arise from NSI utilizing a smaller aperture (~2μm for NSI vs ~300μm for ESI) [93–95], chromatographic flow rates approximately an order of magnitude lower than ESI, and a greatly reduced distance between the nebulizer and MS inlet; the initial droplet size formed from NSI are estimated to be approximately ~180nm in diameter, versus ESI's being in the μm range[93–95]. Concomitantly, NSI has been shown to yield 'cleaner' mass spectra; reduced droplet size increases ionization yields due to a decreased dependency on solvent evaporation, and an increased droplet readiness for fission.

1.2.4. Mass Analysis of Peptide Ions

Following ESI/NSI of peptides into the gaseous state, they are subject to mass analysis. While a myriad of mass analyzers exists, the bottom-up proteomics techniques have made utilization of a select few in recent years; ion-traps, orbitraps, or hybridized systems[80,86]. Within such instruments, ions enter through an ion transfer tube (ITT), and proceed to be filtered and focused through what are known as the mass spectrometer's 'optics'[96]. These optics are typically comprised of several electrodes with applied radio-frequency (RF) voltage, operating in series and serve to accomplish several things; i) focusing and directing of the ion beam through the internal components of the mass spectrometer; ii) remove neutral-species generated through the ESI/NSI process that contaminate the ion beam; iii) filtering out of ions with undesired m/z ratios; and iv) accelerate or decelerate the ions in the ion beam. The culmination of processes ultimately increases the signal-to-noise ratio of the resultant mass spectra. Ions that successfully pass through the optics are finally 'trapped' and scanned by the analyzer. However, the way linear ion traps and orbitraps perform scans for, and detect ion species greatly differs.

1.2.4.1. Quadrupole Linear Ion Trap Mass Analyzers

Quadrupole ion traps – invented by Wolfgang Paul in 1953[97], earning him the Nobel Prize in Physics in 1989[98] – trap ions via oscillating electric fields generated through a combination of radio frequency (RF) and static direct current (DC) voltages. The original 3D quadrupole ion trap (also known as a 'Paul' trap) consists of two hyperbolic electrodes – charged with DC potential – separated by a ringed electrode with an applied RF potential (**Figure 1.9**). A variant of this original 3D Paul trap, linear ion traps adhere to the same physical principles, albeit with the use of three series of four electrodes (quadrupoles) running in parallel[99].

Ions entering the linear ion trap are initially slowed and ultimately deflected back by a DC 'capping' voltage applied to all 4 electrodes of the terminal quadrupole; after entrance, the initial quadrupole's voltage is also 'capped' with a static DC potential. Thus, the ions become trapped in the central quadrupole, as movement in the z plane becomes inhibited. To understand ion movement within the x and y planes, we must define the electric potential over time within a two-dimensional quadrupole electric field $[\phi_2(x, y, t)]$[99].

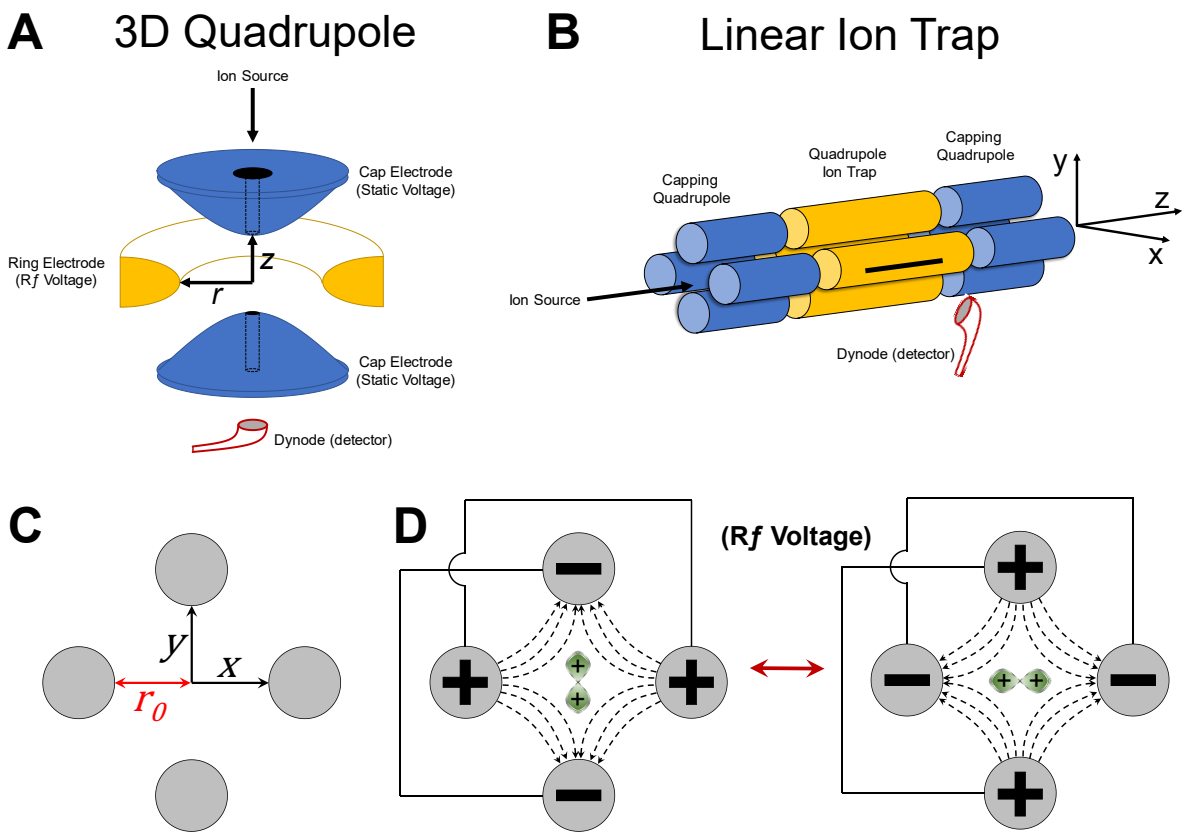


Figure 1.9 Quadrupole ion traps.

Configuration of (**A**) 3D quadrupole ion trap and (**B**) Linear quadrupole ion trap. Capping electrodes (static DC voltage) are depicted in blue, while quadrupole electrodes (RF voltage) are depicted in yellow. (**C**) Cross-section of trapping electrodes in a linear ion trap. (**D**) Depiction of alternating electric field induced by RF voltage, with a trapped ion cloud depicted in green.

The 2D quadrupole potential within the x and y plane is defined, for an individual axis, as:

$$\boldsymbol{\phi_2(x,y) = \frac{(x^2 - y^2)}{r_o^2}}$$

Equation 1.6

Where x and y are the Cartesian co-ordinates within the electric field (with the origin located at the center of the electrodes), and r_o is the radius of the field produced by each electrode.

An alternating applied potential $[\Psi(t)]$ across the electrodes over period t, is described by the function:

$$\boldsymbol{\Psi(t) = \pm(U - V_{RF}\cos\Omega t)}$$

Equation 1.7

where U is DC voltage, V_{RF} is the RF voltage and Ω is the radial frequency of V_{RF}. (Note due to the oscillating nature of the electrodes, when the potential in one plane is positive, the potential in the orthogonal plane is negative, hence the ±).

Applying $[\Psi(t)]$ to the quadrupoles produces a product of quadrupole potential and applied potential over time, described by:

$$\boldsymbol{\Phi_2(x, y, t) = \phi_2(x, y)\Psi(t)}$$

Equation 1.8

Or:

$$\boldsymbol{\phi_2(x, y, t) = \frac{(x^2 - y^2)}{r_o^2}(U - V_{RF}\cos\Omega t)}$$

Equation 1.9

How this affects ion motion is dependent upon the forces this potential generates on charged particles. These forces can be described using the following laws of motion:

$$\boldsymbol{\vec{F} = -ze\vec{E} = -ze\nabla\phi_2(x, y, t)}$$

Equation 1.10

$$\boldsymbol{\vec{F} = m\vec{a} = m\frac{d\vec{v}}{dt} = m\frac{d^2(x \text{ or } y)}{dt^2}}$$

Equation 1.11

The first describes Lorentz's law of motion for a positively charged particle in an electric field, where z indicates an integer number of charges on the ion, e represents the fundamental charge

of an electron, and $\vec{E}$ represents the electric field. As $\vec{E}$ can be described as the rate of change of potential, it is substituted for the divergence (flux over volume) of the potential described earlier $[\nabla\phi_2(x,y,t)]$. The latter equation describes Newton's second law of motion, where m represents the mass of the ion, and $\vec{a}$ is its acceleration. Equating these two laws results in:

$$\boldsymbol{m\frac{d^2(x \text{ or } y)}{dt^2} = -ze\nabla\phi_2(x,y,t)}$$

Equation 1.12

The resulting laws governing ion motion in either the *x* or *y* direction, respectively, are determined by:

$$\boldsymbol{m\frac{d^2x}{dt^2} = -ze\frac{2x}{r_o^2}(U - V_{RF}\cos\Omega t);\ m\frac{d^2y}{dt^2} = ze\frac{2y}{r_o^2}(U - V_{RF}\cos\Omega t)}$$

Equation 1.13

These can be rewritten as general Mathieu equations[100] – which have finite solutions, and can be used to describe ion trajectories – when defining the following variables:

$$\boldsymbol{u = x \text{ or } y}$$

$$\boldsymbol{\xi = \frac{\Omega t}{2}}$$

$$\boldsymbol{a_x = -a_y = \frac{8eU}{mr_o^2\Omega^2}}$$

$$\boldsymbol{q_x = -q_y = \frac{-4eV_{RF}}{mr_o^2\Omega^2}}$$

Equation 1.14 (Mathieu variables)

When substituting these variables into the previously defined equations, the result is:

$$\boldsymbol{\frac{d^2u}{d\xi^2} + (a_u - 2q_u\cos 2\xi)u = 0}$$

Equation 1.15

The variables a_u and q_u are colloquially known as "trapping" parameters and are modulated by the quadrupoles' DC voltage and RF voltage, respectively. Only certain values of a_u and q_u provide ion stability within the trap; other values result in the ion becoming 'unstable' and being ejected. The regions in which ion stabilities are determined as a function (β_u) of the relationship between a_u and q_u, and generally represents ion stability in the oscillating quadrupole electric field (an ion is stable when $0<\beta_u<1$)[101]:

$$\boldsymbol{\beta_u \cong \left[a_u - \left(\frac{(a_u-1)q_u^2}{2(a_u-1)^2 - q_u^2}\right) - \left(\frac{(5a_u+7)q_u^4}{32(a_u-1)^3(a_u-4)}\right) - \left(\frac{(9a_u^2+58a_u+29)q_u^6}{64(a_u-1)^5(a_u-4)(a_u-9)}\right)\right]^{\frac{1}{2}}}$$

Equation 1.16

Ions must be stable in both the x and y dimensions to remain trapped; because of this, the 'trapping region' is where β_x and β_y stability regions intersect; while multiple intersections exist, most mass linear ion traps utilize the first stability region (first intersection of β_x and β_y) due to its low requirement for both DC and RF voltages (**Figure 1.10**).

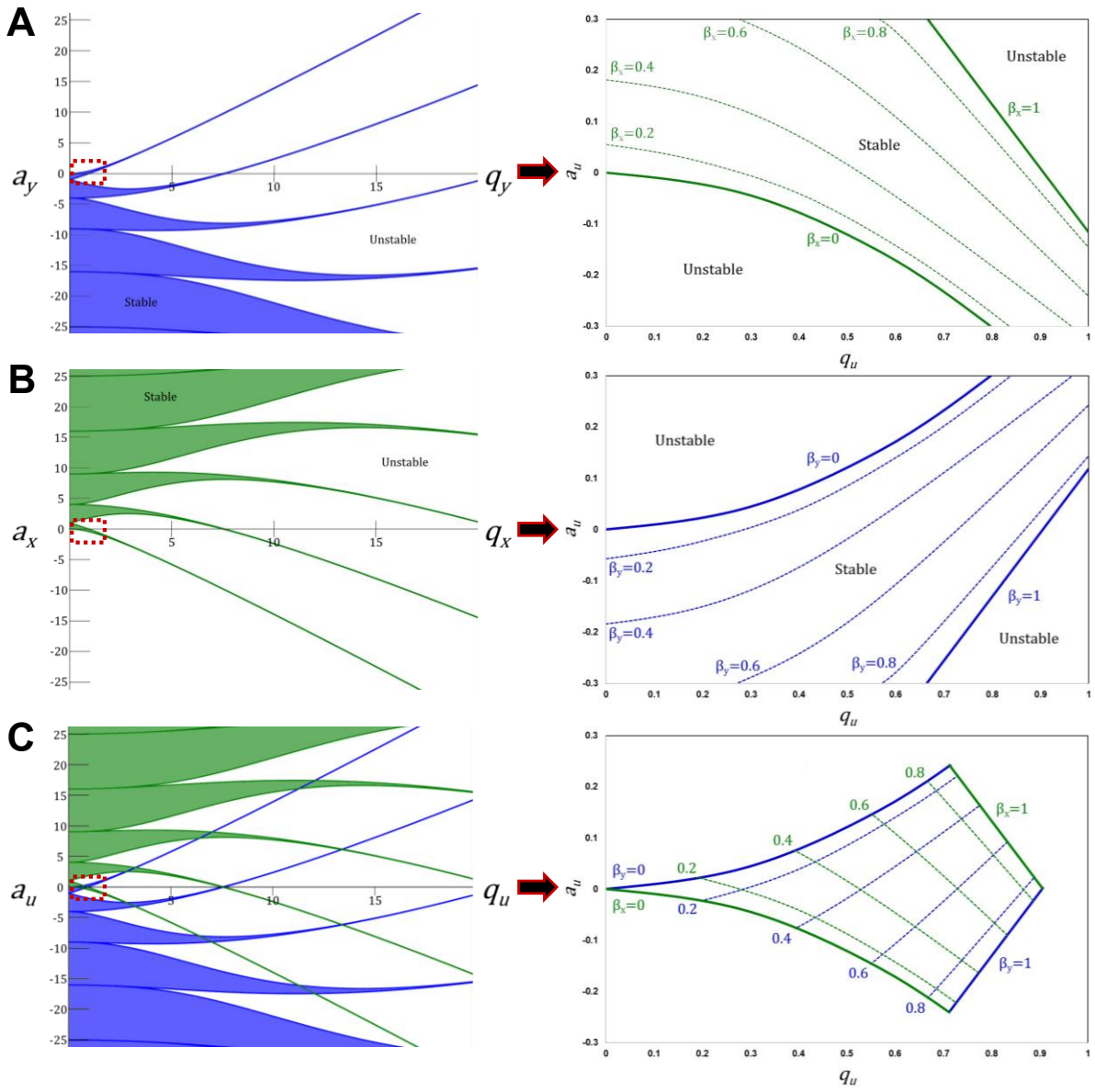


Figure 1.10 Stability regions as defined by Mathieu equations.

Stability regions in the x-plane (**A**), y-plane (**B**), and both planes (**C**) of a linear ion trap; green-shaded regions represent solutions for a_u and q_u where ions are stable in the x-plane and blue-shaded regions where ions are stable in the y-plane. For ions to remain trapped, they must be stable in both x- and y-planes. The lowest-energy intersection where this occurs is $0 < a_u < 1$, which is known as the 'first stability region', represented on the left.

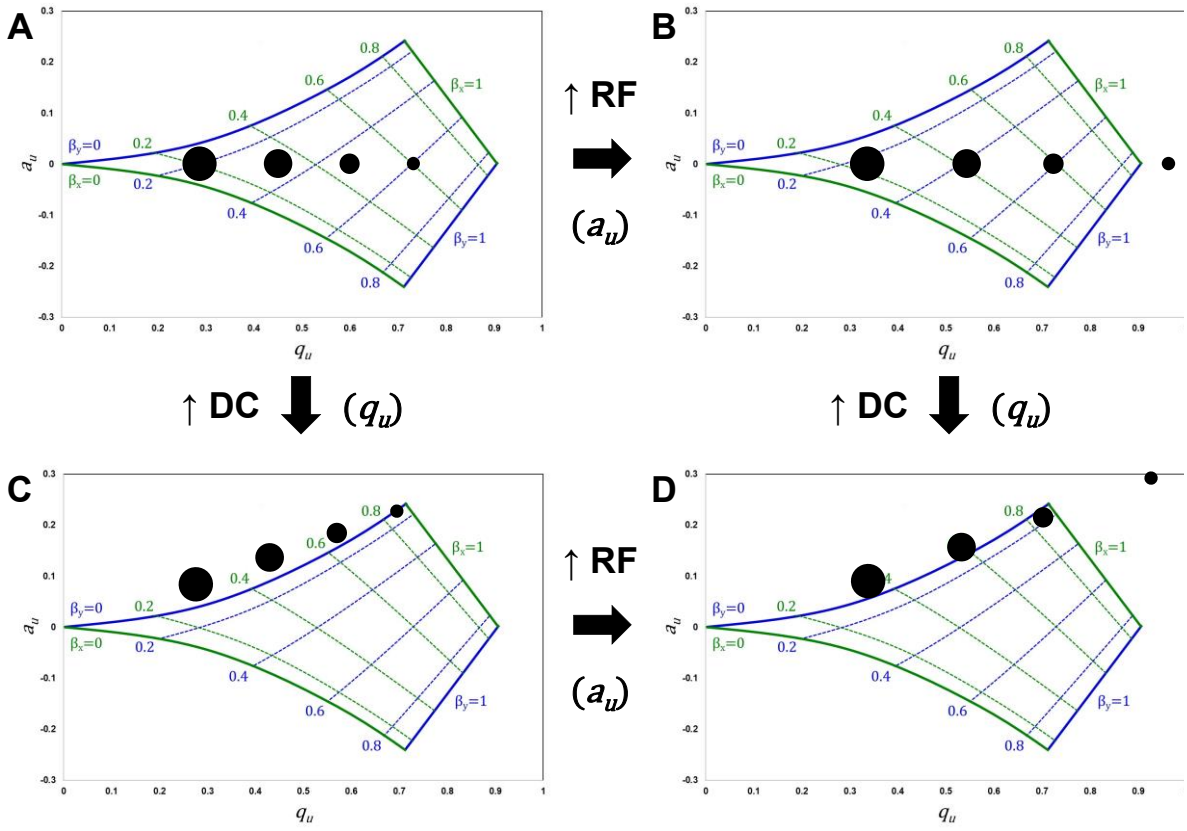


Figure 1.11 Mass selection inside first stability region.

Black dots represent ions of differing m/z ratios, with larger dots corresponding to larger m/z ratios. (**A**) With zero applied DC potential, stable ions orient themselves within the first stability zone of the ion trap. (**B**) By increasing RF voltage, only q_u is modulated, and ions slide through the stability zone in the y-plane, until they are ejected at $q_u = 0.908$. (**C**) Modulation of a_u through ramped DC voltage results in ejection of ions in the x-plane. (**D**) By modulating both a_u and q_u, ions with specific m/z ratios can be selected.

When performing scans of ions present in the ion trap, DC and RF voltages can be modulated, making trapped/stable ions unstable (**Figure 1.11**). This results in ions with specific m/z ratios being ejected from the ion trap, where they contact the mass spectrometer's detector(s).

In linear ion traps the detectors are dynode based. Contact of a dynode by a charged particle results in the release of electrons from the dynode's surface toward a second or secondary point on the dynode, and so on. This cascade effect is read as a current at the terminal end of the series,

and the strength of this current is directly proportional to the ejected ions' intensity within the ion trap.

1.2.4.2. Kingdon Trap ('Orbitrap™') Mass Analyzers

Another form of mass analyzer that has come to prominence in recent years is the Thermo Fisher™ 'Orbitrap™'[75,76,102], a modern variant of a Kingdon trap. Originally described in 1923 by KH Kingdon as a method of generating and studying ion species[103], Kingdon traps consist of four electrodes – a central spindle wire (cathode) surrounded by a cylindrical electrode (anode), and two endcap electrodes at either end of the cylinder to 'imprison' ions generated within the interior space. DC voltage applied between the central wire and outer cylinder electrodes produces a radial logarithmic potential, defined by[104]:

$$\boldsymbol{\Phi = A\ \ln r + B}$$

Equation 1.17

Where A and B are constants at a defined voltage, and r is the radial coordinate from the central wire. However, the capping electrodes on either side of the cylinder (disregarded in ***Equation 1.17***), while serving their purpose to 'imprison' ions within the center of the trap, produced complicated and difficult to map potentials in the distal regions of the trap. Ions generated and stored in Kingdon's original trap had incredibly short lives of ~1.4 milliseconds[103]. In 1981, R.D. Knight improved upon Kingdon's design by modifying the outer electrode to be made of two conical electrodes placed together to have a large central radius with decreased radii at the trap's terminal ends[105]. Applying a DC potential between the outer and inner electrodes produces a harmonic (symmetric) axial potential, described by:

$$\boldsymbol{\Phi = A\left(z^2 - \frac{r^2}{2}\right)}$$

Equation 1.18

Wherein A is a constant, and z and r describe the cylindrical coordinates of the trapping space. This harmonic axial potential effectively confines ions along the center of the axial electrode – in

addition to the logarithmic potential described previously, to give the combined potential known as a 'quadro-logarithmic' potential:

$$\boldsymbol{\Phi = A\left(z^2 - \frac{r^2}{2} + B \ln r\right)}$$

Equation 1.19

This improved the duration of trapped ion species by a factor of ~100. However, neither Kingdon nor Knight's traps were reported to produce mass spectra. In 2000, Alexander Makarov developed what has become known as an 'orbitrap'[102] – a variant of Knight's modified Kingdon trap.

The orbitrap consists of 3 electrodes; two outer, cup-shaped electrodes, electrically insulated from each other that form a 'barrel' around a central spindle electrode. Due to the shape of the orbitrap, applying a DC voltage between the outer and inner electrodes produces the quadro-logarithmic potential described by[75,76,102,106,107]:

$$\boldsymbol{\Phi_{z,r} = \frac{k}{2}\left(z^2 - \frac{r^2}{2}\right) + \frac{k}{2} R_m^2 \ln\left(\frac{r}{R_m}\right) + C}$$

Equation 1.20

And the shape of the electrodes described by:

$$\boldsymbol{z_{1,2}(r) = \sqrt{\left(\frac{r^2}{2}\right) - \left(\frac{\left(R_{1,2}\right)^2}{2}\right) + R_m^2 \ln\left[\frac{R_{1,2}}{r}\right]}}$$

Equation 1.21

where k is the curvature of the electric field (the restorative force towards the center of the trap, analogous to a spring's constant), indexes 1 and 2 represent the spindle electrode and outer electrode respectively, $R_{1,2}$ denote the maximum radius of each respective electrode, R_m is the characteristic radius of the electric field (defined as the radius from R_1 at which stationary ions stop being attracted towards the central axis and start being repelled; $R_m \gtrsim R_2\sqrt{2}$)[107], C is a constant, while z and r are the cylindrical coordinates within the trap.

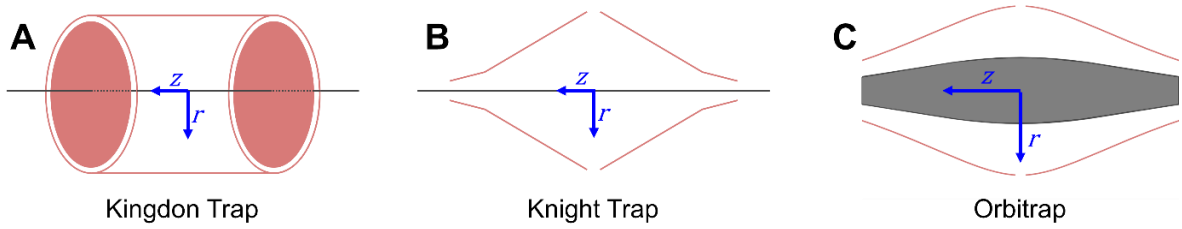


Figure 1.12 Kingdon-style ion traps.

(**A**) The original Kingdon trap consisted of a spindle electrode passing through a cylinder, between which a DC voltage was applied. Cap electrodes (pink disks) were applied to either end of the cylindrical electrode to minimize ion loss. (**B**) Knight trap profile. A derivation of the Kingdon trap, but the outer electrodes (cylinder and caps) have been replaced with two electrically isolated cones to produce a quadro-logarithmic potential. (**C**) Modern orbitrap profile, with a central spindle electrode and two electrically isolated outer electrodes defined by *Equation 1.20* to produce a quadro-logarithmic potential.

Ion motion within the orbitrap depends on both orbital motion around the spindle electrode, and axial oscillations along the z-axis (the ions' secular frequency). Interestingly, the potential arising from the orbitrap's shape ***(Equation 1.19)***, shows that motion in the z-axis is independent of motion around the central spindle. Utilizing ***Equation 1.12*** but for the divergence of potential in the z direction produces the following equation:

$$m\frac{d^2(z)}{dt^2} = -qkz \rightarrow \frac{d^2(z)}{dt^2} = -\left(\frac{q}{m}\right)kz$$

Equation 1.22

with q here representing the molecular charge to avoid confusion (previously denoted z). ***Equation 1.22*** takes the form of a simple harmonic motion equation. As such, the frequency of z-axial oscillations (ω) can be defined as:

$$\omega = \sqrt{\left(\frac{q}{m}\right)k}$$

Equation 1.23

This allows an ion species' m/z to be determined in an orbitrap by monitoring its axial oscillations in the z-direction, assuming all ions of a given species are 'bunched' and oscillating

together. Ion 'bunching' is achieved prior to orbitrap analysis in what is known as a 'C-trap'; the C-trap is a curved quadrupole (in the shape of a C) which forces ions into very tight bunches prior to their injection into the orbitrap. This ensures all ions of a given species have the same starting conditions inside the orbitrap, and thus, oscillate together as a group. Due to an ion species' m/z being determined as a function of time, higher resolution (separation) of ions' masses can be achieved with longer scan-times.

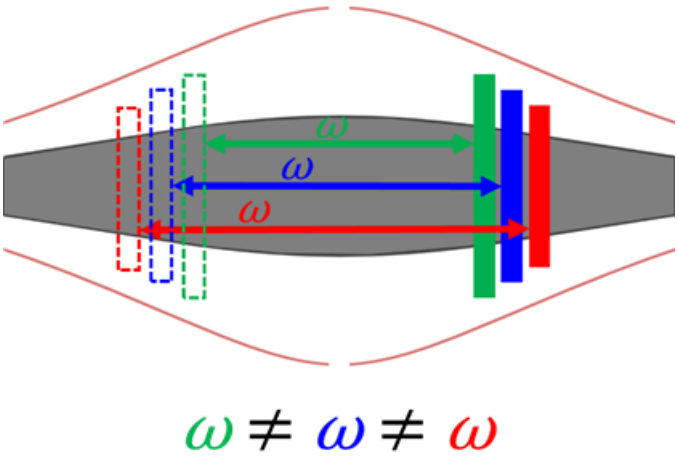


Figure 1.13 Simple harmonic motion of ion packets within orbitrap.

Ions present in the orbitrap oscillate with frequency ω, defined in *Equation 1.18*. As each ion's ω is different, ions separate out over time, allowing both the m/z and number of ions to be determined.

Ions are detected through 'image current detection' (ICD)[75,106]; as a packet of ions approaches an end-plate surface, it causes surface polarization of electrons, which is measured as an induced AC current proportional to the number of ions within the packet:

$$I(t) \approx -qN\omega \frac{\Delta z}{\lambda(r)} \sin(\omega t)$$

Equation 1.24

Where the image current I at time t is dependent on the total charge q of N ions with frequency ω displaced Δz from the center of the trap. $\lambda(r)$ is the 'effective gap' between the outer and inner electrodes, and varies due to the shape of the orbitrap.

ICD monitors the currents induced by all species of ions oscillating within the orbitrap simultaneously; as such, fast Fourier-transforms[75,108] must be made of the total current detected to identify individual species oscillatory frequencies (thus m/z ratios) and intensities.

1.2.4.3. Peptide Ion Excitation and Fragmentation

Once a full scan of all ion species has been made, an individual ion species can be singled out, fragmented, and have its fragments analyzed to determine its composition. This process is colloquially referred to as 'tandem mass spectrometry'[80,86,99,106], MS/MS, or MS^2.

In linear ion traps, this is most-often achieved through collision-induced dissociation (CID). During CID, DC and RF potentials are modulated to make the trap stable for only an individual species of m/z. Subsequently flooding the ion trap with an inert collision gas, most often helium, the isolated ion species is 'excited' via the addition of a supplemental AC potential along one set of axial rods. The frequency of the applied AC potential matches the ion's secular frequency ($\omega_{u,n}$)[99], defined by:

$$\boldsymbol{\omega_{u,n} = \left(n + \frac{\beta_u}{2}\right)\Omega; 0 \leq \beta_u \leq 1; n = 0, \pm 1, \pm 2, \ldots}$$

Equation 1.25

While an infinite number of frequencies exist for a given ion, the most common is the fundamental frequency ($n = 0$). Therefore, for an isolated ion species at a given value of β_u and RF voltage operating at Ω, most ions will oscillate within the ion trap with a frequency of $\frac{\beta_u \Omega}{2}$. A supplemental RF voltage applied to a single axis of rods with frequency $\frac{\beta_u \Omega}{2}$ causes these ions to pick up kinetic energy and collide with the collision gas, inducing fragmentation.

In orbitrap-based mass spectrometers, fragmentation is most-often achieved in a nitrogen-filled 'high-energy collisional dissociation' (HCD) cell[106]. After a full scan performs an inventory of all the ion species entering the mass spectrometer, a filter is applied to the optics at the front-end of the instrument, allowing only a single species of m/z through to the C-trap, from where it is forced with high velocity into the HCD cell, inducing rapid fragmentation. The fragments are then shuttled back into the C-trap where they are directed into the orbitrap for analysis.

When performing an MS/MS series on peptide ions, fragmentation occurs at characteristic sites, usually corresponding to the region with the highest surface area; the peptide backbone[56,57]. As the peptide backbone consists of three types of atomic bonds – N-C_α, C_α-C_o, and C_o-N, these are

the bonds with the highest frequency of dissociation. The resulting nomenclature for the daughter ions (fragments) is based on the site of fragmentation, where the ionic charge is retained following fragmentation, and the orientation of the peptide sequence being read (i.e. N-terminus to C-terminus or vice versa)[109,110].

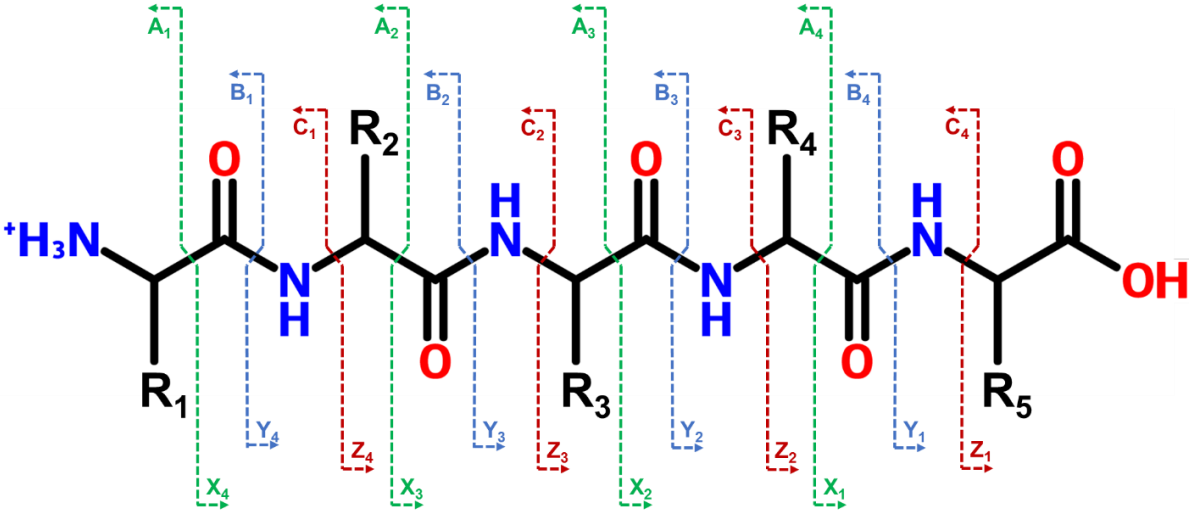


Figure 1.14 Patterns of peptide-backbone fragmentation.

Fragmentation of the peptide backbone produces 3 families of ions; **A/X-series** (**green**) with cleavage of C_α-C_o, **B/Y-series** (**blue**) with cleavage of C_o-N, and **C/Z-series** (**red**) with cleavage of N-C_α. Subscripts are used to denote the number of residues retained on the product ion. Due to the planar structure of the peptide bond, B/Y-series ions are the most prominent.

Fragmentation of a C_α-C_o bond produces A and X daughter ions if the charge is retained on the N-terminal or C-terminal fragment, respectively. Likewise, fragmentation of the C_o-N bond produces B and Y daughter ions, and dissociation of the N-C_α bond produces C and Z daughter ions. Subscripts are used to indicate the number of residues contained on the daughter ion towards its respective terminus of numbering origin. By far the most common type of fragmentation occurs at the peptide bond (C_o-N) due to it's large, rigid, planar structure. This is also the most favourable site of fragmentation, due to the daughter B and Y ions retaining their relative AA compositional mass (± a few atomic mass units); B ion masses are typically equal to the sum of their constituent neutral AA masses less 17 mass units (an OH group, due to cleavage at an amide bond and it only retaining the carbonyl portion), while the masses of the

corresponding Y ions are equal to the sum of their constituent neutral AA masses plus 1 mass unit. This discrepancy in mass on Y-ions results from internal solvation of protons along the peptide backbone during CID; fragmentation results in the formation of an amidogen (HN^{-}-R) group on the Y-ion's N-terminus. Incredibly reactive, this amidogen group is solvated to an amino (H_3N^{+}-R) group. It is important to note that fragmentation of an individual peptide ion typically results in the formation of either a B-ion, or Y-ion; depending on which fragment retains the charge, the other often rapidly decomposes and will not be detected. Fragmentation resulting in the formation of both requires a surplus of protons to be available within the vicinity of the site of breakage, which is a rare occurrence in the gas phase.

Secondary in abundance to B- and Y-series ions are A-series ions, largely due to B-series ions undergoing degradation following CID; loss of a B-ion's C-terminal carbonyl (C=O) group results in the formation of an A-ion. As a result, A-ions are typically more abundant than their counterpart X-ions, which in turn are more abundant than C- and Z-series ions, which are only produced following incredibly high-energy collisions. Due to the relative rates of fragmentation along the peptide backbone, mass spectrometry based *de novo* peptide sequencing primarily utilizes B-, Y-, and A-series ions. For detailed reviews and tutorials see (www.ionsource.com) and refs[111–113].

1.2.5. Peptide Identification Following Fragmentation

1.2.5.1. *De Novo* Peptide Sequencing via LC-MS/MS

De novo peptide sequencing allows the amino acid sequence of a peptide chain to be determined via the parent ion's mass, in conjunction with the observed B-, Y-, and sometimes A-series daughter ions. Unfortunately, peptide fragmentation via LC-MS/MS doesn't occur sequentially as it does with Edman degradation allowing for the callout of a sequence per residue identified; peptide fragments are generated randomly and are measured simultaneously. However, by applying biochemical constraints of protein chemistry, it is possible to deduce the sequence of the parent peptide; for any given spectrum, only certain combinations of residues and/or

modifications can produce the mass for a given peak. Therefore, by collectively analyzing all peaks from a given spectrum, all of which originated from a single parent peptide (unique m/z) and are primarily B- and Y-series ions, it is possible to "stitch" together the peptide's sequence. This is achieved by first identifying a terminal residue of the parent peptide. For peptides generated via proteolytic cleavage such as trypsinization, this process can be much easier, as one can simply search the spectrum for the corresponding R/K ion, which forms the first Y-series ion, and the corresponding penultimate B-ion. For **singly-charged** species, this can be described as[111–113]:

$$\left(\frac{\mathrm{m}}{z}\right)_{PenB} = (\mathrm{M}+\mathrm{H})^{1+}_{Parent} - 18 - [156_{Arg}\ \mathrm{OR}\ 128_{Lys}]$$

Equation 1.26

$$\left(\frac{\mathrm{m}}{z}\right)_{PenY} = (\mathrm{M}+\mathrm{H})^{1+}_{Parent} - \mathrm{AA}_{N-term}$$

Equation 1.27

Following identification of the first B-/Y-series ion, it's corresponding Y-/B-series (respectively) ion can be determined and found using the following formula[111–113]:

$$\left(\frac{\mathrm{m}}{z}\right)_{B-ion} = (\mathrm{M}+\mathrm{H})^{1+}_{Parent} - \left(\frac{\mathrm{m}}{z}\right)_{Y-ion} + 1$$

Equation 1.28

By identifying a B-/Y-series ion, the process is then to simply 'walk' along the spectrum looking for peaks differing from the identified ion's mass by that of an AA residue. An example of this is depicted in **Figure 1.15**, using the peptide DAVIDK with an $(\mathrm{M}+\mathrm{H})^{1+}_{Parent} = 660.3563$.

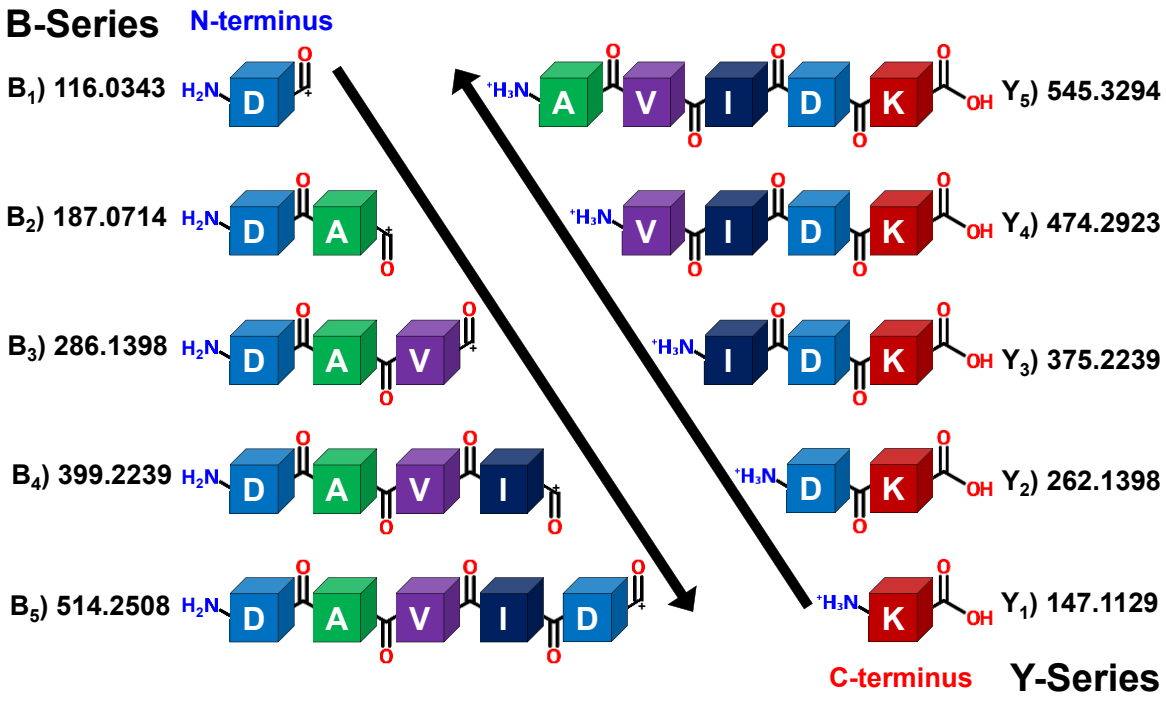


Figure 1.15 *De novo* sequencing schematic for peptide DAVIDK.

The respective theoretical B- and Y-series daughter ions' masses (when singly charged) following fragmentation. B-series ions begin with the first residue at the N-terminus, and extend toward the C-terminus, while Y-series ions begin with the first residue at the C-terminus and extend toward the N-terminus. For both series generated, positive charge is retained at the site of fragmentation.

While *de novo* sequencing is incredibly fast and powerful compared to older techniques such as Edman degradation, it still requires a relatively large amount of starting material that is free from impurities. The process of *de novo* sequencing is also notoriously complicated; a myriad of rules dictates the process for searching spectra to determine and stitch together sequences – a process which becomes compounded when determining large sequences. While modern computing has made the call-out of peptide sequences easier, the process as a whole still remains quite difficult[114]. Because of this, *de novo* sequencing's application in the study of complex, whole-proteome derived peptide mixtures has become somewhat limited in recent years.

Following the completion of the human genome project[7], an incredible amount of information was extracted pertaining to open reading frames (ORFs), and conversely, predicted, putative, and

known protein sequences. Using this information, a multitude of protein sequence databases were – and continue to be – constructed, updated, and curated. One such example, and possibly the most robust protein database resource, is the Universal Protein Resource Knowledgebase – also known as UniProtKB[115] – the product of a consortium started in December 2003[116] between the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). With public-access databases such as UniprotKB available for reference, newer techniques for protein identification such as peptide fingerprinting and spectral matching have reimplemented *de novo* sequencing.

1.2.5.2. Peptide Fingerprinting and Spectral Matching via LC-MS/MS

Peptide fingerprinting[117] and spectral matching[118] utilize the same core principles as *de novo* sequencing, albeit in a much more efficient way. Following isolation and fragmentation of a parent peptide, the daughter ions' masses are measured, just as in *de novo* sequencing. However, instead of requiring careful analysis of the spectrum to stitch together a sequence, these techniques map the observed parent ion mass and fragmentation spectra to a reference database[117–120]; utilizing sequence constraints specified by the user, such as the type of protease used during front-end handling and static or dynamic modifications expected on any specific residues[119], the fragment spectra are scored against theoretical fragmentation spectra derived from the reference database[117–120]. While this might seem to confound the identification process, it improves the positive identification of peptides, as it avoids one of the largest caveats of *de novo* sequencing – misinterpretation of daughter ions with multiple possible compositional permutations. Such instances are referred to as 'conflicting masses', the simplest example being the ions of leucine and isoleucine or diglycine and asparagine; both leucine and isoleucine have identical atomic composition and thus identical m/z values of 113.08406, while diglycine is identical to asparagine. As the number of residues in an observed peptide ion increases, or post-translational modifications are present on specific residues, conflicting masses become more common, which in turn confound the *de novo* sequencing process. Therefore, peptide

fingerprinting and spectral matching are capable of discerning conflicting masses by matching the observed spectra to the only sequences possible to be observed within a specified database.

In a traditional LC-MS/MS shotgun proteomics pipeline, the mitigation of mass spectra misinterpretation occurs at several stages during data analysis. The first step occurs following the matching of an observed peptide fragment to a theoretical peptide fragment originating from the reference database. These assignments are termed 'peptide-spectrum matches'[83], abbreviated as PSMs, with each PSM being assigned a score (S_i) based on the search algorithm's test statistic (and corresponding p-value)[121]:

$$S_i = -10 \log_{10}(p_i)$$

Equation 1.29

However, with proteome databases often exceeding 50,000 protein sequences for an individual eukaryotic organism, many of which have conserved peptide sequences – termed 'degenerate peptides' – peptide-spectrum matching often experiences the challenge of assigning fragment spectra correctly to proteins within their reference database[122–124]. To filter out incorrect PSM assignments and only retain those that are correct, a false-discovery proportion (FDP) is applied during data processing[121]. In statistics, the FDP can be defined as the estimated proportion of a selected number of 'significant' observations – significance being a threshold defined by the user – for a given statistical test which have occurred by chance (i.e. are not significant). In proteomics, for any given pairing of a theoretical peptide and an observed spectrum – of all possible peptide-spectrum pairings – the PSM is extremely likely to be incorrect. To estimate the null-distribution, a 'decoy' database is used[125]; a database of roughly the same size of peptide sequences as the reference, or 'target', database is randomly generated, albeit without any of the sequences observed in the reference database. Peptide spectra observed from the dataset are then matched and scored against this decoy database. Using this method, the distribution of PSMs identified in the decoy database are assumed to be equivalent to incorrect PSMs identified in the target database[121]. Using the decoy PSM scores as a threshold (the applied false-discovery proportion or FDP, colloquially referred to as the false-discovery rate or FDR), incorrect PSMs identified in the

target database are filtered out and minimized. For a more robust review of statistics, please refer to **Section 1.6**.

Unfortunately, having a high-stringency for peptide or PSM assignments does not always translate to the protein level. To further combat the issue of incorrect protein inference, as is often the case when scoring and assigning degenerate peptides, several approaches exist. One of the earliest methods created was the "two-peptide rule"[121,123,126], in which the identity of an inferred protein is only considered 'real' if at least two peptides have been assigned to it. While this decreases sensitivity, as many proteins are likely identified with a single unique peptide, it greatly increases specificity in the proteins being reported.

Another method is the 'minimum set cover' or MSC approach[123]. MSC algorithms utilize the parsimony principle, or Occam's razor, to deduce which proteins are present in a dataset given the presence of certain high-confidence peptides. By using a list of high-confidence peptides from a reference database, the algorithm generates the smallest possible list of proteins to which these peptides can be assigned. Though MSC provides a high degree of specificity for inferred proteins, it is unable to distinguish between proteins co-identified via high-confidence degenerate peptides. As a result, proteins identified exclusively via degenerate peptides are often reported as a family rather than individual proteins.

Lastly, there exists probabilistic inference algorithms (PIAs)[123,127]. While similar to MSC algorithms, PIAs first convert PSM scores to probabilities, which are then used to determine the probability of a protein's presence in the dataset. The most widely utilized of PIAs is PeptideProphet[127,128], which formulates the current framework for which the popular proteomics search engine SEQUEST[118] is formulated. PeptideProphet first determines a peptide's probability by utilizing the highest PSM probability observed for that peptide. Following the assignment of a peptide's probability, the algorithm predicts the theoretical number of sibling peptides (NSPs) from the parent protein, and whether these are observed. Using the highest-probability peptide as a starting point, the protein's probability is gradually approximated based on the presence of NSPs and their respective test statistics. Degenerate peptides which map to several proteins in the dataset are 'weighted' based on the protein probabilities dictated by unique peptides. Through

successive iterations of this process, an expectation probability for individual proteins within the dataset is generated. One caveat with this process, however, lies with proteins identified exclusively via degenerate peptides; such proteins cannot be distinguished from each other probabilistically, and as a result, are treated as a single protein 'grouping'. This can be seen in datasets utilizing SEQUEST, where multiple proteins identified exclusively through their shared degenerate peptides are assigned the same, often, low probabilities and scores.

However, PeptideProphet and its employment in SEQUEST constitutes only a single peptide MS/MS search engine; a myriad of peptide MS/MS search engines have been – and continue to be – developed, each employing a unique statistical approach and attempting to increase peptide identification rates with improved sensitivity and specificity. While more than twenty alternative search engines to SEQUEST exist, several of the most popular (listed in order of release) include: Mascot[129], an engine based upon the MOWSE (molecular weight search) peptide-mass database[130], albeit incorporating probability-based scoring for the correlation of calculated and measured fragment masses; X!Tandem[131], an open-source algorithm which matches and identifies observed peptides/fragments through a multistep process involving gradual refinement of potential candidate sequences; OMSSA (open mass spectrometry search algorithm)[132], an open-source matching algorithm based upon the BLAST framework for sequence identification[133]; MaxQuant's Andromeda[134], an open-source PIA-based search engine developed by Jürgen Cox which builds upon Mascot's probability-based scoring method; PEAKS DB, a search tool incorporating *de novo* sequencing results into database search results to both increase confidence and validate identifications[135]; Comet, an open-source variant of the SEQUEST search engine[136]; and MS-GF+ (mass spectra generating function-plus)[137], a self-described 'universal' mass spectrometry database search tool that utilizes unique scoring parameters dependent on how the spectra were generated[126,138].

1.3. Protein Quantitation & Comparative Proteomics

One of the most alluring features of mass spectrometry in the study of proteomics is its ability to implement quantitation in tandem with protein identification. This ability primarily stems from the mass spectrometers' measurement of the ion current being observed for both unique and total peptide ions (see sections 1.3.4.1 and 1.3.4.2 for an overview of dynodes and induced ion currents). While absolute quantitation is achievable using 'spike-in' standards – often synthetic isotopologues of peptides of interest – of a known concentration, most proteomic quantitation is achieved through relative comparisons.

One of the earliest forms of quantitative proteomics was with 2DE in tandem with mass spectrometric protein identification[139]. By running two 2DE gels identically – one experimental sample and one control – one could see which 'spots' on the gel changed in size and position. Differences in a protein spot's area were used to determine the change in abundance, and the spot's protein was identified via mass spectrometry. As previously mentioned, utilization of 2DE gels gradually lost favor to more high-throughput methods such as LC-MS/MS, but in doing so, scientists had to utilize a different method for comparisons of protein abundance.

1.3.1. Stable Isotope Labelling

One of the most widely implemented techniques for quantitating differential protein expression between proteomic datasets within an experiment is via stable isotopic labels[80,140–142]. Several derivations of this approach exist, with their difference primarily relying on the stage at which the labels are incorporated into the sample, and what is detected for quantification. Popular isotopic labelling techniques include isotope-coded affinity tags (ICAT)[143], isotope-coded protein labels (ICPL)[144], isobaric tags for relative and absolute quantification (iTRAQ)[145], tandem-mass-tags (TMT)[146], N-terminal labelling[147], terminal-amine isotopic labeling of substrates (TAILS)[148], enzymatic labelling techniques[149], and the stable-isotopic labelling of amino acids in cell-culture (SILAC)[150].

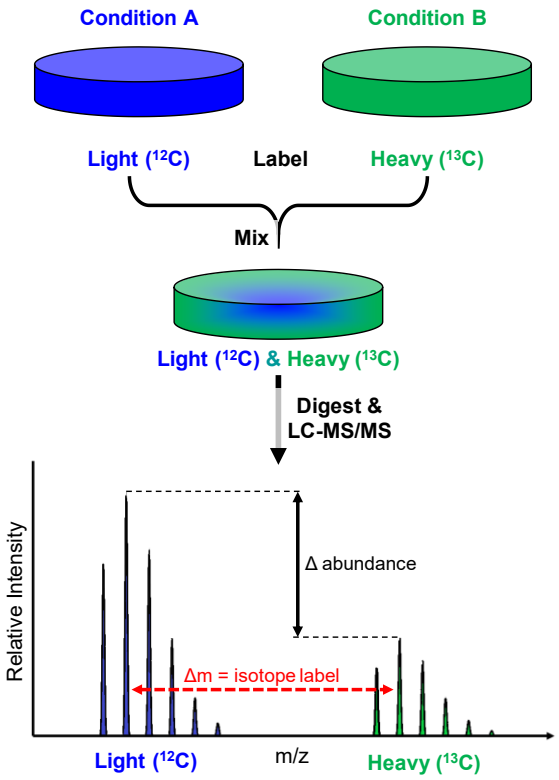


Figure 1.16 Direct comparison and quantitation of isotope-labelled peptides. Identical peptides originating from different experimental conditions can be directly compared following labelling with either **heavy** (*e.g.* ^{13}C) or **light** (*e.g.* ^{12}C) isotopic labels. The difference in ion intensities for peptides originating from different samples can be directly compared, allowing the user to quantitatively determine a protein's change in abundance.

Using these techniques, samples originating from different experimental conditions are labelled with tags of identical chemical structure containing either heavy (^{2}H, ^{13}C, or ^{15}N) or light (^{1}H, ^{12}C, or ^{14}N) isotopes. Following an experiment, samples each containing their respective heavy or light tags are mixed together prior to mass spectrometric analysis. Due to the tags being of identical chemical makeup apart from heavy or light isotopes, it stands to reason that for any given analyte, the chemical reaction affixing the isotopic label would have proceeded in the same way, irrespective of the isotopes contained in the tag[80,140–142]. This reasoning likewise holds true (for the most part) during the ensuing liquid chromatographic separation, with identical analytes from either (heavy/light) sample co-eluting from the column; subsequent MS analysis is performed on the co-eluted species, with both appearing on the same spectra, only separated by the difference in mass of the isotopic label. This allows for a direct comparison of the intensities for a given ion from two experimental conditions, providing a quantitative measure for how a given protein's abundance changes (**Figure 1.16**).

While incredibly powerful for relativistic quantitation, isotopic labelling techniques do possess some caveats[141,142]. Perhaps one of the biggest hindrances to the use of isotopic labels is the increased requirement for sample handling; the chemical addition of chemical groups to

proteomic samples requires further purification and recovery steps, reducing the amount of available starting material for mass spectrometric analysis and increasing the time required to perform experiments. In addition, heavy isotopes, if present in great abundance, can alter the hydrophobicity of an analyte, altering its chromatographic properties[151,152]. If working with chemical tags, incomplete labelling reactions or peptide ion fragmentation during MS/MS can often result in the loss of the chemical label or reporter ions, making data interpretation extremely difficult. Lastly, the reagents required to perform isotopic labelling are incredibly expensive; the sheer cost of isotopic reagents is often enough to prevent their practical use in a laboratory setting.

1.3.2. Label-Free Proteomics

While stable isotope labelling remains the gold-standard for protein quantitation in mass spectrometry-based comparative proteomics, these techniques require a large amount of sample handling, are incredibly time-intensive, and expensive. Attempting to address these caveats which often serve as hurdles to many scientists, in addition to technological advances in both mass spectrometry and liquid chromatography, recently there has been a surge in 'label-free' comparative proteomics techniques[153,154]. Using careful front-end standardization – often including automated sample handling, identical amounts of starting material and protein/peptide concentrations, identical chromatographic gradients, and any of various methods for 'normalization' of proteins' ion intensities[155–157] – label-free proteomics techniques can provide information regarding the compositional protein abundance of a sample or how relative protein abundances change with respect to an experimental stimuli/condition. At the fundamental level, two approaches for comparative label-free proteomics exist; these approaches use either 'spectral counting'[158–160], or peptide-ion chromatogram extraction to determine a protein's absolute or relative abundance in the sample[153,154,161,162].

1.3.2.1. Spectral Counting as a Measurement of Protein Abundance

One of the easiest and earliest methods to infer a protein's quantity from mass spectra is through spectral counting. Spectral counting refers to the process of 'counting' the number of high-confidence mass spectra matched to the protein of interest; because the number of times a peptide ion is fragmented is directly proportional to the abundance of that peptide in the sample being analyzed, spectral counting is – in essence – the utilization of proteins' PSMs as a measure of abundance[158–160]. However, spectral counting is affected by the same difficulties facing PSM-assignment; ion suppression and degenerate peptides can lead to an under- or over-representation of a protein of interest, respectively. To counter these potential pitfalls, PSMs are adjusted based on their proteins' amino acid length; it stands to reason that proteins with longer sequences are theoretically capable of producing more peptides following tryptic digestion, therefore being over-represented by raw PSMs alone. Several methods for adjusting PSMs exist, and all follow the same general principle (Equation 1.29). These include the protein abundance index (PAI)[163], the exponentially modified protein abundance index (emPAI)[164], the normalized spectral abundance factor (NSAF)[165,166], the absolute protein expression (APEX)[167], and normalized spectral index quantitation (SINQ)[168]. Generally,

$$Adj.Measurement_k = \frac{Measurement_k}{\sum_{i=1}^{N} Measurement_i}$$

Equation 1.30

where the adjusted measurement for protein k is equal to the raw measurement of k (typically PSMs) divided by the sum of measurements from all N proteins observed in the dataset.

With so many methods available, there exists a lack of consensus as to which method is best, but APEX, NSAF, and emPAI have gained the most popularity. At the time of writing this, it has been demonstrated that while APEX generally produces the most accurate quantitation profiles, its use requires many training datasets necessitating computing power and time[160,169]. In terms of reproducibility of results, it has been found that NSAF > emPAI > APEX[170].

Because of NSAF's high reproducibility and therefore reliability, in addition to its ease of use with virtually any post-processed proteomic dataset, it has become our laboratory's preferred method of spectral counting. The equation used to determine a protein's NSAF is detailed below[170]:

$$NSAF_k = \frac{(\#\,\mathrm{PSMs}/L)_k}{\sum_{i=1}^{N}(\#\,\mathrm{PSMs}/L)_i}$$

Equation 1.31

where the # PSMs is the number of observed PSMs for a given protein k, L is the length of the protein in terms of amino acid residues, and N is the total number of observed proteins in the dataset. Following determination of a protein's adjusted spectral counts for an individual sample or experimental condition, this process can then be repeated for additional samples to be compared to.

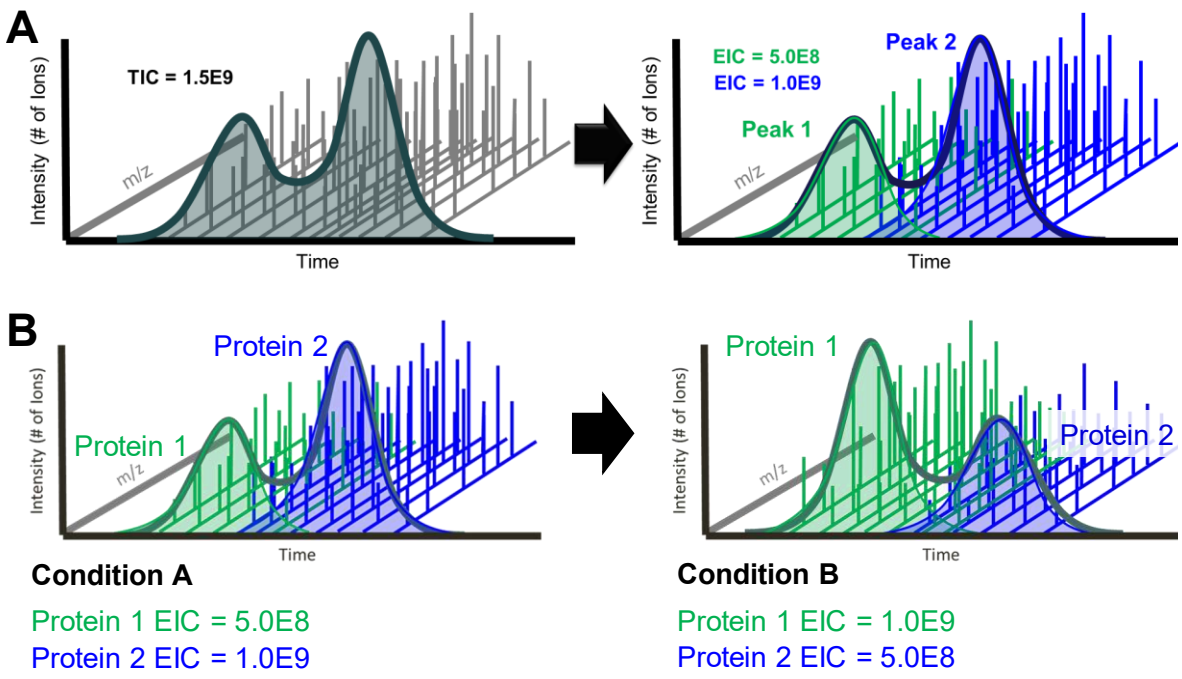
One of the caveats with this method becomes apparent when setting up the mass spectrometer's operating parameters. Often, mass spectrometers are operated using what is known as an 'exclusion list'; following the successful fragmentation of a peptide ion observed within the ion trap, that ion's m/z is ignored (or excluded) with respect to further fragmentation for a user-specified amount of time. This is beneficial when wanting to identify low-abundance ions present in the sample, but can be detrimental when utilizing spectral counting as a form of quantification if the period of exclusion is excessive.

1.3.2.2. Peptide-Ion Chromatogram Extraction as a Measure of Protein Abundance

With respect to relative protein abundance determination achieved via between-sample comparisons, perhaps the most prominent of the emerging techniques utilize peptide-ion chromatogram extraction. This technique is analogous to the 'area under-the-curve' technique used in chromatography[153,154,161,162]. To understand this process more easily, it helps to think of the mass spectrometer in the LC-MS/MS pipeline as the detector for the LC. For the duration of a chromatographic run, as peptide ions elute from the LC into the mass spectrometer, the total

number of ions are catalogued as a 'total ion current' (TIC) (for more information on how ion currents are determined, see **Sections 1.2.4.1** and **1.2.4.2**). Following the successful identification of a peptide, its abundance is determined by integrating its area under-the-curve (for all observed charge states, i.e. +1, +2, +3, …) for the duration of the chromatographic run; this is known as the peptide/proteins' extracted ion chromatogram (EIC). Peptides' EICs corresponding to a specific protein are then combined giving an average representation of that protein's abundance. However, not all peptides ionize with the same efficiency; reagents used during chromatography in conjunction with spray voltages and the amino acid composition of the peptide itself can all contribute to what is known as 'ion suppression'[95]. Ion suppression is the under-representation or lack of detection of an ion's presence in the mass spectra. Therefore, inclusion of such suppressed ions in the determination of a protein's abundance can lead to an under-representation of a protein's abundance. To curtail this issue, the program SEQUEST determines a protein's abundance by averaging its three most intense unique peptides' EICs – this is known as the 'Top 3' Method[171,172]. A popular alternative to this method is the intensity based absolute quantification 'iBAQ' method[172–174], in which the total ion intensity of the identified protein is normalized by dividing it by all possible tryptic peptides that can originate from said protein that have a length between 6 and 30 residues.

Following determination of proteins' EICs from ions eluted from a chromatographic run(s), relative quantification can be achieved by generating a ratio of an individual protein's EIC from two different samples; often, proteins' EICs from biological replicates (i.e. replicates from a singular experimental condition) are averaged prior to the generation of the ratio. Often, this ratio is depicted as a fold-change, generated by taking the $\log_2$ function of the ratio to normalize the difference about zero, as depicted in **Figure 1.17**.



C

$$\text{Protein 1: } \log_2\left(\frac{EIC_A:TIC_A}{EIC_B:TIC_B}\right) = \log_2\left(\frac{5.0E8/1.5E9}{1.0E9/1.5E9}\right) = -1$$

$$\text{Protein 2: } \log_2\left(\frac{EIC_A:TIC_A}{EIC_B:TIC_B}\right) = \log_2\left(\frac{1.0E9/1.5E9}{5.0E8/1.5E9}\right) = +1$$

Figure 1.17 Label-free relative quantitation by ion intensities.

(**A**) The total ion current (TIC), defined as the integrated area under the curve (AUC) for all ion species during a chromatographic run, is composed of different species of unique ions, whose individual integrated AUCs are defined as extracted ion chromatograms (EIC). (**B**) Determination of individual proteins' EIC-to-TIC ratios during an experiment comparing two conditions can allow for the calculation of the relative fold-change of a protein's abundance, (**C**).

However, when averaging protein abundance between biological replicates, it is assumed that the parameters leading up to the replicates' data are identical, when often this is not the case. Two commonly observed phenomena are drift in the chromatographic elution time for a specific peptide of interest or the entire chromatographic run, and a decrease in total analyte intensity. To address the former problem of chromatographic drift, many programs 'normalize' the length of replicates' chromatographic runs to reduce between-sample variability[175,176], but relying on this alone is often insufficient. To address the latter problem of reduced analyte intensities, as

mentioned earlier, many scientists have resorted to carefully standardized front-end procedures. To further normalize biological replicates, our lab has begun calculating protein abundance within a biological sample as a proportion of a reference protein[24] known to be in (relatively) constant abundance, or to that sample's representative TIC[177] (the summation of all proteins' EICs identified in that biological sample). This method minimizes the effect of variations in sample loading between biological replicates, in addition to reducing the impact of proteins assigned an EIC of '0' in only a subset of the replicates.

Perhaps the largest caveat with the utilization of extracted ion chromatograms is the mis-assignment of an m/z to a protein of interest. The chromatogram extraction of peptide ions occurs for the entire chromatographic run, not only when the peptide of interest was fragmented. Because of this, there is the possibility of false assignment of ion species possessing an identical m/z ratio as that of the peptide of interest, which were identified in pre-fragmentation MS scans at various points of the LC gradient. While possible, this phenomenon is incredibly rare due to modern mass spectrometer's high sensitivities and accuracies, and as such, is often ignored.

1.3.2.3. Quantification of Compositional Protein Abundance

The label-free techniques previously mentioned describe relative between-sample protein abundance quantification. However, both spectral counting and extracted ion chromatogram techniques can be utilized to determine the relativistic proportions of various proteins comprising an individual sample. With respect to spectral counting, this process is relatively straightforward[160]; the derivation of the 'adjusted measurement' such as the NSAF is, for all intents and purposes, an individual sample adjusted measurement of abundance relative to all other observed proteins in that sample. When utilizing extracted ion chromatograms, the process is similar; the EIC-to-TIC ratio is the proportion of detectable ion current for a specific protein that comprises the total ion current of the entire sample. However, due to some proteins' EICs being comprised of low-confidence peptides, degenerate peptides assigned to multiple proteins, or falsely-assigned ions, between EIC-to-TIC ratios and normalized spectral counting, spectral counting is historically more utilized, but this trend is beginning to shift.

While EIC-to-TIC ratios and spectral counts can be utilized for determining relative compositional abundance, their use is limited when wanting to determine the absolute abundance of a peptide within a sample. Selected reaction monitoring (SRM), multiple reaction monitoring (MRM), or parallel reaction monitoring (PRM) are techniques that can be utilized to quantitatively determine a peptide (or multiple peptides') abundance within a sample[63,178,179]. SRM, MRM, and PRM techniques utilize mass filters, selecting voltage parameters to stabilize only very precise m/z ratios. Once this mass filter is set, the mass spectrometer counts the number of ions it observes for the specified species of parent ion – in addition to either user-specified m/z daughter ion(s) (SRM/MRM) or the full MS/MS spectrum of the parent ion (PRM) – relative to a control ion (usually the most abundant peptide observed). However, being a targeted approach, to determine which m/z ratio will be monitored, a regular LC-MS/MS run must first be run, and peptides of interest pulled from the post-processed data. This has made SRM, PRM, and MRM incredibly impractical for large numbers of individual proteins, resulting in them being preferred for validation experiments.

1.4. Statistical Approaches to Data Interpretation

Modern mass spectrometry-based proteomics experiments are often capable of identifying several hundreds, if not thousands, of proteins within a single sample. Due the sheer number of observations complete with measurements, in addition to proteomic experimental designs being comparative (i.e. cause and effect), the application of statistical analysis techniques is necessary for coherent conclusions to be drawn from the data. This section will focus primarily on the application of statistics in post-processed data interpretation. For information regarding how proteomics software determines test statistics (and their corresponding p-values) for peptides and proteins from raw mass spectra, see **Section 1.2.5**.

1.4.1. Determination of the Probability Value (p-value)

Most often, proteomics experiments are performed by comparing the measured abundance of individual proteins identified within two or more samples (or groups of samples) making up different experimental conditions. The goal of such comparisons is to identify proteins whose abundances change with respect to the experimental condition. Regrettably, there is no test in existence capable of determining whether or not the measured abundance for a given protein between two or more conditions is truly different (true discovery) or indifferent (false discovery) in a completely binary way[121]. Because of this, scientists rely on probabilistic approaches which describe the likelihood that the difference observed, if any, for a given protein is significant.

When making comparisons on populations with numerical values, one can choose between two families of statistical testing – parametric or nonparametric. Parametric tests are defined as tests which make assumptions about the underlying parameters (such as mean and variance) for a population's distribution, while nonparametric tests do not. The most widely utilized nonparametric tests are those of the rank and order statistical families, where, rather than directly utilizing a population's parameters for statistical inference, these parameters first undergo a transformation into rank/order. This allows inferences to be made regarding different populations, even when population distributions are non-normal, but at the cost of statistical

power. Frequently used statistical tests include those from the parametric *t*-test (two populations)[180] and analysis of variance (ANOVA; two or more populations)[181] families, and their corresponding nonparametric tests, the Mann-Whitney *U*-test (MWU)[182] and Kruskal-Wallis *H*-test (KWH), respectively[183].

Application of a given statistical test will return a test-statistic (e.g., *t*-statistic from a *t*-test, *U*-statistic from a MWU-test), which is subsequently used to calculate a probability known as a p-value. P-values are often misinterpreted as the probability that the observation in question is false. In fact, the p-value for a given comparison should be interpreted as the proportion of false-discoveries that are less similar to the 'norm of false-discoveries' than the given comparison[121]. For example, when comparing protein X's abundance between two conditions, a p-value of 0.05 indicates that, relative to the difference observed for protein X, 5% of all comparisons that are false-discoveries are more different than the observation for protein X. Because of this, when dealing with large numbers of comparisons as proteomics experiments often do, large proportions of 'significant' observations are incorrect. This is known as the multiple comparisons problem (MCP)[121,184], and while still an active area of research, several methods to circumvent it have been proposed.

There is a common misconception – primarily with those new to the field of proteomics – that bigger protein lists represent better, or more successful, scientific experiments. However, this is not always the case; with mass spectrometers becoming increasingly sensitive, bigger lists are often burdened with quantified proteins that are most likely contaminants[24,184,185]. While certainly impressive in terms of coverage, experiments identifying and comparing larger lists of proteins are thusly more prone to the MCP. As a first-line defense, it is generally agreed upon to do two things; manually curate post-processed data to remove known contaminants prior to application of a statistical comparison test[186], and to apply a highly discriminatory statistical test during comparisons[185]. While manual removal of contaminants is straight-forward, the selection of an appropriate statistical test can often be troublesome.

1.4.1.1. Comparison of Two Populations Using *t*-Tests

For binary comparisons of samples or groups, assuming normal population distribution, the most appropriate and preferred statistical tests belong to those of the *t*-test (tests following Student's *t*-distribution)[180]; in proteomics specifically, these tests are incredibly useful when comparing two experimental conditions due to the inability to analyze large numbers of biological samples in a practical way. However, there are three variants of this test, each suited to the comparison of different types of populations, and therefore with differing levels of discrimination as to what is 'significant'; paired (Student's *t*-test)[187], unpaired-homoscedastic, or unpaired-heteroscedastic (Welch's *t*-test)[188]. Paired *t*-tests assume both populations are equal, fit a normal distribution, and both distributions have equal variance. In assuming such equal measures between two sample populations, paired *t*-tests are best suited to cause-and-effect studies where a population is measured prior to, and after, application of the experimental condition. However, the assumptions made with paired *t*-tests also dismiss many variables often at play in health science experiments, which results in a large amount of leniency with respect to the test's outcome. Conversely, unpaired *t*-tests dismiss several of the assumptions paired t-tests make, the most notable being that the populations being compared are equal. While all unpaired *t*-tests assume unequal populations, those of the homoscedastic category still assume that the variance exhibited by the measurements of each population are equal. Heteroscedastic tests, on the other hand, assume unequal populations in addition to unequal variance. Because unpaired *t*-tests treat each experimental condition as a unique population, the permitted 'randomness' of measurements for each population is much higher. This results in much larger differences being necessitated to produce statistically significant results.

Even with our current breadth of knowledge pertaining to the complexities of biological systems, it is not possible to determine with absolute certainty whether any two biological samples – even if replicates of one another – are equal in terms of variance. As such, unless one has exhibited the utmost control over their experimental conditions, the best choice of statistical test is that which has the strictest parameters pertaining to what is deemed 'significant' – Welch's *t*-test[185,188,189].

1.4.2. Statistical Means to Increase Confidence for Significant Observations

When dealing with comparative proteomic studies, the removal of contaminants from proteomic datasets, in addition to choosing strict and appropriate statistical tests are excellent procedures to curate a reliable and high-confidence list of proteins whose abundances change significantly. A technique borrowed from transcriptomics[190,191] and routinely applied in smaller studies with noisy data, these procedures are combined with a user-defined 'threshold' for what determines a significant 'fold-change' in abundance[192]; often this is more than sufficient for the generation of a high-confidence list of proteins that change in abundance and can easily be verified with molecular techniques. However, with extremely large datasets often reporting hundreds of significant ($p<0.05$) changes, these techniques become insufficient, and validation via molecular tests of such large lists becomes impractical.

Because of the nature of the MCP, several methods, each aiming to control a different aspect of statistical testing outcomes, have been proposed[121]. To understand these techniques and what they are trying to control, the possible outcomes of a statistical test must first be defined.

When making a statistical comparison, there are two assumed possible outcomes; the *null hypothesis* (H_0), which assumes for a given comparison, there is no difference between the two groups, and the *alternative hypothesis* (H_A), which rejects H_0. Based on these two outcomes, there becomes the possibility of four, as outlined in **Table 1.2**.

Table 1.2 Summary of statistical test outcomes and error types.

		Null Hypothesis (H_0)	
		True	**False**
Decision about Null Hypothesis (H_0)	**Reject**	Type I Error (False-Positive; α)	True Negative
	Accept	True Positive	Type II Error (False-Negative; β)

1.4.2.1. The Bonferonni Correction

When performing statistical tests for a series – or 'family' – of comparisons, the primary concern is with the occurrence of type I errors (false-positives; α; 'critical' p-value); the probability of occurrence for such an error in a series of comparisons is known as the 'family-wise error rate' (FWER). One of the most popular techniques, the Bonferonni correction[193,194], addresses the MCP by controlling the FWER through adjustment of reported p-values. This is achieved by taking the critical, or acceptable type-I error rate, p-value (i.e. $\alpha = 0.05$) and dividing it by the total number of hypotheses being tested (m). This procedure generates a new, smaller critical value of α/m, such that the null hypothesis is rejected for a comparison in the series if $p \leq \alpha/m$.

1.4.2.2. The False Discovery Rate

An alternative method to the Bonferonni correction is the false-discovery rate (FDR) approach. Rather than adjusting the critical p-value, as in the Bonferonni method, FDRs attempt to predict, at specific thresholds of p-values, the proportion of the encompassed p-values that are likely part of the null-distribution (π_0). The null-distribution can be defined as[121]:

$$\pi_0 = \frac{\#\text{ false discoveries}}{\#\text{ putative discoveries}}$$

Equation 1.32

Therefore, the null distribution, as described in **Section 1.4.1.**, is the total proportion of false, or insignificant, discoveries within all possible putative (false and true) discoveries for a dataset. Intuitively, one might expect the p-values for insignificant/false discoveries to cluster in a way that is *opposite* the way significant/true discoveries cluster; the p-values for true discoveries are often focused in a small region near 0 in the [0,1] interval of probability. However, because of what a p-value represents (the proportion of π_0 that is less similar than the comparison in question) the idea that false discovery p-values may cluster near 1 in the [0,1] interval is untrue. Instead, the distribution for π_0 is uniform along the [0,1] interval, as depicted in **Figure 1.18B**. Therefore, for a theoretical series of statistical comparisons containing [π_0] false discoveries, and [1- π_0] true discoveries, the distribution should look like the histogram depicted in **Figure 1.18C**.

While the absolute determination of π_o is not possible without a near limitless dataset, it can be approximated.

The approximated value for π_o can then be utilized for the determination of the false-discovery proportion (FDP):

$$\mathrm{FDP} = \frac{(\text{false discoveries} \cap \text{selected discoveries})}{(\text{selected discoveries})} \approx \frac{\alpha \times m \times \pi_0}{k}$$

Equation 1.33

where α is the user-defined p-value threshold, m is the total number of putative discoveries in the dataset, and k is the integer number of selected discoveries (i.e. the number of discoveries corresponding to α).

Unfortunately, however, p-value histograms from multiple comparisons of biological samples are almost never as 'clean' as the histograms depicted in **Figure 1.18C**. When utilizing real data, it is incredibly difficult to distinguish the behaviour of true discoveries and false discoveries, and because of this, makes the determination of α and π_0 nearly impossible. As a result, the FDP is often estimated and referred to as the false discovery rate (FDR).

As the FDR is an estimate of the FDP, it should possess two qualities; FDRs should be *conservative* (i.e. they should not under-estimate the number of false-discoveries in the dataset), and they should be *asymptotically convergent* (i.e. with increasing data quantity and quality, repeated calculation of the FDR should converge on the real value of the FDP).

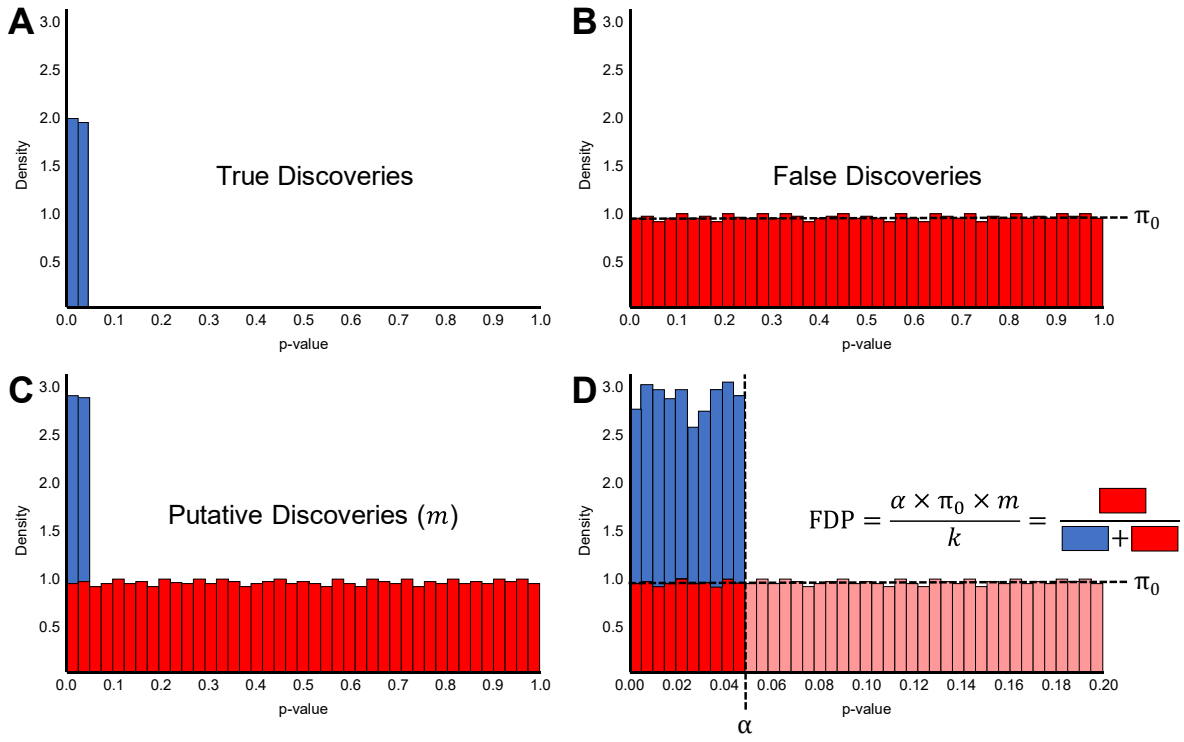


Figure 1.18 Distribution of p-values.

Ideal distribution of p-values for (**A**) true discoveries, (**B**) false discoveries, and (**C**) all (putative) discoveries. (**D**) Using the average 'density' of false discoveries as depicted by the red bars in B and C, the FDP can be calculated for a critical p-value threshold. Adapted from [121]

The determination of FDRs is a step-wise procedure. First, p-values corresponding to comparisons are ordered from lowest to highest. Once sorted, p-values are assigned an integer rank, i, such that $i \in [1, m]$. For a given p-value k in the list, the FDP is approximated by setting α to p_k and using the following equation:

$$\widehat{\mathrm{FDP}}_k = \frac{p_k \times m \times \pi_0}{k}$$

Equation 1.34

An intermediate table is then utilized, and $\widehat{\mathrm{FDP}}_k$ is stored in the kth cell. This process is repeated until all m FDPs have been approximated. Next, for the k best p-values (i.e. 1,2,3,..., k), the FDR is calculated as the smallest $\widehat{\mathrm{FDP}}$ in the intermediate list that is found in a cell greater than, or equal to, the kth cell. Mathematically, this is represented as:

$$\mathrm{FDR}(\{\mathrm{protein}_{(1)} \rightarrow \mathrm{protein}_{(k)}\}) = \min_{i \geq k}(\widehat{\mathrm{FDP}}(i))$$

Equation 1.35

Finally, one simply walks through the list of FDRs and stops when they reach the limit for what is an acceptable rate of false-discovery; the null hypothesis is rejected for all proteins with a calculated FDR up to this point.

This procedure was originally described by Benjamini and Hochberg (BH) in 1995[195], however the FDR was a user-defined critical value denoted as q^*, and π_0 was equal to 1. Additionally, the mathematical definition of the FDR was dependent on the probability of the null hypothesis being rejected. As a result, the BH-procedure is slightly simpler; p-values are ordered and ranked as described above, but with the null hypothesis being rejected for all p-values up to the maximum k, where:

$$p_k \leq \frac{k}{m} q^*$$

Equation 1.36

Likewise, in 2001 Storey and Tibshirani (ST) mathematically redefined the FDR as the 'positive' FDR (pFDR) by removing the term describing the probability of the null hypothesis being rejected from the equation[196]. Because the probability of the null hypothesis being rejected is incredibly high when comparing biological samples, BH's FDR and ST's pFDR are remarkably similar.

1.4.2.3. Adjusted p-values and q-values

When calculating FDRs, both BH's and ST's procedures allow for the generation of *adjusted p-values*[197–199] and *q-values*[200–202], respectively. The generation of these *adjusted p-/q-values* originates from the substitution of ***Equation 1.34*** into ***Equation 1.35***, giving:

$$\mathrm{FDR}(\{\mathrm{protein}_{(1)} \rightarrow \mathrm{protein}_{(k)}\}) = \min_{i \geq k}\left(\frac{m \times \pi_0}{i} \times p_i\right)$$

Equation 1.37

Due to the similarity between FDRs and pFDRs when dealing with biological datasets, these two terms are often colloquially used interchangeably. However, in theory, the logical arguments that make up the differences between the two – specifically with respect to the utilized value of π_0 – q-values are ultimately less biased than their counterpart adjusted p-values. It is also important to note that FDRs, along with their adjusted p-/q-values, are properties of the raw p-value set. Raw p-values retain their statistical meaning if their order in the list is changed, whereas this is not the case for FDRs and adjusted p-/q-values; any shuffling or removal of the raw p-values makes their respective FDRs unmeaningful.

1.4.2.4. Local-False Discovery Rate

FDRs, adjusted p-values, and q-values are incredibly powerful and useful tools, allowing the user to have some level of quality control when considering the test statistics for all comparisons made in a dataset. However, the fact that these metrics are properties pertaining to the entire set is also a slight hindrance; these metrics are unable to provide control with respect to individual proteins within the dataset. Due to the 'blending' nature of true and false discoveries as previously mentioned, it is not possible to provide any sort of quality control for individual proteins. Yet, with a large enough dataset (thousands of comparisons), it is possible to define subsets of the entire dataset that share similar properties. By calculating the metrics described in **Sections 1.6.2.2.** and **1.6.2.3.** for each subset of the data, one can produce local-FDRs[121,203].

When subsets of large datasets are defined by their p-value rank, the calculation of local-FDRs utilizes only p-values belonging to specific intervals and their neighbours. By performing calculations in this way, a 'sliding scale' emerges, rather than clear-cut groupings, thus allowing for incredibly refined estimations of the proportion of false discoveries for specific p-value intervals. However, this technique is reserved for incredibly large datasets, where p-value histograms have a smooth profile. For small datasets with irregular histogram profiles, the 'sliding scale' calculation of the local-FDR becomes unstable which can result in poor estimates of the behaviour of the data.

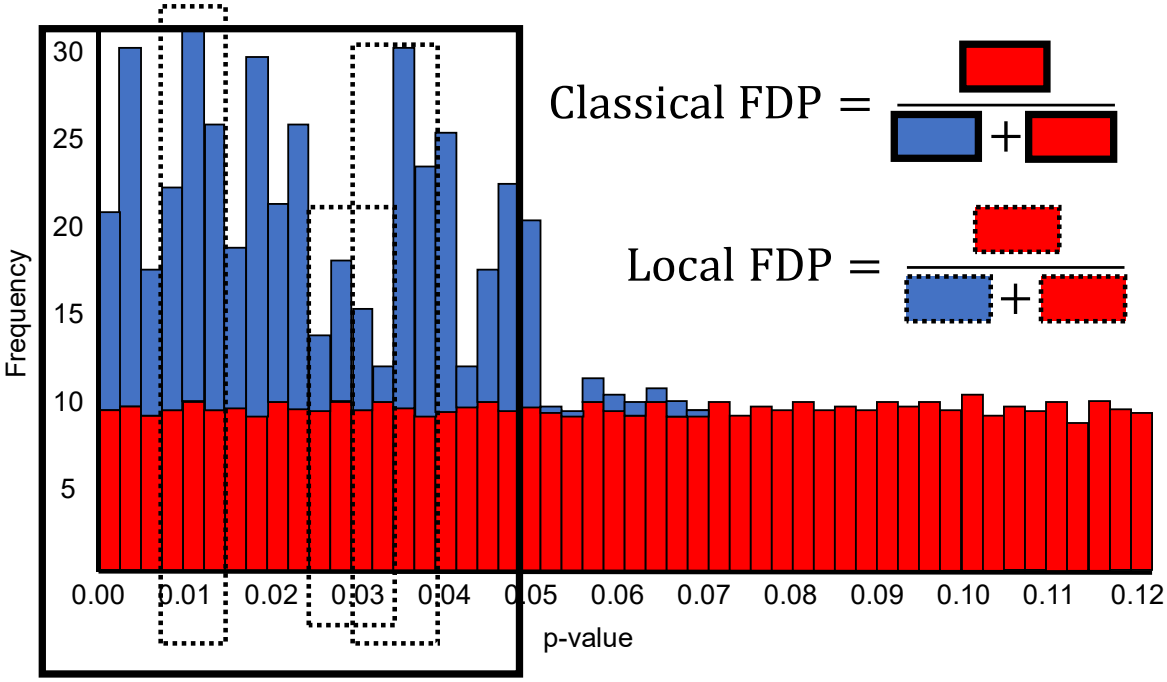


Figure 1.19 Calculation of the local-FDP (FDR).

Classically, the FDP refers to the false-discovery proportion for all values in the range from [o,α]. With a large enough dataset, 'local-FDPs' can be approximated for small regions of the distribution (dotted borders). By performing these calculations for successive regions overlapping each other, incredibly precise FDPs can be determined for very small ranges of p-values. (Adapted from [121])

1.5. Post-Processing and Data Interpretation

Until now, all previous sections have discussed how proteomics data is generated, compared, and curated for statistical significance. Often in comparative proteomics experiments, following statistical analysis, the reported lists of statistically significant changes between experimental conditions can consist of several hundred proteins, if not upwards of a thousand. One of the largest caveats with the reporting of such massive lists is to determine the biological meaning of the results; where does one even begin to interpret what is happening at the cellular/tissue level with so much reportedly changing?

1.5.1. Functional and Locational Protein Annotation

Thankfully, a plethora of online databases exist to help with the question posed above, curating information with respect to proteins' genetic families, molecular functions, cellular locations, and biological pathways/processes. The two most prominent of these databases are Gene Ontology (GO)[204,205], and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[206,207]. GO and KEGG often work to complement each other; GO identifiers are primarily descriptors used to identify proteins' molecular functions (MF), the biological processes (BP) it is involved with, and the cellular component (CC) it can be found in. KEGG identifiers are relational, used to annotate signalling pathways, diseases, and interacting partners. Because of the dynamic nature of proteins, most of which possess multiple functions and locations in the cell, there exists an incredible amount of redundancy; proteins are often given multiple GO and KEGG identifiers to signify this diversity. However, even with the convenience of databases annotating proteins' location, function, and biological significance, the manual retrieval and cross-reference of this information for every protein identified to change significantly within a dataset would be incredibly arduous.

1.5.2. Inference of Biological Meaning

A plethora of web utilities and programs have been designed to make the curation of large lists of 'omics' data more manageable. In recent years, the most prominent of these utilities and programs for 'omics' studies are the web-based DAVID[208] (the database for annotation, visualization, and integrated discovery), PANTHER[209–211] (protein analysis through evolutionary relationships), and STRING[212] (search tool for the retrieval of interacting genes/proteins), in addition to the desktop-based application Cytoscape[213], designed for visual representation of biological networks, but also allowing for the utilization of plugins for functional enrichments such as ClueGO[214]. These utilities provide graphical, tabulated, and/or pictorial representations of which GO and KEGG identifiers are 'enriched' within the user-provided dataset, allowing the user to infer biological meaning to the experimental outcome in a quick and efficient manner.

The enrichment process begins with interpretation of the protein input list. Ontological and pathway identifiers for each protein present in the list are compiled, and the total number of each identifier observed is recorded. From this, a ratio of Σ(# unique identifier) to Σ(# proteins) is generated, giving the proportional representation of each identifier in the input list. This proportion is then compared to that identifier's frequency in the species' genome. Interpretation of this comparison – either through the implementation of Jaccard's similarity coefficient, a binomial test, a hypergeometric probability function (i.e. Fisher's exact test), the chi-squared test, or some combination of these (as implemented in gene set enrichment analysis)[215–217] – a 'fold-enrichment' value can be determined for individual identifiers within the dataset. Confidence estimates for enriched identifiers are often displayed as p-values; following the p-value determination for all represented GO/KEGG identifiers in a dataset, the procedure described in **Section 1.4.2.** can be utilized to determine the FDR corresponding to specific identifications within the identifier list.

Additional to identifier enrichments within these datasets, some web utilities and applications allow for the visualization of protein-protein interactions and functional ontology associations. Through visualization of such networks found within the input data, protein pathways up-/down-regulated with respect to an experimental condition can be mapped, and individual

proteins/processes correlating to the increase/decrease in said pathways identified, providing insight to components critical to the biological response associated with the experimental condition. Furthermore, the strength of evidence/confidence for a given protein-protein interaction or functional ontological association can be depicted directly on these interactome and functional association networks. The most common implementation of this usually takes the form of the thickness of the edge (line) connecting two nodes (proteins/ontologies); the thicker the line, the more evidence exists in the scientific community to support the validity of the interaction. Additionally, when visualizing ontological associations, the size of the node is indicative of its statistical significance, with larger nodes corresponding to increasing confidence. Example protein-protein interaction and functional ontology networks generated using proteins annotated to be involved in the human tricarboxylic acid (TCA) cycle are illustrated in **Figure 1.20A** and **B**, respectively.

While immensely powerful, these tools do possess a critical caveat; utilization of such tools biases data interpretation for what is already known and present in the database. In other words, we are limited to what we already know. Ontological identifiers and utilities showing pathway enrichment and protein-protein interactions are vital for practical analyses of large proteomic datasets, but the pathways and protein-protein interactions identified using only these methods can never be novel; such analyses may instill knowledge on a more macro-level, for example which pathways may be affected in cancer cells by condition X, yet they are only capable of identifying what has already been annotated. Nevertheless, enrichment and interaction analyses in large datasets have become a pillar in the formulation of novel hypotheses which can be experimentally tested.

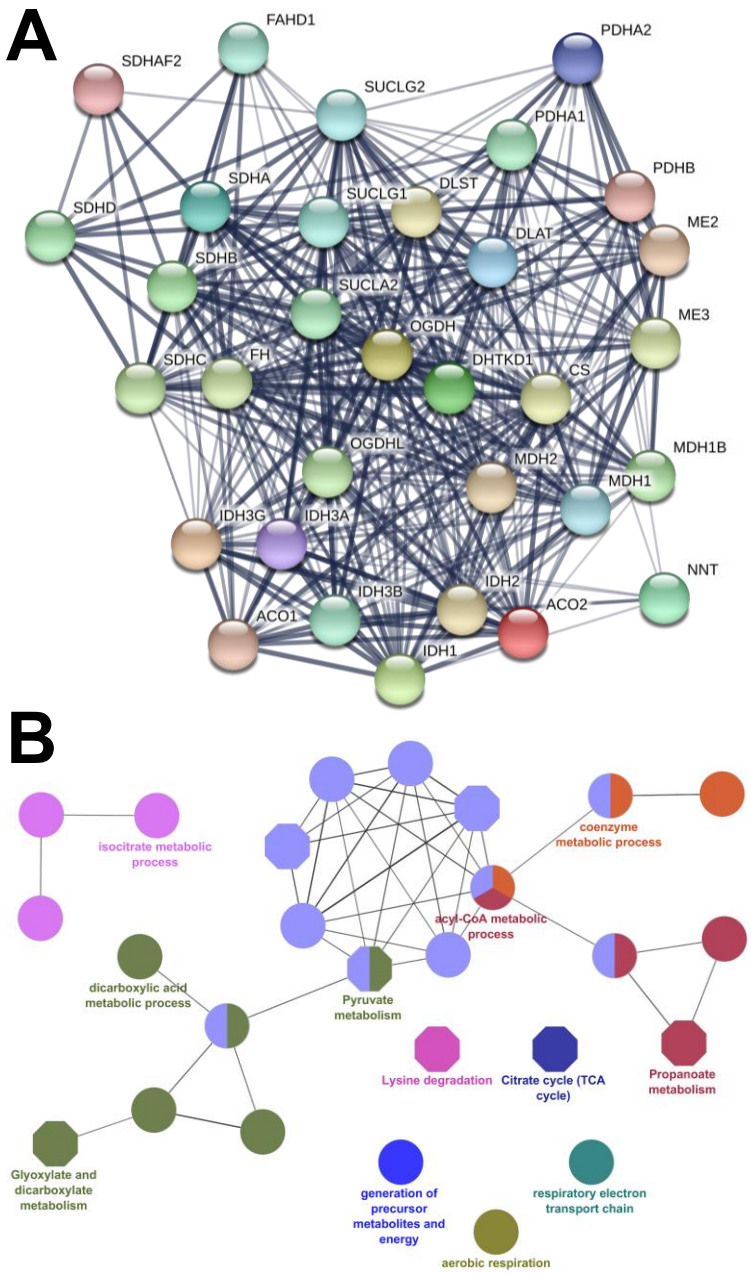


Figure 1.20 Example visualization of protein-protein interaction and functional ontology networks.

Example protein-protein interaction (**A**; STRING v.10.5) and functional ontology networks (**B**; ClueGO v.2.5.1 plugin for Cytoscape v.3.6.1) for proteins annotated to be involved in the human tricarboxylic acid (TCA) cycle. (**A**) Lines indicate known interactions; thicker lines indicate stronger evidence of interacting pairs, while thinner lines indicate interactions with less evidence. (**B**) Functional ontology network generated using GO-biological process (circles) and KEGG pathway (octagons) terms; linkages indicate known interaction between biological functions, while the size of the nodes indicates statistical significance.

1.6. Thesis Objectives

The objectives of this thesis are oriented towards the demonstration of a relatively novel, conservative approach with multiple applications in the field of mass spectrometry-based proteomics. As previously mentioned, the technique to be demonstrated in this thesis does not solely rely on the comparisons of raw mass spectrometric data. In addition to imposing strict front-end sample handling, we implement methods of data normalization *per sample* prior to averaging and comparison to ensure the most accurate, reliable, and reproducible results.

While incredibly useful when applied to the study of cellular and organellar proteomes in response to a specific stimulus or experimental condition, we believe this technique has unprecedented potential in the study of whole organs and tissues – in particular, cancer. By studying how a tissue's proteome changes in response to a stimulus, it allows for the development of a more thorough understanding of disease pathogenesis, progression, and maintenance. In addition, how cancerous tissue responds to anti-cancer therapies such as chemotherapeutics can be studied; pathways activated following chemotherapeutic administration causing cancerous tissue to die or survive can be mapped, allowing for the development of combination treatments that are complementary in their mechanisms of action. Likewise, analysis of patient-derived tumour samples, in conjunction with clinical outcomes, can allow for the identification of prognostic indicators[218].

The following chapters will demonstrate this technique's reliable application in: the small-scale study of organellar proteomes such as lipid droplets (**Chapter 2**[24]); the analysis of whole-tumours from animal models, pre- and post-chemotherapeutic administration (**Chapter 3**[177]); and finally, how this technique can be applied to clinical tumour samples of human origin in the search for novel biomarkers indicative of patients' disease prognosis (**Chapter 4**).

Chapter 2 : Proteomic Analysis of Murine Hepatic Lipid Droplets Following Dietary Stress

2.0. Proem

During fasting, the liver increases lipid storage as a mean to reserve and provide energy for vital cellular functions. After re-feeding, hepatocytes rapidly decrease the amount of triacylglycerol that is stored in lipid droplets (LDs), visible as the size of hepatic LDs significantly decreases after re-feeding. Little is known about the changes in the liver LD proteome that occur during the fasting/re-feeding transition. This study aimed to investigate the hepatic LD proteome in fasted and re-fed conditions in mice using a comparative label-free LC-MS/MS analysis, allowing us to achieve relative quantitation between experimental conditions.

A version of this chapter has been published as:

Kramer, D. A., Quiroga, A. D., Lian, J., Fahlman, R. P., & Lehner, R. (2018). Fasting and refeeding induces changes in the mouse hepatic lipid droplet proteome. *Journal of Proteomics, 181*, 213–224. http://doi.org/10.1016/J.JPROT.2018.04.024

Supplementary data to this chapter can be found online with the published version of this chapter, or at the following link:

Supplemental Tables

2.0.1. Acknowledgements

I would like to thank Drs. Ariel Quiroga, Jihong Lian, and Richard Lehner for their expertise in performing protein purification, and for performing all microscopy, RT-qPCR, and immunoblot validation experiments. Additionally, their open-door policy with respect to communication proved invaluable in the completion of this project.

2.1. Introduction

Lipid Droplets (LDs) are cytoplasmic organelles ubiquitous to all cells and species (for reviews see Refs.[219,220]. LDs have been proposed to play important roles in many cellular processes including lipid metabolism, signal transduction, protein storage and lipid trafficking[221]. Hepatic LDs contain mainly triacylglycerol (TG), with some cholesteryl ester and retinyl ester as a neutral lipid core and are central to abnormal lipid accumulation during hepatic steatosis[222]. This core is surrounded by a monolayer of amphipathic lipids (phospholipids and free cholesterol) and LD-associated proteins of the PAT (Perilipin-1, ADRP/Perilipin-2, TIP47/Perilipin-3) family (for reviews see [219–221,223]). In addition, proteomic studies have shown that a variety of proteins, other than PAT proteins, interact with LDs (either embedded or associated), thus enabling the multiple functions of this organelle[224–226]. Moreover, based on the characteristics of the identified LD proteins, LDs are now known to interact with various other cellular compartments including the endoplasmic reticulum (ER), mitochondria, peroxisomes, endosomes, and the cytoskeleton (reviewed in [227,228]), suggesting LDs possess highly dynamic functions within the cell.

During fasting, the liver enters a state of physiological steatosis, increasing lipid storage in LDs as a mean to reserve and provide energy for vital cellular functions. The source of fatty acids for hepatic TG synthesis are non-esterified fatty acids derived from hydrolysis of TG stored in the adipose tissue, dietary fatty acids from intestinal chylomicron remnants, and fatty acids newly synthesized through *de novo* lipogenesis. Because fatty acids exert deleterious effects on cellular functions when in their free form (biological detergents at neutral pH), excess fatty acids are esterified into TG. In the liver, TG can be either stored in LDs or secreted in TG-rich apoB-containing lipoproteins into the bloodstream. TG can also be hydrolyzed and fatty acids directed toward mitochondrial β-oxidation. As determined empirically by others[229], after re-feeding, hepatocytes rapidly decrease the number and size of LDs; however, little is known about the physiology of this process and the changes in the proteome during the fasting/re-feeding transition that allow for this process to occur.

This study aimed to investigate the hepatic LD proteome in fasted and re-fed conditions in the mouse using gel-LC-MS/MS analysis. Our findings reveal unexpected changes in the LD proteome in fasting versus re-fed livers.

2.2. Experimental Procedures

2.2.1. Animals and Feeding Conditions

We utilized 4-month-old male C57BL/6 mice, with each individual LD sample prep comprised of the livers of 3 animals. Mice were fed *ad libitum* a chow diet (LabDiet, PICO Laboratory Rodent Diet 20, 23.9% protein, 5% fat, 48.7% carbohydrates). Mice were randomly split into two groups: fasted (24 h fast) and re-fed (24 h fast followed by 6 h re-feeding). All animal procedures were approved by the University of Alberta's Animal Care and Use Committee and were in accordance with guidelines of the Canadian Council on Animal Care. Mice, housed three to five per cage, were exposed to a 12 h light/dark cycle beginning with light at 8:00 a.m.

2.2.2. Lipid Droplet Fractionation

At the end of each feeding period mice were sacrificed by cardiac puncture, livers were harvested, rinsed in ice-cold PBS and immediately subjected to homogenization with a motor-driven potter in a hypotonic lysis medium (HLM, 20 mM Tris·Cl, pH 7.4, 1 mM EDTA). All solutions included protease inhibitors (EDTA-free Complete protease inhibitors, Roche Diagnostics) and phosphatases inhibitors (PhosSTOP, Roche Diagnostics). LDs were isolated according to Brasaemle and Wolins,[230], with some modifications[231]. Homogenates were spun at 500x *g* for 10 min. Supernatants were then spun at 15,000x *g* for 10 min to remove mitochondria and to allow fat cake separation. Fat cakes were transferred into new tubes and washed twice at same speed and length of centrifugation. Fat cakes were then diluted 1/3 in 60% sucrose in order to obtain 20% density adjusted suspensions. These were layered at the bottom of ultracentrifuge tubes and overlaid with double the volume HLM-5% sucrose, followed by careful overlay with same volume of HLM. Samples were centrifuged at 28,000 x *g* for 30 min and fat cakes were carefully recovered and analyzed.

2.2.3. Solubilization of Lipid Droplet–Associated Proteins for Western blot and LC-MS/MS Analysis.

To solubilize LD-associated proteins for subsequent Western blot and LC-MS/MS procedures, fresh LD fractions prepared from fasted (n=3) and refed (n=3) mice were mixed with 10% sodium dodecyl sulfate (SDS) (1:1, v/v) and incubated for 1 h at 37°C in a sonicating water bath with constant agitation. Then, samples were microcentrifuged 10 min at maximum speed, at room temperature and the infranatants containing the solubilized proteins were collected from beneath the floating lipid layer. Equivalent volumes of 2× SDS sample buffer were added to the samples, which were then boiled for 10 min prior to loading equivalent amounts of total protein onto a discontinuous SDS-PAGE gel.

2.2.4. Sample Preparation and Mass Spectrometry

SDS-PAGE gels were visualized with R-250 coomassie blue protein stain (SigmaAldrich). Once visualized, protein bands were excised in segments, as outlined in **Supplemental Figure 2.8**. Each gel section was individually treated to in-gel tryptic digestion as previously described[232].

2.2.4.1. LC-MS/MS of Lipid Droplet-Associated Proteins

Fractions containing tryptic peptides dissolved in aqueous 5% v/v ACN and 0.2% v/v formic acid were resolved and ionized by using nanoflow HPLC (Easy-nLC II, Thermo Scientific) coupled to a LTQ XL-Orbitrap hybrid mass spectrometer (Thermo Scientific). Nanoflow chromatography and electrospray ionization were accomplished with a PicoFrit fused silica capillary column (ProteoPepII, C18) with 100μm inner diameter (300Å, 5μm, New Objective). Peptide mixtures were resolved at 500 nL/min using 60 min linear ACN gradients from 0 to 45% v/v aqueous ACN in 0.2% v/v formic acid. The mass spectrometer was operated in data-dependent acquisition mode, recording high-accuracy and high-resolution survey Orbitrap spectra using external mass calibration, with a resolution of 60 000 and m/z range of 400–2000. The ten most intense multiply charged ions were sequentially fragmented by using collision induced dissociation, and

spectra of their fragments were recorded in the linear ion trap; after two fragmentations, all precursors selected for dissociation were dynamically excluded for 60 seconds. Raw data was processed using Proteome Discoverer 1.4.1.14 (Thermo Scientific) and a reviewed, non-redundant *Mus musculus* complete proteome FASTA index (UniprotKB – retrieved October 2015) protein database was searched using SEQUEST (Thermo Scientific). Search parameters included a precursor mass tolerance of 10ppm and a fragment mass tolerance of 0.8Da. Peptides were searched with static modifications as we have previously described[177]. The 'Precursor Ion Area Detector' node was implemented in the data processing workflow to determine relative extracted ion chromatograms (EICs) for each protein identified. Processed data was then filtered using a minimum of 2 medium-confidence (FDR<0.05) peptides per protein, the data from this analysis is listed in **Supplemental Table 2.1**. False discovery rate thresholds for peptide confidence were as follows; Fasted-1: Strict FDR=0.0097; Fasted-2: Strict FDR=0.0097; Fasted-3: Strict FDR=0.0097; Re-Fed-1: Strict FDR=0.0097; Re-Fed-2: Strict FDR=0.0097; Re-Fed-3: Strict FDR=0.0092. Protein lists from were then exported and compared using Microsoft Excel.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[233,234] partner repository with the dataset identifier PXD005977.

2.2.4.2. Mass Spectrometry Data Analysis and Network Analysis

To simplify data analysis, notable contaminants (keratins) were removed from the data set. Likewise, proteins with an observed EIC in ≤1 sample across all n=6 samples were removed. The remaining data was corrected by normalizing identified proteins' EICs to the observed perilipin-2 EIC in a sample-specific manner; being a coat-protein, the surface-area to volume ratio of perilipin-2 to LD volume should remain constant. The corrected protein abundances (EIC-*corr*) for each protein were averaged among samples within their respective experimental condition (fasted versus re-fed). To determine relative changes in average protein abundance between experimental conditions, $\log_2(\text{EIC-}corr_{\text{fasted}}/\text{EIC-}corr_{\text{refed}})$ ratios were generated. To determine the significance of the changes observed, a two-tailed heteroscedastic *t*-test[188] was applied to each protein's EIC-*corr* array between experimental conditions; once grouped, proteins unique and

significantly up-regulated ($p<0.05$ and/or ≥6-fold-change in abundance) in each condition were searched using the STRING[212] 10.0 web-utility and enriched for KEGG pathways and the Gene Ontologies (GO): Molecular Function, Biological Processes, and Cellular Components. Output lists were exported to Microsoft Excel and only ontologies and pathways with an FDR≤0.05 were utilized. Additionally, GO analyses were performed using the PANTHER v13.1[209–211] database and Cytoscape v3.5.1[213] with the ClueGO v.2.5.1 plugin[214].

While the utilization of label-free proteomic quantification is most effective when comparing the abundance of individual proteins between samples, as described above, the comparison of proteins' absolute abundances within a sample can provide valuable information with respect to total protein composition. Bias exists when comparing peptides from different proteins within a sample due to the different ionization efficiencies of the various tryptic peptides, in conjunction with a large dynamic range of protein abundance; while this hinders absolute quantitative comparisons within a sample, it has been observed that label-free proteomic quantification methods do correlate with global protein abundance[172]. As such, in addition to our comparative analysis using relative ion intensities, we performed compositional analysis of LD protein abundance across all datasets using normalized peptide spectral matches (PSMs), relative to Plin2.

2.2.5. Western Blotting and Immunostaining of Membranes

Proteins were resolved by SDS-PAGE (10%) based either on the same triacylglycerol or protein concentrations and were transferred onto PVDF membranes. Specific primary antibodies were incubated with the membranes overnight after blocking the membrane with 5% w/v skimmed milk for 1 hour at room temperature. The following primary antibodies, with their working dilutions from the stock solutions obtained from the supplier in 3% w/v BSA in TBST, were used: acyl-CoA synthetase/ligase 1 (Acsl1) (1:1000, Cell Signaling, #4047), perilipin 2 (Plin2) (1:1000, Abcam #108323), perilipin 5 (Plin5) (1:2000, Progen #GP31), calnexin (Cnx) (1:1000, Enzo Life Sciences #SPA-865), β-actin (1:1000, Cell Signaling, #4967), carboxylesterase 1d (Ces1d, also

called Ces3 or TGH) that also reacts with carboxylesterase 1g (Ces1g, also called Ces1 or Es-x) (1:30000, generated in-house[235]), major urinary protein 1 (MUP1*) (1:300, Santa Cruz Biotechnology #SC-66976), glyercaldehyde 3-phosphate dehydrogenase (GAPDH) (1:5000, Abcam #ab8245), and phosphatidylethanolamine-*N*-methyltransferase (Pemt) (1:1000, generous gift from Dr. Dennis Vance[236]). The following secondary antibodies, diluted 1:5000 in 5% w/v skimmed milk in TBST, were incubated for 1 h at room temperature: HRP-labelled donkey anti-guinea pig IgG (Fitzgerald #43R-ID039hrp) and HRP-labelled goat anti-rabbit IgG (Invitrogen #31460). Immunoreactive proteins were detected by enhanced chemi-luminescence (GE Healthcare, UK) using HRP-labelled secondary antibodies.

2.2.6. Histological Analysis

Livers collected from mice from both fasted and re-fed conditions were embedded in OCT and frozen for further histological analysis. Frozen liver sections were stained with 2 μg/mL BODIPY 493/503 (Invitrogen, USA) in PBS for 1h at room temperature to visualize LDs. Images were collected with a laser scanning confocal microscope (Leica TCS SP5, software version Leica LAS AF 2.6.0, Leica, Germany). Quantification of LD number and size was done with ImageJ software (NIH, USA).

2.2.7. RNA Isolation and Real-Time qPCR Analysis

Liver total RNA was isolated using Trizol reagent (Invitrogen, USA). First-strand cDNA was synthesized from 2μg total RNA using Superscript III reverse transcriptase (Invitrogen) primed by Oligo $(dT)_{12-18}$ (Invitrogen) and random primers (Invitrogen). Real-time qPCR was performed with Power SYBR® Green PCR Master Mix kit (Life Technologies, UK) using the StepOnePlus-Real time PCR System (Applied Biosystems, Canada). Real-time qPCR primers, whose sequences are outlined below, were synthesized by Integrated DNA Technologies (USA). Data were analyzed with the StepOne software (Applied Biosystems). A standard curve was used to calculate mRNA abundance relative to that of a control gene, cyclophilin.

Primer sequences: Cyclophilin (F: 5'-TCCAAAGACAGCAGAAAACTTTCG-3', R: 5'-TCTTCTTGCTGGTCTTGCCATTCC-3'), Plin2 (F: 5'-CACTCCACTGTCCACCTGATT-3', R: 5'-TCCTGAGCACCCTGAATTTT-3'), Plin3 (F: 5'-GGAGGAACCTGTTGTGCAG-3', R: 5'-ACCATCCCATACGTGGAACT-3'), Plin5 (F: 5'-TGTGTGTAGTGTGACTACCTGTGC-3', R: 5'-GGCAAGATCATTCACTGTGG-3'), and ACSL1 (F: 5'-CCACCATCTTCCCTGTGG-3', R: 5'-GGAAGTGTTTGCTTGTCCAAA-3').

2.3. Results

2.3.1. Liver Morphology During Fasted and Re-Fed States

The liver accumulates TG after 24 h fast due to the high fatty acid flux from the adipose tissue leading to significant increase in individual LD size and total area (**Figure 2.1A-C**), reflecting what has been reported previously[229]. Consequential of re-feeding the animals for 6 h after a 24 h fast, the average area of an individual LD decreased by 66% while the total LDs area decreased by 27% (**Figure 2.1A-C**). Concurrently, the number of LDs increased by 126% after re-feeding (**Figure 2.1C**), possibly due to the increased nascent LDs generated from induced *de novo* lipogenesis. Accordingly, hepatic TG levels were 1.5-fold lower after re-feeding the animals for 6 h compared with fasted mice (**Figure 2.1D**).

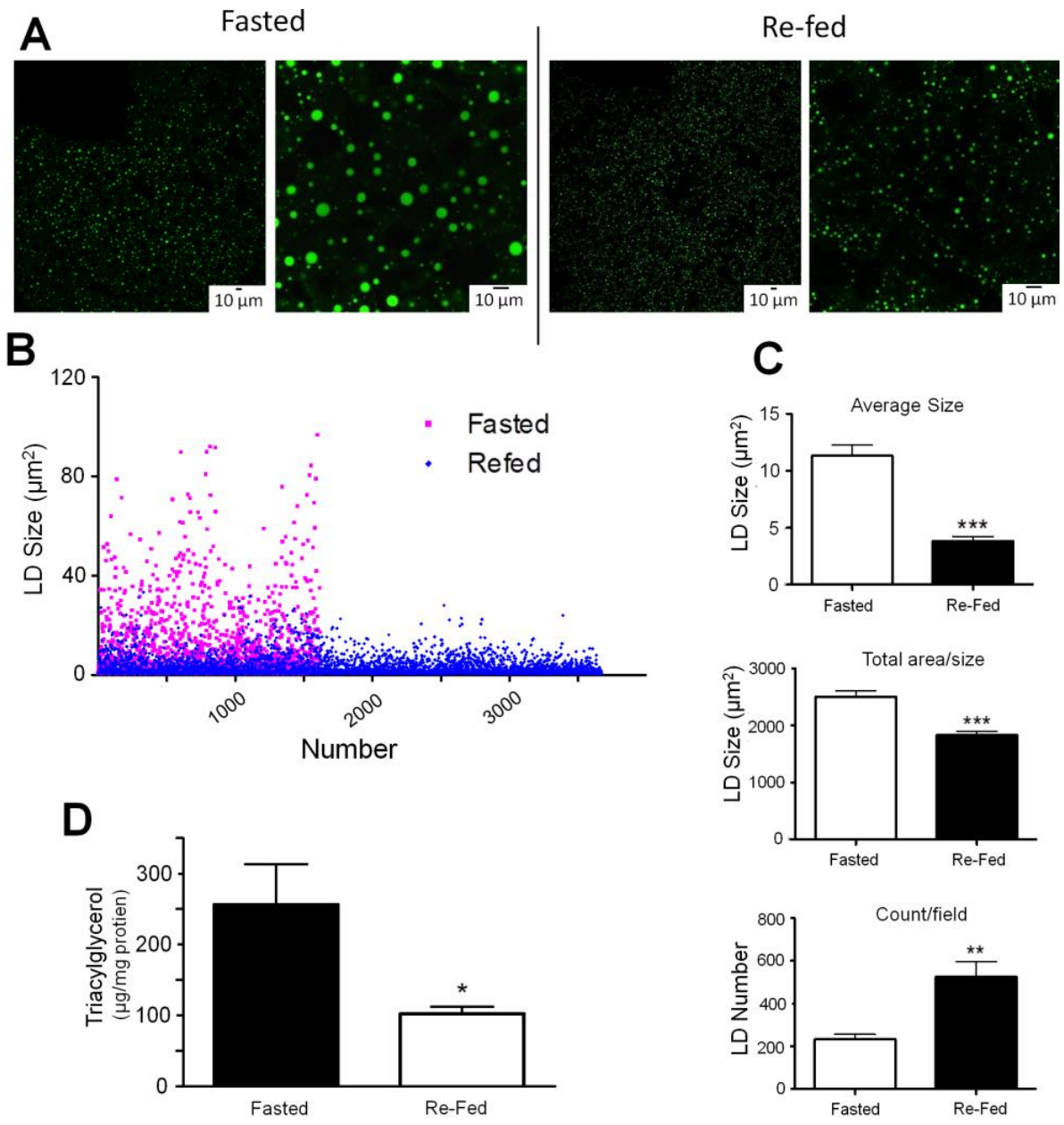


Figure 2.1 Liver LD morphology during fasting and re-feeding.

(**A**) Representative images of hepatic LDs. Bar = 10 µm. (**B**) Distribution of hepatic LDs from fasted and re-fed mice. Seven images were taken from four sections of each condition (100x objective at a zoom-factor of 2), sizes and numbers of LDs were analyzed and pooled. Each data point represents an individual LD. (**C**) Average area of an individual LD, total area of LDs per image field, and number of LDs per image field. (**D**) Liver triacylglycerol concentration in fasted and re-fed states. *P<0.05, **P<0.01, ***P<0.001.

2.3.2. Preparation and Purity of LDs

Livers from fasted and re-fed animals were manually dissected, and LDs were released from the tissue and subsequently purified by sucrose density centrifugation. LD-associated proteins were delipidated, resolved by SDS-PAGE, where gel sections were individually treated to in gel tryptic digestion and analyzed by LC-MS/MS as outlined in **Supplemental Figure 2.8**. The purity of LDs was evaluated by immunoblot analysis. As shown in **Figure 2.2**, LD-associated protein perilipin-2 (Plin2) was found in LD fractions from all samples. Calnexin (Cnx), a resident ER protein, was also present in all the studied samples. One particular interest is the partition pattern of ER carboxylesterases. Carboxylesterases Ces1d [also called triacylglycerol hydrolase, previously annotated as Ces3] and Ces1g [also called esterase-x, previously annotated as Ces1] are related carboxylesterases present in the lumen of the ER. Interestingly, Western blot analysis revealed Ces1d (lower band) visibly partitioned to LDs, while Ces1g (upper band) was absent (**Figure 2.2**). To additionally verify the identity of the lower band being Ces1d, LDs isolated from Ces1d knockout mice[237] were included for comparison (**Figure 2.2**). An ER resident polytopic membrane protein phosphatidylethanolamine *N*-methyltransferase (Pemt) was essentially absent from isolated LDs (**Figure 2.2**).

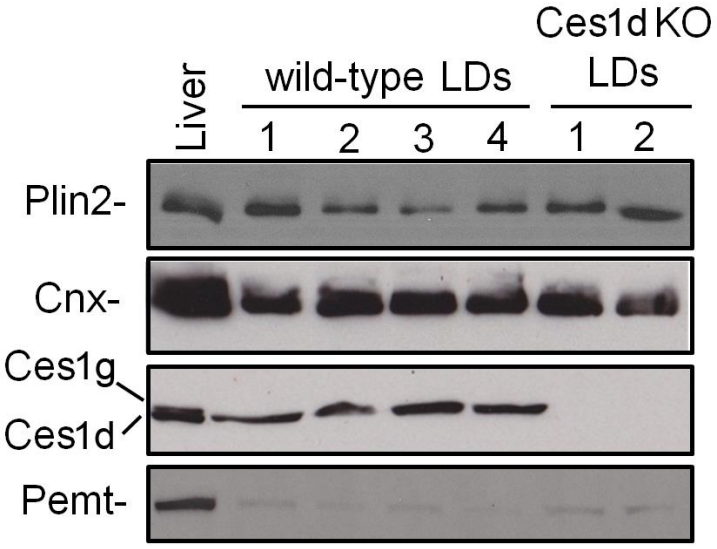


Figure 2.2 Analysis of protein markers in purified LDs.

Purified LDs from livers of four wild type mice or 2 Ces1d knockout mice were analyzed by Western Blotting for the indicated proteins. Liver homogenate from wild type mice (Liver) was used as a control.

2.3.3. LD-Associated Proteins in the Liver

The lipid droplets purified from the livers of nine fasted mice and nine ref-fed mice were then analyzed by gel-LC-MS/MS and the data was refined as described in the Materials and Methods (**Section 2.2**). From this analysis (**Supplemental Table 2.2**), 810 proteins were identified in the LDs isolated from the fasting mice and 784 were identified in the LDs from the re-fed mice (**Figure 2.3A**). Of these, 777 were common to both datasets, 7 were unique to the re-fed LDs, and 33 unique to the fasted LDs. In comparison, previous investigations of the liver LD proteome have identified a number of LD-associated proteins ranging from 1520[238] to only 134[239].

To further investigate the quality of the data set, all the proteins observed were queried by GO analysis using the PANTHER database[209–211]. For this analysis, an enrichment is determined by comparing the ontological frequency of identified proteins to that of the entire mouse proteome. GO analysis of all the proteins observed in the fasted and fed states revealed an enrichment for proteins that have been annotated to various organelles such as the ER, mitochondria, nuclear envelope, and peroxisomes, while being depleted for nuclear and plasma membrane proteins (**Figure 2.3B**). While several proteins observed may reflect molecular contaminants resulting from purification procedure limitations, many of the enrichments observed in our data reflect the current perception of the dynamic nature of LDs; there is an increasing understanding of LD association to the various organelles within a cell[228]. Further GO analysis for biological functions of the identified proteins using Cytoscape v3.5.1[213] with the ClueGO plugin[214] reflects the diversity observed with respect to the cellular localization of the annotated proteins. As summarized in **Supplemental Figure 2.9**, in addition to the major expected networks of metabolism and macromolecular complex assembly, significant numbers of proteins have been annotated for a variety of RNA metabolic processes such as heterocycle and aromatic biosynthesis, nucleoside phosphate metabolism and gene expression.

Quantitative analysis of the EIC intensities of the proteins identified in both the fasted and re-fed LD datasets revealed changes in proteins abundance upon re-feeding, the resulting data of which is summarized in the volcano plot depicted in **Figure 2.3C**. To facilitate the depiction of proteins uniquely observed in a single experimental condition on the log scale plot and provide an estimate

of their minimal fold-change, proteins with an average EIC=0 for an experimental condition were deemed to be missing not at random (MNAR) and assigned values of the global minimum observed within their dataset. The abundance and distribution of all identified LD-associated proteins are listed in **Supplemental Table 2.2** and those observed to change in abundance following data refinement are listed in **Supplemental Table 2.3**.

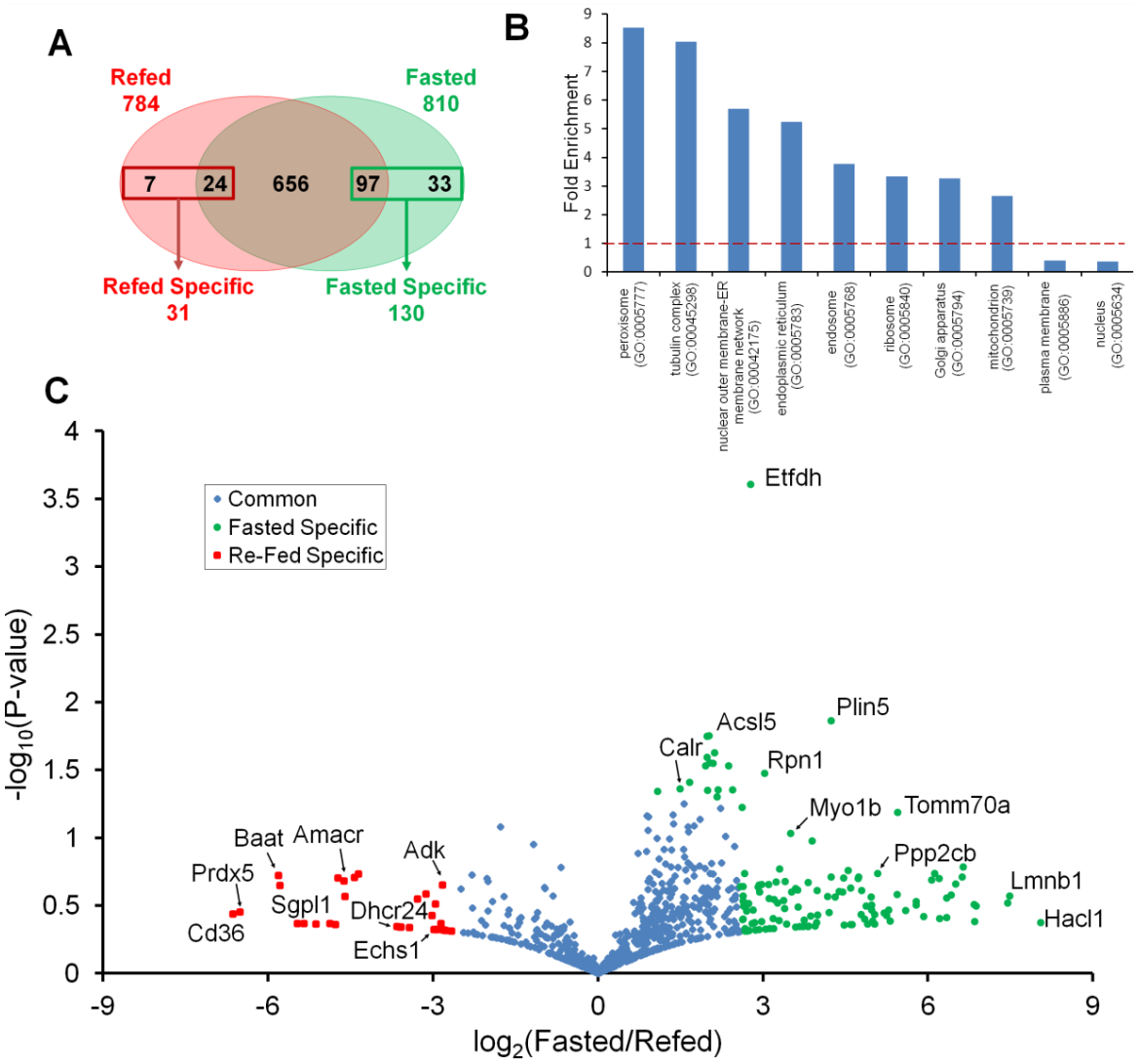


Figure 2.3 Distribution of hepatic LD proteins during fasting and re-feeding.

(**A**) Venn diagram showing the cross-correlation of identified hepatic LD-associated proteins in fasted (**green**) and re-fed (**red**) mice. Numbers enclosed in boxes represent proteins unique (not in central overlap) or determined to be more abundant (~≥6-fold-change), by comparison of the ion intensities of the peptides derived from the proteins (within central overlap), to a specific feeding condition. (**B**) GO analysis enrichment for cell localization for all proteins identified on LDs in both the fasted and re-fed datasets. (**C**) Volcano plot of proteins identified among both energetic states. The x-axis represents the fold-change in average protein abundance observed between fasted and re-feeding conditions as the function $\log_2$ of the corrected extracted ion chromatograms. The y-axis plots the statistical significance of the fold-difference observed between the states as the function $-\log_{10}$(P-Value).

2.3.4. Pathway Analysis of Dynamic LD-Associated Proteins

Utilizing STRING 10.0's functional enrichment tool, KEGG pathway analyses[212] of the 130 LD proteins determined to be significantly more abundant in the fasted experimental condition revealed thirteen distinct pathways (excluding KEGG ID: 1100 – 'Metabolic Pathways') to be up-regulated; these include: ER protein processing, peroxisomal proteins, propanoate metabolism, valine/leucine/isoleucine degradation, fatty acid metabolism, TCA cycle, carbon metabolism, peroxisome proliferator-activated receptor (PPAR) signaling pathway, pyruvate metabolism, arginine and proline metabolism, microbial metabolism in diverse environments, antigen processing and metabolism, and fatty acid degradation. The complete list of these pathways and the proteins identified within these pathways are listed in **Table 2.1**. A complete list of all proteins observed to change in abundance are in **Supplemental Table 2.3**. Interestingly, significantly less pathways were identified using this technique for the 31 proteins determined to be significantly more abundant during the re-fed state; only two pathways were observed (excluding KEGG ID: 1100 – 'Metabolic Pathways'); peroxisomal proteins and primary bile acid biosynthesis (**Table 2.1**). While several of the pathways in both feeding states are populated by individual proteins, we believe this provides additional evidence of the diverse functions these organelles are capable of within the cell.

As pathway analysis revealed an enrichment of different peroxisome-associated proteins in both fasted and re-fed datasets (**Table 2.1**), in addition to peroxisomal proteins exhibiting the highest fold-enrichment in the entire dataset (**Figure 2.3B**), we further evaluated the data for these protein changes. The quantified data from the replicate analysis of the 12 peroxisomal proteins identified by KEGG pathway analysis, in addition to the mitochondrial protein Acsl1 involved in fatty acid oxidation[240,241], as shown in **Figure 2.4,** reveals the magnitude of change for each proteins' observed LD-association upon fasting and re-feeding. While Acsl1 exhibited an approximately 30% decrease in abundance upon re-feeding, this did not meet the cut-off criteria used to list proteins that significantly change. This apparent dichotomy of differential peroxisomal protein association is suggestive of unique roles for LD-peroxisome interactions

during fasting or feeding. It must also be noted that while the analyses revealed an enrichment for peroxisomes based on these proteins' annotations, many are not exclusively peroxisomal and have additional sub-cellular localizations as well.

Table 2.1 KEGG pathway identifiers populated from the proteins enriched in LDs isolated from livers of either fasted or re-fed mice with an FDR <0.01.

Proteins identified in each pathway are listed.

Pathway ID	Pathway Description	Protein Count in Dataset	FDR
Enriched in the Fasted LDs			
1100	**Metabolic pathways**	29	9.70×10^{-9}
	Aass, Man2a1, Man1a, Asah1, Hsd17b2, Rpn1, Pigs, Dld, Acaca, Scp2, Acsl5, Uroc1, Slc27a5, Ndufv2, Dlat, Hibch, Tecr, Maoa, Acadvl, Suclg2, Aldh4a1, Oat, Abat, Prodh, Pccb, Pcx, Hmgcl, Gpam, Lpin1		
4141	**Protein processing in endoplasmic reticulum**	10	5.27×10^{-6}
	Pdia4, Hspa8, Sec63, Rpn1, Calr, Hspa5, Pdia3, Txndc5, Lman2, Man1a		
4146	**Peroxisome**	7	3.81×10^{-5}
	Pex6, Pex1, Pex26, Scp2, Acsl5, Hmgcl, Hacl1		
640	**Propanoate metabolism**	5	5.91×10^{-5}
	Abat, Pccb, Suclg2, Acaca, Hibch		
280	**Valine, leucine and isoleucine degradation**	5	4.51×10^{-4}
	Abat, Pccb, Hmgcl, Dld, Hibch		
1212	**Fatty acid metabolism**	5	4.51×10^{-4}
	Acaca, Acsl5, Tecr, Acadvl, Cpt2		
20	**Citrate cycle (TCA cycle)**	4	1.13×10^{-3}
	Dld, Dlat, Suclg2, Pcx		
1200	**Carbon metabolism**	6	1.21×10^{-3}
	Dld, Dlat, Hibch, Suclg2, Pccb, Pcx		
3320	**PPAR signaling pathway**	5	2.64×10^{-3}
	Scp2, Cpt2, Acsl5, Slc27a5, Apoa5		
620	**Pyruvate metabolism**	4	2.69×10^{-3}
	Dld, Acaca, Dlat, Pcx		
330	**Arginine and proline metabolism**	4	8.08×10^{-3}
	Aldh4a1, Oat, Prodh, Maoa		
1120	**Microbial metabolism in diverse environments**	6	9.93×10^{-3}
	Dld, Dlat, Suox, Suclg2, Pccb, Pcx		
4612	**Antigen processing & presentation**	4	1.67×10^{-2}
	Hspa8, Calr, Hspa5, Pdia3		
71	**Fatty acid degradation**	3	4.35×10^{-2}
	Cpt2, Acadvl, Acsl5		
Enriched in the Re-Fed LDs			
1100	**Metabolic pathways**	13	6.70×10^{-7}
	Idh2, Baat, Ndufb9, Adk, Atp6v1e1, Amacr, Dhcr24, Sgpl1, Echs1, Aox3, Ftcd, Dhrs4, Dpm1		
4146	**Peroxisome**	5	1.15×10^{-5}
	Idh2, Baat, Prdx5, Amacr, Dhrs4		
120	**Primary bile acid biosynthesis**	2	2.03×10^{-2}
	Baat, Amacr		

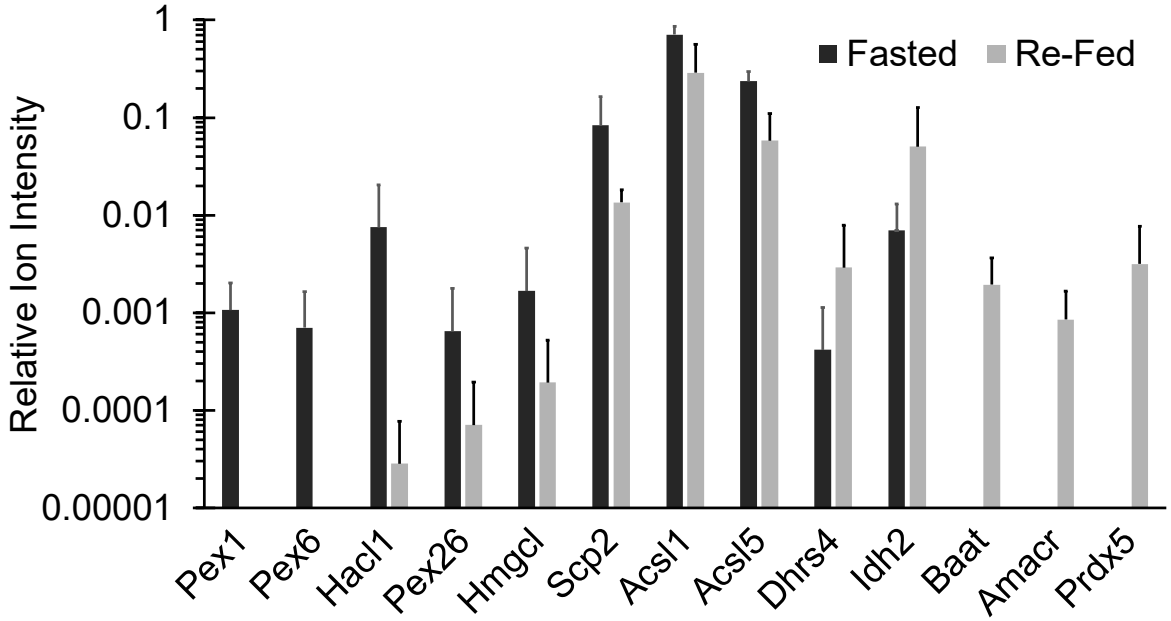


Figure 2.4 Relative abundance of peroxisomal proteins.

Relative abundance of peroxisomal proteins associated with LDs isolated from the livers of mice after fasting (**black**) or re-feeding (**light grey**). Quantification is the average relative EICs, relative to Plin2, for each indicated protein from triplicate analysis.

2.3.5. Global Protein Abundance

A challenge arising from proteomic investigations of sub-cellular purified components is determining what proteins are most relevant from those that are likely contaminants that undoubtedly remain even after purification. To address this, at least in part, we next investigated the proteomic data with respect to PSMs for all proteins across all data sets to obtain estimates of the most abundant proteins in the LD samples and less likely to be low abundance contaminants. The PSMs, normalized to protein length to compensate for the increased number of potential tryptic peptides generated from longer proteins, were compared to estimate which proteins are most abundant with respect to LD composition. From the raw data, **Supplemental Table 2.1**, it is noted that Plin2 is the protein which makes up the largest fraction of the normalized PSMs, indicating this to be one of, if not the most abundant protein(s) in the purified LDs. The average normalized spectral counts across all datasets were then plotted in **Figure 2.5A**

in decreasing abundance relative to Plin2. The data reveals that only 57 proteins of the 817 identified make up 50% of the spectral counts in the experiment, which leads to an estimate that these proteins make up half of the protein abundance in the samples. Correspondingly, 314 and 633 proteins make up 90% and 99% of the spectral counts, respectively. The proteins that comprise 50% of the normalized spectral counts likely represent the most abundant proteins in the purified LDs and their identities are summarized in **Figure 2.5B**.

The unexpected observation of the murine specific secreted major urinary proteins (Mups) populating the list of the 57 most abundant proteins, led to validation experiments of the association of these proteins in the hepatic LDs. Fractions from a sucrose gradient purification of LDs from the livers of fasted mice were resolved by SDS-PAGE and analyzed for total protein abundance by coomasie staining or immunoblotting for Plin2 or Mups (with an antibody specific for a range of Mup isoforms). As seen in **Figure 2.6**, while much of the observed Mup isoforms we observed in the high density heavy membrane associated fraction of the gradient, a significant amount of the Mups are also observed in the low-density fraction containing the LDs.

With the spectral counts and ion intensities observed ranging over several orders of magnitude for the different proteins identified in purified LDs, we evaluated proteins that were determined to significantly change in abundance upon fasting and re-feeding; if only very low abundance proteins were observed to significantly change, it would suggest many of our observed differences may simply reflect statistical noise as a result of proteins being at or near the limit of their detection. The data representing LD-associated proteins that significantly change in abundance during fasting and feeding with respect to their relative ion intensities is overlaid on the plot in **Figure 2.5A**. While there is a trend for proteins with low normalized PSMs to be more frequently observed significantly changing in abundance between feeding states using relative ion intensities, a large proportion of proteins with high normalized PSMs are also observed to significantly change based on their relative ion intensities (**Figure 2.5A**). Acadvl, Cpt2, and Hspa5, identified to significantly change in abundance between feeding states, populate within the list of 57 proteins that make up 50% of the normalized spectral counts.

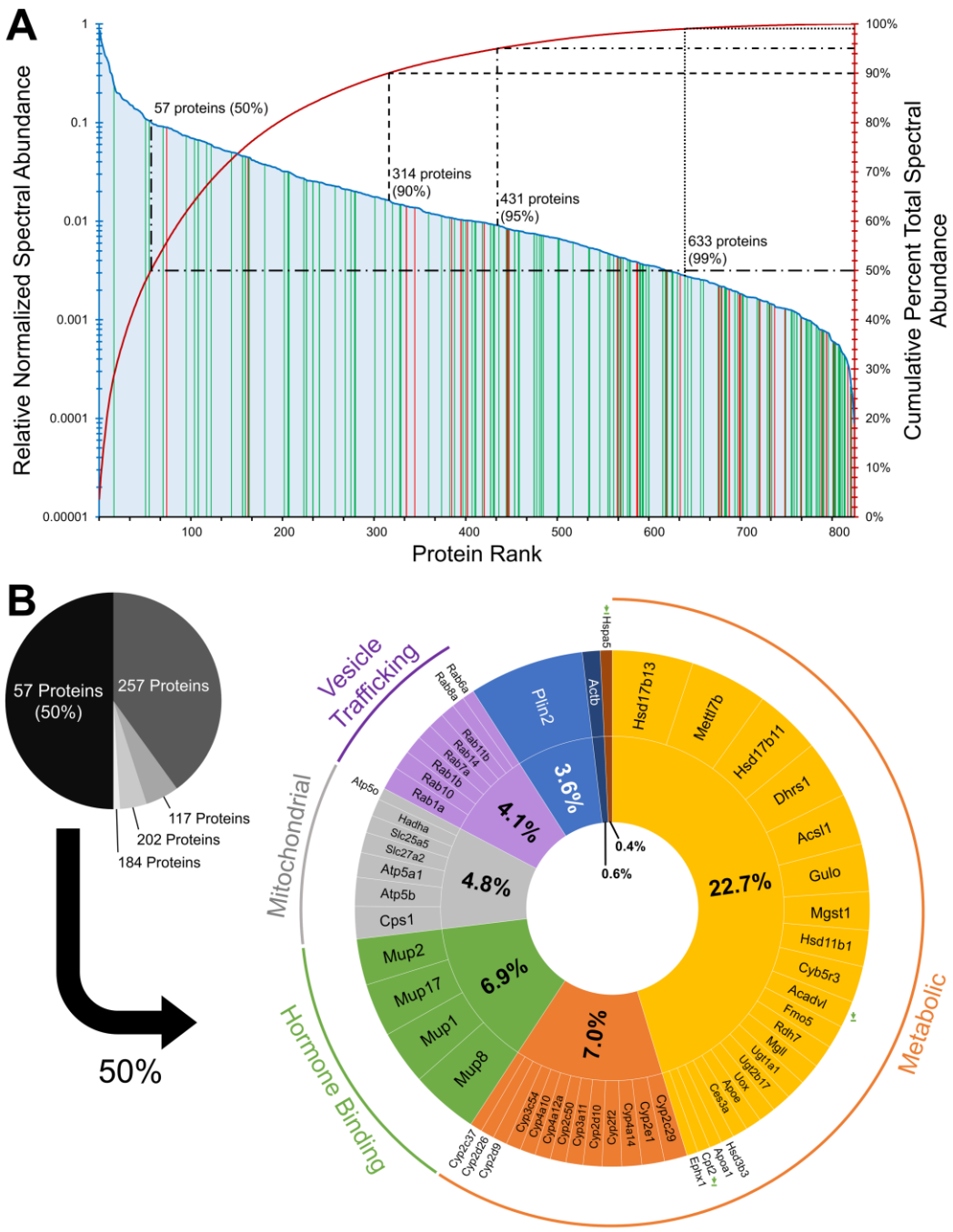


Figure 2.5 Analysis of LD-associated proteins by spectral count abundance.

A) All proteins detected were sorted by abundance (normalized spectral counting, relative to Plin2). The number of proteins that make the total quantified signal, as determined by spectral counting, are indicated. Green horizontal bars indicate proteins with higher abundance in the hepatic LDs of fasted mice, while red bars indicate protein that are higher in abundance in re-fed mice. **B**) A graphic depiction and classification of the 57 proteins that make up 50% of the normalized spectral counts are shown. Green arrows indicate the proteins that are lower in abundance in the liver LDs of re-fed mice.

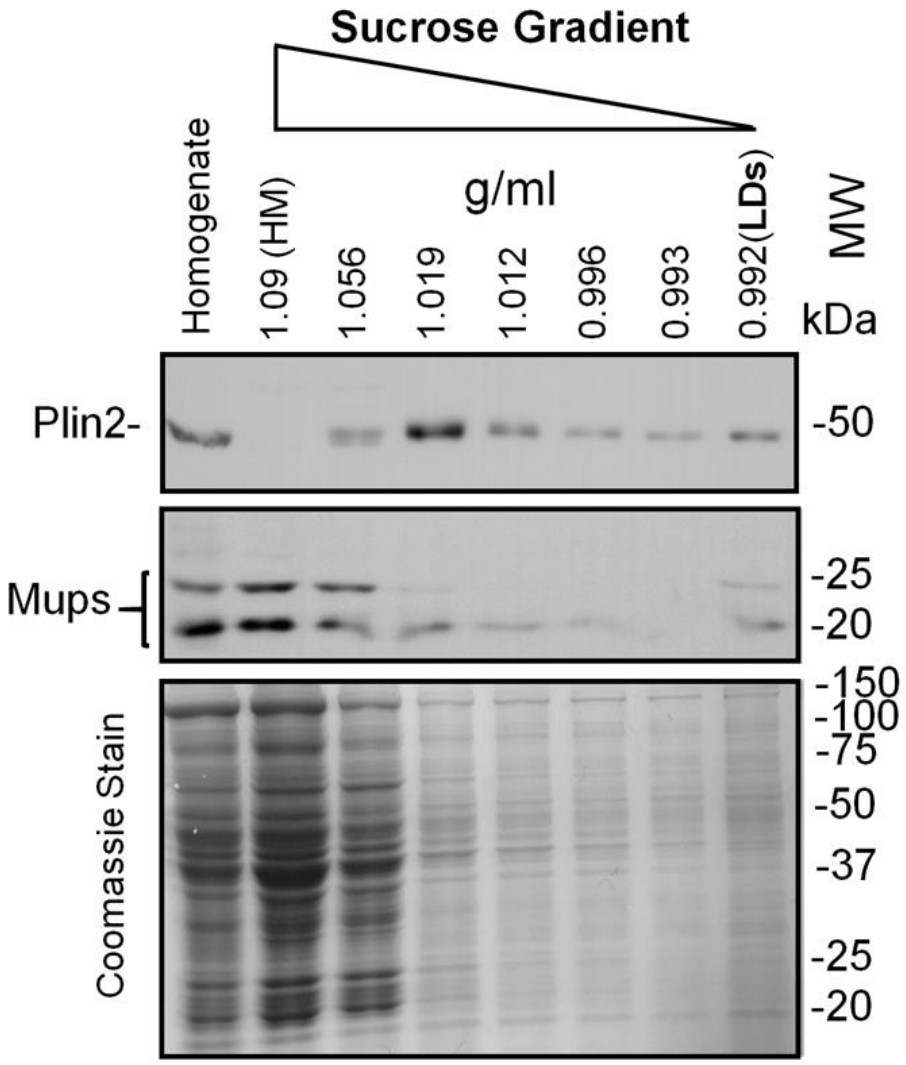


Figure 2.6 Immunoblot analysis of sucrose gradient fractions reveal Mup association to LDs.

Fractions from a sucrose gradient for the purification of LDs were resolved by SDS-PAGE and analyzed by: immunoblotting for Plin2 (**upper panel**), total protein by coomassie staining (**lower panel**) and, immunoblotting for Mups with an antibody targeting a range of Mup isoforms. The fractions containing the heavy membranes (HM) and lipid droplets (LDs) are indicated.

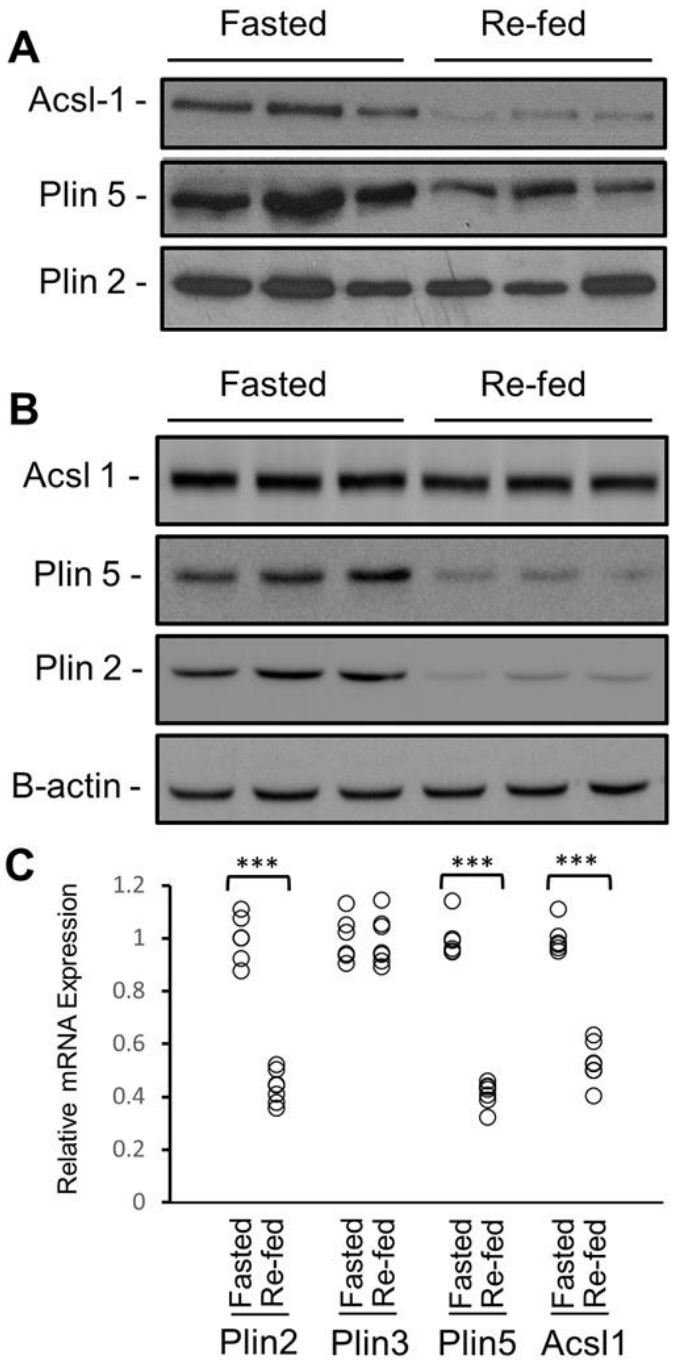


Figure 2.7 Immunoblot and RT-qPCR analysis of representative LD-associated proteins from isolated LDs.

Immunoblots of LD-associated proteins from (**A**) isolated LDs and (**B**) whole liver homogenates from fasted and re-fed mice. Samples were analyzed on the basis of equal triacylglycerol content (**A**) or equal protein (10 μg) content (**B**) from fasted and re-fed mice. (**C**) RT-qPCR analysis of representative LD-associated proteins in livers from fasted and re-fed mice (N=6-7). ***P<0.001 vs fasted group, the significance is based on two-tailed *t*-tests.

2.3.6. Immunoblot Validation of LD-Associated Proteins

To confirm the LD-association of some of the identified proteins, and their changes upon feeding state, we assessed their abundance in purified LDs and whole liver homogenates by immunoblotting. To control for equal loading of isolated LDs between feeding states, sample loading was normalized to the LDs' TG content. Whole liver homogenates were used to evaluate whether the changes observed in proteins associated with LDs is a result of changes in a protein's global cellular abundance, or a change in LD-association; whereas LD loading was normalized to TG content, whole liver homogenates were normalized to total protein.

Plin2 and Plin5 are well characterized LD-associated proteins that play an important role in the regulation of LD turnover by preventing adipose triglyceride lipase (ATGL) - catalyzed lipolysis[242,243]. Plin2/TG ratio did not significantly change in the purified LDs isolated from livers of fasted/refed mice (**Figure 2.7A**). In agreement with our mass spectrometry data for Plin5 (**Supplemental Table 2.2**), LDs purified from fasting livers contained higher Plin5 abundance compared to those from re-fed livers, consistent with diminished lipolysis and increased TG storage (**Figure 2.7A**). Furthermore, both Plin2 and Plin5 exhibit larger increases in abundance, relative to total protein, in the whole liver homogenate upon fasting (**Figure 2.7B**) consistent with increased expression of Plin2 and Plin5 mRNAs during fasting (**Figure 2.7C**). In contrast, Acsl1 revealed an increase in abundance by mass spectrometry (**Figure 2.4**) and Western blot analysis (**Figure 7A**), but no changes were observed in the whole liver homogenates (**Figure 2.7B**), indicating a change in protein localization and not protein abundance in the cell. Interestingly, the expression of Acsl1 mRNA was decreased during refeeding (**Figure 2.7C**) but this change was not translated into lower protein abundance suggesting a long half-life of the protein.

2.4. Discussion

2.4.1. Fasted and Re-Fed Liver LDs

The study presented here was designed to systematically identify the LD-associated proteome from the mouse liver during fasting and re-feeding. Two key issues when analyzing organelle proteomes are assessing the purity of the preparations, as well as how to correct datasets to reduce between-sample variability upon comparison. With respect to sample purity, we expectedly found the canonical LD-marker proteins in our LD preparations, including Plin2 and Plin3. Cnx, an ER-localized protein was present in all studied LD fractions (**Figure 2.2B**); although the presence of this protein on the LD surface was previously reported[244], we analyzed the presence of another ER membrane protein Pemt in order to evaluate LD purity and possible ER contamination. Pemt was nearly completely absent from the isolated LD fractions (**Figure 2.2B**), suggesting minimal contamination of our LD preparations with the ER. Luminal ER proteins (Bip, Pdi) have been previously found in LD preparations from various cells and tissues and a possible mechanism how such proteins could be targeted to LDs has been proposed[245,246].

Ces1d (Ces3/TGH) and Ces1g (Ces1/Es-x) are ER luminal resident lipid hydrolases containing the C-terminal ER-retrieval motif –HVEL[247–249]. Interestingly, we found by immunoblotting that Ces1d partitioned to the LDs, while its close family member Ces1g did not. The mechanism for such selectivity is unclear because both Ces1d and Ces1g were identified via LC-MS/MS on purified LDs from both fasted and refed conditions. Our studies showed that hepatocytes lacking Ces1d contain an increased number of smaller LDs compared to wild-type hepatocytes, suggesting that Ces1d plays a role in LD maturation[237]. The role of Ces1g in LD growth and maturation has not been evaluated, however; unlike Ces1d, lack of Ces1g leads to increased size and number of cytosolic LDs[250], showing functional differences between these two ER-localized carboxylesterases.

2.4.2. Global Proteome Analysis of Liver LDs

Shotgun proteomics was used to characterize the murine hepatic LD proteome of fasted and re-fed mice. Our analysis identified a complete set of 817 proteins that are associated with purified with LD in both energetic states (**Figure 2.3A**). GO analysis for cellular localization of the entire data set revealed a significant enrichment for proteins that have been annotated to a variety of sub-cellular localizations (**Figure 2.3B**). The greatest enrichment observed was for peroxisomal proteins, in agreement with previous work demonstrating a close association between LDs and peroxisomes[251]. As both LDs and peroxisomes are key players in the lipid metabolic flux, their close interactions have been proposed to be key for bidirectional lipid trafficking[252]. As predicted, a depletion of nuclear and plasma membrane proteins was also observed. The enrichment of peroxisomal, ER, mitochondrial, and Golgi proteins was expected as there is a growing appreciation of the interaction and function of LDs with these organelles[227,228]. At first glance, other observed enrichments may appear to be contaminants, such as ribosomal proteins (**Figure 2.3B**) and proteins involved in RNA metabolism (**Supplemental Figure 2.2**); however, these have been previously reported to be associated to the LD proteome[253,254] and have been verified by ultrastructural investigations[254]. The role of ribosomes in LDs is still unclear and remains an open question in LD biology[255]. However, early investigations with model organisms have suggested that LDs may also function as protein storage organelles[224]. Alternatively, the presence of highly abundant proteins in LDs may simply be an artifact of these proteins being trapped in LDs during their rapid formation and growth in the cell. Nonetheless, the agreement of our hepatic LD dataset with previous studies only strengthens these observations.

With ongoing proteomic studies of hepatic LDs, there is an increasing number of proteins identified as being LD-associated[238,239,256,257]. A challenge with these growing lists, as there is when performing proteomics on any organelle, is the analysis and follow-up of the data. As mass spectrometers become more sensitive, the limits of detection will continue to decrease to a point that far exceeds the organelle's biochemical purity. As a result, the instrument's high sensitivity can detect the low abundant contaminants that are undoubtedly present in the sample. To address this, we evaluated the data to identify the most abundant proteins in the LD proteome

with respect to composition (**Figure 2.5**). This perspective is reminiscent of early investigations of the LD proteome, where samples were resolved by SDS-PAGE and individual bands were excised and identified by mass spectrometry[258,259]. The identification of the most abundant proteins in the organelle proteome is predicted to validate well-known proteins already associated with the organelle or identify major components that have been previously overlooked. The 57 proteins that are estimated to make up 50% of the LD proteome (**Figure 2.5B**) fall within both groups. In addition to the expected proteins such as the PAT domain family member, Plin2, and the most abundant group of proteins involved in metabolism (such as Acsl1), several other protein families are observed in high abundance. These include a series of Rab proteins which have been well described with respect to LD trafficking[227,258,260–266], in addition to a number of other enzymes also observed in high abundance from cell culture models such as a series of mitochondrial proteins[267]. Cytochrome P450 proteins have not historically been considered to be associated with LDs, nor were they significantly observed in LDs from hepatic cell lines[259]. However, they have been observed in multiple reports from liver-derived LDs and have been observed to increase in abundance in liver LDs during diet-induced hepatic steatosis[238,239,256].

The most unexpected protein family in our dataset observed in high compositional abundance were the murine specific Mups, a family of rodent-specific proteins with roles in the transport and excretion of pheromones and other lipophilic molecules[268]. With the complexities of the metabolic reactions carried out by LD-associated enzymes, it is not surprising to observe an abundance of these lipophilic carrier proteins; the LD-association of Mups may potentially be where they are loaded with their cargo prior to secretion from the liver. Our observation of Mups being associated to LDs is not the first; supplemental data from other proteomic investigations of murine liver-derived LDs have also identified these proteins[238,239], but they were not discussed as they were not observed to change in abundance between experimental conditions. With our analysis, the Mups only stood out when we queried the data for the most compositionally-abundant proteins in the liver LD proteome. Being so highly abundant, the role of Mups' association with LDs warrants further investigation.

2.4.3. Dynamics of the LD Proteome Upon Feeding

As the focus of our investigations was on the dynamics of the hepatic LD proteome upon fasting and feeding, the quantitative analysis of the data was explored in greater detail. Analysis of our collected datasets (**Supplemental Table 2.2**) revealed that 130 and 31 proteins were found to be more abundant in the fasted and re-fed LDs, respectively (**Figure 2.3A**). Of these changes, it is noteworthy that the catabolic proteins on LDs (**Table 2.1**) change between fasting and re-fed states, in agreement with the fact that lipid accumulation decreases after re-feeding (**Figure 2.1**).

Previous LD proteomic screens with other cellular systems such as yeast[269], mammalian tissue culture[258,259] and germline cells[270] reveal hundreds of proteins with moderate overlap regarding protein composition. Comparing the LD proteomes originating from different tissues provides some insight into the core essential LD proteins and those which may be tissue specific. A comparison of our presented liver LD proteomes with that from the testes of mice[270] reveals an overlap of 159 proteins (~47% of the LD proteome from mice testes; **Supplemental Table 2.4**), including the PAT domain family members Plin2 and Plin3, various lipid metabolizing enzymes, esterases, vesicular trafficking-associated proteins of the Rab- and chaperone-families. Recent reports on murine liver LDs, focusing on diet and disease models[239,256,257] such as hepatic steatosis[238] reveal even higher similarities to our dataset. In the context of these previous investigations, our presented work clearly highlights the dynamic changes undergone by the LD proteome simply upon feeding. As such, our work demonstrates how conditions must be carefully controlled when investigating the LDs of various model systems. Additionally, it is interesting to note from these results how the LD proteome changes during different metabolic states to reflect functions other than fat storage. This provides a clearer picture of the dynamic nature of these organelles.

As the proteins that where observed in both higher and lower abundance upon re-feeding included proteins annotated to be linked to peroxisomes (**Table 2.1**), these data were focused on with interest. Not all peroxisome proteins were observed to change in abundance, as no significant changes in abundance were observed for Pxmp4 and Pex5 (**Supplemental Table 2.2**). While at first the increase and decrease of peroxisomal proteins might seem contradictory, a

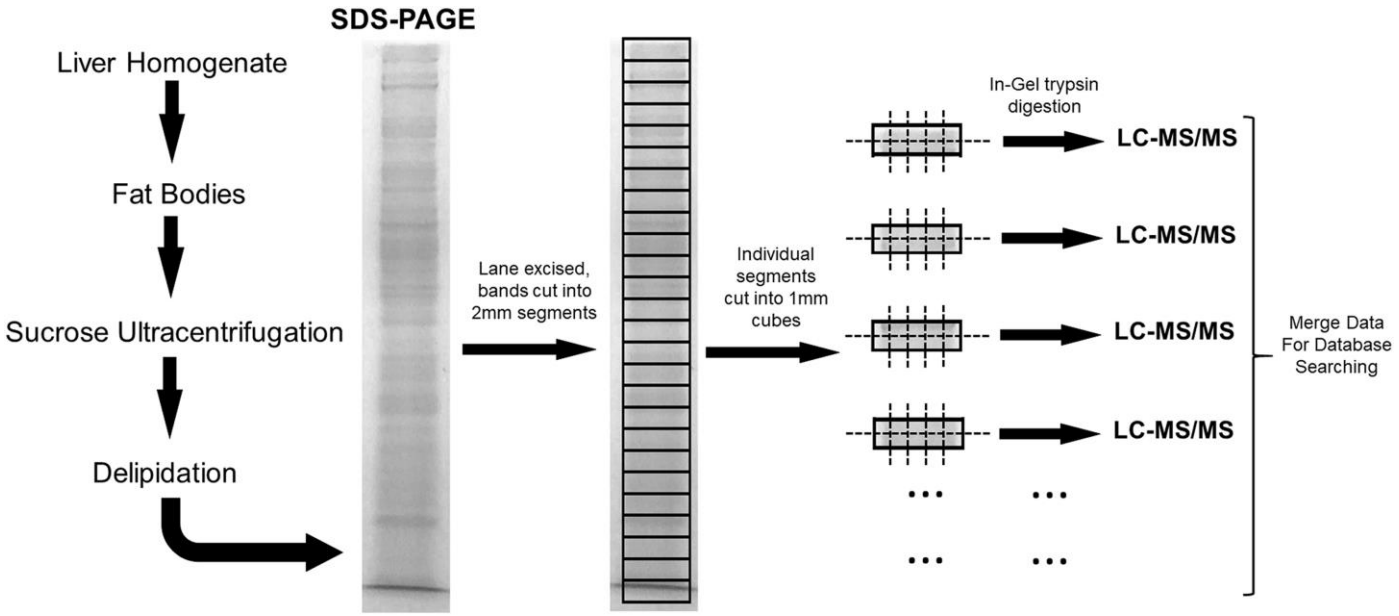
review of their functions is revealing. For example, Baat is observed to be associated in higher amounts with LDs upon re-feeding which is consistent with the function of this enzyme in bile acid synthesis[271].

To validate some of the changes to the LD proteome we observed via LC-MS/MS, several candidate proteins were validated by Western blot analysis of purified LDs. In addition, to ensure the changes observed were due to an increase/decrease in LD-association and a change in global cellular protein abundance, these proteins were also investigated in whole liver homogenates. The proteins for these investigations were Plin5 and Acsl1. In both cases, the changes observed by LC-MS/MS analysis were mirrored by Western blot analysis; for instance, the quantified LC-MS/MS data for Acsl1 (**Figure 2.4**) reveals a similar drop in abundance upon re-feeding as is observed by Western blotting (**Figure 2.7B**).

Here we found that during the fasting state there is enrichment in Plin5 on LDs, along with an even greater increase in global Plin5 abundance in the liver (**Figure 2.7**). Plin5 is mainly expressed in tissues with high levels of fatty acid oxidation, including the heart, liver, and skeletal muscle. It was recently demonstrated that Plin5 antagonizes lipase activities in the heart and is essential for maintaining LDs at detectable sizes[272]. In hepatocytes, Plin5 assists TG accumulation in LDs through a reduction of lipolysis and therefore release of fatty acids that could be substrates for β-oxidation[242]. Increased abundance of Plin5 on LDs during fasting (a catabolic state) is counterintuitive, given that Plin5 is inhibitory to lipolysis. The plausible explanation is that during fasting there is increased flux of adipose tissue-derived fatty acids to the liver with some being directly delivered to mitochondria for oxidation and the excess being stored in LDs. Plin5 could play a regulatory role of modulating lipolysis and hence prevent unregulated flux of fatty acids into mitochondria.

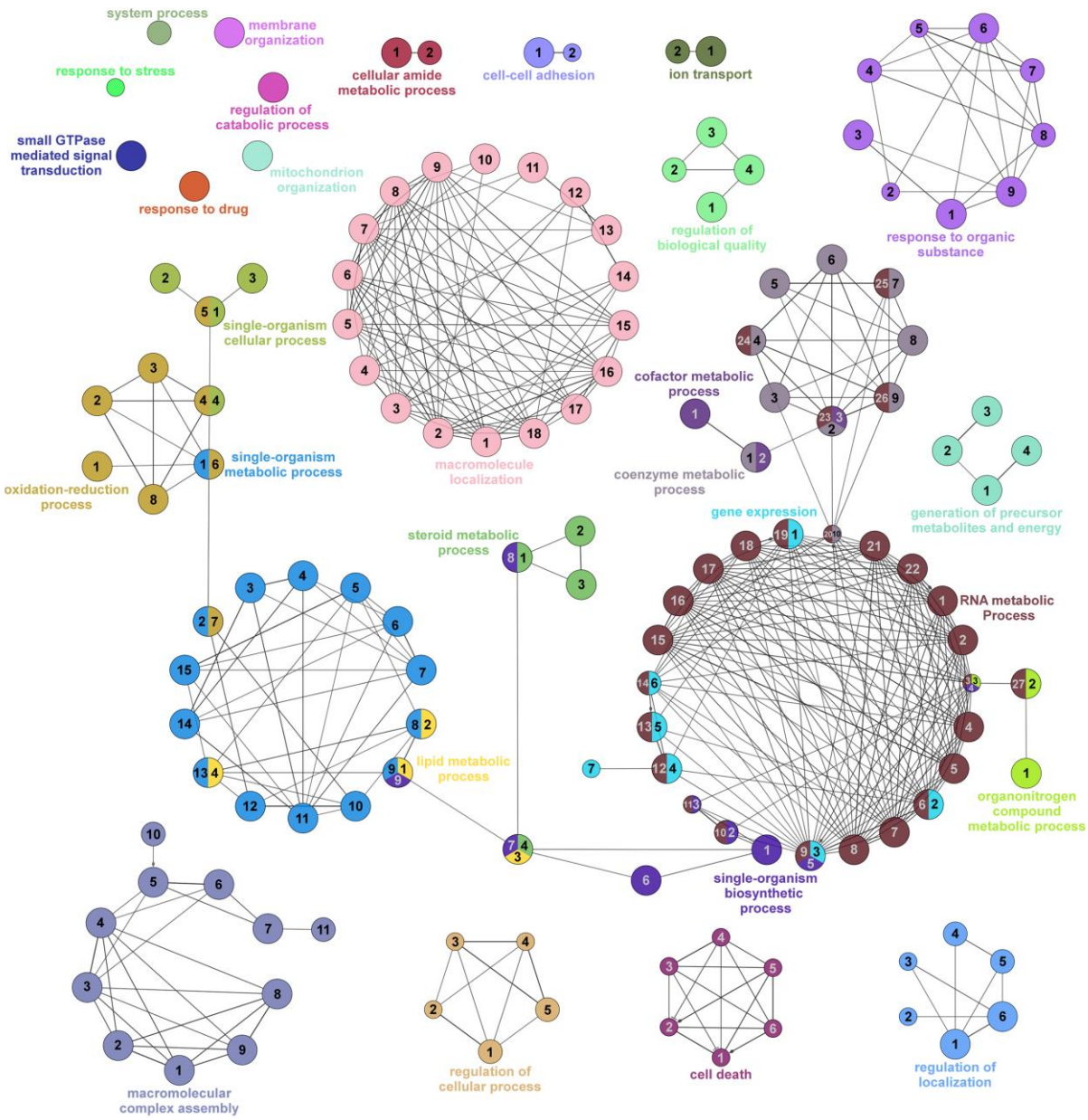
2.5. Conclusions

To our knowledge, this is the first comprehensive study describing the changes in the hepatic LD proteome in response to the physiological demands of fasting and re-feeding. Our analysis of the entire LD proteome adds to the growing understanding regarding the complexities of LD interactions with other cellular organelles. Analysis of LDs' compositional protein abundance has revealed that the major urinary proteins are highly abundant LD-associated proteins in hepatocytes that have been overlooked in previous studies. Additionally, our quantitative analysis revealed that the identified LD proteome is highly dynamic, with 161 proteins, nearly 20% of the observed 817 hepatic LD proteins, significantly varying in abundance during fasting and re-feeding. Immunoblotting validation of selected proteins also revealed that the LD proteome changes occur independent of global cellular protein abundance; large cellular changes in protein abundance observed for Plin2 and Plin5 were not reflected in the LD proteome, while Acsl1 was increased in LDs during fasting, even though its cellular abundance remained unchanged. We believe that the impact of this work is crucial for understanding the general hepatic physiology and particularly the complexity of LD metabolism. Our findings reveal the significance of well-controlled feeding in experimental design when investigating this cellular organelle.



***Supplemental* Figure 2.8 Schematic of the experimental workflow for LD isolation and LD-associated protein identification.**

Livers from three mice were used to make each liver homogenate and the analysis for each condition (Fasted or Refed) was performed in triplicate (18 mice total).



***Supplemental* Figure 2.9 Global GO-Biological Process analysis of fasted and refed lipid droplet proteomes.**

Proteins common to both Fasted and Refed datasets were analyzed with the Cytoscapes's ClueGO application and subject to analysis using Global network specificity, and showing only pathways/terms with a $p<0.05$. From the 791 proteins uploaded, 763 had functional annotations within the database, comprising 140 GO terms under 29 parental group terms. *Legend below*.

Label	GO-Term	GO-ID	Term P-Value*	Group P-Value*
	system process	GO:0003008	780.0E-6	50.0E-6
	response to stress	GO:0006950	14.0E-3	570.0E-6
	mitochondrion organization	GO:0007005	360.0E-12	38.0E-12
	small GTPase mediated signal transduction	GO:0007264	4.8E-9	500.0E-12
	regulation of catabolic process	GO:0009894	7.3E-6	690.0E-9
	response to drug	GO:0042493	12.0E-12	1.3E-12
	membrane organization	GO:0061024	2.9E-6	290.0E-9
1	**cellular amide metabolic process**	GO:0043603	7.8E-6	680.0E-9
2	peptide metabolic process	GO:0006518	2.7E-3	680.0E-9
1	**ion transport**	GO:0006811	1.3E-6	8.1E-9
2	transmembrane transport	GO:0055085	1.1E-3	8.1E-9
1	**cell-cell adhesion**	GO:0098609	41.0E-6	540.0E-6
2	biological adhesion	GO:0022610	11.0E-3	
1	**cofactor metabolic process**	GO:0051186	4.7E-24	1.3E-24
2	coenzyme metabolic process	GO:0006732	1.8E-21	
3	nucleoside phosphate metabolic process	GO:0006753	390.0E-15	
1	**organonitrogen compound metabolic process**	GO:1901564	3.6E-36	460.0E-6
2	cellular nitrogen compound biosynthetic process	GO:0044271	18.0E-3	
3	organonitrogen compound biosynthetic process	GO:1901566	190.0E-9	
1	**generation of precursor metabolites and energy**	GO:0006091	99.0E-33	77.0E-27
2	carbohydrate metabolic process	GO:0005975	300.0E-12	
3	single-organism carbohydrate metabolic process	GO:0044723	98.0E-12	
4	energy derivation by oxidation of organic compounds	GO:0015980	86.0E-30	
1	**regulation of biological quality**	GO:0065008	150.0E-18	16.0E-18
2	chemical homeostasis	GO:0048878	1.2E-3	
3	cellular homeostasis	GO:0019725	210.0E-6	
4	homeostatic process	GO:0042592	9.0E-9	

Label	GO-Term	GO-ID	Term P-Value*	Group P-Value*
1	**lipid metabolic process**	GO:0006629	540.0E-66	
2	cellular lipid metabolic process	GO:0044255	1.8E-51	68.0E-66
3	lipid biosynthetic process	GO:0008610	5.3E-24	
4	fatty acid metabolic process	GO:0006631	54.0E-51	
1	**single-organism cellular process**	GO:0044763	2.5E-42	
2	single-organism process	GO:0044699	410.0E-39	210.0E-30
3	cellular process	GO:0009987	330.0E-12	
4	cellular metabolic process	GO:0044237	2.0E-21	
1	**regulation of cellular process**	GO:0050794	1.2E-3	
2	regulation of biological process	GO:0050789	10.0E-3	
3	single organism signaling	GO:0044700	12.0E-3	180.0E-6
4	cell communication	GO:0007154	30.0E-3	
5	signal transduction	GO:0007165	3.1E-3	
1	**regulation of localization**	GO:0032879	36.0E-6	
2	secretion	GO:0046903	17.0E-3	
3	positive regulation of transport	GO:0051050	8.2E-3	1.4E-6
4	regulation of cellular localization	GO:0060341	1.6E-3	
5	regulation of protein localization	GO:0032880	2.1E-3	
6	regulation of transport	GO:0051049	49.0E-6	
1	**cell death**	GO:0008219	11.0E-3	
2	programmed cell death	GO:0012501	24.0E-3	
3	negative regulation of cell death	GO:0060548	38.0E-3	630.0E-6
4	negative regulation of programmed cell death	GO:0043069	16.0E-3	
5	regulation of programmed cell death	GO:0043067	13.0E-3	
6	regulation of cell death	GO:0010941	12.0E-3	
1	**gene expression**	GO:0010467	470.0E-9	250.0E-9
2	cellular macromolecule biosynthetic process	GO:0034645	8.8E-6	

Label	GO-Term	GO-ID	Term P-Value*	Group P-Value*
3	macromolecule biosynthetic process	GO:0009059	24.0E-6	
4	cellular macromolecule metabolic process	GO:0044260	2.0E-6	
5	macromolecule metabolic process	GO:0043170	2.4E-6	
6	regulation of macromolecule metabolic process	GO:0060255	940.0E-6	
7	macromolecule modification	GO:0043412	12.0E-3	
1	**oxidation-reduction process**	GO:0055114	59.0E-135	
2	primary metabolic process	GO:0044238	2.6E-12	
3	metabolic process	GO:0008152	1.4E-30	
4	cellular metabolic process	GO:0044237	2.0E-21	48.0E-33
5	single-organism cellular process	GO:0044763	2.5E-42	
6	single-organism metabolic process	GO:0044710	2.0E-129	
7	small molecule metabolic process	GO:0044281	38.0E-108	
8	organic substance metabolic process	GO:0071704	4.6E-18	
1	**single-organism biosynthetic process**	GO:0044711	670.0E-42	
2	biosynthetic process	GO:0009058	4.0E-3	
3	organic substance biosynthetic process	GO:1901576	13.0E-3	
4	cellular nitrogen compound biosynthetic process	GO:0044271	18.0E-3	
5	macromolecule biosynthetic process	GO:0009059	24.0E-6	17.0E-18
6	small molecule biosynthetic process	GO:0044283	730.0E-33	
7	lipid biosynthetic process	GO:0008610	5.3E-24	
8	steroid metabolic process	GO:0008202	6.7E-30	
9	lipid metabolic process	GO:0006629	540.0E-66	
1	**response to organic substance**	GO:0010033	220.0E-9	
2	response to lipid	GO:0033993	12.0E-3	
3	response to chemical	GO:0042221	430.0E-6	32.0E-6
4	response to hormone	GO:0009725	3.2E-3	
5	response to nitrogen compound	GO:1901698	9.5E-3	

Label	GO-Term	GO-ID	Term P-Value*	Group P-Value*
6	response to organic cyclic compound	GO:0014070	2.4E-6	
7	response to organonitrogen compound	GO:0010243	2.2E-3	
8	response to endogenous stimulus	GO:0009719	1.4E-3	
9	response to oxygen-containing compound	GO:1901700	1.0E-6	
1	**steroid metabolic process**	GO:0008202	6.7E-30	
2	alcohol metabolic process	GO:0006066	14.0E-27	12.0E-39
3	organic hydroxy compound metabolic process	GO:1901615	170.0E-27	
4	lipid biosynthetic process	GO:0008610	5.3E-24	
1	**coenzyme metabolic process**	GO:0006732	1.8E-21	
2	nucleoside phosphate metabolic process	GO:0006753	390.0E-15	
3	carbohydrate derivative metabolic process	GO:1901135	5.4E-9	
4	purine-containing compound metabolic process	GO:0072521	160.0E-15	
5	phosphorus metabolic process	GO:0006793	990.0E-9	140.0E-3
6	ribose phosphate metabolic process	GO:0019693	25.0E-12	
7	phosphate-containing compound metabolic process	GO:0006796	62.0E-6	
8	organophosphate metabolic process	GO:0019637	6.3E-15	
9	nucleobase-containing small molecule metabolic process	GO:0055086	6.5E-15	
10	nucleobase-containing compound metabolic process	GO:0006139	24.0E-3	
1	**macromolecular complex assembly**	GO:0065003	640.0E-12	
2	macromolecular complex subunit organization	GO:0043933	7.9E-9	
3	cellular component biogenesis	GO:0044085	650.0E-9	
4	cellular component assembly	GO:0022607	740.0E-9	
5	cellular component organization	GO:0016043	11.0E-6	920.0E-9
6	cellular component organization or biogenesis	GO:0071840	11.0E-6	
7	organelle organization	GO:0006996	3.5E-6	
8	protein complex subunit organization	GO:0071822	18.0E-9	
9	protein complex assembly	GO:0006461	3.6E-9	

Label	GO-Term	GO-ID	Term P-Value*	Group P-Value*
10	regulation of cellular component organization	GO:0051128	1.6E-3	
11	single-organism organelle organization	GO:1902589	4.3E-3	
1	**single-organism metabolic process**	GO:0044710	2.0E-129	
2	small molecule metabolic process	GO:0044281	38.0E-108	
3	catabolic process	GO:0009056	250.0E-45	
4	cellular catabolic process	GO:0044248	790.0E-42	
5	cellular lipid catabolic process	GO:0044242	6.3E-36	
6	organic acid catabolic process	GO:0016054	13.0E-54	
7	small molecule catabolic process	GO:0044282	15.0E-57	
8	cellular lipid metabolic process	GO:0044255	1.8E-51	7.8E-117
9	lipid metabolic process	GO:0006629	540.0E-66	
10	organic acid metabolic process	GO:0006082	1.8E-96	
11	single-organism catabolic process	GO:0044712	490.0E-63	
12	oxoacid metabolic process	GO:0043436	450.0E-96	
13	fatty acid metabolic process	GO:0006631	54.0E-51	
14	organic substance catabolic process	GO:1901575	190.0E-39	
15	lipid catabolic process	GO:0016042	350.0E-33	
1	**macromolecule localization**	GO:0033036	320.0E-33	
2	protein transport	GO:0015031	31.0E-24	
3	establishment of localization	GO:0051234	5.8E-30	
4	localization	GO:0051179	280.0E-24	
5	intracellular transport	GO:0046907	360.0E-24	31.0E-24
6	establishment of protein localization	GO:0045184	2.6E-24	
7	establishment of localization in cell	GO:0051649	7.6E-18	
8	protein localization	GO:0008104	130.0E-24	
9	cellular localization	GO:0051641	170.0E-21	
10	vesicle-mediated transport	GO:0016192	3.3E-18	

Label	GO-Term	GO-ID	Term P-Value*	Group P-Value*
11	single-organism intracellular transport	GO:1902582	4.8E-6	
12	single-organism transport	GO:0044765	52.0E-18	
13	single-organism cellular localization	GO:1902580	4.1E-9	
14	single-organism localization	GO:1902578	28.0E-18	
15	cellular protein localization	GO:0034613	4.8E-15	
16	nitrogen compound transport	GO:0071705	1.3E-21	
17	intracellular protein transport	GO:0006886	570.0E-15	
18	organic substance transport	GO:0071702	86.0E-30	
1	**RNA metabolic process**	GO:0016070	1.2E-15	
2	regulation of macromolecule biosynthetic process	GO:0010556	220.0E-12	
3	cellular nitrogen compound biosynthetic process	GO:0044271	18.0E-3	
4	regulation of biosynthetic process	GO:0009889	140.0E-6	
5	nucleobase-containing compound biosynthetic process	GO:0034654	1.0E-6	
6	cellular macromolecule biosynthetic process	GO:0034645	8.8E-6	
7	nucleic acid metabolic process	GO:0090304	4.7E-15	
8	regulation of nucleobase-containing compound metabolic process	GO:0019219	3.8E-9	
9	macromolecule biosynthetic process	GO:0009059	24.0E-6	
10	biosynthetic process	GO:0009058	4.0E-3	27.0E-3
11	organic substance biosynthetic process	GO:1901576	13.0E-3	
12	cellular macromolecule metabolic process	GO:0044260	2.0E-6	
13	macromolecule metabolic process	GO:0043170	2.4E-6	
14	regulation of macromolecule metabolic process	GO:0060255	940.0E-6	
15	aromatic compound biosynthetic process	GO:0019438	3.9E-6	
16	heterocycle biosynthetic process	GO:0018130	14.0E-6	
17	regulation of cellular biosynthetic process	GO:0031326	400.0E-9	
18	regulation of gene expression	GO:0010468	54.0E-9	
19	gene expression	GO:0010467	470.0E-9	

Label	GO-Term	GO-ID	Term P-Value*	Group P-Value*
20	nucleobase-containing compound metabolic process	GO:0006139	24.0E-3	
21	regulation of nitrogen compound metabolic process	GO:0051171	1.6E-6	
22	regulation of RNA metabolic process	GO:0051252	60.0E-15	
23	nucleoside phosphate metabolic process	GO:0006753	390.0E-15	
24	purine-containing compound metabolic process	GO:0072521	160.0E-15	
25	phosphate-containing compound metabolic process	GO:0006796	62.0E-6	
26	nucleobase-containing small molecule metabolic process	GO:0055086	6.5E-15	
27	organonitrogen compound biosynthetic process	GO:1901566	190.0E-9	

Supplemental Figure 2.9 Legend

Denotes Bonferroni Step-down Correction

Chapter 3 : Proteomic Characterization of the *In Vivo* Chemotherapeutic Response of EL4 Lymphoma-Derived Tumours

3.0. Proem

In the previous chapter, we successfully investigated how the murine hepatic-tissue derived lipid droplet proteome changed in response to dietary stress associated with a period of fasting, or fasting followed by refeeding. With an interest in oncology and desiring to extrapolate this technique to determine its suitability to the study of whole-tissues, we opted to perform a study into how the proteomes of EL4-derived lymphoma tumours – grown *in vivo* – changed in response to a chemotherapeutic insult. The murine EL4 tumour model is an established *in vivo* model for the investigation of novel cancer imaging agents and immunological treatments due to the rapid and significant response of the EL4 tumours to cyclophosphamide and etoposide combination chemotherapy. Despite the utility of this model system in cancer research little is known regarding the molecular details of the *in vivo* tumour cell death. Here we report the first in-depth quantitative proteomic analysis of the changes that occur in these tumours upon cyclophosphamide and etoposide treatment *in vivo*.

A version of this chapter has been published as:

Kramer, D. A., Eldeeb, M. A., Wuest, M., Mercer, J., & Fahlman, R. P. (2017). Proteomic characterization of EL4 lymphoma-derived tumours upon chemotherapy treatment reveals potential roles for lysosomes and caspase-6 during tumour cell death in vivo. Proteomics, 17(12), 1700060. http://doi.org/10.1002/pmic.201700060

Supplementary data to this chapter can be found online with the published version of this chapter, or at the following link:

Supplemental Tables

3.0.1. Acknowledgements

I would like to thank Drs. Melinda Wuest and John Mercer for performing the animal work and providing us with the tumour samples used for this study. Additionally, I would like to thank Dr. Mohamed Eldeeb for performing immunblot validation of caspases 3 and 6, and Jack Moore, Paul Semchuk, and Audric Moses of the Alberta Proteomics and Mass Spectrometry Facility (APM) and Lipidomics Core Facility at the University of Alberta for guidance on equipment operation.

3.1. Introduction

Pre-clinical models are valuable tools in cancer research as they provide the foundation for the discovery of new treatments and novel diagnostic agents, while providing insights into fundamental molecular processes associated with cancer initiation, progression and response to therapy. A detailed understanding of specific model systems is essential to fully understand their strengths and limitations. The focus of this investigation is the EL4 lymphoma derived tumour model in C57BL/6 mice. The murine EL4 lymphoma cell line was first established in 1945, following treatment of C57BL/6 mice with 9:10-dimethyl-1:2-benzanthracene[273] and has since become a widely utilized model for investigating lymphoma tumours and tumour apoptosis *in vivo*. Having originated in C57BL/6 mice, EL4 cellular suspensions can be injected into these animals without inducing immunological tissue rejection[274–276]. As a result, the EL4 lymphoma tumour model has been utilized for investigations on novel immunotherapy developments[277,278], novel cancer treatment strategies[279], and investigations into the molecular mechanisms of tumour clearance[280].

The murine EL4 tumour model has also been valuable in oncologic imaging research as tumour-burdened mice typically experience a ≥50% decrease in tumour mass following cyclophosphamide-etoposide combination chemotherapy over the span of a few days[281]. This potent response to treatment has been demonstrated, at least in part, to be a result of tumour cell apoptosis as indicated by the presence of extracellular phosphatidylserine and caspase 3/7 activation[281–283]. This rapid response to chemotherapy and the presence of classical markers for apoptosis has made this mouse model suitable for the investigation of novel tumour-death imaging reagents[275,281,284,285] and novel applications for tumour imaging[286,287].

Despite the wide utilization of the murine EL4 tumour model, much remains unknown regarding the molecular changes that occur within the tumour cells following cyclophosphamide and etoposide co-treatment. The use of large scale -omic technologies in modern pharmacological investigations[288] allows for the identification of potential novel mechanisms of drug-induced cancer cell death[289] or resistance[290] in cancer models. Using modern proteomic analysis, we have quantified the whole proteome changes that occur in EL4 cell derived tumours in response to

cyclophosphamide and etoposide treatment; these changes reflect the contextual cellular milieu of the tumour, such as resident macrophages[291] and vasculature[292], in addition to those cells potentially recruited to the tumour upon treatment[293]. In addition to providing several confirmatory observations for aspects previously reported in this tumour model, our proteomic analysis revealed many proteomic changes that may have utility as novel markers for molecular imaging while providing unique insights into the molecular mechanisms of *in vivo* tumour cell death.

3.2. Experimental Procedures

3.2.1. Animal Work

Animal experiments were performed according to the guidelines of the Canadian Council on Animal Care (CCAC) and were approved by the Cross Cancer Institute Animal Ethics Committee (protocol number AC10171). EL4 cells were obtained from ATCC. A suspension of $1x10^6$ EL4 cells was injected subcutaneously into the left flank of wild-type female C57BL/6 mice and allowed to grow for 7 days. Cell death was induced in these tumours through intraperitoneal injection of cyclophosphamide (50mg/mL) and etoposide (12.5mg/mL) in 50% DMSO/saline, corresponding to doses of 100 mg/kg cyclophosphamide and 38 mg/kg etoposide, on days 7 and 8. Control mice were injected with a 50% DMSO/saline vehicle control of equivalent volume. The mice were sacrificed, and the tumours were excised on day 9 at 28 hours following the second chemotherapy treatment. The excised tumour tissue was flash frozen in a glycerol suspension.

3.2.2. Sample Preparation and Mass Spectrometry

3.2.2.1. Tumour Homogenization and Protein Extraction

Frozen tumours were thawed and washed in ice-cold PBST prior to tissue homogenization. Tumours were mechanically homogenized in homogenization buffer [100mM Tris-HCl; 8% glycerol; 4.8% SDS; 100mM β-mercaptoethanol; 1mM phenylmethylsulfonyl fluoride; 10ng/mL leupeptin; 1X PhosSTOP™ phosphatase inhibitor (Roche)]. A ratio of 4 mL of homogenization buffer to 1 g of tumour was used. Tumours were homogenized over ice using a teflon-piston homogenizer. Homogenates were centrifuged at 1000× *g* for 5 minutes at 4°C, and the resulting supernatants were then subjected to micro-tip sonication with a Q-Sonica sonicator (80% amplitude for 1 minute of total sonication using 5 second bursts). Sample were then clarified by centrifugation at 14,000× *g* for 15 minutes at 4°C.

3.2.2.2. Electrophoresis and In-Gel Protein Digestion

Homogenates (20µL) were resolved by SDS-PAGE using 10% polyacrylamide gels. Protein lanes were visualized by Coomassie Blue staining prior to whole-lane excision. Each lane was subsequently cut into 15 equal bands, with each band corresponding to a region of the gel containing proteins of a distinct molecular weight range as described in **Chapter 2**. Each of the gel fractions was subjected to in-gel tryptic digestion as previously described[294]; the resulting peptides were extracted in three stages (*i* - 1% formic acid/2% acetonitrile in water; *ii* – 1% formic acid/50% acetonitrile in water; *iii* – 1% formic acid/25% water in acetonitrile), pooled, then dried and resuspended in 60µL of 0.2% formic acid in 5% acetonitrile.

3.2.2.3. Mass Spectrometry & Database Search Parameters

Digested peptides were analyzed by LC-MS/MS using a ThermoScientific Easy nLC-1000 in tandem with a Q-Exactive Orbitrap mass spectrometer. 5 µL of each sample was subject to a 120-minute gradient (0% to 45% buffer B; buffer A: 0.2% formic acid; buffer B: 0.2% formic acid in acetonitrile) on a 2 cm Acclaim 100 PepMap Nanoviper C18 trapping column in tandem with a New Objective PicoChip reverse-phase analytical LC column. For data dependent analysis, the top 15 most abundant ions were analyzed for MS/MS analysis while +1 ions were excluded from MS/MS analysis. Additionally, a dynamic exclusion of 10 seconds was applied to prevent continued re-analysis of abundant peptides. For the analysis, a resolution of 35,000 was used for full scans that ranged from 400 to 2000 m/z and a resolution of 17,500 was used for MS/MS analysis. For data analysis, raw data files corresponding to samples comprising an entire gel lane were grouped together and searched using Proteome Discoverer 1.4.1.14's SEQUEST search algorithm using the reviewed, non-redundant *Mus musculus* complete proteome retrieved from UniprotKB on October 16, 2015. Parameters were set as follows: event detector mass precision = 2ppm; spectrum selector minimum precursor mass = 350Da, maximum precursor mass = 5000Da; maximum collision energy = 1000; input data digestion enzyme = trypsin (full) with maximum missed cleavage sites = 2; precursor mass tolerance = 10ppm with fragment mass tolerance =0.01Da; dynamic modifications to peptides = oxidation of methionine (+15.995Da),

deamidation of asparagine and glutamine (+0.984Da); static modifications to peptides = carbamidomethylation of cysteine (+57.021Da). During data processing, the 'Precursor Ion Area Detector' node of Proteome Discoverer 1.4.1.14's SEQUEST workflow editor was implemented to determine the relative extracted ion chromatogram for each protein identified from the raw data. Searched results were filtered using a minimum of 2 medium confidence peptides per protein.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[233,234] repository with the dataset identifier PXD005592.

3.2.2.4. Statistics & Data Analysis

False discovery rates for the resultant searched samples were as follows; "treated" samples' actual relaxed FDRs for individual peptides were 0.0490, 0.0493, 0.0470, 0.0429, and 0.0462 (for replicates 1-5 respectively), while the actual strict FDRs were 0.0097, 0.0099, 0.0097, 0.0081, and 0.0089 (for replicates 1-5 respectively); "untreated" samples' actual relaxed FDRs were 0.0447, 0.0467, 0.0458, 0.0439, and 0.0440 (for replicates 1-5 respectively), while the actual strict FDRs were 0.0092, 0.0092, 0.0093, 0.0082, and 0.0086 (for replicates 1-5 respectively). Protein lists were exported to Microsoft Excel. Protein abundance was determined by looking at each protein's extracted ion chromatogram (EIC). EICs for an entire lane were totalled to comprise the relative total ion current (TIC), then each protein's EIC was divided by the TIC to give a 'proportion-of-total' value per sample. Untreated and treated proteins were compared, and only proteins with an observed EIC≥0 in ≥1 sample(s) were used in comparative data analysis. To determine proteins that changed in abundance between the two sample sets, a two-tailed, heteroscedastic (Welch's) *t*-test[188] was applied to proteins observed in both data sets. Sorted p-values were uploaded to the 'q-value estimation for FDR control' web utility (qvalue.princeton.edu) [200–202,295] to generate false-discovery rates (q-values). To determine fold-changes between protein abundances, a $\log_2$ function was applied using the standardized average EICs for each protein. The complete set of the proteomic data collected is provided in **Supplemental Table 3.1**.

For functional analysis of the proteins that exhibit altered expression, the proteins of interest were analyzed with DAVID[208] v6.7 (https://david.ncifcrf.gov/home.jsp) and enriched for KEGG pathway identifiers.

3.2.3. Western Blot Analysis

For validation of protein expression by Western Blot analysis, tumour lysates were resolved by SDS-PAGE on a 12% gel. Proteins were transferred to a nitrocellulose membrane (LI-COR) and the membranes were blocked with 2.5% fish skin gelatin blocking buffer (0.5% of Cold Water Fish Skin Gelatin (Truin Science) in 1× phosphate buffered saline - pH 7.4 with 0.1% Triton X-100) and probed with primary and secondary antibodies and imaged with an Odyssey infrared imaging system using the manufacturer's recommended procedures (LI-COR). The rabbit anti-β-actin antibody (I-19, sc-1616-R) was purchased from Santa Cruz Biotechnology. The rabbit anti-caspase-3 antibody (2H334, ab17819) was purchased from Abcam and the rabbit anti-caspase-6 antibody (#9762) was purchased from Cell Signalling Technologies. The secondary goat anti-rabbit antibody labelled with IRDyes was purchased from LI-COR.

3.3. Results and Discussion

3.3.1. Tumour Treatment and Collection

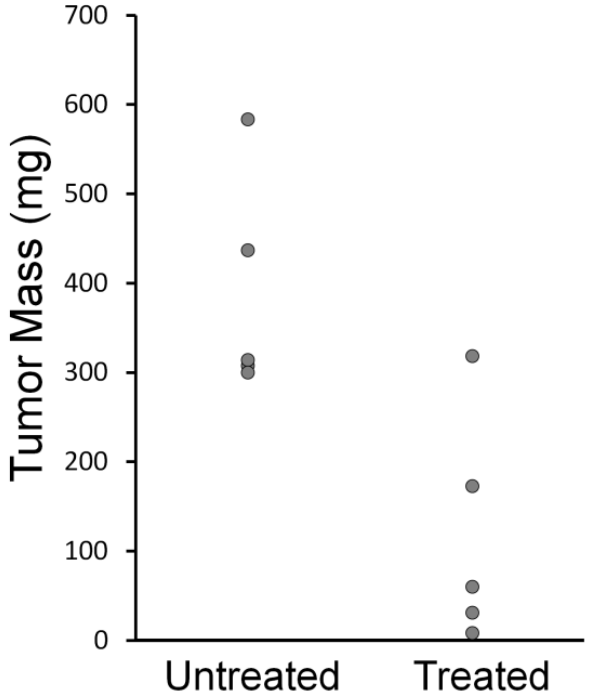


Figure 3.1 EL4 tumour mass reduction after treatment with cyclophosphamide and etoposide combination therapy.

Following injection of EL4 cells, tumours developed for 7 days, after which the tumour bearing mice were treated with cyclophosphamide and etoposide (Treated) or a DMSO/saline control (Untreated) on days 7 and 8. The mice were sacrificed, and tumours excised 28 hours following the second treatment. Untreated tumours had an average mass of 388 ± 123mg, while treated tumours had an average mass of 118 ± 128mg (each n=5; p=0.009), equating to a ~70% decrease in size.

Ten tumours were formed in ten C57BL/6 mice by injecting cultured EL4 cells. As summarized in **Figure 3.1**, the untreated tumours had an average mass of 388±123 mg while the treated tumours had an average mass of 118±128 mg (n=5; $p<0.01$). This statistically significant decrease in tumour mass upon treatment with cyclophosphamide and etoposide mirrors previous reports utilizing this *in vivo* tumour model[281,282,285]. The isolated tumours were then homogenized and analyzed by Gel/LC-MS/MS analysis for quantitative proteomic analysis.

3.3.2. Proteome Analysis

Whole proteome analysis of the tumours by Gel/LC-MS/MS analysis resulted in the identification of a total of 5838 unique proteins, using a minimum criteria of two unique peptides per protein, across all ten tumours analyzed (**Supplemental Table 3.1**). Of these identified proteins, 5687

were identified with quantifiable extracted ion chromatograms (EIC), for label-free quantification, in either the treated or control tumours. Of these 5687 proteins with EICs above the limit of detection, 5038 were observed in the untreated group of which 1271 proteins were uniquely observed. In the treated group, 4416 proteins were observed of which 649 proteins were uniquely observed.

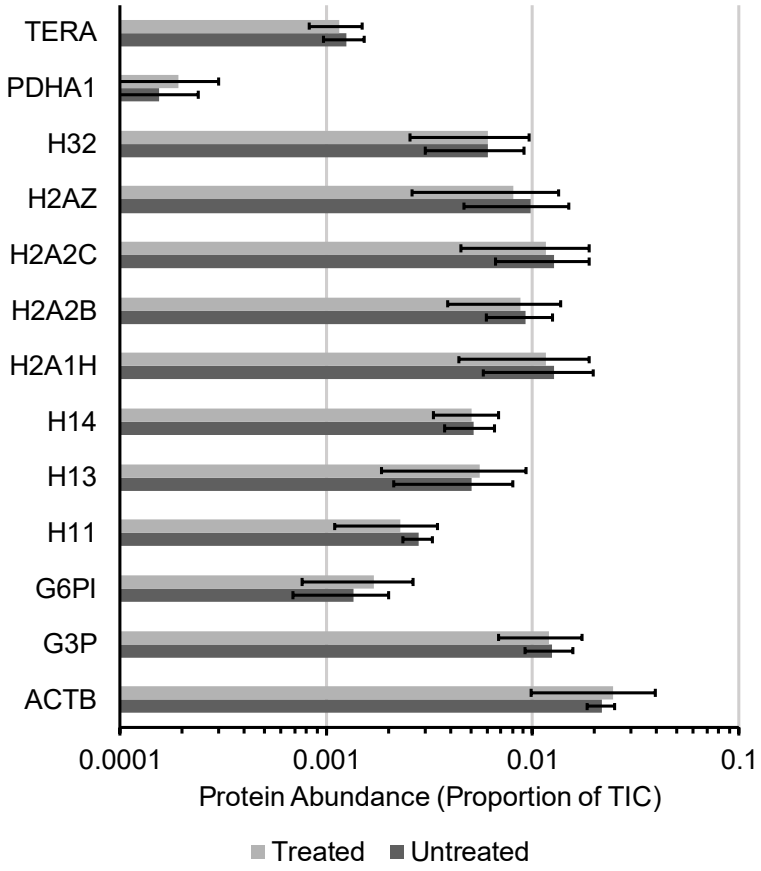


Figure 3.2 Data normalization controls.

Several housekeeping proteins were selected to ensure our EIC/TIC data correction technique was valid for the treated (**light grey**) and untreated tumours (**dark grey**). These include cytoplasmic actin (ACTB), transitional endoplasmic reticulum ATPase (TERA), glyceraldehyde-3-phosphate dehydrogenase (G3P), and the histones H1.3, H1.4, H3.2, H2A-2B, H2A-2C, and H2A-1H.

For quantitative comparisons of the individual samples using the EIC intensities, each dataset was first internally normalized for total protein abundance as described in section 2.5. To evaluate the result of data normalization, the relative EICs for several proteins frequently used as loading controls are compared in **Figure 3.2**. These include actin, histones H1.3, H1.4, H3.2, H2A-2B, H2A-2C, H2A-1H, GAPDH, and TERA. For these proteins, no statistically significant differences in abundance were determined.

For label-free comparison of the proteomes of tumours from mice either treated with combination cyclophosphamide-etoposide or a DMSO/saline control, the EIC intensities of each

protein were compared by individual Welch's *t*-tests. The complete set of statistical comparisons of the normalized data are listed in **Supplemental Table 3.2**. The resulting data, with the fold-change of each protein's average EIC intensities, is depicted as a volcano plot in **Figure 3.3**. To facilitate the depiction of proteins uniquely observed in a single experimental condition and provide an estimate of their minimal fold-change, proteins with an average EIC=0 for an experimental condition were assigned values of the global minimum observed within their dataset. The comparative analysis reveals many statistically significant changes to the proteome and an even distribution of the data points. Because of the multiple comparisons problem when comparing large proteomic data sets by *t*-test comparisons[184], the q-values were determined as an estimate of the FDR for a given p-value cut off. The graphical insert of **Figure 3.3** summarizes the number of proteins that were observed to change in abundance with p-value cut-offs of 0.0001, 0.001, 0.01, and 0.05, with corresponding q-values, as an estimate of FDRs of 0.0164, 0.0202, 0.0460, and 0.0921 respectively. With these cut-offs 19 proteins are determined to have changes at a p-value <0.0001 (17 proteins higher in the untreated tumours and 2 higher in the treated tumours), 77 proteins change at a p-value <0.001 (63 higher in the untreated tumours and 14 higher in the treated tumours), 348 proteins change at a p-value <0.01 (291 higher in the untreated tumours and 57 higher in the treated tumours), and 875 proteins change with a p-value <0.05 (679 higher in the untreated tumours and 196 in the treated tumours).

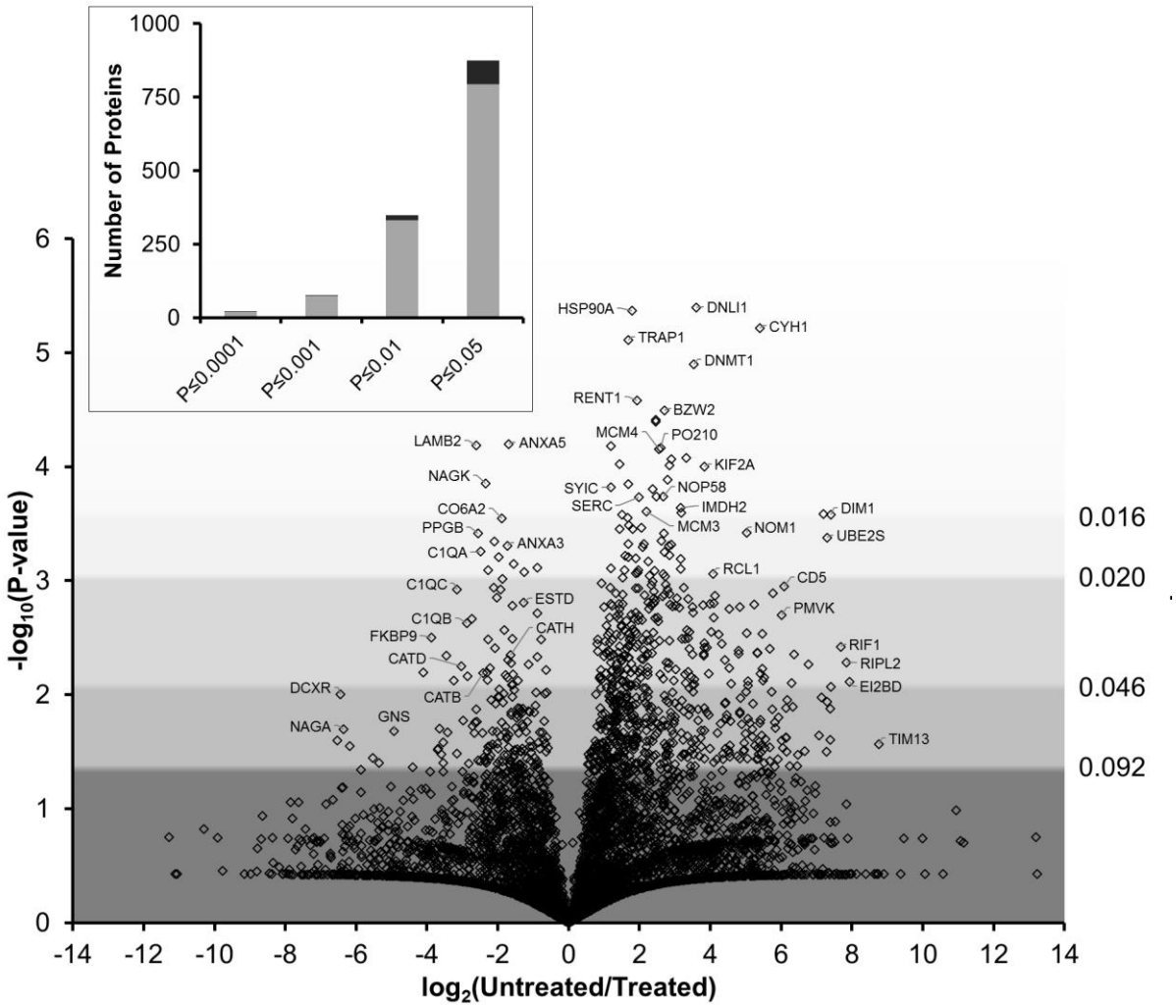


Figure 3.3 Volcano plot of the protein changes observed between untreated and treated tumours.

The $-\log_{10}$ of p-values generated by *t*-test comparisons between corrected protein abundance from each experimental condition was plotted against the $\log_2$ of the ratio of averaged corrected EIC for protein abundance. In addition, the corresponding q-values as an estimate for the FDR of the p-values are listed on the left vertical axis. Proteins not observed in an experimental condition were assigned the global minimal EIC for the sample to facilitate their representation in the plot. (**Insert**) The number of proteins observed to change in abundance upon cyclophosphamide-etoposide treatment are plotted for each p-value cut-off. For each p-value, the q-value estimation for the FDR is shown as the black shading as the fraction of potential false positive proteins determined to change in abundance at a given p-value cut off.

3.3.3. Functional Analysis of the Altered Proteome

For a global analysis of the proteome changes in the EL4 derived tumours upon cyclophosphamide-etoposide treatment, we bioinformatically investigated the function of the proteins that changed in abundance. For this analysis, proteins which met the criteria of a p-value cut-off of ≤0.01 with the corresponding q-value of ≤0.046 were investigated. The proteins that were either up or down regulated in the tumour upon treatment were analyzed with the DAVID v6.7 functional annotation tool for KEGG pathway identifiers[208]. For the proteins with decreased abundance following cyclophosphamide-etoposide treatment, higher in the untreated tumours, 154 of the 291 proteins populated a list of 11 pathways with a $p<0.05$ (**Table 3.1**). Not unexpectedly, the pathways identified are known to be active in cells undergoing active proliferation and include; DNA replication and repair, splicing, translation, purine and pyrimidine metabolism, ribosomal proteins, and protein processing in the ER. For the proteins upregulated in the treated tumours, 40 of the 57 proteins populated a list of 8 pathways with a $p<0.05$ and include; the lysosome, amino acid metabolism, complement and coagulation cascades, and a few pathways involved in the response to pathogens (**Table 3.2**). While the tumours in our investigation were not infected, the enrichment of various pathogenic response pathways resulted from the up regulation of the compliment C1q subcomponent proteins; CIQA, CIQB, and CIQC are involved in various immunological responses. Therefore, the increased abundance of these proteins is suggestive of an immunological response occurring within the tumour upon treatment. These observations are in accordance with previously reported roles for infiltrating macrophages and lymphocytes in EL4 tumourigenesis[291], and with findings suggesting functions for the complement cascade and infiltrating neutrophils following radiation therapy in this model[293,296].

Table 3.1 KEGG pathway identifiers enriched in the untreated tumours.

Proteins with higher observed abundance in the Untreated dataset at p<0.01 were analyzed for KEGG pathway enrichment. The proteins identified in each group are listed.

KEGG Identifier	Term	Proteins	P-Value
mmu03030	**DNA replication**	13	1.3E-12
	DNLI1, DPOLA, DPOD1, MCM2, MCM3, MCM4, MCM5, MCM7, RFC2, RFC3, RFC4, RFC5, RFA1		
mmu03013	**RNA transport**	19	5.4E-09
	F4A2, EIF1B, EIF3A, EIF3B, EIF3F, EIF3G, EI3JA, XPO5, NU107, NU133, PO210, NU188, NUP53, PININ, DDX20, RENT1, STRAP, EI2BB, EI2BD		
mmu03430	**Mismatch repair**	7	3.2E-06
	DNLI1, DPOD1, RFC2, RFC3, RFC4, RFC5, RFA1		
mmu03420	**Nucleotide excision repair**	8	2.2E-05
	CUL4B, DNLI1, DPOD1, RFC2, RFC3, RFC4, RFC5, RFA1		
mmu03040	**Spliceosome**	12	6.1E-05
	HSP72, HNRPU, AQR, NH2L1, PRP8, PR38A, DDX5, SRSF2, SRSF3, SF3A1, SF3B3, RU17		
mmu00230	**Purine metabolism**	13	2.1E-04
	5NT3B, NUDT5, DPOLA, DPOD1, RPB1, RPB2, RPAC1, GUAA, IMDH1, IMDH2, PUR6, PUR4, PUR2		
mmu03008	**Ribosome biogenesis in eukaryotes**	9	2.2E-04
	NHP2, NAT10, NH2L1, NOP56, NOP58, PWP2, RCL1, UTP15, WDR43		
mmu00240	**Pyrimidine metabolism**	8	4.2E-03
	5NT3B, PYRG1, PYRD, DPOLA, DPOD1, RPB1, RPB2, RPAC1		
mmu03010	**Ribosome**	9	8.0E-03
	RM19, RM09, Rps14, Rps15A, Rps21, Rps26, RplA0, Rpl22L, Rpl4		
mmu00970	**Aminoacyl-tRNA biosynthesis**	6	9.7E-03
	SYAC, SYRC, SYG, SYIC, SYSC, SYTC		
mmu04141	**Protein processing in endoplasmic reticulum**	8	4.9E-02
	CALX, ERO1A, HS105, HS90A, HS90B, HSP72, NSF1C, UB2D3		

Table 3.2 KEGG pathway identifiers enriched in the cyclophosphamide-etoposide treated tumours.

Proteins with higher observed abundance in the Untreated dataset at $p<0.01$ were analyzed for KEGG pathway enrichment. The protein identified in each group are listed.

KEGG Identifier	Term	Proteins	P-Value
mmu05150	**Staphylococcus aureus infection**	6	4.5E-06
	C1QA, C1QB, C1QC, CFAB, ITAM, FCGR2		
mmu04142	**Lysosome**	6	3.4E-04
	NAGAB, CATB, CATD, MA2B1, PPGB, CATH		
mmu05133	**Pertussis**	4	6.0E-03
	C1QA, C1QB, C1QC, ITAM		
mmu04610	**Complement and coagulation cascades**	4	6.5E-03
	C1QA, C1QB, C1QC, CFAB		
mmu01200	**Carbon metabolism**	4	2.0-E02
	CATA, IDHC, ESTD, TPIS		
mmu00380	**Tryptophan metabolism**	3	2.3E-02
	ALDH2, AOFB, CATA		
mmu00330	**Arginine and proline metabolism**	3	2.5E-02
	ALDH2, AOFB, KCRB		

3.3.4. Down Regulation of Ribosomes

The reduction in many proteins involved in normal cell growth and proliferation was expected in the cyclophosphamide-etoposide treated tumours, as this treatment leads to a significant drop in tumour mass (**Figure 3.1**). A reduction in ribosome abundance has been well documented to be associated with a variety of cellular stress responses[297,298] and has been rationalized in many ways, including the high metabolic cost of ribosome biosynthesis[299]. Our initial detailed analysis of the proteome was thusly focused on all the constituent ribosomal proteins within the data. For this we examined all the large and small ribosomal subunit proteins identified in the complete data set. A global down regulation of ribosomes is predicted to result in the depletion of all ribosomal proteins. This view is a little oversimplified as a number of ribosomal proteins are known to have alternative 'off-the-ribosome' functions[300], but the general trend is nonetheless predicted to hold true.

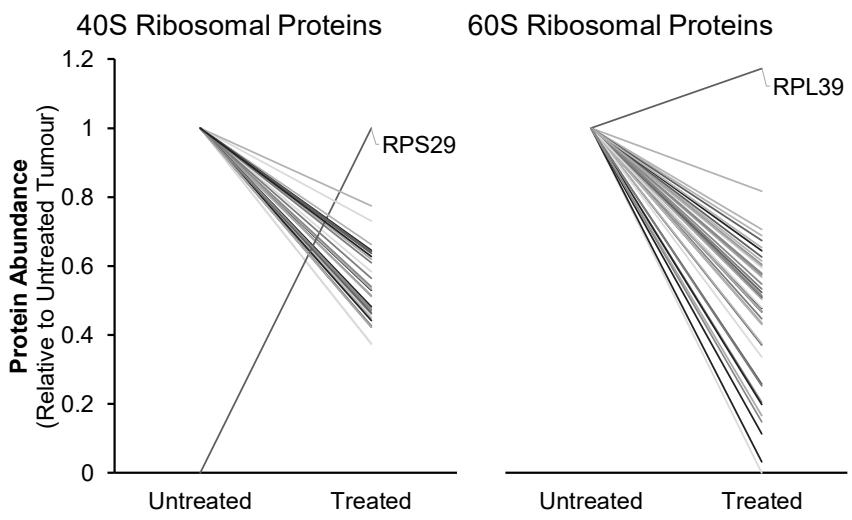


Figure 3.4 Relative change in ribosomal protein abundance following cyclophosphamide-etoposide treatment.

The general trend observed for both 40S (**Left panel**) and 60S (**Right panel**) ribosomal proteins was a decrease in abundance following tumour cyclophosphamide-etoposide combination therapy. The two exceptions that deviate from this trend are 40S ribosomal protein S29 (RPS29) and 60S ribosomal protein L39 (RPL39).

The relative EICs for all the observed ribosomal proteins, 32 small ribosomal subunit proteins and 45 large ribosomal subunit proteins, in the untreated and treated tumours reveals an almost uniform trend in the down regulation both the large and small ribosomal subunit proteins with two obvious exceptions (**Figure 3.4**). This trend is even observed for the ribosomal proteins that had not been initially identified to be down regulated using statistical cut-off criteria. The two exceptions to the trend are RPL39 and RPS29 which have been linked to cancer development[301,302] and Diamond Blackfan anemia[303], respectively. The complete list of the ribosomal proteins quantified in the figure are listed in **Supplemental Table 3.3**. While these changes in ribosomal protein amounts were expected, they provide some validation for the label-free proteomic analysis to detect changes in the proteome of the tumours upon cyclophosphamide-etoposide treatment.

3.3.5. Caspase- and Granzyme-Family Protease Expression

As cyclophosphamide-etoposide treatment of EL4 cell derived tumours in C57BL/6 mice results in classical apoptotic markers, such as the surface localization of phosphatidyl serine[285], it has been assumed apoptosis plays a major role in the efficacy of the drug treatment. We next

evaluated the data for the presence of both the caspase family proteases and granzymes. **Figure 3.5A** summarizes the quantified relative EICs for both the caspases (caspase 1, 2, 3, 6, 7, 8, and 9) and granzymes (granzyme A, B, C, D, E, F, and G) observed. Of these, the only statistically significant changes observed were for caspase-3 and caspase-6, where caspase-3 was down regulated two-fold ($p=0.036$) and caspase-6 was up regulated 13-fold ($p=0.03$) in response to treatment. As caspases, like many proteases, are synthesized as zymogens, the quantification of protein levels cannot be directly interpreted to activity. We had queried our data for signature peptides that may reflect the active state of the protease[304], but these were not observed when analyzing the entire dataset nor when analyzing the region of the gel corresponding to the active caspases.

To validate the changes in protein abundance quantified by mass spectrometry, lysates for two control tumours and two treated tumours were analyzed by Western Blot for both caspase-3 and -6, using actin as a loading control. As seen in **Figure 3.5B**, Western Blot analysis confirmed results we observed with mass spectrometry. In concordance with our mass spectrometry findings, Western Blot analysis was also unable to detect any significant accumulation of either active caspase-3 or -6, which is typically observed as p17/p12 for caspase-3 or p18/p10 for caspase-6.

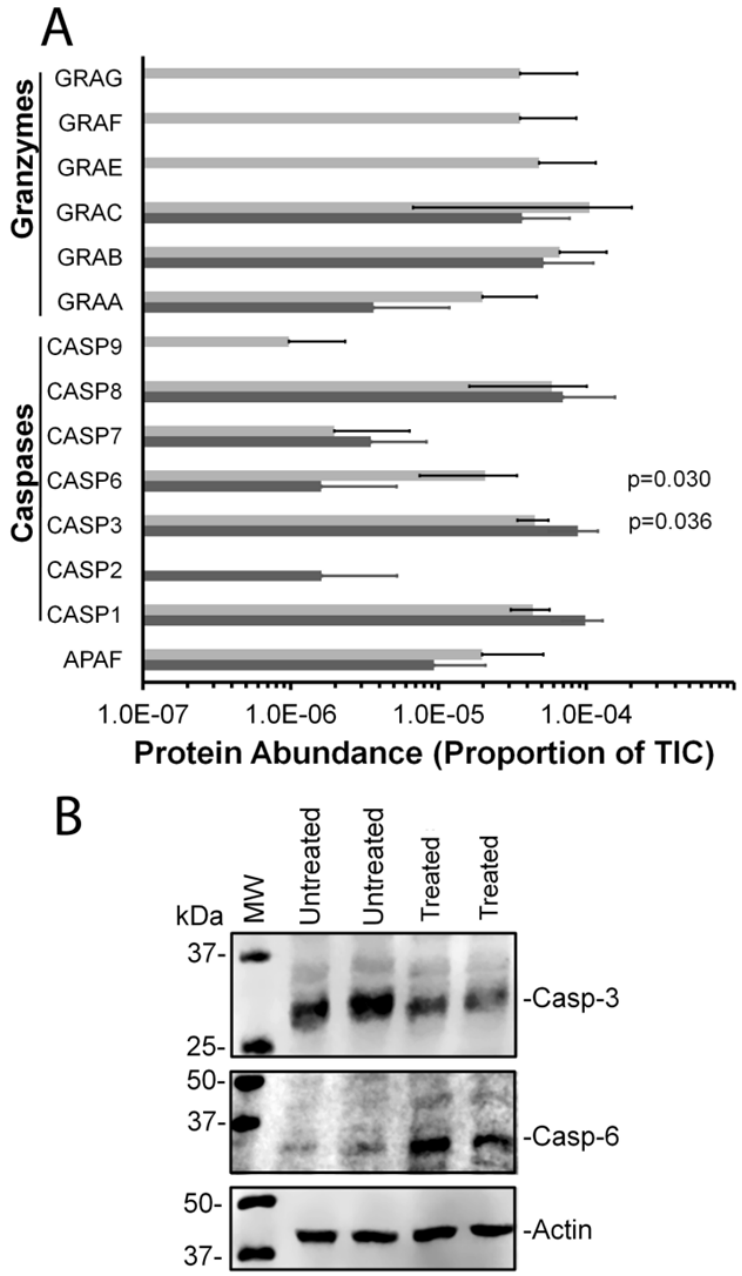


Figure 3.5 Changes in protein abundance of caspase- and granzyme-family proteases.

(**A**) Average measured protein abundance levels from the treated (**light grey**) and untreated (**dark grey**) tumours for apoptotic protease activating factor-1 (APAF), caspase-family proteases (CASP1/2/3/6/7/8/9), and granzyme-family proteases (GRAA/B/C/E/F/G) were examined. (**B**) Western-blot analysis was performed on two randomly selected tumour homogenates from each experimental condition to validate the observed changes in caspase-3 and caspase-6 abundance. The bands reflect full-length caspases. The fragments corresponding to active caspase-3 (p17 and p12) and caspase-6 (p18 and p11) were not observed in the corresponding regions of the gel (data not shown). In addition, Western-blot analysis was performed for actin as a loading control.

Our inability to detect active caspases in the bulk treated tumour extracts is consistent with previous investigations where microscopy performed on treated tumours revealed only a minor number of cells containing active caspase-3[282]. The lack of accumulation of cells with active caspase-3 follows the understanding of apoptosis *in vivo*; cells undergoing active apoptosis are rapidly cleared by surrounding tissue[305]. Prior to full activation of the apoptotic program it is known that active caspase-3 is difficult to detect because of its metabolic instability and it being rapidly degraded in cells[306]. Taken together it is suggestive that our data reflects the global state of the tumour upon treatment and not specifically that of the dying tumour cells. The lack of detectable apoptotic markers indicates that the majority of cells have not committed to the full activation of apoptosis, and the proteomes may be reflecting survival mechanisms in response to treatment, cellular responses leading cell death, and/or the recruitment of infiltrating immune cells[293,296].

With a focus on known proteins involved in cell death, we also queried our proteomic dataset for Bcl-2 family members because of their well-known roles in apoptosis[307]. While proteins such as Bax, Bak, Bid, and Bcl-2 like protein 13 were detected in the analysis, the quantified data did not reveal statistically confident differences for any of them (**Supplemental Table 3.2**).

3.3.6. Lysosomal Protein Accumulation

Upon observing the lysosome as being the second highest ranking KEGG pathway identifier in the treated tumours at the $p<0.01$ cut-off, we chose to further investigate this family of organellar proteins using more lenient criteria. At $p<0.05$, the number of lysosomal proteins identified via KEGG is 19. With this relaxed criteria, additional lysosomal proteins are identified; N-acetylglucosamine-6-sulfatase (GNS) with an abundance 30-fold higher in the treated tumours ($p=0.021$), and group XV phospholipase A2 (PAG15) with a 21-fold higher abundance in the treated tumours ($p=0.044$). In addition, several cathepsin family proteases are observed. As lysosomal cathepsins play an integral role in lysosomal function during autophagy and cell death[308,309], we closely examined the relative EICs for the nine cathepsins observed in the dataset.

The cathepsins, on average, were approximately 3.9-fold more abundant in the tumours responding to cyclophosphamide-etoposide treatment (**Figure 3.6**). While the changes for several of the cathepsins was not statistically significant due to EIC variability, several exhibited statistically significant increases. Cathepsin D exhibited an 8.2-fold (p=0.0056) increase upon treatment and cathepsins A and B exhibited 5.9-fold (p=0.00039) and 4.9-fold (p=0.0064) increases respectively.

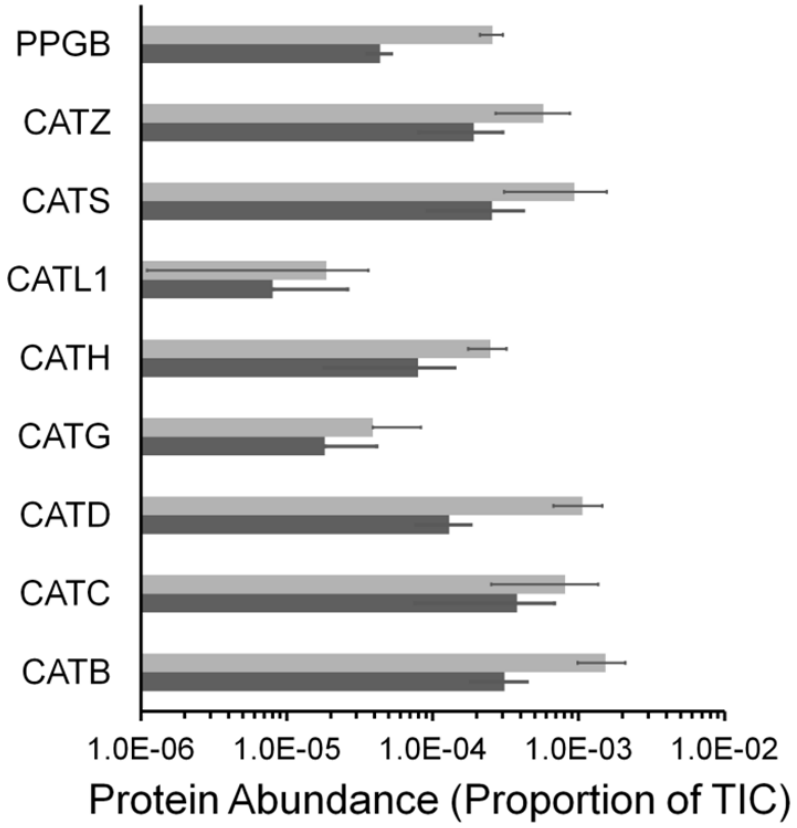


Figure 3.6 Cathepsin-family protease expression.

We looked at global cathepsin-family proteases observed within the treated (**light grey**) and untreated (**dark grey**) tumours. All cathepsin-family proteases were observed with a q-value ≤0.15, and displayed increased abundance in the treated dataset. On average, cathepsins had a ~3.9-fold increase in abundance in the chemotherapy-treated dataset compared to the untreated. The most significant increases were observed in cathepsin B (CATB, 4.9-fold increase, p=0.0064), cathepsin D (CATD, 8.2-fold increase, p=0.0056), and cathepsin A (PPGB, 5.9-fold increase, p=0.00039).

With the observed up regulation of cathepsins, there is the potential of interplay with caspase-dependent apoptosis as previous reports have demonstrated the cleavage and inhibition of XIAP by cathepsin B[310], in addition to its direct proteolytic processing of several caspases[311]. Nonetheless, our data does not reveal detectable active caspase-3 or caspase-6 despite cathepsin upregulation, which may be due the metabolic instability of caspase-3[306] as previously mentioned. Alternatively, the lysosomal membranes may be sufficiently intact to prevent cathepsin leakage and subsequent apoptotic activation[312]; this idea fits with a model where the global upregulation of lysosomal proteins in the bulk tumour is occurring due to the increased phagolysosomal activity of cells clearing apoptotic bodies. Taken together, this upregulation of lysosomal proteins is suggestive of a potential role for lysosomes in the EL4 derived tumours in C57BL/6 mice in response to etoposide and cyclophosphamide treatment. In the context of the growing understanding of autophagy in the cell survival/cell death axis of cellular fates[313,314], our data reveals the potential utility of this model system to investigate novel imaging applications focused on this biological process.

3.4. Conclusions

As our understanding of the complexities of cell death increases beyond proteolytic cascade activation to include the interplay with cellular processes such as autophagy[314] and changes to the complex network of protein-protein interactions[315], quantitative whole-system data of model systems is needed to further investigate both the complexities of these processes. Furthermore, additional evaluation of model systems' appropriateness in the context for which they are studied should be considered in the future. Here we report our findings on the proteome analysis of the murine EL4 tumour model in response to high concentration cyclophosphamide-etoposide treatment.

For tumours comprised of various cell types, our reported robust dataset reveals a large number of changes at the global protein level, with varying degrees of statistical confidence that can be queried for any protein of interest. This is supportive of the technique's applicability in the comparative analysis and characterization of other solid tissues and tumour types. The utility of this dataset is demonstrated with our initial analysis of the data revealing potential novel roles for caspase 6 and lysosomes in the response of these tumours to chemotherapeutic drug treatment.

Chapter 4 : Characterization of Proteomic Signatures in Human Estrogen-Receptor Positive Breast Cancer Tumours

4.0. Proem

In the previous two chapters, we explored the applications of label-free, relativistic quantitation, mass spectrometry-based comparative proteomics on the proteomes of *i*) tissue-derived organelles, and *ii*) entire 'bulk' tumours. In both applications, we showcased the ability to resolve a high degree of temporal and/or conditional proteome changes, leading to novel insights into the mechanistic processes occurring in response to the specified condition.

This application is particularly suited towards clinical studies; with an interest in cancer pathogenesis, specifically disease prognosis, we decided to attempt to identify proteomic signatures – biomarkers – unique to specific breast cancer subtypes (estrogen receptor positive Luminal A/B tumours). These tumours are classically difficult to distinguish from one another and patients have drastically different prognoses. Identification of even a handful of proteins capable of distinguishing these tumour types could provide clinicians with an invaluable ability to determine the most effective treatment regimen at the time of diagnosis.

Supplementary data for this chapter can be found at:

Supplemental Tables

4.0.1. Acknowledgements

The work in the following chapter was only made possible with the aid of Drs. Judith Hugh and Sambasivarao Damaraju of the Department of Laboratory Medicine and Pathology at the University of Alberta. Additionally, if it were not for Daryl Glubrecht of Dr. Roseline Godbout's lab, I would not have been able to complete the immunohistochemistry presented. Furthermore, the visual scoring of stained tissue-microarray slides performed by Dr. Wei-Feng Dong of the Cross Cancer Institute was invaluable in reaching some of our conclusions.

4.1. Introduction

In Canadian women, breast cancers account for approximately 26% of new cancer diagnoses and 13% of cancer-related deaths each year[316]. While breast cancer survivorship has increased over the past several decades with the onset of better treatment and detection strategies, there is still a crucial need for better indicators of patient prognosis at the time of diagnosis[317,318]. Breast cancers are currently assessed and treated based on their 'receptor status' - these being the presence of estrogen and/or progesterone receptors (ER+/PgR+), the HER2 receptor (HER2+), or the absence of all three (TN)[318,319].

ER+ breast cancers are the most common grouping, accounting for nearly 75% of all breast cancers, and can be further divided into Luminal A and B subtypes. Luminal A breast cancers account for half of all breast cancers diagnosed and have the best prognosis; these tumours have high expression of ER, and relatively low amounts of Ki-67 protein[320,321], a marker for cellular proliferation. Luminal A patients typically present later in life with well-defined and localized tumours that tend to respond well to ER-targeting anti-hormone therapies such as tamoxifen, resulting in high rates of survival with low rates of disease recurrence [319–322]. Conversely, Luminal B ER+ breast cancer patients present at a young age, often with large, poorly defined tumours and are node-positive; these tumours tend to be associated with lower ER abundance and a high abundance of Ki-67 protein, in addition to the possibility of being HER2 receptor positive. This makes Luminal B tumours more resistant to anti-hormone therapies, and treatment often requires systemic adjuvant chemotherapy [319–322]. Currently, distinguishing Luminal A from HER2-negative Luminal B in the clinical setting is challenging[317,318,320]. As the first-line treatment for all ER+ breast cancer patients is tamoxifen (adjuvant chemotherapy is only warranted if the patient's is HER2+ or node-positive), Luminal B patients tend to have higher rates of recurrence and lower rates of survival compared to Luminal A patients[318].

Aside from patients' node status at the time of diagnosis, few screening tools can accurately predict the likelihood a tumour is either Luminal A or B. The current gold-standard, Oncotype DX, screens tumours by assaying 21 genes of interest but is expensive, time-intensive, and not at all definitive [318,323,324]. Additionally, a recent review[318] of clinical practices of the management of

early-stage breast cancers highlighted a need for better diagnostic differentiation techniques for Luminal-type breast cancers, specifically with respect to disease prognosis and chemotherapeutic effectiveness.

With the advent of mass spectrometry-based proteomics, we are capable of identifying and quantitating thousands of proteins from a single tissue sample. Characterization of the proteomic profiles of patient-derived tumours, whose long-term outcomes are known, has been shown to identify biological markers (biomarkers) that correlate with patients' disease prognosis and therefore outcome.

4.2. Experimental Procedures

4.2.1. Tumour Selection

Flash-frozen human breast tumours from the Alberta Cancer Research Biobank (ACRB) repository, with corresponding de-identified patient data were obtained with sponsorship from the Alberta Cancer Foundation (ACF), with approval from the University of Alberta Health Research Ethics Board – Biomedical Panel ("Alberta Cancer Proteome Platform", ID# Pro-00049009, 08/26/2014) by Dr. Judith Hugh, of the Department of Laboratory Medicine & Pathology at the University of Alberta. Nineteen tumours were selected based on their pre-treatment pathological classifications of being primary invasive ER+, HER2-negative tumours, with patient follow-up data available to approximately 80 months. An approximately equal proportion of Luminal A (10) and Luminal B (9) classified tumours were included.

4.2.2. Tumour Preparation for Mass Spectrometry

The frozen tumour's masses were measured, and subject to the homogenization protocol using the buffers and procedures as previously described[177] in **Section 3.2.2.**

4.2.2.1. Gel Electrophoresis and In-Gel Protein Digestion

Tumour homogenates (20µL) were resolved using 1-dimensional SDS-PAGE in 1.5mm thick 10% polyacrylamide gels. Gels were visualized using colloidal Coomassie blue (G-250) stain, and whole lanes were excised, sectioned, and subject to in-gel trypsin digestion as described in **Section 3.2.3.** Resultant extracted peptides were dried and resuspended in 60µL 0.2% formic acid in 5% acetonitrile.

4.2.2.2. Mass Spectrometry and Database Search Parameters

Digested peptides were subject to LC-MS/MS analysis utilizing the protocol outlined in **Section 3.2.2.**, albeit shortening most LC gradients to 75 minutes, to accommodate data collection.

Following data collection, raw mass spectra files corresponding to samples comprising an entire gel lane were grouped and searched together using ProteomeDiscoverer 1.4.1.14's SEQUEST search algorithm. Data was searched using a reviewed, non-redundant *Homo sapiens* proteome retrieved from UniprotKB in February 2015. The search parameters and tolerances used for protein identification were identical to those outlined in **Section 3.2.2.3.** Search results were filtered using a minimum of 2 medium-confidence peptides per protein. Additionally, distinct protein isoforms from a single gene product were merged. Protein quantitation was determined using the 'Precursor Ion Area Detector' node in SEQUEST, quantitating only the top 3 most abundant peptides passing the filter parameters.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE repository[233,234] with the dataset identifier PXD009827.

4.2.2.3. Statistics and Data Analysis

Peptide FDRs for the searched samples were as follows; actual relaxed FDRs ranged from 0.032 to 0.0451, while actual strict FDRs ranged from 0.072 to 0.093. Protein lists were exported to Microsoft Excel. Protein abundance was determined by adjusting each protein's reported EIC relative to a correction factor calculated per sample via histone H4 abundance. This was achieved by finding the average reported EIC for histone H4 and determining a correction factor for each sample to make the measured H4 abundances equal across all samples. Based on peptide coverages per sample, in addition to their observed proportions, missing values (MVs) were determined to be missing not at random (MNAR)[325,326], and the global minimum adjusted-EIC for the entire dataset (half of the limit of detection; LOD) was imputed[325]. Corrected EICs for each protein were then normalized using a $\log_2$ transformation[156,157].

For initial data comparison, tumours were grouped based on patients' disease recurrence status. Data was refined by only including proteins identified in n=3 samples of at least one of the patient groupings. Following data refinement, p-values were generated by applying a two-tailed, heteroscedastic *t*-test (Welch's *t*-test)[188] to protein abundances observed in each grouping. Sorted

p-values were uploaded to the 'q-value estimation for FDR control' web utility (qvalue.princeton.edu)[200–202,295] for the generation of q-values.

Following cluster analysis (see **Section 4.2.7.**), data was further refined, using the same threshold criteria as described for two-group analysis. Additionally, differential protein expression was determined using a one-way analysis of variance (ANOVA), with p-values being controlled through q-value estimation[200–202,295]. Following ANOVA analysis, proteins with a $q<0.05$ were subject to a *post-hoc* Tukey-Kramer Honest Significant Difference (TK-HSD) test to determine between which pairs the differences lie[327–329]. TK-HSD tests were performed courtesy of Dr. John Maringa using MATLAB R2017b.

4.2.3. Functional Analysis of Proteins

Proteins determined to have increased abundance with respect to a tumour grouping were subject to functional analysis using both STRING[212] v.10.5 and PANTHER[209–211] v.13.1 web-utilities. Protein lists determined to be specific to a tumour group were subject to GO-molecular function (GO-MF) and -biological process (GO-BP), in addition to pathway (KEGG) and protein class (PANTHER) analyses. Statistical overrepresentation using was determined utilizing 'slim' annotation sets with Fisher's Exact test, reporting only identifiers with an $FDR<0.05$.

4.2.4. Hierarchical Clustering and Principal Component Analysis of Tumours

Hierarchical clustering of tumour samples was performed using the software PermutMatrix v1.9.4[330]. Treatment options were set as follows; dissimilarity was measured using Euclidean distance, cluster linkage was determined through 'unweighted pair-group method using arithmetic averages' (UPGMA; average-linkage method), rows' values (individual protein's abundances) were normalized as z-scores, and tree-seriation was determined using a multiple-fragment heuristic method. Principal component analysis (PCA) was performed using Perseus[331] v1.6.0.3. on z-score normalized $\log_2$-transformed data.

4.2.5. Immunoblotting

Equivalent volumes of seven whole tumour homogenates were loaded onto 10% polyacrylamide gels and subject to SDS-PAGE. Samples were transferred to 0.22µm nitrocellulose membranes (LICOR) and blocked using 3% skim milk in PBS. Blocked membranes were incubated with biotinylated goat anti-human PIGR (1:1000; R&D Systems #BAF2717) in 3% skim milk in PBST overnight at 4°C. Membranes were subsequently incubated with IRDye-680-labelled streptavidin (1:2000; LICOR #925-68079) in 3% skim milk in PBST for 1 hour at room temperature and visualized using a LICOR Odyssey Fc Imager's 700nm channel. To determine non-specific binding of labelled streptavidin, a second blot was performed without the use of primary antibody.

4.2.6. Immunohistochemistry, Scoring, and Survival Analysis

Breast tumour tissue microarrays (TMAs) were obtained from Dr. Sambasivarao Damaraju, of the Department of Laboratory Medicine & Pathology at the University of Alberta. De-identified patient data corresponding to the TMAs were obtained with permission from the University of Alberta Health Research Ethics Board – Biomedical Panel ("Alberta Cancer Proteome Platform", ID# Pro-00049009, 08/26/2014). Slides obtained had been microtome sectioned 4µm thick. Slides were baked at 60°C for 1 hour to remove paraffin, and subsequently rehydrated: 3 × 10-minute incubations in 100% xylene, 4 × ~30s washes in 100% EtOH, then ~30s washes in 80% EtOH, 50% EtOH, and deionized water. Rehydrated samples were subject to antigen retrieval by incubation in 600mL boiling 0.05% citraconic anhydride (pH=7.50) for 6 minutes in a pressure-cooker. Slides were cooled, and subsequently rinsed in TBS.

Endogenous peroxidase activity was degraded by incubating tissue with 50µL of 3% H_2O_2 for 15 minutes. Endogenous biotin was blocked by incubating each slide with 50µL of 0.001% unconjugated streptavidin (Sigma) in TBS for 15 minutes at 37°C with gentle agitation. Slides were rinsed 3 × in TBS, subsequently incubated with 50µL of 0.005% biotin (Sigma) in TBS for 15 minutes at 37°C with gentle agitation and rinsed an additional 3 × in TBS. Tissue was blocked with blocking buffer (1% BSA, 0.5% FSG, 0.5% triton X-100, 0.05% sodium azide in TBS) for 30

minutes at room temperature with gentle agitation. Slides were incubated with 50µL of primary antibody (biotinylated goat anti-human PIGR, R&D Systems #BAF2717, 1:350 in blocking buffer) at 4°C, washed 3 × 10-minutes in TBST, and incubated with 50µL of 2µg/mL of streptavidin-HRP conjugate (Jackson Immunoresearch #016-030-084) in azide-free blocking buffer for 30 minutes at room temperature with gentle agitation. Slides were then rinsed 3 × 5-minutes in TBST. Colourization was achieved by adding 100µL of 3,3'-diaminobenzidine (DAB; DAKO #K3467) and allowing colour development for 6 minutes. Following colourization quenching in TBS, slides were incubated in 1% $CuSO_4$ for 5 minutes, rinsed in deionized water, and subsequently incubated in hematoxylin stain for 15 seconds. Slides were then rinsed with water and incubated in saturated Li_2CO_3 for 3 minutes. Slides were subsequently dehydrated (rehydration procedure above, in reverse-order), and allowed to dry prior to addition of coverslips and visualization. As a negative control, the above procedure was repeated, less incubation with primary antibody, on extra TMA slides.

Stained slides were scanned using an Aperio ScanScope CS slide scanner. Digital files were visualized and scored using Aperio ImageScope v12.3.2.8013. Scoring was achieved through visual assessment of diseased-tissue staining by Dr. Wei-Feng Dong (Pathology), using a 0 to 3+ classification system. Additionally, entire tissue staining was assessed by manually annotating tissue sections and determining pixel-density analysis using Aperio's Positive Pixel Count v9 algorithm (default settings). Following scoring, tissue sections corresponding to a single patient's disease were averaged and correlated with clinical information supplied to us by Dr. Damaraju. Survival analysis was performed with NCSS 12 statistical software, using only patient data corresponding to those with ER-positive tumours who did not receive therapy prior to surgery/biopsy. Receiver operating characteristic (ROC) curve analysis was used to determine 'low' and 'high' thresholds for PIGR scores; cut-off points were chosen by minimizing the distance to the corner at which sensitivity and specificity = 1 on the ROC curve which maximized the AUC. Using these thresholds, Kaplan-Meier curves were generated, assessing disease-free survival with clinical endpoints corresponding to 60 and 120 months post-surgery/biopsy.

4.2.7. mRNA Survival Analysis

Genes corresponding to proteins of interest were investigated with the KMplotter[332,333] web-utility (kmplot.com/analysis/) for assessing gene expression microarray data and disease survivorship. Probe sets were limited to the JetSet[334] best probe set, while only assessing expression in patients with ER+, HER2- breast cancers. For quality control, redundant samples were removed from analysis, and biased arrays excluded.

4.3. Results

4.3.1. Experimental Design Rationale

In the previous two chapters, the proteomes of whole-tissue derived organelles (lipid droplets)[24] or tumours (EL4-lymphomas)[177] were analyzed for changes following exposure to an experimental condition. For these types of analyses, the experimental design is rather straight-forward; a condition which is hypothesized to elicit a biological change in a specified organelle/cell-type/tissue is chosen, and the specified biological samples are exposed to the condition. Through the utilization of tightly controlled biological replicates – as is the case when dealing with tissue culture or animal work – an average proteomic profile per condition can be determined, and statistical analyses can be carried out for proteins' abundances between experimental conditions[335]. However, when attempting to apply this experimental technique to clinical samples, several problems emerge, precisely because clinical samples often originate under vastly different conditions from vastly different individuals[336,337]. Thus, replicates of clinical origin are difficult to classify as biological replicates, as defined by the replication which can be achieved in a laboratory.

4.3.1.1. Grouping of Patient Data

Exercising tight control over biological replicates makes the life of a typical scientist immensely easier than it could be otherwise[335]. Between-sample variability, if samples are obtained at random from a population, can be widely diverse. One of the caveats of complex system analysis is that, to fully understand a given biological system, the variables at play within the system must be fully understood. The use of such biological replicates circumvents this issue; by utilizing samples from sources that are (or nearly are) identical, between sample variability is minimized. By applying this methodology to groups of samples that are subject to differing – though specific and highly-controlled – experimental conditions, scientists can infer causality between experimental conditions and observed outcomes[336,337].

In the case of our 19 patient-derived ER+ breast cancer tumour samples, we tried to control these as much as possible with respect to tumour biology; Dr. Judith Hugh, a clinical pathologist at the University of Alberta, selected tumours based on their overall tumour grade (accounting for size, mitotic activity, nuclear dysmorphia, receptor status, adjuvant and neoadjuvant treatments, *et cetera*), providing us with a relatively equal proportion of pathologist-diagnosed Luminal A and Luminal B tumour samples.

However, even with careful selection of tumour specimens for inclusion in the study, it has been shown pathological diagnoses of ER+ tumours are not always accurate[318,338]. In a recent study[338], it was demonstrated that the most definitive way to determine the Luminal A vs B subtyping for an ER+ tumour was to subject the patient to anti-estrogen chemotherapy, and observe the tumour response. Tumours that were determined to be non-responsive over the course of a 4-month treatment were declared to have Luminal B subtypes, while patients who responded quickly, or slowly (with eventual tumour volume reduction) have Luminal A. When comparing clinical outcomes to the pre-treatment clinical diagnoses, pathological classification of a tumour as Luminal B was shown to correlate with only ~30% of patients. In agreement with these general trends, of the 19 patients in our study, 10 were diagnosed as Luminal A status, and 9 as Luminal B. When judging 'true' Luminal status by disease outcome, a Luminal A classification was 90% predictive of disease-free survival, while Luminal B classification was ~67% predictive of disease-recurrence. While clearly indicative of a requirement for better classifiers, these data also provided us with some rationale for our experimental design; rather than rely on pre-existing definitions of Luminal subtypes for sample grouping and comparison, tumours in our cohorts can be grouped based on their respective patients' clinical outcomes as these are known.

Comparing the proteomes of pre-treatment-defined Luminal A and Luminal B tumours, both of which possess patients who experienced disease-free and disease-recurrent survival, confounds our ability to reach meaningful biological conclusions; some differences may be a result of the subtyping (which may not even be correct in some cases), with others a result of the patients' recurrence-status. By grouping tumours solely on the patients' clinical outcomes, we mitigate some of the variability between sample populations.

4.3.1.2. Data Refinement and Normalization

However, more convenient groupings do not negate the fact that these samples have high variability. The number of biological variables at play – including a patient's age, weight, race, reproductive status, previous diagnoses, family history, presence of genetic mutation or viral infection, activity level, diet, *et cetera* – and how these can affect protein expression and cellular/protein architecture in a localized tumour[25], makes the search for meaningful 'differences' between sample groups incredibly difficult. Compounding this issue are technical variables such as ion suppression, the instrument's limit of detection, and detector signal saturation limiting the number of identifiable ion species (for reviews see [339–341]). Therefore, to minimize the effect such variables can have during analysis, it is necessary to both normalize the protein abundances being compared, and refine the dataset to remove biased comparisons[156,157]. To illustrate this point, the p-value distributions from an initial set of comparative analyses using unrefined data are depicted in the left panel of **Figure 4.1**. Regardless of the normalization procedure used, the p-value distributions are indicative of biased/noisy data; a relatively low amount of p-values is observed at the lower boundary of the [0,1) range, with the highest density incredibly focused within the 0.33-0.36 range. Such distributions are indicative of low statistical power (a high proportion of false-negatives), which can be the result of: poor sampling size (low *n*); the critical value used as an indicator of significance (α); high variability between measurements within and between populations (large standard deviations, i.e. noisy data); and a generally minimal difference between the observed means for the two populations being compared[342]. Because our study is limited with respect to sample sizes, methods to increase the statistical power of our study are limited to data normalization and refinement as previously mentioned.

With respect to data refinement, we opted to include *only* proteins with an EIC>0 in *at least* n=3 samples for *at least* a single patient grouping; this removes noisy data (i.e. proteins identified in the dataset with measurable abundance in only a select few samples), while increasing the likelihood that population means being compared will be different. The effect of this refinement can be observed in the right panel of **Figure 4.1A**. Following our method of data refinement, the

biased region populating the 0.33-0.36 interval is markedly reduced, allowing a truer representation of the dataset's statistical properties to be observed.

Additionally, we required a method for data normalization that was not reliant on the EIC-to-TIC ratio. To circumvent this issue, we opted to normalize each protein identified within an individual tumour's proteome to a reference protein, with constant cellular abundance, ubiquitously identified in all tumours. The reference protein we selected was Histone H4, based loosely on the rationale behind Matthias Mann's 'proteomic ruler' methodology[343]; for any given cell or tissue, the proportion of DNA to histones should be relatively constant. Conveniently, even though cancer is prone to aneuploidy (aberrant amounts of nuclear chromatin), ER+ breast cancers (especially those which are ER+/PgR+; 16/19 tumours analyzed) rarely exhibit such nuclear dysmorphia[344–346]. Additionally, while all histones are well-conserved in their sequences, histone H4 is remarkably so, with more than 95% of various evolutionary sequences conserved[347]. Considering this, in addition to its consistently high ion abundance (within the top 1% of all EICs – **Supplemental Table 4.1**) across all analyzed ER+ tumour samples, histone H4 makes an excellent reference protein for sample standardization and normalization.

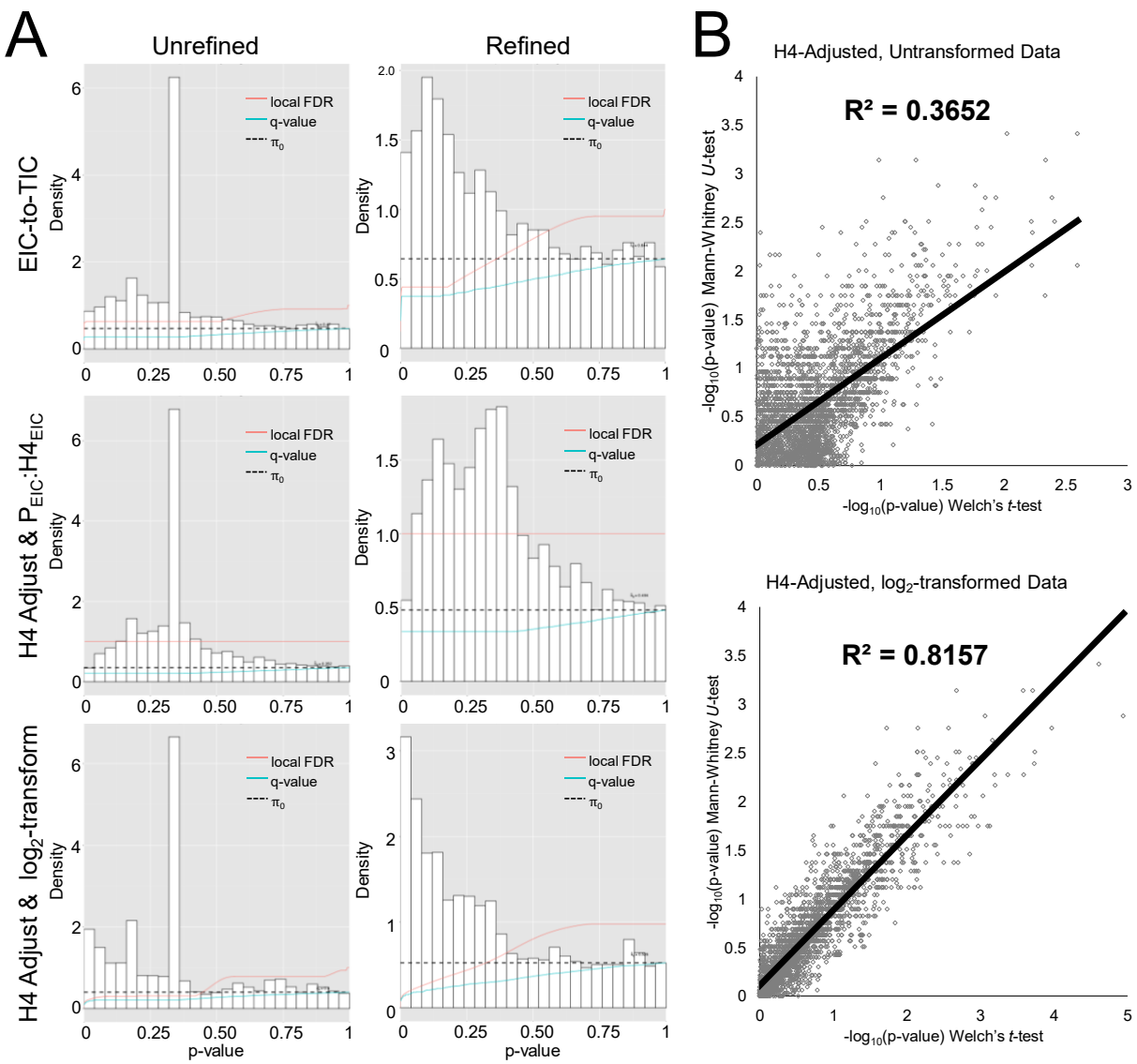


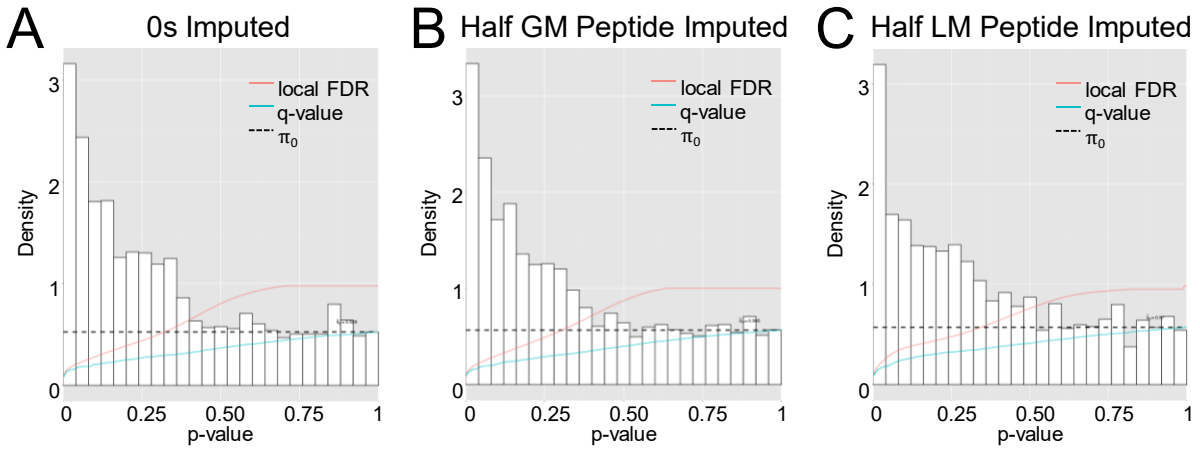
Figure 4.1 Selection of optimal data refinement and normalization treatments for clinical samples.

(**A**) P-value histograms corresponding to unrefined (**left**) and refined (n≥3 samples in ≥1 group; **right**) in addition to protein abundance correction methods. Data refinement resulted in removal of biased samples populating the (0.33-0.36) interval. Due to different amounts of total protein between tumours, EIC-to-TIC normalization (**top**) yielded poor results as local sample TICs were subject to variation. Protein correction to histone H4 (**middle**) yielded similar results, indicating non-normal data distribution. Log_2-tranformations of H4-corrected data (**bottom**) resulted in normalization of data. (**B**) Correlation of parametric (Welch's *t*-test) and non-parametric (Mann-Whitney *U*-test) test results of H4-adjusted non-log_2-transformed (**top**) and H4-adjusted log_2-transformed (**bottom**) datasets. For normally-distributed data, the Mann-Whitney *U*-test performs similarly to the Welch's *t*-test, as indicated by a larger R^2 value

Two methods of normalization utilizing histone H4 were attempted (**Figure 4.1A**). These included normalizing individual proteins' abundances to H4's abundance within a sample (P_{EIC}:$H4_{EIC}$), or adjusting ***all*** EICs within a sample to equate H4 EICs between samples (i.e. $H4_{sample1}$ = $H4_{sample2}$ = $H4_{sample3}$...). While both techniques are mathematically similar, the latter method allows for easier subsequent transformations if required; when using the latter method, the magnitude of measured ion abundances ranges by several orders of magnitude, having distributions which do not conform to that of Gaussian normal. As a result, -omics datasets are often log_2 transformed prior to statistical analysis[156,157]. Due to low sampling (i.e. a maximum of 12 observations per grouping) preventing us from determining the normality of untransformed versus log_2-transformed data, normality was assessed based on the correlation of different statistical tests' outcomes. The statistical tests whose outcomes were correlated were the parametric Welch's *t*-test (an unpaired, heteroscedastic *t*-test)[188], and the non-parametric Mann-Whitney *U*-(MWU) test[182]; parametric tests function under assumptions made about the sampling populations fitting a normal distribution, while non-parametric tests do not. While possessing less statistical power than its parametric counterpart, the MWU test is known to perform similarly to Welch's test if the data fits a normal distribution[348]. **Figure 4.1B** shows the correlation of p-values resulting from MWU and Welch's tests performed on untransformed and log_2-transformed datasets.

As illustrated, global adjustment of proteins' abundances via equating histone H4 followed by log_2-transformation proved to be the most desirable treatment. However, even with a preferred method of data normalization, one of the remaining issues was how to deal with MVs in the dataset[157,349,350]. As the dataset is comprised of clinical replicates, the problem of MVs becomes a complicated one; without experimentally-controlled replicates from which an individual's expected abundance can be observed, we are unable to determine whether a MV is due to an absolute absence from the dataset or below the LOD (missing not at random; MNAR) versus being present but missing due to stochastic variation (missing completely at random; MCAR)[157,326,351].

Some studies have chosen to avoid this issue by performing extreme data refinement procedures, often removing proteins which do not have a determinable abundance in at least 90% of the dataset[352,353]. However, if such procedures were to be performed on our dataset, we would lose a substantial amount of valuable information; when searching for biomarkers, it becomes desirable to identify proteins whose expression – or lack thereof – correlates with disease outcome. The removal of proteins with missing values in a sample grouping would leave only proteins whose expression *differs* between disease outcomes; while still valuable, such differences become incredibly hard to discern when being screened for using traditional techniques such as immunohistochemistry.



D

Significance Threshold	0s Imputed	Half Global Min. Imputed	Half Local Min. Imputed
p≤0.05	387	408	407
p≤0.01	110	115	119
q≤0.10	47	46	52
local-FDR≤0.10	39	33	30

Figure 4.2 Comparison of data imputation methods for missing values.

Due to the number of MVs present within sample groups, MVs were determined to be MNAR, and various values were imputed for MVs to approximate the limit of detection. (**A-C**) P-value histograms correlating to statistical outcomes following MVs imputed as (**A**) zeroes, (**B**) half of the global minimum peptide abundance, or (**C**) half of the sample-specific (local) minimum peptide abundance. (**D**) A summary of various FDR-approximations resulting from different methods of data imputation. Half GM peptide imputation was chosen based on the lowest number of proteins with q≤0.05.

Due to the proportion of MVs in our dataset – an average of ~46% of values missing in disease-free tumours and ~28% missing in disease-recurrent tumours - missing values were treated as MNAR. To determine how best to deal with these missing values, in addition to assigning missing values as zeros (as in **Figure 4.1**), two common methods of data imputation were compared[326,351] as illustrated in **Figure 4.2**; half of the adjusted global minimum peptide EIC (half of the smallest observed adjusted-EIC for an individual peptide in the entire dataset), or half of the adjusted local minimum peptide EIC (half of the smallest observed adjusted-EIC for an individual peptide in the corresponding sample). **Figure 4.2D** summarizes the outcomes of each MV imputation method.

4.3.2. General ER+ Tumour Proteome Analysis

Our proteomic analysis of 19 human-derived ER+ breast tumours identified a total of 4477 proteins across all samples. Placement of tumours into Disease-Free (*DF*; n=12) and Disease-Recurrent (*DR*; n=7) groups revealed 3539 and 4020 proteins observed in DF and DR tumours, respectively. However, following data refinement and correction of protein abundance as described above, the total number of identified proteins in the dataset was reduced to 2729. In tumours originating from disease-free survival patients, 2633 proteins were observed, of which 32 were unique to this patient grouping. In tumours originating from patients experiencing disease-recurrence, 2697 proteins were observed, 96 of which were unique.

Following data refinement and missing value imputation, application of a two-tailed Welch's *t*-test between our defined groupings with subsequent q-value calculation revealed 46 proteins to have a q-value at our designated significance threshold of $q \leq 0.10$. Interestingly, of these proteins selected, only 3 displayed an increased abundance in patients with *DF* survival: polymeric immunoglobulin receptor (PIGR), phytanoyl-CoA dioxygenase domain-containing protein 1 (PHYHD1), and tenascin-N (TENN), while the remaining 43 displayed an increased abundance in patients experiencing *DR*. **Supplemental Table 4.2** summarizes the statistical significance and $\log_2$-fold-changes observed for *DF* and *DR* comparisons.

To assess the biological function of proteins identified possessing both unique and statistically significant expression profiles in *DF* and *DR* tumours, protein lists (*DF*=33 proteins; *DR*=139 proteins; proteins in black boxes in **Figure 4.3** insert) were subject to gene ontology enrichment analysis using STRING v10.5 (**Supplemental Table 4.4**).

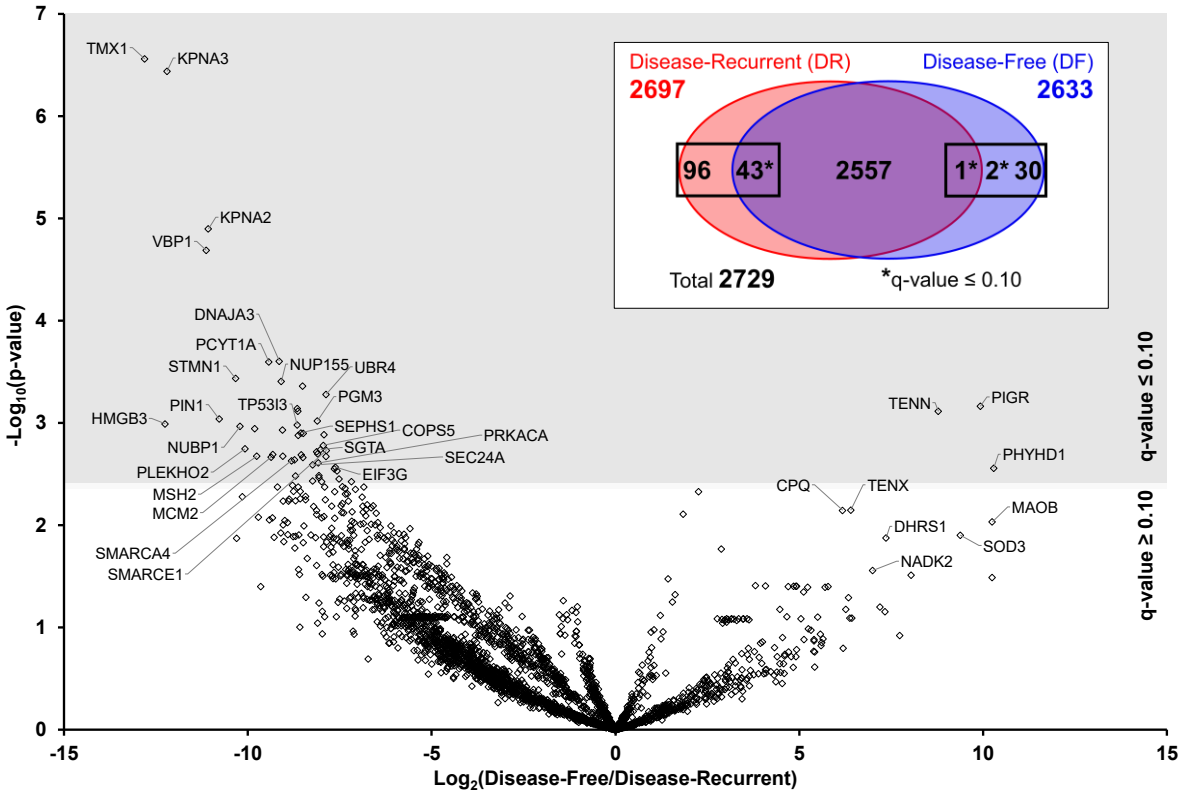


Figure 4.3 Volcano plot of disease-free *versus* disease-recurrent protein abundance.

The -$\log_{10}$ of p-values generated from Welch's *t*-test performed on protein abundances between disease-free (DF) and disease-recurrent (DR) tumours was plotted against the difference of the average $\log_2$ transformed DF and DR protein abundances. Additionally, a grey-scale threshold indicating the level of significance correlating to q≤0.10 is indicated. Clustering of data points is artefactual of proportions of missing values imputed in each sample grouping. (**Insert**) Venn diagram depicting total number of proteins identified among all samples and each grouping following data refinement. A total of 2697 proteins were identified in disease-recurrent (**red**) samples, of which 96 were uniquely observed, and an additional 43 significantly up-regulated. For disease-free (**blue**) samples, a total of 2633 proteins were identified, of which 32 were uniquely observed. Only 3 proteins were determined to be significantly upregulated in DF tumours, 2 of which were only observed in DF tumours.

Analysis of *DR*-abundant proteins revealed a plethora of processes to be enriched in these tumours – and therefore depleted in *DF* tumours – including those involved in protein biosynthesis, cell-cycle progression, and DNA mismatch repair. Analysis of *DF*-tumour specific proteins on the other hand revealed little in the way of metabolic processes. Rather, KEGG pathway analysis revealed an increase in proteins belonging to the extracellular matrix (ECM), involved in focal adhesion and ECM-receptor binding. To verify these results, protein lists were subject to statistical overrepresentation analysis using PANTHER[209–211] v.13.1. Unfortunately, *DF*-specific proteins yielded no significant enrichments using PANTHER. However, in agreement with our STRING findings, *DR*-specific proteins were found to be enriched for proteins involved in translational initiation and elongation, in addition to DNA mismatch repair (**Figure 4.4**).

To determine whether any correlational abundances existed for the 46 proteins identified with a $q \leq 0.10$, in addition to the effectiveness of our sample grouping (recurrence status), hierarchical clustering analysis (HCA) was performed. Interestingly, rather than the two expected groupings of DF and DR patients, three groupings emerged (**Figure 4.5**). *DR* tumours formed a single cluster, while DF tumours clustered into two distinct groups; 7 *DF* tumours formed a unique seriation (hereby denoted as DF_O), while the remaining 5 shared a seriation – and thereby more similarity – with the DR tumour cluster. Likewise, principle component analysis (PCA) on these 46 proteins illustrated the same pattern; 66% of the variance observed within our data was capable of being explained by two variables. plotting samples according to these two variables resulted in tumour samples grouping as observed in our HCA analysis (**Figure 4.5C)**. To distinguish this new cluster of DF tumours, we have applied the distinction 'disease-free, recurrent-like' (DF_{RL}).

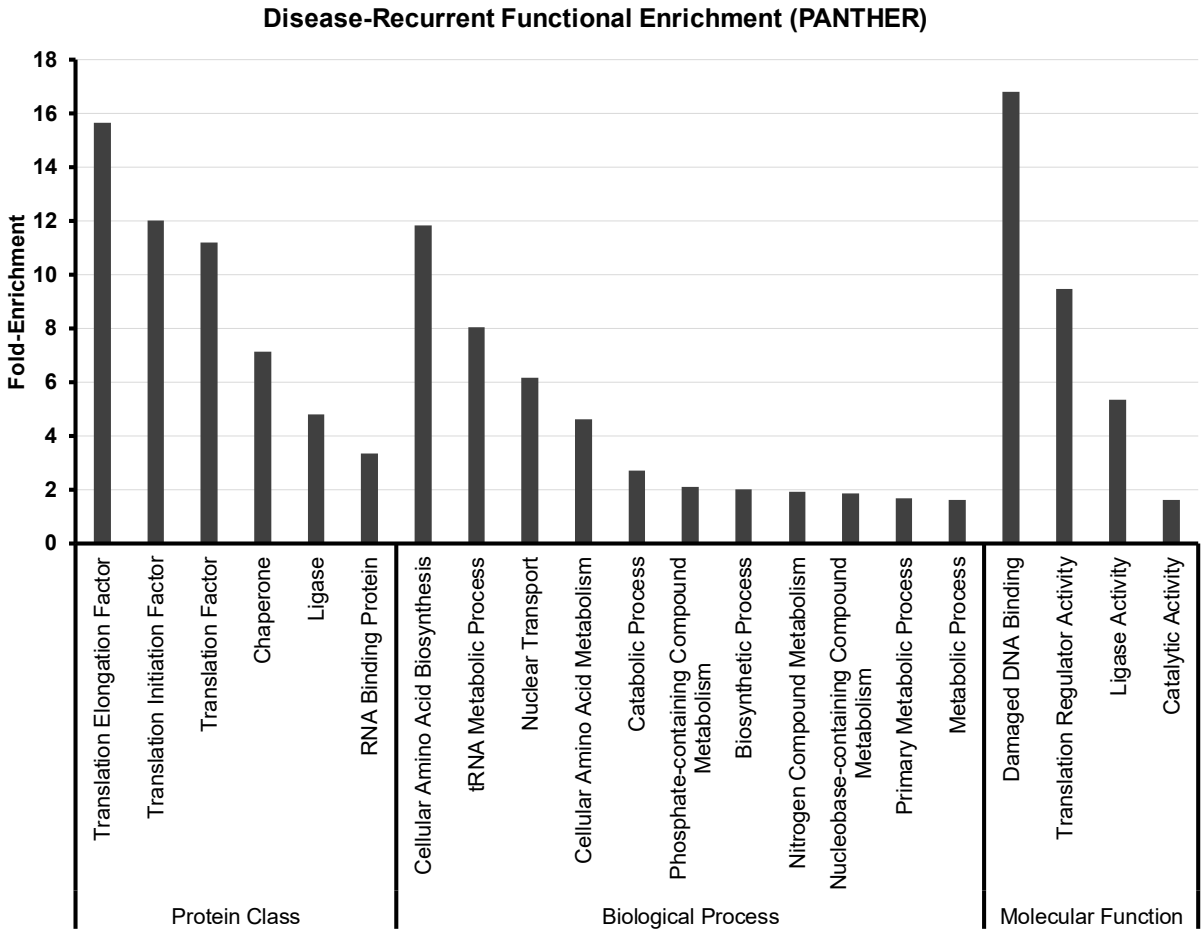


Figure 4.4 Overrepresentation analysis of DR-specific proteins.

Proteins determined to be significantly up-regulated (q≤0.10 or unique expression; proteins within left-sided black box of **Figure 4.3** insert) in DR tumours were subject to statistical over-representation analysis using PANTHER v.13.1 for identifiers for Protein Class (**left**), GO-Biological Process (**middle**), and GO-Molecular Function (**right**). Enriched proteins were primarily identified to be involved in protein biosynthesis and turnover, in addition to DNA damage repair.

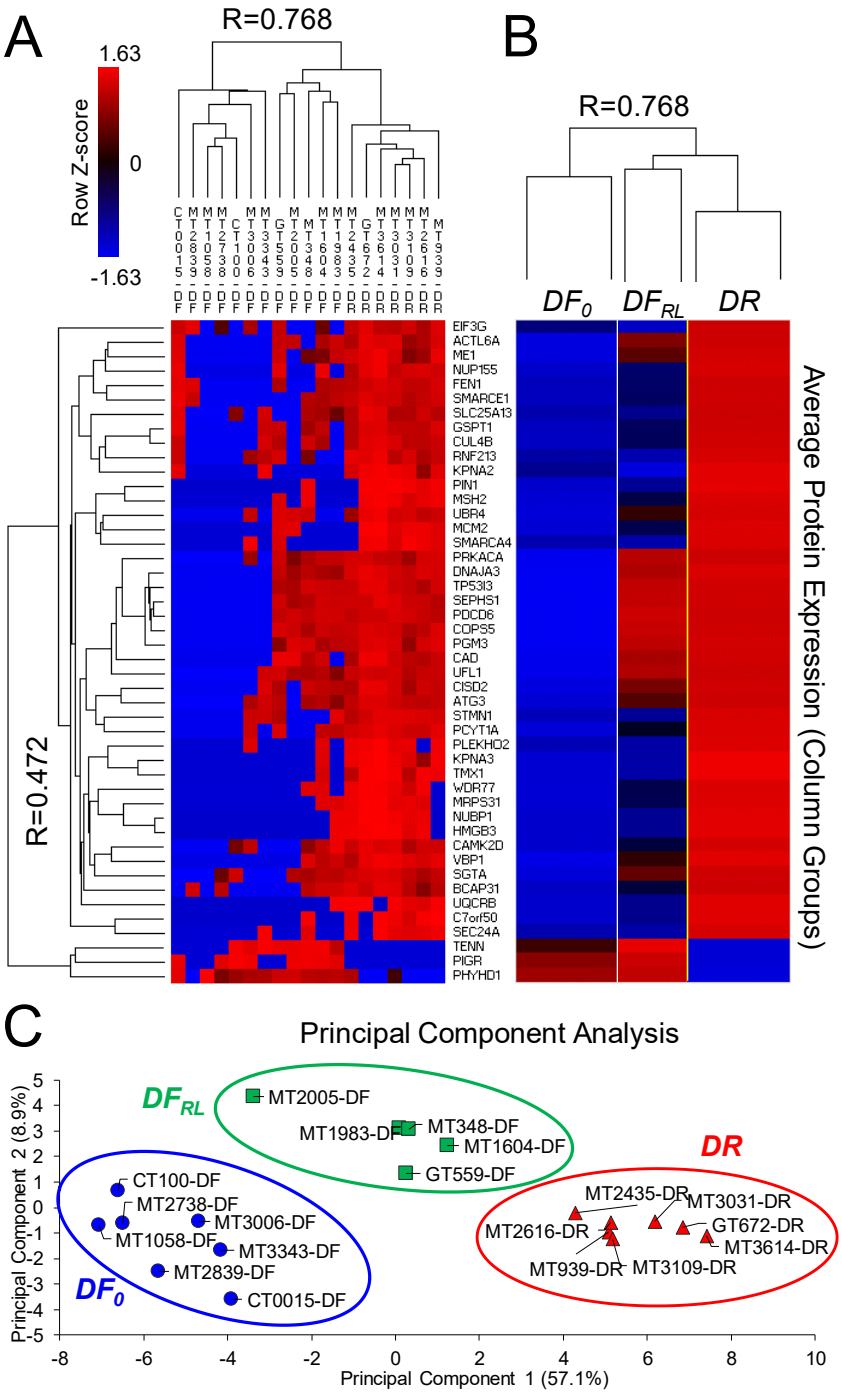


Figure 4.5 Hierarchical clustering and principal component analysis of q<0.10 proteins.

Proteins identified to have a q<0.10 following DF and DR tumour comparison were subject to hierarchical clustering (**A** and **B**) and principal component analysis (**C**). (**A**) HCA of the 46 q<0.10 proteins with (**B**) average cluster expression illustrating 3 clusters (*DF_0*, *DF_{RL}*, and *DR)*. (**C**) PCA determined 66% of the variability could be explained by 2 components. Plotting samples with these components produced the same clustering result observed via HCA.

4.3.3. Identification of Biological Differences Between DF_0, DF_{RL}, and *DR* Tumour Types

To assess the cluster-analysis defined groupings' likenesses and dissimilarities, our original dataset was re-defined; tumours were grouped based on clusters as defined by HCA/PCA analysis, and only proteins with a measurable abundance of n≥3 samples in ≥1 group were included, resulting in a dataset consisting of 2631 proteins. To identify proteins whose abundance differed between the newly defined groups, a one-way ANOVA was performed for each protein's abundance between all three sample groupings, with q-values being calculated to control the FWER. Using a refined significance threshold of q≤0.05 to control for type I error resulting from multiple ANOVA comparisons, 426 proteins were identified to have different means between *any* of the groupings. *Post-hoc* application of a TK-HSD test was used to determine between which groups the differences in protein abundance lie, as indicated via ANOVA testing.

To garner some insight into the biological differences between these three tumour groupings, these 426 proteins were queried for their relative abundance within each sample. Proteins with a mean positive z-score-normalized abundance in each tumour group were determined to have increased abundance, and lists of proteins determined to be increased with respect to a tumour grouping were subject to functional annotation analysis using STRING[212] v10.5 and PANTHER[210,211,209] v13.1 as described above. As expected, DF_0 tumours were characterized based on a general lack of protein abundance relative to DF_{RL} and *DR* tumours, with no significant enrichments identified via STRING or PANTHER; only 12 proteins of the original 426 had a positive z-score abundance in DF_0 tumours. For DF_{RL} and *DR* tumours, 329 and 377 proteins were identified to have a positive z-score abundance, respectively, of which 280 were common. Functional analysis via both web-utilities illustrated a high degree of protein turnover – specifically translation, proteolysis, and transport/localization – in both tumour types (**Figure 4.6**). Interestingly however, were the differences observed between the two groupings; DF_{RL} tumours displayed >2-fold increase in carbohydrate metabolism, while *DR* tumours displayed a >3-fold increase in fatty acid metabolism in addition to a nearly 6-fold increase in oxidative phosphorylation (**Figure 4.6A**). Additionally, DF_{RL} tumours showcased an increase in ECM-

receptor interactions and proteins involved in viral oncogenesis, while *DR* tumours displayed enrichment for DNA replication and mismatch repair (for full lists of STRING analysis with proteins, see **Supplemental Tables 4.5** for $\boldsymbol{DF_{RL}}$ and **4.6** for ***DR***).

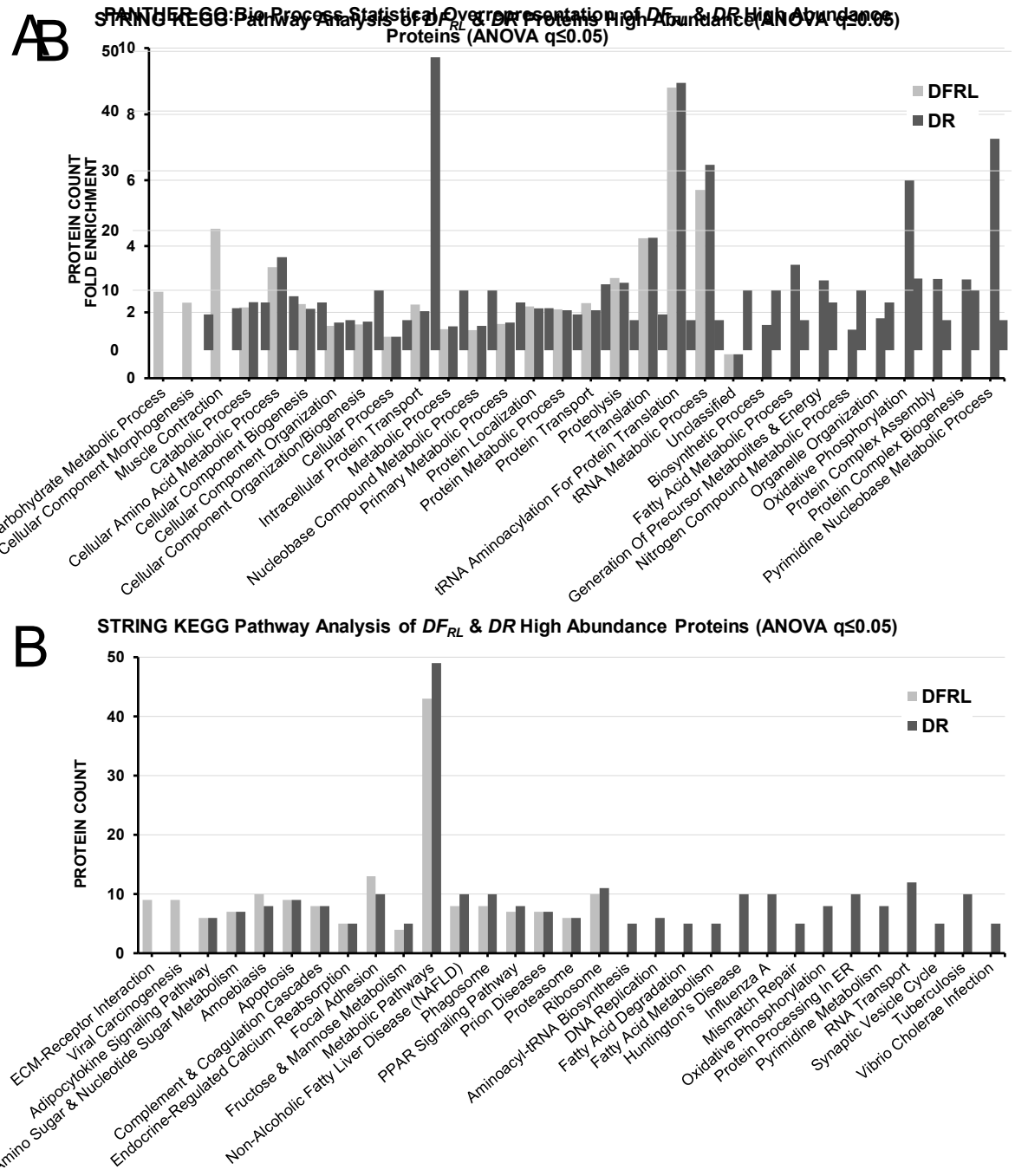


Figure 4.6 Functional enrichment and comparison of *DF*$_{RL}$ and *DR* tumour groups.

Proteins identified to have above average abundance in *DF*$_{RL}$ (**light grey**) and *DR* tumours (**dark grey**), with a q≤0.05 following a one-way ANOVA were subject to functional enrichment using PANTHER or STRING databases, with reported pathways limited to those with an FDR<0.05. (**A**) PANTHER statistical overrepresentation analysis for GO-BP identifiers (**B**) STRING functional annotation analysis for KEGG pathway identifiers.

4.3.4. Selection of Proteins Suitable as Biomarkers

While the 426 proteins identified to be significantly different between the three tumour groups are informative from a biological standpoint, when considered for diagnostic, predictive, or prognostic purposes, it would be more desirable for this list to be truncated. Protein-based clinical assays are still largely histological[354]; testing for large numbers of proteins requires both large amounts of tissue and time for scoring and interpretation. Therefore, the most idyllic protein-based assays can communicate large amounts of information regarding a patient's disease with reliance on only a select few biomarkers.

Desiring biomarkers that are both diagnostic and prognostic, our list of 426 proteins differentially abundant between tumour groups was significantly refined. Proteins' significance from pairwise comparisons were limited to those with a TK-HSD result of $p<0.05$. However, because each pairwise comparison yielded different levels of overall statistical significance for protein abundance between tumour groupings, our selection was limited to proteins observed to be the most statistically significant and differentially abundant for each pairwise comparison, as determined through visual assessment of these data's graphical representation (**Figure 4.7A)**. Following selection of proteins meeting the above criteria, the list was manually curated; immunoglobulin kappa variable 3-20 (IGKV3-20), a sequence-variable protein responsible for immunoglobulins' ability to bind various antigens (for reviews regarding antibody maturation, see [355,356]), was removed due to its poor suitability as biomarker which could easily be screened for. Utilizing this strategy, a minimalist list of 20 proteins was generated, possessing high correlation between both measured protein abundance ($R=0.804$) and tumour groupings ($R=0.785$), capable of discerning DF_O, DF_{RL}, and *DR* tumour groupings from each other if measuring a few select markers' abundances (**Figure 4.7B-D)**.

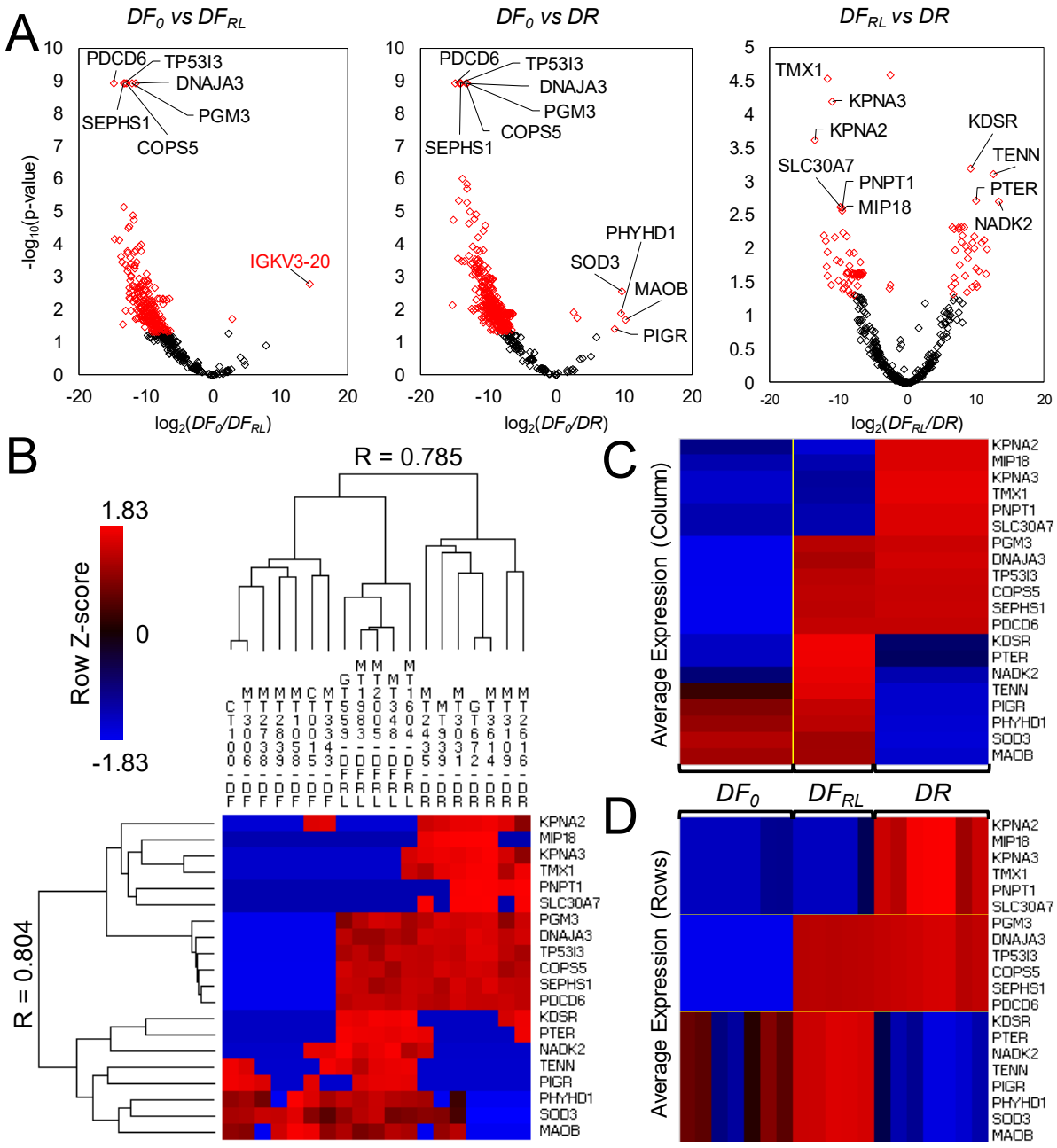


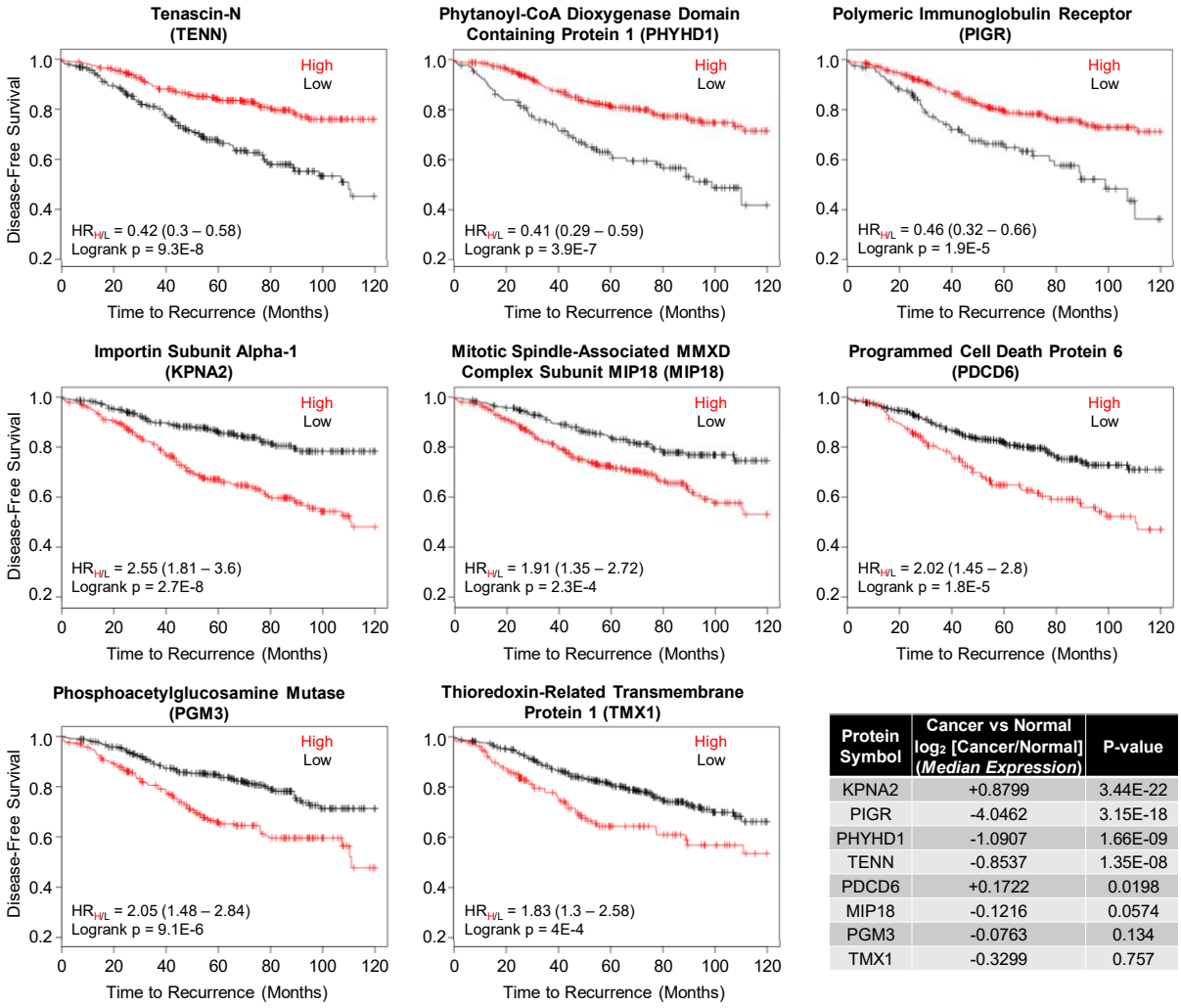
Figure 4.7 Selection of primary biomarker candidates.

(**A**) Volcano plots were generated for each pair-wise comparison of tumour groupings. The -$\log_{10}$ of the p-value originating from *post-hoc* TK-HSD tests were plotted against the $\log_2$-fold-change in protein abundance. Only proteins corresponding to the most statistically significant and differentially abundant per group per comparison were utilized. (**B**) HCA of refined biomarker list, with (**C**) average tumour grouping protein expression, and (**D**) average protein cluster grouping expression.

4.3.5. Correlation of Protein Abundance with mRNA Expression for Proteins of Interest

While it has been shown in several instances that absolute mRNA copy-numbers do not always correlate with their measured proteins' abundance in mammals[357], it is nevertheless accepted and understood that there is a positive correlation between the general trends in these two biomolecules' abundances[358]. Considering this, we were curious how these two parameters correlated with respect to patient survival for proteins in our refined biomarker list. Proteins whose transcript abundance positively correlated with the trends observed for proteins in our list would be ideal candidates for verification and follow-up studies.

As an initial investigation into these trends, proteins from our refined list were searched using the KM-plotter web-utility[332–334]. Proteins considered for follow-up validation were refined by only retaining those which produced an FDR<0.05 (following optimal separation and survival analysis) via KM-plotter, in addition to keeping with the trend we observed via LC-MS/MS for disease-free survival. This resulted in elimination of 12 potential biomarkers from our dataset: DNAJA3, TP53I3, COPS5, SEPHS1, KPNA3, PNPT1, SLC30A7, KDSR, PTER, NADK2, SOD3, and MAOB (data not shown). The remaining list of well-correlated biomarker candidates' (PGM3, PDCD6, TENN, KPNA2, MIP18, TMX1, PIGR, and PHYHD1) microarray plots are illustrated in **Figure 4.8**.



Protein Symbol	Cancer vs Normal log_2 [Cancer/Normal] (*Median Expression*)	P-value
KPNA2	+0.8799	3.44E-22
PIGR	-4.0462	3.15E-18
PHYHD1	-1.0907	1.66E-09
TENN	-0.8537	1.35E-08
PDCD6	+0.1722	0.0198
MIP18	-0.1216	0.0574
PGM3	-0.0763	0.134
TMX1	-0.3299	0.757

Figure 4.8 Investigation of mRNA abundance as indicators of disease prognosis.

KM-plotter[332–334] was utilized to investigate microarray data for mRNA transcripts corresponding to our proteins of interest. Proteins whose transcriptional abundance provided significant (FDR≤0.05) separation with respect to patient disease-recurrence were selected for initial follow-up and validation. This produced 8 initial candidates. (**Bottom-right**) transcript levels for candidate proteins were compared for differential abundance with respect to healthy tissue.

4.3.6. PIGR as a Marker for Disease-Free Survival via Immunohistochemistry (IHC)

Following identification of these 8 proteins with a positive correlation between mRNA and protein abundance with respect to patient recurrence-free survival, PIGR was selected for further investigation for utility as a prognostic and predictive biomarker for disease outcome. Belonging

to the immunoglobulin superfamily, PIGR is a ~100-120 kDa (variably glycosylated)[359–361]transmembrane Fc-receptor primarily expressed on the basil lamina of mucosal epithelial cells[362,363]. Responsible for the transcytosis of polymeric-immunoglobulins (poly-Igs; IgAs and IgMs) across epithelial surfaces[362–364], PIGR – also known as the transmembrane secretory component (SC) – binds poly-Igs in a J-chain (linker peptide between monomeric Igs) dependent fashion[365,366]. Following binding to PIGR, poly-Igs are transcytosed across the epithelium to the apical membrane, where PIGR is cleaved via endoproteases. This proteolysis results in the extracellular release of the poly-Ig, still bound to PIGRs extracellular domain (SC-domain; ~80 kDa). Because of this, PIGR is integral to the proper functioning of mucous membrane immunity[362–366].

As IHC utilizes antibodies for protein detection, we wanted to first verify the changes observed via mass spectrometry could be observed using antibody-based methods; often, mass spectrometers' high sensitivity outperforms that of antibody-based methods[367]. To assess whether this was the case, whole tumour homogenates from seven patients were subject to immunoblot detection. PIGR had been detected in two of the seven samples via LC-MS/MS analysis. As shown in **Figure 4.9**, presence of PIGR as determined by LC-MS/MS analysis of tumour homogenates, correlated well with antibody-based detection. Interestingly however, full-length PIGR was not detected; rather, a fragment correlating to the ~80kDa cleaved extracellular 'secretory component' of PIGR was detected, suggesting an extracellular origin.

Following confirmation that PIGR – as detected or not via LC-MS/MS – could be confirmed with our antibody, tissue microarrays containing breast tumours were subject to IHC analysis. The amount of cancerous-tissue staining was determined visually by Dr. Wei-Feng Dong (Pathology). Additionally, global staining for tissue, including healthy stroma, was performed using Aperio ImageScope's Positive Pixel Count v9 algorithm.

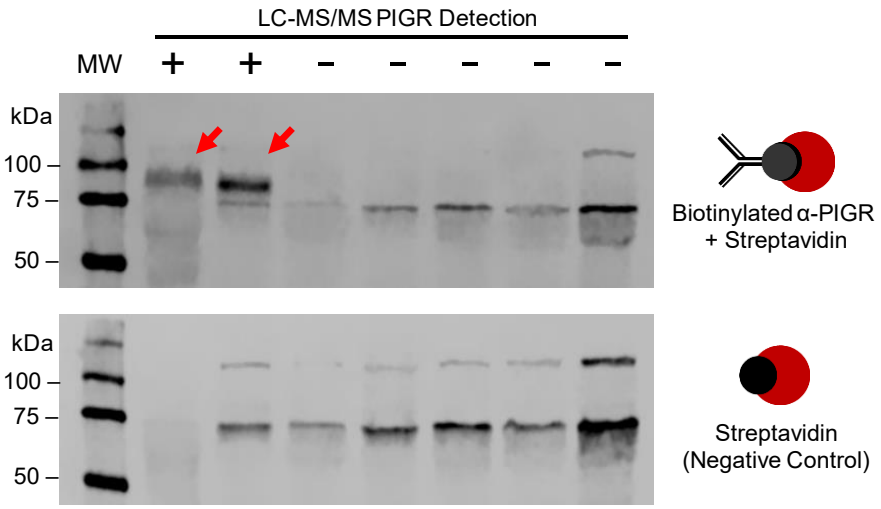


Figure 4.9 Immunoblot confirmation of PIGR abundance as predicted via MS.

Whole tumour lysates used for our initial mass spectrometry analysis were subject to Western blotting using a biotinylated α-human PIGR antibody. (**Top**) PIGR was only detectable in lysates reported to have PIGR in the mass spec data. (**Bottom**) A negative control blot with streptavidin was performed to clarify non-specific/background binding of labelled streptavidin to endogenously biotinylated proteins; these bands were ignored during the interpretation of the blot utilizing biotinylated anti-PIGR.

The ability of PIGR staining via IHC to predict overall patient disease-recurrence was determined through the implementation of ROC curves (**Figure 4.10**)[368,369]. Interestingly, when plotting staining intensities as determined via visual Pathologist-assessment of strictly cancerous cells originating from ER+ tumours, it was found that PIGR was only slightly better at predicting disease-recurrence than flipping a coin, regardless of the timeframe in which recurrence was measured. Interestingly however, when assessing whole-tissue for staining (positive pixel count analysis), PIGR was found to consistently correlate (moderately) well with disease-recurrence, regardless of the reference timeframe. For determination of optimal cut-points for PIGR staining intensities of cancerous and whole tissues, we utilized the 'minimal distance to corner (0%,100%)' method[370,371] on the ROC curve providing the largest reported AUC. Due to the structure of the ROC curve unit square, the upper-left corner (0%,100%) coincides with a perfect test (sensitivity and specificity = 100%). By minimizing the distance on the ROC curve to correlate with this point, an even balance between a test's sensitivity and specificity is achieved. While maximization of Youden's J-statistic[372] (point where sensitivity + specificity is maximal)

was also considered when deciding a cut-point[373], it has recently been shown that the minimal distance to corner (0%,100%) method, as it produces the cut-point with the least bias, and maximally distributes sample populations[374].

Both pathologist-scoring and positive pixel count ROCs displayed the largest AUC when measuring disease-recurrence within 60 months post-surgery/biopsy as shown in **Figure 4.10** (AUC_{Path}=0.5800 ± 0.0753, $CI_{95\%}$=0.4135 – 0.7089; AUC_{Pixel}=0.6921 ± 0.0703, $CI_{95\%}$=0.5282 – 0.8063), resulting in a cut-point of ≤0.3 for pathologist-scored cancerous tissue staining, and ≤0.035 (35 positive pixels per 1000) for positive pixel count scored whole tissue staining. Subsequent survival analysis utilizing our ROC-defined cut-points displayed minimal utility in the ability of visually-assessed PIGR abundance in cancerous tissues, at both 60 and 120 months post-surgery, to determine disease-recurrence (***60 months:*** logrank p=0.0685; $HR_{low:high}$=2.44, $CI_{95\%}$=0.71 – 8.37; ***120 months:*** logrank p=0.1959; $HR_{low:high}$=1.69, $CI_{95\%}$=0.66 – 4.31). However, this was not the case with respect to global tissue staining for PIGR, as determined through pixel-density analysis. At both 60 and 120 months post-surgery, the surviving fractions of defined *high* and *low* populations displayed significant differences in separation (***60 months:*** logrank p=0.0155; $HR_{low:high}$=3.36, $CI_{95\%}$=1.28 – 8.81; ***120 months:*** logrank p=0.0135; $HR_{low:high}$=2.36, $CI_{95\%}$=1.16 – 4.82). The average representative positive pixel counts for dichotomized PIGR abundance, as determined through ROC analysis, is depicted in **Figure 4.11.**

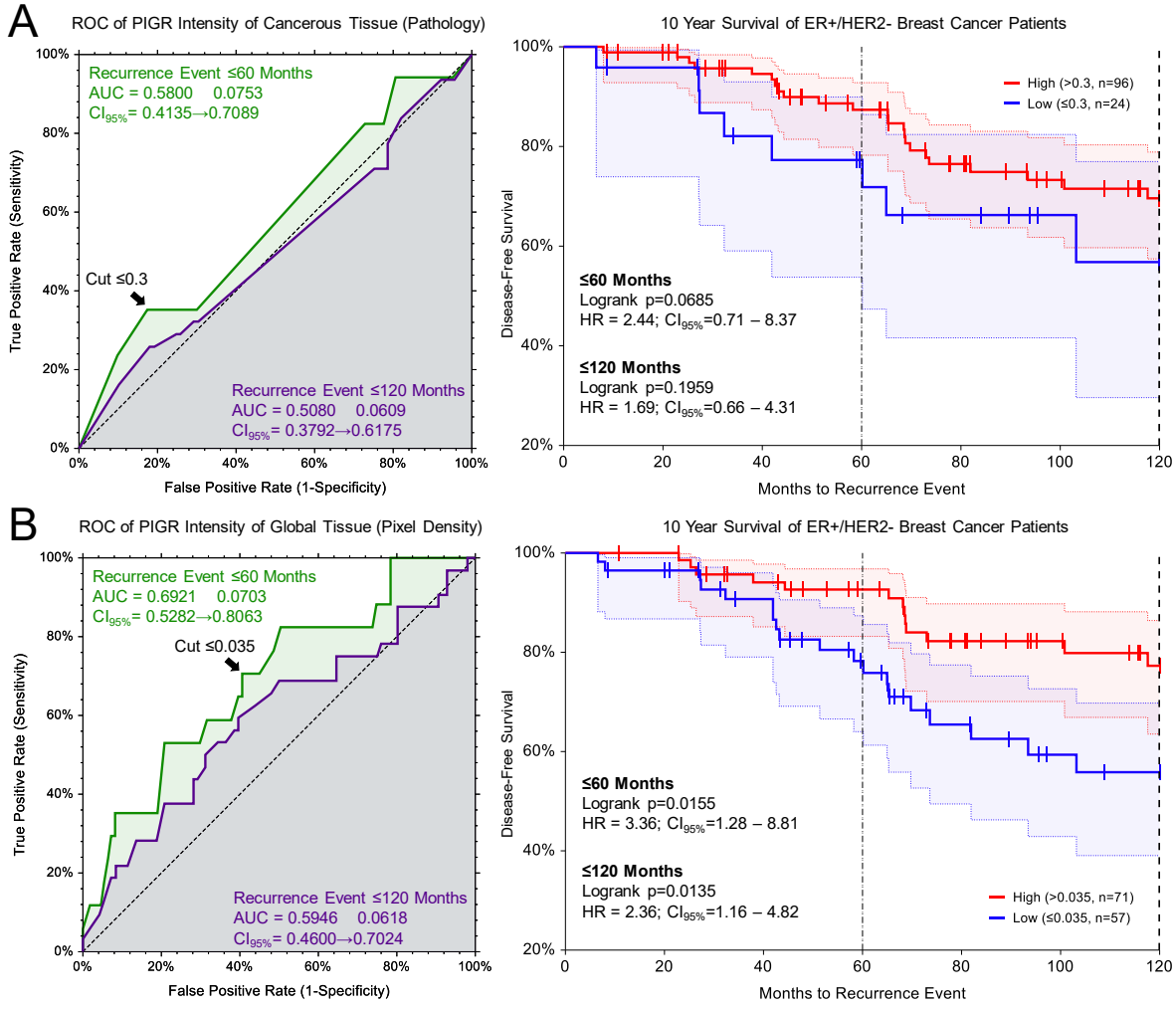


Figure 4.10 Dichotomization and subsequent survival analysis of PIGR abundance as determined via immunohistochemistry.

PIGR was investigated as for its ability to predict disease outcome. ROC curves and corresponding survival plot for (**A**) Pathologist-derived visual scores of cancerous-tissue staining and (**B**) positive pixel counts for staining of whole tissue sections. Optimal dichotomization was for both scoring methods was determined using the shortest-distance to the upper-left corner of the ROC curves (**left**) measuring disease-recurrence within 60 months post-surgery/biopsy (**green**). Survival analysis post-dichotomization of PIGR abundance at both 60 and 120 months post-biopsy/surgery (**right**) illustrated little difference for exclusively cancerous tissue staining (visually scored), while high global PIGR abundance (positive pixel count scored) was indicative of positive patient outcomes at both 60 and 120 months (60 months, p=0.0155; 120 months, p=0.0135).

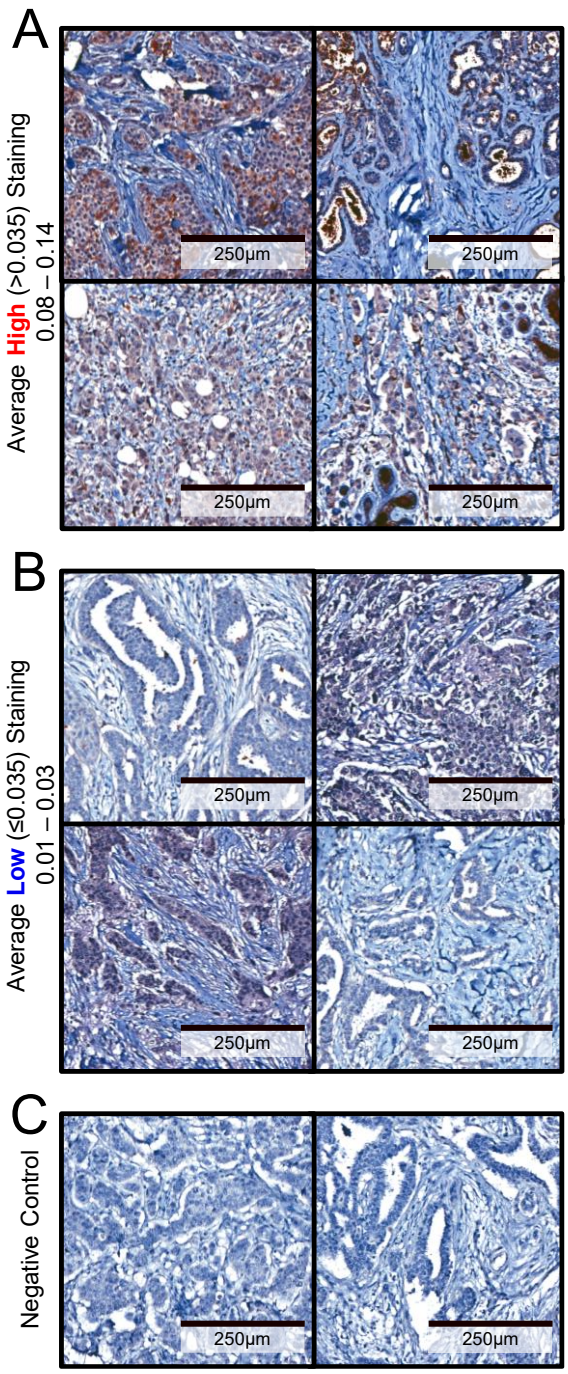


Figure 4.11 Representative average IHC staining of dichotomized positive pixel count PIGR abundance.

Visual representation of average positive pixel counts corresponding to (**A**) **high** and (**B**) **low** groupings relative to tissue without staining (**C**).

This seeming discrepancy between diseased-cellular versus overall tissue staining for PIGR – in conjunction with our proteomics data – can easily be explained when considering the origin of our samples; the clinical tumour samples utilized for this study were unable to be micro-dissected prior to homogenization and LC-MS/MS analysis. Taking this into consideration, it is likely the majority of PIGR observed via LC-MS/MS analysis originated from SC fragments contained in mammary ducts within the bulk tumour. As previously illustrated in **Figure 4.9**, PIGR was detected at ~80kDa via immunoblotting, adding support to this notion. To further investigate this idea, peptides' sequences correlating to PIGR identification and quantification were queried for their point-of-origin within PIGR. Of 26 peptides sequenced from PIGR, only 1 (with a corresponding PSM across all 19 samples = 1) peptide originated from the transmembrane/intracellular domain, while the remaining 25 peptides corresponded to the SC domain.

4.4. Discussion

One of the largest issues affecting the effective management of breast cancer patients is the ability to distinguish between the 'Luminal' subtypes of ER+ breast cancers[317,318,320,338]. To better characterize Luminal-type breast tumours on the proteomic scale and identify better indicators of disease prognosis, we performed a label-free quantitative proteomic analysis of 19 ER+/HER2- clinical breast cancer tumours. Using this technique, we identified a total of 4477 proteins across all tumours. Rather than group tumours based on their pre-treatment diagnosis of Luminal-type, we opted to perform comparative analysis of tumours based on their patients' disease-recurrence status.

4.4.1. Proteomic Characterization of Global Tumour Traits

Interestingly, analysis of tumours in this way revealed an incredibly small proportion of proteins (46 at q<0.10) to correlate both positively and negatively with patient disease-recurrence. Using these 46 proteins, we showcased these tumours can be re-classified into three groups, each with a distinct proteomic fingerprint; these groups were deemed 'disease-free' (DF_o), 'disease-free, recurrent-like' (DF_{RL}), and 'disease-recurrent' (*DR*). It is interesting to note, ER+ breast cancers have been previously classified into three groups based on their response to treatment[338]. The existence of an intermediary classification of tumours in both studies is supportive of the existence of non-binary populations of luminal-type breast cancers.

While performing functional analyses on a refined list of proteins determined to significantly (q<0.05) differ between any of the three groups, *DR* tumours were consistently characterized as having signatures indicative of metabolically and proliferatively-aggressive cellular activity; molecular machinery involved in protein translation, DNA replication, and mismatch repair were found to be in high abundance. These tumours were also found to have a significantly overrepresented propensity for fatty-acid metabolism and oxidative phosphorylation. This could suggest an increase in fatty-acid synthesis for membrane growth during proliferation and cellular division[375–377] – an idea supported by the presence of PPARα-signalling (detailed in reviews [378–

[380]) in these tumours – and/or a high cellular energetic demand – likely due to their highly proliferative state. One protein, of particular interest to our lab, that was observed in high abundance in these tumours was UBR4 – an E3-ubiquitin ligase that is an actuator of the N-end rule[381], a pathway responsible for the proteolytic degradation of proteins with 'destabilizing' residues at their N-terminus[382,383]. Our lab has demonstrated that the N-end rule is responsible for the degradation of both anti-[384,385] and pro-apoptotic[385,386] protein fragments generated via caspase-mediated proteolysis.

Curiously, *DF$_{RL}$* tumours, while originating from patients without disease-recurrence, bore a high degree of similarity to those of the *DR* grouping; of the 426 proteins identified as being significantly differentially expressed ($q<0.05$) between any of the three tumour groups, approximately two-thirds (280) had abundances on par with those observed in *DR* tumours. Most of these shared proteins belonged to pathways related to protein synthesis and turnover – specifically ribosomal proteins, translation initiators and elongators, and factors involved in proteolysis. Additionally, proteins involved in lipogenesis such as those involved in PPARα-signalling were observed, as well as proteins involved in apoptosis such as the Bcl-2-family proteins BAX and BID, the tumour necrosis factor receptor TRADD, cytochrome c (CYCS), and caspase-6 (CASP6) (reviewed in [387,388]). Interestingly, several anti-apoptotic and pro-inflammatory proteins were also observed such as NFκB, IκB-kinase[389], protein kinase A (PKA)[390], and cartilage oligomeric matrix protein (COMP)[391].

However, several distinctions were also observed. *DF$_{RL}$* tumours showed an increase in carbohydrate metabolism rather than the fatty acid metabolism and oxidative phosphorylation observed in *DR* tumours. This could be suggestive of *DF$_{RL}$* tumours experiencing the 'Warburg effect' – cancer cells' propensity to inefficiently metabolize glucose via anaerobic glycolysis, even in the presence of an adequate oxygen supply[392,393]. Alternatively, *DF$_{RL}$* tumours' absence of oxidative phosphorylation relative to *DR* tumours could be indicative of their being in a hypoxic state[394,395]. Additionally, *DF$_{RL}$* tumours displayed an enrichment for ECM-receptor interactions. Among these were several structural/anchoring proteins such as laminin subunit α-4 (LAMA4), integrin β-5 (ITB5), and various isoforms of collagen. Interestingly however, two molecules with

anti-angiogenic properties – collagen α-1 (IV)-chain (COL4A1)[396] – the parent protein of arrestin – and thrombospondin-2 (THBS2)[397], co-populated this list, further supporting the notion that these tumours could potentially be characterized as hypoxic.

This enriched presence of proteins involved in opposing pathways in DF_{RL} tumours could quite possibly be what differentiates these tumours as a separate group from those originating from *DR* patients. The presence of many pro-apoptotic and anti-angiogenic factors could be keeping an otherwise aggressive tumour phenotype under some level of constraint.

Lastly, DF_{o} tumours were characterized by a general absence of proteins observed in DF_{RL} and *DR* groups, indicating their existence in a relatively benign state. From a global perspective, the features unique to each tumour grouping are potentially indicative of individual tumour groups populating different stages along the continuum of oncogenic transformation[398,399], from benign tumours to aggressively proliferative disease.

4.4.2. Identification of Prognostic Biomarkers for Disease Recurrence

In the search for potential predictive and prognostic biomarkers for disease recurrence in luminal-type breast cancers, we opted for a serial-refinement approach, selecting only the most significant and differentially expressed proteins. Through group-wise comparisons of *DF* vs *DR* tumours we initially identified a list of 46 high-confidence proteins which could potentially serve to differentiate the disease-recurrence. Following cluster analysis and the realization that we were in possession of 3 distinct tumour groupings rather than 2, this list of 46 grew to 426 proteins. However, by selecting only the most significant proteins with the largest differential abundance in each group, we produced a list of 20 proteins that could differentiate between three tumour profiles for ER+/HER2- tumours.

To further increase the confidence in our potential biomarkers, we further refined the list of candidates to include only proteins with transcriptional trends corresponding to those we observed via proteomic analysis. Of the resultant 8 proteins, 5 had high expression levels

positively correlated with disease-recurrence. These proteins were: the importin subunit α-1 (KPNA2), a protein responsible for the recognition of proteins' nuclear localization sequences and their subsequent translocation into the nucleus[400]; the mitotic spindle-associated MMXD complex subunit MIP18 (MIP18/FAM96B), involved in iron-sulfur (Fe/S) cluster incorporation[401,402] into various proteins/complexes during chromosomal segregation[403]; thioredoxin-related transmembrane protein 1 (TMX1), a redox-sensitive oxidoreductase present on mitochondrial-associated membranes recently demonstrated to possess tumour-suppressive qualities[404]; phosphoacetylglucosamine mutase (PGM3/AGM1), an isomerase responsible for converting N-acetylglucosamine-6-phosphate into N-acetylglucosamine-1-phosphate, a compound required for multiple protein glycosylation processes[405–407]; and the pro-apoptotic programmed cell death protein 6 (PDCD6), a calcium-binding protein[408] responsible for stabilizing protein-protein complexes (important during apoptosis)[409], and potentially possessing anti-angiogenic properties[410].

Among this subset of proteins, KPNA2 and PGM3 have previously been demonstrated to correlate with poor cancer outcomes. KPNA2 has been correlated with poor breast cancer prognosis[411,412] due to its role in the cytoplasmic retention of enzymes responsible for actuating the DNA-damage response (DDR)[412], while PGM3 has been demonstrated to act as an immunosuppressant[405–407] and be significantly upregulated in clinical prostate cancer in what appears to be an androgen-dependent manner[413,414]. Additionally, the presence of MIP18 as a marker of disease-recurrence is unsurprising; known to play an important role in chromosomal segregation during mitosis[403], its presence is supportive of disease-recurrence tumour cells undergoing cellular division. Interestingly, MIP18 has also been implicated in the down-regulation of E2-2 transcriptional regulation, resulting in an increase in several processes involved in the angiogenic response[415].

Perhaps the most interesting findings were the inclusion of TMX1 and PDCD6 in the list of proteins specific for poor prognosis; TMX1 has recently been implicated in the negative regulation of the sarco/endoplasmic reticulum Ca^{2+} ATPase (SERCA) pump[404]. This interaction results in cytosolic Ca^{2+} retention in addition to increased mitochondrial Ca^{2+} abundance, thereby increasing mitochondrial energy production through oxidative phosphorylation, while also

increasing the effectiveness of the mitochondria's involvement during apoptosis. In line with TMX1's functionality, PDCD6 – a cytosolic Ca^{2+} binding protein[408] – utilizes cytosolic Ca^{2+} to enhance weak protein-protein interactions during cell death[409]. While these data are supportive of the enhanced oxidative phosphorylation observed within *DR* tumours, their correlation as indicators of poor prognosis is slightly perplexing, and warrants further investigation.

The remaining 3 proteins possessed the opposing correlation; high expression positively correlated with disease-free survival. These proteins were: tenascin-N (TENN/TNW), the smallest[416] of a family of extracellular matrix proteins implicated in cellular adhesion and migration[417,418]; phytanoyl-CoA dioxygenase domain-containing protein 1 (PHYHD1), an α-ketoglutarate-dependent dioxygenase closely related to the lipid-metabolizing peroxisomal phytanoyl-CoA dioxygenase (PHYH)[419]; and polymeric immunogloblulin receptor (PIGR), a protein responsible for the transcytosis and secretion of poly-immunoglobulins (IgAs /IgMs) into luminal spaces, in addition to secretory component, forming a crucial component in proper immune function[362–366].

Curious to this list was TENN; the tenascin-family proteins are large ECM proteins primarily serving as ligands for integrins, and are thought to be involved in cell growth and migration[417,418]. TENN and its sibling tenascin-C (TNC) have been demonstrated to be up-regulated in breast cancers[420,421]. TENN has been demonstrated to play a positive role in angiogenesis by modulating endothelial cell growth[422]. Additionally, TENN has been suggested to have a role similar to TNC in metastasis[421]. However, TENN is known to inhibit osteoblastic proliferation[423]. Furthermore, in healthy tissues, TENN and TNXB (also observed in our dataset specific to *DF* tumours, but at slightly lower confidence – see **Supplemental Table 4.2** and **4.3**) are known to occupy opposing areas and tissues to TNC[418,424], suggesting these proteins, while originating from the same protein family, may possess different and opposing functions.

Interestingly, the presence of both PHYHD1 and PIGR are supportive of the role of immunosurveillance[425] in disease-free survival. PHYHD1 has been demonstrated to be up-regulated in T-cells following their immunological stimulation[426]. Our follow-up investigation of PIGR in cancerous tissue via IHC indicated that while PIGR expression in strictly cancerous tissue

was a poor indicator of disease outcome, global PIGR abundance – including its secreted form SC – in bulk tumour tissue correlates incredibly well with patient outcome. We believe this may be suggestive of the immune system's function with aiding in tumour clearance.

4.5. Conclusion and Future Directions

While our understanding of what drives oncogenic transformation has grown exponentially over the past several decades, the complex interactions of systems at play in bulk tumours – including both cellular and extracellular environments – are just beginning to be adequately understood. By increasing our understanding of the biological processes at play, we are better able to classify cancers and predict their natural course of progression, allowing for increasingly better means to manage patients afflicted with this terrible disease. With this also comes the responsibility to reassess the classification systems in implementation.

Here we present the comprehensive analysis of 19 human estrogen receptor positive breast tumours using a label-free quantitative approach. Through reliance on only disease-outcome as an initial means of tumour classification, we ultimately separated these tumours into 3 distinct populations based on protein expression, suggesting and supporting previous observations of luminal subtypes' behaviour when treated with anti-estrogen chemotherapies[338]. Additionally, during our initial analysis, we identified several proteins that can serve as indicators of both disease type and outcome.

Moving forward, we hope to increase the power of our study through inclusion of additional tumours meeting the criteria of this dataset. Through increased statistical power, we believe distinct sub-populations of luminal-subtype breast cancers will become more clearly discernable, yielding both proteomic rationale for their distinctness from one another as well as potential correlative biomarkers of their prognosis. Additionally, the continued validation of potential biomarkers from this study, with potential inclusion of proteins in less-strict q-value ranges, could be utilized to develop a reliable, easy-to-use, protein-based screening tool for scoring and therefore predicting the likelihood of disease recurrence. Ultimately, we hope to enable the better management of breast cancer patients from the time of diagnosis.

Chapter 5 : Current Challenges, Emerging Techniques, and Concluding Remarks

Situated at an intersect between analytical chemistry and cellular biology, the field of proteomics is unique in its ability to shed light – both qualitatively[427] and quantitatively[239] – on the complexities of protein biochemistry. In recent years this field has experienced an explosion in popularity. This explosion can be attributed, partly, to advances in the fields of genomics and transcriptomics, providing comprehensive databases on variable protein sequences and expression[428]. This integration of -omics fields has been dubbed 'proteogenomics'[428,429], and is responsible for greater accuracy in peptide identifications, allowing for deeper proteomic analyses to be performed on biological tissues. Additionally, contributing to this explosion are technological advances in mass spectrometry and in computational performance. While instrument sensitivity has dramatically increased over the last decade, capable of providing incredible accuracy with respect to ion quantification, most quantitative proteomic studies continue to rely on the use of stable-isotopic labelling systems for quantification. Evidence of this is apparent using a key-word search of PubMed Central's database for both 'label-free proteomics' and 'isotope proteomics'; at the time of writing this, approximately 50% more studies utilizing isotopic labels have been published (~12800) compared to those utilizing label-free methods (~8100). While this continued preference for the use of stable-isotopes is somewhat understandable (as samples can be directly compared during data collection), in addition to being expensive and time-consuming, it is becoming increasingly unwarranted with the advent of modern mass spectrometers.

My initial attraction to the field of proteomics was a direct result of both observing and studying real biological systems in the latter years of my undergraduate program. I had found, and continue to find, it perplexing that autonomous, self-replicating and -regulating characteristics, can become emergent properties from a seemingly chaotic mixture of organic molecules. When I discovered that most of these processes were mediated by proteins, I immediately gravitated towards proteomics due to its ability to deconvolute complex proteinaceous mixtures and explain – to some degree – how these emergent properties arise. Because of this, I vehemently believe that the development of approaches to make mass spectrometry-based proteomics more accessible to basic scientists would greatly enable research, and the pursuit of knowledge.

5.1. Practical Application of Quantitative Label-Free Comparative Proteomics

The majority of this thesis is focused on utilization of our label-free proteomic technique for the comparison and relative quantification of proteins originating from a variety of biological samples. While the overall premise of this technique is by no means novel (for reviews see [65,158,430–432]), our technique utilizes a sample-specific normalization procedure for all observed proteins in conjunction with commonly used individual protein quantification methods[24,177]. It has been previously proposed that such a technique can become problematic with respect to pre-fractionation of proteins present within a biological sample[161]. In theory, different combinations of proteins/peptides in each sub-fraction can behave differently during chromatographic separation, leading to poorly reproducible results. However, several steps can be taken to mitigate this sub-fraction variation. By carefully controlling sample pre-fractionation – including how samples are fractionated – and subsequent front-end preparation for replicate samples, in addition to choosing an adequate per-sample normalization procedure (such as normalization to a specified marker or total signal), we have repeatedly demonstrated reliable semi-quantitative comparisons can be made which are verifiable when tested visually through immunoblotting techniques[24,177].

The second chapter[24] of this thesis focused on application of this technique in the characterization of hepatic lipid droplets. Specifically, we investigated how these organelles' proteomes changed with respect to dietary stresses induced by fasting, or fasting followed by a period of refeeding. Our study was the first proteomic analysis of murine hepatic lipid droplets which carefully controlled for animals' feeding states, which in turn demonstrated the dynamic nature of these organelles. Arguably, this type of comparative experiment is the most suitable for application of our technique; samples obtained and analyzed are purified biological samples, replicates of one another, and experimental groups differ by a singular variable. However, due to the accuracy of modern instruments, such experiments are also plagued by noise; due to samples' origin as subcellular organelles, sample preparation often results in varying levels of quantifiable protein

contamination. While such contaminants are easily recognized, they are difficult to conclusively prove as untrue, and as a result can reduce the statistical power of comparative tests.

Chapters 3 and 4 served as deviations from application of our technique in a sub-cellular context. With interests rooted in the study of cancer, we sought to perform proteomic analysis of entire tissues of *in vivo* origin. Chapter 3[177] illustrated one of the first attempts to assess the biological applicability and suitability of a widely and routinely utilized animal tumour model in the field of Experimental Oncology through its proteomic characterization pre- and post-chemotherapeutic administration. Interestingly, as a model widely utilized in the development of imaging agents specific to the activation of apoptosis during chemotherapy treatments, we discovered tumour death may, in fact, be due primarily to the activation of or involvement of cellular pathways such as autophagy or lysosome-mediated cell death. Likewise, due to the successful implementation of our technique with respect to both sub-cellular and whole tissue applications, as described in Chapters 2 and 3, we sought to extrapolate this technique to clinical breast cancer tissues in the search for novel prognostic biomarkers. Chapter 4 thoroughly explored this idea, with the successful identification and preliminary validation of PIGR in breast tissues as an indicator of prognosis, in addition to several other promising candidates. Additionally, we were capable of illustrating – through cluster analysis on proteins differentially abundant with respect to disease outcome – that the tumours analyzed, classically defined by their 'luminal-subtype' consisted of 3 distinct populations, challenging the standards of current clinical disease classifiers.

However, our analysis of clinical tumour samples stressed the importance of both dealing with missing values in datasets, and choosing an appropriate method of data normalization with respect to sample type and/or origin. For both our lipid droplet[24] and EL4-tumour[177] analyses, samples pertaining to each experimental condition were biological replicates of one another. Additionally, experimental conditions were incredibly controlled, typically corresponding to a singular variable. Because of this, individual protein and between-sample normalization, achieved via a specific reference protein or a sample's TIC, are functional solutions. Replicates by nature allow for specific assumptions to be made, including that which claims protein abundance and distribution should be similar. Therefore, normalization of data to a reference protein or ion

count (TIC) within a sample is warranted. Additionally, such techniques minimize the impact of missing values; referencing each protein's abundance to either a protein consistently observed with the highest intensity or to a sample's TIC, returns small values. As most programs impute missing values as zeroes by default, these techniques reduce the variation present in values representing individual proteins' abundance across several replicates. Therefore, this reduces the impact of the missing value during subsequent statistical analysis.

Unfortunately, the same cannot be said for samples with an uncommon origin. While it is possible to rationalize and identify a suitable reference protein for data normalization, care must be taken in assessing the data's distribution/goodness of fit. Often, techniques for normalization of data possessing non-normal (non-Gaussian) distributions are not always effective, resulting in difficulties when performing subsequent statistical analyses. As illustrated in Chapter 4, normalization of tumours' protein abundance to that of histone H4's was insufficient to produce adequate statistical power; normalized protein abundances were also required to be log-transformed in order to make adequate statistical inferences.

5.2. Current Challenges in Mass Spectrometry-based Proteomics

5.2.1. Missing Values

The past several years of my studies have been devoted to the development and application of techniques for, what I consider to be, a *crucial* technology in the study of biological systems via proteomics. However, the field of mass spectrometry-based proteomics is by no means without its difficulties. As previously discussed in detail, issues regarding both data normalization and transformation continue to exist[157,161,350]. However, perhaps one of the largest issues with persistent community discord, affecting but not limited to proteomics, is that of the proper handling of data with missing values (MVs)[157,349,350].

The presence of MVs in large datasets poses a rather unique problem, as the user is never capable of determining with exact certainty whether the value is missing due to complete absence/presence below the limit of detection (missing not at random; MNAR), or presence but with missing-ness due to the culmination of stochastic variability/error (missing at random; MAR or, missing completely at random; MCAR)[157,325,326,433]. Additionally, leaving values missing can result in difficulty performing and interpreting subsequent statistical analyses. As the reasons for MVs vary significantly from one another, so too do the methods devised to deal with them, making discernment of their source crucial.

5.2.1.1. Imputation of Missing Values

Several methods for dealing with MVs have been proposed. For values determined to be MNAR, popular values to impute for MVs are zeroes, the local minimum (LM) value observed for a data range or sample, and the global minimum (GM) observed value for a data range or sample (or some variation of these minimum values, as depicted in Chapter 4)[325,326]. Unfortunately, no one method functions best. Imputation of zeroes and GMs enables maximal differentiation of proteins' abundances, but in datasets with large proportions of MVs these techniques can introduce massive errors in measurements, reducing statistical power. Additionally, techniques

imputing strictly zeroes can completely impede calculations and statistical analyses due to division by zero. On the contrary, imputation of MVs with samples' LMs tends to avoid the problems experienced using zeroes/GMs. However, LM imputation can introduce artificial differences, as LMs may differ significantly between samples, leading to skewed statistical outcomes and false conclusions.

For methods dealing with MVs that are MCAR, algorithms for both simple and complex imputations methods exist. Simple imputation techniques refer to those which perform menial calculations, such as replacing a variables MVs with the mean of its observed values, or estimating a variable's natural distribution, and randomly drawing the values for MVs from said distribution[325,326]. More popular however, are techniques reliant on characteristics defined by the non-MVs of all observed variables to determine the imputed value for MVs. Several popular techniques of this variety exist, including: *k* Nearest Neighbours (*k*NN) estimation[434,435], maximum likelihood estimation (MLE)[436,437], and multivariate imputation by chained equations (MICE)[438–442]. For *k*NN, a data array is converted to a matrix, and for a variable containing MVs, the *k* most similar variables in the matrix (determined through a user-defined distance measurement such as Euclidean distance) are averaged and used to derive a value for MVs[434,435]. For maximum likelihood estimation (MLE) imputation, variables in a dataset are assumed to conform to a function, which is approximated from available data and used to estimate the value of MVs[436,437]. Lastly, multivariate imputation by chained equations (MICE)[438–442] uses a series of successive iterations to accurately estimate MVs. Initially, MICE utilizes a simple MV-imputation technique using a variable's non-MVs; subsequent iterations utilize values (non-MV and MV alike) produced by the preceding round of imputation to generate a multivariate normal distribution, from which, MVs values are estimated. This process is then repeated a user-specified number of times.

Unfortunately, real label-free proteomic datasets often contain MVs in the range of 10-50%[326,443], usually existing as some combination of MCAR and MNAR. This makes choosing an appropriate method of MV imputation incredibly difficult. However, several studies have shown that imputation techniques' performance is directly related to the proportion of MVs; MNAR

techniques are recommended for high proportions of MVs[326], while MCAR techniques are recommended for low proportions of MVs[326]. Logically this makes sense, as MCAR techniques impute random variables from a distribution modelled from available data; if the available data is insufficient to estimate a distribution used for MV imputation, the results could be incredibly biased; imputation from an insufficient distribution could lead to introduction of either a false inflation of meaningful differences, or just the opposite[436,440,442]. However, if the available data is sufficient in estimating a variable's distribution, as it is in datasets with small proportions of MVs, imputation is likely to have a positive impact during subsequent analyses by benefiting statistical power. Contrarily, MNAR techniques are typically more conservative, having little effect on the statistical power of datasets with small proportions of MVs, and greatly diminish the statistical power of datasets with large proportions of MVs[325,326]. Selection of imputation method therefore has direct implications particularly when performing analyses on samples with high variability, such as those of clinical origins.

Compounding this problem is a current trend for the reporting of increasingly large 'significant' lists of proteins. With newer comparative studies reporting increasingly large numbers of significant differences between experimental conditions – which can be partially, but not completely, attributed to improvements in instrument sensitivity – discretion must be taken when interpreting the results. Adding to the list of data normalization/transformation, statistical comparison, and FDR control as sources of bias is method of MV imputation; all can greatly affect the outcome of comparative studies. If any of these factors are inappropriate for the dataset, it can lead to the inference of false or misleading conclusions. While conservative approaches reduce statistical power, they still allow for detection and selection of the most robust differences, which are often of the most interest biologically. Because of this, the data analyzed in this thesis has been analyzed using conservative approaches. With respect to MV imputation, in Chapters 2 and 3, when MV imputation using MCAR methods were warranted, MVs were imputed as zeroes, while in Chapter 4 we imputed half of the detectable global minimum as this was demonstrated to be the most conservative approach following FDR correction.

5.2.2. Identification of Signal Source in Biological Tissue

Another challenge facing the field of proteomics lies with the identification of a peptide/protein's biological source. This can take the form of determining whether a signal is due to high background/noise (as discussed in Chapter 2), or determining the cellular/tissue origin of an identified protein from an extract or homogenate (as referred to in Chapters 3 and 4). While determination of noise can be relatively easy to determine using a sample's number of PSMs per protein (with lower values corresponding to proteins very rarely observed and therefore at or near the limit of detection), the latter has become a somewhat more difficult problem to address. Furthermore, as proteomic characterizations of whole tissues become more common, specifically in the characterization of biosignatures, reliable means of determining proteins' cellular, extracellular, or subcellular origins within tissues are paramount. While the information presented in this thesis – specifically that in Chapters 3 and 4 – is informative of processes happening within the tissue(s) being analyzed, classically samples have often been either too small or not capable of being micro-dissected (i.e. ossification of tissue). This, in conjunction with the requirement of samples to be homogenized prior to protein digestion has drawn comparisons of samples analyzed in this way being more akin to 'fruit smoothies' versus their original 'fruit salad' composition. Fortunately, however, recent advances in technologies have enabled more accurate means of micro-dissecting tissue (via laser-capture microdissection), or have enabled the *post-hoc* correlation of protein localization *in-situ* following homogenate-based identification (via MALDI-imaging).

5.2.2.1. Laser-Capture Microdissection

Laser-capture microdissection (LCM), a technique originally pioneered in 1996[444], utilizes laser-light to microscopically subsection/dissect microtomed tissue samples, allowing for the isolation of specific cellular populations. In the context of determining *in situ* protein localization, LCM is particularly attractive due to its ability to preserve microscopic tissue architecture[445], therefore providing spatial context to proteomic analyses. However, until recently, LCM has been limited in application due to its poor protein recovery (typically less than 1µg)[13]. A high degree of sample

handling – protein extraction, labelling, and digestion – in conjunction with previous-generation mass spectrometers possessing low sensitivity, restricted LCM use in proteomics. Nevertheless, advances in quantitative label-free techniques and instrumentation have both reduced the amount of starting material lost due to handling and required for detection, respectively. Recent reports have illustrated analyses using LCM in conjunction with label-free proteomics are capable of quantitatively identifying thousands of proteins from several thousand cells in a reproducible fashion[13,446,447].

5.2.2.2. MALDI-imaging

An alternative technique that has been gaining momentum is that of MALDI-imaging. A form of mass spectrometry imaging (MSI) originally devised in 1997[448,449], MALDI-imaging utilizes tissue sections that have been coated in an ionizable matrix, allowing for the ionization of biomolecules directly from their point of origin in tissue. Resulting MS spectra can be mapped to an image of the tissue being analyzed, providing an accurate portrayal of biomolecule localization *in situ*[450]. While other forms of MSI exist[451], MALDI-imaging has recently taken precedence due to its ability to ionize large biomolecules (up to ~150kDa) with moderate spatial resolution (~20μm), without being overly destructive to tissue compared to other techniques[451]. Because of this, MALDI-imaging is well-suited for a variety of *in situ* applications, ranging from the identification of small metabolites to large intact proteins. One of the current limitations however is the poor ability to resolve individual molecules out of the high-complexity spectra; few databases exist which adequately address the intact masses (complete with PTMs) of large biomolecules such as proteins. Because of this, MALDI-imaging is still in its infancy with primary applications in the mapping of cellular metabolites[452,453]. However, due to its inherent suitability to diagnostic fields such as pathology, the number and sizes of available databases for proteins identified through MALDI-imaging is steadily increasing[454].

Unfortunately, technological availability remains a major limitation in the use of LCM and MALDI-imaging. As a result, many scientists – ourselves included – have been reliant on classical approaches such as *post-hoc* IHC to determine the spatial *in situ* localization of proteins

discovered through whole-tissue homogenization. However, due to proteomics' applicability in fields of medicine studying the pathological basis of disease, it is foreseeable that techniques capable of spatially resolving unique protein signatures in tissue, such as LCM and MALDI-imaging, will replace classical techniques such as IHC, setting a new benchmark in the process.

5.3. Emerging Proteomics Techniques

5.3.1. Targeted Approaches

In recent years there has been a surge in the development and application of targeted (data-dependent) proteomic techniques, based around the principles of selected[63] reaction monitoring (SRM), in which a peptide of interest's m/z (precursor ion) is targeted for fragmentation, using a specified m/z daughter ion as a reporter for the precursor (SRM). The parent-daughter m/z pair(s) is referred to as a 'transition', with transitions allowing for the selective quantitation of parent ions within a sample. Workflows which monitor the generation of multiple transitions are referred to as multiple-reaction monitoring (MRM)[178]. Additionally, through the implementation of orbitraps' superior scan times, mass accuracy, and resolution, this technique has recently been extrapolated to monitor the all daughter ions formed following their parent's fragmentation – a process which has been termed 'parallel reaction monitoring' (PRM)[179]prm. SRM, PRM, and MRM allow for the relative quantification of a peptide in a label-free manner (comparing relative abundances), while introduction of an isotopically-labelled peptide identical to that being targeted allows for absolute quantification. Due to this, it has been postulated that classical means of visual protein detection and validation, such as immunoblotting, be replaced by SRM, PRM, and MRM techniques[455].

5.3.2. Untargeted Approaches

In addition to SRM, MRM, and PRM, a recent technique building on their strengths albeit in a data-independent manner has been developed, termed SWATH-MS[68,456]. Named due to data acquisition resembling swaths, SWATH-MS sequentially cycles through small, incremental precursor isolation windows (i.e. 25Da increments within a defined full-MS range of 400-1200 m/z) in a repeating manner. All precursor ions residing within an isolation window, and present during that window's acquisition scan are fragmented, cataloguing all daughter ions generated. In this way, SWATH-MS serves as a global means of performing SRM/MRM. Following data acquisition, SWATH-MS data is correlated with proteomic spectral libraries containing *a*

priori information regarding peptides' precursor and fragment ion masses along with any additional and informative information such as LC elution time[456,457]. In this way, it is becoming increasingly possible to perform highly accurate, per-sample quantitation of proteins present in complex mixtures.

Lastly, an incredibly recent label-free proteomics technique, similar in principle to SWATH-MS but specific to Orbitrap mass analyzers, dubbed 'BoxCar' acquisition has been described[67]; BoxCar acquisition utilizes narrow, interspaced, boxcar-function m/z acquisition windows to sequentially catalogue an entire m/z range over a series of scans. By limiting the number of ions entering the mass analyzer to those ions existing in specific m/z ranges, the analyzer can resolve and quantify more of the peptides present, thusly increasing the depth of the sampled proteome (up to 10000 proteins)[67] in a highly reproducible and time-efficient manner.

With techniques such as SWATH-MS and BoxCar continually increasing the quantifiable depth of proteomic studies, an increasing amount of care must also be taken by their user(s) during such studies' data analysis. As previously mentioned – and demonstrated in Chapter 2 – an increase in proteomic depth comes with an increase in noise and/or contaminants; as techniques continue to advance the number of proteins quantifiably identified within samples, the ability to distinguish between those which are biologically relevant and those which are not is imperative.

5.4. Concluding Remarks

Over the past several years, my research has focused on the development of robust, reliable, and reproducible techniques for performing bottom-up label-free quantitative proteomics via mass spectrometry. This thesis has explored several aspects of my research, from methods pertaining to reliable data correction and normalization, to suitable applications of this method in the characterization of various biological systems such as: characterization of organellar proteomes; the chemotherapeutic response of lymphoma tumours; and the identification of prognostic classifiers of disease outcome in ER+ breast cancers. While newer techniques and technologies continually advance the depth at which proteomic studies can be performed, their dependence on state-of-the art instrumentation can be limiting. Therefore, in addition to all-else, the techniques described in this thesis are practical, encouraging their further application in biological research.

Certainly, as mass spectrometric techniques for the study of proteomics – and biological systems in general – become further engrained into the practice of medical and biological research, bountiful gains will be observed. Finally, perhaps the most exciting applications of such label-free techniques lie with their direct application in diagnostic medicine, increasing diagnostic accuracy and our ability to understand and manage complex disease processes. With diagnostic fields trending towards making the implementation of label-free proteomic analyses a reality, the future for all aspects of medicine is looking undeniably promising.

References

1. Wasinger VC, Cordwell SJ, Poljak A, et al. Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis*. 1995;16(1):1090-1094. doi:10.1002/elps.11501601185.

2. Polakis ES, Bartley W. Changes in dry weight, protein, deoxyribonucleic acid, ribonucleic acid and reserve and structural carbohydrate during the aerobic growth cycle of yeast. *Biochem J*. 1966;98(3):883-887. http://www.ncbi.nlm.nih.gov/pubmed/5911532. Accessed March 13, 2018.

3. Yamada EA, Sgarbieri VC. Yeast (*Saccharomyces cerevisiae*) Protein Concentrate: Preparation, Chemical Composition, and Nutritional and Functional Properties. *J Agric Food Chem*. 2005;53(10):3931-3936. doi:10.1021/jf0400821.

4. Feijó Delgado F, Cermak N, Hecht VC, et al. Intracellular Water Exchange for Measuring the Dry Mass, Water Mass and Changes in Chemical Composition of Living Cells. Polymenis M, ed. *PLoS One*. 2013;8(7):e67590. doi:10.1371/journal.pone.0067590.

5. Chinwalla AT, Cook LL, Delehaunty KD, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002;420(6915):520-562. doi:10.1038/nature01262.

6. Gibbs RA, Weinstock GM, Metzker ML, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*. 2004;428(6982):493-521. doi:10.1038/nature02426.

7. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science (80-)*. 2001;291(5507):1304-1351. doi:10.1126/science.1058040.

8. Wu C, Nebert DW. Update on genome completion and annotations: Protein Information Resource. *Hum Genomics*. 2004;1(3):229-233. doi:10.1186/1479-7364-1-3-229.

9. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R. UniProt archive. *Bioinformatics*. 2004;20(17):3236-3237. doi:10.1093/bioinformatics/bth191.

10. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res*. 2015;43(D1):D204-D212. doi:10.1093/nar/gku989.

11. Christoforou A, Mulvey CM, Breckels LM, et al. A draft map of the mouse pluripotent stem cell spatial proteome. *Nat Commun*. 2016;7:9992. doi:10.1038/ncomms9992.

12. Marx V. Mapping proteins with spatial proteomics. *Nat Methods*. 2015;12(9):815-819. doi:10.1038/nmeth.3555.

13. Clair G, Piehowski PD, Nicola T, et al. Spatially-Resolved Proteomics: Rapid Quantitative Analysis of Laser Capture Microdissected Alveolar Tissue Samples. *Sci Rep*. 2016;6(1):39223. doi:10.1038/srep39223.

14. Rigbolt KTG, Prokhorova TA, Akimov V, et al. System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. *Sci Signal*. 2011;4(164):rs3. doi:10.1126/scisignal.2001570.

15. Bartke T, Borgel J, DiMaggio PA. Proteomics in epigenetics: New perspectives for cancer research. *Brief Funct Genomics*. 2013;12(3):205-218. doi:10.1093/bfgp/elt002.

16. Grünenfelder B, Rummel G, Vohradsky J, Röder D, Langen H, Jenal U. Proteomic analysis of the bacterial cell cycle. *Proc Natl Acad Sci U S A*. 2001;98(8):4681-4686. doi:10.1073/pnas.071538098.

17. Ly T, Ahmad Y, Shlien A, et al. A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. *Elife*. 2014;2014(3):e01630. doi:10.7554/eLife.01630.

18. Li G-W, Burkhardt D, Gross C, Weissman JS. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*. 2014;157(3):624-635. doi:10.1016/j.cell.2014.02.033.

19. Boisvert F-M, Ahmad Y, Gierliński M, et al. A Quantitative Spatial Proteomics Analysis of Proteome Turnover in Human Cells. *Mol Cell Proteomics*. 2012;11(3):M111.011429. doi:10.1074/mcp.M111.011429.

20. Claydon AJ, Beynon R. Proteome Dynamics: Revisiting Turnover with a Global Perspective. *Mol Cell Proteomics*. 2012;11(12):1551-1565. doi:10.1074/mcp.O112.022186.

21. Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Mol Cell Proteomics*. 2012;11(3):M111.014050. doi:10.1074/mcp.M111.014050.

22. Azimifar SB, Nagaraj N, Cox J, Mann M. Cell-type-resolved quantitative proteomics of murine liver. *Cell Metab*. 2014;20(6):1076-1087. doi:10.1016/j.cmet.2014.11.002.

23. Lindqvist LM, Tandoc K, Topisirovic I, Furic L. Cross-talk between protein synthesis, energy metabolism and autophagy in cancer. *Curr Opin Genet Dev*. 2018;48:104-111. doi:10.1016/j.gde.2017.11.003.

24. Kramer DA, Quiroga AD, Lian J, Fahlman RP, Lehner R. Fasting and refeeding induces changes in the mouse hepatic lipid droplet proteome. *J Proteomics*. 2018;181:213-224. doi:10.1016/J.JPROT.2018.04.024.

25. Minamoto T, Mai M, Ronai Z. Environmental factors as regulators and effectors of multistep carcinogenesis. *Carcinogenesis*. 1999;20(4):519-527. doi:10.1093/carcin/20.4.519.

26. Newman JRS, Ghaemmaghami S, Ihmels J, et al. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*. 2006;441(7095):840-846. doi:10.1038/nature04785.

27. Tomanek L. Environmental Proteomics: Changes in the Proteome of Marine Organisms in Response to Environmental Stress, Pollutants, Infection, Symbiosis, and Development. *Ann Rev Mar Sci*. 2011;3(1):373-399. doi:10.1146/annurev-marine-120709-142729.

28. Klose J. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik*. 1975;26(3):231-243. http://www.ncbi.nlm.nih.gov/pubmed/1093965. Accessed March 14, 2018.

29. O'Farrell PH. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem*. 1975;250(10):4007-4021. http://www.ncbi.nlm.nih.gov/pubmed/236308. Accessed March 14, 2018.

30. Rabilloud T, Vaezzadeh AR, Potier N, Lelong C, Leize-Wagner E, Chevallet M. Power and limitations of electrophoretic separations in proteomics strategies. https://arxiv.org/ftp/arxiv/papers/0909/0909.4158.pdf. Accessed March 14, 2018.

31. Görg A, Boguth G, Köpf A, Reil G, Parlar H, Weiss W. Sample prefractionation with Sephadex isoelectric focusing prior to narrow pH range two-dimensional gels. *Proteomics*. 2002;2(12):1652-1657. doi:10.1002/1615-9861(200212)2:12<1652::AID-PROT1652>3.0.CO;2-3.

32. LAEMMLI UK. Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature*. 1970;227(5259):680-685. doi:10.1038/227680a0.

33. Schägger H, von Jagow G. Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal Biochem*. 1987;166(2):368-379. doi:10.1016/0003-2697(87)90587-2.

34. Edman P, Högfeldt E, Sillén LG, Kinell P-O. Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chem Scand*. 1950;4:283-293. doi:10.3891/acta.chem.scand.04-0283.

35. Edman P, Begg G. A protein sequenator. *Eur J Biochem*. 1967;1(1):80-91. http://www.ncbi.nlm.nih.gov/pubmed/6059350. Accessed March 14, 2018.

36. Niall HD. Automated Edman degradation: the protein sequenator. *Methods Enzymol*. 1973;27:942-1010. http://www.ncbi.nlm.nih.gov/pubmed/4773306. Accessed March 14, 2018.

37. Goldstein E. Ueber eine noch nicht untersuchte Strahlungsform an der Kathode inducirter Entladungen. *Ann Phys*. 1898;300(1):38-48. doi:10.1002/andp.18983000105.

38. Thomson JJ. Bakerian Lecture: Rays of Positive Electricity. *Proc R Soc A Math Phys Eng Sci*. 1913;89(607):1-20. doi:10.1098/rspa.1913.0057.

39. Wien W. Untersuchungen über die electrische Entladung in verdünnten Gasen. *Ann Phys*. 1898;301(6):440-452. doi:10.1002/andp.18983010618.

40. Thomson JJ. XL. *Cathode Rays*. *London, Edinburgh, Dublin Philos Mag J Sci*. 1897;44(269):293-316. doi:10.1080/14786449708621070.

41. Thomson JJ. LVIII. *On the masses of the ions in gases at low pressures*. *London, Edinburgh, Dublin Philos Mag J Sci*. 1899;48(295):547-567. doi:10.1080/14786449908621447.

42. Thomson JJ. XIX. *Further experiments on positive rays*. *London, Edinburgh, Dublin Philos Mag J Sci*. 1912;24(140):209-253. doi:10.1080/14786440808637325.

43. Downard KM. Francis William Aston: The Man Behind the Mass Spectrograph. *Eur J Mass Spectrom*. 2007;13(3):177-190. doi:10.1255/ejms.878.

44. Dempster AJ. A new Method of Positive Ray Analysis. *Phys Rev*. 1918;11(4):316-325. doi:10.1103/PhysRev.11.316.

45. Aston FW. LXXIV. *A positive ray spectrograph*. *London, Edinburgh, Dublin Philos Mag J Sci*. 1919;38(228):707-714. doi:10.1080/14786441208636004.

46. Fenn JB. Electrospray Wings for Molecular Elephants (Nobel Lecture). *Angew Chemie Int Ed*. 2003;42(33):3871-3894. doi:10.1002/anie.200300605.

47. Karas M, Bachmann D, Hillenkamp F. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Anal Chem*. 1985;57(14):2935-2939. doi:10.1021/ac00291a042.

48. Dreisewerd K. The Desorption Process in MALDI. *Chem Rev*. 2003;103(2):395-426. doi:10.1021/cr010375i.

49. Karas M, Krüger R. Ion Formation in MALDI: The Cluster Ionization Mechanism. *Chem Rev*. 2003;103(2):427-440. doi:10.1021/cr010376a.

50. Tanaka K, Waki H, Ido Y, et al. Protein and polymer analyses up tom/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom*. 1988;2(8):151-153. doi:10.1002/rcm.1290020802.

51. Yamashita M, Fenn JB. Electrospray ion source. Another variation on the free-jet theme. *J Phys Chem*. 1984;88(20):4451-4459. doi:10.1021/j150664a002.

52. Rayleigh, Lord. XX. *On the equilibrium of liquid conducting masses charged with electricity*. *London, Edinburgh, Dublin Philos Mag J Sci*. 1882;14(87):184-186. doi:10.1080/14786448208628425.

53. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. *Science*. 1989;246(4926):64-71. http://www.ncbi.nlm.nih.gov/pubmed/2675315. Accessed March 14, 2018.

54. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization-principles and practice. *Mass Spectrom Rev*. 1990;9(1):37-70.

doi:10.1002/mas.1280090103.

55. Banerjee S, Mazumdar S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *Int J Anal Chem*. 2012;2012:1-40. doi:10.1155/2012/282574.

56. Biemann K. Mass Spectrometric Methods for Protein Sequencing. *Anal Chem*. 1986;58(13):1288 A-1300 A. doi:10.1021/ac00126a001.

57. Hunt DF, Yates JR, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci*. 1986;83(17):6233-6237. doi:10.1073/pnas.83.17.6233.

58. Vandermarliere E, Stes E, Gevaert K, Martens L. Resolution of protein structure by mass spectrometry. *Mass Spectrom Rev*. 2016;35(6):653-665. doi:10.1002/mas.21450.

59. Larsen MR, Trelle MB, Thingholm TE, Jensen ON. Analysis of posttranslational modifications of proteins by tandem mass spectrometry. *Biotechniques*. 2006;40(6):790-798. doi:10.2144/000112201.

60. Witze ES, Old WM, Resing KA, Ahn NG. Mapping protein post-translational modifications with mass spectrometry. *Nat Methods*. 2007;4(10):798-806. doi:10.1038/nmeth1100.

61. Ewing RM, Chu P, Elisma F, et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*. 2007;3(1). doi:10.1038/msb4100134.

62. Smits AH, Vermeulen M. Characterizing Protein–Protein Interactions Using Mass Spectrometry: Challenges and Opportunities. *Trends Biotechnol*. 2016;34(10):825-834. doi:10.1016/j.tibtech.2016.02.014.

63. Lange V, Picotti P, Domon B, Aebersold R. Selected reaction monitoring for quantitative proteomics: A tutorial. *Mol Syst Biol*. 2008;4(1):222. doi:10.1038/msb.2008.61.

64. Liebler DC, Zimmerman LJ. Targeted Quantitation of Proteins by Mass Spectrometry. *Biochemistry*. 2013;52(22):3797-3806. doi:10.1021/bi400110b.

65. Bantscheff M, Lemeer S, Savitski MM, Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem*. 2012;404(4):939-965. doi:10.1007/s00216-012-6203-4.

66. James P. Protein identification in the post-genome era : the rapid rise of proteomics. 2017.

67. Meier F, Geyer PE, Virreira Winter S, Cox J, Mann M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Methods*. May 2018:1. doi:10.1038/s41592-018-0003-5.

68. Gillet LC, Navarro P, Tate S, et al. 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*. 2012;11(6):O111.016717. doi:10.1074/mcp.O111.016717.

69. Murray KK, Boyd RK, Eberlin MN, Langley GJ, Li L, Naito Y. Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013). *Pure Appl Chem*. 2013;85(7):1515-1609. doi:10.1351/PAC-REC-06-04-06.

70. Winkler R. ESIprot: a universal tool for charge state determination and molecular weight calculation of proteins from electrospray ionization mass spectrometry data. *Rapid Commun Mass Spectrom*. 2010;24(3):285-294. doi:10.1002/rcm.4384.

71. resolution in mass spectroscopy. In: *IUPAC Compendium of Chemical Terminology*. Research Triagle Park, NC: IUPAC. doi:10.1351/goldbook.R05318.

72. Todd JFJ. Recommendations for nomenclature and symbolism for mass spectroscopy (including an appendix of terms used in vacuum technology). (Recommendations 1991). *Pure Appl Chem*. 1991;63(10):1541-1566. doi:10.1351/pac199163101541.

73. Co. AT. Mass Accuracy and Mass Resolution in TOF MS. 2011;(October):1-32. papers3://publication/uuid/02661834-3D64-4516-9F0B-9B61DFBC04A5.

74. resolving power in mass spectrometry. In: *IUPAC Compendium of Chemical Terminology*. Research Triagle Park, NC: IUPAC. doi:10.1351/goldbook.R05321.

75. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Cooks RG. The Orbitrap: A new mass spectrometer. *J Mass Spectrom*. 2005;40(4):430-443. doi:10.1002/jms.856.

76. Eliuk S, Makarov A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu Rev Anal Chem*. 2015;8(1):61-80. doi:10.1146/annurev-anchem-071114-040325.

77. Strupat K, Scheibner O, Bromirski M, Fisher Scientific T. High-Resolution, Accurate-Mass Orbitrap Mass Spectrometry – Definitions, Opportunities, and Advantages. https://assets.thermofisher.com/TFS-Assets/CMD/Application-Notes/TN-64287-LC-MS-Orbitrap-MS-Terminology-Advantages-TN64287-EN.pdf. Accessed May 14, 2018.

78. Ribeiro D. A Short Overview of the Components in Mass Spectrometry Instrumentation for Proteomics Analyses. In: *Tandem Mass Spectrometry - Molecular Characterization*. ; 2013. doi:10.5772/54484.

79. Bogdanov B, Smith RD. Proteomics by FTICR mass spectrometry: Top down and bottom up. *Mass Spectrom Rev*. 2005;24(2):168-200. doi:10.1002/mas.20015.

80. Zhang Z, Wu S, Stenoien DL, Paša-Tolić L. High-Throughput Proteomics. *Annu Rev Anal Chem*. 2014;7(1):427-454. doi:10.1146/annurev-anchem-071213-020216.

81. Kelleher NL. Peer Reviewed: Top-Down Proteomics. *Anal Chem*. 2004;76(11):196 A-203 A. doi:10.1021/ac0415657.

82. Toby TK, Fornelli L, Kelleher NL. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu Rev Anal Chem*. 2016;9(1):499-519. doi:10.1146/annurev-anchem-071015-041550.

83. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR. Protein analysis by shotgun/bottom-up proteomics. *Chem Rev*. 2013;113(4):2343-2394. doi:10.1021/cr3003533.

84. Giansanti P, Tsiatsiani L, Low TY, Heck AJR. Six alternative proteases for mass spectrometry–based proteomics beyond trypsin. *Nat Protoc*. 2016;11(5):993-1006. doi:10.1038/nprot.2016.057.

85. Mant CT, Hodges RS. [1] Analysis of peptides by high-performance liquid chromatography. *Methods Enzymol*. 1996;271:3-50. doi:10.1016/S0076-6879(96)71003-0.

86. Pitt JJ. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *Clin Biochem Rev*. 2009;30(1):19-34. http://www.ncbi.nlm.nih.gov/pubmed/19224008. Accessed May 14, 2018.

87. Mant CT, Lorne Burke TW, Hodges RS. Optimization of peptide separations in reversed-phase HPLC: Isocratic versus gradient elution. *Chromatographia*. 1987;24(1):565-572. doi:10.1007/BF02688546.

88. Mant CT, Parker JM, Hodges RS. Size-exclusion high-performance liquid chromatography of peptides. Requirement for peptide standards to monitor column performance and non-ideal behaviour. *J Chromatogr*. 1987;397:99-112. http://www.ncbi.nlm.nih.gov/pubmed/3654835. Accessed May 14, 2018.

89. Mant CT, Hodges RS. Separation of peptides by strong cation-exchange high-performance liquid chromatography. *J Chromatogr A*. 1985;327(C):147-155. doi:10.1016/S0021-9673(01)81643-5.

90. Lorne Burke TW, Mant CT, Black JA, Hodges RS. Strong cation-exchange high-performance liquid chromatography of peptides. Effect of non-specific hydrophobic interactions and linearization of peptide retention behaviour. *J Chromatogr A*. 1989;476(C):377-389. doi:10.1016/S0021-9673(01)93883-X.

91. Yoshida T. Peptide Separation in Normal Phase Liquid Chromatography. *Anal Chem*. 1997;69(15):3038-3043. doi:10.1021/ac9702204.

92. Aguilar M-I. Reversed-Phase High-Performance Liquid Chromatography. *HPLC Pept Proteins Methods Protoc*. 2004;251:9-22. doi:10.1385/1-59259-742-4:9.

93. Wilm M, Mann M. Analytical Properties of the Nanoelectrospray Ion Source. *Anal Chem*. 1996;68(1):1-8. doi:10.1021/ac9509519.

94. Juraschek R, Dülcks T, Karas M. Nanoelectrospray—more than just a minimized-flow electrospray ionization source. *J Am Soc Mass Spectrom*. 1999;10(4):300-308.

doi:10.1016/S1044-0305(98)00157-3.

95. Wilm M. Principles of Electrospray Ionization. *Mol Cell Proteomics*. 2011;10(7):M111.009407. doi:10.1074/mcp.M111.009407.

96. Wollnik H. Ion optics in mass spectrometers. *J Mass Spectrom*. 1999;34(10):991-1006. doi:10.1002/(SICI)1096-9888(199910)34:10<991::AID-JMS870>3.0.CO;2-1.

97. Paul W, Steinwedel H. Notizen: Ein neues Massenspektrometer ohne Magnetfeld. *Zeitschrift für Naturforsch A*. 1953;8(7):448-450. doi:10.1515/zna-1953-0710.

98. Paul W. Electromagnetic Traps for Charged and Neutral Particles(Nobel Lecture). *Angew Chemie Int Ed English*. 1990;29(7):739-748. doi:10.1002/anie.199007391.

99. Douglas DJ, Frank AJ, Mao D. Linear ion traps in mass spectrometry. *Mass Spectrom Rev*. 2005;24(1):1-29. doi:10.1002/mas.20004.

100. Richards JA. The Mathieu Equation. In: *Analysis of Periodically Time-Varying Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg; 1983:93-107. doi:10.1007/978-3-642-81873-8_6.

101. Dawson PH. *Quadrupole Mass Spectrometry and Its Applications*. Elsevier Science; 2013. https://books.google.ca/books?id=A603BQAAQBAJ.

102. Makarov A. Electrostatic Axially Harmonic Orbital Trapping : A High-Performance Technique of Mass Analysis. 2000;72(6):1156-1162. doi:10.1021/ac991131p.

103. Kingdon KH. A Method for the Neutralization of Electron Space Charge by Positive Ionization at Very Low Gas Pressures. *Phys Rev*. 1923;21(4):408-418. doi:10.1103/PhysRev.21.408.

104. Perry RH, Cooks RG, Noll RJ. ORBITRAP MASS SPECTROMETRY : INSTRUMENTATION , ION MOTION AND APPLICATIONS. 2008;(May):661-699. doi:10.1002/mas.

105. Knight RD. Storage of ions from laser-produced plasmas. *Appl Phys Lett*. 1981;38(4):221-223. doi:10.1063/1.92315.

106. Perry RH, Cooks RG, Noll RJ. Orbitrap mass spectrometry: Instrumentation, ion motion and applications. *Mass Spectrom Rev*. 2008;27(6):661-699. doi:10.1002/mas.20186.

107. Makarov A, Denisov E, Lange O. Performance Evaluation of a High-field Orbitrap Mass Analyzer. *J Am Soc Mass Spectrom*. 2009;20(8):1391-1396. doi:10.1016/j.jasms.2009.01.005.

108. Scigelova M, Hornshaw M, Giannakopulos A, Makarov A. Fourier Transform Mass Spectrometry. 2011:1-19. doi:10.1074/mcp.M111.009431.

109. Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biol Mass Spectrom*. 1984;11(11):601-601. doi:10.1002/bms.1200111109.

110. Jankowski K, Paré JR, Bélanger J. Comments on a “Proposal for a common nomenclature for sequence ions in mass spectra of peptides.” *Biomed Mass Spectrom*. 1985;12(10):631. http://www.ncbi.nlm.nih.gov/pubmed/2933086. Accessed May 15, 2018.

111. Medzihradszky KF. Peptide Sequence Analysis. *Methods Enzymol*. 2005;402:209-244. doi:10.1016/S0076-6879(05)02007-0.

112. Coon JJ, Syka JEP, Shabanowitz J, Hunt DF. Tandem mass spectrometry for peptide and protein sequence analysis. *Biotechniques*. 2005;38(4):519, 521, 523. http://www.ncbi.nlm.nih.gov/pubmed/15884666. Accessed May 15, 2018.

113. Papayannopoulos IA. The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrom Rev*. 1995;14(1):49-73. doi:10.1002/mas.1280140104.

114. Tran NH, Zhang X, Xin L, Shan B, Li M. De novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A*. 2017;114(31):8247-8252. doi:10.1073/pnas.1705691114.

115. Bateman A, Martin MJ, O’Donovan C, et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45(D1):D158-D169. doi:10.1093/nar/gkw1099.

116. Boeckmann B, Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. 2003;31(1):365-370. http://www.ncbi.nlm.nih.gov/pubmed/12520024. Accessed May 15, 2018.

117. Cottrell JS. Protein identification by peptide mass fingerprinting. *Pept Res*. 1994;7(3):115-124. http://www.ncbi.nlm.nih.gov/pubmed/8081066. Accessed May 15, 2018.

118. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*. 1994;5(11):976-989. doi:10.1016/1044-0305(94)80016-2.

119. Yates JR, Eng JK, McCormack AL, Schieltz D. Method to Correlate Tandem Mass Spectra of Modified Peptides to Amino Acid Sequences in the Protein Database. *Anal Chem*. 1995;67(8):1426-1436. doi:10.1021/ac00104a020.

120. Yates JR. Mass spectrometry and the age of the proteome. *J Mass Spectrom*. 1998;33(1):1-19. doi:10.1002/(SICI)1096-9888(199801)33:1<1::AID-JMS624>3.0.CO;2-9.

121. Burger T. Gentle Introduction to the Statistical Foundations of False Discovery Rate in Quantitative Proteomics. *J Proteome Res*. 2018;17(1):12-22. doi:10.1021/acs.jproteome.7b00170.

122. Li YYF, Arnold RJ, Li YYF, Radivojac P. A Bayesian Approach to Protein Inference Problem. *J Comput Biol*. 2009;16(8):1183-1193. doi:10.1089/cmb.2009.0018.

123. Li YF, Radivojac P. Computational approaches to protein inference in shotgun proteomics. *BMC Bioinformatics*. 2012;13 Suppl 1(Suppl 16):S4. doi:10.1186/1471-2105-13-S16-S4.

124. Huang T, Wang J, Yu W, He Z. Protein inference: a review. *Brief Bioinform*. 2012;13(5):586-614. doi:10.1093/bib/bbs004.

125. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007;4(3):207-214. doi:10.1038/nmeth1019.

126. Gupta N, Pevzner PA. False discovery rates of protein identifications: A strike against the two-peptide rule. *J Proteome Res*. 2009;8(9):4173-4181. doi:10.1021/pr9004794.

127. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*. 2003;75(17):4646-4658. http://www.ncbi.nlm.nih.gov/pubmed/14632076. Accessed May 15, 2018.

128. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*. 2002;74(20):5383-5392. doi:10.1021/ac025747h.

129. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. In: *Electrophoresis*. Vol 20. Wiley-Blackwell; 1999:3551-3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

130. Pappin DJC, Hojrup P, Bleasby AJ. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol*. 1993;3(6):327-332. doi:10.1016/0960-9822(93)90195-T.

131. Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom*. 2003;17(20):2310-2316. doi:10.1002/rcm.1198.

132. Geer LY, Markey SP, Kowalak JA, et al. Open mass spectrometry search algorithm. *J Proteome Res*. 2004;3(5):958-964. doi:10.1021/pr0499491.

133. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25(17):3389-3402. http://www.ncbi.nlm.nih.gov/pubmed/9254694. Accessed June 28, 2018.

134. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen J V., Mann M. Andromeda: A peptide search engine integrated into the MaxQuant environment. *J Proteome Res*. 2011;10(4):1794-1805. doi:10.1021/pr101065j.

135. Zhang J, Xin L, Shan B, et al. PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Mol Cell Proteomics*. 2012;11(4):M111.010587. doi:10.1074/mcp.M111.010587.

136. Eng JK, Jahan TA, Hoopmann MR. Comet: An open-source MS/MS sequence database search tool. *Proteomics*. 2013;13(1):22-24. doi:10.1002/pmic.201200439.

137. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for

proteomics. *Nat Commun*. 2014;5(1):5277. doi:10.1038/ncomms6277.

138. Kim S, Mischerikow N, Bandeira N, et al. The Generating Function of CID, ETD, and CID/ETD Pairs of Tandem Mass Spectra: Applications to Database Search. *Mol Cell Proteomics*. 2010;9(12):2840-2852. doi:10.1074/mcp.M110.003731.

139. Delahunty C, Yates III JR. Protein identification using 2D-LC-MS/MS. *Methods*. 2005;35(3):248-255. doi:10.1016/j.ymeth.2004.08.016.

140. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003;422(6928):198-207. doi:10.1038/nature01511.

141. Becker GW. Stable isotopic labeling of proteins for quantitative proteomic applications. *Briefings Funct Genomics Proteomics*. 2008;7(5):371-382. doi:10.1093/bfgp/eln047.

142. Chahrour O, Cobice D, Malone J. Stable isotope labelling methods in mass spectrometry-based quantitative proteomics. *J Pharm Biomed Anal*. 2015;113:2-20. doi:10.1016/J.JPBA.2015.04.013.

143. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*. 1999;17(10):994-999. doi:10.1038/13690.

144. Schmidt A, Kellermann J, Lottspeich F. A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics*. 2005;5(1):4-15. doi:10.1002/pmic.200400873.

145. Ross PL, Huang YN, Marchese JN, et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*. 2004;3(12):1154-1169. doi:10.1074/mcp.M400129-MCP200.

146. Thompson A, Schäfer J, Kuhn K, et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem*. 2003;75(8):1895-1904. http://www.ncbi.nlm.nih.gov/pubmed/12713048. Accessed May 15, 2018.

147. Koehler CJ, Arntzen MØ, Strozynski M, Treumann A, Thiede B. Isobaric Peptide Termini Labeling Utilizing Site-Specific N-Terminal Succinylation. *Anal Chem*. 2011;83(12):4775-4781. doi:10.1021/ac200229w.

148. Kleifeld O, Doucet A, Prudova A, et al. Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. *Nat Protoc*. 2011;6(10):1578-1611. doi:10.1038/nprot.2011.382.

149. Rashidian M, Dozier JK, Distefano MD. Enzymatic Labeling of Proteins: Techniques and Approaches. *Bioconjug Chem*. 2013;24(8):1277-1294. doi:10.1021/bc400102w.

150. Ong S-E, Blagoev B, Kratchmarova I, et al. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol Cell Proteomics*. 2002;1(5):376-386. doi:10.1074/mcp.M200025-MCP200.

151. Kakihana H, Oi T, Nomura T. Theory of Chromatographic Separation of Isotopes. *J Nucl Sci Technol*. 1977;148:572-581. doi:10.1080/18811248.1977.9730805.

152. Tanaka N, Yamaguchi A, Hashizume K, Araki M, Wada A, Kimata K. Separation of isotopic compounds by reversed-phase liquid chromatography. Effect of pressure gradient on isotope separation by ionization control. *J High Resolut Chromatogr*. 1986;9(11):683-687. doi:10.1002/jhrc.1240091119.

153. Fabre B, Lambour T, Bouyssié D, et al. Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry. *EuPA Open Proteomics*. 2014;4:82-86. doi:10.1016/J.EUPROT.2014.06.001.

154. Wong JWH, Cagney G. An Overview of Label-Free Quantitation Methods in Proteomics by Mass Spectrometry. In: Humana Press; 2010:273-283. doi:10.1007/978-1-60761-444-9_18.

155. Ranjbar MRN, Tadesse MG, Wang Y, Ressom HW. Bayesian Normalization Model for Label-Free Quantitative Analysis by LC-MS. *IEEE/ACM Trans Comput Biol Bioinforma*. 2015;12(4):914-927. doi:10.1109/TCBB.2014.2377723.

156. Välikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform*. 2016;19(1):bbw095. doi:10.1093/bib/bbw095.

157. Karpievitch Y V, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*. 2012;13(Suppl 16):S5. doi:10.1186/1471-2105-13-S16-S5.

158. Nahnsen S, Bielow C, Reinert K, Kohlbacher O. Tools for Label-free Peptide Quantification . *Mol Cell Proteomics*. 2013;12(3):549-556. doi:10.1074/mcp.R112.025163.

159. Arike L, Peil L. Spectral Counting Label-Free Proteomics. In: Humana Press, New York, NY; 2014:213-222. doi:10.1007/978-1-4939-0685-7_14.

160. Zhu W, Smith JW, Huang CM. Mass spectrometry-based label-free quantitative proteomics. *J Biomed Biotechnol*. 2010;2010:840518. doi:10.1155/2010/840518.

161. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol Cell Proteomics*. 2014;13(9):2513-2526. doi:10.1074/mcp.M113.031591.

162. Al Shweiki MHDR, Mönchgesang S, Majovsky P, Thieme D, Trutschel D, Hoehenwarter W. Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. *J Proteome Res*. 2017;16(4):1410-1424. doi:10.1021/acs.jproteome.6b00645.

163. Rappsilber J, Ryder U, Lamond AI, Mann M. Large-scale proteomic analysis of the human

spliceosome. *Genome Res*. 2002;12(8):1231-1245. doi:10.1101/gr.473902.

164. Ishihama Y, Oda Y, Tabata T, et al. 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*. 2005;4(9):1265-1272. doi:10.1074/mcp.M500061-MCP200.

165. Paoletti AC, Parmely TJ, Tomomori-Sato C, et al. Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc Natl Acad Sci U S A*. 2006;103(50):18928-18933. doi:10.1073/pnas.0606379103.

166. Florens L, Carozza MJ, Swanson SK, et al. Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods*. 2006;40(4):303-311. doi:10.1016/J.YMETH.2006.07.028.

167. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*. 2007;25(1):117-124. doi:10.1038/nbt1270.

168. Trudgian DC, Ridlova G, Fischer R, et al. Comparative evaluation of label-free SINQ normalized spectral index quantitation in the central proteomics facilities pipeline. *Proteomics*. 2011;11(14):2790-2797. doi:10.1002/pmic.201000800.

169. Braisted JC, Kuntumalla S, Vogel C, et al. The APEX Quantitative Proteomics Tool: Generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics*. 2008;9(1):529. doi:10.1186/1471-2105-9-529.

170. McIlwain S, Mathews M, Bereman MS, Rubel EW, MacCoss MJ, Noble WS. Estimating relative abundances of proteins from shotgun proteomics data. *BMC Bioinformatics*. 2012;13:308. doi:10.1186/1471-2105-13-308.

171. Silva JC, Gorenstein M V, Li G-Z, Vissers JPC, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics*. 2006;5(1):144-156. doi:10.1074/mcp.M500230-MCP200.

172. Ahrné E, Molzahn L, Glatter T, Schmidt A. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics*. 2013;13(17):2567-2578. doi:10.1002/pmic.201300135.

173. Schwanhausser B, Busse D, Li N, et al. Global quantification of mammalian gene expression control. *Nature*. 2011;473(7347):337-342. http://dx.doi.org/10.1038/nature10098.

174. Schwanhausser B, Busse D, Li N, et al. Corrigendum: Global quantification of mammalian gene expression control. *Nature*. 2013;495(7439):126-127. http://dx.doi.org/10.1038/nature11848.

175. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing mass

spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*. 2006;78(3):779-787. doi:10.1021/ac051437y.

176. Prince JT, Marcotte EM. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal Chem*. 2006;78(17):6140-6152. doi:10.1021/ac0605344.

177. Kramer DA, Eldeeb MA, Wuest M, Mercer J, Fahlman RP. Proteomic characterization of EL4 lymphoma-derived tumors upon chemotherapy treatment reveals potential roles for lysosomes and caspase-6 during tumor cell death in vivo. *Proteomics*. 2017;17(12):1700060. doi:10.1002/pmic.201700060.

178. Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc Natl Acad Sci U S A*. 2007;104(14):5860-5865. doi:10.1073/pnas.0608638104.

179. Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ. Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Mol Cell Proteomics*. 2012;11(11):1475-1488. doi:10.1074/mcp.O112.020131.

180. Weisstein EW. Student's t-Distribution. http://mathworld.wolfram.com/Studentst-Distribution.html. Accessed May 15, 2018.

181. Weisstein EW. ANOVA. http://mathworld.wolfram.com/ANOVA.html. Accessed May 15, 2018.

182. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann Math Stat*. 1947;18(1):50-60. doi:10.1214/aoms/1177730491.

183. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. *J Am Stat Assoc*. 1952;47(260):583-621. doi:10.1080/01621459.1952.10483441.

184. Diz AP, Carvajal-Rodríguez A, Skibinski DOF. Multiple Hypothesis Testing in Proteomics: A Strategy for Experimental Work. *Mol Cell Proteomics*. 2011;10(3):M110.004374. doi:10.1074/mcp.M110.004374.

185. Serang O, Käll L. Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less. *J Proteome Res*. 2015;14(10):4099-4103. doi:10.1021/acs.jproteome.5b00568.

186. Feist P, Hummon A. Proteomic Challenges: Sample Preparation Techniques for Microgram-Quantity Protein Analysis from Biological Samples. *Int J Mol Sci*. 2015;16(2):3537-3563. doi:10.3390/ijms16023537.

187. Statistician William Sealy Gosset (Student). the Probable Error of a Mean. *Biometrika*. 1908;6(1):1-25. doi:10.1093/biomet/6.1.1.

188. Welch BL. The Generalization Of "Student"s' Problem When Several Different Population Variances Are Involved. *Biometrika*. 1947;34(1-2):28-35. doi:10.1093/biomet/34.1-2.28.

189. Delacre M, Lakens D, Leys C. Why Psychologists Should by Default Use Welch's <i>t</i>-test Instead of Student's <i>t</i>-test. *Int Rev Soc Psychol*. 2017;30(1):92. doi:10.5334/irsp.82.

190. Mutch DM, Berger A, Mansourian R, Rytz A, Roberts M-A. The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics*. 2002;3(1):17. doi:10.1186/1471-2105-3-17.

191. Dalman MR, Deeter A, Nimishakavi G, Duan Z-H. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics*. 2012;13(Suppl 2):S11. doi:10.1186/1471-2105-13-S2-S11.

192. Serang O, Cansizoglu AE, Käll L, Steen H, Steen JA. Nonparametric Bayesian evaluation of differential protein quantification. *J Proteome Res*. 2013;12(10):4556-4565. doi:10.1021/pr400678m.

193. Dunn OJ. Multiple Comparisons among Means. *J Am Stat Assoc*. 1961;56(293):52-64. doi:10.1080/01621459.1961.10482090.

194. Bland JM, Altman DG. Multiple significance tests: The Bonferroni method. *BMJ*. 1995;310(6973):170. doi:10.1136/bmj.310.6973.170.

195. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*. 1995;57(1):289-300. doi:10.2307/2346101.

196. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical bayes analysis of a microarray experiment. *J Am Stat Assoc*. 2001;96(456):1151-1160. doi:10.1198/016214501753382129.

197. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*. 2003;19(3):368-375. doi:10.1093/bioinformatics/btf877.

198. Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*. 2006;93(3):491-507. https://pdfs.semanticscholar.org/7155/80a7be4c1945b2ab608bd43dd4f718587643.pdf. Accessed May 15, 2018.

199. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29(4):1165-1188. doi:10.1214/aos/1013699998.

200. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci*. 2003;100(16):9440-9445. doi:10.1073/pnas.1530509100.

201. Storey JD. The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann Stat*. 2003;31(6):2013-2035. doi:10.1214/aos/1074290335.

202. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J R Stat Soc Ser B Stat Methodol*. 2004;66(1):187-205. doi:10.1111/j.1467-

9868.2004.00439.x.

203. Efron B. Local false discovery rates. *Discovery*. 2005:63-79. doi:10.1198/016214507000000941.

204. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556.

205. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res*. 2017;45(D1):D331-D338. doi:10.1093/nar/gkw1108.

206. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 1999;27(1):29-34. doi:10.1093/nar/27.1.29.

207. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. 2016;44(D1):D457-D462. doi:10.1093/nar/gkv1070.

208. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211.

209. Mi H, Thomas P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol*. 2009;563:123-140. doi:10.1007/978-1-60761-175-2_7.

210. Mi H, Huang X, Muruganujan A, et al. PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res*. 2017;45(D1):D183-D189. doi:10.1093/nar/gkw1138.

211. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc*. 2013;8(8):1551-1566. doi:10.1038/nprot.2013.092.

212. Szklarczyk D, Morris JH, Cook H, et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res*. 2017;45(D1):D362-D368. doi:10.1093/nar/gkw937.

213. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303.

214. Bindea G, Galon J, Mlecnik B. CluePedia Cytoscape plugin: Pathway insights using integrated experimental and in silico data. *Bioinformatics*. 2013;29(5):661-663. doi:10.1093/bioinformatics/btt019.

215. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: A network-based method for gene-set enrichment visualization and interpretation. *PLoS One*. 2010;5(11):e13984. doi:10.1371/journal.pone.0013984.

216. Glass K, Girvan M, Huang DW a. W, et al. Annotation Enrichment Analysis: An Alternative Method for Evaluating the Functional Properties of Gene Sets. *Sci Rep*. 2014;4(1):1-9. doi:10.1038/srep04191.

217. Alhamdoosh M, Ng M, Wilson NJ, et al. Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*. 2017;33(3):414-424. doi:10.1093/bioinformatics/btw623.

218. Qin X-J, Ling BX. Proteomic studies in breast cancer (Review). *Oncol Lett*. 2012;3(4):735-743. doi:10.3892/ol.2012.573.

219. Farese R V, Walther TC. Lipid Droplets Finally Get a Little R-E-S-P-E-C-T. *Cell*. 2009;139(5):855-860. doi:10.1016/j.cell.2009.11.005.

220. Murphy DJ. The biogenesis and functions of lipid bodies in animals, plants and microorganisms. *Prog Lipid Res*. 2001;40(5):325-438. doi:10.1016/S0163-7827(01)00013-3.

221. Olofsson SO, Boström P, Andersson L, Rutberg M, Perman J, Borén J. Lipid droplets as dynamic organelles connecting storage and efflux of lipids. *Biochim Biophys Acta - Mol Cell Biol Lipids*. 2009;1791(6):448-458. doi:10.1016/j.bbalip.2008.08.001.

222. Gluchowski NL, Becuwe M, Walther TC, Farese R V. Lipid droplets and liver disease: from basic biology to clinical implications. *Nat Rev Gastroenterol Hepatol*. 2017;14(6):343-355. doi:10.1038/nrgastro.2017.32.

223. Okumura T. Role of lipid droplet proteins in liver steatosis. *J Physiol Biochem*. 2011;67(4):629-636. doi:10.1007/s13105-011-0110-6.

224. Cermelli S, Guo Y, Gross SP, Welte MA. The Lipid-Droplet Proteome Reveals that Droplets Are a Protein-Storage Depot. *Curr Biol*. 2006;16(18):1783-1795. doi:10.1016/j.cub.2006.07.062.

225. Casanovas A, Sprenger RR, Tarasov K, et al. Quantitative analysis of proteome and lipidome dynamics reveals functional regulation of global lipid metabolism. *Chem Biol*. 2015;22(3):412-425. doi:10.1016/j.chembiol.2015.02.007.

226. D'Aquila T, Sirohi D, Grabowski JM, et al. Characterization of the proteome of cytoplasmic lipid droplets in mouse enterocytes after a dietary fat challenge. Blachier F, ed. *PLoS One*. 2015;10(5):e0126823. doi:10.1371/journal.pone.0126823.

227. Murphy S, Martin S, Parton RG. Lipid droplet-organelle interactions; sharing the fats. *Biochim Biophys Acta - Mol Cell Biol Lipids*. 2009;1791(6):441-447. doi:10.1016/j.bbalip.2008.07.004.

228. Gao Q, Goodman JM. The lipid droplet—a well-connected organelle. *Front Cell Dev Biol*. 2015;3:49. doi:10.3389/fcell.2015.00049.

229. Haas JT, Miao J, Chanda D, et al. Hepatic insulin signaling is required for obesity-

dependent expression of SREBP-1c mRNA but not for feeding-dependent expression. *Cell Metab*. 2012;15(6):873-884. doi:10.1016/j.cmet.2012.05.002.

230. Brasaemle DL, Wolins NE. Isolation of lipid droplets from cells by density gradient centrifugation. *Curr Protoc Cell Biol*. 2016;2016(1):3.15.1-3.15.13. doi:10.1002/cpcb.10.

231. Wang H, Quiroga AD, Lehner R. Analysis of lipid droplets in hepatocytes. *Methods Cell Biol*. 2013;116:107-127. doi:10.1016/B978-0-12-408051-5.00007-3.

232. Al-Saikan B, Tredget E, Fahlman R, Ding J, Metcalfe P. Proteomic profile of an acute partial bladder outlet obstruction. *Can Urol Assoc J*. 2015;9(3-4):114. doi:10.5489/cuaj.2267.

233. Vizcaíno JA, Deutsch EW, Wang R, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol*. 2014;32(3):223-226. doi:10.1038/nbt.2839.

234. Vizcaíno JA, Csordas A, Del-Toro N, et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res*. 2016;44(D1):D447-D456. doi:10.1093/nar/gkv1145.

235. Wang H, Gilham D, Lehner R. Proteomic and lipid characterization of apolipoprotein B-free luminal lipid droplets from mouse liver microsomes: Implications for very low density lipoprotein assembly. *J Biol Chem*. 2007;282(45):33218-33226. doi:10.1074/jbc.M706841200.

236. Cui Z, Vance JE, Chen MH, Voelker DR, Vance DE. Cloning and expression of a novel phosphatidylethanolamine N- methyltransferase. A specific biochemical and cytological marker for a unique membrane fraction in rat liver. *J Biol Chem*. 1993;268(22):16655-16663. http://www.ncbi.nlm.nih.gov/pubmed/8344945. Accessed May 10, 2018.

237. Wang H, Wei E, Quiroga AD, Sun X, Touret N, Lehner R. Altered Lipid Droplet Dynamics in Hepatocytes Lacking Triacylglycerol Hydrolase Expression. Parton RG, ed. *Mol Biol Cell*. 2010;21(12):1991-2000. doi:10.1091/mbc.E09-05-0364.

238. Khan SA, Wollaston-Hayden EE, Markowski TW, Higgins L, Mashek DG. Quantitative analysis of the murine lipid droplet-associated proteome during diet-induced hepatic steatosis. *J Lipid Res*. 2015;56(12):2260-2272. doi:10.1194/jlr.M056812.

239. Crunk AE, Monks J, Murakami A, et al. Dynamic Regulation of Hepatic Lipid Droplet Properties by Diet. Davis J, ed. *PLoS One*. 2013;8(7):e67631. doi:10.1371/journal.pone.0067631.

240. Ellis JM, Mentock SM, DePetrillo MA, et al. Mouse Cardiac Acyl Coenzyme A Synthetase 1 Deficiency Impairs Fatty Acid Oxidation and Induces Cardiac Hypertrophy. *Mol Cell Biol*. 2011;31(6):1252-1262. doi:10.1128/MCB.01085-10.

241. Ellis JM, Li LO, Wu PC, et al. Adipose Acyl-CoA synthetase-1 directs fatty acids toward β-oxidation and is required for cold thermogenesis. *Cell Metab*. 2010;12(1):53-64.

doi:10.1016/j.cmet.2010.05.012.

242. Li H, Song Y, Zhang LJ, et al. LSDP5 enhances triglyceride storage in hepatocytes by influencing lipolysis and fatty acid β-oxidation of lipid droplets. Ko BCB, ed. *PLoS One*. 2012;7(6):e36712. doi:10.1371/journal.pone.0036712.

243. Conte M, Franceschi C, Sandri M, Salvioli S. Perilipin 2 and Age-Related Metabolic Diseases: A New Perspective. *Trends Endocrinol Metab*. 2016;27(12):893-903. doi:10.1016/j.tem.2016.09.001.

244. Zhang H, Wang Y, Li J, et al. Proteome of skeletal muscle lipid droplet reveals association with mitochondria and apolipoprotein A-I. *J Proteome Res*. 2011;10(10):4757-4768. doi:10.1021/pr200553c.

245. Ploegh HL. A lipid-based model for the creation of an escape hatch from the endoplasmic reticulum. *Nature*. 2007;448(7152):435-438. doi:10.1038/nature06004.

246. Mishra S, Khaddaj R, Cottier S, Stradalova V, Jacob C, Schneiter R. Mature lipid droplets are accessible to ER luminal proteins. *J Cell Sci*. 2016;129(20):3803-3815. doi:10.1242/jcs.189191.

247. Dolinsky VW, Gilham D, Alam M, Vance DE, Lehner R. Triacylglycerol hydrolase: role in intracellular lipid metabolism. *Cell Mol Life Sci*. 2004;61(13):1633-1651. doi:10.1007/s00018-004-3426-3.

248. Gilham D, Alam M, Gao W, Vance DE, Lehner R. Triacylglycerol Hydrolase Is Localized to the Endoplasmic Reticulum by an Unusual Retrieval Sequence where It Participates in VLDL Assembly without Utilizing VLDL Lipids as Substrates. *Mol Biol Cell*. 2005;16(2):984-996. doi:10.1091/mbc.e04-03-0224.

249. Ko KWS, Erickson B, Lehner R. Es-x/Ces1 prevents triacylglycerol accumulation in McArdle-RH7777 hepatocytes. *Biochim Biophys Acta - Mol Cell Biol Lipids*. 2009;1791(12):1133-1143. doi:10.1016/j.bbalip.2009.07.006.

250. Quiroga AD, Li L, Trötzmüller M, et al. Deficiency of carboxylesterase 1/esterase-x results in obesity, hepatic steatosis, and hyperlipidemia. *Hepatology*. 2012;56(6):2188-2198. doi:10.1002/hep.25961.

251. Binns D, Januszewski T, Chen Y, et al. An intimate collaboration between peroxisomes and lipid bodies. *J Cell Biol*. 2006;173(5):719-731. doi:10.1083/jcb.200511125.

252. Lodhi IJ, Semenkovich CF. Peroxisomes: A nexus for lipid metabolism and cellular signaling. *Cell Metab*. 2014;19(3):380-392. doi:10.1016/j.cmet.2014.01.002.

253. Beller M, Riedel D, Jänsch L, et al. Characterization of the Drosophila Lipid Droplet Subproteome. *Mol Cell Proteomics*. 2006;5(6):1082-1094. doi:10.1074/mcp.M600011-MCP200.

254. Wan H-C, Melo RCN, Jin Z, Dvorak AM, Weller PF. Roles and origins of leukocyte lipid

bodies: proteomic and ultrastructural studies. *FASEB J*. 2007;21(1):167-178. doi:10.1096/fj.06-6711com.

255. Ohsaki Y, Suzuki M, Fujimoto T. Open questions in lipid droplet biology. *Chem Biol*. 2014;21(1):86-96. doi:10.1016/j.chembiol.2013.08.009.

256. Liu M, Ge R, Liu W, et al. Differential proteomics profiling identifies LDPs and biological functions in high-fat diet-induced fatty livers. *J Lipid Res*. 2017;58(4):681-694. doi:10.1194/jlr.M071407.

257. Baumeier C, Kaiser D, Heeren J, et al. Caloric restriction and intermittent fasting alter hepatic lipid droplet proteome and diacylglycerol species and prevent diabetes in NZO mice. *Biochim Biophys Acta - Mol Cell Biol Lipids*. 2015;1851(5):566-576. doi:10.1016/j.bbalip.2015.01.013.

258. Brasaemle DL, Dolios G, Shapiro L, Wang R. Proteomic analysis of proteins associated with lipid droplets of basal and lipolytically stimulated 3T3-L1 adipocytes. *J Biol Chem*. 2004;279(45):46835-46842. doi:10.1074/jbc.M409340200.

259. Fujimoto Y, Itabe H, Sakai J, et al. Identification of major proteins in the lipid droplet-enriched fraction isolated from the human hepatocyte cell line HuH7. *Biochim Biophys Acta - Mol Cell Res*. 2004;1644(1):47-59. doi:10.1016/j.bbamcr.2003.10.018.

260. Bartz R, Li W-H, Venables B, et al. Lipidomics reveals that adiposomes store ether lipids and mediate phospholipid traffic ,. *J Lipid Res*. 2007;48(4):837-847. doi:10.1194/jlr.M600413-JLR200.

261. Liu P, Bartz R, Zehmer JK, et al. Rab-regulated interaction of early endosomes with lipid droplets. *Biochim Biophys Acta - Mol Cell Res*. 2007;1773(6):784-793. doi:10.1016/j.bbamcr.2007.02.004.

262. Martin S, Driessen K, Nixon SJ, Zerial M, Parton RG. Regulated localization of Rab18 to lipid droplets: Effects of lipolytic stimulation and inhibition of lipid droplet catabolism. *J Biol Chem*. 2005;280(51):42325-42335. doi:10.1074/jbc.M506651200.

263. Turró S, Ingelmo-Torres M, Estanyol JM, et al. Identification and characterization of associated with lipid droplet protein 1: A novel membrane-associated protein that resides on hepatic lipid droplets. *Traffic*. 2006;7(9):1254-1269. doi:10.1111/j.1600-0854.2006.00465.x.

264. Ohsaki Y, Cheng J, Suzuki M, Shinohara Y, Fujita A, Fujimoto T. Biogenesis of cytoplasmic lipid droplets: From the lipid ester globule in the membrane to the visible structure. *Biochim Biophys Acta - Mol Cell Biol Lipids*. 2009;1791(6):399-407. doi:10.1016/j.bbalip.2008.10.002.

265. Ozeki S, Cheng J, Tauchi-Sato K, et al. Rab18 localizes to lipid droplets and induces their close apposition to the endoplasmic reticulum-derived membrane. *J Cell Sci*. 2005;118(Pt 12):2601-2611. doi:10.1242/jcs.02401.

266. Liu P, Ying Y, Zhao Y, Mundy DI, Zhu M, Anderson RGW. Chinese Hamster Ovary K2 Cell Lipid Droplets Appear to be Metabolic Organelles Involved in Membrane Traffic. *J Biol Chem*. 2004;279(5):3787-3792. doi:10.1074/jbc.M311945200.

267. Bouchoux J, Beilstein F, Pauquai T, et al. The proteome of cytosolic lipid droplets isolated from differentiated Caco-2/TC7 enterocytes reveals cell-specific characteristics. *Biol Cell*. 2011;103(11):499-517. doi:10.1042/BC20110024.

268. Zhou Y, Rui L. *Major Urinary Protein Regulation of Chemical Communication and Nutrient Metabolism*. Vol 83.; 2010. doi:10.1016/S0083-6729(10)83006-7.

269. Athenstaedt K, Zweytick D, Jandrositz A, Kohlwein SD, Daum G. Identification and characterization of major lipid particle proteins of the yeast Saccharomyces cerevisiae. *J Bacteriol*. 1999;181(20):6441-6448. http://www.ncbi.nlm.nih.gov/pubmed/10515935. Accessed May 11, 2018.

270. Wang W, Wei S, Li L, et al. Proteomic analysis of murine testes lipid droplets. *Sci Rep*. 2015;5(1):12070. doi:10.1038/srep12070.

271. Ferdinandusse S, Denis S, Faust PL, Wanders RJA. Bile acids: the role of peroxisomes. *J Lipid Res*. 2009;50(11):2139-2147. doi:10.1194/jlr.R900009-JLR200.

272. Kuramoto K, Okamura T, Yamaguchi T, et al. Perilipin 5, a lipid droplet-binding protein, protects heart from oxidative burden by sequestering fatty acid from excessive oxidation. *J Biol Chem*. 2012;287(28):23852-23863. doi:10.1074/jbc.M111.328708.

273. Gorer PA. Studies in antibody response of mice to tumour inoculation. *Br J Cancer*. 1950;4(4):372-379. doi:10.1038/bjc.1950.36.

274. Sakurai H, Mitsuhashi N, Murata O, et al. Early radiation effects in highly apoptotic murine lymphoma xenografts monitored by31P magnetic resonance spectroscopy. *Int J Radiat Oncol Biol Phys*. 1998;41(5):1157-1162. doi:10.1016/S0360-3016(98)00158-8.

275. Zhao M, Beauregard DA, Loizou L, Davletov B, Brindle KM. Non-invasive detection of apoptosis using magnetic resonance imaging and a targeted contrast agent. *Nat Med*. 2001;7(11):1241-1244. doi:10.1038/nm1101-1241.

276. Krawczyk CM, Verstovšek S, Ujházy P, Maccubbin D, Ehrke MJ. Protective specific immunity induced by cyclophosphamide plus tumor necrosis factor α combination treatment of EL4-lymphoma-bearing C57BL/6 mice. *Cancer Immunol Immunother*. 1995;40(6):347-357. doi:10.1007/BF01525385.

277. Yoshizaki Y, Yuba E, Komatsu T, Udaka K, Harada A, Kono K. Improvement of peptide-based tumor immunotherapy using pH-sensitive fusogenic polymer-modified liposomes. *Molecules*. 2016;21(10):1284. doi:10.3390/molecules21101284.

278. Yamazaki T, Pitt JM, Vétizou M, et al. The oncolytic peptide LTX-315 overcomes resistance of cancers to immunotherapy with CTLA4 checkpoint blockade. *Cell Death*

Differ. 2016;23(6):1004-1015. doi:10.1038/cdd.2016.35.

279. Yoshimoto Y, Suzuki Y, Mimura K, et al. Radiotherapy-induced anti-tumor immunity contributes to the therapeutic efficacy of irradiation and can be augmented by CTLA-4 blockade in a mouse model. Shiku H, ed. *PLoS One*. 2014;9(3):e92572. doi:10.1371/journal.pone.0092572.

280. Noh KT, Son KH, Jung ID, Kang TH, Choi CH, Park YM. Glycogen synthase kinase-3β (GSK-3β) inhibition enhances dendritic cell-based cancer vaccine potency via suppression of interferon-γ-induced indoleamine 2,3-dioxygenase expression. *J Biol Chem*. 2015;290(19):12394-12402. doi:10.1074/jbc.M114.628578.

281. Wuest M, Perreault A, Kapty J, et al. Radiopharmacological evaluation of (18)F-labeled phosphatidylserine-binding peptides for molecular imaging of apoptosis. *Nucl Med Biol*. 2015;42(11):864-874. doi:10.1016/j.nucmedbio.2015.06.011.

282. Al-Ejeh F, Darby JM, Tsopelas C, Smyth D, Manavis J, Brown MP. APOMAB®, a La-specific monoclonal antibody, detects the apoptotic tumor response to life-prolonging and DNA-damaging chemotherapy. Boswell A, ed. *PLoS One*. 2009;4(2):e4558. doi:10.1371/journal.pone.0004558.

283. Chen DL, Engle JT, Griffin EA, et al. Imaging Caspase-3 Activation as a Marker of Apoptosis-Targeted Treatment Response in Cancer. *Mol Imaging Biol*. 2015;17(3):384-393. doi:10.1007/s11307-014-0802-8.

284. Al-Ejeh F, Darby JM, Pensa K, Diener KR, Hayball JD, Brown MP. In vivo targeting of dead tumor cells in a murine tumor model using a monoclonal antibody specific for the La autoantigen. *Clin Cancer Res*. 2007;13(18 Pt 2):5519s-5527s. doi:10.1158/1078-0432.CCR-07-0964.

285. Perreault A, Richter S, Bergman C, Wuest M, Wuest F. Targeting Phosphatidylserine with a64Cu-Labeled Peptide for Molecular Imaging of Apoptosis. *Mol Pharm*. 2016;13(10):3564-3577. doi:10.1021/acs.molpharmaceut.6b00666.

286. Keating JJ, Nims S, Venegas O, et al. Intraoperative imaging identifies thymoma margins following neoadjuvant chemotherapy. *Oncotarget*. 2016;7(3):3059-3067. doi:10.18632/oncotarget.6578.

287. Lai CP, Kim EY, Badr CE, et al. Visualization and tracking of tumour extracellular vesicle delivery and RNA translation using multiplexed reporters. *Nat Commun*. 2015;6(1):7029. doi:10.1038/ncomms8029.

288. Khan SR, Baghdasarian A, Fahlman RP, Michail K, Siraki AG. Current status and future prospects of toxicogenomics in drug discovery. *Drug Discov Today*. 2014;19(5):562-578. doi:10.1016/j.drudis.2013.11.001.

289. Matondo M, Marcellin M, Chaoui K, et al. Determination of differentially regulated proteins upon proteasome inhibition in AML cell lines by the combination of large-scale

and targeted quantitative proteomics. *Proteomics*. 2017;17(7):1600089. doi:10.1002/pmic.201600089.

290. Paul D, Chanukuppa V, Reddy PJ, et al. Global proteomic profiling identifies etoposide chemoresistance markers in non-small cell lung carcinoma. *J Proteomics*. 2016;138:95-105. doi:10.1016/j.jprot.2016.02.008.

291. Komohara Y, Takemura K, Lei XF, et al. Delayed growth of EL4 lymphoma in SR-A-deficient mice is due to upregulation of nitric oxide and interferon-γ production by tumor-associated macrophages. *Cancer Sci*. 2009;100(11):2160-2166. doi:10.1111/j.1349-7006.2009.01296.x.

292. Ruan J, Luo M, Wang C, et al. Imatinib disrupts lymphoma angiogenesis by targeting vascular pericytes. *Blood*. 2013;121(6):5192-5202. doi:10.1182/blood-2013-03-490763.

293. Elvington M, Scheiber M, Yang X, et al. Complement-dependent modulation of antitumor immunity following radiation therapy. *Cell Rep*. 2014;8(3):818-830. doi:10.1016/j.celrep.2014.06.051.

294. Khan SR, Baghdasarian A, Nagar PH, et al. Proteomic profile of aminoglutethimide-induced apoptosis in HL-60 cells: Role of myeloperoxidase and arylamine free radicals. *Chem Biol Interact*. 2015;239:129-138. doi:10.1016/j.cbi.2015.06.020.

295. Storey JD. A direct approach to false discovery rates. *J R Stat Soc B*. 2002;64(3):479-498. http://genomics.princeton.edu/storeylab/papers/directfdr.pdf. Accessed March 6, 2018.

296. Takeshima T, Pop LM, Laine A, Iyengar P, Vitetta ES, Hannan R. Key role for neutrophils in radiation-induced antitumor immune responses: Potentiation with G-CSF. *Proc Natl Acad Sci*. 2016;113(40):11300-11305. doi:10.1073/pnas.1613187113.

297. Marion RM, Regev A, Segal E, et al. Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc Natl Acad Sci*. 2004;101(40):14315-14322. doi:10.1073/pnas.0405353101.

298. Guerra-Moreno A, Isasa M, Bhanu MK, et al. Proteomic analysis identifies ribosome reduction as an effective proteotoxic stress response. *J Biol Chem*. 2015;290(50):29695-29706. doi:10.1074/jbc.M115.684969.

299. Warner JR. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci*. 1999;24(11):437-440. doi:10.1016/S0968-0004(99)01460-7.

300. Zhou X, Liao WJ, Liao JM, Liao P, Lu H. Ribosomal proteins: Functions beyond the ribosome. *J Mol Cell Biol*. 2015;7(2):92-104. doi:10.1093/jmcb/mjv014.

301. Dave B, Granados-Principal S, Zhu R, et al. Targeting RPL39 and MLF2 reduces tumor initiation and metastasis in breast cancer by inhibiting nitric oxide synthase signaling. *Proc Natl Acad Sci*. 2014;111(24):8838-8843. doi:10.1073/pnas.1320769111.

302. Dave B, Gonzalez DD, Liu Z Bin, et al. Role of RPL39 in Metaplastic Breast Cancer. *J Natl*

Cancer Inst. 2017;109(6):djw292. doi:10.1093/jnci/djw292.

303. Doherty L, Sheen MR, Vlachos A, et al. Ribosomal Protein Genes RPS10 and RPS26 Are Commonly Mutated in Diamond-Blackfan Anemia. *Am J Hum Genet*. 2010;86(2):222-228. doi:10.1016/j.ajhg.2009.12.015.

304. Fahlman RP, Chen W, Overall CM. Absolute proteomic quantification of the activity state of proteases and proteolytic cleavages using proteolytic signature peptides and isobaric tags. *J Proteomics*. 2014;100:79-91. doi:10.1016/j.jprot.2013.09.006.

305. Poon IKH, Lucas CD, Rossi AG, Ravichandran KS. Apoptotic cell clearance: Basic biology and therapeutic potential. *Nat Rev Immunol*. 2014;14(3):166-180. doi:10.1038/nri3607.

306. Tawa P, Hell K, Giroux A, et al. Catalytic activity of caspase-3 is required for its degradation: Stabilization of the active complex by synthetic inhibitors. *Cell Death Differ*. 2004;11(4):439-447. doi:10.1038/sj.cdd.4401360.

307. Luna-Vargas MPA, Chipuk JE. The deadly landscape of pro-apoptotic BCL-2 proteins in the outer mitochondrial membrane. *FEBS J*. 2016;283(14):2676-2689. doi:10.1111/febs.13624.

308. Repnik U, Stoka V, Turk V, Turk B. Lysosomes and lysosomal cathepsins in cell death. *Biochim Biophys Acta - Proteins Proteomics*. 2012;1824(1):22-33. doi:10.1016/j.bbapap.2011.08.016.

309. Kaminskyy V, Zhivotovsky B. Proteases in autophagy. *Biochim Biophys Acta - Proteins Proteomics*. 2012;1824(1):44-50. doi:10.1016/j.bbapap.2011.05.013.

310. Droga-Mazovec G, Bojič L, Petelin A, et al. Cysteine cathepsins trigger caspase-dependent cell death through cleavage of bid and antiapoptotic Bcl-2 homologues. *J Biol Chem*. 2008;283(27):19140-19150. doi:10.1074/jbc.M802513200.

311. Vancompernolle K, Van Herreweghe F, Pynaert G, et al. Atractyloside-induced release of cathepsin B, a protease with caspase-processing activity. *FEBS Lett*. 1998;438(3):150-158. doi:10.1016/S0014-5793(98)01275-7.

312. Zhou XY, Luo Y, Zhu YM, et al. Inhibition of autophagy blocks cathepsins-tBid-mitochondrial apoptotic signaling pathway via stabilization of lysosomal membrane in ischemic astrocytes. *Cell Death Dis*. 2017;8(2):e2618. doi:10.1038/cddis.2017.34.

313. Lin L, Baehrecke EH. Autophagy, cell death, and cancer. *Mol Cell Oncol*. 2015;2(3):e985913. doi:10.4161/23723556.2014.985913.

314. Ojha R, Ishaq M, Singh SK. Caspase-mediated crosstalk between autophagy and apoptosis: Mutual adjustment or matter of dominance. *J Cancer Res Ther*. 2015;11(3):514-524. doi:10.4103/0973-1482.163695.

315. Scott NE, Rogers LD, Prudova A, et al. Interactome disassembly during apoptosis occurs independent of caspase cleavage. *Mol Syst Biol*. 2017;13(1):906.

doi:10.15252/msb.20167067.

316. Canadian Cancer Statistics. http://www.cancer.ca/~/media/cancer.ca/CW/cancer information/cancer 101/Canadian cancer statistics/Canadian-Cancer-Statistics-2017-EN.pdf?la=en. Accessed May 13, 2018.

317. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet*. 2005;365(9472):1687-1717. doi:10.1016/S0140-6736(05)66544-0.

318. Coates AS, Winer EP, Goldhirsch A, et al. Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann Oncol*. 2015;26(8):1533-1546. doi:10.1093/annonc/mdv221.

319. Voduc KD, Cheang MCU, Tyldesley S, Gelmon K, Nielsen TO, Kennecke H. Breast Cancer Subtypes and the Risk of Local and Regional Relapse. *J Clin Oncol*. 2010;28(10):1684-1691. doi:10.1200/JCO.2009.24.9284.

320. Dowsett M, Nielsen TO, A'Hern R, et al. Assessment of Ki67 in Breast Cancer: Recommendations from the international Ki67 in breast cancer working Group. *J Natl Cancer Inst*. 2011;103(22):1656-1664. doi:10.1093/jnci/djr393.

321. Feeley LP, Mulligan AM, Pinnaduwage D, Bull SB, Andrulis IL. Distinguishing luminal breast cancer subtypes by Ki67, progesterone receptor or TP53 status provides prognostic information. *Mod Pathol*. 2014;27(4):554-561. doi:10.1038/modpathol.2013.153.

322. Lee SK, Bae SY, Lee JH, et al. Distinguishing Low-Risk Luminal A Breast Cancer Subtypes with Ki-67 and p53 Is More Predictive of Long-Term Survival. Ahmad A, ed. *PLoS One*. 2015;10(8):e0124658. doi:10.1371/journal.pone.0124658.

323. Harris LN, Ismaila N, McShane LM, et al. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline. *J Clin Oncol*. 2016;34(10):1134-1150. doi:10.1200/JCO.2015.65.2289.

324. Krop I, Ismaila N, Andre F, et al. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update. *J Clin Oncol*. 2017;35(24):2838-2847. doi:10.1200/JCO.2017.74.0472.

325. Webb-Robertson B-JM, Wiberg HK, Matzke MM, et al. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res*. 2015;14(5):1993-2001. doi:10.1021/pr501138h.

326. Lazar C, Gatto L, Ferro M, Bruley C, Burger T. Accounting for the Multiple Natures of

Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J Proteome Res*. 2016;15(4):1116-1125. doi:10.1021/acs.jproteome.5b00981.

327. Tukey JW. Comparing Individual Means in the Analysis of Variance. *Biometrics*. 1949;5(2):99. doi:10.2307/3001913.

328. Kramer CY. Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications. *Biometrics*. 1956;12(3):307. doi:10.2307/3001469.

329. Driscoll WC. Robustness of the ANOVA and Tukey-Kramer statistical tests. *Comput Ind Eng*. 1996;31(1-2):265-268. doi:10.1016/0360-8352(96)00127-1.

330. Caraux G, Pinloche S. PermutMatrix: A graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics*. 2005;21(7):1280-1281. doi:10.1093/bioinformatics/bti141.

331. Tyanova S, Temu T, Sinitcyn P, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods*. 2016;13(9):731-740. doi:10.1038/nmeth.3901.

332. Lánczky A, Nagy Á, Bottai G, et al. miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. *Breast Cancer Res Treat*. 2016;160(3):439-446. doi:10.1007/s10549-016-4013-7.

333. Győrffy B, Lanczky A, Eklund AC, et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat*. 2010;123(3):725-731. doi:10.1007/s10549-009-0674-9.

334. Li Q, Birkbak NJ, Gyorffy B, Szallasi Z, Eklund AC. Jetset: Selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics*. 2011;12. doi:10.1186/1471-2105-12-474.

335. Blainey P, Krzywinski M, Altman N. Points of Significance: Replication. *Nat Methods 2014 119*. August 2014.

336. Gündisch S, Hauck S, Sarioglu H, et al. Variability of Protein and Phosphoprotein Levels in Clinical Tissue Specimens during the Preanalytical Phase. *J Proteome Res*. 2012;11(12):5748-5762. doi:10.1021/pr300560y.

337. Tripathi NK, Everds NE, Schultze AE, et al. Deciphering Sources of Variability in Clinical Pathology. In: *Toxicologic Pathology*. Vol 45. ; 2017:90-93. doi:10.1177/0192623316675766.

338. Turnbull AK, Arthur LM, Renshaw L, et al. Accurate prediction and validation of response to endocrine therapy in breast cancer. *J Clin Oncol*. 2015;33(20):2270-2278. doi:10.1200/JCO.2014.57.8963.

339. Molloy MP, Brzezinski EE, Hang J, McDowell MT, VanBogelen RA. Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics*.

2003;3(10):1912-1919. doi:10.1002/pmic.200300534.

340. Karp NA, Lilley KS. Design and analysis issues in quantitative proteomics studies. *Proteomics*. 2007;7 Suppl 1(S1):42-50. doi:10.1002/pmic.200700683.

341. Piehowski PD, Petyuk VA, Orton DJ, et al. Sources of technical variability in quantitative LC-MS proteomics: Human brain tissue sample analysis. *J Proteome Res*. 2013;12(5):2128-2137. doi:10.1021/pr301146m.

342. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet*. 2014;15(5):335-346. doi:10.1038/nrg3706.

343. Wiśniewski JR, Hein MY, Cox J, Mann M. A "Proteomic Ruler" for Protein Copy Number and Concentration Estimation without Spike-in Standards. *Mol Cell Proteomics*. 2014;13(12):3497-3506. doi:10.1074/mcp.M113.037309.

344. Shackney SE, Singh SG, Yakulis R, et al. Aneuploidy i n Breast Cancer: A Fluorescence In Situ Hybridization Study. *Cytometry*. https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.990220404. Accessed May 13, 2018.

345. Arpino G, Weiss H, Lee A V., et al. Estrogen receptor-positive, progesterone receptor-negative breast cancer: Association with growth factor receptor expression and tamoxifen resistance. *J Natl Cancer Inst*. 2005;97(17):1254-1261. doi:10.1093/jnci/dji249.

346. Xu J, Huang L, Li J. DNA aneuploidy and breast cancer: a meta-analysis of 141,163 cases. *Oncotarget*. 2016;7(37):60218-60229. doi:10.18632/oncotarget.11130.

347. Sullivan S. The Histone Database. *Nucleic Acids Res*. 2002;30(1):341-342. doi:10.1093/nar/30.1.341.

348. Ziegel ER, Lehmann EL. Elements of Large-Sample Theory. In: *Technometrics*. Vol 42. ; 2000:176. doi:10.2307/1271493.

349. Stead DA, Paton NW, Missier P, et al. Information quality in proteomics. *Brief Bioinform*. 2007;9(2):174-188. doi:10.1093/bib/bbn004.

350. Chawade A, Alexandersson E, Levander F. Normalyzer: A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets. *J Proteome Res*. 2014;13(6):3114-3120. doi:10.1021/pr401264n.

351. Webb-Robertson B-JM, Wiberg HK, Matzke MM, et al. Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *J Proteome Res*. 2015;14(5):1993-2001. doi:10.1021/pr501138h.

352. Seyfried NT, Dammer EB, Swarup V, et al. A Multi-network Approach Identifies Protein-Specific Co-expression in Asymptomatic and Symptomatic Alzheimer's Disease. *Cell Syst*. 2017;4(1):60-72.e4. doi:10.1016/j.cels.2016.11.006.

353. Umoh ME, Dammer EB, Dai J, et al. A proteomic network approach across the ALS-FTD disease spectrum resolves clinical phenotypes and genetic vulnerability in human brain. *EMBO Mol Med*. 2018;10(1):48-62. doi:10.15252/emmm.201708202.

354. Matos LL de, Trufelli DC, de Matos MGL, da Silva Pinhal MA. Immunohistochemistry as an important tool in biomarkers detection and clinical practice. *Biomark Insights*. 2010;5:9-20. http://www.ncbi.nlm.nih.gov/pubmed/20212918. Accessed May 13, 2018.

355. Schroeder HW, Cavacini L. Structure and function of immunoglobulins. *J Allergy Clin Immunol*. 2010;125(2):S41-S52. doi:10.1016/j.jaci.2009.09.046.

356. McHeyzer-Williams M, Okitsu S, Wang N, McHeyzer-Williams L. Molecular programming of B cell memory. *Nat Rev Immunol*. 2012;12(1):24-34. doi:10.1038/nri3128.

357. Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett*. 2009;583(24):3966-3973. doi:10.1016/j.febslet.2009.10.036.

358. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*. 2016;165(3):535-550. doi:10.1016/j.cell.2016.03.014.

359. Mizoguchi A, Mizuochi T, Kobata A. Structures of the carbohydrate moieties of secretory component purified from human milk. *J Biol Chem*. 1982;257(16):9612-9621. http://www.ncbi.nlm.nih.gov/pubmed/7107583. Accessed June 14, 2018.

360. Mostov KE, Blobel G. The transmembrane precursor of secretory component. The receptor for transcellular transport of polymeric immunoglobulins. *J Biol Chem*. 1982;257(19):11816-11821. http://www.ncbi.nlm.nih.gov/pubmed/7118912. Accessed June 14, 2018.

361. Zhang J-R, Mostov KE, Lamm ME, et al. The Polymeric Immunoglobulin Receptor Translocates Pneumococci across Human Nasopharyngeal Epithelial Cells. *Cell*. 2000;102(6):827-837. doi:10.1016/S0092-8674(00)00071-4.

362. Johansen F-E, Kaetzel CS. Regulation of the polymeric immunoglobulin receptor and IgA transport: new advances in environmental factors that stimulate pIgR expression and its role in mucosal immunity. *Mucosal Immunol*. 2011;4(6):598-602. doi:10.1038/mi.2011.37.

363. Emmerson CD, van der Vlist EJ, Braam MR, et al. Enhancement of Polymeric Immunoglobulin Receptor Transcytosis by Biparatopic VHH. Bansal GP, ed. *PLoS One*. 2011;6(10):e26299. doi:10.1371/journal.pone.0026299.

364. Kaetzel CS, Robinson JK, Chintalacharuvu KR, Vaerman JP, Lamm ME. The polymeric immunoglobulin receptor (secretory component) mediates transport of immune complexes across epithelial cells: a local defense function for IgA. *Proc Natl Acad Sci U S A*. 1991;88(19):8796-8800. doi:10.1073/pnas.89.2.792.

365. Brandtzaeg P, Prydz H. Direct evidence for an integrated function of J chain and secretory component in epithelial transport of immunoglobulins. *Nature*. 1984;311(5981):71-73. doi:10.1038/311071a0.

366. Johansen, Braathen, Brandtzaeg. Role of J chain in secretory immunoglobulin formation. *Scand J Immunol*. 2000;52(3):240-248. doi:10.1046/j.1365-3083.2000.00790.x.

367. Kumar M, Joseph SR, Augsburg M, et al. MS Western, a Method of Multiplexed Absolute Protein Quantification is a Practical Alternative to Western Blotting. *Mol Cell Proteomics*. 2018;17(2):384-396. doi:10.1074/mcp.O117.067082.

368. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007;115(5):654-657. doi:10.1161/CIRCULATIONAHA.105.594929.

369. Pletcher MJ, Pignone M. Evaluating the clinical utility of a biomarker: a review of methods for estimating health impact. *Circulation*. 2011;123(10):1116-1124. doi:10.1161/CIRCULATIONAHA.110.943860.

370. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*. 1988;44(3):837. doi:10.2307/2531595.

371. Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochem medica*. 2016;26(3):297-307. doi:10.11613/BM.2016.034.

372. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32-35. doi:10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

373. Perkins NJ, Schisterman EF. The Inconsistency of "Optimal" Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve. *Am J Epidemiol*. 2006;163(7):670-675. doi:10.1093/aje/kwj063.

374. Rota M, Antolini L, Valsecchi MG. Optimal cut-point definition in biomarkers: The case of censored failure time outcome. *BMC Med Res Methodol*. 2015;15(1):24. doi:10.1186/s12874-015-0009-y.

375. Santos CR, Schulze A. Lipid metabolism in cancer. *FEBS J*. 2012;279(15):2610-2623. doi:10.1111/j.1742-4658.2012.08644.x.

376. Currie E, Schulze A, Zechner R, Walther TC, Farese R V. Cellular fatty acid metabolism and cancer. *Cell Metab*. 2013;18(2):153-161. doi:10.1016/j.cmet.2013.05.017.

377. Röhrig F, Schulze A. The multifaceted roles of fatty acid synthesis in cancer. *Nat Rev Cancer*. 2016;16(11):732-749. doi:10.1038/nrc.2016.89.

378. Roberts RA, Chevalier S, Hasmall SC, James NH, Cosulich SC, Macdonald N. PPAR alpha and the regulation of cell division and apoptosis. *Toxicology*. 2002;181-182:167-170.

http://www.ncbi.nlm.nih.gov/pubmed/12505304. Accessed May 13, 2018.

379. Tachibana K, Yamasaki D, Ishimoto K, Doi T. The role of PPARs in cancer. *PPAR Res*. 2008;2008:102737. doi:10.1155/2008/102737.

380. Youssef J, Badr M. Peroxisome proliferator-activated receptors and cancer: challenges and opportunities. *Br J Pharmacol*. 2011;164(1):68-82. doi:10.1111/j.1476-5381.2011.01383.x.

381. Tasaki T, Mulder LCF, Iwamatsu A, et al. A Family of Mammalian E3 Ubiquitin Ligases That Contain the UBR Box Motif and Recognize N-Degrons. *Mol Cell Biol*. 2005;25(16):7120-7136. doi:10.1128/MCB.25.16.7120-7136.2005.

382. Eldeeb M, Fahlman R. The-N-End Rule : The Beginning Determines the End. *Protein Pept Lett*. 2016;23(4):1-6. doi:10.2174/0929866523666160108115809.

383. Eldeeb MA, Leitao LCA FR. Emerging Branches of the N-End Rule Pathways are Revealing the Sequence Complexities of N-Termini Dependent Protein Degradation. *Biochem Cell Biol*. December 2017:1-6. doi:10.1139/bcb-2017-0274.

384. Eldeeb MA, Fahlman RP. The anti-apoptotic form of tyrosine kinase Lyn that is generated by proteolysis is degraded by the N-end rule pathway. *Oncotarget*. 2014;5(9):2714-2722. http://europepmc.org/abstract/MED/24798867.

385. Eldeeb MA, Fahlman RP. Phosphorylation Impacts N-end rule degradation of the proteolytically activated form of BMX kinase. *J Biol Chem*. 2016;291(43):22757-22768. doi:10.1074/jbc.M116.737387.

386. Xu Z, Payoe R, Fahlman RP. The C-terminal Proteolytic Fragment of the Breast Cancer Susceptibility Type 1 Protein (BRCA1) Is Degraded by the N-end Rule Pathway. *J Biol Chem*. 2012;287(10):7495-7502. doi:10.1074/jbc.M111.301002.

387. Elmore S. Apoptosis: A Review of Programmed Cell Death. *Toxicol Pathol*. 2007;35(4):495-516. doi:10.1080/01926230701320337.

388. Ouyang L, Shi Z, Zhao S, et al. Programmed cell death pathways in cancer: A review of apoptosis, autophagy and programmed necrosis. *Cell Prolif*. 2012;45(6):487-498. doi:10.1111/j.1365-2184.2012.00845.x.

389. Hoesel B, Schmid JA. The complexity of NF-κB signaling in inflammation and cancer. *Mol Cancer*. 2013;12:86. doi:10.1186/1476-4598-12-86.

390. Sapio L, Di Maiolo F, Illiano M, et al. Targeting protein kinase A in cancer therapy: an update. *EXCLI J*. 2014;13:843-855. http://www.ncbi.nlm.nih.gov/pubmed/26417307. Accessed May 14, 2018.

391. Gagarina V, Carlberg AL, Pereira-Mouries L, Hall DJ. Cartilage oligomeric matrix protein protects cells against death by elevating members of the IAP family of survival proteins. *J Biol Chem*. 2008;283(1):648-659. doi:10.1074/jbc.M704035200.

392. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science*. 2009;324(5930):1029-1033. doi:10.1126/science.1160809.

393. Liberti M V, Locasale JW. The Warburg Effect: How Does it Benefit Cancer Cells? *Trends Biochem Sci*. 2016;41(3):211-218. doi:10.1016/j.tibs.2015.12.001.

394. Denko NC. Hypoxic regulation of metabolism offers new opportunities for anticancer therapy. *Expert Rev Anticancer Ther*. 2014;14(9):979-981. doi:10.1586/14737140.2014.930345.

395. Jose C, Bellance N. Choosing between glycolysis and oxidative phosphorylation: A tumor's dilemma? *Biochim Biophys Acta - Bioenerg*. 2011;1807(6):552-561. doi:10.1016/J.BBABIO.2010.10.012.

396. Nyberg P, Xie L, Sugimoto H, et al. Characterization of the anti-angiogenic properties of arresten, an α1β1 integrin-dependent collagen-derived tumor suppressor. *Exp Cell Res*. 2008;314(18):3292-3305. doi:10.1016/j.yexcr.2008.08.011.

397. Koch M, Hussein F, Woeste A, et al. CD36-mediated activation of endothelial cell apoptosis by an N-terminal recombinant fragment of thrombospondin-2 inhibits breast cancer growth and metastasis in vivo. *Breast Cancer Res Treat*. 2011;128(2):337-346. doi:10.1007/s10549-010-1085-7.

398. Berger AH, Knudson AG, Pandolfi PP. A continuum model for tumour suppression. *Nature*. 2011;476(7359):163-169. doi:10.1038/nature10275.

399. Leedham S, Tomlinson I. The Continuum Model of Selection in Human Tumors: General Paradigm or Niche Product? *Cancer Res*. 2012;72(13):3131-3134. doi:10.1158/0008-5472.CAN-12-1052.

400. Goldfarb DS, Corbett AH, Mason DA, Harreman MT, Adam SA. Importin α: a multipurpose nuclear-transport receptor. *Trends Cell Biol*. 2004;14(9):505-514. doi:10.1016/j.tcb.2004.07.016.

401. Stehling O, Mascarenhas J, Vashisht AA, et al. Human CIA2A-FAM96A and CIA2B-FAM96B Integrate Iron Homeostasis and Maturation of Different Subsets of Cytosolic-Nuclear Iron-Sulfur Proteins. *Cell Metab*. 2013;18(2):187-198. doi:10.1016/j.cmet.2013.06.015.

402. van Wietmarschen N, Moradian A, Morin GB, Lansdorp PM, Uringa E-J. The Mammalian Proteins MMS19, MIP18, and ANT2 Are Involved in Cytoplasmic Iron-Sulfur Cluster Protein Assembly. *J Biol Chem*. 2012;287(52):43351-43358. doi:10.1074/jbc.M112.431270.

403. Ito S, Tan LJ, Andoh D, et al. MMXD, a TFIIH-independent XPD-MMS19 protein complex involved in chromosome segregation. *Mol Cell*. 2010;39(4):632-640. doi:10.1016/j.molcel.2010.07.029.

404. Raturi A, Gutiérrez T, Ortiz-Sandoval C, et al. TMX1 determines cancer cell metabolism as a thiol-based modulator of ER-mitochondria Ca2+ flux. *J Cell Biol*. 2016;214(4):433-444. doi:10.1083/jcb.201512077.

405. Stray-Pedersen A, Backe PH, Sorte HS, et al. PGM3 mutations cause a congenital disorder of glycosylation with severe immunodeficiency and skeletal dysplasia. *Am J Hum Genet*. 2014;95(1):96-107. doi:10.1016/j.ajhg.2014.05.007.

406. Zhang Y, Yu X, Ichikawa M, et al. Autosomal recessive phosphoglucomutase 3 (PGM3) mutations link glycosylation defects to atopy, immune deficiency, autoimmunity, and neurocognitive impairment. *J Allergy Clin Immunol*. 2014;133(5):1400-1409, 1409-5. doi:10.1016/j.jaci.2014.02.013.

407. Sassi A, Lazaroski S, Wu G, et al. Hypomorphic homozygous mutations in phosphoglucomutase 3 (PGM3) impair immunity and increase serum IgE levels. *J Allergy Clin Immunol*. 2014;133(5):1410-1419, 1419-13. doi:10.1016/j.jaci.2014.02.025.

408. Okumura M, Ichioka F, Kobayashi R, et al. Penta-EF-hand protein ALG-2 functions as a Ca2+-dependent adaptor that bridges Alix and TSG101. *Biochem Biophys Res Commun*. 2009;386(1):237-241. doi:10.1016/j.bbrc.2009.06.015.

409. Lee JH, Rho SB, Chun T. Programmed Cell Death 6 (PDCD6) Protein Interacts with Death-Associated Protein Kinase 1 (DAPk1): Additive Effect on Apoptosis via Caspase-3 Dependent Pathway. *Biotechnol Lett*. 2005;27(14):1011-1015. doi:10.1007/s10529-005-7869-x.

410. Rho SB, Song YJ, Lim MC, Lee S-H, Kim B-R, Park S-Y. Programmed cell death 6 (PDCD6) inhibits angiogenesis through PI3K/mTOR/p70S6K pathway by interacting of VEGFR-2. *Cell Signal*. 2012;24(1):131-139. doi:10.1016/j.cellsig.2011.08.013.

411. Gluz O, Wild P, Meiler R, et al. Nuclear karyopherin α2 expression predicts poor survival in patients with advanced breast cancer irrespective of treatment intensity. *Int J Cancer*. 2008;123(6):1433-1438. doi:10.1002/ijc.23628.

412. Alshareeda AT, Negm OH, Green AR, et al. KPNA2 is a nuclear export protein that contributes to aberrant localisation of key proteins and poor prognosis of breast cancer. *Br J Cancer*. 2015;112(12):1929-1937. doi:10.1038/bjc.2015.165.

413. Lee C-H, Jeong S-J, Yun S-M, et al. Down-regulation of phosphoglucomutase 3 mediates sulforaphane-induced cell death in LNCaP prostate cancer cells. *Proteome Sci*. 2010;8(1):67. doi:10.1186/1477-5956-8-67.

414. Munkley J. Glycosylation is a global target for androgen control in prostate cancer cells. *Endocr Relat Cancer*. 2017;24(3):R49-R64. doi:10.1530/ERC-16-0569.

415. Yang W, Itoh F, Ohya H, et al. Interference of E2-2-mediated effect in endothelial cells by FAM96B through its limited expression of E2-2. *Cancer Sci*. 2011;102(10):1808-1814. doi:10.1111/j.1349-7006.2011.02022.x.

416. Tucker RP, Drabikowski K, Hess JF, Ferralli J, Chiquet-Ehrismann R, Adams JC. Phylogenetic analysis of the tenascin gene family: Evidence of origin early in the chordate lineage. *BMC Evol Biol*. 2006;6(1):60. doi:10.1186/1471-2148-6-60.

417. Hsia HC, Schwarzbauer JE. Meet the tenascins: Multifunctional and mysterious. *J Biol Chem*. 2005;280(29):26641-26644. doi:10.1074/jbc.R500005200.

418. Valcourt U, Alcaraz LB, Exposito JY, Lethias C, Bartholin L. Tenascin-X: Beyond the architectural function. *Cell Adhes Migr*. 2015;9(1-2):154-165. doi:10.4161/19336918.2014.994893.

419. Zhang Z, Kochan GT, Ng SS, et al. Crystal structure of PHYHD1A, a 2OG oxygenase related to phytanoyl-CoA hydroxylase. *Biochem Biophys Res Commun*. 2011;408(4):553-558. doi:10.1016/J.BBRC.2011.04.059.

420. Degen M, Brellier F, Schenk S, et al. Tenascin-W, a new marker of cancer stroma, is elevated in sera of colon and breast cancer patients. *Int J Cancer*. 2008;122(11):2454-2461. doi:10.1002/ijc.23417.

421. Scherberich A, Tucker RP, Degen M, Brown-Luedi M, Andres A-C, Chiquet-Ehrismann R. Tenascin-W is found in malignant mammary tumors, promotes alpha8 integrin-dependent motility and requires p38MAPK activity for BMP-2 and TNF-alpha induced expression in vitro. *Oncogene*. 2005;24(9):1525-1532. doi:10.1038/sj.onc.1208342.

422. Martina E, Degen M, Rüegg C, et al. Tenascin-W is a specific marker of glioma-associated blood vessels and stimulates angiogenesis in vitro. *FASEB J*. 2010;24(3):778-787. doi:10.1096/fj.09-140491.

423. Kimura H, Akiyama H, Nakamura T, Crombrugghe B de. Tenascin-W inhibits proliferation and differentiation of preosteoblasts during endochondral bone formation. *Biochem Biophys Res Commun*. 2007;356(4):935-941. doi:10.1016/j.bbrc.2007.03.071.

424. Scherberich A, Tucker RP, Samandari E, Brown-Luedi M, Martin D, Chiquet-Ehrismann R. Murine tenascin-W: a novel mammalian tenascin expressed in kidney and at sites of bone and smooth muscle development. *J Cell Sci*. 2004;117(Pt 4):571-581. doi:10.1242/jcs.00867.

425. Galon J, Angell HK, Bedognetti D, Marincola FM. The Continuum of Cancer Immunosurveillance: Prognostic, Predictive, and Mechanistic Signatures. *Immunity*. 2013;39(1):11-26. doi:10.1016/J.IMMUNI.2013.07.008.

426. Furusawa Y, Kubo T, Fukazawa T. Phyhd1, an XPhyH-like homologue, is induced in mouse T cells upon T cell stimulation. *Biochem Biophys Res Commun*. 2016;472(3):551-556. doi:10.1016/j.bbrc.2016.03.039.

427. Cardoza JD, Parikh JR, Ficarro SB, Marto JA. Mass spectrometry-based proteomics: Qualitative identification to activity-based protein profiling. *Wiley Interdiscip Rev Syst Biol Med*. 2012;4(2):141-162. doi:10.1002/wsbm.166.

428. Sheynkman GM, Shortreed MR, Cesnik AJ, Smith LM. Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. *Annu Rev Anal Chem*. 2016;9(1):521-545. doi:10.1146/annurev-anchem-071015-041722.

429. Nesvizhskii AI. Proteogenomics: Concepts, applications and computational strategies. *Nat Methods*. 2014;11(11):1114-1125. doi:10.1038/NMETH.3144.

430. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem*. 2007;389(4):1017-1031. doi:10.1007/s00216-007-1486-6.

431. Listgarten J, Emili A. Statistical and Computational Methods for Comparative Proteomic Profiling Using Liquid Chromatography-Tandem Mass Spectrometry. *Mol Cell Proteomics*. 2005;4(4):419-434. doi:10.1074/mcp.R500005-MCP200.

432. Domon B, Aebersold R. Mass Spectrometry and Protein Analysis. *Science (80-)*. 2006;312(5771):212-217. doi:10.1126/science.1124619.

433. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592. doi:10.1093/biomet/63.3.581.

434. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*. 2001;17(6):520-525. doi:10.1093/bioinformatics/17.6.520.

435. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: A critical evaluation. *BMC Med Inform Decis Mak*. 2016;16(S3):74. doi:10.1186/s12911-016-0318-z.

436. Schafer JL, Graham JW. Missing data: Our view of the state of the art. *Psychol Methods*. 2002;7(2):147-177. doi:10.1037//1082-989X.7.2.147.

437. Ibrahim JG, Chen M-H, Lipsitz SR, Herring AH. Missing-Data Methods for Generalized Linear Models. *J Am Stat Assoc*. 2005;100(469):332-346. doi:10.1198/016214504000001844.

438. Buuren S van, Groothuis-Oudshoorn K. **mice** : Multivariate Imputation by Chained Equations in *R*. *J Stat Softw*. 2011;45(3). doi:10.18637/jss.v045.i03.

439. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? *Int J Methods Psychiatr Res*. 2011;20(1):40-49. doi:10.1002/mpr.329.

440. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30(4):377-399. doi:10.1002/sim.4067.

441. Royston P, White I. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *J Stat Softw*. 2011;45(4):1-20. doi:10.18637/jss.v045.i04.

442. Schafer JL. Statistical Methods in Medical Research. *Stat Methods Med Res*. 1999;8(1):3-

15. doi:10.1177/096228029900800102.

443. Albrecht D, Kniemeyer O, Brakhage AA, Guthke R. Missing values in gel-based proteomics. *Proteomics*. 2010;10(6):1202-1211. doi:10.1002/pmic.200800576.

444. Emmert-Buck MR, Bonner RF, Smith PD, et al. Laser capture microdissection. *Science (80-)*. 1996;274(5289):998-1001. doi:10.1126/science.274.5289.998.

445. Espina V, Heiby M, Pierobon M, Liotta LA. Laser capture microdissection technology. *Expert Rev Mol Diagn*. 2007;7(5):647-657. doi:10.1586/14737159.7.5.647.

446. Staunton L, Tonry C, Lis R, et al. Profiling the tumor microenvironment proteome in prostate cancer using laser capture microdissection coupled to LC-MS-A technical report. *EuPA Open Proteomics*. 2016;10:19-23. doi:10.1016/j.euprot.2015.11.001.

447. Longuespée R, Alberts D, Pottier C, et al. A laser microdissection-based workflow for FFPE tissue microproteomics: Important considerations for small sample processing. *Methods*. 2016;104:154-162. doi:10.1016/j.ymeth.2015.12.008.

448. Caprioli RM, Farmer TB, Gile J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Anal Chem*. 1997;69(23):4751-4760. http://www.ncbi.nlm.nih.gov/pubmed/9406525. Accessed May 13, 2018.

449. Schwamborn K, Caprioli RM. MALDI Imaging Mass Spectrometry – Painting Molecular Pictures. *Mol Oncol*. 2010;4(6):529-538. doi:10.1016/J.MOLONC.2010.09.002.

450. Aichler M, Walch A. MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Lab Investig*. 2015;95(4):422-431. doi:10.1038/labinvest.2014.156.

451. McDonnell LA, Heeren RMA. Imaging mass spectrometry. *Mass Spectrom Rev*. 2007;26(4):606-643. doi:10.1002/mas.20124.

452. Dilillo M, Ait-Belkacem R, Esteve C, et al. Ultra-High Mass Resolution MALDI Imaging Mass Spectrometry of Proteins and Metabolites in a Mouse Model of Glioblastoma. *Sci Rep*. 2017;7(1):603. doi:10.1038/s41598-017-00703-w.

453. Patel E, Cole LM, Bradshaw R, et al. MALDI-MS imaging for the study of tissue pharmacodynamics and toxicodynamics. *Bioanalysis*. 2015;7(1):91-101. doi:10.4155/bio.14.280.

454. Maier SK, Hahne H, Gholami AM, et al. Comprehensive Identification of Proteins from MALDI Imaging. *Mol Cell Proteomics*. 2013;12(10):2901-2910. doi:10.1074/mcp.M113.027599.

455. Aebersold R, Burlingame AL, Bradshaw RA. Western Blots versus Selected Reaction Monitoring Assays: Time to Turn the Tables? *Mol Cell Proteomics*. 2013;12(9):2381-2382. doi:10.1074/mcp.E113.031658.

456. Anjo SI, Santa C, Manadas B. SWATH-MS as a tool for biomarker discovery: From basic research to clinical applications. *Proteomics*. 2017;17(3-4):1600278. doi:10.1002/pmic.201600278.

457. Meier F, Geyer PE, Virreira Winter S, Cox J, Mann M. Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Nat Methods*. 2018;11(6):1. doi:10.1038/s41592-018-0003-5.