Early stage diagnosis of colorectal cancer using a multi-variate blood-based test

By

Samridhi Sharma BPharm MPharm

Principal Supervisor: Professor Mark Baker

Co-Supervisor: Dr Seong Beom Ahn

MACQUARIE
University

A thesis submitted to Macquarie University in fulfilment of
the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Sciences
Faculty of Medicine and Health Sciences
Macquarie University
Sydney. 2109. New South Wales. Australia

January 2019

Dedicated to all my teachers,
first of whom were my parents

Declaration of originality

I hereby declare that the work presented in this PhD thesis entitled *"Early stage diagnosis of colorectal cancer using a multi-variate blood-based test"* under supervision of Professor Mark Baker was carried out at Department of Biomedical Sciences, Macquarie University. This thesis is not been submitted for higher degree to any other university or institution. The contents of this thesis are original and to the best of my knowledge contains no material previously published or written by another person, except where due reference is stated otherwise with permission taken from respective journals and authors (attached in appendix).

Samridhi Sharma
44549997

Department of Biomedical Sciences,
Faculty of Medicine and Health Sciences
Macquarie University.

2/1/2019

Acknowledgements

First and foremost, I would like to express my deepest gratitude towards my supervisor Professor Mark Baker without whose constant support and encouragement this thesis could not have been completed. Thanks for giving me the opportunity to carry out this interesting project and keeping me motivated through to completion. I am especially thankful for the freedom to drive this research whilst receiving valuable advice and pushing me to strive for the best. I would also like to thank Professor Baker for believing in me and providing me with the exposure to share my research on an international platform with other eminent scientists.

I am eternally grateful to our head of department Professor Helen Rizos for being forthcoming in helping me and inspiring me through this journey. I am grateful to her for actively revising my thesis correction reports.

I would like to thank my co-supervisor, Dr Seong Beom Ahn, for supporting me and helping me towards the completion of this thesis. Dr Ahn helped me troubleshoot complex experiments with ease. Scientific discussions with him regarding best practices to analyse data is a lesson, that I have taken for life. I would also like to thank my ex-co-supervisor Dr Susan Fanayan for interviewing me and accepting as her and Mark's PhD student. Dr Fanayan had not only provided scientific support in a short 6-month time but braced me emotionally and helped me adjust in the new city and lab environment. I will forever be grateful.

Many thanks to Professor Edward Nice, Dr David Cantor, Dr William Redmond and Dr Abidali Mohamedali for their patient reading of my thesis chapters and providing their valuable scientific inputs. I would extend my thanks to my current and past colleagues, Mr Subash Adhikari, Ms Sachini Fonseka, Dr Harish Cheruku, for being forthcoming, and camaraderie shared during this roller coaster ride of 3 years. I would like to thank Australian Proteomics Analysis Facility (APAF) for their constant support and providing access to important mass spectrometry machines. I am grateful for the help received from Dr Mathew Mckay, Dr Karthik Kamath, Dr Dana Pascovici, Thiri Zaw and Dr Jemma X Wu. A special thanks to Professor Shoba Ranganathan and her student Ms Zainab Noor for scientific discussions regarding SWATH™-MS data and creating iSWATHX for easy data conversions.

I would like to extend my thanks to pillars of Faculty of Medicine and Health Sciences (FMHS), Ms Laura Newey and Ms Viviana Bong for patiently answering academic documentation related queries. I also acknowledge the support received from lab operations team, Ms Louise Marr, Ms Tamara Leo, Ms Lucy Lu and Mr Mitchel Borton for smooth

running of my experiments. I am thankful to Dr Jennifer Rowland for her help in queries regarding scientific writing and formatting of this thesis.

A special thanks goes to Dr Nitin Chitranshi, Dr Vivek Gupta, Mrs Preeti Manandhar, and Dr Sumudu Gangoda for their constant moral support. I will never forget the unconditional love and support from my best friends Ms Shabarni Gupta, Dr Vineet Vaibhav, Ms Surbhi Chabbra, Dr Yogita Dheer, Dr Nikita Gahoi, Dr Darpan Malhotra, Dr Rohith Basu and Mr Ravi Sharma who are always with me in thick and thin of life. Words cannot express how grateful I am to my friends for being on my side and making it look like a smooth ride. I thank them for their friendship, objective criticism, and motivating me to grow professionally and personally.

I would like to thank my parents Mrs Renuka Sharma and Mr Rama Kant Sharma and my sister Ms Saumya Sharma for inspiring my moral conduct, and character. Thanks for their constant love and support, that sustains me through tribulations and testing times.

Financial Acquisition for this project comes from departmental research funds from Faculty of Medicine and Health Sciences for PhD students and Cancer Council Australia and Sydney Vital Scholarship Funding 50468/00 79730226.

Contribution Statement

Chapter 1

*Publication 1: **Baker MS**, Ahn S B, Mohamedali A, Islam M T, Cantor D, Verhaert PD, Fanayan S, **Sharma S**, Nice E C, Connor M and Ranganathan S. Accelerating the search for the human proteome's missing proteins, Nat Comms. 2017 Jan 24; 8:14271. doi:10.1038/ncomms14271*

Sharma S conducted the bioinformatic data analyses and interpretation required to generate Figures 1-3 of this manuscript. **Sharma S** also contributed to the preparation and review of this manuscript. Contribution of the other authors was as follows: Ahn S, Cantor D and Mohamedali A contributed to the conception and design of the three displays, Box 1, Box 2 and Box 3. Islam T, Ranganathan S, Mohamedali A and **Baker MS** designed and conceived the idea behind the MissingProteinPedia. Ranganathan S and Islam MT performed the bioinformatics analysis for the MissingProteinPedia. Cantor D and Verhaert PD prepared Figure 6. Fanayan S and Nice E contributed to drafting the manuscript. Baker MS was corresponding author who conceived, designed, organised, collated information and wrote the initial and final drafts of the manuscript. All authors contributed to multiple revisions of the manuscript.

*Publication 2: Adhikari S, **Sharma S**, Ahn S B and **Baker MS**. How Much of the Human Olfactory Receptor Proteome is Findable Using High-Stringency Mass Spectrometry? Journal of Proteome Research, 2018 (Accepted)*

Sharma S contributed to the conception of this project from the previous analysis done for the figure 2, Publication I of this thesis. **Sharma S** wrote the parts pertaining to figure 2 in Material and Method section, result section. **Sharma S** helped in revision of the manuscript. Adhikari S conceived the idea to write this publication II. He performed the analysis, wrote the code to design the software for this manuscript and wrote and revised the manuscript. Ahn S B helped in supervision and revision of the manuscript. **Baker MS** conceived the idea, wrote the manuscript and supervised the software.

Chapter 2

Introduction to CRC

Sharma S contributed to the conception, content and design of this chapter. Baker M contributed to the revision of this chapter.

Chapter 3

*Publication 3: **Sharma S**, Ahn SB, Redmond W, Mohamedali A, Vaibhav V, Pascovici D, Wu J X, Zaw T, Nice E, and **Baker, MS**. Potential early clinical stage colorectal cancer diagnosis using a proteomics blood test" submitted in journal, Clinical Proteomics on 30th May 2019 (Ref #: CLIP-19-00034)* .

Sharma S contributed to the conception and design of the experiments, performed all experiments, analysed the data, interpreted the results and was the major contributor to the manuscript. Ahn SB and Mohamedali A, designed the experiments assisted with analysing the data, interpreting the results and in writing plus revising the manuscript. Redmond W designed and created machine learning model. He also wrote the material and method section, results and discussion section of machine learning model of this project. Vaibhav V performed western blotting experiment. Pascovici D and Wu JX and Zaw T performed statically analysis of mass spectrometry data. **Baker MS** contributed to the conception and design of the experiments, interpretation of the results and revision of the manuscript. Nice E contributed to the revision of the manuscript.

Chapter 4

Verification of a multi-variate signature assay for early diagnosis using Parallel Reaction Monitoring (PRM) assay.

Sharma S contributed to the conception and design of the experiments, performed all mass spectrometry-related experiments, analysed the data and wrote the first draft of the manuscript. Fonseka S contributed by co-optimising PRM experiment method in Chapter 5 with **Sharma S**, Ahn SB and McKay M supervised the technical aspects of the PRM experiments. Nice E contributed to the revision of the manuscript. **Baker MS** contributed to the conception and design of the experiments and writing of the manuscript draft.

Chapter 5

***Sharma S, Fonseka S**, Ahn SB, Mckay M, and **Baker MS**, μPAR and αvβ6 as metastatic marker of colorectal cancer quantified using parallel reaction monitoring. (In preparation)*

Sharma S contributed to the conception and design of the experiments along with Fonseka S. Fonseka S performed initial experiments, analysed and interpreted results involving recombinant proteins and cell lysates independently. **Sharma S** performed verification experiments, analysed and interpreted results on recombinant proteins and patient plasma

independently. The writing of this manuscript was carried out by the candidate. A detailed itemised contribution is presented in the beginning of this chapter. Ahn S B and McKay M contributed by supervising the technical aspects of the PRM experiments. **Baker MS** contributed to the conception and design of the experiments and writing of the manuscript draft.

All original publications have been reproduced with permission of the authors and copyright holders. Additional previously unpublished data are also presented in this thesis.

International presentations

1. Oral/poster presentations; **Sharma S** with co-authors Ahn S B, Redmond W, Mohamedali A, Vaibhav V, Pascovici D, Wu J X, Zaw T, Nice E, and Baker, MS. *"Early stage diagnosis of colorectal cancer using multi-analyte blood-based test"* at the Sao Paulo School of Advanced Science on Mass spectrometry, Aug 8th -15th, Brazil. Travel was funded and sponsored by the Brazilian National Research Centre for Energy and Materials (CNPEM).
2. Poster presentation; Human Proteome Organisation (HUPO) 2016 World Congress, Dublin, Ireland entitled *"Early stage diagnosis of colorectal cancer using multi-analyte blood-based test"* sponsored by Postgraduate Research Travel Award, Faculty of Medicine and Health Sciences, Macquarie University.

Patent

SWATH™-MS based discovery of early stage CRC proteomic biosignatures using ultradepleted plasma. #APPA 2016902484. In communication with industrial partners around the world.

Awards

1. **2017-2018**: Awarded $AUD 10,000 from Sydney Vital Translational Cancer Research, Australia for research
2. **2017**: Travel Award covering registration, airfare and accommodation oral presentation from Sao Paulo School of Advanced Science on Mass spectrometry, Brazil sponsored by CNPEM
3. **2016**: $AUD 5,000 was received from Post Graduate Research Funds sponsored by Macquarie University, Australia

Media Outreach (Attached as Appendix IV)

1. Off-site research program initiated ICPC Pilots International Student Training under Cancer Moonshot Project at Fred Hutch Cancer Research Centre.
2. Published in Local Newspaper of Sydney and NCI website, USA.

List of Abbreviations

Abbreviation	Abbreviation in full
ADAMDEC1	A Distintegrin Metalloprotease Protein
AFAP	Attenuated Familial Adenomatous Polyposis
AJCC	American Joint Committee on Cancer
APC	Adenomatous Polyposis Coli
APOA4	Apolipoprotein A-Iv
APOLLO	Applied Proteogenomics Organisation
ASR	Age Standardised Rate
αV	αv Integrin Subunit
αVB6	αvβ6 Integrin Heterodimer
β6	β6 Integrin Subunit
B/D-HPP	Biology/ Disease-Driven Human Proteome Project
C-HPP	Chromosome Centric Human Proteome Project
Chr	Chromosome
C1qc	Complement C1q Subcomponent Subunit C
cfDNA	Cell-Free DNA
ctDNA	Circulating Tumour DNA
CEA	Carcinoembryonic Antigen
CFD	Complement Factor D
CIN	Chromosomal Instability
CO-IP	Co-Immunoprecipitation
COMP	Cartilage Oligomeric Matrix Protein
CPTAC	Clinical Proteomic Tumour Analysis Consortium
CRC	Colorectal Cancer
CST3	Cystatin-C
DDA	Data-Dependent Acquisition
DIA	Data-Independent Acquisition
ECM	Extracellular Matrix
EDTA	Ethylenediaminetetraacetic Acid
EGFR	Epidermal Growth Factor Receptor
ELISA	Enzyme-Linked Immune Sorbent Assay

EMT	Epithelial-to-Mesenchymal Transition
ESI	Electrospray Ionisation
FA	Formic Acid
FAP	Familial Adenomatous Polyposis
FDR	False Discovery Rate
FIT	Faecal Immunochemical Test
FOBT	Faecal Occult Blood Test
FDA	Food and Drug Administration
GENIE	Genomics Evidence Neoplasia Information Exchange
GPMDB	Global Proteome Machine Database
GPM	Global Proteome Machine
GPR56	G-Protein Coupled Receptor 56
GPX3	Glutathione Peroxidase 3
gFOBT	Guaiac Faecal Occult Blood Test
HDI	Human Development Index
HNPCC	Hereditary Nonpolyposis Colorectal Cancer
HPH	High pH Reversed Phased C18 Peptide Fractionation
HPP	Human Proteome Project
HPLC	High Pressure Liquid-Chromatography
HS-gFOBT	High-Sensitivity Guaiac FOBT
HUPO	Human Proteome Organisation
IBD	Inflammatory Bowel Disease
IARC	International Agency for Research on Cancer
ICAT	Isotope Coded Protein Labels
IgY	Immunoglobulin Y
IT	Ion Trap
iTRAQ	Isobaric Tag for Relative and Absolute Quantitation
KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog
LC	Liquid Chromatography
LS	Lynch Syndrome
LoD	Limit of Detection
LoH	Loss of Heterozygosity

LYG	Life Years Gained
mt-sDNA	Multi-Target Stool DNA Test
MALDI	Matrix-Assisted Laser Desorption/Ionisation
MAPK	Mitogen-Activated Protein Kinase
MARCO	Macrophage Receptor
MARS	Multiple Affinity Removal System
MET	Mesenchymal-To-Epithelial Transition
mSEPT9	Methylated Septin9 Gene
MS	Mass Spectrometry
MYH/ MUTYH	Muty Dna Glycosylase
miRNA	Micro RNA
MMP	Matrix Metalloprotease
MMR	Mismatch Repair
MPP	MissingProteinPedia
MRC1	Macrophage Mannose Receptor 1
MSI	Microsatellite Instability
MS/MS	Tandem Mass Spectrometry
m/z	Mass-to Charge Ratio
nanoLC	Nanoscale Reversed-Phase Liquid Chromatography
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NGS	Next-Generation Sequencing
NMR	Nuclear Magnetic Resonance
nrdG4	NDRG Family Member 4
NSAIDS	Non-Steroidal Anti-Inflammatory Drugs
NSAF	Normalised Spectral Abundance Factor
P4	Personalised, Precision, Preventative and Participatory Medicine
PCA	Principal Component Analysis
PE	Protein Evidence
PON1	Serum Paroxonase/Arylesterase 1
PPV	Positive Predictive Value

PRIDE	Proteomics Identifications Database
PRKDC	Protein Kinase, DNA-Activated, Catalytic Polypeptide
PRM	Parallel Reaction Monitoring
PTMS	Post-Translational Modification(S)
PX	ProteomeXchange
RNA	Ribonucleic Acid
S100A8	Protein S100-A8
SAX	Strong Anion Exchange
SEC	Size Exclusion Chromatography
SEM	Standard Error of the Mean
SCX	Strong Cation Exchange Chromatography
SILAC	Stable Isotope Labelling by Amino Acids in Culture
SRM/MRM	Single/Multiple Reaction Monitoring
SWATH™-MS	Sequential Window Acquisition of All Theoretical Spectra Mass Spectrometry
TCGA	The Cancer Genome Atlas
TFPI2	Tissue Factor Pathway Inhibitor 2
TNM	Tumour, Nodes, Metastasis (Staging System)
TOF	Time-Of-Flight
uPAR	Urokinase-Type Plasminogen Activator Receptor
VIM	Vimentin
WHO	World Health Organisation

Overview of thesis

Over the past nine years, the Human Proteome Project (HPP) has made enormous progress in achieving two goals. These are: mapping the human protein repertoire under the Chromosome-centric human proteome project (C-HPP) and, understanding the pathophysiology of human biology and disease under the Biology/Disease-driven human proteome project (B/D-HPP) (Figure 1).

This thesis contributes towards these two HPP goals by;

ii) cataloguing human proteins, particularly the "missing proteins' (Chapter 1) and

ii) diagnosing early stage colorectal cancer (CRC) through the development of a multi-variate blood-based assay (Chapters 2-5; with results contained across Chapters 3, 4 and 5).

Chapter 1 of this thesis summarises contributions in developing an investigative approach and community-centric resource (MissingProteinPedia) to accelerate the discovery and understanding of 'missing proteins' (Publication I). Out of ~ 20,000 human protein coding genes, 2,319 proteins (at time of writing) still lack mass spectrometry-based evidence for reliable high-stringency protein existence data. These are referred to as the PE2-4 proteins or colloquially as the 'missing proteins'. This chapter explores the common characteristics of missing proteins, including assignment to groups of proteins, topological distribution and structural composition, and it details challenges associate with identifying these proteins. This is exemplified by investigation of an elusive class/family of GPCR proteins, known as the olfactory receptors (ORs). The chapter provides potential approaches to facilitate identification of missing proteins in order to accelerate successful and timely completion of the HPP.

Chapter 2 introduces CRC by reviewing incidence, stages, mortality rate, as well as the factors governing susceptibility and pathophysiology, staging and survival. The chapter describes existing screening tests, and CRC biomarkers in context unmet clinical needs. One major requirement is to develop a minimally invasive, blood-based markers for routine CRC screening is explored in this chapter with particular reference to technological and strategic advances across the field of proteomics. The complexities and challenges associated with plasma proteomic biomarker discovery are also discussed.

Chapter 3 employs state-of-the-art proteomic techniques to provides an in-depth analysis of the CRC plasma proteome in order to develop such blood-based diagnostic test. A combination of

patented, in-house ultradepletion methods, commercial immunodepletion approaches and multiple peptide fractionations was used to overcome the challenges of detecting low-abundance proteins in plasma.

Furthermore, SWATH™-MS was employed for specific and reliable exploration of protein biomarkers using 100 clinically staged EDTA-plasma samples (n=20 per AJCC stage for CRC (i.e. stages I, II, III and IV)). A total of 37 potential protein candidates were identified by comparing differential expression in CRC stages (I-IV) against the plasma pool of healthy controls (n=20) with stringent statistical analysis (fold change ≥ 1.5, unique peptide ≥ 1, $p < 0.05$). A literature search indicated some of these putative CRC biomarkers were novel while others had been previously associated with CRC. The capability of 7 of these 37 candidates to distinguish CRC early stages (I/II) from healthy controls were confirmed using Western blotting and/or ELISA-based method. In addition, a machine learning model was then utilised to confirm the potential of 5 protein candidates (from list of 37 candidates) as a putative panel to distinguish early stage (I/II) CRC from healthy controls using a 5000 synthetic patient cohort.

Chapter 4 then develops and verifies a first-pass parallel reaction monitoring (PRM) assays to interrogate plasma samples for two target peptides for Complementary factor D and ADAMDEC1, both were identified as potential candidates from Chapter 3. Despite encouraging initial data, this work remained preliminary and requires further characterisation and validation studies, which are described in detail.

Penultimately, Chapter 5 develops a proof-of-concept PRM assay to quantify the expression of two markers of the epithelial-to-mesenchymal transition (uPAR and αvβ6) in HCT116 (colon carcinoma) cell line and CRC plasma samples. Both (uPAR and αvβ6) are of low abundance in plasma and have been extensively studied by our group. Increased levels of uPAR has been established as a stage II prognostic marker in CRC tissues, and equally, αvβ6 expression has shown to be elevated in early stage of CRC. The objective of this chapter was to explore the increased expressions of circulating uPAR and αvβ6 as potential biomarkers in development of CRC in using plasma samples. The proof-of-concept PRM assay developed, showed a 33% decrease in uPAR expression between antisense ($HCT116^{AS}$) and wildtype ($HCT116^{WT}$). This study is ongoing and requires optimisation to verify and quantitate both uPAR and αvβ6 peptide from CRC plasma samples.

Finally, Chapter 6 summarises the thesis in the context of the current literature providing a discussion of possible future directions and the clinical significance and limitations of the studies outlined. I describe the importance of population-based studies of the selected CRC protein biomarkers and the need for direct comparisons to other existing screening methodologies.

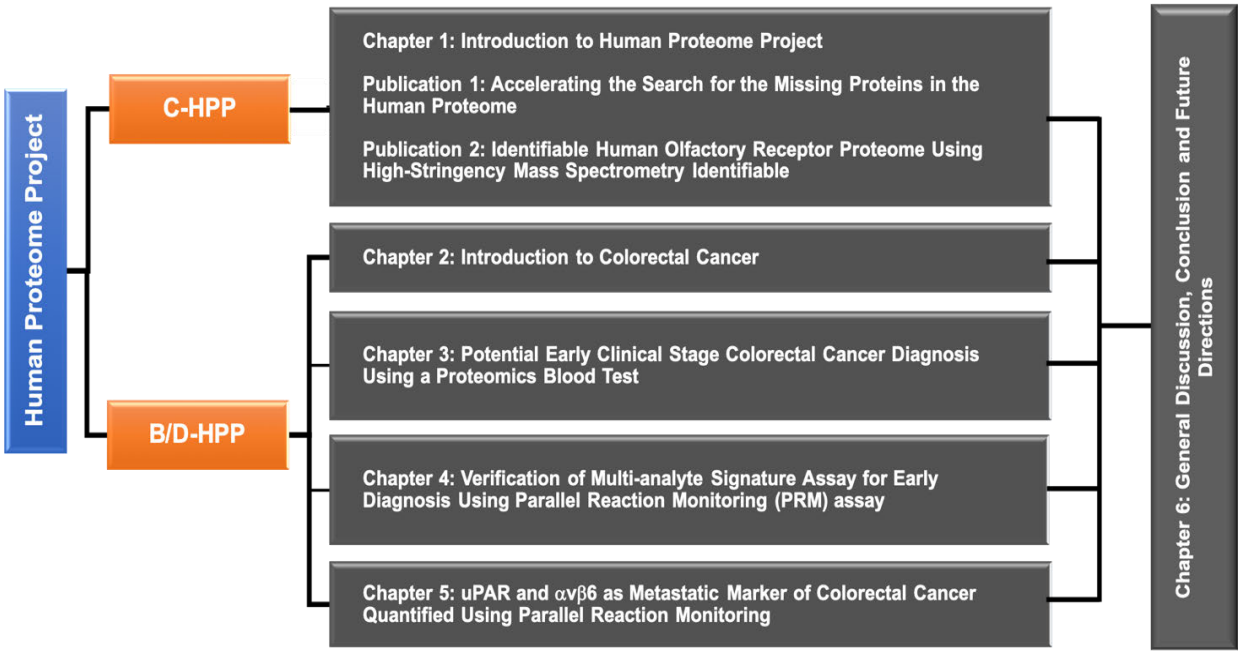


***Figure 1**: **Thesis Outline Flowchart:** Thesis chapters are organised around the two key central streams/programs of the Human Proteome Project programs namely – the Chromosome-centric Human Proteome Project (C-HPP) and the Biology/Disease-Driven Human Proteome Project (B/D-HPP)*

Table of Contents

Chapter 1

The Human Proteome Project

1.1 The Human Proteome Project – Positioning Involvement in the Journey

The central paradigm of molecular biology emphasises the ordered flow of information between three biomolecules, namely DNA, RNA and proteins. These govern every intricate aspect of cellular function (Li and Xie, 2011). The initial publication of the human genome in 2000 (Venter et al., 2001), (Lander et al., 2001) provided the DNA blueprint for all molecules that enabled any cell to carry out biochemical processes. This was a landmark achievement in the field of life science research and an important step towards understanding human health and physiology at the grass roots - its DNA.

However, the genome is relatively static in nature, while human cellular function, physiology and disease are highly dynamic (Aebersold and Mann, 2003). It was quickly realised that other technologies like proteomics, epigenomics, transcriptomics, and metabolomics would be essential for understanding human biology (Bilello, 2005), *in toto*. Researchers have also been quick to harness the power of next-generation sequencing (NGS) to explore the dynamic human transcriptome in our efforts to understand complexities in cellular phenotype (Wang et al., 2009). Importantly, genomics and transcriptomics made remarkable advances in this area with technological breakthroughs in both NGS (improved speed, cost and accuracy) and the development of sensitive and specific, high throughput, microarrays (Shendure and Ji, 2008). With a minimum of ~230 different cell human types comprising the human body (Legrain et al., 2011) transcriptomics complemented genomics in its power to ascertain the profile of RNA levels expressed in each cell type, increasing our knowledge of the cellular landscape (Wang et al., 2009).

In our endeavours to fill the gap in genomics and transcriptomics, proteomics has emerged as a key tool to decipher cellular phenotypes, principally by studying the most crucial determinant of biochemical function, namely proteins expression, turnover, activity and post-translational modification (Cox and Mann, 2007). Biomolecules, however, do not act in isolation. It is the interaction between various classes of biomolecules (interactomics) in time and space, their regulation through post-translational modifications and response to external factors that

provide not only a holistic, but also a more simplistic understanding of cellular function. For example, it is possible to observe a host of genetic alterations or transcript profiles in a number of disease-related genes which can all, in effect, be mapped onto a single biochemical pathway or pathways that interlink to give rise to a certain phenotype (Aebersold et al., 2013).

The Human Proteome Organisation (HUPO) has played a pivotal role in revolutionising the understanding of human proteomics by providing an organised framework for globally sharing experimental protocols, data and research techniques. Various research initiatives, pillars and teams started under HUPO have been amalgamated and collectively termed the global Human Proteome Project (HPP). The HPP was officially launched by HUPO at its 2010 World Congress in Sydney, Australia and further extensions have been developed and launched since then (Aebersold et al., 2013; Omenn, 2012). The HPP broadly provides a global effort to identify proteins produced by protein-coding genes, in respect to protein abundance, subcellular localisation, interaction with other biomolecules and function (Omenn, 2017). In contrast to the genome which is relatively static (~20,000 predicted protein-coding genes) (Aebersold et al., 2018), the human proteome is vast, nebulous and dynamic. The complexity of the total human proteome is contributed to by many factors, including, but not restricted to, alternative splice variation, post-translational modifications, variation in protein activity and cellular location. These result in a dynamic proteome, where not only does the repertoire of proteins expressed differs between cell types but expression levels of a subsets of proteins may change with time and in response to external factors within any cell type.

To address this dynamism, a three-pronged approach was proposed by the HPP to unravel the human proteome (Legrain et al., 2011). The first was to harness the technological and analytical prowess of mass spectrometry to identify peptides and proteins from human cells and tissues with high reproducibility and stringent data analysis. The second aimed at generating specific antibodies against each of the ~20,000 proteins in order to identify spatio-temporal location in normal and diseased human cells, tissues and organs. The third aspect was directed towards developing state-of-the-art bioinformatics for the project in the form of a knowledge database (i.e., knowledgebase) to compile, curate and organise information obtained from the first two approaches. This framework, termed the "*two streams and three pillars of the HPP*" (Figure 1), has received substantial support from the proteomics community. Over ensuing years, efforts involving numerous laboratories spread across the globe and these have contributed to

realising many of the goals set by the HPP and are working in synergy towards its completion (Omenn et al., 2018).

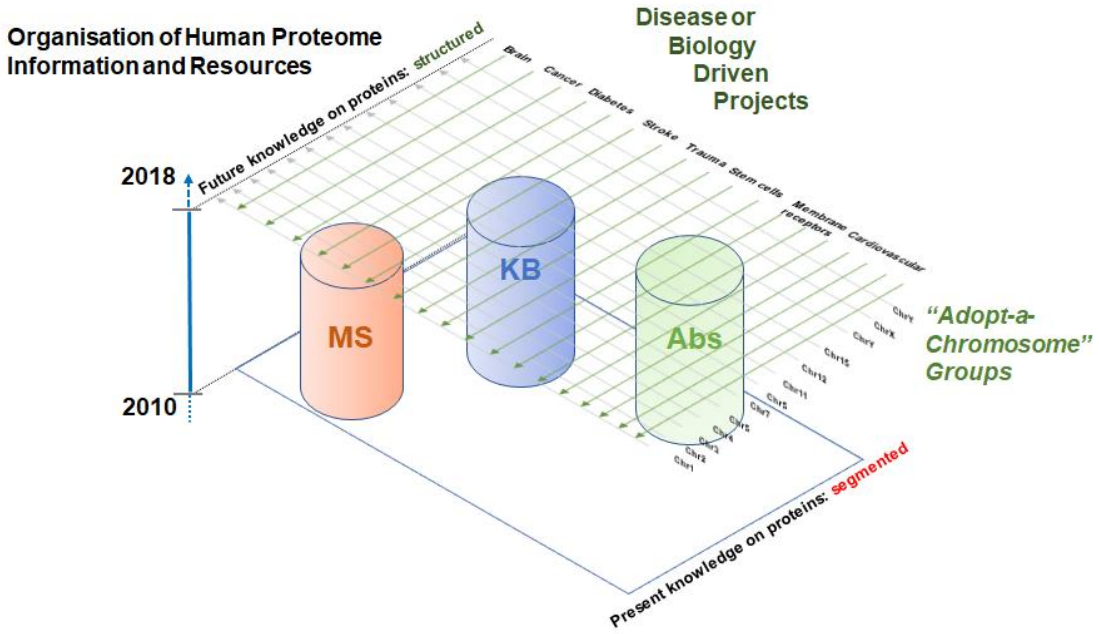


Figure 1.1: Scheme demonstrating the two foundation streams (incorporating many existing initiatives; C-HPP and B/D-HPP) and three foundation resource pillars of HUPO's HPP*: mass spectrometry (MS), antibody (Abs) and knowledge database (KB) on which the Human Proteome Project (C-HPP and B/D-HPP) is based. Adapted from https://hupo.org/human-proteome-project*

1.2 Three Resource Pillars of HPP

The first pillar of the HPP comprises mass spectrometry-based approaches to analytically identify and quantitate the human proteome. This approach typically involves shotgun-based identification and quantitation of peptides with protein inference from biospecimens generating data for the building of a repository of fragment ion spectra for each of these proteins. These serve as an invaluable asset for developing validation assays using targeted selected reaction monitoring (SRM) or parallel reaction monitoring (PRM) assays through the use of stable isotope-labelled peptide standards. The data obtained from these assays can be stored and shared using databases like PeptideAtlas, SRMAtlas, PASSEL (Kusebauch et al., 2014). This pillar relies on the mass spectrometer's sensitivity and high-throughput capabilities to not only quantify proteins that are known, but more importantly identify proteins whose expression has not been validated, making it the force behind unravelling the human proteome.

The second pillar focuses on connecting protein expression to tissue distribution and subcellular location using affinity/antibody reagents. This resource pillar is also meant to serve as a resource for protein-specific capture reagents for human proteins to support related immunochemistry and affinity assays. This pillar has already resulted in a comprehensive tissue-based human map showing protein expression, termed the Human Protein Atlas (HPA). As of 2018, the HPA represented proteins encoded by 17,000 human genes accounting for ~87% of the human proteome and using 26,009 antibodies (Thul and Lindskog, 2018). The HPA has sub-categorised its data into a range of atlases, including the Tissue Atlas, Cell Atlas, Pathology Atlas and The Cancer Genome Atlas integrating RNA-Seq data alongside immunochemistry data to provide an invaluable and well-curated resource for all life scientists (Uhlen et al., 2010).

Proteomics being an intensive technology-driven science requires robust bioinformatics support to analyse and curate the large amount of high-throughput data being generated by the mass spectrometry (MS) and antibody (Ab) pillars. The HPP knowledgebase compiles, curates, organises and shares data from the other two pillars and both HPP initiatives (C-HPP and B/D-HPP). From the very outset, the HPP drew upon databases like UniProt/SwissProt and this approach was rapidly supplemented by many other repositories. Prominent among these were neXtProt, PRIDE, PeptideAtlas, GPMdb and the Human Protein Atlas (Legrain et al., 2011). Updated datasets and meta-datasets are now routinely shared through ProteomeXchange and other repositories (Legrain et al., 2011), with ProteomeXchange datasets assigned a specific PDX identifier. The overall knowledgebase (KB) is indispensable for interrogating and understanding data derived from streams, initiatives, resource pillars and merging teams.

Together, the pillars comprised the technological resource foundation on which the HPP was built. Recently, (HUPO Orlando 2018 World Congress), the pathology resource pillar was added as a 4th pillar, recognising the translational role/s that clinical pathology is playing in proteomics. Since inception, the HPP has made significant inroads towards accomplishing two overarching goals, namely;

1. deciphering the proteins that compose the human proteome and map protein parts to their respective gene, by quantitating, locating and validating their presence; and

2. integrating proteomics with other –omics technologies like genomics, transcriptomics, metabolomics to empower the biomedical and research community to translate proteomics knowledge into clinical solutions (Omenn et al., 2018).

HPP was initially launched under two central major parts (called streams) to help achieve each of the above goals. The two streams were the Chromosome Centric Human Proteome Project (C-HPP) focused on Aim 1 (Paik et al., 2012), while the Biology/ Disease-driven Human Proteome Project (B/D-HPP) focussed on Aim 2 (Aebersold et al., 2013).

1.3 C-HPP

In its formative years, the HPP allocated each human chromosome (Chr 1-22, X, Y and mitochondrial DNA) to members of a global consortium formed from international research teams with expertise in proteomics and aligned research. This scheme closely aligned with the approach used by its predecessor, the Human Genome Project (Venter et al., 2001), without which much in human proteomics would be impossible to achieve.

The C-HPP initially aimed at identifying and mapping all the protein-coding genes from a human chromosomal complement using physical and functional perspectives. This was achieved by "dividing" the human proteome into chromosomal landmarks. In addition, there was an additional task of characterising protein expression at both the tissue and cellular level alongside identifying protein variants and post translation modifications (PTMs). This part of the HPP was termed the Chromosome Centric Human Proteome Project or C-HPP. The primary goal of C-HPP was to rapidly identify at least one representative protein of then (2011) estimated ~20,300 human genes (Aebersold et al., 2018) predicted to encode proteins. C-HPP also aimed to independently categorise and study proteins based on subcellular characteristics like membrane protein, protein variants, or proteins based on abundance (Aebersold et al., 2013). An ultimate motive behind chromosome-based protein data curation was to increase the overall "uptake" of proteomics alongside multi-omics work by the general research community by integrating datasets into a compatible and easily understandable format (Paik et al., 2012).

The C-HPP was tightly integrated with the three pillars (Figure 1.1). This multinational program aimed at characterising the proteome by mapping the proteome through mass spectrometry-based SRMAtlas, antibody-based Human Proteome Atlas and the bioinformatics knowledgebase and was supported through ProteomeXchange by developments like the

PRoteomics **IDE**ntification database (PRIDE), Tranche, PeptideAtlas, Global Proteome Machine Database (GPMDB), UniProt and finally neXtProt (Paik et al., 2012), which was assigned the official KB reporting role/s for the HPP. To assess the progress of HPP, it was decided to communally categorise proteins in neXtProt based on what was termed "protein existence" or PE1-5 scores. neXtProt initiated the classification of proteins identified in the HPP based on the range of PE scores from 1-5 (Table 1.1). The PE score was a measure of protein evidence based on credible MS data, partial/complete Edman sequencing data, X-ray/NMR structure, protein interaction data and/or detection of proteins using affinity reagents as reviewed by us previously (Baker et al., 2017).

The HPP KB pillar (with inputs from all parts of the HPP consortia) also established high stringency metrics in terms of guidelines for submission of MS data. These included a deliberate effort to bring high stringency data to the fore, by requiring protein identification to only come from two or more uniquely-coding non-nested 9 or more amino acid-containing, peptides that could be confirmed by spectra of cognate synthetic peptides if they were previous missing (i.e., PE2-4) proteins (Paik et al., 2018). These standards extend to data curation by MS portals reflecting the annotation of a protein surpassing the established metrics to qualify as a PE1 protein. The success of the HPP project Phase 1 and all composite supportive HPP parts is directly related to the number of human proteins currently annotated as being PE1 (Baker et al., 2017).

Table 1.1: Protein evidence (PE) as defined by neXtProt database in 2018 (**neXtProt**, 2018)			
PE Group	**Definition**	**Number of Proteins**	**Status**
PE1	Evidence at protein level: Strong evidence of detection by MS or other methods like affinity reagents and/or sequencing by Edman degradation or 3-D structure defined by X-ray/NMR	17,487	Annotated
PE2	Evidence at transcript level but no experimental evidence as in PE1	1728	Missing
PE3	Evidence in homologous species only with no evidence at transcript or protein level as PE1/2	515	
PE4	Evidence at theoretical level only, protein predicted to be expressed by a gene, no homology, transcript or protein expression evidence as in PE1/2/3	76	
PE5	Evidence is dubious due to lack of essential features for transcription and/or mutations in the sequence in the cases of pseudogenes	571	N/A
N/A - Not Applicable			

At the commencement of my doctoral studies, it was curious that over 2013-2016, transition of proteins from PE2-4 to PE1 increased by only 5%. One of the first endeavours during my PhD, was to review the possible reasons for this data and to explore measures to accelerate the missing proteome discovery. An extensive bioinformatics analysis was performed to find trends in the most prolific PE1 proteins and the class to which they map and compared them to similar analysis for missing PE2-4 protein classifications on basis of neXtProt protein descriptions, as a surrogate grouping for protein families. Our collaborative observations are detailed in the next published thesis sub-section.

1.4 Study I: Accelerating the Search for the Missing Proteins in the Human Proteome (Publication I)

nature COMMUNICATIONS

REVIEW

Received 16 May 2016 | Accepted 6 Dec 2016 | Published 24 Jan 2017

DOI: 10.1038/ncomms14271 OPEN

Accelerating the search for the missing proteins in the human proteome

Mark S. Baker[1], Seong Beom Ahn[1], Abidali Mohamedali[1,2], Mohammad T. Islam[2], David Cantor[1], Peter D. Verhaert[3], Susan Fanayan[1], Samridhi Sharma[1], Edouard C. Nice[4], Mark Connor[1] & Shoba Ranganathan[2]

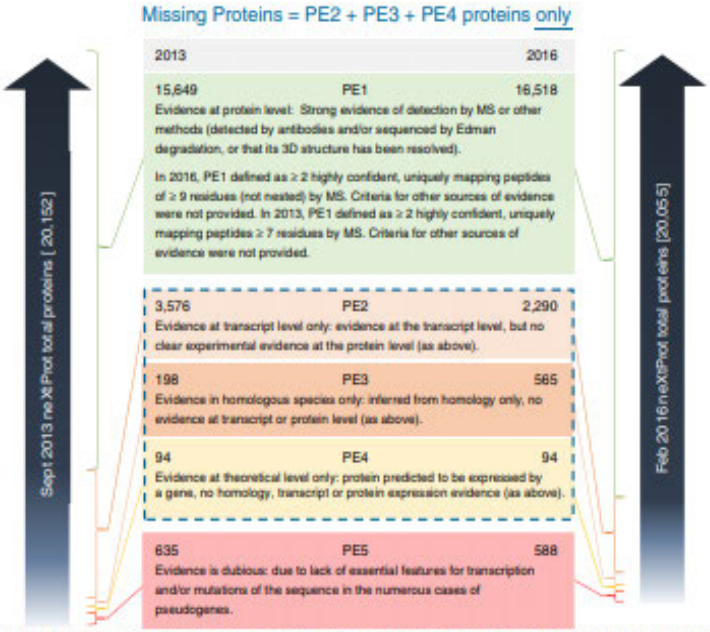
The Human Proteome Project (HPP) aims to discover high-stringency data for all proteins encoded by the human genome. Currently, ~18% of the proteins in the human proteome (the missing proteins) do not have high-stringency evidence (for example, mass spectrometry) confirming their existence, while much additional information is available about many of these missing proteins. Here, we present MissingProteinPedia as a community resource to accelerate the discovery and understanding of these missing proteins.

The Human Proteome Project (HPP) supports defining what it is to be human in molecular terms. It strives to 'know thyself' by finding high-stringency evidence for the ~20,000 proteins encoded by the human genome. Here, we focus on what has been termed the human proteome's 'missing proteins', discuss what renders them currently unobservable using high-stringency proteomic approaches, and outline a road-map that aims to accelerate the HPP. We review milestones and the progress of this global scientific effort to accurately identify and understand the biology of genome-coded human proteins. We focus on what has been achieved to date and we identify some areas where progress may be made. We provide a comprehensive survey of the characteristics of the so-called 'missing proteins', a term initially coined by Hancock and colleagues defined in Box 1 (refs 1,2), and we emphasize why they may be difficult to detect using mass spectrometry (MS) and/or validated antibody (Abs) techniques. Our re-analysis of publicly available MS data for the largest family of missing proteins (olfactory receptors), viewed in conjunction with other specific missing protein examples reveals a need for the community to capture as much complementary evidence as possible about missing proteins, in addition to high-stringency MS data. With this aim, we launch MissingProteinPedia (http://www.missingproteins.org), a community biological database that is complementary to the high-stringency HPP methodologies currently underway. MissingProteinPedia is a low-stringency communal database that will increase our understanding of the spatiotemporal biology of missing proteins, and accelerate their discovery by high-stringency MS.

[1] Department of Biomedical Sciences, Faculty of Medicine & Health Sciences, Macquarie University, New South Wales 2109, Australia. [2] Department of Chemistry & Biomolecular Sciences, Macquarie University, New South Wales 2109, Australia. [3] Department of Biology, Antwerp University, Antwerpen 2020, Belgium. [4] Department of Biochemistry and Molecular Biology, Monash University, Victoria 3800, Australia. Correspondence and requests for materials should be addressed to M.B. (email: mark.baker@mq.edu.au).

Box 1 | neXtProt protein evidence (PE) definitions and 2013→2016 PE data comparison.

neXtProt assigns every one of the 20,055 human proteome proteins as either PE1-5, using evolving communal metrics that have become stricter to improve identification confidence. HPP PE status from 2013→2016 is shown below (with protein numbers indicated top left and right of boxes).



The PE2-4 proteins are now considered as the missing proteins since insufficient evidence has been produced as per the HPP metrics. The criteria for categorizing PE status, using data other than MS, remains to be communally defined. neXtProt protein data are constantly updated, so PE numbers vary with each version release.

Human Proteome Project (HPP) goals and progress

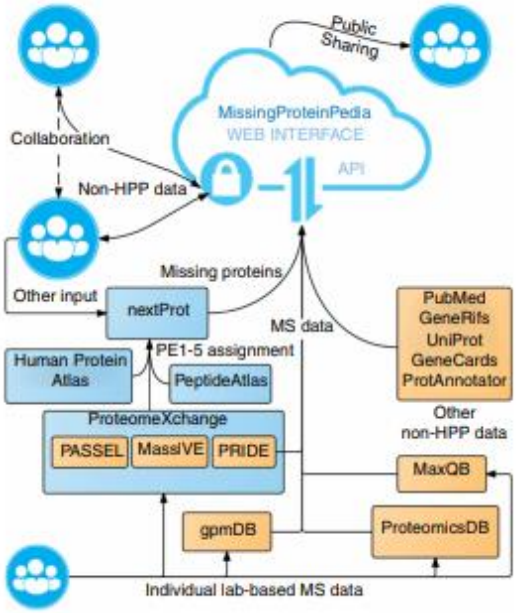
Science is rapidly becoming a global endeavour, with high-quality curation and annotation of data becoming the responsibility of the whole scientific community. Despite the Delphic maxim 'know thyself' being inscribed on the forecourt of the Temple of Apollo in ancient Greece during the sixth century BC, we still do not have a comprehensive description of what it means to be human in strictly molecular terms (that is, genome + epigenome + transcriptome + proteome + peptidome + metabolome). In 2010, the Human Proteome Organization (HUPO) formally initiated a flagship project called the Human Proteome Project (HPP). This ambitious project contributes to humans knowing themselves by collecting credible, high-stringency MS and other evidence for the ~20,000 or so proteins coded by human genes. The long-term aims of HPP are twofold. First, it aims to complete the protein 'parts list' of *Homo sapiens* by identifying and characterizing at least one protein product and as many post-translational modifications, single amino acid polymorphisms and splice variant isoforms as possible for each protein-coding gene. Second, it aims to transform proteomics so it becomes complementary to genomics across clinical, biomedical and life sciences, through technological advances and creation of knowledgebases for the identification, quantitation and characterization of the functionally networked human proteome.

In order to ensure all encoded proteins would be revealed and that all important biology and diseases would be represented, the HPP was amalgamated under two distinct but overlapping streams called the chromosome-centric (C-HPP) and Biology/Disease (B/D-HPP) Human Proteome Projects[3]. These are underpinned by three resource pillars; (i) MS, (ii) Affinity Reagents (for example, Abs), and (iii) a Knowledgebase. In addition to re-analysing and reporting HPP data, a number of complementary groups (PeptideAtlas; http://www.peptideatlas.org, neXtProt; http://www.neXtProt.org, GPMDB; http://www.gpmdb.org and Human Protein Atlas (HPA); http://www.proteinatlas.org) work cooperatively to provide annual HPP updates, present chromosome-by-chromosome tabulations, evolve high-stringency HPP data analysis metrics[4,5], and supply HPP data deposition guidelines for all researchers[6]. Critically, the HPP consortium encourages concurrent raw data deposition through standardized MS portals (for example, ProteomeXchange; shown as a schema in Box 2). The HPP also undertakes critical, annual re-analyses and reporting of the growing MS dataset with accompanying metadata using community-approved, high-stringency metrics.

The desire to build a reproducible, definable, metrics-driven, annotated HPP of the highest quality necessitated the imposition of terms defining the categories of evidence obtained. To enable this, it was communally agreed that the protein-centric knowledge platform neXtProt[7,8] would classify HPP proteins by protein existence (PE), based on partial/complete Edman sequencing, identification by MS, 3D structure (X-ray/NMR), good quality protein–protein interaction data and/or detection of a protein by validated Abs (for example, in the HPA[9]). Metrics, guidelines and/or PE categories have been agreed on and revised through community forums, facilitated by HUPO. Since the HPP was launched in 2010, we have learned many lessons. The importance of 'speaking the same language' with regard to MS analysis metrics and data submission guidelines has been prominent. Kim *et al.*[10] and Wilhelm *et al.*[11] proposed drafts of the human

Box 2 | Integration of MissingProteinPedia with HPP.

The *MissingProteinPedia* is a publicly available protein data and information sharing web system that aims to collate any relevant data pertaining to any PE2-4 protein. At its core is a flexible schema-based database-driven web system allowing captures of all PE2-4 protein PubMed data, based upon gene and protein including synonyms. The database also allows unpublished, preliminary or proprietary data (for example, antibody, MS, cell biology and genetic studies) to be shared with collaborators via a protected interface.



Schema 1: The *MissingProteinPedia* collates and displays protein information from existing databases using various web services and application programming interfaces. Furthermore, the web interface allows researchers to collaborate and share data not available through other databases. The schema includes the recent illustration of the high-stringency HPP metrics engine[9].

MissingProteinPedia facilitates HPP cross-disciplinary collaboration by providing a complementary, unfiltered, lower stringency perspective to both the HPP metrics and guidelines approaches, enabling community evaluation and scrutiny. *MissingProteinPedia* incorporates text mining technology to fetch and search accumulated UniProt, GeneCards, GeneRifs, PubMed and ProtAnnotator PE2-4 data. In addition, *MissingProteinPedia* summarize publicly available MS data from PRIDE, GPMDB, ProteomicsDB and MaxQB for relevant PE2-4 proteins. It also allows community users to annotate data and administrators to curate information before web publication.

proteome in 2014. These studies challenged the imposition of communal metrics, including previously agreed consensus regarding protein target-decoy false discovery rates (FDRs) and requisite minimum proteotypic peptide length ($\geq$7 amino acids in 2014). The term proteotypic in this context refers to a human peptide sequence of any length found by MS that is uniquely derived from a single known human protein expressed by the genome. The term is often used interchangeably with the commonly used terms, uniquely expressed and unitypic. In the HPP, proteotypic peptides (that is, two proteotypic peptides of suitable length) are employed to identify the expression of a human protein by MS methods. Discussion around the impact of single amino acid variation on application of the term proteotypic are currently underway.

Conclusions from both the human proteome drafts[10,11] were considered contentious[12,13] because they chose to interrogate MS findings using different metrics to those established by the HPP after communal agreement. Because of debate around these publications, large-scale heterogeneous datasets were recognized as raising questions related to assumptions around FDR protocols[12]. Encouragingly, positive, collaborative, communal efforts (for example, revised data deposition guidelines and clear diagrammatic representations of data re-analysis workflows and metrics) are underway and will resolve many of the issues raised. In response, the HPP Knowledgebase pillar proposed more rigorous metrics for substantiating claims of the identification of previously unobserved proteins (that is, PE2-5 proteins; Box 1). It has been proposed that datasets should be culled at 1% protein FDR with additional estimates of peptide and peptide spectral match (PSM) level FDRs and notification of the numbers of proteins, peptides and spectra passing/failing these thresholds. In late 2015, PeptideAtlas proposed increasing the minimum thresholds to two proteotypic peptides of $\geq$9 amino acids with raw spectra to be made publicly available (downgrading 432 previously validated PE1 proteins)[4]. Some exceptions included predicted proteins that are unable to be cleaved to form at least two tryptic proteotypic peptides of required length[4]. While neXtProt initially retained less stringent criteria thresholds of two proteotypic peptides of $\geq$7 amino acids or one proteotypic peptide of $\geq$9 amino acids (that is, with consequent downgrading of 20 PE1 proteins), in February 2016 they aligned with the more stringent PeptideAtlas metrics. These developments were incorporated into both the 2016 HPP metrics and HPP guidelines for data submission that have been recently

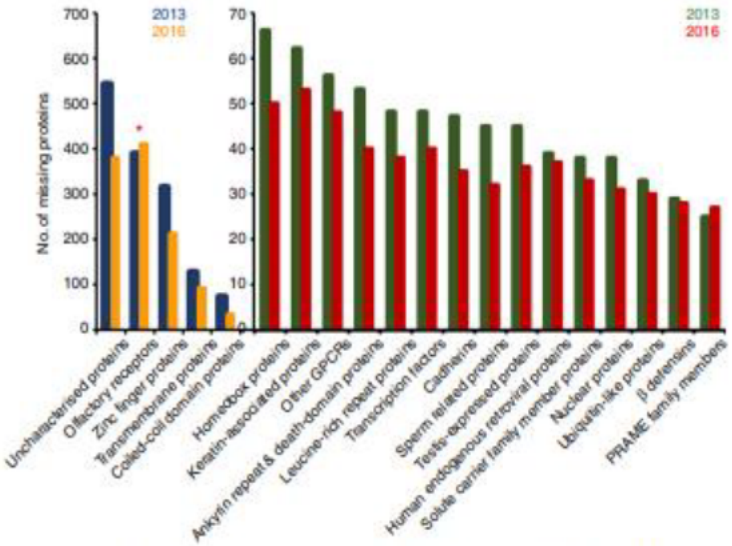


Figure 2 | Top 20 missing protein families to determine protein families enriched in the February 2016 neXtProt PE2-4 report list. According to these data, olfactory receptors (ORs; marked with a red asterisk *) represent the largest family of PE2-4 proteins. The olfactory receptors also show the largest increase between 2013 and 2016 (that is, 15% in 2016 from 10% in 2013) when compared to the other families. The scale '0–70' represents a magnified axis scale for protein descriptors having <70 missing proteins. Blue and green colours represent PE2-4 proteins from 2013 whereas orange and red colours represent 2016 missing proteins.

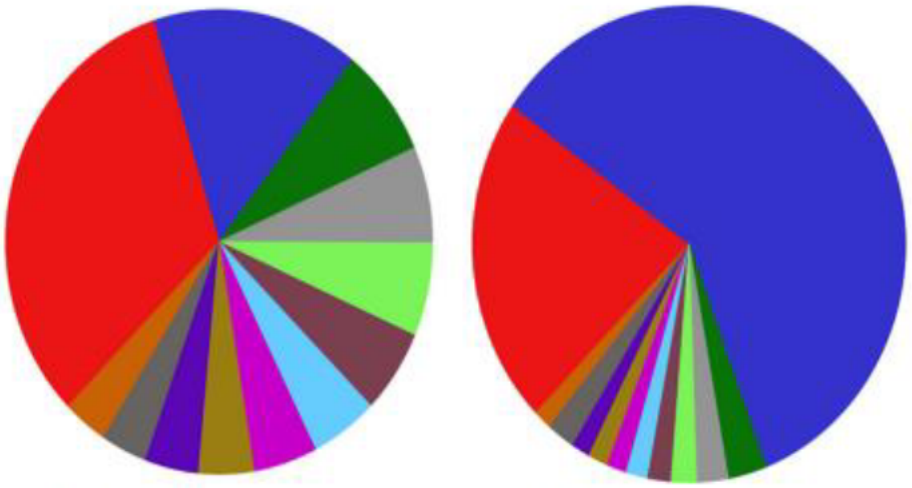


Figure 3 | Most prolific PE1 and 12 PE2-4 UniProt protein families represented in the HPP neXtProt February 2016 release. The most represented PE1 families (left hand side) are the Krueppel zinc-finger protein family followed by the G-protein coupled receptor 1 family. These two families are also at the top of the PE2-4 category (right hand side) with the order reversed.

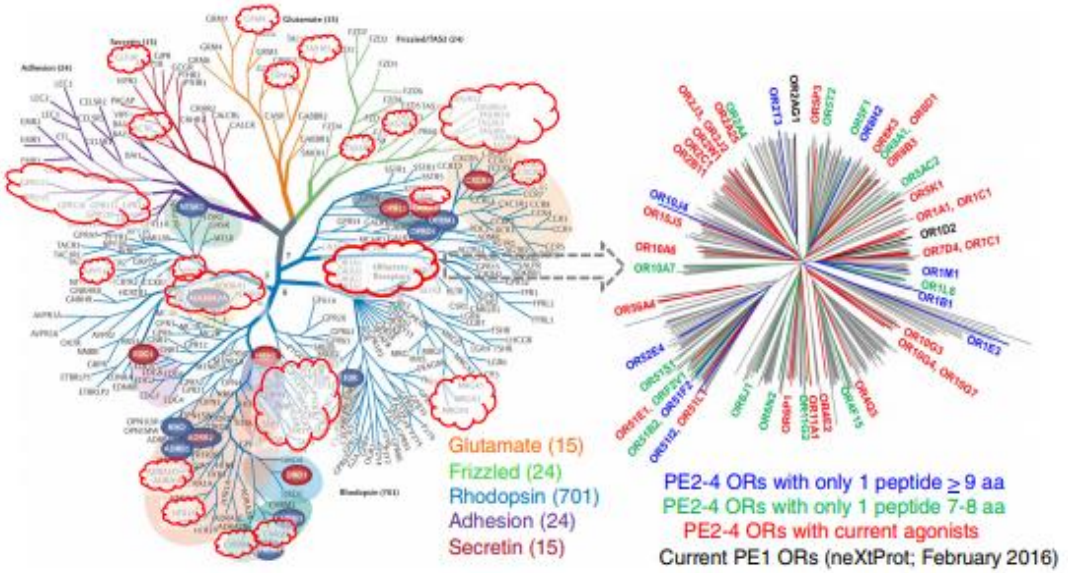


Figure 4 | Phylogenetic analysis of PE distribution across GPCRs and olfactory receptors. In this composite figure, GPCR (left) family branches (largest 'receptor' subset of all human and the PE2-4 proteins) are shown in an unrooted phylogenetic tree from Panther analyses with PE2-4 GPCRs highlighted inside red clouds, and an unrooted GCPR subset phylogenetic tree showing olfactory receptors (right) was produced using iTOP[56], from neXtProt February 2016 PE1 olfactory receptors or best available, manually validated proteotypic MS evidence for olfactory receptor was retrieved. olfactory receptors with functional activity (known agonists) are shown in red in the left figure, as from Mainland *et al.*[16]. GPCR figure modified with permission from Macmillan Publishers Ltd: *Nature Reviews. Drug Discovery*, Stevens *et al.*[57] copyright 2013.

of HPP PE1 calls, the statistical analysis of neXtProt is particularly telling, with a recent lag/hiatus evident. Equally, extrapolating PeptideAtlas data alone suggests 95% completion somewhere around 2030–40.

Orthogonal efforts to find missing proteins

A major outcome from the C-HPP effort to date has been that researchers have been made to consider possible reasons why PE2-4 proteins have not been found by MS, Ab-based or other methods. This has now inspired the development of novel strategies to find the PE2-4 proteins, or understand why they are missing. Some approaches, envisaged to date, include subcellular enrichment of families, groups, clades or classes (for example, membrane proteins); more extensive protein and peptide fractionation before MS; increased MS accuracy, sensitivity and throughput; more reliable, specific and accurately validated Ab technologies, which are currently underway with collaborative efforts by the HPP Ab technology pillar; scrutiny of proteins not amenable to tryptic digestion, those failing to yield 'flying' tryptic peptides or those outside observable mass range detection settings[14]; analysis of cross-linked or otherwise insoluble proteins; examination of rare human tissues/cells under differing spatiotemporal conditions or differentiation states; exposure of tissues to pathophysiological and/or environmental cues, and finally; broadening the capture of data from solely MS and Ab-based data streams.

Bioinformatics efforts to understand missing proteins

Given the current scientific and protein informatics data detailed in Supplementary Table 1 and with a view to finding more PE2-4 proteins, we additionally undertook bioinformatics analyses of all PE2-4 proteins according to their families, sub-families, clades, groups, ontologies, pathways and networks. Figures 2–4 summarize these analyses with increasing depth across neXtProt descriptors (Fig. 2), comparison of protein biologies between PE1 and PE2-4 (Fig. 3), and PE2-4 G protein-coupled receptor (GPCR) family (Fig. 4, left) and OR* (Fig. 4, right) clade phylogenetic tree analyses, focussing on the most populous protein families from Figs 2 and 3.

Analyses of major descriptors (that is, protein subfamilies, classes, domain-type) for neXtProt 2016 PE2-4s indicated that five groups of proteins were highly represented. The PE2-4 groups with greater than 50 members in decreasing order are: olfactory receptors (red * in Fig. 2), zinc finger proteins, non-GPCR transmembrane proteins, coil-coil domain proteins and homeobox proteins (Fig. 2). Encouragingly, our analysis demonstrates a decrease in the percentage of HPP PE2-4 proteins assigned as 'uncharacterized' by neXtProt over the 2013–16 period. These data also demonstrate the substantial success made across all major (that is, the top 20) protein groups, with the sole exception of the enigmatic olfactory receptors. In agreement with these data, Panther Protein Class analysis of 2,491 classifiable genes confirmed the major PE2-4 protein types were: receptors (PC00197), transcription factors (PC00218), transferases (PC00220), transporters (PC00227), membrane traffic proteins (PC00150), enzyme modulators (PC00095) and signalling molecules (PC00207), with other groups represented at low percentages.

Analysis of the top 12 UniProt families found in the 2016 PE2-4 and the PE1 lists (Fig. 3) demonstrates a highly significant enrichment of GPCR type 1 family missing proteins, and a reduction in the % of zinc finger proteins in the PE2-4 proteins list. Furthermore, we note that when the highest 12 families are

examined in the PE2-4 list, the vast majority of those families' members are found to be 'missing', with relatively few PE1 representatives. Only three families (that is, Kruppel C2H2-type zinc finger, GPCR type 1 and Peptidase C19 protein families) were common to both the major PE1 and the major PE2-4 families. Interestingly, PE1 assignments account for only 22% of all GPCR type 1 proteins while it accounts for 59% of the Kruppel zinc finger proteins. If one considers only the PE2-4 'missing' proteins, GPCR type 1 members represent 25% and zinc finger family members 9%. On a family-by-family basis, apart from Kruppel zinc finger (34%) and peptidase C19 (31%) proteins, the remainder of the top 12 families are noticeably composed of missing proteins (that is, range 50–95% of the total family membership). This implies that when a major family is 'missing' by current HPP metrics, extremely limited high-stringency MS knowledge exists for any member of that protein family (for example, of 22 known PRAME proteins 19, 86% are assigned as PE2-4 and re-analysis of olfactory receptor MS data summarized in Supplementary Table 2 shows all (100%) are currently missing).

Box 3 | Accelerating discovery of the complete human proteome.

We recognize the tremendous achievement the Human Proteome Project has made since its 2010 launch by making available high-quality, communal MS (and other) data for ~82% of the human proteome (February 2016).

To accelerate discovery of the 15% of the human proteome defined as the missing PE2-4 proteins, we recommend and encourage the following:

1. All proteomics practitioners, human researchers and human biology/medicine journals renew their efforts to observe current high-stringency HPP re-analysis *metrics* and researcher data submission *guidelines*.
2. All MS data should be incorporated into a single database (for example ProteomeXchange), including MS databases not currently captured, where data are provided transparently for any claim for a current PE2-4 protein.
3. The HPP should communally develop metrics and guidelines for processes by which they deal with all non-MS data sources. In particular, transparency around how protein evidence scoring for non-MS data needs to be communally accepted and reported.
4. An annual jamboree to evaluate and approve both MS and non-MS protein evidence reclassification proposals.
5. All possible biological data concerning the PE2-4 missing proteins to be comprehensively captured in *MissingProteinPedia*.

The olfactory receptor family missing proteins

Subsequently, we examined the largest PE2-4 family, namely human GPCRs (shown in dark blue in Fig. 3). These are responsible for cellular responses to everything from protons and photons to hormones of >30 kd, metals, nutrients, small molecules including volatiles and neurotransmitters through many of our major senses (that is, sight, olfaction and taste). GPCRs also are the most important pharmaceutical drug target and largest family (>800) in the human proteome, as well as the largest membrane receptor family. They instigate signalling through nucleotide exchange involving heterotrimeric G-proteins and can be classified into five major families and subdivided into subfamilies based on sequence homology, to (1) rhodopsin (class A), (2) secretin, (3) adhesion (class B), (4) glutamate (class C), and (5) Frizzled/taste receptor 2 (TAS2). Phylogenetic analysis of GPCR PE2-4 proteins demonstrates that although singleton representatives and a few clusters are distributed across all five major subfamily branches/classes (Fig. 4), by far the highest proportion of missing proteins ($n=400$; ~15% of all human PE2-4 proteins) emanate from the rhodopsin branch of the unrooted GPCR phylogenetic tree where the olfactory receptors reside. Note that family members with determined crystal structures are highlighted on the phylogenetic tree in coloured ovals (including ADORA2A, which has been recently re-classified by neXtProt as PE1).

Discovering functionality of the complete missing human olfactory receptor repertoire has proved difficult with only 49/~400 human olfactory receptors having known ligands before the recent studies of Mainland *et al.*[16]. Using high-throughput screens of human olfactory receptors against 73 potential ligands they identified agonists for 27 receptors (coloured red in Fig. 4, right), including 18 that were previously orphan receptors. Their dataset addressed a bottleneck in research around functionality of human olfactory receptors by showing how physical olfaction stimuli can signal post-receptor activation. Correlating odorant ligands to olfactory receptors provides a valuable database, identifying functional olfactory receptors with potential to be strategically targeted through proteomic approaches and subsequent conversion to PE1 proteins.

The recent studies by Kim *et al.*[10] and Wilhelm *et al.*[11] generated intense interest in MS evidence for the expression of the chemosensory olfactory receptor family, as they claimed to have 'unearthed' a surprisingly high number of 108 and 200 PE2-4 olfactory receptors, respectively. Of the human genome's 480 olfactory receptor genes in the latest version of neXtProt, 12 are considered hypothetical or putative (PE5). The remaining 468 olfactory receptor genes code for 411 unique proteins, with only two classified as PE1, and the remaining 409 classified as PE2-4. The claims for finding missing olfactory receptors by the draft human proteome papers above were rapidly critiqued by Ezkurdia *et al.*[12] and Deutsch *et al.*[13] on the basis of marginal spectral quality, deficiency of stringent protein/peptide 1% FDR criteria, use of short peptides, and erroneous or potentially ambiguous peptide identification, with the suggestion that these claims represent 'the cream of false positives'. Collectively, these errors led Ezkurdia *et al.*[12] and Deutsch *et al.*[13] to conclude that there was little evidence for even a single olfactory receptor (including the two listed in previous releases of PeptideAtlas). Incidentally, 10 olfactory receptors were considered 'found' by Choong *et al.*[17] in the 2015 release of neXtProt with MS and Ab evidence. However, this evidence was considered insufficient for all these 10 olfactory receptors, suggesting that currently known olfactory receptor proteins may not possess sufficiently documented protein evidence in neXtProt.

From the amazing repertoire of 411 unique olfactory receptor proteins, only two are currently considered PE1 in the neXtProt 2016 release (namely, OR2AG1 and OR1D2; coloured black in Fig. 4, right). For OR1D2, no MS or Ab evidence is available, with three publications cited as functional evidence. For OR2AG1, neXtProt reports a single peptide 7 amino acids long, with no Ab evidence and functional evidence from two publications[18,19]. One of these studies[18] equally reports function for another olfactory receptor, namely OR1F12 but this remains classified by neXtProt as PE4, whose status is based upon sequence homology. Thus, it appears that both these PE1 olfactory receptor proteins do not actually conform to HPP MS-based metrics and require

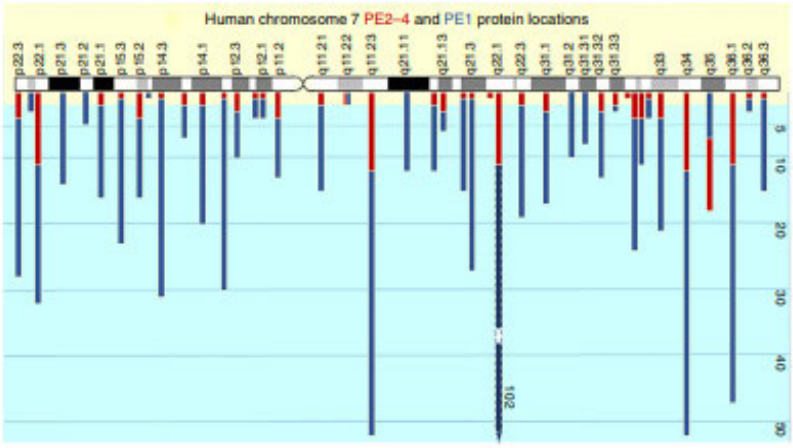


Figure 5 | Positional mapping of the PE1 (757) and PE2-4 (139) proteins along human Chr 7. The data show random distribution of both along the complete length of human Chr 7. However, Giemsa banding patterns of light (GC-rich) and dark (GC-poor) bands are shown that debatably correspond to regions of gene density from light (higher gene density) to dark (lower gene density)[58].

closer community examination (Box 3), as does the way we consider functional/biological data as evidence for PE.

Olfactory receptors are involved under most physiological situations with odour recognition but have recently been shown to be expressed in multiple epithelial tissues with many potential chemosensory roles[20–22]. Criticisms of olfactory receptor restriction to nasal epithelial tissue are ill-advised[11] and appear erroneous[12,20–22]. Given these data and the comprehensive olfactory receptor functional studies conducted by Mainland *et al.*[16], we believe that a systematic capture of non-MS data and a communal re-assessment of all olfactory receptor PE assignments would be timely. To bring additional perspective to the olfactory receptor mêlée and to emphasize the challenges we face in finding the missing olfactory receptors by high-stringency MS, we undertook an analysis of all currently available raw olfactory receptor spectra from public repositories. This re-analysis reinforces that the best available MS data fail to provide high-stringency PE1 level proof for any GPCR olfactory receptor members using current metrics (Supplementary Fig. 1 and Supplementary Table 2). Despite 2,361 manuscripts revealed by an 'olfactory receptor and human' PubMed keyword search, only piecemeal MS evidence for any human olfactory receptor is currently available.

To verify the *status quo*, we trawled public MS proteomic repositories (including GPMDB, PRIDE, ProteomicsDB, MAXQB and Human ProteinPedia), and aggregated 122,717 peptide MS entries (PSMs of length $\geq$7 aa), including many with multiple PE2-4 olfactory receptor observations. This collective dataset was processed through a semi-automated workflow (Supplementary Fig. 1), including manual spectral validation to filter reliable peptide assignments, with consideration of leucine/isoleucine ambiguity and BLAST analysis to account for possible single amino acid variations coding for peptides, as detailed elsewhere[23]. Briefly, the data (using Batch Peptide Match) identified 4,751 proteotypic olfactory receptor peptides (3.9%), following removal of non-proteotypic and decoy peptides. Of the proteotypic peptides, only 286 (6%) were tagged with a high search engine confidence value score by either SEQUEST, Mascot or MaxQuant. Finally, manual spectral validation (taking into consideration, noise, error rates (to matched peptide sequence), the run of B and Y singly charged ions, unassigned peaks and relative intensity of the spectrum) allowed us to sift out 64 high quality spectra for 24 peptides. As two overlapping peptides could be merged for a single olfactory receptor, this culminated in 23 unique olfactory receptor peptides. In summary, this analysis provided MS evidence for 23 of 409 missing olfactory receptors (5.6%).

The best available MS evidence for these 23 olfactory receptors is shown in Supplementary Table 2, and it includes peptides from GPMDB (1 green, 1 yellow and 5 red peptides), PRIDE (10 peptides) and ProteomicsDB (7 peptides). It should be noted that 14 PSMs represent a single 7–8 amino acid peptide, while 9 possess a single PSM of >9 amino acids. Proteins derived from matches were cross-referenced against HPA with no (zero) olfactory receptors found in the current (May 2016) high confidence HPA premium dataset. In addition, 13 peptides (Supplementary Table 2) were found to have complete or partial matches with 14 SRM peptides listed in the current version of SRMAtlas.

In summary, we demonstrate that many missing PE2-4 olfactory receptors possess single high-confidence PSM evidence, although best available MS spectra are insufficient to meet current HPP metrics. These could be considered as PE2-4 proteins 'waiting in the wings', requiring confirmatory proteotypic PSM identifications at the required length to reach high-stringency requirements.

Chromosome 7 example missing proteins

Under the C-HPP, the proteomic information found across chromosomes 1-22, X, Y and mitochondrial DNA are being studied by country-based or regional cluster teams. Australia and New Zealand undertook analysis of the proteins coded by human chromosome 7 (Chr 7)[24,25]. As part of our ongoing efforts, we demonstrate that current PE2-4 proteins are located across the length of the long and short arms, approximately equally dispersed across the length of Chr 7 (Fig. 5). This holds true for the majority (but not all) chromosomes examined to date. At one chromosomal location, namely 7q35, a significantly greater number of PE2-4 proteins (18/25) were found than PE1 proteins (7/25). Interestingly, however, when Giemsa

(that is, reported relative gene richness) staining patterns along Chr 7 were compared for PE2-4 and PE1 distribution, we observed that 56% PE2-4s emanate from high gene density Chr 7 regions, 12% from moderate, 25% from low-moderate and only 1.5% from regions of low gene density. PE1 proteins generally distribute across Chr 7 locations with PE2-4 proteins, with few regions (only p22.2, p21.3, p21.2, p15.1, q21.11, q31.2 and q31.31) not having both PE classifications represented. Chr 7 PE2-4 proteins do not emanate from gene-poor regions and hence it is reasonable to suspect that other factors (for example, low spatiotemporal expression) are more likely to explain why they have not been found by high-stringency MS to date. These observations need to be replicated for all chromosomes by other C-HPP teams.

Of the 134 Chr 7 PE2-4 proteins, 27 are known to be GPCRs. The majority of these encode olfactory (15) or taste-related (six) receptors, with only four 'orphan' GPCRs and two well-described GPCRs (5-HT_{5A} and mGluR8). There are many reasons why these proteins may still be considered missing. First, they all have restricted anatomical expression. In particular, the receptors for odours and ingested chemicals, which are likely expressed in only a few cells in specific regions of the body. Further, many missing proteins may be localized to a few discrete cells and/or difficult to access cellular compartments, like axon terminals, inner/outer hair cells (OHCs) or cilia on olfactory sensory neurones. Second, receptor expression may be extremely low even where they are physiologically active. Finally, it is possible that gene products are not translated/transcribed under normal physiological situations, or indeed at all. Their absence from proteomic databases suggests they are not highly abundant but it does not mean they are not important or not expressed. Indeed, a cursory examination of Chr 7 PE2-4 GPCR proteins reveals many non-proteomic studies show these GPCRs represent a very active part of the human proteome. Using the BPS/IUPHAR Concise Guide to Pharmacology (http://www.guidetopharmacology.org/index.jsp)[26] as a starting point for analysis, we provide some examples. First, HTR5A is part of the large family of receptors for the neurotransmitter serotonin (5-HT). When expressed, 5-HT_{5A} receptors stimulate G protein activity resulting in inhibition of adenylyl cyclase[27], indicating it is a functional GPCR. mRNA for 5HT_{5A} receptor has been detected in the human brain by *in situ* hybridization[28] and PCR[29]. However, our search shows no reports of protein localization by immunohistochemistry or identification by western blot in any human tissue. Mice with a 5-HT_{5A} receptor deletion have altered behaviour and a distinct response to the serotonin receptor ligand LSD[30], indicating the protein is functional. It is likely that low levels of protein and restricted anatomical localization preclude identification of 5-HT_{5A} receptors by MS.

A second receptor we considered is GRM8 (metabotropic glutamate receptor 8, $mGlu_8$), which is part of the large family of receptors for the prominent neurotransmitter glutamate. In a heterologous expression system, activation of $mGlu_8$ receptors results in inhibition of adenylyl cyclase[31], indicating it is a functional GPCR. *In situ* hybridization reveals discrete but low levels of mRNA in human brain[32,33], while $mGlu_8$ mRNA has been reported in cancer cell lines[34], hippocampal cells[35], astrocytes[36] and in patient tissue in epilepsy or multiple sclerosis. Murine deletion of $mGlu_8$ affects hippocampal synaptic transmission[37], suggesting function under physiological conditions. Low levels and restricted anatomical localization may preclude identification of $mGlu_8$ receptors by MS, although the receptor is also large and has a complex genetic structure, which probably leads to alternatively splice transcripts, and potentially several protein species[33,38].

Finally, GPR22 (Probable G-protein coupled receptor 22) is a class A GPCR, with mRNA expressed in human heart and brain[39–42]. Interestingly, GPR22 has an unusually AT-rich mRNA, and only when enrichment is artificially rectified by introduction of G-C bases can signalling be restored in heterologous expression systems (Gi/o-mediated stimulation of G protein activity and constitutive inhibition of AC activity[41]). No ligand has been identified for GPR22, and GPR22 knockouts seem physiologically unremarkable. However, GPR22 mRNA is significantly reduced by aortic banding, a procedure that mimics cardiac hypertrophy produced by high blood pressure, and in GPR22 knockouts heart failure follows more rapidly than in wild type animals, implying a role for responses to cardiac stress[41]. There is no peer-reviewed report of GPR22 immunoreactivity in human tissues, although several corporate sites show neurons and other cells displaying putative GPR22 immunoreactivity. Sera from mice immunized against a human GPR22 peptide label cells in rat heart, although staining suggests GPR22 is restricted to subsets of myocytes[41]. The lack of an identified ligand for GPR22 has dampened enthusiasm for further pursuing functional studies through conventional biochemistry, and coupled with lack of neuronal phenotype in GPR22 null mice, it is not surprising no further attention has been paid to it. Unlike $5HT_{5A}$ and $mGlu_8$ receptors, which likely have roles in normal physiology (even if understudied), there is little evidence to speak for or against function of GPR22, despite mRNA being detected by multiple investigators. However, for even the most obscure (non-olfactory) PE2-4 GPCRs, some evidence exists, suggesting that they are expressed in some tissues under certain conditions.

While we can learn much from an analysis of the Chr 7 PE2-4 GPCR proteins, the reasons for other proteins apparently 'falling through the cracks' and having PE2-4 assignments may be legion. Below, we examine two current PE2-4 examples that appear to have strong biological non-HPP evidence that, combined with the olfactory receptor data above, argue for a broader, community-based, open data base strategy. We propose that opening up the HPP to consider other sources of data might concomitantly accelerate re-classification of PE2-4 proteins to PE1 status through the existing high-stringency HPP workflow.

In an orthogonal approach to understand the Chr 7 PE2-4 proteins, an example was randomly selected. Prestin (gene name *SLC26A5*) retrieved 91 peer-reviewed PubMed manuscripts, with the oldest in 2000 entitled 'Prestin is the motor protein of cochlear outer hair cells'[43], while another was a recent review of structural and functional properties[44]. Antibodypedia unearthed 83 anti-prestin Abs from 15 different vendors (http://www.antibodypedia.com/explore/prestin). Though not listed on the Therapeutic Target database, Drugbank or Binding DB, prestin's substrates are listed as Cl^- and HCO^-_3 by the IUPHAR-DB (pharmacological targets) database[45]. Additionally, the Human Gene Mutation Database lists two prestin missense/nonsense mutations that produce deafness/autism phenotypes (CM075015 and CM124551), with one splice-variant linked with deafness (CS030995). Furthermore, the gene is known to have 15 transcripts. Equally, 12 patients with overlapping copy number variants are listed in DECIPHER: Database of Genomic variants and phenotype in Humans Using Ensembl Resources. Additionally, zebrafish studies captured in ZFIN include several CRISPR targeting agents (http://zfin.org/ZDB-GENE-030131-1566) directed against prestin. In conclusion, this randomly selected Chr 7 PE2-4 protein shows there is copious public functional evidence at the protein level available, despite there being zero high-stringency MS or acceptable Ab evidence.

Particular physiological, cell and molecular factors make prestin intractable to being found by MS. First, it is a bullet-shaped membrane protein that is localized only on the OHCs of

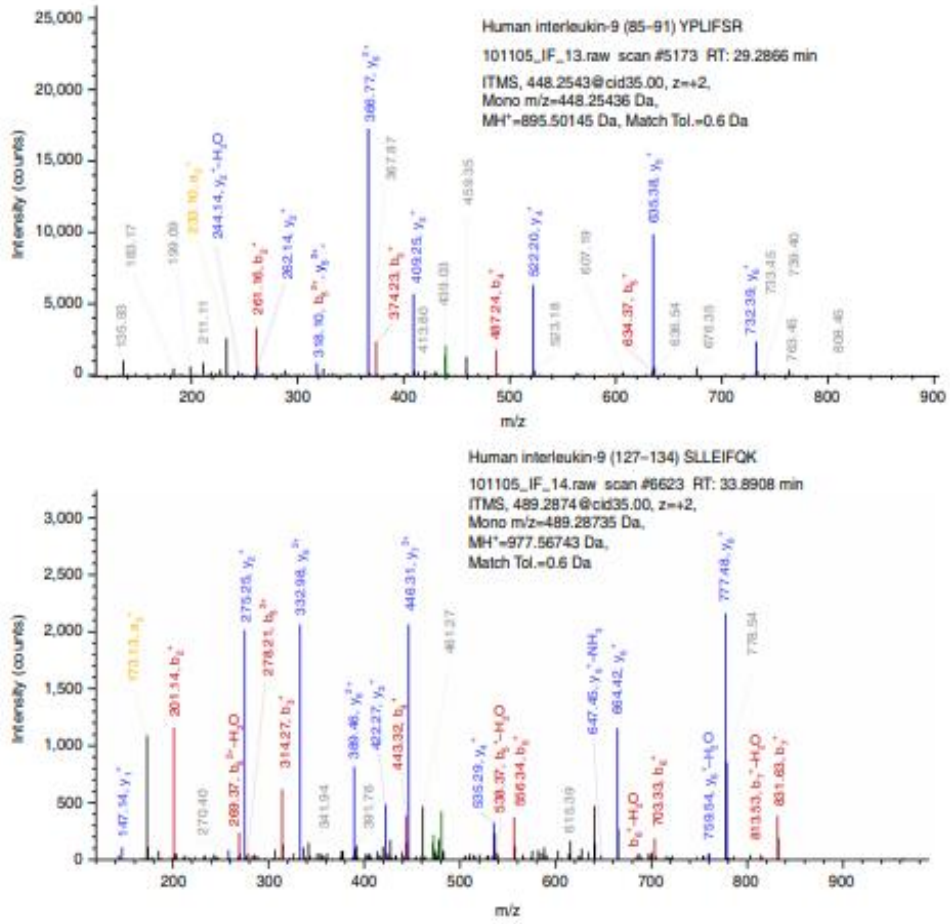


Figure 6 | Fragmentation spectra of two IL-9 proteotypic peptides detected in the secretome of activated T-cells. Although not yet observed in any publicly available MS databases, both of these peptides are predicted to be proteotypic by neXtProt Unicity checker (https://search.nextprot.org/viewers/unicity-checker/app/index.html).

the mammalian inner ear[46]. This presents three challenges; highly specific tissue of origin, low copy number and membrane localization. OHCs are relatively few in number and are in the minority of the cells of the cochlea[47], requiring specialized techniques such as laser capture microdissection to capture cells from very thin tissue sections. Each cochlear microdissection performed by Anderson *et al.*[47] found only 200–300 OHCs per human being, far below the number required for routine proteomic analysis, let alone those involving OHC plasma membrane preparations. Equally, we know that membrane proteins are notoriously resistant to purification and identification by traditional techniques; requiring specialized enrichment strategies due to low copy number per cell, high-hydrophobicity and potential shielding of tryptic cleavage sites by either co-localized membrane proteins or the lipid bilayer itself. It is understandable why prestin is currently a PE2 (transcript evidence only) protein, even though 10 synthetic 10-28mer proteotypic peptides have been reported in neXtProt[45], but no endogenous peptides have yet been captured experimentally by MS.

Interleukin 9 an example missing protein

A number of small biologically active secretory proteins risk being overlooked primarily because of their typical low abundance *in vivo* (in particular relative to the extremely high level of extracellular 'background' proteins), in combination with a specific spatiotemporal expression/secretion profile, a very limited number of predicted potential proteotypic peptides and a relatively high ratio of post-translationally modified residues. One obvious example is the MS detection of interleukin-9 (IL-9) in secretome analysis of post-activation primary cultured T cells. Previous studies of the secretome of cells *ex vivo* had never identified IL-9, as they typically involve only short culture times. To facilitate secretome analysis, typical studies analyse cells grown in serum-free media, inevitably generating considerable cellular stress (with many stress- and apoptosis-related proteins detected). When we analysed cells grown for several days in the presence of foetal bovine serum (described in Supplementary Note 1), a very high percentage ($\approx$95%) of detected tryptic peptides from the conditioned media proteins are evidently of bovine serum origin. After exclusion of bovine proteins and human T cell secretory proteins released from control 'resting' (non-activated) cells, many other secretory proteins (for example, missing interleukins) are now exclusively detected from activated cells. Among these is the 125 amino acid residue, currently PE2 protein, IL-9. MS analyses reveal that IL-9 generates two proteotypic peptides of 7 and 8 residues, respectively (Fig. 6). Subsequent deposition of this and similar data into ProteomeXchange with annual communal re-analysis

with stringent criteria will result in the re-classification of IL-9 as PE1. Similar discoveries accompanied with appropriate MS data deposition are expected to result in the re-classification of PE2-4 missing proteins that are unable to generate any proteotypic peptides acceptable to the HPP metrics, yielding a dramatic increase in the rates of discovery of missing proteins.

Complementary efforts to characterize missing proteins

At present, there are also unrelated efforts (for example, Antibodypedia) to capture standardized, non-HPA affinity reagent data. Abs represent the main thrust one of the three pillars of the HPP initiative, and Ab-based techniques (for example Ab-enrichment, immunohistochemistry, western blot) support the search for the PE2-4 missing proteins[48]. However, issues around validity of Ab data have recently been raised across many forums, including this journal[49]. Key problems revolve around selectivity, acceptability and suitability for a given specific application. To facilitate resolving these issues, efforts are being made (for example, Antibodypedia, HPA) to collect, in searchable databases, detailed information concerning Ab validation and their use, and in some cases, literature performance review. Clearly, careful validation of all Abs is mandatory to allow researchers to make informed choices about suitable reagents with the knowledge that they are specific, selective, fit-for-purpose and reproducible in the context for which they are required[50]. Such validation should include western blot, immunohistochemistry, immunofluorescence, flow cytometry and microarrays, and ideally also Surface Plasmon Resonance data with detailed kinetic information. Where possible, the use of gene knockout/gene silencing (for example RNAi, CRISPR/Cas9) to confirm specificity has also been proposed[51]. Both polyclonal Abs (ideally affinity-purified) and monoclonal Abs have their advantages and disadvantages in the search for the PE2-4 missing proteins. Multiple epitopes, accessible by polyclonal Abs, can facilitate targeting specific proteins in complexes where some epitopes may be masked. They do, however, often have higher non-specific background and cannot be replaced once stocks are depleted. Monoclonal Abs, by contrast, are a renewable resource and typically have high affinity, high specificity and reduced non-specific binding[52], while binding only a single epitope. Furthermore, monoclonal Ab libraries against target proteins can be readily generated[53]. For the missing proteins, a further dilemma is how to obtain an appropriate antigen for immunization. A potentially generic approach is the use of a proteospecific recombinant protein fragment and Protein Epitope Signature Tags (PrESTs)[54]. In a recent study, this approach has successfully generated a panel of monoclonal Abs and affinity purified polyclonal Abs against a number of targets, including some missing proteins[48].

MissingProteinPedia

The availability of large volumes of published, peer-reviewed, credible scientific data for PE2-4 proteins outside of high-stringency PE1 MS and Ab-based evidence (for example, IL-9 and prestin) struck us as a resource we could further exploit. Given the need to accelerate the HPP, we contend that the acquisition of such additional data streams concerning the biology of all PE2-4 proteins is self-evident. This has inspired us to explore, create and launch a communal database called MissingProteinPedia. This database assembles in one repository the vast amounts of publicly available, complementary data about all the current PE2-4 proteins that sit outside of the well-justified, high-stringency HPP pipeline. We contend that the knowledge captured by MissingProteinPedia will accelerate the communal HPP effort, as we seek strategies to allow the generation of high confidence MS evidence for as many PE2-4 proteins as possible. In addition, by providing an assembly of all available biological clues in one repository about every single current PE2-4 protein, it is likely that the MissingProteinPedia database may assist C-HPP chromosomal teams that have accepted the 'Top 50 Missing Protein Marathon Challenge' launched recently at the 15th HUPO 2016 World Congress in Taipei to successfully identify an additional 50 PE2-4 proteins per chromosome to those already found by high stringency methods.

MissingProteinPedia is an open, comprehensive, communal, evidence-based, searchable and sortable (by chromosome, tissue and keywords) community knowledgebase, addressing the HPP's PE2-4 proteins. The launch of MissingProteinPedia aims to capture the broadest level of scientific data necessary to increase the rate at which PE2-4 proteins are validated. MissingProtein-Pedia represents a new community-based proteomics tool, analogous to human genome annotation jamborees[55], where open big data contributions are invited from the broader scientific community regarding evidence for the existence of any missing protein. Unlike the high-stringency HPP data re-analysis, MissingProteinPedia makes no attempt to edit or judge the quality of submitted data, rather utilizing data to expose hidden possibilities not deposited into the current HUPO-accredited databases, including legacy lab books, unpublished works and data found in commercial/protected environments. It is anticipated that MissingProteinPedia collation will reveal clues that will contribute to an acceleration of high quality MS and qualified Ab data that allow confirmation beyond reasonable doubt of many of the current PE2-4 missing proteins. We believe MissingProteinPedia can cooperate and be easily integrated with high-stringency HPP data re-analysis, assisting the completion of the first phase of the HPP on schedule.

In summary, MissingProteinPedia aims to define, summarize and discuss all available data (including single proteotypic MS spectra) for the so-called missing proteins, emphasizing why they may be currently difficult to observe/find, using standard proteomics MS and Ab-based techniques.

Conclusions and the way forward

The HPP was launched in 2010 and since then has grown organically with a general initial phase aimed at providing knowledge about the human proteome parts list. Progress has entailed the formation of a two-pronged strategy (C-HPP and B/D-HPP) culminating in the creation of guidelines and repositories (for example, ProteomeXchange) for MS and Ab-based (for example, HPA) data deposition; metrics for communal, annual MS re-analysis (for example, PeptideAtlas); categorization of the ~20,000 basal components of the human proteome into PE levels (PE1-5; neXtProt); and forums for discussion and communication between research teams (for example, annual HUPO Congresses and HHP workshops).

The controversial release of the two draft human proteome papers[10,11] has compelled researchers to recognize that the HPP is still in its infancy and much remains to be done. This is especially so with regard to the absence of a universally agreed long-term strategy for piloting the project into the future the capture of high-stringency data from all potential MS and Ab sources, capture of the breadth of other scientific human protein data to searchable knowledgebases, and finally the dissemination of the impact and success of the HPP to the public.

Of 20,055 human proteins (neXtProt, February 2016), 16,518 are PE1 (known), a further 2,949 are currently PE2-4 proteins (missing), while 588 PE5 proteins are considered only to be hypothetical. Current PE1-5 assignment strategies do not take into account all other alternative data streams available from the

broader scientific community, preferentially relying on high-stringency MS data.

Analysis undertaken herein demonstrates that the rate of progress of the HPP in finding PE1 proteins needs to be accelerated in order to meet proposed HPP decadal plans. To hasten the progress of the current high-stringency HPP engine, we propose to capture other credible scientific data focussing on the PE2-4 missing proteins. This complementary engine is called the MissingProteinPedia and provides clues in the search for missing proteins, learning more about proteins that fall through the cracks of current data re-analysis. It is our hope that the communal MissingProteinPedia tool will allow researchers to better understand where, how, when and why PE2-4 proteins can be found. Capture of high-stringency data will populate the pool of PE1 proteins more readily and efficiently, building our knowledge of what it is to be human in strictly molecular terms.

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD005656.

References

1. Paik, Y. K. *et al.* The Chromosome-centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **30,** 221–223 (2012).
   Aims to define full set of human proteins encoded by ~20,300 genes, chromosome-by-chromosome including tissue localization, isoforms and PTMs using MS and Abs. First coined term 'missing proteins'.
2. Paik, Y. K. *et al.* Standard guidelines for the Chromosome-centric Human Proteome Project. *J. Proteome Res.* **11,** 2005–2013 (2012).
3. Legrain, P. *et al.* The Human Proteome Project: current state and future direction. *Mol. Cell Proteomics* **10,** M111.009993 (2011).
4. Omenn, G. S. *et al.* Metrics for the Human Proteome Project 2015: progress on the Human Proteome and Guidelines for High-confidence Protein Identification. *J. Proteome Res.* **14,** 3452–3460 (2015).
5. Omenn, G. S. *et al.* Metrics for the Human Proteome Project 2016: progress on identifying and characterizing the human proteome, including post-translational modifications. *J. Proteome Res.* **15,** 3951–3960 (2016).
   Update on HPP annual communal data re-analyses that adopted higher stringency MS metrics for protein evidence (PE1 = two unitypic peptides > 9 residues). HPP (neXtProt version 2016-02) has 16,518 PE1 proteins, with 2,949 PE2-4 missing proteins and 485 reclassified by higher stringency HPP Guidelines v2.0 to reduce false positives.
6. Deutsch, E. W. *et al.* Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* **15,** 3961–3970 (2016).
7. Gaudet, P. *et al.* neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* **12,** 293–298 (2013).
8. Lane, L. *et al.* neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res.* **40,** D76–D83 (2012).
   Describes neXtProt the human protein-centric knowledge platform that supports and reports the HPP.
9. Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).
10. Kim, M. S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581 (2014).
11. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).
12. Ezkurdia, I., Vazquez, J., Valencia, A. & Tress, M. Analyzing the first drafts of the human proteome. *J. Proteome Res.* **13,** 3854–3855 (2014).
13. Deutsch, E. W. *et al.* State of the human proteome in 2014/2015 as viewed through PeptideAtlas: enhancing accuracy and coverage through the AtlasProphet. *J. Proteome Res.* **14,** 3461–3473 (2015).
14. Elguoshy, A. *et al.* Why are they missing?: bioinformatics characterization of missing human proteins. *J. Proteomics* **149,** 7–14 (2016).
   Recent physicochemical analysis of missing proteins, erroneously including PE5 along with the current PE2-4 missing protein definition. Claim 24% PE2-4 proteins possess hydrophobic transmembrane domains and a significant number do not generate suitable unitypic tryptic peptides.
15. Gillet, L. C. *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* **11,** O111.016717 (2012).
16. Mainland, J. D. *et al.* The missense of smell: functional variability in the human odorant receptor repertoire. *Nat. Neurosci.* **17,** 114–120 (2014).
17. Choong, W. K. *et al.* Informatics view on the challenges of identifying missing proteins from shotgun proteomics. *J. Proteome Res.* **14,** 5396–5407 (2015).
18. Neuhaus, E. M., Mashukova, A., Zhang, W., Barbour, J. & Hatt, H. A specific heat shock protein enhances the expression of mammalian olfactory receptor proteins. *Chem. Senses* **31,** 445–452 (2006).
19. Mashukova, A., Spehr, M., Hatt, H. & Neuhaus, E. M. Beta-arrestin2-mediated internalization of mammalian odorant receptors. *J. Neurosci.* **26,** 9902–9912 (2006).
20. Kang, N. & Koo, J. Olfactory receptors in non-chemosensory tissues. *BMB Rep.* **45,** 612–622 (2012).
21. Flegel, C., Manteniotis, S., Osthold, S., Hatt, H. & Gisselmann, G. Expression profile of ectopic olfactory receptors determined by deep sequencing. *PLoS ONE* **8,** e55368 (2013).
22. Ferrer, I. *et al.* Olfactory receptors in non-chemosensory organs: the nervous system in health and disease. *Front Aging Neurosci.* **8,** 163 (2016).
23. Islam, M. T. *et al.* A systematic bioinformatics approach to identify high quality MS data and functionally annotate proteins and proteomes. *Methods Mol. Biol.* **1549,** 163–176 (2016).
   A simple and intuitive MS evidence workflow for verifying peptides from proteins, along with in silico functional annotation from ProtAnnotator that is integrated into MissingProteinPedia.
24. Ranganathan, S., Khan, J. M., Garg, G. & Baker, M. S. Functional annotation of the human chromosome 7 'missing' proteins: a bioinformatics approach. *J. Proteome Res.* **12,** 2504–2510 (2013).
25. Islam, M. T. *et al.* Protannotator: a semiautomated pipeline for chromosome-wise functional annotation of the 'missing' human proteome. *J. Proteome Res.* **13,** 76–83 (2014).
26. Alexander, S. P. *et al.* The Concise Guide to PHARMACOLOGY 2015/16: Overview. *Br. J. Pharmacol.* **172,** 5729–5743 (2015).
27. Hurley, P. T. *et al.* Functional coupling of a recombinant human 5-HT5A receptor to G-proteins in HEK-293 cells. *Br. J. Pharmacol.* **124,** 1238–1244 (1998).
28. Pasqualetti, M. *et al.* Distribution of the 5-HT5A serotonin receptor mRNA in the human brain. *Brain Res. Mol. Brain Res.* **56,** 1–8 (1998).
29. Rees, S. *et al.* Cloning and characterisation of the human 5-HT5A serotonin receptor. *FEBS Lett.* **355,** 242–246 (1994).
30. Grailhe, R. *et al.* Increased exploratory activity and altered response to LSD in mice lacking the 5-HT(5A) receptor. *Neuron* **22,** 581–591 (1999).
31. Wu, S. *et al.* Group III human metabotropic glutamate receptors 4, 7 and 8: molecular cloning, functional expression, and comparison of pharmacological properties in RGT cells. *Brain Res. Mol. Brain Res.* **53,** 88–97 (1998).
32. Berthele, A. *et al.* Expression of metabotropic glutamate receptor subtype mRNA (mGluR1-8) in human cerebellum. *Neuroreport* **10,** 3861–3867 (1999).
33. Malherbe, P. *et al.* Cloning and functional expression of alternative spliced variants of the human metabotropic glutamate receptor 8. *Brain Res. Mol. Brain Res.* **67,** 201–210 (1999).
34. Stepulak, A. *et al.* Expression of glutamate receptor subunits in human cancers. *Histochem. Cell Biol.* **132,** 435–445 (2009).
35. Tang, F. R. & Lee, W. L. Expression of the group II and III metabotropic glutamate receptors in the hippocampus of patients with mesial temporal lobe epilepsy. *J. Neurocytol.* **30,** 137–143 (2001).
36. Geurts, J. J. *et al.* Expression patterns of Group III metabotropic glutamate receptors mGluR4 and mGluR8 in multiple sclerosis lesions. *J. Neuroimmunol.* **158,** 182–190 (2005).
37. Zhai, J. *et al.* Modulation of lateral perforant path excitatory responses by metabotropic glutamate 8 (mGlu8) receptors. *Neuropharmacology* **43,** 223–230 (2002).
38. Scherer, S. W., Soder, S., Duvoisin, R. M., Huizenga, J. J. & Tsui, L. C. The human metabotropic glutamate receptor 8 (GRM8) gene: a disproportionately large gene located at 7q31.3-q32.1. *Genomics* **44,** 232–236 (1997).
39. O'Dowd, B. F. *et al.* Cloning and chromosomal mapping of four putative novel human G-protein-coupled receptor genes. *Gene* **187,** 75–81 (1997).
40. Lee, J., Hever, A., Willhite, D., Zlotnik, A. & Hevezi, P. Effects of RNA degradation on gene expression analysis of human postmortem tissues. *Faseb J.* **19,** 1356–1358 (2005).
41. Adams, J. W. *et al.* Myocardial expression, signaling, and function of GPR22: a protective role for an orphan G protein-coupled receptor. *Am. J. Physiol. Heart Circ. Physiol.* **295,** H509–H521 (2008).
42. Raine, E. V. *et al.* Gene expression analysis reveals HBP1 as a key target for the osteoarthritis susceptibility locus that maps to chromosome 7q22. *Ann. Rheum. Dis.* **71,** 2020–2027 (2012).
43. Zheng, J. *et al.* Prestin is the motor protein of cochlear outer hair cells. *Nature* **405,** 149–155 (2000).

44. He, D. Z., Lovas, S., Ai, Y., Li, Y. & Beisel, K. W. Prestin at year 14: progress and prospect. *Hear. Res.* **311,** 25–35 (2014).
45. Mistrik, P., Daudet, N., Morandell, K. & Ashmore, J. F. Mammalian prestin is a weak Cl(−)/HCO(3)(−) electrogenic antiporter. *J. Physiol.* **590,** 5597–5610 (2012).
46. Mio, K. *et al.* The motor protein prestin is a bullet-shaped molecule with inner cavities. *J. Biol. Chem.* **283,** 1137–1145 (2008).
47. Anderson, C. T. & Zheng, J. Isolation of outer hair cells from the cochlear sensory epithelium in whole-mount preparation using laser capture microdissection. *J. Neurosci. Methods* **162,** 229–236 (2007).
48. Horvatovich, P. *et al.* Quest for missing proteins: update 2015 on Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **14,** 3415–3431 (2015).
49. Baker, M. Antibody anarchy: a call to order. *Nature* **527,** 545–551 (2015).
50. Bordeaux, J. *et al.* Antibody validation. *Biotechniques* **48,** 197–209 (2010).
51. Barrangou, R. *et al.* Advances in CRISPR-Cas9 genome engineering: lessons learned from RNA interference. *Nucleic Acids Res.* **43,** 3407–3419 (2015).
52. Colwill, K. & Graslund, S. A roadmap to generate renewable protein binders to the human proteome. *Nat. Methods* **8,** 551–558 (2011).
53. Layton, D., Laverty, C. & Nice, E. C. Design and operation of an automated high-throughput monoclonal antibody facility. *Biophys. Rev.* **5,** 47–55 (2012).
54. Larsson, K. *et al.* Multiplexed PrEST immunization for high-throughput affinity proteomics. *J. Immunol. Methods* **315,** 110–120 (2006).
55. Thiele, I. & Palsson, B. Ø. Reconstruction annotation jamborees: a community approach to systems biology. *Mol. Syst. Biol.* **6,** 361–361 (2010).
56. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44,** W242–W245 (2016).
57. Stevens, R. C. *et al.* The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. *Nat. Rev. Drug Discov.* **12,** 25–34 (2013).
58. Niimura, Y. & Gojobori, T. *In silico* chromosome staining: reconstruction of Giemsa bands from the whole human genome sequence. *Proc. Natl Acad. Sci. USA* **99,** 797–802 (2002).

Acknowledgements
We thank Dr M Shaikh for assistance with web engineering the MissingProteinPedia web interface. MTI acknowledges the award of a Macquarie University Research Excellence Scholarship (MQRES). We thank Professor G Omenn, Dr L Lane, Dr E Deutsch, Dr H Cheruku, Ms I Nawar and the HUPO community for helpful discussions during the course of this work. Digital storage and computing was provided by Intersect Australia Space and Time, with acknowledgement to the Australian Government NCRIS Research Data Services (RDS) and National eResearch Collaboration Tools and Resources (Nectar).

Author contributions
M.S.B. conceived MissingProteinPedia. M.S.B., S.R. and E.C.N. planned this study. S.R. and M.S.B. named MissingProteinPedia. S.R. and M.T.I. organized all necessary MissingProteinPedia compute resources. M.T.I. designed, developed and implemented the MissingProteinPedia database, all automated workflows and the community web portal. S.B.A., A.M., M.T.I., M.S.B. and S.R. assembled, interrogated spectra manually, re-analysed and reported olfactory receptor data. E.C.N. coordinated the Australia–New Zealand Chr7 initiative. P.V. contributed analysis of small peptides and new IL-9 MS data. D.C. contributed to the prestin analysis. M.C. provided missing protein pharmacological data review. S.B.A., A.M., M.T.I., D.C., S.S., S.F., S.R. and M.S.B. produced graphics, formatting and referencing. All authors contributed to the writing/reviewing of each version of this manuscript.

Additional information
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Baker, M. S. *et al.* Accelerating the search for the missing proteins in the human proteome. *Nat. Commun.* **8,** 14271 doi: 10.1038/ncomms14271 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

In the bioinformatic analysis presented in the above study (**Study I**), one of the most interesting observations was that the largest PE2-4 classification mapped onto the transmembrane G-protein coupled receptor family. What was more astonishing, was the fact that progress in the discovery of putative proteins belonging to a subfamily of these receptors encompassing receptors for taste and olfaction had seen no progress since 2012 (Baker et al., 2017).

It has been postulated that one of the possible reasons for poor translation of "missing proteins" to PE1 could be in their inadequacy for tryptic digestion. Not all proteins are amenable to tryptic digestion due to their sequence, chemical properties or expression in limited biospecimens (Paik et al., 2012). Thus, they fail to yield tryptic peptides that suitably "fly" in the mass spectrometer or simply lie outside observable mass range detection settings. Based on this, the question stands; "How much of the Human Olfactory Receptor Proteome is Findable using High Stringency Mass Spectrometry". The reasons for the absence of MS-based data on this family of proteins and proposed orthogonal approaches were investigated to attempt to answer this question. This is summarised in the following review (Study II) as part of this thesis. Our hope is that adoption of the proposed strategies in this manuscript alongside innovation across proteomics technical avenues will contribute towards finding missing proteins as one of the goals of HPP.

1.5 Study II: In Silico Peptide Repertoire of Human Olfactory Receptor Proteome on High-Stringency Mass Spectrometry (Publication II)

Subash Adhikari[†], Samridhi Sharma[†], Seong Beom Ahn[†] and Mark S. Baker[†]*

[†] *Department of Biomedical Sciences, Faculty of Medicine & Health Sciences, Macquarie University, NSW, 2109, Australia.*

Abstract

Human olfactory receptors (ORs) are 7-pass transmembrane G-protein coupled receptors (GPCR) involved in smell perception and many other signalling pathways. They are primarily expressed in the olfactory epithelium and ectopically expressed in several other organs/tissues. neXtProt contains four (4) PE1 (protein existence 1 - evidence at protein level) ORs, based either on protein interaction data (i.e., OR1D4, OR2AG1) or convincing genetic, haplotype or biochemical data (i.e., OR1D2, OR2J3). Not a single OR currently qualifies for neXtProt PE1 status based on mass spectrometry (MS) evidence. Many reasons for this absence of MS-based identification have been proposed, including (i) confined or spatiotemporal or developmental expression, (ii) low copy number, (iii) OR repertoire gene silencing, and/or (iv) limited tissue availability. OR transmembrane domains (TMDs) inherently limit MS identification as the hydrophobic nature restricts the access of trypsin to potential cleavage sites. Equally, the extremely low frequency or lack of accessible arginine (R) and lysine (K) residues in TMDs renders trypsin cleavage ineffective. Here, an analytical approach specifically focussed on the hydrophilic (trypsin-accessible) domains of ORs (i.e., with all transmembrane segments (TMDs) and anchored-peptides excluded) is demonstrated. The ability of OR soluble (hydrophilic) domains to yield two or more > 9 amino acids (aa) length uniquely mapping (unique to a protein only), non-nested (peptides with varying length at the N or C terminal but contains the same core sequence), leucine/isoleucine (I/L) switch examined (I and L have same mass and can't be distinguished by MS) tryptic peptides. Our analysis showed that ~58% of the human OR proteome could potentially generate tryptic peptides that satisfy current Human Proteome Project (HPP) data interpretation guidelines (v 2.1), when no missed cleavage is allowed and increases to ~78 % when one missed cleavage is allowed. Utilisation of current biological data (adjuvant genomics, expression profile, transcriptomics, epigenome silencing data etc.) and adoption of non-conventional proteomics approach (e.g., Confetti multi-protease digestion, CNBr cleavage of TMDs, more extreme chromatographic/MS methods) could aid in the detection of the remaining ORs.

Keywords: olfactory receptors, missing proteins, transmembrane proteins, high-stringency mass spectrometry, *in-silico* trypsin digestion, HPP metrics, HPP data interpretation guidelines, uniquely mapping non-nested peptides, membrane hydrophobicity, trypsin activity.

Introduction

Human ORs are a family of 404 (UniProt release: 24th March 2018) seven TMD-containing GPCR signalling proteins.[1] They are involved in human olfaction[2] and several other human biologies.[3] ORs sit on the rhodopsin branch of the unrooted GPCR phylogenetic tree.[4] They are responsible for initiating signalling in response to a range of ligands, including protons, photons, low molecular weight (<30kD) hormones, metals, nutrients, volatiles and neurotransmitters.[5, 6, 7] Elucidating OR function is progressing, coincident with advances in structural and physiological analysis, signalling models and other interactomic methodologies.[8] ORs are implicated in many ectopic physiologies with escalating chemosensory roles, independent of nasal epithelial tissues.[6, 9, 10] ORs have restricted expression in ectopic sites such as brain, breast, colon, liver, lung, testis, thyroid etc, usually with FKPM (fragments per kilobase per million mapped fragments) value of less than 1. For reference, ß-actin gene yields an expression value at a range of 500-5000 FPKM whereas the TATA box binding protein has an expression value of 1.6-21 FPKM.[6]

Each OR contains one free N-terminal strand exposed extracellularly (i.e., ecto-), one C-terminal strand exposed to the cytoplasm (i.e., endo-), 7 TMDs, and 3 ecto- and 3 endo-domain loops between these TMDs, respectively. Each of these domains varies in length, sequence and R/K composition that makes them heterogeneously-susceptible to tryptic digestion.[1] For example, while hydrophilic OR loops and free N- and C- strands contain many R/K residues that make them readily available, ORs TMD domains are notably deficient in R/K residues. In fact, these positively-charged AAs are most likely located at the extremities of the hydrophobic/hydrophilic membrane interfaces or not located in the TMDs at all.[11], [12] In addition, after tryptic digestion hydrophobic TMDs remain embedded within the plasma membrane and can only be removed/solubilised using more extreme sample preparation strategies[13] and utilisation of heated chromatography.[14]

In our previous study, all 122,717 "stranded" OR peptide spectra ($\geq$ 7 aa in length) were concatenated from the publicly-available database.[15] The analysis included studies that previously claimed identification of a significant number of ORs[16, 17] despite what has now

been confirmed as reliance on the marginal spectral quality, a lack of stringent applicable FDRs, inclusion of shorter non-proteotypic observed peptides (≤ 8 aa) and numerous erroneous or ambiguous identifications.[18] Our study (Suppl. Data) concluded that at very best there was patchy/unconvincing MS evidence for ~6% (i.e., 23) of the 404 ORs.[15] The data concludes that no human OR currently met the high-stringency MS criteria set by the HPP data interpretation guidelines.[19]

Indisputably, the community faces ongoing difficulties identifying OR family members by high-stringency MS. Plausible explanations for the paucity of identifications include, that;

- they have nil/low transcription;
- there are few OR-expressing cells;
- they have limited tissue/sample availability;
- they have restricted spatiotemporal/differentiation-dependent expression;
- gene expression is inactivated in olfactory sensory neurons for all but a single OR;
- there is a lack of availability/solubility of trypsin-accessible sites in many ORs; and/or

Here, the ability of ORs to generate peptides from trypsin accessible domains was analysed exclusively. Since it is extremely unlikely that membrane protein TMDs contribute to MS data collations unless specifically enriched, *in silico* digestion of both the full-length ORs was undertaken and concatenations of the exposed hydrophilic OR domains (free N- and C-termini strands plus 3 each of the ecto-domain and endo-domain loops). Approximately ~58% of the human OR proteome could potentially generate non-missed cleaved tryptic peptides qualifying the current HPP PE1 guidelines.[19]

Methods

Grouping of missing proteins based on NeXtProt descriptors: Analysis was performed on chromosomal reports available from neXtProt protein dataset release for the year 2013, 2014-2019 (ftp://ftp.nextprot.org/pub/current_release/). The chromosomal reports (1-22, X, Y and MT) were downloaded and protein descriptions of PE2-4 category proteins were sorted into protein groups based on neXtProt descriptors, e.g. zinc finger proteins were pooled together and counted. Subsequently, the 20 most populous proteins 'descriptors' according to neXtProt were plotted on a bar graph using graph pad prism (version 7). neXtProt chromosomal reports contain proteins that are assigned as putative or probable, e.g., Probable G-protein coupled receptor 63 or Putative olfactory receptor 2B8. Grouping of these proteins into the pool of

GPCRs or ORs was performed based on Pfam for protein families and UniProtKB and GeneCards for putative/probable status. Cluster of uncharacterised proteins were not included in the figure, owing to the lack of their functional annotation. (Figure 2).

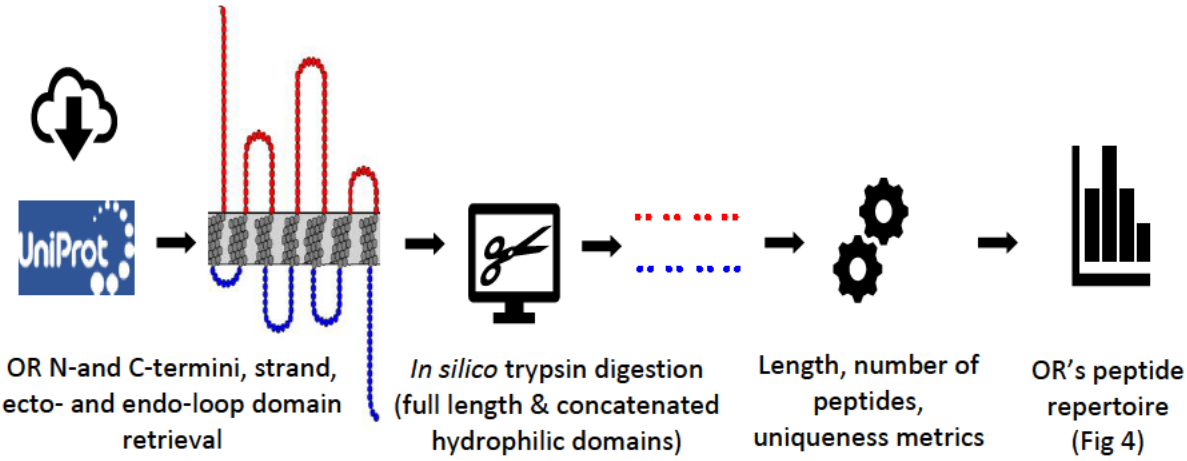


Figure 1: Summary of methods used for identifying ORs at different metrics stringency. *UniProt-derived soluble OR domains were trypsin digested in silico, with no missed cleavages allowed. These peptides were checked for uniqueness using neXtProt's peptide uniqueness checker. Uniquely mapping non-nested peptides from the digest were matched to respective ORs to count their ability to generate peptides at different stringencies.*

Retrieval of current ORs from UniProt: Human Swiss-Prot entries with the term "Olfactory receptor" were retrieved from UniProt (release: 24th March 2019). These entries were filtered to exclude (i) non-olfactory receptors (protein name not starting with "olfactory receptor") and (ii) putative olfactory receptor proteins, resulting in 404 ORs.

Relative R/K residue location along OR TMDs: The relative distances of identified TMD R/K residues from either the cytoplasmic or extracellular hydrophilic interface to the mid-TMD hydrophobic location was normalised. The ratio of R/K residue location from hydrophilic/hydrophobic interface over total TMD length was calculated so that 0-0.2 and 0.8-1.0 reflected positioning in the external hydrophilic region, whilst values of 0.2-0.8 reflected positioning within hydrophobic TMDs (Figure 3).

Identification of ORs capable of generating peptides qualifying PE1 Status: UniProt definitions were used for topological classification of all OR ecto- and endo-domain loop and strand regions (release: 24th March 2019). Whole OR protein sequence and hydrophilic (soluble) domains concatemers were digested *in silico* using the cleaver R package,[20] allowing

for no (zero) missed cleavages. Uniquely mapping non-nested peptides were subsequently selected using the neXtProt's peptide uniqueness checker (neXtProt release v2.21.0).[21] These peptides were then matched to respective ORs. Both uniquely mapping non-nested peptide lists (i.e., derived from full length OR and concatenated hydrophilic domains) were analysed (Figure 1). The percentage of the 404 ORs as full-length ORs or non-TMD-containing OR concatemers capable of producing peptides at different MS stringency levels was calculated and is illustrated in Figure 4. This shows 58% (235) of 404 ORs could generate tryptic peptides qualifying the current PE1 HPP data interpretation guidelines[19] with no missed cleavage allowed. The workflow was repeated allowing one tryptic missed cleavage for the remaining 42% ORs, leading to an inclusion of additional 80 (~78% total) ORs.

OR GPCR topology cartoon representations and sequence were generated using Protter.[22] An interactive HTML file containing a list of observable peptides from concatenated hydrophilic OR domains (supplementary file) was prepared using the DT R package.[23]

Results

Figure 2 contains past and current neXtProt PE datasets from 2013 to 2019. This represents an update of analyses previously undertaken.[15] Positive trends demonstrating increased neXtProt PE1 assignment are observed across most (18/20) of the top 20 protein group based on neXtProt descriptors. This demonstrates that the HPP has successfully identified more human proteins at high stringency than ever before, with few protein family exceptions. For example, noteworthy progress has been made in identifying zinc finger, keratin-associated, leucine-rich repeats and sperm and testes-related proteins. Despite this unquestionable progress, Figure 2 also demonstrates that membrane protein identification at high stringency remains problematic, including across neXtProt protein groups covering the olfactory receptors, other transmembrane non-GPCR transmembrane proteins, non-OR GPCRs, taste receptors and solute carrier proteins.

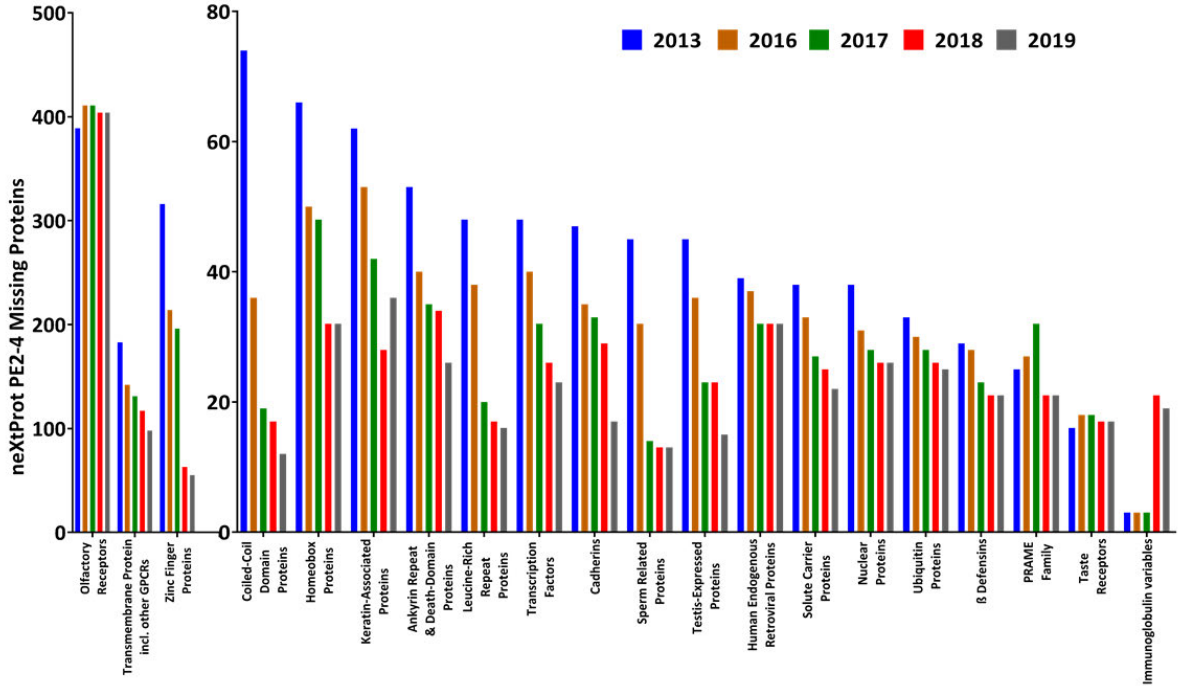


Figure 2: Updated (2013-2019) most abundant 20 neXtProt PE2-4 missing protein descriptors: *neXtProt protein datasets were captured from neXtProt chromosomal download reports as previously described. PE2-4 proteins were sorted by neXtProt protein descriptions under the term "descriptions".*

In summary, the only *neXtProt protein descriptor* group where **absolutely zero** progress has been made since 2013 are the ORs. Figure 2 also shows that the 2019 neXtProt release demonstrates that ORs continue to be the major PE2-4 missing protein family, representing a massive ~19% (400 out 2,129) of all missing PE2-4 proteins. *In silico* digestion of trypsin-available regions of the ORs was performed to calculate ORs capable of generating peptides as per the current HPP PE1 guidelines.[15]

Trypsin digestion is widely adopted in shotgun proteomic approaches because of its stability, high activity, and specificity, often resulting in the generation of "flyable" y-ion high mass series peptides.[24] The availability/accessibility of R/K residues in domains is crucial for determining susceptibility to cleavage. Limited MS-based detection of TMD peptides has been previously established, owing to hydrophobicity[25] and poor MS signal.[26] Unusual LC and MS conditions are required to comprehensively analyse peptides generated from membrane protein

TMDs.[27], [28] An initial structural analysis of the OR proteome to determine where all OR TMD R/K residues resided was undertaken.

Our UniProt-based analysis (Figure 3) indicated that 31% of OR TMDs are completely deficient (126/404; data not shown) in either R or K residues and are unable to produce any tryptic peptides. Figure 3 illustrates that occasionally R/K residues (278/404 ORs with 1 reside and 57/404 ORs with $\geq$ 2 residues) are always positioned (>85%) proximal to the interface between the hydrophobic and hydrophilic environments – in other words at the inside (cytoplasmic) and outside (extracellular) membrane boundaries. Of those R/Ks present in OR TMDs, ~ 50 % of Rs and ~ 85 % of Ks are located at the TMD extremities. Recognising the paucity of TMD R/K residue locations and considering the TMD hydrophobicity, it is extremely unlikely that TMD tryptic peptides could ever significantly contribute to OR HPP data by relying on conventional MS approaches.

The whole sequence *in silico* digestion many do not correlate experimental peptide yield from membrane proteins containing multiple TMDs (e.g., 7 TMDs in all ORs and GPCRs), as peptide cleavage and release is restricted. This is because the soluble ecto- and endo-domain loops between adjacent hydrophobic membrane-embedded TMDs **need to be cleaved at a minimum of two (2) locations before tryptic peptides can be released**.

When **only** a single one tryptic cleavage is present, both nascent ends of ecto- or endo-domain loops remain anchored through their adjacent single membrane-embedded TMDs. This simple observation results in a far lower likelihood of soluble tryptic peptide release from ecto- and endo- domain loops in multi-TMD containing membrane proteins (e.g., ORs/GPCRs), suggesting that membrane proteomics requires careful consideration of the repertoire of proteolytic peptides necessary to contribute to PE1 assignments.

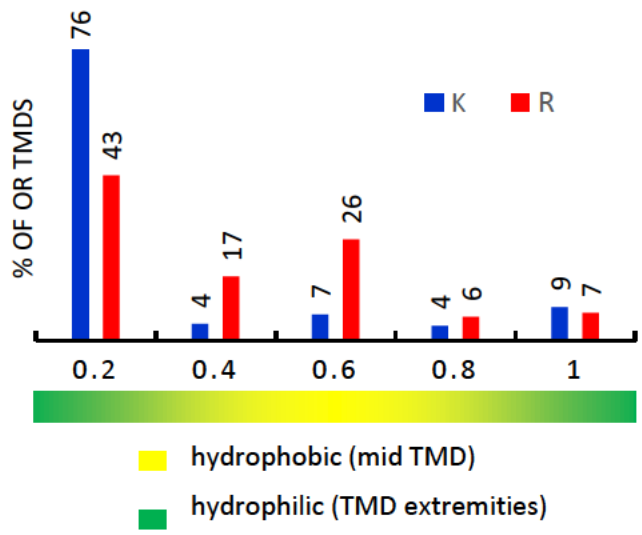


Figure 3: Relative location of R/K residues in TMDs of 278 human OR subset that contains R and/or K residues. *The relative location of OR TMD R and K residues found in the 278 of a total 404 human ORs that contain these residues was analysed. Amongst those 278 proteins, only 57 contain ≥2 R/K residues. The majority of K (~85%) and R (~50%) residues are found proximal to hydrophobic/hydrophilic TMD interfaces. Most R/K residues were located at interfaces with extremely sparse distribution within TMD hydrophobic regions, presumably due to the presence of negatively charged hydrophobic residues disrupts membrane stability*

An *in-silico* digestion of concatenated OR soluble domains alone should provide a superior prediction of the potential peptide complement to those obtained from previous full length OR analyses (i.e., including TMDs) is argued. Analysis of TMD-containing peptides warrants the more generalised adoption of non-conventional, extreme sample preparation, peptide cleavage (e.g., CNBr) and heated column LC methodologies to allow digestion and release of measurable peptides.[13, 14]

To explain the paucity of MS-based OR identifications (0/404) in the 2019 neXtProt HPP, it was important to determine if ORs soluble domains could produce enough tryptic peptides for reliable, high-quality HPP identification by MS. Figure 4 shows the data from *in silico* OR digestion of; (i) full length, (ii) trypsin-accessible soluble domains from the 404 ORs (i.e., N-termini, C- termini, ecto- and endo-domain loops) allowing no missed cleavage and (iii) ORs soluble domains allowing one missed cleavage (Figure 4). This analysis indicates how many ORs are theoretically capable of generating peptides at the number, length and uniqueness stringency required under the HPP PE1 guidelines or at reduced stringency.

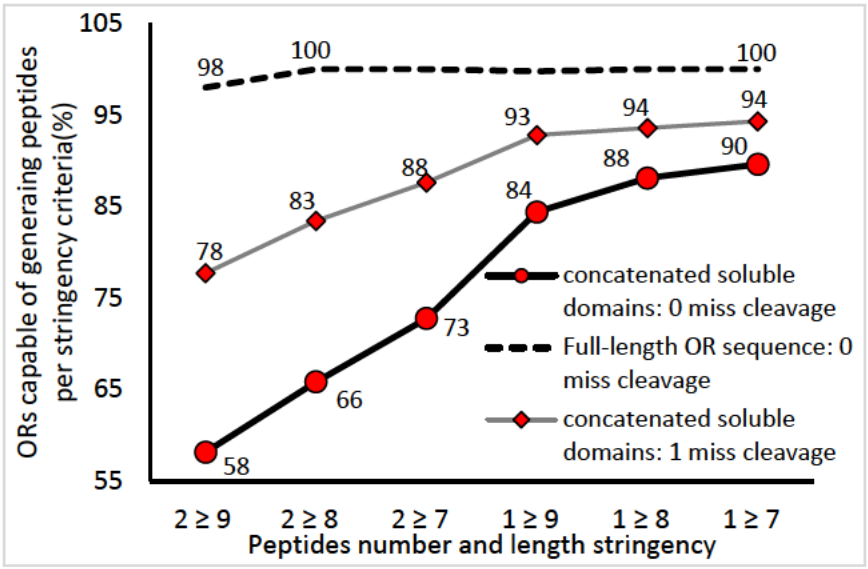


Figure 4: ORs capable of generating peptide as per the current HPP guidelines for missing proteins and lower stringency: *Uniquely mapping non-nested tryptic peptides (only) produced by in silico digestion (non-missed cleaved) of full-length OR sequences (inclusive of TMDs) and concatenated soluble domains restricted to ecto-, endo-loops and N- and C-termini strands (at zero and one tryptic miss cleavages) were utilised to predict the number of ORs that potentially could be identified at different stringencies.*

Figure 4 shows that ~ 58% (235) ORs could generate peptides as per the HPP PE1 guidelines, based on non-missed cleaved tryptic peptides generated from the concatenated endo- and ecto-domains. ~ 78% (315) ORs meet the guidelines upon the inclusion of one tryptic missed cleavage. Relaxation of the metrics resulted in a gradual increase in ORs identification up to ~94% (378). Upon whole sequence ORs *in silico* digestion, ~2% (8) ORs could not generate enough peptides as per PE1 requirement on non-missed cleavage allowed tryptic digestion but were able to generate peptide as per the HPP requirement upon allowing digestion with one missed cleavage.

In summary, although 396 ORs (~98%) can theoretically generate peptides meeting the HPP PE1 guidelines of 2 or more ≥ 9 aa uniquely mapping non-nested peptides upon complete sequence digestion, this calculation includes peptides that have been derived from consideration of TMD-containing peptides which are not easily detected. In contrast, ~ 58% of ORs trypsin-accessible domains (N-termini strands, C- termini strands, ecto- and endo-domain loops) qualify for the HPP PE1 guidelines, the number extends to ~ 78 % upon including missed-cleavages. By default, this presents ~22% of human ORs may not meet PE1 requirements based on their soluble domains. Should relaxation of metrics (either decreasing number or length of required OR peptides or allowing missed cleavages) occur[18] this will significantly increase chances of detecting human ORs, albeit at lower stringency.

Low occurrence of R/K residues within TMDs and restriction of trypsin proteolytic activity against these residues if any, within hydrophobic TMD leads to the generation of most MS identifiable peptides from ecto- or endo- domains using a conventional MS approach. Figure 5 presents two representative topological distributions of ORs along plasma membrane that can (OR4K5_HUMAN, UniProt accession: Q8NGD3; Figure 5a) or cannot (OR5MA_HUMAN, UniProt accession: Q6IEU7; Figure 5b) meet the current the HPP PE1 criteria, based on tryptic peptide generation when no missed cleavage is allowed. OR4K5 generates 4 uniquely mapping non-nested peptides (green), 3 from its endo- domain (YVAICKPLYYVVIMSR, IVNHYLRPR and ISEMSLVVR) and 1 from ecto- domain (SNSSVVSEFVLLGLCSSQK). OR4K5 can generate peptides as per the HPP PE1 guidelines, with an ability to yield two or more uniquely mapping non-nested peptides that are nine or more aa in length. By contrast, OR5MA generates 3 non-uniquely mapping non-nested tryptic peptides (yellow), 1 from ecto-domain (MLSPNHTIVTEFILLGLTDDPVLEK) and 2 from endo- domain (YVAICSPLHYSSR, DVILAIQQMIR). Additionally, OR5MA contains 3 "hanging" peptides (light blue) (NVTPNMLHNFLSEQK, LLTFHLSFCGSLEINHFYCADPPLIMLACSDTR, YLFIFAAIFR). These nascent peptides remain anchored to the plasma membrane as their respective domains lack two (or more) cleavage sites for peptide release leading to the conclusion that OR5MA could never reach PE1 status, solely based on tryptic peptides.

Eight (8) ORs that **could not** generate peptides as per PE1 assignment criteria were identified, upon whole sequence digestion. This observation directed us to probe if any evidence of experimental uniquely mapping non-nested OR peptide exists. PeptideAtlas (http://www.peptideatlas.org/)[29] and neXtProt (https://www.nextprot.org/)[21] were thoroughly analysed in search of publicly available experimental OR peptide evidences. Current PeptideAtlas build does not contain any evidence for experimental OR peptide spectra. (personal communication, Eric Deutsch) and the HPP does not permit the use of synthetic peptides or data from bait-proteins for immune-precipitations as evidence for PE1 assignment.

The most current 2019 neXtProt release has assigned PE1 status to four ORs based upon non-MS evidence i.e., either protein-protein interaction evidence (OR1D4_HUMAN, UniProt accession: P47884 and OR2AG1_HUMAN, UniProt accession: Q9H205) or other relevant genomics and biochemical evidence (OR1D2_HUMAN, UniProt accession: P34982[30] and OR2J3_HUMAN, UniProt accession: O76001. Among the remaining 400 ORs, **neXtProt** classifies 227 ORs as PE2 and 173 ORs as PE3.[31]

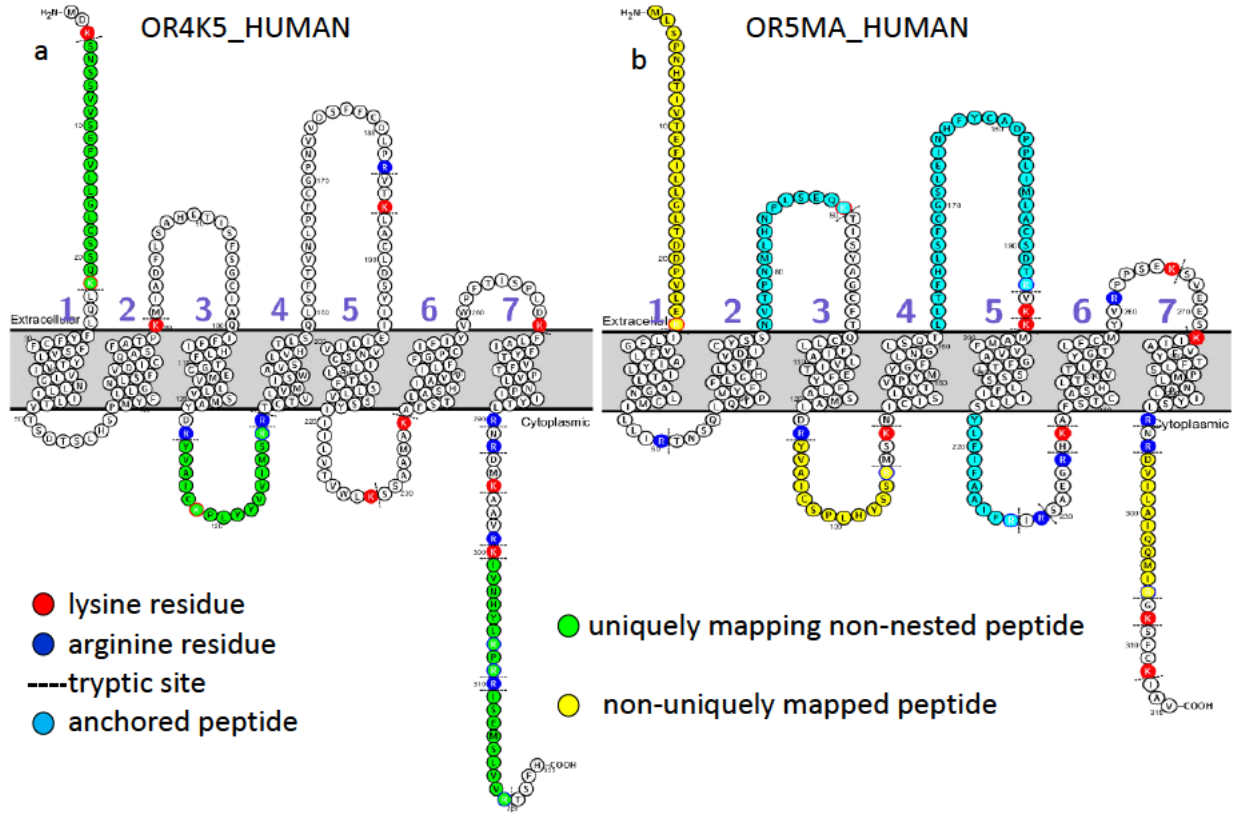


Figure 5: Topological distribution of exemplar OR (OR4K5 and OR5MA) N-termini strands, C-termini strands, intra-TMD loops, TMDs, R/Ks, residue number and trypsin cleavage sites. *ORs contain seven TMDs with 3 each of ecto- and endo-domains (loops) and 1 N-terminal and 1 C-terminal strands. The figure illustrates one example of an OR that can (OR4K5_HUMAN, UniProt accession: Q8NGD3;* ***Figure 5a****) or cannot (OR5MA_HUMAN, UniProt accession: Q6IEU7;* ***Figure 5b****) generate peptides as per the HPP PE1 criteria. Q8NGD3 can generate 4 uniquely mapping non-nested (green) peptides that are ≥ 9 aa in length. On the other hand, Q6IEU7 generates 3 non-uniquely mapped peptides (yellow) and 3 strands that remain anchored to the plasma membrane (light blue).*

DISCUSSION

There has been no progress (i.e., zero PE1 identifications) in MS-based ORs identification since the 2011 inception of neXtProt.[32] Multiple (~22%) ORs fail to generate tryptic peptides from their soluble domains at the stringency levels set by the HPP PE1 guideline requirements. Apart from inherent technical challenges attributed by limited abundance,[33] ORs topology plays a significant role in the paucity of OR identification.[28] Restriction of trypsin proteolytic

activity on membrane-embedded R/K residues, if any and requirement of minimum two (2) cleavage sites within any ecto- and endo-domain loops, generally results in low tryptic peptide yield. Peptides generated by a potential PE1 candidate OR requires LC-MS compatibility and "flyability" to be identified on any MS platform,[34] further limiting identification. An interactive HTML file (supplementary file, requires file type conversion to HTML) containing tryptic peptides generated from ORs endo- and ecto- domain, with associated annotations is provided. This file could be used to query uniquely mapping non-nested peptides obtainable from any of the 404 OR concatenated hydrophilic domains.

Should alternate proteolytic enzyme systems be used, additional missed cleavage allowed, TMD considered or change in UniProt topology definition observed, our current prediction number will change on a case by case basis. Analysis of MS-identifiable ORs taking these aspects into account was not the aim of this analysis. Given the lack of OR identification over recent years, assignment of any ORs as PE1 proteins based solely on the current HPP guidelines seems farfetched. These ORs require adjuvant genomics, expression profile, transcriptomics or epigenome silencing data complementing MS evidence for PE1 assignment. This corollary holds true for other missing proteins containing multiple TMDs such as taste receptors or other GPCRs.

Supplementary information

The supplementary file contains a list of tryptic peptides obtainable from concatenated OR ecto- and endo- domains by *in silico* digestion allowing no missed cleavage (S1) or one-missed cleavage (S2). ORs capable of generating peptides at different length and number stringency can be obtained from column "Matches to PMID" with corresponding peptide length stringency on immediate right columns.

Author information

corresponding author
Mark S. Baker
email: mark.baker@mq.edu.au

ORCHID IDs

Subash Adhikari: 0000-0001-5945-7804
Samridhi Sharma: 0000-0002-1167-6511
Seong Beom Ahn: 0000-0001-5907-3544

Mark Baker: 0000-0001-5858-4035

Author contributions

MSB conceived the idea, SA performed *in silico* analysis, SS contributed on neXtProt missing proteins update. SA, SBA and MSB prepared the manuscript. All authors read and approved the final manuscript.

Funding sources

SA and SS acknowledge iMQRES funding from Macquarie University, SBA on Cancer Institute NSW grant (15/ECF/1-38) and MSB on NHMRC Project Grant APP1010303.

Acknowledgments

Our deep thanks to Eric Deutsch and Lydie Lane for confirmation of current OR peptides information from PeptideAtlas and neXtProt respectively.

Supplementary data

Supplementary data contains interactive HTML files containing a list of theoretical tryptic peptides from concatenated endo- and ecto- OR domains (in text format). The file could be accessed by common browsers upon changing the file extension to HTML.

The authors declare no competing financial interest.

References

1.Malnic, B.; Godfrey, P. A.; Buck, L. B., The human olfactory receptor gene family. *Proc. Natl. Acad. Sci. U S A* **2004,** 101, (8), 2584-9.

2.Buck, L. B., Olfactory receptors and odor coding in mammals. *Nutrs. Rev.* **2004,** 62, (11 Pt 2), S184-8; discussion S224-41.

3.Feldmesser, E.; Olender, T.; Khen, M.; Yanai, I.; Ophir, R.; Lancet, D., Widespread ectopic expression of olfactory receptor genes. *BMC Genomics* **2006,** 7, 121.

4.Attwood, T. K.; Findlay, J. B., Fingerprinting G-protein-coupled receptors. *Protein Eng.* **1994,** 7, (2), 195-203.

5.Aisenberg, W. H.; Huang, J.; Zhu, W.; Rajkumar, P.; Cruz, R.; Santhanam, L.; Natarajan, N.; Yong, H. M.; De Santiago, B.; Oh, J. J.; Yoon, A. R.; Panettieri, R. A.; Homann, O.; Sullivan, J. K.; Liggett, S. B.; Pluznick, J. L.; An, S. S., Defining an olfactory receptor function in airway smooth muscle cells. *Sci. Rep.* **2016,** 6, 38231.

6.Flegel, C.; Manteniotis, S.; Osthold, S.; Hatt, H.; Gisselmann, G., Expression profile of ectopic olfactory receptors determined by deep sequencing. *PLoS One* **2013,** 8, (2), e55368.
7.Horowitz, L. F.; Saraiva, L. R.; Kuang, D.; Yoon, K. H.; Buck, L. B., Olfactory receptor patterning in a higher primate. *J. Neurosci.* **2014,** 34, (37), 12241-52.
8.Mainland, J. D.; Keller, A.; Li, Y. R.; Zhou, T.; Trimmer, C.; Snyder, L. L.; Moberly, A. H.; Adipietro, K. A.; Liu, W. L.; Zhuang, H.; Zhan, S.; Lee, S. S.; Lin, A.; Matsunami, H., The missense of smell: functional variability in the human odorant receptor repertoire. *Nat. Neurosci.* **2014,** 17, (1), 114-20.
9.Kang, N.; Koo, J., Olfactory receptors in non-chemosensory tissues. In *BMB Rep.*, 2012; Vol. 45, pp 612-22.
10.Ferrer, I.; Garcia-Esparcia, P.; Carmona, M.; Carro, E.; Aronica, E.; Kovacs, G. G.; Grison, A.; Gustincich, S., Olfactory Receptors in Non-Chemosensory Organs: The Nervous System in Health and Disease. *Front. Aging Neurosci.* **2016,** 8, 163.
11.Ulmschneider, M. B.; Sansom, M. S., Amino acid distributions in integral membrane protein structures. *Biochim. Biophys. Acta* **2001,** 1512, (1), 1-14.
12.Hildebrand, P. W.; Preissner, R.; Frommel, C., Structural features of transmembrane helices. *FEBS Lett.* **2004,** 559, (1-3), 145-51.
13.Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001,** 19, (3), 242-7.
14.Blackler, A. R.; Speers, A. E.; Wu, C. C., Chromatographic benefits of elevated temperature for the proteomic analysis of membrane proteins. *Proteomics* **2008,** 8, (19), 3956-64.
15.Baker, M. S.; Ahn, S. B.; Mohamedali, A.; Islam, M. T.; Cantor, D.; Verhaert, P. D.; Fanayan, S.; Sharma, S.; Nice, E. C.; Connor, M.; Ranganathan, S., Accelerating the search for the missing proteins in the human proteome. *Nat. Commun.* **2017,** 8, 14271.
16.Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.;

Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A., A draft map of the human proteome. *Nature* **2014,** 509, (7502), 575-81.

17. Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A. M.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B., Mass-spectrometry-based draft of the human proteome. *Nature* **2014,** 509, (7502), 582-7.

18.Ezkurdia, I.; Vazquez, J.; Valencia, A.; Tress, M., Analyzing the first drafts of the human proteome. *J. Proteome Res.* **2014,** 13, (8), 3854-5.

19.Deutsch, E. W.; Overall, C. M.; Van Eyk, J. E.; Baker, M. S.; Paik, Y. K.; Weintraub, S. T.; Lane, L.; Martens, L.; Vandenbrouck, Y.; Kusebauch, U.; Hancock, W. S.; Hermjakob, H.; Aebersold, R.; Moritz, R. L.; Omenn, G. S., Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* **2016,** 15, (11), 3961-3970.

20.Gibb, S. *Cleaver: Cleavage of Polypeptide Sequences*, **2015**. R package version 1.12.0, https://github.com/sgibb/cleaver/.

21.Gaudet, P.; Michel, P. A.; Zahn-Zabal, M.; Britan, A.; Cusin, I.; Domagalski, M.; Duek, P. D.; Gateau, A.; Gleizes, A.; Hinard, V.; Rech de Laval, V.; Lin, J.; Nikitin, F.; Schaeffer, M.; Teixeira, D.; Lane, L.; Bairoch, A., The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* **2017,** 45, (Database issue), D177-82.

22.Omasits, U.; Ahrens, C. H.; Muller, S.; Wollscheid, B., Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics* **2014,** 30, (6), 884-6.

23.Xie, Y. DT: A Wrapper of the JavaScript Library 'Data Tables', **2018**. R package version 0.4, https://CRAN.R-project.org/package=DT

24.Ong, S. E.; Foster, L. J.; Mann, M., Mass spectrometric-based approaches in quantitative proteomics. *Methods* **2003,** 29, (2), 124-30.

25.Eichacker, L. A.; Granvogl, B.; Mirus, O.; Muller, B. C.; Miess, C.; Schleiff, E., Hiding behind hydrophobicity. Transmembrane segments in mass spectrometry. *J. Biol. Chem.* **2004,** 279, (49), 50915-22.

26.Bagag, A.; Jault, J. M.; Sidahmed-Adrar, N.; Refregiers, M.; Giuliani, A.; Le Naour, F., Characterization of hydrophobic peptides in the presence of detergent by photoionization mass spectrometry. *PLoS One* **2013,** 8, (11), e79033.

27.Kar, U. K.; Simonian, M.; Whitelegge, J. P., Integral membrane proteins: bottom-up, top-down and structural proteomics. *Expert Rev. Proteomics* **2017,** 14, (8), 715-723.

28.Vit, O.; Petrak, J., Integral membrane proteins in proteomics. How to break open the black box? *J. Proteomics* **2017,** 153, 8-20.
29.Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R., The PeptideAtlas project. *Nucleic Acids Res.* **2006,** 34, (Database issue), D655-8.
30.Neuhaus, E. M.; Mashukova, A.; Barbour, J.; Wolters, D.; Hatt, H., Novel function of beta-arrestin2 in the nucleus of mature spermatozoa. *J. Cell Sci.* **2006,** 119, (Pt 15), 3047-56.
31.McRae, J. F.; Mainland, J. D.; Jaeger, S. R.; Adipietro, K. A.; Matsunami, H.; Newcomb, R. D., Genetic variation in the odorant receptor OR2J3 is associated with the ability to detect the "grassy" smelling odor, cis-3-hexen-1-ol. *Chem. Senses* **2012,** 37, (7), 585-93.
32.Omenn, G. S.; Lane, L.; Overall, C. M.; Corrales, F. J.; Schwenk, J. M.; Paik, Y. K.; Van Eyk, J. E.; Liu, S. Q.; Snyder, M.; Baker, M. S.; Deutsch, E. W., Progress on Identifying and Characterizing the Human Proteome: 2018 Metrics from the HUPO Human Proteome Project. *Journal of Proteome Research* **2018,** 17, (12), 4031-4041.
33.Fonslow, B. R.; Carvalho, P. C.; Academia, K.; Freeby, S.; Xu, T.; Nakorchevsky, A.; Paulus, A.; Yates, J. R., Improvements in Proteomic Metrics of Low Abundance Proteins through Proteome Equalization Using ProteoMiner Prior to MudPIT. *Journal of Proteome Research* **2011,** 10, (8), 3690-3700.
34.Sanders, W. S.; Bridges, S. M.; McCarthy, F. M.; Nanduri, B.; Burgess, S. C., Prediction of peptides observable by mass spectrometry applied at the experimental set level. *BMC Bioinformatics* **2007,** 8 Suppl 7, S23.

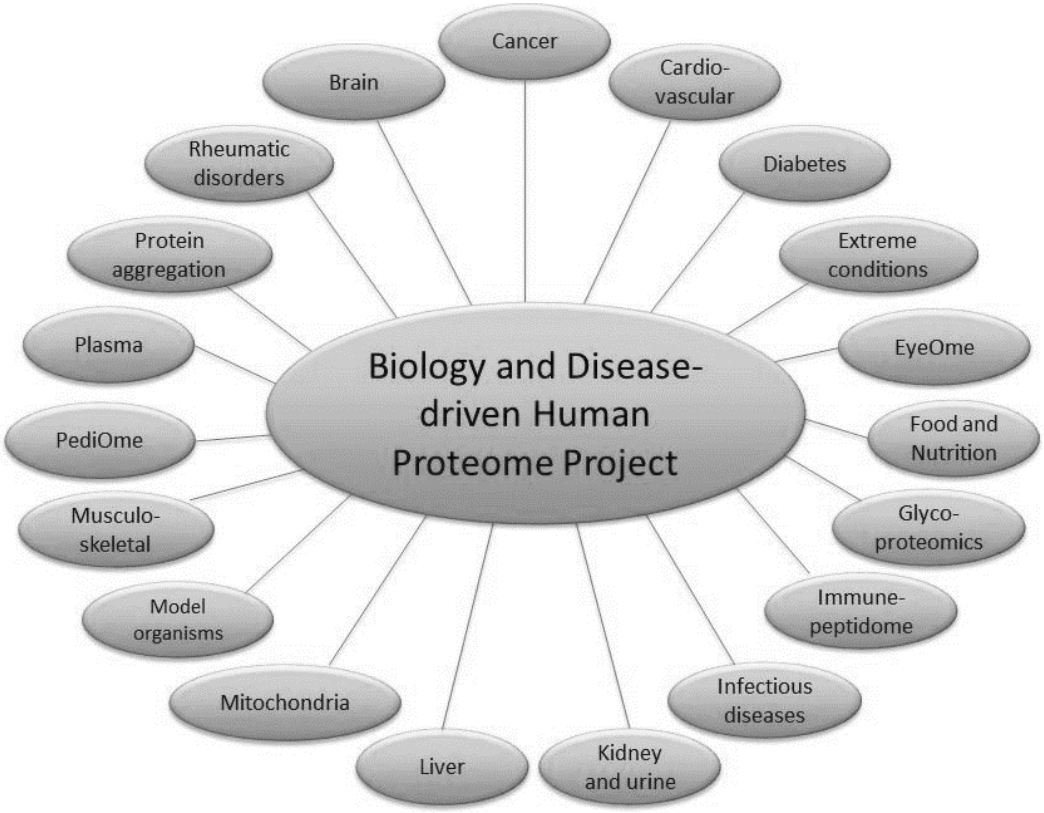


Figure 1.2: Constituents of the Biology and Disease oriented initiative of the Human Proteome Project (B/D-HPP). *Adapted and reproduced from https://hupo.org/human-proteome-project*

1.6. B/D-HPP

The B/D-HPP was established in order to obtain a functional insight into the proteome and to develop tools, technologies, and informatics pipelines addressing biomedical and translational needs. It focussed on the identifying the trends in the proteome by investigating a specific biological problem or disease rather than segmenting the proteome as in the C-HPP. The road map to achieve this involved the following steps. Firstly, a focus area central to a biological problem or disease was selected. This enabled researchers across the globe with specific interests in a bioprocess or disease to self-organise to form a typically multidisciplinary group to effectively contribute to the project. A target list involving hypothetical proteins or proteins known to get altered could then be generated. Assays to robustly detect and quantify proteins involved in the disease process could be established to probe for potential targets for further characterisation. SOPs and data emerging from these assays could be shared using the knowledge base (Aebersold et al., 2013). Currently, the B/D HPP (Figure 1.2) includes a

repertoire of initiatives including the eye, brain, cancer, musculo-skeletal, paediatric, cardiovascular, diabetes, epigenetics and chromatin, glyco-proteomics, infectious disease, kidney and urine, liver, mitochondria, model organisms, plasma, and stem cells proteome. B/D-HPP groups are also studying proteomic signatures following exposure to extreme conditions including nutrition, the immunopeptidome, protein aggregation, and toxicoproteomics, reproductive health, membrane proteins and neurodegenerative (and protein misfolding) disorders (Van Eyk et al., 2016). The B/D HPP interlinks with the C-HPP to help populate the individual chromosomes and understand specific diseases at the molecular level (Figure 1.2).

The B/D-HPP, like the C-HPP, strongly relies on the three HPP pillars (Figure 1.1). Technological advances in all three pillars have enhanced the pace at which the B/D-HPP is accomplishing its various goals. One such branch of B/D-HPP has enabled researchers to obtain a new dimension of understanding into cancer biology in the post-genomic era, the cancer-HPP.

Cancer arises from genomic aberrations that result in dysregulation in signalling or protein activity/function (Hanahan and Weinberg, 2011). Whilst over the past decade or so genomics has fuelled many ambitious projects aimed at identifying cancer-related mutations, this approach has not been without challenges. For example, only ~10% of tumours result from actionable mutations with a majority of tumours harbouring alterations where functional outcomes are difficult to interpret (Jimenez et al., 2018). It is now realised that proteomics may be central to facilitating the HPP visions of P4 (i.e., personalised, precision, preventative and participatory) medicine (Nice, 2016). Tumours are highly heterogenous and as a result clinical response differs from case to case. Another unmet clinical challenge is that resistance to therapy frequently arises during prolonged treatment (Jimenez et al., 2018). To overcome the aforementioned unmet clinical needs of tumour heterogeneity and resistance to therapies, comprehensive proteomic analysis of cancer is needed to improve the understanding of the disease and improve clinical outcomes, fitting well into the P4 framework and complementing genomics (Nice, 2016). This led to the inception of the Human Cancer Proteome Project, part of the B/D-HPP.

Like most other teams within the B/D-HPP, the cancer HPP aims to benefit from multidisciplinary international collaborations. Cancer covers a vast spectrum of malignancies affecting almost every organ of the body. Moreover, cancer of a particular organ can often comprise several types of tumours. For example, there are over 120 different types of brain and

CNS tumours alone, each with its own pattern of progression and characteristics (NBTS, 2018). To obtain useful insights into such a complex disease requires a multidisciplinary approach. Global projects studying cancer specific studies have resulting in resources like The Cancer Genome Atlas (TCGA) (Vogelstein et al., 2013), Clinical Proteomic Tumour Analysis Consortium (CPTAC) (Ellis et al., 2013) and the Genomics Evidence Neoplasia Information Exchange (GENIE) (2017) which provides access to datasets with either genomic landscapes, proteome profiles or NGS data in a cohort of patients with specific cancers. In order to build a platform that integrates genomic and proteomic datasets from a large cohort of patients with high sample quality, an initiative called the Applied Proteogenomics Organisation (APOLLO) launched with the Cancer Moonshot project under the auspices of the NCI (Fiore et al., 2017). The Human Cancer Proteome Project fits well within this scheme to delineate and grade proteomes from normal to malignant, identify tumour specific signatures to dissect tumour heterogeneity and to develop assays and reagents with translational utility to be distributed into the community. The Cancer-HPP strongly encourages researchers to conform to the high stringency metrics and standards set by the HPP, especially in the context of MS data collection and analysis and in promoting the post-publication submission of cancer proteomics datasets for global meta- and pan-cancer initiatives (Jimenez et al., 2018).

In this thesis, the state-of-art proteomics was used to find translational solutions for one of the biggest clinical problems plaguing our world, namely colorectal cancer (CRC) (Bray et al., 2018) discussed in detail in **Chapter 2** "Introduction to CRC".

References

(2017). Project GENIE Goes Public. Cancer Discov *7*, 118.

Aebersold, R., Agar, J.N., Amster, I.J., Baker, M.S., Bertozzi, C.R., Boja, E.S., Costello, C.E., Cravatt, B.F., Fenselau, C., Garcia, B.A., *et al.* (2018). How many human proteoforms are there? Nature chemical biology *14*, 206-214.

Aebersold, R., Bader, G.D., Edwards, A.M., van Eyk, J.E., Kussmann, M., Qin, J., and Omenn, G.S. (2013). The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. J Proteome Res *12*, 23-27.

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

Baker, M.S., Ahn, S.B., Mohamedali, A., Islam, M.T., Cantor, D., Verhaert, P.D., Fanayan, S., Sharma, S., Nice, E.C., Connor, M., *et al.* (2017). Accelerating the search for the missing proteins in the human proteome. Nat Commun *8*, 14271.

Bilello, J.A. (2005). The agony and ecstasy of "OMIC" technologies in drug development. Current molecular medicine *5*, 39-52.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians *68*, 394-424.

Cox, J., and Mann, M. (2007). Is proteomics the new genomics? Cell *130*, 395-398.

Ellis, M.J., Gillette, M., Carr, S.A., Paulovich, A.G., Smith, R.D., Rodland, K.K., Townsend, R.R., Kinsinger, C., Mesri, M., Rodriguez, H., *et al.* (2013). Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumour Analysis Consortium. Cancer Discov *3*, 1108-1112.

Fiore, L.D., Rodriguez, H., and Shriver, C.D. (2017). Collaboration to Accelerate Proteogenomics Cancer Care: The Department of Veterans Affairs, Department of Defense, and the National Cancer Institute's Applied Proteogenomics OrganizationaL Learning and Outcomes (APOLLO) Network. Clin Pharmacol Ther *101*, 619-621.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674.

Jimenez, C.R., Zhang, H., Kinsinger, C.R., and Nice, E.C. (2018). The cancer proteomic landscape and the HUPO Cancer Proteome Project. Clin Proteomics *15*, 4.

Kusebauch, U., Deutsch, E.W., Campbell, D.S., Sun, Z., Farrah, T., and Moritz, R.L. (2014). Using PeptideAtlas, SRMAtlas, and PASSEL: Comprehensive Resources for Discovery and Targeted Proteomics. Curr Protoc Bioinformatics *46*, 13.25.11-28.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., Bergeron, J., Borchers, C.H., Corthals, G.L., Costello, C.E., *et al.* (2011). The human proteome project: current state and future direction. Mol Cell Proteomics *10*, M111.009993.

Li, G.W., and Xie, X.S. (2011). Central dogma at the single-molecule level in living cells. Nature *475*, 308-315.

NBTS (2018). Tumour Types: Understanding Brain Tumours (National Brain Tumour Society).

neXtProt (2018). neXtProt. In Explopring the universe of human proteins.

Nice, E.C. (2016). From proteomics to personalized medicine: the road ahead. Expert review of proteomics *13*, 341-343.

Omenn, G.S. (2012). The HUPO Human Proteome Project (HPP), a Global Health Research Collaboration. Cent Asian J Glob Health *1*, 37.

Omenn, G.S. (2017). Advances of the HUPO Human Proteome Project with broad applications for life sciences research. Expert review of proteomics *14*, 109-111.

Omenn, G.S., Lane, L., Overall, C.M., Corrales, F.J., Schwenk, J.M., Paik, Y.K., Van Eyk, J.E., Liu, S., Snyder, M., Baker, M.S., *et al.* (2018). Progress on Identifying and Characterizing the Human Proteome: 2018 Metrics from the HUPO Human Proteome Project. J Proteome Res.

Paik, Y.-K., Overall, C.M., Corrales, F., Deutsch, E.W., Lane, L., and Omenn, G.S. (2018). Toward Completion of the Human Proteome Parts List: Progress Uncovering Proteins That Are Missing or Have Unknown Function and Developing Analytical Methods. Journal of Proteome Research *17*, 4023-4030.

Paik, Y.K., Jeong, S.K., Omenn, G.S., Uhlen, M., Hanash, S., Cho, S.Y., Lee, H.J., Na, K., Choi, E.Y., Yan, F., *et al.* (2012). The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat Biotechnol *30*, 221-223.

Peng, L., Cantor, D.I., Huang, C., Wang, K., Baker, M.S., and Nice, E.C. (2018). Tissue and plasma proteomics for early stage cancer detection. Molecular omics *14*, 405-423.

Schubert, O.T., Gillet, L.C., Collins, B.C., Navarro, P., Rosenberger, G., Wolski, W.E., Lam, H., Amodei, D., Mallick, P., MacLean, B., *et al.* (2015). Building high-quality assay libraries for targeted analysis of SWATH™-MS data. Nat Protoc *10*, 426-441.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat Biotechnol *26*, 1135-1145.

Thul, P.J., and Lindskog, C. (2018). The human protein atlas: A spatial map of the human proteome. Protein Sci *27*, 233-244.

Thygesen, C., Boll, I., Finsen, B., Modzel, M., and Larsen, M.R. (2018). Characterizing disease-associated changes in post-translational modifications by mass spectrometry. Expert review of proteomics *15*, 245-258.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., *et al.* (2010). Towards a knowledge-based Human Protein Atlas. Nat Biotechnol *28*, 1248-1250.

Van Eyk, J.E., Corrales, F.J., Aebersold, R., Cerciello, F., Deutsch, E.W., Roncada, P., Sanchez, J.C., Yamamoto, T., Yang, P., Zhang, H., *et al.* (2016). Highlights of the Biology and Disease-driven Human Proteome Project, 2015-2016. J Proteome Res *15*, 3979-3987.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science (New York, NY) *339*, 1546-1558.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet *10*, 57-63.

Chapter 2

Introduction to CRC

Even though cancer can simplistically be described as a genetic disease of somatic cells, it underlays a deceptively complex interplay of cellular processes leading to its manifestation. Despite relentless progress in biomedical science, the multi-layered complexity governing tumourigenesis poses significant resistance to a majority of currently available cancer treatment modalities. This is particularly true after a late stage diagnosis, where many complications (including metastasis) occur, eventually leading to breakdown of most physiological processes. This occurs to such an extent that medical practitioners are compelled to switch from curative to palliative cancer treatment.

In their seminal paper in 2000, Hanahan and Weinberg predicted that *"those researching the cancer problem will be practicing a dramatically different type of science than we have experienced over the past 25 years"* (Hanahan and Weinberg, 2000). Our understanding of tumour heterogeneity, molecular landscapes, microbiome, tumour microenvironment, epigenetic factors and immune checkpoints are indeed shaping the way cancer is now seen and treated (Flavahan et al., 2017; Gopalakrishnan et al., 2018; Meacham and Morrison, 2013; Pickup et al., 2013; Reya et al., 2001; Vesely et al., 2011; Vogelstein et al., 2013). Research accompanied by conceptual and technical advancements will continue to shape the future of cancer therapy and treatment (Hanahan and Weinberg, 2000). However, the path to a "cure" remains a long and arduous one and begs the question - what actionable changes can be implemented relatively immediately that will reduce the mortality burden from cancer?

One central tenet to combating diseases like cancer is philosophically *"nipping it in the bud"*. This is true for almost all cancers where curative procedures or treatments are in place. In most types of cancer, survival of patients improves significantly when diagnosis is made earlier rather than later in disease progression (WHO, 2018). One classical example is cancer of the colon or rectum – or CRC. It results from the accumulation of genetic mutations that lead to neoplasia (de la Chapelle, 2004). Its histopathological progression can be demarcated into clinical stages where the tumour progresses, allowing researchers to understand cancer biology in greater depth (de la Chapelle, 2004). Perhaps the most interesting observation regarding CRC is how stage of diagnosis directly impacts patient prognosis (Kolligs, 2016). It seems

apparent that the major driving factor of CRC-related mortality is the lack of an effective early clinical stage diagnostic test, that could convert the third largest number of global cancer deaths into a combatable ailment.

This chapter provides an introductory overview about CRC incidence, susceptibility, basic biology and a clinical perspective on CRC pathogenesis. It specifically focuses on current early diagnosis tests used in clinical practise, as well as tests that are currently in development. Shortcomings are highlighted to plan likely solutions and form the rationale for this thesis.

2.1 Colorectal Cancer Incidence and Mortality

CRC is the fourth most common malignancy in the world today, with an age standardised rate (ASR) of 19.7 per 100,000 individuals. In 2018, 1,849,518 new CRC cases and 880,792 deaths were recorded globally, making it the third most common cause of death from malignancy (Figure 2.1). Of these, 21,217 mapped to Australasian incidence cases, and an estimated 7,424 deaths making it the region with the largest percentage of recorded CRC cases globally (CRC fact sheet Globocon) (Ferlay et al., 2018). The estimated number of cases in Australasia are predicted to increase by ~48% and number of deaths predicted to increase by 83% by 2040, making CRC a serious public health problem (IARC Global Cancer Observatory web site (http://gco.iarc.fr/) (Bray et al., 2018; Ferlay J, 2018).

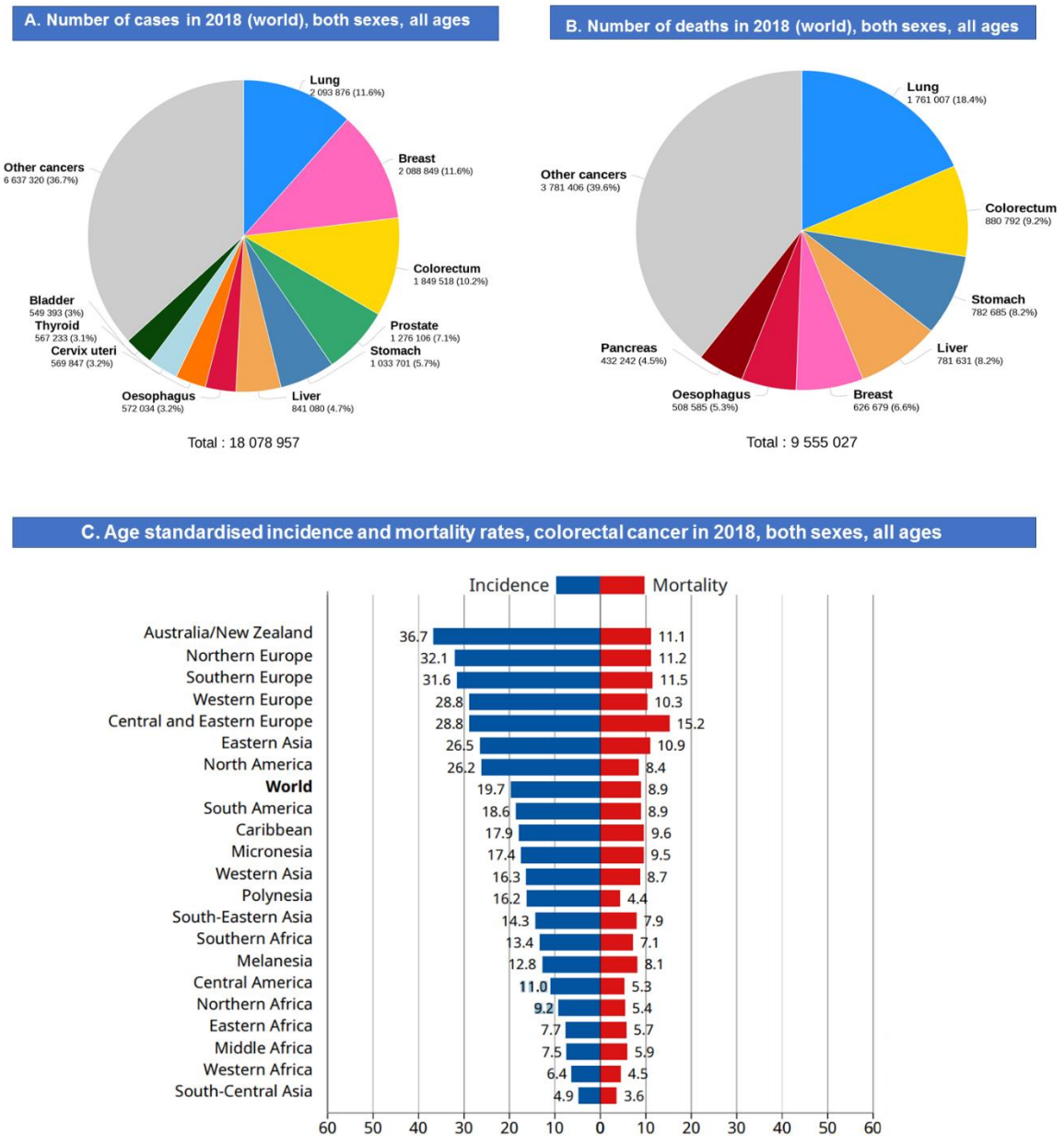


Figure 2.1: *Estimated age-standardised A) CRC new cases, B) deaths (world 2018) worldwide for both sexes and all ages, and C) countries with highest incidence and mortality. Image from GLOBOCON 2018, Global Cancer Observatory (*http://gco.iarc.fr/*)(Ferlay J, 2018).*

2.2 Factors Governing CRC Susceptibility

Temporal CRC trends have been studied in the context to various factors that contribute to aetiology. These can broadly be segregated as epidemiological, genetic, dietary/lifestyle and microbiome risk factors.

2.2.1 Epidemiological Factors

2.2.1.1 Gender and Age

CRC affects men and women almost equally. However, incidence and mortality are particularly high in individuals over 55 years of age. In 2018, CRC incidence and mortality ratios affecting individuals over 55 years compared to those under 55 years was ~7:1 and ~11:1, respectively (Arnold et al., 2017; Bray et al., 2018). However, temporal trends suggest an increase of 22% in CRC incidence in young adults in the period 2000-2012 (Bray et al., 2018).

2.2.1.2 Geographic Prevalence

Epidemiological studies show clear patterns of CRC incidence. The burden of CRC seems to be borne particularly by those 1st world countries with high scores on the human development index (HDI), accounting for 60% of global incidence. Europe, Oceania, Canada, Korea and Japan contribute largely to this incidence (Arnold et al., 2017; Bray et al., 2018) (Figure 2.2). Rates of mortality in these countries have declined from 2000 and this has been attributed to the increased uptake of (bowel cancer) screening programs and improved treatment modalities. However, countries undergoing societal and economic transformation from low to middle/high incomes are demonstrating dramatic increases in mortality. This pattern accentuates findings from studies that show CRC risk aligned to incorporation of a "Western" lifestyle combined with poor early screening programs in those countries compared to high HDI counterparts.

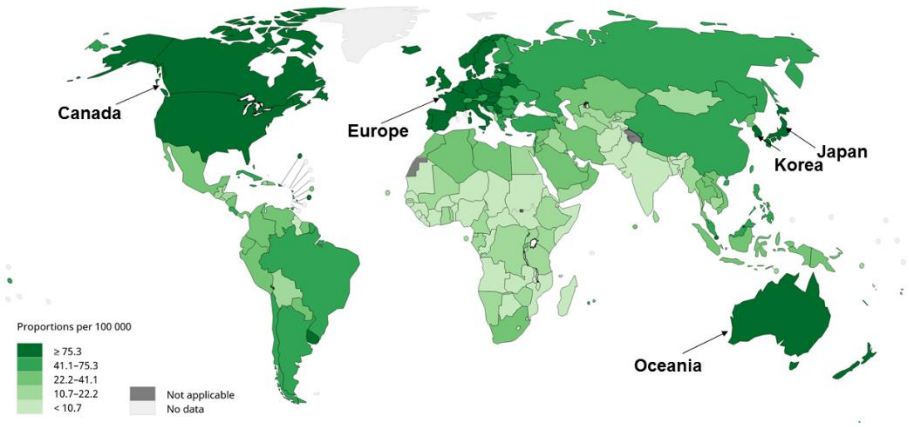


Figure 2.2: Estimated CRC prevalence (5-year) as a proportion in 2018 for both sexes and all ages. *Image from GLOBOCON 2018, Global Cancer Observatory (*http://gco.iarc.fr/*)(Ferlay J, 2018).*

2.2.2 Genetic Factors

CRC is primarily driven by genetic mutations. These mutations can be segregated based on pattern of occurrence as familial, inherited or sporadic. Approximately 75% of CRC results from sporadic gene mutations. The remaining ~25% of cases occurs in patients with a family history of disease (Amersi et al., 2005). About 2-5% of these arise from germline mutations that are inherited. However the aetiology of the remaining 20% of familial CRCs remain elusive (Jasperson et al., 2010).

2.2.2.1 Inherited CRC

A patient who has a first degree relative diagnosed with CRC or colonic polyps before 60 years of age or two or more relatives diagnosed with CRC or colon polyps at any age is considered to have significant history of CRC (Amersi et al., 2005). The magnitude of this risk is considered to be elevated if affected family members are diagnosed before 55 years (Amersi et al., 2005). The risk is considered substantially greater if affected family members are diagnosed before 45 years of age (Amersi et al., 2005). This category is further subclassified into CRCs derived from polyposis and non-polyposis backgrounds.

2.2.2.1.1 Polyposis variant

About 1-2% of diagnosed CRC cases are attributed to *familial adenomatous polyposis* (FAP). This autosomal dominant form of CRC is one the best characterised hereditary polyposis syndromes, where a germline mutation in the adenomatous polyposis coli (APC) gene drives the syndrome. APC is a tumour suppressor gene located on the long arm of chromosome 5q21. It is considered to be a gene that primarily regulates colonic cell neoplastic transition via the β-catenin cell signalling pathway (Amersi et al., 2005). APC promotes apoptosis in colon epithelial cells by regulating the degradation of β-catenin. A mutation in the APC gene results in loss of function that leads to rapid β-catenin-induced proliferation of cells (Stoffel and Kastrinos, 2014). This results in formation of numerous adenomatous polyps throughout the gastrointestinal tract, especially colon. To date, over 1,000 different mutations in APC gene leading to formation of a truncated APC protein have been reported in association with FAP (Rowan et al., 2000). Although APC is the main driver mutation occurring in FAP, ~7-8% of FAP cases result from a mutation in MYH gene (Amersi et al., 2005). MYH is located on chromosome 1p and is a known base excision repair gene. A less aggressive mutation in APC can lead to what is termed as Attenuated FAP (AFAP), that is characterised by 10-100 adenomatous polyps found predominantly in the proximal colon and sometimes rectum.

Besides FAP and AFAP, MUTYH-associated polyposis (MAP), Peutz-Jeghers syndrome (PJS), serrated polyposis syndrome (SPS) are well described conditions associated with the polyposis variant of CRC (Bogaert and Prenen, 2014).

2.2.2.1.2 Non-polyposis variant

Hereditary non-polyposis CRC (HNPCC) or Lynch syndrome constitutes the non-polyposis variant of CRC (Aarnio et al., 1995). This is an autosomal dominant condition and accounts for about 5-10% of CRC cases. It is characterised by a mutation in any of the DNA mismatch repair genes (MMR), including hMSH2, hMLH1, hPMS1, hPMS2, hMSH3 and hMSH6. Defects in MMR genes lead to accumulation of spontaneous point mutations or insertion/deletions of short tandem repeats termed microsatellites. This results in genomic instability called microsatellite instability (MSI). Although mutations can occur in any of the MMR genes, mutations in hMSH2 and hMLH1 are most likely in HNPCC cases (Katballe et al., 2002).

2.2.2.2 Spontaneous

Most CRC tumours develop following a succession of mutations that allow the morphological transition of an adenoma to carcinoma. The primary driver mutation in these cases is a mutation in APC. This is followed by sequential mutations in the KRAS, TP53 and DCC genes. Besides these mutations, three major genomic alterations have been identified that trigger CRC. Chromosomal instability (CIN) leading to chromosomal changes and translocations cause aneuploid tumours or loss of heterozygosity (LOH) that likely affects the APC, KRAS, PI3K and TP53 gene (Armaghany et al., 2012). These aberrations invariably affect central signalling and cell proliferation pathways like WNT, MAPK/PI3K and TGF-b. CpG island methylation CRC phenotype (CIMP) is associated with epigenetic instability. These are characterised by hypermethylation of oncogenes which lead to silencing of critical genes that trigger CRC pathogenic pathways. The third genomic alteration is due to microsatellite instability caused by a hypermutable phenotype due impairment of DNA repair mechanisms. Loss of DNA repair mechanisms impair the cell's ability to repair short DNA chains or inserts of tandem repeats leading to accumulation of deleterious mutations (Amersi et al., 2005).

2.2.3 Risk Factors

2.2.3.1 Diet

The so-called "Western" lifestyle largely refers to a diet which influences CRC in many ways (Arnold et al., 2017). Numerous studies assessing the effect of dietary factors like dietary fibre (vegetables, fruits), type of meat, dietary fat and micronutrient (Calcium, vitamins) intake have been undertaken across the globe (Potter, 2009). The relationship of dietary fibre and total dietary fat have inconclusive outcomes in the context of CRC incidence, however meat-eating (particularly red &/or processed meat) has been substantially associated with elevated CRC risk (Potter, 2009; Sugimura and Sato, 1983). Heterocyclic amines, PAHs, heme, nitrosylation and O6 carboxymethyl guanine have been proposed as possible carcinogens found in cooked meat (Bingham et al., 2002; Cross et al., 2003). Calcium has been shown to reduce the proliferation of adenomas (Bostick et al., 1995; Hyman et al., 1998) and in conjugation with Vitamin D has been shown to associate with a reduced risk for CRC (Potter, 2009).

2.2.3.2 Lifestyle

Apart from diet, populations across high- and middle-income countries tend have a sedentary lifestyle (Marmol et al., 2017). This is possibly the reason why CRC risk is elevated in individuals in *"white collar"* professions. Obesity has been shown to be associated with an elevated risk of CRC in men (Potter, 2009), whilst the data on women is variable. Alcohol consumption has been consistently shown to associate with elevated CRC risk. Alcohol acts as a potent risk factor by inhibiting DNA repair pathways (Marmol et al., 2017; WRCF, 1987). Smoking has also been shown to be a CRC risk factor and is associated with microsatellite unstable (MSI-H) types of CRC (Slatter et al., 2001). Heterocyclic amines in tobacco are likely to trigger pathogenesis in CRC as they have been shown to cause specific mutations in APC in rat models (Kakiuchi et al., 1995).

2.2.3.3 Microbiome and Infections

Among factors that influence CRC risk, a healthy gut microbiome is associated with low CRC risk. A case study by Want *et al*., 2012 has shown individuals with CRC have a predominance of gut microbes from the genera *Bacteroides fragilis, Enterococcus, Escherichia/Shigella, Klebsiella, Streptococcus* and *Peptostreptococcus* spp. when compared to healthy individuals (Wang et al., 2012). It has also been established that a healthy gut microbiome reduces the incidence of an animal's strains genetically vulnerable to CRC (TCRβ/p53, IL-10, Gpx1/Gpx2,

Gαi2, Smad3, Muc2, Tgfβ1/Rag2, IL-2/β2m knock-out strains) when compared to germ-free animals from the same strains (Antonic et al., 2013).

On the flipside, CRC is notoriously associated with aggravated risk in individual post exposure to certain pathogenic infection. There are a handful of case studies reporting likely associations with Schistosomiasis, a communicable trematode infection prevalent in the tropics (Potter, 2009). *In vitro* studies provide evidence that *Streptococcus bovis* or *Streptococcus galloliticus* infections result in formation of pre-cancerous lesions, making affected individuals more vulnerable to CRC. *Helicobacter pylori* has also been identified as a class I carcinogen by the International Agency for Research in Cancer owing to its association to gastric cancer. There are several studies correlating *H. pylori* risk to CRC (Fireman et al., 2000; Meucci et al., 1997; Potter, 2009) . Infections from bacteria belonging to *Fusobacterium spp*. have been strongly correlated to CRC development. Several viruses like human polyoma virus (John Cunningham virus (JC virus)), BK virus, human cytomegalovirus (CMV), human papilloma viruses (HPV; predominantly type 16 and 18) may be potent risk factors (Antonic et al., 2013). These virus highjack important regulatory pathways like Notch, JNK, cyclin-CDK pathway by dysregulating tumour suppressor genes like p53 and Rb (Antonic et al., 2013) . They could induce chromatin remodelling, modulate cellular transcription and/or elicit inflammatory responses potentiating CRC development.

2.2.3.4 Other medical conditions

Conditions like cholecystectomy affecting bile flow in the intestinal tract have been shown to have an underlying 30% higher CRC risk (Todoroki et al., 1999). An increase of 30-40% was also observed in individuals diagnosed with diabetes mellitus (Larsson et al., 2005). Among other medical conditions, inflammatory bowel disease is known to have a pronounced CRC risk associated with it. This risk is reported to be ~7-14% in individuals with IBD affected for ≥25 years. However, risk is more pronounced (~30%) in individuals suffering from IBD for >35 years (Amersi et al., 2005). Researchers have been unable to find genetic links between ulcerative colitis or Crohn's disease (comprising IBD) and CRC. However, it has been proposed that a link could be due to anatomical manifestation of IBD involving he loss of brush border lining of the intestine (Potter, 2009). The recent decline in IBD-associated CRC cases have been attributed to increased use of anti-inflammatory medications prescribed to control IBD (Potter, 2009).

2.2.3.5 Medication

Non-steroidal anti-inflammatory drugs (NSAIDs) have been consistently shown to have a protective role in CRC development (Potter, 2009). There are numerous case control studies associated with reduced CRC incidence/mortality in individuals administered aspirin (Bigler et al., 2001; Kune et al., 1988; Rosenberg et al., 1991; Ruder et al., 2011; Suh et al., 1993), sulindac (Moorghen et al., 1988), piroxicam (Reddy et al., 1987), celecoxib (Kawamori et al., 1998) and indomethacin (Pollard and Luckert, 1980). All have been shown to inhibit carcinogenesis in rodent CRC models. Effects are attributed to potent inhibition of inflammatory cyclooxygenase (COX) enzymes that are highly up-regulated in CRC (Potter, 2009).

2.3 CRC Pathophysiology

2.3.1 Anatomical Standpoint

The human colon (large intestine) measures approximately 1.5m and comprises five anatomic sections: caecum, followed by ascending column, transverse colon, descending colon, sigmoid column and finally rectum. The caecum, transverse column and sigmoid colon are enveloped by peritoneum whereas ascending colon, descending colon, sigmoid colon, and rectum are retroperitoneal. The histological and cellular composition of the large intestine is identical throughout these five sections. The colonic wall is made up of five layers, the innermost layer starting with the mucosa, muscularis mucosa, submucosa, muscularis externa, and serosa that is the outermost layer (Berrocal et al., 1999).

Understanding the arrangement of multiple layers of colonic wall and their respective characteristic properties sheds light on CRC development and progression. Colonic polyps initiate from the inner-most epithelial mucosal surface. Aggressive polyps invade the next layer, muscularis mucosa, which is made of a thin layer of muscle fibres consisting of lymphocytes and solitary lymphatic nodules. Further progression of cancer penetrates submucosa and functionally this layer is responsible for venous and lymphatic drainage. Aggressive cancers spread into the lymphatic system thereby causing cancer to metastasise to distant organs (Jung, 2013). A representative figure shows the anatomy of colon and rectum (Figure 2.3 A), with a second figure showing the colonic wall, divided into multiple layers and progression of polyps within that mucosal layer (Figure 2.3 B).

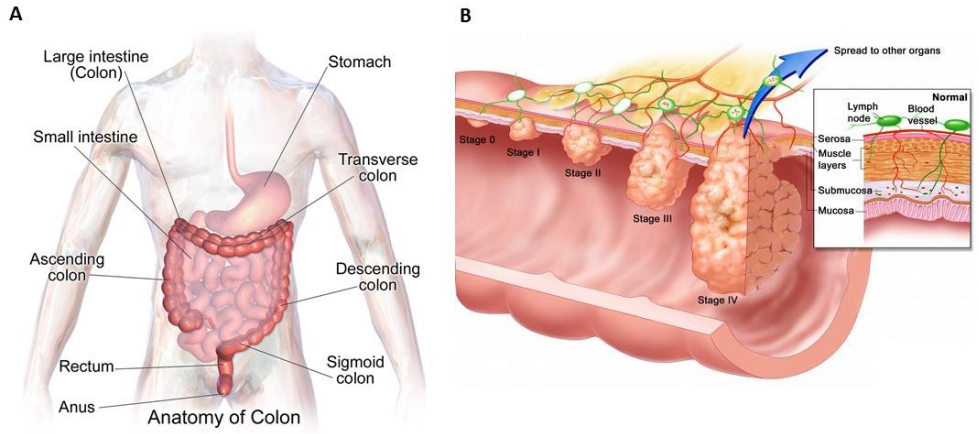


Figure 2.3: Anatomy and histological layers of CRC: *A) Anatomy of human large intestine and rectum (Blausen.comstaff, 2014) and B) Five layers of colonic wall involved in the progression of colon cancer. Image from National Cancer Institute, National Institute of Health (*https://www.nih.gov/*)*

2.3.2 CRC Staging and Survival

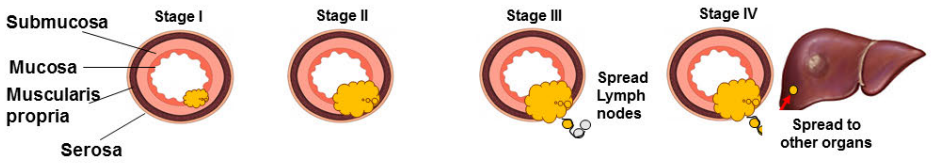
Stage is a strong predictor of patient survival and tumour characteristics such as size and location of tumour and its progression in a patient's body. An accurate cancer staging is imperative to evaluate patient treatment options. This, therefore, warrants a globally recognised cancer staging system for exchange and comparison between hospitals and as basis for translational research. Broadly, a staging system is generally divided into clinical and pathologically staging. Clinical staging is based on imaging (e.g., x-rays, CT-scans and more), physical exams, tumour biopsies and blood tests as part of diagnosis/monitoring (i.e., before surgery or post therapy or recurrence). A staging system is particularly used in deciding best treatment options by clinicians. On the other hand, a pathological staging system is performed after surgery and is derived from clinical staging information complemented/revised by pathological evaluation of the resected tumour specimens and other post-operative findings (Donna M. Gress, 2017).

The most elaborated and useful staging system developed so far is the American Joint Committee on Cancer (AJCC) staging system developed in collaboration with the Union of International Cancer Control (UICC). It is broadly referred to as AJCC staging system and refers to Tumour Node Metastasis (TNM) staging for both pathological and clinical staging. This system elaborates and classifies three properties 1) the size and contiguous spread of the

primary tumour (T), 2) the presence and absence of cancer in regional lymph node (N), and 3) the absence and presence of distant metastasis in sites/organs outside local tumour area (M). The combined values of TNM determines overall cancer stage. CRC staging primarily determines the size and the depth at which tumour has penetrated the bowel wall (T), whether it has spread to lymph node vessels (N) and/or further metastasised to other organs (M) (Charanjeet Singh, 2017; Stephen B. Edge 2010).

Before TNM staging was introduced, CRC was classified under the Dukes' staging system. It classified CRC into four different stages, A, B, C, and D. In Dukes' A, a tumour is confined within the mucosal and sub-mucosal wall. In Dukes' stage B, the tumour penetrates tumour penetrates through the muscularis propria but does not infiltrate further. In Dukes' C, tumour cells spread to local lymph node vessels, whilst Dukes' D marks metastasis to distant organs. Figure 2.4 summarises the AJCC TNM and Dukes' staging systems.

All biospecimens used in studies referred to in this thesis were originally classified under Dukes' staging system. They were later re-classified using the AJCC-TNM staging system (Dukes, 1932). It should be noted that since CRC is a heterogenous disease identification of new mutations in individuals and varied morphological observations by pathologists warrants periodic updating of staging. This will help in more accurate diagnosis, treatment, prognosis and development of new targeted therapies.



Dukes' staging system	AJCC-TNM stages	T (extent of primary tumour invasion)	N (presence and extent of lymph node involvement)	M (presence of metastasis)	5- year survival
--	0	Tis	N0	M0	
A	I	T1	N0	M0	92%
		T2	N0	M0	
B	IIa	T3	N0	M0	87%
	IIb	T4a	N0	M0	65%
	IIC	T4b	N0	M0	65%
C	IIIA	T1-T2	N1/N1c	M0	90%
		T1	N2a	M0	
	IIIB	T3-T4a	N1/N1c	M0	72%
		T2-T3	N2a	M0	
		T1-T2	N2b	M0	
	IIIC	T4a	N2a	M0	53%
		T3-T4a	N2b	M0	
		T4b	N1-N2	M0	
D	IVA	Any T	Any N	M1a	12%
	IV B	Any T	Any N	M1b	
	IV C	Any T	Any T	M1c	

Key:

NX: regional lymph nodes cannot be assessed
N0: no regional lymph node metastasis
N1: metastasis in 1 - 3 regional lymph nodes
N1a: metastasis in 1 regional lymph node
N1b: metastasis in 2 - 3 regional lymph nodes
N1c: no regional lymph nodes are positive but there are tumour deposits in the subserosa, mesentery or non peritonealized pericolic or perirectal / meso-rectal tissues
N2: metastasis in 4 or more regional lymph nodes
N2a: metastasis in 4 - 6 regional lymph nodes
N2b: metastasis in 7 or more regional lymph nodes

NX: regional lymph nodes cannot be assessed
N0: no regional lymph node metastasis
N1: metastasis in 1 - 3 regional lymph nodes
N1a: metastasis in 1 regional lymph node
N1b: metastasis in 2 - 3 regional lymph nodes
N1c: no regional lymph nodes are positive but there are tumour deposits in the subserosa, mesentery or non-peritonealized pericolic or perirectal / meso-rectal tissues
N2: metastasis in 4 or more regional lymph nodes
N2a: metastasis in 4 - 6 regional lymph nodes
N2b: metastasis in 7 or more regional lymph nodes

M0: no distant metastasis by imaging; no evidence of tumour in other sites or organs
M1: distant metastasis
M1a: metastasis confined to 1 organ or site without peritoneal metastasis
M1b: metastasis to 2 or more sites or organs is identified without peritoneal metastasis
M1c: metastasis to the peritoneal surface is identified alone or with other site or organ metastases

Figure 2.4: Staging and survival rate of CRC: *The CRC staging system is a predictor of patient survival and prognostic assessment of disease. The currently most utilised staging system is the AJCC Tumour-Node-Metastasis (TNM) staging system, where; Stage I: penetration of tumour into sub mucosa, and mucosa; stage II: tumour invasion through muscularis propria; stage III: tumour invasion to lymph nodes via visceral peritoneum; and stage IV tumour invasion to adjacent distant organs. The TNM and Dukes' staging with respective survival data from the American cancer society and AJCC 8th edition pathology outliners: colon tumour staging. Figure concept adapted from Cantor, 2016 (Cantor et al., 2015).*

2.3.3 Signs and Symptoms:

The most common, although diffuse and non-specific symptoms of CRC are traces of blood on/in stools expelled from the anus, changes in bowel habits, abrupt weight loss, anaemia and the loss of appetite. However, not all CRCs display these symptoms, nor do all these symptoms necessarily signify the onset of CRC. Notably, early CRC stages (I/II) are notoriously asymptomatic, presenting a major diagnostic challenge. Later CRC stages (III/IV), however, display severe symptoms leading to prompt diagnosis but poor patient survival.

On the other hand, some symptoms of CRC overlap with common ailments experienced by a healthy population. For example, rectal bleeding could occur due to rectal fissures, haemorrhoids and both are common causes of false FOBT positives. The ambiguity and lack of specific symptoms complicates and delays accurate CRC diagnosis. Another important factor to be considered is delay caused in seeking medical help by patients due to non-recognition of the seriousness of symptoms. Some patients choose to adopt a "wait and see" strategy. Misdiagnosis due to symptoms attributed to alternate benign conditions which happen has been estimated to occur in 31-34% of cases (Emery, 2015). Therefore, greater public awareness and a standard fast track diagnosis pathway and further research on the basis of epidemiological evidence could add predictive value contributing to earlier diagnosis (Fijten et al., 1995).

2.4 CRC Population Screening Tests

CRC is a disease associated with significantly low morbidity if diagnosed in its earlier clinical stages (AJCC I/II). However, disease caught at a later stage (AJCC III/IV) present with particularly poor prognosis.

To date, CRC detection tests have incorporated traditional technologies. However, with the evolution of technology, CRC detection modalities have moderately improved. Traditionally, CRC screening modalities can be categorised into either;

- Stool-based, and/or
- Structural (visual) examination-based tests (Wolf et al., 2018).

2.4.1 Stool-based tests

One of the first stool-based tests shown to be effective in the detection of CRC were the guaiac-based faecal occult blood (gFOBT) tests. These tests detect trace amounts of blood haemoglobin in a stool specimen by virtue of the peroxide activity between heme and guaiac (Wolf et al., 2018). However, both low and high sensitivity gFOBT tests lead to false-positive results, as blood can be detected in stool because of other reasons like gastro-intestinal bleeding caused due to non-steroidal anti-inflammatory drugs or red meat, and dietary peroxidases and therefore are neither highly specific. False negative results also occur after consumption of antioxidants like Vitamin C (Wolf et al., 2018).

On the other hand, the faecal immunochemical test (FIT) affords a suitable alternative (Wolf et al., 2018) to gFOBT tests. FIT (earlier known as iFOBT), like gFOBT, screens for occult blood caused as a result of bleeding from polyps or CRC. The antibody used in FIT is highly specific for the globulin component of human haemoglobin. FIT is basically unreactive to upper GIT bleeding caused by other agents as the globulin component of any released hemoglobulin is totally degraded by digestive enzymes. Therefore, FIT preferentially detects bleeding from the lower GI tract caused due to adenoma, polyps, inflammatory diseases and CRC (Wolf et al., 2018) and other non-neoplastic lower GI tract lesions.

gFOBT has shown a sensitivity of 12.9%-79.4% with a specificity of 86.7%-97.7% across a range of different population studies (Song and Li, 2016). FIT, on the other hand, has better performance with sensitivity at 79% and specificity at 94% (Lee et al., 2014). Few high sensitivity FOBT kits termed HSgFOBT have shown to have a sensitivity ranging from 62-79% and specificity ranging from 87-96% placing this test at par, as well as being a cost-effective alternative to FIT (Wolf et al., 2018). HSgFOBT and FIT have superseded gFOBT as the preferable non-invasive stool-based haemoglobin tests to screen CRC in populations (Wolf et al., 2018).

However, performance characteristics of gFOBT and FIT tests have been found to be inconsistent. This primarily stems from the US Food and Drug Administration (US FDA) clearance process which clears them for detection of occult blood not CRC screening *per se* (Rex et al., 2017). Thus, information on sensitivity and specificity of tests detecting CRC are not a prerequisite for FDA clearance. It is because of this reason that clinicians continue to seek a high sensitivity and specific test for CRC or adenoma population screening.

A different perspective on why the gFOBT and FIT test fail to capture a true spectrum of CRC incidence can be attributed to how samples are collected. Most manufacturers of these tests recommend that stool samples be collected at home. However, physicians often collect small amount of stool samples during digital rectal examinations which is sufficient volume to run gFOBT or FIT tests in the clinic. This practice fails to detect CRC 90% of the time (Wolf et al., 2018).

Besides the above two stool-based tests, a third DNA-inclusive multi-target stool (mt-sDNA) test has become available for screening CRC. This test combines FIT and assays for aberrantly methylated BMP3, NDRG, NDRG4 genes and mutations in KRAS and b-actin genes from cells exfoliated from colonic neoplasms (Wolf et al., 2018). In a large comparative trial comparing mt-sDNA to FIT with colonoscopy as the gold reference standard for CRC detection, mt-sDNA was found to be 92.3% sensitive for detecting CRC as compared to 73.8% for FIT (Wolf et al., 2018). Both tests were found to be equally specific at 86.6%. The sensitivity of mt-sDNA was found to be particularly high for individuals with advanced adenomas and sessile serrated polyps at 42.4% when compared to FIT which detected these tumours with only a 23.8% sensitivity. Specificity was, however, significantly lower (89.9%) when compared to FIT (96.4%) when confirmed with colonoscopy (Rex et al., 2017; Wolf et al., 2018). The interpretation of mt-sDNA data is elusive considering that commercial kits do not specify if a positive test emanates from the FIT component or which DNA components of the kit. Clinical follow-up may differ depending upon the source of the positive result on a case-by-case basis (Wolf et al., 2018). However, mt-sDNA continues to be a diagnostic test recommended to be taken every 3 years which yields 88% of Life Years Gained (LYG) from colonoscopy every 10 years (Lansdorp-Vogelaar et al., 2011; Peterse et al., 2018). The FDA has recently approved the costly mt-sDNA test for CRC screening in average risk individuals above the age of 50.

Amidst these established stool-based tests currently in clinical practice, DNA and DNA methylation markers from stool have also been proposed as a promising early diagnostic non-invasive marker for CRC. Specific DNA alterations from the neoplasm are released into the lumen continuously via mechanism of tumour exfoliation (Ahlquist and Gilbert, 1996) rather solely on bleeding which is often intermittent. Moreover, DNA is stable in stool and PCR amplification methods enables detection of analyte present in limited amounts (Ahlquist et al., 2000). Many studies have reported the identification of DNA methylated markers from the

stool of the patients diagnosed with CRC for example, methylated fibrillin-1 (mFBN1), vimentin (VIM), methylated progesterone receptor (mPGR) and O6-methylguanine DNA methyl transferase (mMGMT). (Nguyen and Weinberg, 2016). However, DNA testing for CRC screening has been reported with limited sensitivity owing to poor quality of DNA detected and tumour inter/intra heterogeneity and specific genetic alterations selected for detection (Richter, 2008).

2.4.2 Structural (visual) examination-based tests

2.4.2.1 Colonoscopy

Colonoscopy is a medical procedure to visualise the bowel and detect cancer, colon polyps or other abnormalities. It is the most frequently used physical test to diagnose CRC in the USA. The principle advantage of colonoscopy is the direct visualisation of the colon accompanied by detection, biopsy and removal of any underlying polyps in a single session. Its sensitivity for detecting adenomas ≥6mm in diameter ranges from 75% to 93% with specificity of 94%, while for adenomas ≥1cm in diameter sensitivity ranged from 89% to 98% with specificity of 89%. The US Preventive Services Task Force (USPSTF) estimated that colonoscopy-based screening performed once every ten years in individuals between the ages of 50-75 years would reduce CRC incidence by 62% and mortality by 79-90%. It is for these reasons that colonoscopy is considered the gold standard diagnostic modality for CRC detection (Rex et al., 2017; Wolf et al., 2018).

Despite, the merits offered by colonoscopy, it has several disadvantages. Although colonoscopy is a high sensitivity test, it does tend to miss colonic polyps around the sharp anatomical hepatic and sigmoid flexure turns. Even though colon cancers are rarely missed due to sheer size difference when compared to adenomatous polyps, flat adenomas which tend to be small and discoid resembling erythematous plaques with significant risk of CRC or high-grade dysplasia are occasionally missed during colonoscopy (Rex et al., 2017).

Logistics involved in performing colonoscopy are far from a *"point and shoot"* approach. It requires a rigorous patient bowel cleansing regime, the efficiency of which is paramount before conducting colonoscopy. It has also been found that the length of time between last dose of bowel cleansing agent (e.g., GoLYTELY, MoviPrep, SUPREP, OsmoPrep or Prepopik) and initiation of colonoscopy correlate directly with quality of evidence obtained from tests.

Colonoscopy is considered relatively painless but needs to be conducted under sedation and requires at least a day of recovery. Perhaps the biggest drawback of colonoscopy is due to bowel perforations and/or bleeding occurring at the rate of 4 and 8 events per 10,000 colonoscopies, respectively. Such iatrogenic injuries because of colonoscopy procedures follow a non-linear but significant rise with increasing age and have a significant co-morbidity burden. In addition to this, there is minor risk of splenic injury requiring splenectomy. Further, sedative drugs used in colonoscopic procedures often leads to cardiopulmonary complications such as minor fluctuations in heart rate to cardiac arrhythmias, myocardial infarction and respiratory arrests. *In toto,* even though there are associated risks with colonoscopy, it is by far the most specific, sensitive and effective screening method for CRC in use today (Wolf et al., 2018).

2.4.2.2 Computer Tomography Colonoscopy (CTC)

CTC involves generation of thin-slice computed tomography images through that can be analysed in two-dimensions and/or reconstructed in three-dimensions. Effectively, it reconstructs what can potentially be observed through a colonoscopy. In such a test, a small tube-like instrument is inserted a short distance into the rectum, where images of the colon and the rectum are taken. CTC has a sensitivity of 96.1% for CRC detection and 73-98% for detection of adenomas ≥6mm with a specificity ranging from 89-91%. The risk of bowel wall perforation is comparatively lower for CTC when compared to colonoscopy. CTC is far more effective than barium enema for colorectal imaging. The disadvantages of CTC are that less rigorous patient bowel preparation strongly affects the quality of the test. CTC also performs poorly when it comes to detecting polyps ≤1cm, as well as flat or serrated lesions (Wolf et al., 2018). The USPSTF advocates a CTC screening every 5 years between the ages of 50-75 years as a model recommendable strategy. Strong evidence of CTC reducing CRC mortality has yet to be established. Radiation exposure like with other radiological tests needs to be considered as a potential risk. A positive result CTC result would inevitably be recommended for standard follow-up colonoscopy (Rex et al., 2017; Song and Li, 2016).

2.4.2.3 Flexible Sigmoidoscopy (FS)

As the name suggests, a FS scope allows imaging of the rectum and the distal end of the sigmoid colon spanning about 61cm. This was one of the first structural imaging examinations shown to be efficient in detecting CRC. The procedure does not require extensive bowel preparation or sedation in patients. In a study reported by USPSTF, CRC mortality was reduced

by 27% after 11-12 years of follow-up. This reduction in mortality was found to be significant in distal CRC but did not translate to proximal CRCs. The MISCAN modelling recommends FS once every five years in individuals aged from 45-75 years. FS scores in its underlying cost and risk associated when compared to colonoscopy. The procedure involves less rigorous bowel preparation that could be of benefit, whilst the lack of sedation leads to an unpleasant procedure for patients. For this reason, patients are often non-compliant to follow-up or repeat the procedure. Moreover, the fact that only a part of the colon is tested makes it unpopular in the USA. It has been reported in 2010 that only 2.5% of individuals aged 50-75 years underwent FS whilst 60% underwent colonoscopy (Nguyen and Weinberg, 2016; Wolf et al., 2018).

2.4.2.4 Capsule Colonoscopy

Capsule colonoscopy is a procedure derived from capsule endoscopy, that allows the imaging of the entire GI tract. It involves ingestion of a capsule with a camera placed on both ends of it. As the capsule passes through the GI tract, it provides evidence of polyps or CRC in the colon or rectum as recorded images stored in an external device and later analysed by a medical professional. The test is considered complete when the capsule is passed in the stool. The diagnostic accuracy of this procedure on individuals with signs/symptoms of CRC or individuals having a high-risk of developing CRC has been assessed. It was found that the pooled sensitivity and specificity for capsule colonoscopy was 87% and 76% respectively for polyps ≥6mm. These numbers improved for larger polyps with a sensitivity and specificity of 89% and 91% respectively (Wolf et al., 2018).

Capsule colonoscopy is an attractive procedure for CRC detection because it is non-invasive, and it circumvents hazards associated with colonoscopy. Thus, capsule colonoscopy is a promising technology available for cases where despite adequate preparation, incomplete optical colonoscopy was achieved, and/or complete colon examination was not possible. Capsule colonoscopy is also an ideal technique in patients who are not candidates for receiving colonoscopy or sedation due to other ailments. The disadvantages associated with this procedure involves the burden of a more extensive bowel preparation than standard colonoscopy. Moreover, there is a logistical barrier in performing same-day colonoscopy for individuals testing positive in capsule colonoscopy. Adverse events associated with this procedure included nausea, vomiting, abdominal pain and fatigue in <4% of patients who were administered capsule colonoscopy (Rex et al., 2017). The most serious issue reported was capsule retention in 0.8% of patients. The FDA approves the use of capsule colonoscopy for

detecting colon polyps in patients suffering from lower GI bleeding. However, this procedure has not yet received clearance as a diagnostic modality for CRC screening (Song and Li, 2016; Wolf et al., 2018). Various recommended CRC screening methods and respective sensitivities and specificities is summarised in Table 2.1.

Table 2.1: Recommended CRC screening methods

Screening tests	Type	Screening intervals	Performance	Preparation	Limitations	Patient burden	Costs and further tests requirement	Sensitivity and Specificity (respectively)	References
Faecal Immunochemical Test (FIT)	Stool-based	Every Year	Equivalent or superior performance compared with gFOBT and tests performance can vary between brands	No	Nonadherence in annual testings Poor detection for advanced adenomas	Usually done at home, or clinics	Low cost but to be followed by colonoscopy if positive.	73%-92% and 91%-97%	(Nguyen and Weinberg, 2016; Wolf et al., 2018)
High-Sensitivity gFOBT (HSgFOBT)	Stool-based	Every Year	Good evidence in incidence and mortality Performance varies by different versions of the test.	No	High nonadherence to annual testing Limited data available from various FDA cleared tests.	Usually done at home, or clinics Requires dietary and medication restrictions Higher false-positive rate than FIT	Low cost but to be followed by colonoscopy if positive.	62%-79% and 87%-96%	(Nguyen and Weinberg, 2016; Wolf et al., 2018)
MT-sDNA	Stool-based	In Every 3 years	Performance monitoring is needed over time	No	limited data on screening outcomes	Can be done at home Higher false-positive rate than FIT	Expensive than other stool-based tests To be followed up by colonoscopy if positive	92.3% compared to 73.8% FIT with specificity of 89.80%	(Nguyen and Weinberg, 2016; Wolf et al., 2018)

Colonoscopy	Structural and visual examination	In Every 10 years	This test is suitable for both early detection and prevention of CRC through polypectomy (removing polyps from inside the colon)	Yes	Risk of bowel perforation- 4 in 10,000 major bleeding- 8 in 10,000; and cardiopulmonary complications of anesthesia-2-4 in 10,000. Adherence data to colonoscopy 10-year	Laxative preparation and bowel cleaning	Most expensive test Polypectomy and anaesthesia may be expensive	75%-93%; for adenomas ≥ 6mm and 89%-98% for adenomas ≥ 1cm Specificity: 94% for adenomas ≥ 6mm and 89% for adenomas ≥ 1cm	(Nguyen and Weinberg, 2016; Wolf et al., 2018)
Computed Tomography Colonoscopy (CTC)	Structural and visual examination	In Every 5 years	Comparable performances to colonoscopy in identifying advanced adenomas without procedural risks of colonoscopy and Exposure to low-dose radiations	Yes	Incidental extracolonic findings may require workup, with unclear benefit-burden balance	Laxative preparation and bowel cleaning Colonoscopy required if test positive, sometimes on the same day	Relatively expensive To be followed up by colonoscopy for positive test	75%-98%; for adenomas ≥ 6mm and 89%-91%; for adenomas ≥ 6mm	(Wolf et al., 2018)
Flexible Sigmoidoscopy (FS)	Structural and visual examination	In Every 5 years	Best evidence among structured examination for reducing mortality and incidence	Yes	Risk of bowel perforation- 1 in 10,000; major bleeding- 2 in 10,000 Suboptimal visualisation due to poor bowel cleaning Poor sensitivity for CRC in proximal colon	Pain and discomfort Self-administration of enema prior to procedure Abnormal findings require second endoscopic procedure (colonoscopy)	To be followed up by colonoscopy for positive test	-	(Wolf et al., 2018)
Septin9	Blood-based (DNA methylation)	Not determined	Serum assay potentially more convenient for patients	No	Marked inferior than FIT Inability to detect adenomas Low cost-effective in comparison to other tests	Patient preferred septin9 test over colonoscopy and FIT	To be followed up by colonoscopy for positive test Test is expensive than FIT	77% and 80%	(Nguyen and Weinberg, 2016)

2.5 Minimally invasive CRC biomarkers for early screening

Biomarkers can be defined as any *"biological molecules found in blood, other body fluids or tissues that are a sign of a normal or abnormal process, or of a condition or disease. Biomarkers may be used to see how well the body responds to treatment for a disease or condition"* (Mayeux, 2004).

The US FDA is the main regulatory body that approves the clinical use of biomarkers. The FDA categorises biomarkers into seven categories, namely, susceptibility/risk, diagnostic, monitoring, prognostic, predictive, pharmacodynamic/response, and safety. It approves use through a stringent, multistage regulatory process (Group, 2016). Numerous research endeavours have shed light on the basic biology of CRC, and these have led to the identification of some promising CRC biomarkers candidates. These markers represent potential anomalies in physiological molecules, including DNA, RNA, proteins, glycans and other post-translational modifications and/or metabolites within biospecimens.

In an attempt to, shift CRC diagnosis from procedure-centric to minimally invasive, cost-effective tests, researchers have tested several putative DNA, RNA and protein candidates in faecal, tissue, urine and blood specimens for potential as a diagnostic, prognostic and/or therapeutic marker for CRC. Putative CRC biomarkers from stool samples have been described earlier. Notable biomarker candidates proposed from, blood and urine in the past decades will be discussed here.

2.5.1 Urine Biomarkers

Tumour-derived DNA at circulating levels in plasma that are high enough to exceed renal absorption enters the urine. Alongside proteins and volatile compounds, tumour-derived DNA has been deemed as a useful source of non-invasive biomarkers for the early detection of CRC. These include DNA for arylsufatase, lysosomal exoglycosidases and cathepsin D (Altobelli et al., 2016); prostanoids metabolites, Prostaglandin E2 and M (PGE2, PGEM) (Davenport et al., 2016); (N(1),N(12)-diacetylspermine) and DiAcSpd (N(1),N(8)-diacetylspermidine (Umemori et al., 2010), epigenetic markers involving changes in DNA methylation in genes like vimentin, Wif-1 and ALX-4 (Amiot et al., 2014); mutation in genes such as KRAS (Su et al., 2008). In addition, elevated levels of volatile organic compounds like citrate, hippurate, p-cresol, 2-aminobutyrate, myristate, putrescine, and kynurenate, and nucleosides including adenosine, cytidine, N(2),N(2)- dimethylguanine, 8-hydroxy-2'-deoxyguanosine and uridine have been

proposed as putative indicators of CRC, in other studies(Cheng et al., 2012; Hsu et al., 2009). Urinary PGE-M seem to be promising candidates for detection of adenomas and CRC, which has been examined in more than five studies one of which evaluates 1000 individual patients (Altobelli et al., 2016) (Cheng et al., 2012). Therefore, evidence suggests that urine biomarkers can potential be used to screen early CRC; however, collection of urine samples can encounter the challenge of patient adherence to testing by the populace.

2.5.2 Serological (Blood-based) Biomarkers

The dynamic nature of blood vasculature and its constituents is reflective of individuals' physiological and pathological state. In addition, the ease of sampling makes blood not an only a logical choice but also studies have shown that more than 85% of patients prefer blood-based diagnosis over stool-based tests. Blood components that provide an indication of cancer status include various cellular elements such as proteins, peptides, metabolites, circulating tumour cells, cell-free DNA and RNA. Thousands of publications have explored the use of various components of the blood to diagnose cancer early (Table 2.2). In the following section details about the protein markers, and nucleic acid markers is discussed.

2.5.2.1 Protein biomarkers

There are numerous serological biomarkers currently under investigation for CRC early diagnosis. However, only a handful of these have proceeded to be available in the form of diagnostic blood tests.

For example, carcinoembryonic antigen (CEA) is an oncofetal protein whose level is found to be elevated in the serum of late stage CRC patients. Being highly specific (87%) but not so sensitive (36%), CEA is an FDA recommended CRC recurrence monitoring marker. In more than one study, elevated levels of CEA delineate with later stages of CRC rather than earlier (Fakih and Padmanabhan, 2006).

Carbohydrate antigen 19-9 (CA 19-9) is a protein elevated in serum of individuals with CRC or other malignancies related to the gastrointestinal tract. Like CEA, CA 19-9 is a late stage CRC marker with a lower specificity than CEA (Mahboob et al., 2015; Vukobrat-Bijedic et al., 2013). Tumour associated glycoprotein (TAG 72) is one such protein marker recommended in combination to other CRC biomarkers in a panel. The sensitivity of TAG 72 as a stand-alone

CRC marker has been found to range between 28% to 67% (Yanqing et al., 2018). Tissue polypeptide specific antigen (TPS) is released from cells after mitosis, thus serving as an appropriate marker for the rapidly dividing oncogenic cells. TPS can serve as an early diagnosis marker in a panel of CRC marker (Fiala et al., 2015). Further, of other protein markers proposed to date, two protein markers have been extensively studied, 1) Tumour specific M2 isoform of pyruvate kinase (PKM2) and 2) tissue inhibitor of matrix metalloproteinase 1 (TIMP1) (Fung et al., 2014). PKM2 is measured both in plasma and stool and has a reported sensitivity of more than 90% in stool but specificity in plasma is unknown. TIMP1 is reported to be elevated in CRC in comparison to healthy controls with reported sensitivity and specificity of 63% and 98% (Nielsen et al., 2008). However, sensitivity for early stage CRC (Dukes stage A and B) is poor at 56% (Holten-Andersen et al., 2002). In addition, both PKM2 and TIMP2 are less sensitive in comparison to FOBT (Tao et al., 2012; Wild et al., 2010).

Additional serological protein biomarkers under investigation include cytokeratins, dermokine (DK), melanotransferrin, N methyltransferase, neutrophil elastase, cathepsin D and lysosomal exoglycosidases such as isoenzymes of N-acetyl-β-D-hexosaminidase (HEX) (Rasmussen et al., 2013), HEX A and B, β-D-galactosidase (GAL), α-fucosidase (FUC), α-mannosidase (MAN), cathepsin D (Waszkiewicz et al., 2012), insulin-like growth factor binding protein 2 (IGFBP2), and matrix metalloproteinase 9 (MMP9).

However, none of these identified markers have proved to effectively detect early stages of CRC.

2.5.2.2 Nucleic Acid Biomarkers

Recent advances in technology have allowed researchers to monitor DNA shed from cells that enter circulation. These are commonly termed as cell-free DNA (cfDNA) and were described in times as early as 1948 (Volik et al., 2016). It is only now, that researchers have been able to exploit this for screening genetic aberrations or specific DNA fragments derived from tumour cells to aid diagnosis through NGS and RNA-Seq (Volik et al., 2016). These fragments of DNA are termed as circulating tumour DNA (ctDNA) could potentially provide a snapshot of the pathology of an individual and are therefore often referred to as “liquid biopsies” (Gorgannezhad et al., 2018).

In the context to CRC, numerous studies focus on viable ctDNA and ctRNA markers. Exploiting epigenetic aberrations triggering CRC, several serological DNA methylation markers have been investigated for their potential to diagnose CRC at an early stage. In one study, mBCAT1 and mIKZF1 measured together detected CRC with a sensitivity and specificity of 77% and 92.4% (Symonds and Young, 2015). The extent of methylation of BCAT and IKZF1 closely corelated with CRC stages I to IV at 50%, 68%, 87% and 100%, respectively (Jedi et al., 2018). Methylation abnormalities of DNA for Neuro D3/Neurogenin 1/NGN1 (NEUROG1), TAC1, EYA4, RUNX3, S100P promoter, p16 and THBD have also been found to be relevant in the context of CRC diagnosis (Vijeta Pamudurthy, 2016). Amongst these, the methylated Septin9 gene (mSEPT9) was recently approved by the FDA as the first serum-based nucleic acid assay for CRC screening. It is manufactured by Epigenomics, Seattle (Song et al., 2017). This non-invasive test screens *mSEPT9* shed by tumours and is a preferred test by patients not willing to undergo (or repeatedly refuse) other forms of CRC screening. This test diagnoses all stage CRCs and adenomas at a sensitivity and specificity of 68% and 80%, respectively, making performance inferior to tests like FIT and colonoscopy (Song et al., 2017). Moreover, compliance of patients after receiving a positive test to undergo colonoscopy for confirmation of diagnosis is uncertain. Being a novel blood test, information on mortality and incidence reduction or other critical CRC outcomes is obscure. This test is more expensive than standard, better-performing modalities like FIT. Because of these limitations, FDA clearance of mSEPT9 has not been forthcoming as a routine CRC screening modality for average at-risk individuals (Nguyen and Weinberg, 2016).

Further, various forms of RNA markers like messenger RNA (mRNA), miRNA and long non-coding RNA (lncRNA) have been studied for potential to allow early diagnosis of CRC. Marshall *et al.,* in 2010 introduced a 7-gene serological mRNA biomarker panel targeted at ANXA3, CLEC4D, LMNB1, PRRG4, TNFAIP6, VNN1, and IL2RB, from gene expression profiling data of CRC patients across various stages (Marshall *et al.,* 2010). This was further validated in a larger patient cohort and has demonstrated some clinical utility in determination of patient's risk of developing CRC, with a sensitivity and specificity of 78% and 66%, respectively (Ganepola *et al.,* 2014). This blood test for CRC diagnosis is currently called ColonSentry® and is manufactured by Innovative Diagnostic Laboratory (Ganepola et al., 2014).

The role of miRNAs has been increasing appreciated in the context of human malignancies. Numerous serological miRNA markers, including miR-7, miR-15b, miR-17-5p, miR-17-3pmiR-18a, miR-19a, miR-19b, miR-20a, miR-21, miR-29a, miR-92a, miR-96, miR-106b, miR-133a, miR-142-3p,miR-143, miR-145, miR-183, miR-195, miR-196a, miR-214, miR-221, miR-331, miR-335, miR-532-5p,miR-532-3p, miR-652, miR-1246 have been shown to be up-regulated in primary CRC (Schee et al., 2012), whilst miR-124, miR-127-3p, miR-138, miR-143, miR-146a, miR-222, miR-601 and miR-760 have been found to be down-regulated in primary CRC (Mitchell et al., 2008). Additionally, miR-15b, miR-29a, miR-139-3p, miR-141, miR-431 were found to be up-regulated in the metastatic disease.

Recently a broad spectrum, multi-analyte blood test called CancerSEEK was launched by researchers at John Hopkins Kimmel Cancer Center, Baltimore, USA and is being commercialised by Thrive Earlier Detection Corp. This test was designed to serve as an early pan-cancer diagnostic modality covering eight common types of cancers, including ovarian, liver, stomach, pancreatic, oesophageal, colorectal, lung and breast.

When applied to 1,005 patients with non-metastatic tumours, the test was positive in a median of 70% of all eight cancer types. The sensitivity of this test ranged between 69%-98% for detection of five cancer types (i.e., ovarian, liver, stomach, pancreatic and oesophageal) with the specificity of the CancerSEEK test over 99%. This test comprised two components. The first focused on ctDNA mutations in 16 candidate genes using 61 primer pairs, whilst the second component sought changes in expression of a 41-protein panel. An algorithm was then utilised on data obtained from the test to provide a specific diagnosis (Cohen et al., 2018). Careful analysis of this *Science* publication laid out a promising proof-of-concept multi-analyte blood test. However, the study had several limitations. The biggest limitation was that the subjects used in this study were pre-diagnosed with cancer. Therefore, they were distinct from the general “blinded” population on whom this test was intended for use. Moreover, the test seems to perform better for the five tumours ovary, liver, stomach, pancreas, and oesophagus but not as well for CRC.

More to this, CancerSEEK tests relies on a set of five literature derived plasma protein biomarkers CA125, CA19-9, CEA, HGF, MPO, OPN, PRL, TIMP-1. Additionally, CancerSEEK’s use of CA 19-9, CEA and TIMP1 is puzzling, as these markers have restricted FDA approval **specifically** for stage III/IV recurrence and tumour burden applications. CEA

does not differentiate adenomas or early stage I/II from healthy (Halford et al., 2013) and CA 19-9 and TIMP1 have also shown poor sensitivity for early CRC screening as discussed in detail above (section 2.5.2.1).

2.6 Unmet Clinical Needs:

Ostensibly, the holy grail for decreasing CRC mortality is the development of a new asymptomatic population screening/diagnostic test with crucial characteristics. The test must be:

- Specific - which means true negatives identified to prevent costly unnecessary interventions or psychological distress from false positive results
- Sensitive - which means true positives must be detected to ensure confirmatory coloscopy is prescribed and early enough for surgery to be curable
- have a high positive predictive value (PPV) - which depends on sensitivity, specificity and whether disease is common/rare
- be readily performed
- cause minimal discomfort
- be readily taken up - have compliance >90%, and
- be amenable in routine clinical practice and pathology settings.

Currently, after screening patients with existing tests (stool-based tests mentioned above in Table 2.1), adenoma/early stage CRC confirmation is performed by colonoscopy with subsequent surgical resection. Despite this, only 9% and 24% of patients are currently diagnosed at stage I/II respectively, with most diagnosed at stage III (23%) or IV (44%). Survival rates tumble when spread (metastasis) has occurred (e.g., <11% 5-year survival for stage IV patients; Figure 2.5).

Various test sensitivities, specificities are summarised (Table 2.1). Despite education programs, patient compliance for stool-based tests is lower than 40%, which leads to poor adherence to tests and annual testing. Also, the reliance of stool-based tests on detection of Hb in stool samples leads to detection of false positives, which exposes population to adverse effects of colonoscopy (discussed in detail in section 2.4).

On the other hand, blood tests have a patient compliance of ~95%, due to ease-of-use strategies, aimed at relieving discomfort around faeces handling (Elsafi et al., 2015).

Further, many companies such as GRAIL, Freenome, Clinical Genomics and more utilise mutant ctDNA liquid biopsy and/or methylation epigenomic technologies in asymptomatic population early stage screening. Colvera (Clinical Genomics) is now marketed as a clinically validated blood test for improved established CRC recurrence monitoring.

The feasibility of ctDNA for adenoma and early stage CRC screening has been recently discounted around issues, including; 1) capacity of deep sequencing to find early mutations, 2) undetectable or vanishingly low levels of mutant ctDNA released from small early stage tumours, 3) low apoptosis rates in adenomas and early tumours, 4) subsequent low/zero levels of ctDNA leakage from small benign tumours (Bettegowda et al., 2014; Cohen et al., 2017; Diamandis and Fiala, 2017), 5) accumulation of cancer-associated mutations with normal ageing, and 6) potential inability to identify cancer tissue of origin as many signatures overlap (Heitzer et al., 2017).

Significantly, cell of origin analysis surprisingly indicated driver mutations are also acquired in long-lived healthy tissues, stem or progenitor cells (Heitzer et al., 2017). More encouragingly though, a plethora of studies validate ctDNA (after deep sequencing of primary tumours) as enabling better understanding of later stage cancer, tumour heterogeneity, load and detection of patient resistance to therapy or presence of postoperative minimal residual disease (Fan et al., 2017).

The enthusiasm around the CancerSEEK blood test described in section (2.5.2.2) should be tempered due to the above-mentioned reasons. Further, a recent paper inadvisably uses the term early detection to refer to recurrence/relapse (Heitzer et al., 2017), unnecessarily creating confusion with the accepted meaning of early stage screening. Equally, many studies don't benchmark against previous epigenomic alterations or ctDNA mutations (e.g., SEPT92, APC3). Some tests like CancerSEEK (7.5ml), Clinical Genomics BCAT1/IKZF1 (~4ml) and Epigenomics Epi proColon methylated Septin9 (3.5ml) require large plasma volumes, precluding use of many established small draw plasma collections or the performing of preferred duplicate/triplicate assays.

Further, a faecal proteogenomics (genomic/epigenomic and proteomic) test involving Hb, β-actin, KRAS mutation and aberrant NDRG4 and BMP3 methylation (Imperiale et al., 2014) discussed above in section 2.4.1 had a high false positive rate, limited advanced adenoma sensitivity (42%) and suffered low compliance like other faecal tests (e.g., FOBT, gFOBT, FIT and Exact Sciences' Cologuard).

Therefore, there is still an unmet clinical need to find a blood-based biomarker that can screen early CRC patients from healthy controls and hence forms the primary objective of this thesis (**Chapter 3**).

2.7 Novel CRC Biomarkers by Plasma Proteomics

The potential for biomarkers to supplement/facilitate CRC early diagnosis poses a clinical necessity. Enzyme assays and immunoassays have dominated the FDA-approved blood-based diagnostic testing space for decades. However, the turnover of new FDA-approved protein biomarkers and the uptake of new assay platforms has been particularly slow. Over the last decade, proteomics has emerged as a powerful tool for accelerating identification of clinically significant protein biomarkers.

Proteomics is defined as the study of the entire repertoire of proteins (proteome) of a biospecimen under defined conditions at any given time point (Wilkins et al., 1996). It is a technology-driven field that allows the study of protein structure, function, post-translational modification, protein-protein interaction and abundance. Although proteomics is currently dominated by various forms of mass spectrometry (MS), other technologies are incorporated into major efforts like the Human Proteome Project (see **Chapter 1**).

MS has evolved over the years in terms of its technology and workflows to accurately measure protein abundance in biospecimens. Briefly, a mass spectrometer has three essential components: ionization source, mass analyzer and detector. In short, for the most commonly applied LC-MS/MS mode of mass spectrometry, enzymatically digested proteins from biospecimens are separated by high performance liquid-chromatography (HPLC) that is coupled to the electrospray ionization (ESI) source. ESI helps in imparting a charge to the peptides which are then separated as per their m/z ratio. The mass and abundance of each peptide is acquired in the mass spectrometer mass analyzer through a complete MS scan. The next dimension of MS involving peptide fragmentation pattern provides an MS/MS spectrum

which allows peptide identification after subjecting it to mass spectrometry-based data base searches. The details of common workflows involving MS-based proteomics is found in Section 2.9. MS-based proteomics can not only provide a global spectrum of the proteome of a biospecimen but is also well-suited to detect PTMs which could be of immense diagnostic value. (Aebersold and Mann, 2003).

Despite many advantages, MS-based proteomics has several challenges especially when challenged by such a complex biospecimen as human plasma. The primary hurdle in subjecting plasma to MS analysis lies in its relative proteomic composition (Anderson and Anderson, 2002).

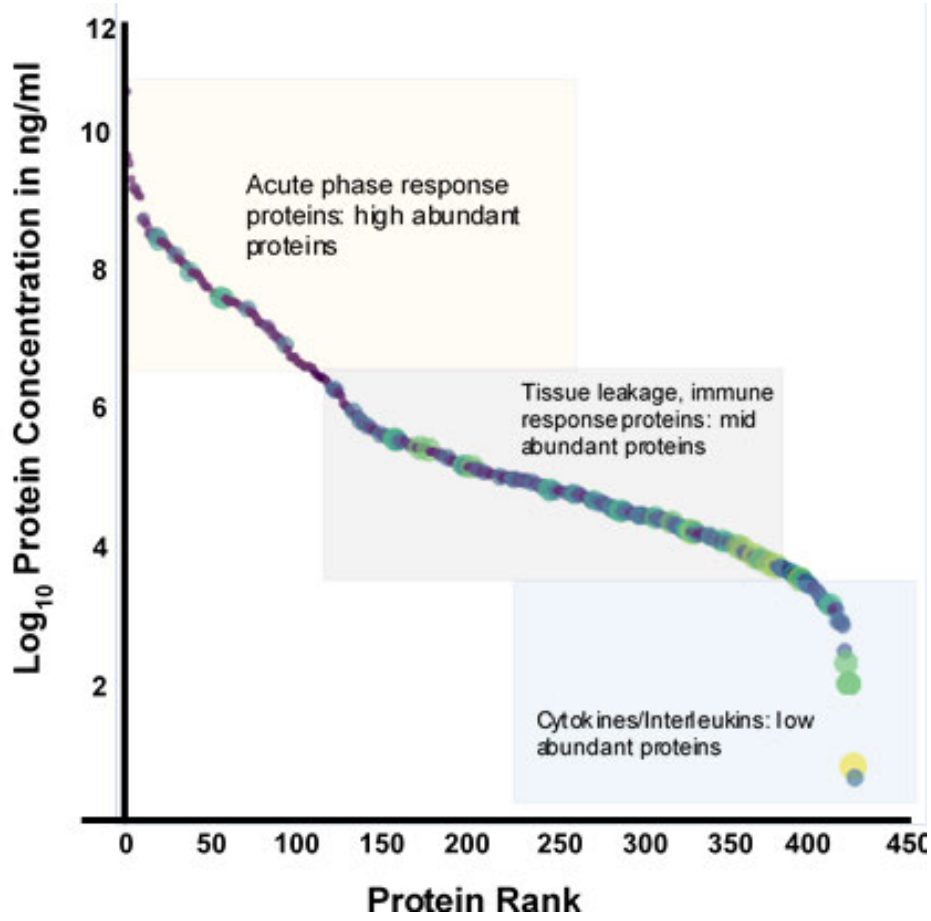


Figure 2.5: Protein Concentration Curve: *Data points represents an individual protein and its relative log concentration found in serum/plasma as measured with diverse methods and as retrieved from either the Plasma Proteome Database (*http://www.plasmaproteomedatabase.org/*) and/or PeptideAtlas (*www.peptideatlas.org/*) searched manually for the list of proteins obtained as a result in* ***Chapter 3****.*

The human plasma proteome can be broadly functionally grouped into immunoglobulins, acute phase response proteins, tissue leakage proteins (proteins secreted or shed from the cell into systemic circulation), and cytokines. Out of 100 or so FDA approved biomarkers, 50% can be functionally identified as acute phase response proteins, 25% as tissue leakage proteins with the remaining can be identified as immunoglobulins or receptor ligands. Many of these biomarkers have been in use for decades but continue to be routinely utilised in clinical care settings for example C-Reactive Proteins is used as an inflammatory marker and Cardiac troponin T as a cardiac test and many more (Geyer et al., 2017).

The dynamic concentration range difference between the most abundant plasma protein (albumin) and interleukins/cytokines (e.g., IL-6) (Figure 2.5) is over 10-12 log orders of magnitude. This fact alone decides which plasma proteome proteins are discoverable by MS. Over 90% of plasma proteome is dominated with systemic and acute phase response proteins found at mg/ml concentration can also be termed as the high abundance plasma proteome whilst mid and low abundance proteins are found in the ug/ml-mg/ml and ng/ml-pg/ml ranges, respectively (Anderson and Anderson, 2002).

Recently, two proteomics studies of unparalleled scale were undertaken. The first reported ~300 differentially-expressed glycoproteins in CRC tissues (parallel to our previous study (Sethi et al., 2016)) from which they derived a 5-protein consensus MRM signature using undepleted CRC plasma cohorts (Surinova et al., 2015a; Surinova et al., 2015b) . The signature consisted of ceruloplasmin (CP), serum paroxonase/arylesterase 1 (PON1), serpin A3 (SERPINA3), leucin-rich α2glycoprotein (LRG1), and tissue inhibitor of metalloproteinase-1 (TIMP1). Notably, all these proteins lay in the medium abundance range (0.8ng/ml→2mg/ml). Their test had a modest sensitivity (70%) and specificity (79%) that was strangely similar across all CRC stage /healthy comparisons. Addition of CEA cut-off of >5ng/mL did not appreciably change sensitivity. In conclusion, this study had lower sensitivity and specificity than colonoscopy, was inferior to FIT and was only marginally better than FOBT.

In the second study, Applied Proteomics extracted 187 CRC-associated proteins from the literature. Each protein was then assayed using multiplex MRM MS on MARS14-depleted plasmas from age and gender-matched patients and controls (Jones et al., 2016). A patented signature of ORM2, SERPINA1, AMY2B, CLU, C9, ECH1, FTL, GSN, TIMP-1, OSTP, SBP1, SEPR and SPON2 was generated and called SimpliPro Colon. This test showed a

comparable sensitivity (81%) but inferior specificity (78%) to FIT (Jones et al., 2016) on CRC stage I-IV cohorts. This is not surprising, given the number of abundant, liver-derived acute phase response proteins and complement factors that are known to be differentially expressed across a multitude of diseases/pathologies. In addition to this, multiple proteomics studies have proposed different liver-derived acute phase response proteins as CRC biomarkers (summarised in Table 2.2). Although, none of these markers have proven diagnostic value.

Table 2.2: List of CRC blood /serum/plasma protein biomarkers identified from proteomic studies

Status	Serum Protein Biomarkers	Sample Types	Ref
In use	CEA	Blood/Serum	(Lech et al., 2016)
Clinical Validation	TIMP-1	Plasma	(Lech et al., 2016)
Pre-Clinical Development	3 protein panel IGFBP2, DKK3 and PKM2	Blood/Serum	(Fung et al., 2015)
	4 protein panel DK-BLY, CEA, Ca 19-9, S-p53	Blood/Serum	(Shah et al., 2014)
	6 protein panel SULF1, NHSL1, MST1, GTF2i, SREBF2, GRN	Blood/Serum	(Babel et al., 2011)
	Alpha 1-antitrypsin	Blood/Serum	(Bujanda et al., 2013)
	Amphiregulin	Plasma	(Mahboob et al., 2015)
	C3a-desArg	Blood/Serum	(Fentz et al., 2007)
	Collagen type X alpha1 (CPL10A1)	Blood/Serum	(Sole et al., 2014)
	CP, PON1, SERPINA3, LRG1, CEA & TIMP-1CA125	Blood/Serum	(Surinova et al., 2015a)
	CXCL11	Plasma	(Mahboob et al., 2015)
	RM2, SERPINA1, AMY2B, CLU, CO9, ECH1, FTL, GSN, TIMP-1, OPN, SBP1, SEPR &	Blood/Serum	(Jones et al., 2016)

	SPON29 (v2 CEA, TFRC, A1AG, C09, DPPIV, MIF, PKM & SAA)		
	CXCL5	Plasma	(Mahboob et al., 2015)
	GRN	Blood/Serum	(Babel et al., 2011)
	GTF2i	Blood/Serum	(Babel et al., 2011)
	IL6	Plasma	(Mahboob et al., 2015)
	IL8	Plasma	(Mahboob et al., 2015)
	MMP9	Blood/Serum	(Mroczko et al., 2010)
	MMP9 + CEA	Blood/Serum	(Mroczko et al., 2010)
	MMP9 + TIMP-1	Blood/Serum	(Mroczko et al., 2010)
	MST1	Blood/Serum	(Babel et al., 2011)
	MUC1 + MUC4	Blood/Serum	(Pedersen et al., 2011)
	NHSL1	Blood/Serum	(Babel et al., 2011)
	RPH3AL auto-antibodies	Blood/Serum	(Chen et al., 2011)
	S100A8	Blood/Serum	(Kim et al., 2009)
	S100A9	Blood/Serum	(Kim et al., 2009)
	sCD26	Blood/Serum	(De Chiara et al., 2010)
	SREBF2	Blood/Serum	(Babel et al., 2011)
	SULF1	Blood/Serum	(Babel et al., 2011)
	TIMP-1	Blood/Serum	(Mroczko et al., 2010)
	Transthyretin	Blood/Serum	(Fentz et al., 2007)
	Transthyretin + C3a-desArg	Blood/Serum	(Fentz et al., 2007)
	uPAR	Tissue/serum	(Ahn et al., 2015; Bujanda et al., 2013)
	STK31	Serum	(Watany et al., 2018; Zhong et al., 2017)
	FBLN1	Serum	(Watany et al., 2018)
	Spondin-2, DcR3, Trail-R2, Reg IV, MIC1	Serum	
	PSME3	Serum	
	NNMT	Serum	
	CRMP-2	Serum	(Reviewed by (Tanaka et al., 2010))

	SELDI (apolipoprotein C1, C3a-desArg, α1-antitrypsin, transferring)	Serum	
	HNP 1–3	Serum	
	MIF	Serum	
	M-CSF	Serum	
	M2-PK	Serum	
	Prolactin	Serum	
	CCSA-2, −3, −4	Serum	
	MMP7	Serum	
	Laminin	Serum	

In this thesis, it was envisaged that the challenge of dynamic concentration range of plasma could be partially overcome by depleting the abundant protein load by passing plasma through columns containing immobilised antibodies against the top 1-200 proteins. There are several commercially available and proof-of-principle strategies available for depletion. In fact, a novel aspect about this thesis was to test the efficacy of an in-house chicken IgY-based ultradepletion strategy against commercial depletion column(Tan et al., 2013) (details section 2.9).

An alternative strategy to enhance plasma proteome coverage is to subject samples to extensive fractionation. Fractionation can be undertaken after plasma depletion and has helped identify several thousand plasma proteins. Further, the new advanced mass spectrometry technology SWATH™-MS is a revolutionary technique that compliments traditional MS by synergising the advantages of shotgun and targeted approach which yields high-throughput along with high reproducibility and consistency. It follows an intricate workflow to effectively capture a complete and permanent record of all fragment ions of detectable peptide precursors found in a biospecimen (section 2.9).

This thesis utilises the DIA method SWATH™-MS for plasma biomarker discovery aimed towards developing an early diagnosis modality to supplement gold standard colonoscopic confirmation or early stage CRC.

2.8 Triangular MS-based Biomarker Discovery Strategy

As discussed, the dynamic concentration range of human plasma poses immense technical challenges for in-depth plasma proteomic studies on population sized cohorts. To circumvent this, a triangular multi-staged biomarker discover approach was employed in this study. Our first step undertook shotgun-based library generation for subsequent SWATH™-MS discovery through non-depleted, depleted and ultradepleted plasma proteomics applied to plasma specimens from small pooled (N=20) cohorts comprising few cases and controls. This leads to the identification of tens-hundreds of differentially-expressed proteins (Geyer et al., 2017) as potential CRC biomarkers. Comprehensive SWATH™-MS data analysis leads subsequently to the identification of putative markers which would be then be verified in the next phase in larger individual or population cohorts using either a set number of multiplexed or individual targeted proteomic assays and/or commercially available ELISA/other assays.

Our lab is familiar with the development of targeted proteomic assays as these allow precise detection and quantification of sets of candidate proteins across an independent cohort by selectively monitoring peptide sequences specific to candidate proteins with high precision (Addona et al., 2009; Picotti et al., 2009) (section 2.9). Finally, the last stage discovery stage involve population-based validation and clinical application of the most robust biomarkers identified using clinical immunoassays on pathology-lab platforms and/or development of high-throughput systems to undertake novel targeted MRM/SRM assays (Anderson and Hunter, 2006). Although, all steps in this workflow was not completed, this thesis adheres to a long-term triangular biomarker discovery framework and has a view to identifying blood-based early-stage CRC screening biomarkers, as represented in Figure 2.6.

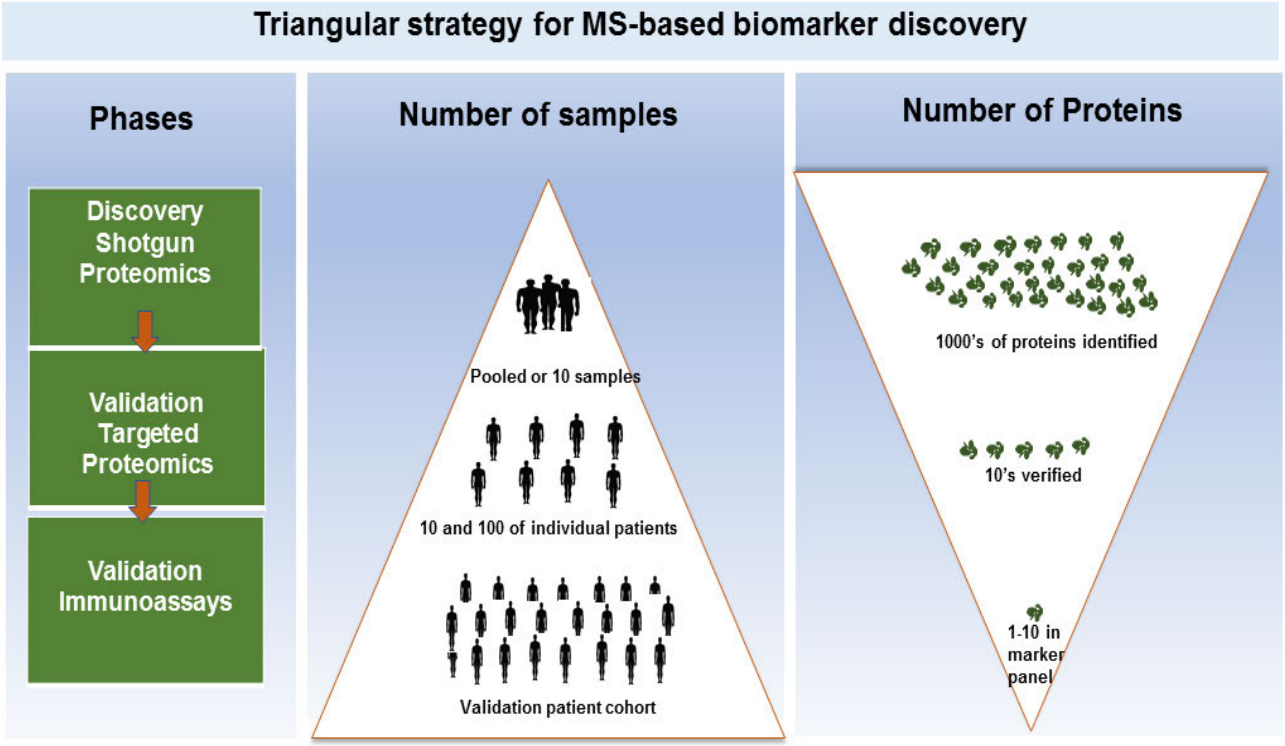


Figure 2.6: Triangular strategy paradigm: *Protein identification is done on a relatively small number or pooled set of samples which yields large datasets. Validation of a statistically significant protein set is performed on a relatively larger data set and final validation assay is performed as a population-based study for potential biomarker. Image adapted from (Geyer et al., 2017)*

2.9 Sample Preparation, Proteomics and Orthogonal tools for Biomarker Identification and Validation

2.9.1 Plasma Sample Preparation

2.9.1.1 Multiple Affinity Removal System (MARS14 - Human 14)

The MARS-14 targeted high-abundance protein depletion system that uses combinations of anti-human plasma protein antibodies attached to chromatographic supports was introduced by Agilent in 2003. Since then it has been employed widely by many labs to identify and characterise low abundant plasma proteins in unbiased biomarker discovery experiments. These MARS-14 columns containing immobilised antibodies can reproducibly and reliably remove14 abundant plasma proteins (i.e., human serum albumin, IgG, antitrypsin, IgA, transferrin, haptoglobin, fibrinogen, α2-macroglobulin, α1-acid glycoprotein, IgM, apolipoprotein AI, apolipoprotein AII, complement C3 and transthyretin) (Tu et al., 2010).The MARS14 columns provides reproducible and specific removal of targeted proteins in human

plasma and enhances the detection of mid-low abundant plasma proteins. However, the non-targeted plasma proteins obtained after enrichment using immunodepletion columns, on subjection to current multidimensional LC-MS/MS technology, detects mid- and high-abundance proteins. The low abundance proteins comprise of 5-6% of depleted plasma in range (Tu et al., 2010). To reach deeper depths of plasma, in-house IgY ultradepletion used in the study discussed in section 2.9.1.2.

2.9.1.2 In-house ultradepletion (API – abundant protein depletion)

Immuno-depletion of high-abundant proteins is widely used before proteome analysis. Despite using these early columns to effectively remove 14 of the most abundant plasma proteins, the dynamic range of plasma protein concentrations remains high and this continues to act as barrier to the discovery of clinically relevant biomarkers.

To address this ongoing problem, our Macquarie University team developed an in-house ultradepletion method that immunodepletes many more additional high- and mid-abundance human plasma proteins (Tan et al., 2013; Tan et al., 2012). This novel, patented immunoaffinity ultradepletion strategy allows for the binding/removal of ~165 human plasma proteins.

In brief, 30 litres of human plasma were fractioned by SCX followed by SAX and where dual IEX flow-through proteins were also collected. This method has been dubbed **P**rotein **R**epetitive **O**rthogonal **O**ffline **F**ractionation (PROOF) (Figure 2.7). Application of this method produced 7 human plasma fractions but maintained the native state of proteins in order to preserve the immunogen status of each. Fractions were injected intramuscularly into chickens that were used as hosts and eggs were collected as a source for chicken-derived, anti-human plasma antibodies (IgY). Yolks were separated, processed, filtered, concentrated, purified, then bound to a solid phase hydrazine bead support and packed into the ultradepletion columns as previously described (Tan et al., 2013). The combination of this novel strategy with use of commercial MARS-14 columns was subsequently shown to significantly increase discovery of lower abundance plasma proteins (Tan et al., 2013).

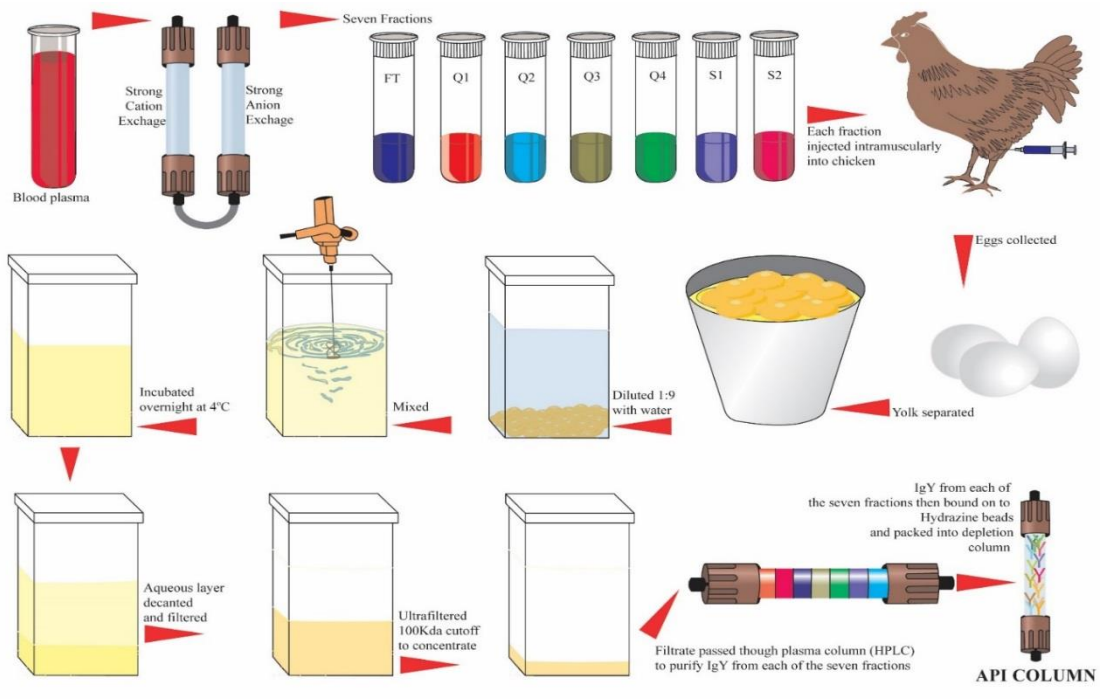


Figure 2.7: In-house API method designed to remove high abundance protein from human plasma- *Non-fractionated human plasma yielded seven fractions after subjecting to dual ion exchange chromatography (PROOF). The resultant seven fractions were used as protein antigens to immunize chickens. Eggs were collected, IgY antibodies extracted and antigen-specifically purified on plasma antigen columns. Specific IgY antibodies were subsequently bound to GE UltraLink affinity media support to make the API ultradepletion column. Image used with permission of the Baker research team and (Tan et al., 2013).*

2.9.2 Proteomics Tools

Proteomics is a congregation of techniques employed to understand system wide changes in the protein complement that takes place in a given biological sample. Initial attempts to comprehend a holistic snapshot of the proteome dates to 1975 from endeavours by O'Farrell, Klose and Scheele who attempted to decipher proteins from various organisms using two-dimensional gel electrophoresis (2-DGE) (Graves and Haystead, 2002). Integration of MS-based techniques has catapulted proteomics to now enable the identification of tens of thousands of proteins in a single experiment. The following section describes advanced proteomics approaches for biomarker discovery and validation of identified candidates

2.9.2.1 Discovery Mass Spectrometry

A mass spectrometer is an analytical instrument that ionises compounds to identify them based on m/z ratios. As described earlier, it is divided into three components. The first component is

the ionisation source. Proteins are labile in nature and the success of mass spectrometers in the field of proteomics can be primarily attributed to the development of soft ionisation techniques like matrix-assisted laser desorption/ionisation (MALDI) and electrospray ionisation (ESI). While MALDI has niche proteomic applications, where the analyte is immobilised on a solid substrate (matrix) and crystallised, most modern LC-MS platforms use ESI as their ionisation source. ESI involves the injection of the liquid analyte through a fine metal capillary tip which generates fine droplets. These droplets are ionised when passed through an electric field and then analysed by the mass analyser, which is the second component of the mass spectrometer. To enhance the resolution of the analyte subjected to mass spectrometry, the liquid analyte is separated by high pressure liquid-chromatography (HPLC) prior to ionisation, therefore the term "LC-MS". Depending on the source of ionisation and downstream objective of the study, several mass analysers have been developed e.g. time-of-flight (TOF), quadrupole, ion traps and Fourier transform ion cyclotron resonance (FT-ICR). TOF can be used in tandem with MALDI, whilst ion traps and FT-ICR are commonly used mass analysers for quantitative LC-MS. Quadrupole-based mass analysers are finding extensive utility in targeted proteomics. The third component of the mass spectrometer is the detector that amplifies signals from ions hitting it to give a readout that aids identification of peptide masses and hence inferred peptide/protein sequence. Commercial mass spectrometers primarily use specific combinations of the above three components to serve applications to many of the downstream proteomic approaches described subsequently in this section.

The choice of approach used in proteomics depends on the end goal of any study. LC MS/MS is the workhorse of proteomics largely due to its sensitivity and high-throughput nature. However, an understated benefit of this technology is its flexibility to integrate and switch between various discovery-based strategies used in proteomics. These strategies can be largely divided into "top-down" and "bottom-up" approaches.

2.9.2.1.1 Top-Down Proteomics

Functional characterisation of proteins is attributed to their structural features, functional domains and motifs. The dynamic post-translational modifications of proteins are key to the proper working of signalling cascades. Top-down preserves the intermolecular complexity of proteins and allows better characterisation of proteins and their post-translational modification. Top-down applies mass spectrometry to study the proteoform level information of an intact protein (Toby et al., 2016). Proteoforms refers to the all the molecular form of protein that is

derived from a single gene, which includes post-translational modifications, genetic variants and alternative splice variants of RNA transcripts (Aebersold et al., 2018; Smith and Kelleher, 2013).

Unlike bottom up where proteins are digested, and peptides subjected to mass spectrometric analysis, top-down involves direct infusion of intact proteins into the mass spectrometer. Proteoforms under analysis are purified or selectively isolated and often pre-fractioned upon isoelectric point or molecular mass that and can be further resolved using different liquid-chromatographic modes. Fractionated samples are ionised or desorbed either through matrix-assisted laser desorption-ionisation (MALDI) or electrospray ionisation (ESI) for detection and identification (Catherman et al., 2014). Ionised intact proteoforms are analysed by either Orbitrap or Fourier transform ion cyclotron resonance (FT-ICR) mass analysers. In order to distinguish between different post-translational modifications and/or allow proper peak resolution, the accurate spectral peak assignment from complex precursor spectra of intact proteoforms and their fragment spectra mass analysers with high resolution, mass accuracy and sensitivity is critical (Compton et al., 2011; Tipton et al., 2011).

2.9.2.1.2 Bottom-up (Shotgun) Proteomics

Complex biological samples often contains thousands of proteins at different concentrations (wide concentration range, the concentration at multiple order of magnitude- relevant if for plasma) (Meier et al., 2018). Shotgun proteomics is widely used for identification and quantification of a wide range of proteins. Shotgun proteomics involves peptide generation upon enzymatic digestion of a protein mixture where these peptides are subsequently used as surrogates for respective protein identification (by inference) when compared with *in silico* generated peptides (Wolters et al., 2001). Resolution of peptides involves chromatographic separation in an LC with peptide spectra generation in tandem-MS (full scan MS1 to measure signal intensity and m/z ratios followed by identification of fragmentation pattern in MS2) and peptide identification and protein inference by search engine upon comparison with protein sequence database (Altelaar et al., 2013). Shotgun proteomics workflow allows quantification in tandem with identification of multiple proteins, and has been adopted using multiple quantification strategies, including label-based, label-free, DDA and DIA as outlined below (Ong and Mann, 2005).

2.9.2.1.3 Label-based Strategies

Label-based quantitative mass spectrometric approaches are commonly used for studying quantitative proteomic alterations produced during various pathological conditions. This approach involves differential labelling of peptides obtained from two or more biological conditions using stable isobaric or isotopic tags. These approaches include stable isotope labelling of amino acids in cell culture (SILAC), isotope-coded affinity tags (ICAT), isobaric tags for relative and absolute quantitation of peptides (iTRAQ) and tandem mass tags (TMT). SILAC is a metabolic labelling method that can be used to study proteomic alterations that occur *in vivo* in cell lines and animal models. In this method, heavy isotope containing arginine and lysine amino acids are added to the culture medium and which are incorporated into newly synthesised proteins in cells. After 100% metabolic labelling, proteins extracted from cells are subjected to tryptic digestion followed by LC-MS/MS. Protein quantification can be performed by calculating the ratio of intensities of precursor ions from different conditions, which differ from each other with a known difference in mass. SILAC can only be used to compare a maximum of three experimental conditions at any given time and cannot be used for *ex vivo* chemical labelling of proteins (Kruger et al., 2008; Seyfried et al., 2010). *In vitro* labelling strategies like iTRAQ, ICAT and TMT are commonly used for comparative proteome profiling.

ICAT is an isotopic label that binds to cysteine residues in proteins (Gygi et al., 1999). The labelled protein extract is digested, and ICAT-labelled peptides enriched using biotin-streptavidin affinity chromatography. Protein quantification is further performed by subjecting the ICAT-labelled peptides to LC-MS/MS. iTRAQ is another *in vitro* labelling technique that allows parallel comparison of either four (iTRAQ-4plex reagent) or eight (iTRAQ-8plex reagent) biological conditions. iTRAQ reagents bind to the free amino terminal of peptides and side chain amino groups of lysine residue (Ross et al., 2004). The pool of labelled peptides from different conditions is subjected to LC-MS/MS based identification. Intensity of the isobaric tags is indicative of relative abundance of proteins in different experimental conditions. TMT-based labelling approach is similar to that of iTRAQ, which allows comparative analysis of ten experimental conditions simultaneously in a single mass spectrometry run (Werner et al., 2014). Label-based quantitative proteomics allows parallel screening of protein alteration in multiple biological conditions thus help in reducing run-run variation.

2.9.2.1.4 Label-Free Strategies

Label-free aims at direct comparison of relative abundance of proteins across mass-spectrometric runs. Label-free quantitation works on the principle that the protein concentration in a sample corroborates well with spectral counts or peak intensities of the peptides that are unique to any specific protein. Spectral counts of the peptide provide the number of tandem spectra obtained for each protein whereas peak intensities are obtained by integrating the area under the curve across retention time windows (Zhu et al., 2010). Unlike labelling methods, label-free approaches are cost-effective, involve minimal sample preparation and allow comparison of multiple biological conditions. However, data analysis for label-free is technically more complicated than labelled-based strategies as the peak areas, m/z and retention time for each peptide should be well-aligned across multiple mass spectrometric runs (Van Riper et al., 2013). Also, the protein sample across different biological conditions is measured separately, which leads to increased variability in data acquisition. Therefore, for label-free strategies includes proper calibration of mass spectrometers and chromatograph between LC-MS/MS runs and multiple technical and biological replicate runs to acquire reliable quantification of proteins with adequate levels of statistical significance (Lin and Garcia, 2012; Zhang et al., 2013).

2.9.2.1.5 Data-Dependent Acquisition (DDA) Shotgun Proteomics

Shotgun proteomics quantifies proteins by indirectly measuring peptides obtained after proteolytic digestion of intact proteins. A typical shotgun experiment comprises of fragmenting a peptide mixture digested via trypsin before subjecting to LC-MS. Peptides are identified by mapping the mass spectra obtained from peptide fragments against protein databases such as Sequest and MASCOT. Peptides are uniquely assigned to a protein but in few cases redundant and homologous protein sequences interfere in the identification of the proteins. This interference issue is one of the major challenges of shotgun approaches. One of the most cited reason quoted is the sequencing speed at which LC-MS platform processes fragment ions. In general, in DDA workflows the first MS1 scan cycle of eluting peptides lasts for 1sec during which peptide intensity is monitored and identified. Following this a series (~10) of MS2 scans occur during which each precursor ion is isolated, fragmented and product ions detected. Precursors ions are fragmented in order of decreasing intensity which implies that the most intense precursor ions are collected in the MS1 survey scan and further subjected to detection leading to inadequate sequencing opportunities. This allows two things; 1) hinders the identification and detection of peptide with low ion signals and 2) peptide interference leading

to identification of homologous proteins. These problems can be largely overcome by data-independent acquisition (DIA) and this is discussed below.

2.9.2.1.6 Sequentially Window Acquisition of Theoretical Spectra (SWATH™-MS)

Recent advancements in the field of qualitative and quantitative mass spectrometry has led to a new era of accurate and reproducible label-free quantification where focus has shifted from enumerating peptides and proteins to consistent quantification of large-scale samples.

This advancement is particularly necessary for fields like proteogenomics, biomarker discovery and drug screening. An emerging technology designed to meet these requirements has been called SWATH™-MS (Sequentially Window Acquisition of Theoretical Spectra). SWATH™-MS is a registered trademark of SCIEX, and it represents one form of Data Independent Acquisition (DIA). SWATH™-MS fundamentally differs from data-dependent acquisition (DDA) MS (described above) by performing repeated cyclical acquisition of precursor ions with fixed isolation windows whilst capturing an entire m/z range (Ludwig et al., 2018). In this manner, SWATH™-MS fragments and collects spectra for all precursor ions within a sample, allowing retrospective examination of all peptides after the generation of a comprehensive spectral library. In conventional DDA-MS methods, fragmentation occurs for a fixed number of abundant precursor ions in a single survey scan. In a general SWATH™-MS workflow, any biospecimen is trypsin digested and fractionated with an LC attached to the tandem mass spectrometer (hybrid full-scans, preferably Q-TOF and Q-Orbitrap). All fragmented precursor ions are isolated in 32 (sometimes more or less) precursor isolation windows in a specific mass range of 25m/z each (Figure 2.8).

Conventional DDA measures proteomic profiles of specimen and is the choice of method for discovering the maximal number of proteins in a biomarker discovery experiment but often it is limited by irreproducible quantification and stochastic precursor ion selection. The variable precursor isolation width of SWATH™-MS or DIA provides an edge and alleviates this limitation of DDA (Gillet et al., 2012).

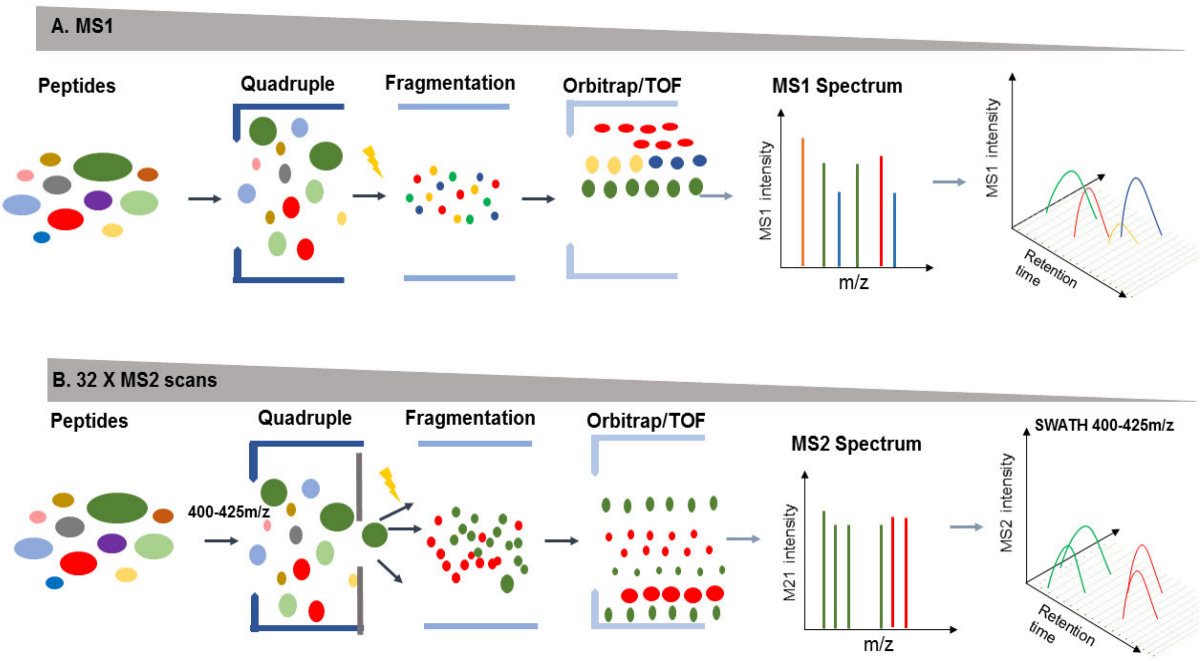


Figure 2.8: Principle of SWATH™-MS *A) measurements are performed on fast scanning hybrid mass spectrometer, classically a quadruple which act as first mass analyser and a TOF or orbitrap as a second mass analyser. In SWATH™-MS mode, typically a single precursor ion (MS1) spectrum is recorded, B) followed by series of fragment ion generating MS2 spectra with wide precursor isolation window (e.g. 25m/z) in repeated cycling using 32 MS2 scans with defined isolation window (400-425m/z).*

Critical Appraisal for SWATH™-MS

SWATH™-MS is a powerful and advanced mass spectrometry technique for in-depth proteome analysis and coverage, although it has some significant limitations (Law et al., 2013). Primary among these is that the data emerging from a SWATH™-MS experiment are incompatible with traditional database searching platforms. Although there have been notable efforts from the community to build algorithms to deconvolute data from SWATH™-MS acquisition, it poses challenges like production of chimeric spectra formed due to residual precursor ions. Next, the Triple TOF, the preferred instrument for SWATH™-MS does not share the high mass accuracy and resolution of Orbitrap or FT-ICR which poses a question of specificity in mass determination. Another problem of SWATH™-MS arises from the 25 theoretical width of the precursor isolation widow which may give rise to interference that is likely to affect the precision of quantification. Hence, validation using SRM/PRM following SWATH™-MS data analyses is imperative.

2.9.2.2 Validation-based Proteomics

Proteomics is a rapidly developing field and with the advent of high-resolution mass spectrometers paradigms have shifted from discovery-based proteomics to validation and quantitation with high precision and accuracy. Targeted approaches allow multiplexing, and hence require minimal clinical sample for validation of multiple candidate proteins (Fortin et al., 2009; Harlan and Zhang, 2014). Unlike discovery phase proteomics that allows monitoring of thousands of protein alteration, targeted-based assays allow repeated monitoring of a few specific scheduled peptides and their fragment ions belonging to proteins of interest. This approach utilises specific properties of peptides like hydrophobicity, mass/charge ratio (m/z) and fragmentation patterns to detect and quantify peptides in complex biological mixtures. These MS-based validation experiments if performed on a triple quadrupole are referred to as selected/multiple reaction monitoring (SRM or MRM) and if performed on Orbitrap mass spectrometers are known as parallel reaction monitoring (PRM) assays.

SRM/MRM-based validation typically monitors multiple transitions (pair of precursor peptide and daughter ions) of the target protein using a triple quadrupole mass spectrometer. The resolved peptides are first selected on the basis their mass/charge (m/z) ratio in the first quadrupole (Q1) which then gets fragmented using collision induced dissociation at second quadrupole (Q2), the third quadrupole (Q3) further isolates the fragment ions of based on specific m/z. Multiple and repeated monitoring of the specific set of peptides and their fragment ion provides selectivity and reproducible quantitative measurements (Gillette and Carr, 2013). The area under the curve obtained for the transitions are compared to those for internal standards and are used for quantitation. Mass-based filtering of peptides at two levels (i.e., Q1 and Q3) effectively excludes most co-eluting interferences, making SRM a highly sensitive technique (Figure 2.9). However, owing to the low resolving power of the quadrupole, interfering peptides with near-isobaric patterns and similar MS/MS fragmentation pattern may co-elute (Gallien et al., 2013). These co-eluting peptides may result in erroneous results.

Moreover, in SRM experiments several optimisation and iterations are needed for defining the optimal set of transitions to be monitored for newly examined proteins (Prakash et al., 2009; Rauniyar, 2015). Thus, the success of SRM-based validation relies on the set of pre-selected transitions that are defined for any candidate proteins. In order to ease basic optimisation steps for SRM experiments, multiple databases are available containing information concerning transitions that can be used to monitor specific proteins.

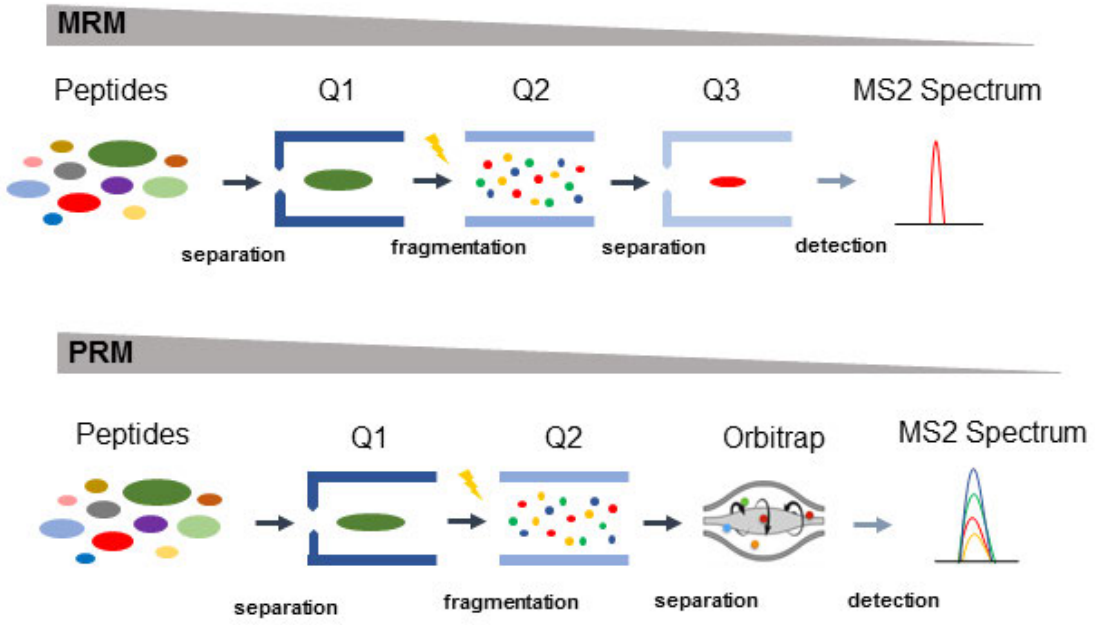


Figure 2.9: Schematic representation of principle of peptide selection in multiple reaction monitoring and parallel reaction monitoring*. In MRM method, selected single precursor ion is measures whereas in PRM's all fragment ions are quantified at the same time.*

Parallel reaction monitoring (PRM) has emerged as an alternative strategy. PRM is also based upon monitoring set of quantotypic peptides for target proteins. However, unlike SRM that monitors a few transitions for each peptide, PRM allows monitoring of all possible fragment ions derived from a specific peptide. Typical instrumentation for PRM includes a quadrupole coupled to the Orbitrap mass analyser (q-OT) (Eliuk and Makarov, 2015; Gallien et al., 2012; Kim et al., 2016). q-OT mass spectrometers can be used for both discovery-based and targeted experiments, thus allowing use of similar parameters used for two data acquisition methods to circumvent the need to optimising multiple parameters. The quadrupole allows isolation of specific peptide based upon the m/z ratio which is then transferred via C-trap to the higher energy collision-induced dissociation (HCD) cell for further fragmentation. The C-trap allows trapping of specified number of precursor ions thereby enhancing the signal to noise ratio (Domon and Gallien, 2015). Fragment ions obtained from the HCD enters the Orbitrap mass analyser and the MS/MS spectra are acquired with high mass accuracy and resolution. In PRM-based targeted experiments, complete MS/MS spectra of the targeted peptides are attained making protein quantitation highly specific and selective (Peterson et al., 2012).

Isotopically-labelled internal standards can be used to enhance precision, accuracy and reproducibility as it has been revealed that run-to-run variation arises due to interference from complex biological matrixes (Gallien et al., 2012). In order to eliminate repeat optimisation steps for MRM experiments, databases have been made available, which contain properly documented SRM coordinates that can be directly deployed for SRM-based validation.

Critical Appraisal for Targeted Proteomics

Although significant progress has been made in the field of targeted proteomics, both in terms of instrumentation and data analyses software, there remains impediments to the clinical application of these techniques. The primary reason for this relates to the specialised training required for the operation of a mass spectrometer and data analyses (Arora et al., 2019). While there are several resources that are freely available for training online, accessibility to instruments is not widespread. Another drawback is the feasibility of running a clinic-based SRM assay which may be more easily/economically performed by antibody-based ELISA techniques. Sample preparation and interpretation of results from immune platforms are also more straightforward/intuitive

SRM assays are also technically challenging. For instance, the development and standardisation of an SRM assay is time consuming and could take up to several months to optimise. Furthermore, several SRM assays require enrichment of proteins in samples using antibodies which can vary depending on the quality of antibody, and the amount of sample available. There are numerous options for setting up a targeted workflow and data analysis (Gonzalez-Galarza et al., 2012) which complicates the establishment of a "universal" SRM test. However, SRM/PRM is a powerful platform for target specifically detecting proteins. This makes it the most sought-after platform for validation experiments and for clinical application.

This thesis utilises targeted assays to verify the ability of candidate biomarkers to differentiate early stage CRC from healthy controls and to validate SWATH™-MS experiments that are described in **Chapter 3**.

2.9.3 Orthogonal Technologies

Immunoassays are a class of binding assays that measure the presence or concentration of an analyte (here a protein) with the use of an antibody. Depending on the application, numerous

immunoassays have been defined over the last 30 years and the two most predominant immune assays are Western blotting (WB) and ELISA.

Western blotting is a technique for visualising the presence of target proteins in a biospecimen through a series of intricate steps. Developed in 1979, WB was inspired by techniques like Southern blotting for DNA and Northern blotting for RNA that revolved around the principle of electrophoretic separation of the biological molecule of interest followed by detecting specific molecules using a suitable probe. In case of Southern and Northern blotting, the target nucleic acids are probed using a labelled sequence of nucleic acid complementary to the target molecule that could detect its presence via probe hybridisation. In case of WB, this is achieved through antibodies targeting the protein of interest that is being probed.

Enzyme-linked immunosorbent assays or ELISA was first described by Engvall and Perlmann in 1971. ELISA, in principle and methodology, is very similar to WB, and has been a preferred mode of detection and quantitation of a wide variety of proteins like cytokines, hormones, antibodies, for which reagents for detection are well characterised and readily available. One of the key advantages of ELISA over western blotting is its throughput. Typically, ELISA is performed in a 96/384-well plate where the antigen being probed is immobilised on a solid surface. The antigen is probed using an antibody that is enzyme linked so as to facilitate a reaction when a substrate is added to it. The product thus formed can be measured colorimetrically indicative of the presence and quantity of the sample (Engvall and Perlmann, 1971; Mahmood and Yang, 2012).
In this thesis, Western blotting was primarily used to validate our putative CRC markers in serum emerging from quantitative MS data. As would be described in subsequent chapters, un-depleted plasma contains a vast repertoire of proteins. This includes proteins like native antibodies produced by the patients, which necessitates running of adequate isotype controls to ensure the efficacy of validation assays.

References

Aarnio, M., Mecklin, J.P., Aaltonen, L.A., Nystrom-Lahti, M., and Jarvinen, H.J. (1995). Life-time risk of different cancers in hereditary non-polyposis colorectal cancer (HNPCC) syndrome. International journal of cancer 64, 430-433.

Addona, T.A., Abbatiello, S.E., Schilling, B., Skates, S.J., Mani, D.R., Bunk, D.M., Spiegelman, C.H., Zimmerman, L.J., Ham, A.J., Keshishian, H., et al. (2009). Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. Nature biotechnology 27, 633-641.

Aebersold, R., Agar, J.N., Amster, I.J., Baker, M.S., Bertozzi, C.R., Boja, E.S., Costello, C.E., Cravatt, B.F., Fenselau, C., Garcia, B.A., et al. (2018). How many human proteoforms are there? Nature chemical biology 14, 206-214.

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

Ahlquist, D.A., and Gilbert, J.A. (1996). Stool markers for colorectal cancer screening: future considerations. Dig Dis 14, 132-144.

Ahlquist, D.A., Skoletsky, J.E., Boynton, K.A., Harrington, J.J., Mahoney, D.W., Pierceall, W.E., Thibodeau, S.N., and Shuber, A.P. (2000). Colorectal cancer screening by detection of altered human DNA in stool: feasibility of a multitarget assay panel. Gastroenterology 119, 1219-1227.

Ahn, S.B., Chan, C., Dent, O.F., Mohamedali, A., Kwun, S.Y., Clarke, C., Fletcher, J., Chapuis, P.H., Nice, E.C., and Baker, M.S. (2015). Epithelial and stromal cell urokinase plasminogen activator receptor expression differentially correlates with survival in rectal cancer stages B and C patients. PLoS One 10, e0117786.

Altelaar, A.F., Munoz, J., and Heck, A.J. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. Nature reviews Genetics 14, 35-48.

Altobelli, E., Angeletti, P.M., and Latella, G. (2016). Role of Urinary Biomarkers in the Diagnosis of Adenoma and Colorectal Cancer: A Systematic Review and Meta-Analysis. Journal of Cancer 7, 1984-2004.

Amersi, F., Agustin, M., and Ko, C.Y. (2005). Colorectal cancer: epidemiology, risk factors, and health services. Clin Colon Rectal Surg 18, 133-140.

Amiot, A., Mansour, H., Baumgaertner, I., Delchier, J.C., Tournigand, C., Furet, J.P., Carrau, J.P., Canoui-Poitrine, F., and Sobhani, I. (2014). The detection of the methylated Wif-1 gene is more accurate than a faecal occult blood test for colorectal cancer screening. PloS one 9, e99233.

Anderson, L., and Hunter, C.L. (2006). Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Molecular & cellular proteomics: MCP 5, 573-588.

Anderson, N.L., and Anderson, N.G. (2002). The human plasma proteome: history, character, and diagnostic prospects. Molecular & cellular proteomics: MCP 1, 845-867.

Antonic, V., Stojadinovic, A., Kester, K.E., Weina, P.J., Brucher, B.L., Protic, M., Avital, I., and Izadjoo, M. (2013). Significance of infectious agents in colorectal cancer development. Journal of Cancer 4, 227-240.

Armaghany, T., Wilson, J.D., Chu, Q., and Mills, G. (2012). Genetic alterations in colorectal cancer. Gastrointestinal cancer research: GCR 5, 19-27.

Arnold, M., Sierra, M.S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017). Global patterns and trends in colorectal cancer incidence and mortality. Gut 66, 683-691

Arora, Anjali, and Kumaravel Somasundaram. "Targeted Proteomics Comes to the Benchside and the Bedside: Is it Ready for Us?." BioEssays 41.2 (2019): 1800042.

Babel, I., Barderas, R., Diaz-Uriarte, R., Moreno, V., Suarez, A., Fernandez-Acenero, M.J., Salazar, R., Capella, G., and Casal, J.I. (2011). Identification of MST1/STK4 and SULF1 proteins as autoantibody targets for the diagnosis of colorectal cancer by using phage microarrays. Mol Cell Proteomics 10, M110 001784.

Berrocal, T., Lamas, M., Gutieerrez, J., Torres, I., Prieto, C., and del Hoyo, M.L. (1999). Congenital anomalies of the small intestine, colon, and rectum. Radiographics: a review publication of the Radiological Society of North America, Inc 19, 1219-1236.

Bettegowda, C., Sausen, M., Leary, R.J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B.R., Wang, H., Luber, B., Alani, R.M., et al. (2014). Detection of circulating tumour DNA in early- and late-stage human malignancies. Science translational medicine 6, 224ra224.

Bigler, J., Whitton, J., Lampe, J.W., Fosdick, L., Bostick, R.M., and Potter, J.D. (2001). CYP2C9 and UGT1A6 genotypes modulate the protective effect of aspirin on colon adenoma risk. Cancer Res 61, 3566-3569.

Bingham, S.A., Hughes, R., and Cross, A.J. (2002). Effect of white versus red meat on endogenous N-nitrosation in the human colon and further evidence of a dose response. J Nutr 132, 3522s-3525s.

Blausen.comstaff (2014). Medical gallery of Blausen Medical 2014. WikiJournal of Medicine 1 (2): 10 1, 10.

Bogaert, J., and Prenen, H. (2014). Molecular genetics of colorectal cancer. Ann Gastroenterol 27, 9-14.

Bostick, R.M., Fosdick, L., Wood, J.R., Grambsch, P., Grandits, G.A., Lillemoe, T.J., Louis, T.A., and Potter, J.D. (1995). Calcium and colorectal epithelial cell proliferation in sporadic adenoma patients: a randomized, double-blinded, placebo-controlled clinical trial. J Natl Cancer Inst 87, 1307-1315.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 68, 394-424.

Bujanda, L., Sarasqueta, C., Cosme, A., Hijona, E., Enriquez-Navascues, J.M., Placer, C., Villarreal, E., Herreros-Villanueva, M., Giraldez, M.D., Gironella, M., et al. (2013). Evaluation of alpha 1-antitrypsin and the levels of mRNA expression of matrix metalloproteinase 7, urokinase type plasminogen activator receptor and COX-2 for the diagnosis of colorectal cancer. PLoS One 8, e51810.

Cantor, D.I., Cheruku, H.R., Nice, E.C., and Baker, M.S. (2015). Integrin alphavbeta6 sets the stage for colorectal cancer metastasis. Cancer Metastasis Rev 34, 715-734.

Catherman, A.D., Skinner, O.S., and Kelleher, N.L. (2014). Top Down proteomics: facts and perspectives. Biochemical and biophysical research communications 445, 683-693.

Charanjeet Singh, M.D. (2017). Staging of colonic carcinoma (AJCC 7th Edition).

Chen, J.S., Kuo, Y.B., Chou, Y.P., Chan, C.C., Fan, C.W., Chen, K.T., Huang, Y.S., and Chan, E.C. (2011). Detection of autoantibodies against Rabphilin-3A-like protein as a potential biomarker in patient's sera of colorectal cancer. Clin Chim Acta 412, 1417-1422.

Cheng, Y., Xie, G., Chen, T., Qiu, Y., Zou, X., Zheng, M., Tan, B., Feng, B., Dong, T., He, P., et al. (2012). Distinct urinary metabolic profile of human colorectal cancer. Journal of proteome research 11, 1354-1363.

Cohen, J.D., Javed, A.A., Thoburn, C., Wong, F., Tie, J., Gibbs, P., Schmidt, C.M., Yip-Schneider, M.T., Allen, P.J., Schattner, M., et al. (2017). Combined circulating tumour DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. Proceedings of the National Academy of Sciences of the United States of America 114, 10202-10207.

Cohen, J.D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A.A., Wong, F., Mattox, A., et al. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science (New York, NY) 359, 926-930.

Compton, P.D., Zamdborg, L., Thomas, P.M., and Kelleher, N.L. (2011). On the scalability and requirements of whole protein mass spectrometry. Analytical chemistry 83, 6868-6874.

Cross, A.J., Pollock, J.R., and Bingham, S.A. (2003). Haem, not protein or inorganic iron, is responsible for endogenous intestinal N-nitrosation arising from red meat. Cancer Res 63, 2358-2360.

Davenport, J.R., Cai, Q., Ness, R.M., Milne, G., Zhao, Z., Smalley, W.E., Zheng, W., and Shrubsole, M.J. (2016). Evaluation of pro-inflammatory markers plasma C-reactive protein and

urinary prostaglandin-E2 metabolite in colorectal adenoma risk. Molecular carcinogenesis 55, 1251-1261.

De Chiara, L., Rodriguez-Pineiro, A.M., Rodriguez-Berrocal, F.J., Cordero, O.J., Martinez-Ares, D., and Paez de la Cadena, M. (2010). Serum CD26 is related to histopathological polyp traits and behaves as a marker for colorectal cancer and advanced adenomas. BMC Cancer 10, 333.

de la Chapelle, A. (2004). Genetic predisposition to colorectal cancer. Nat Rev Cancer 4, 769-780.

Diamandis, E.P., and Fiala, C. (2017). Can circulating tumour DNA be used for direct and early stage cancer detection? F1000Research 6, 2129.

Domon, B., and Gallien, S. (2015). Recent advances in targeted proteomics for clinical applications. Proteomics Clinical applications 9, 423-431.

Donna M. Gress, S.B.E., Frederick L. Greene, Mary Kay Washington, Elliot A.Asare, James D.Brierly, David R. Byrd, Carolyn C.Compton, J. Milburn Jessup, David P.Winchester, Mahul B.Amin and Jeffrey E. Gershenwald (2017). Prinicples of Cancer Staging, Eighth edn.

Dukes, C.E. (1932). The classification of cancer of the rectum. Journal of Pathological Bacteriology 35, 323.

Eliuk, S., and Makarov, A. (2015). Evolution of Orbitrap Mass Spectrometry Instrumentation. Annual review of analytical chemistry (Palo Alto, Calif) 8, 61-80.

Elsafi, S.H., Alqahtani, N.I., Zakary, N.Y., and Al Zahrani, E.M. (2015). The sensitivity, specificity, predictive values, and likelihood ratios of faecal occult blood test for the detection of colorectal cancer in hospital settings. Clinical and experimental gastroenterology 8, 279-284.

Emery, J.D. (2015). The challenges of early diagnosis of cancer in general practice. The Medical journal of Australia 203, 391-393.

Engvall, E., and Perlmann, P. (1971). Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G. Immunochemistry 8, 871-874.

Fakih, M.G., and Padmanabhan, A. (2006). CEA monitoring in colorectal cancer. What you should know. Oncology (Williston Park, NY) 20, 579-587; discussion 588, 594, 596 passim.

Fan, G., Zhang, K., Yang, X., Ding, J., Wang, Z., and Li, J. (2017). Prognostic value of circulating tumour DNA in patients with colon cancer: Systematic review. PloS one 12, e0171991.

Fentz, A.K., Sporl, M., Spangenberg, J., List, H.J., Zornig, C., Dorner, A., Layer, P., Juhl, H., and David, K.A. (2007). Detection of colorectal adenoma and cancer based on transthyretin and C3a-desArg serum levels. Proteomics Clin Appl 1, 536-544.

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D.M., Pineros, M., Znaor, A., and Bray, F. (2018). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. International journal of cancer.

Ferlay J, E.M., Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2018). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. .

Fiala, O., Finek, J., Buchler, T., Matejka, V.M., Holubec, L., Kulhankova, J., Bortlicek, Z., Liska, V., and Topolcan, O. (2015). The Association of Serum Carcinoembryonic Antigen, Carbohydrate Antigen 19-9, Thymidine Kinase, and Tissue Polypeptide Specific Antigen with Outcomes of Patients with Metastatic Colorectal Cancer Treated with Bevacizumab: a Retrospective Study. Targeted oncology 10, 549-555.

Fijten, G.H., Starmans, R., Muris, J.W., Schouten, H.J., Blijham, G.H., and Knottnerus, J.A. (1995). Predictive value of signs and symptoms for colorectal cancer in patients with rectal bleeding in general practice. Family practice 12, 279-286.

Fireman, Z., Trost, L., Kopelman, Y., Segal, A., and Sternberg, A. (2000). Helicobacter pylori: seroprevalence and colorectal cancer. Isr Med Assoc J 2, 6-9.

Flavahan, W.A., Gaskell, E., and Bernstein, B.E. (2017). Epigenetic plasticity and the hallmarks of cancer. Science (New York, NY) 357.

Fortin, T., Salvador, A., Charrier, J.P., Lenz, C., Lacoux, X., Morla, A., Choquet-Kastylevsky, G., and Lemoine, J. (2009). Clinical quantitation of prostate-specific antigen biomarker in the low nanogram/milliliter range by conventional bore liquid chromatography-tandem mass spectrometry (multiple reaction monitoring) coupling and correlation with ELISA tests. Molecular & cellular proteomics: MCP 8, 1006-1015.

Fung, K.Y., Nice, E., Priebe, I., Belobrajdic, D., Phatak, A., Purins, L., Tabor, B., Pompeia, C., Lockett, T., Adams, T.E., et al. (2014). Colorectal cancer biomarkers: to be or not to be? Cautionary tales from a road well-travelled. World J Gastroenterol 20, 888-898.

Fung, K.Y., Tabor, B., Buckley, M.J., Priebe, I.K., Purins, L., Pompeia, C., Brierley, G.V., Lockett, T., Gibbs, P., Tie, J., et al. (2015). Blood-based protein biomarker panel for the detection of colorectal cancer. PLoS One 10, e0120425.

Gallien, S., Duriez, E., Crone, C., Kellmann, M., Moehring, T., and Domon, B. (2012). Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. Molecular & cellular proteomics: MCP 11, 1709-1723.

Gallien, S., Duriez, E., Demeure, K., and Domon, B. (2013). Selectivity of LC-MS/MS analysis: implication for proteomics experiments. Journal of proteomics 81, 148-158.

Ganepola, G.A., Nizin, J., Rutledge, J.R., and Chang, D.H. (2014). Use of blood-based biomarkers for early diagnosis and surveillance of colorectal cancer. World journal of gastrointestinal oncology 6, 83-97.

Geyer, P.E., Holdt, L.M., Teupser, D., and Mann, M. (2017). Revisiting biomarker discovery by plasma proteomics. Molecular systems biology 13, 942.

Gillet, L.C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Molecular & cellular proteomics: MCP 11, O111.016717.

Gillette, M.A., and Carr, S.A. (2013). Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. Nature methods 10, 28-34.

Gonzalez-Galarza, Faviel F., et al. "A critical appraisal of techniques, software packages, and standards for quantitative proteomic analysis." Omics: a journal of integrative biology 16.9 (2012): 431-442.

Gopalakrishnan, V., Helmink, B.A., Spencer, C.N., Reuben, A., and Wargo, J.A. (2018). The Influence of the Gut Microbiome on Cancer, Immunity, and Cancer Immunotherapy. Cancer Cell 33, 570-580.

Gorgannezhad, L., Umer, M., Islam, M.N., Nguyen, N.T., and Shiddiky, M.J.A. (2018). Circulating tumour DNA and liquid biopsy: opportunities, challenges, and recent advances in detection technologies. Lab on a chip 18, 1174-1196.

Graves, P.R., and Haystead, T.A. (2002). Molecular biologist's guide to proteomics. Microbiology and molecular biology reviews: MMBR 66, 39-63; table of contents.

Group, F.-N.B.W. (2016). In BEST (Biomarkers, EndpointS, and other Tools) Resource (Silver Spring (MD): Food and Drug Administration (US)).

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nature biotechnology 17, 994-999.

Halford, M.M., Macheda, M.L., Parish, C.L., Takano, E.A., Fox, S., Layton, D., Nice, E., and Stacker, S.A. (2013). A fully human inhibitory monoclonal antibody to the Wnt receptor RYK. PLoS One 8, e75447.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell 100, 57-70.

Harlan, R., and Zhang, H. (2014). Targeted proteomics: a bridge between discovery and validation. Expert review of proteomics 11, 657-661.

Heitzer, E., Perakis, S., Geigl, J.B., and Speicher, M.R. (2017). The potential of liquid biopsies for the early detection of cancer. NPJ precision oncology 1, 36.

Holten-Andersen, M.N., Christensen, I.J., Nielsen, H.J., Stephens, R.W., Jensen, V., Nielsen, O.H., Sorensen, S., Overgaard, J., Lilja, H., Harris, A., et al. (2002). Total levels of tissue inhibitor of metalloproteinases 1 in plasma yield high diagnostic sensitivity and specificity in patients with colon cancer. Clin Cancer Res 8, 156-164.

Hsu, W.Y., Chen, W.T., Lin, W.D., Tsai, F.J., Tsai, Y., Lin, C.T., Lo, W.Y., Jeng, L.B., and Lai, C.C. (2009). Analysis of urinary nucleosides as potential tumour markers in human colorectal cancer by high performance liquid chromatography/electrospray ionization tandem mass spectrometry. Clinica chimica acta; international journal of clinical chemistry 402, 31-37.

Hyman, J., Baron, J.A., Dain, B.J., Sandler, R.S., Haile, R.W., Mandel, J.S., Mott, L.A., and Greenberg, E.R. (1998). Dietary and supplemental calcium and the recurrence of colorectal adenomas. Cancer Epidemiol Biomarkers Prev 7, 291-295.

Imperiale, T.F., Ransohoff, D.F., and Itzkowitz, S.H. (2014). Multitarget stool DNA testing for colorectal-cancer screening. The New England journal of medicine 371, 187-188.

Jasperson, K.W., Tuohy, T.M., Neklason, D.W., and Burt, R.W. (2010). Hereditary and familial colon cancer. Gastroenterology 138, 2044-2058.

Jedi, M., Young, G.P., Pedersen, S.K., and Symonds, E.L. (2018). Methylation and Gene Expression of BCAT1 and IKZF1 in Colorectal Cancer Tissues. Clin Med Insights Oncol 12, 1179554918775064.

Jones, J.J., Wilcox, B.E., Benz, R.W., Babbar, N., Boragine, G., Burrell, T., Christie, E.B., Croner, L.J., Cun, P., Dillon, R., et al. (2016). A Plasma-Based Protein Marker Panel for Colorectal Cancer Detection Identified by Multiplex Targeted Mass Spectrometry. Clinical colorectal cancer 15, 186-194.e113.

Jung, A. (2013). Abdominal Imaging (Springer).

Kakiuchi, H., Watanabe, M., Ushijima, T., Toyota, M., Imai, K., Weisburger, J.H., Sugimura, T., and Nagao, M. (1995). Specific 5'-GGGA-3'-->5'-GGA-3' mutation of the Apc gene in rat

colon tumours induced by 2-amino-1-methyl-6-phenylimidazo[4,5-b] pyridine. Proceedings of the National Academy of Sciences of the United States of America 92, 910-914.

Katballe, N., Christensen, M., Wikman, F.P., Orntoft, T.F., and Laurberg, S. (2002). Frequency of hereditary non-polyposis colorectal cancer in Danish colorectal cancer patients. Gut 50, 43-51.

Kawamori, T., Rao, C.V., Seibert, K., and Reddy, B.S. (1998). Chemopreventive activity of celecoxib, a specific cyclooxygenase-2 inhibitor, against colon carcinogenesis. Cancer Res 58, 409-412.

Kim, H.J., Kang, H.J., Lee, H., Lee, S.T., Yu, M.H., Kim, H., and Lee, C. (2009). Identification of S100A8 and S100A9 as serological markers for colorectal cancer. J Proteome Res 8, 1368-1379.

Kim, H.J., Lin, D., Lee, H.J., Li, M., and Liebler, D.C. (2016). Quantitative Profiling of Protein Tyrosine Kinases in Human Cancer Cell Lines by Multiplexed Parallel Reaction Monitoring Assays. Molecular & cellular proteomics: MCP 15, 682-691.

Kolligs, F.T. (2016). Diagnostics and Epidemiology of Colorectal Cancer. Visc Med 32, 158-164.

Kruger, M., Moser, M., Ussar, S., Thievessen, I., Luber, C.A., Forner, F., Schmidt, S., Zanivan, S., Fassler, R., and Mann, M. (2008). SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. Cell 134, 353-364.

Kune, G.A., Kune, S., and Watson, L.F. (1988). Colorectal cancer risk, chronic illnesses, operations, and medications: case control results from the Melbourne Colorectal Cancer Study. Cancer Res 48, 4399-4404.

Lansdorp-Vogelaar, I., Knudsen, A.B., and Brenner, H. (2011). Cost-effectiveness of colorectal cancer screening. Epidemiologic reviews 33, 88-100.

Larsson, S.C., Orsini, N., and Wolk, A. (2005). Diabetes mellitus and risk of colorectal cancer: a meta-analysis. J Natl Cancer Inst 97, 1679-1687.

Law, Kai Pong, and Yoon Pin Lim. "Recent advances in mass spectrometry: data independent analysis and hyper reaction monitoring." Expert review of proteomics 10.6 (2013): 551-566.

Lech, G., Slotwinski, R., Slodkowski, M., and Krasnodebski, I.W. (2016). Colorectal cancer tumour markers and biomarkers: Recent therapeutic advances. World J Gastroenterol 22, 1745-1755.

Lee, J.K., Liles, E.G., Bent, S., Levin, T.R., and Corley, D.A. (2014). Accuracy of faecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. Annals of internal medicine 160, 171.

Lin, S., and Garcia, B.A. (2012). Examining histone posttranslational modification patterns by high-resolution mass spectrometry. Methods in enzymology 512, 3-28.

Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B.C., and Aebersold, R. (2018). Data-independent acquisition-based SWATH™-MS for quantitative proteomics: a tutorial. Molecular systems biology 14, e8126.

Mahboob, S., Ahn, S.B., Cheruku, H.R., Cantor, D., Rennel, E., Fredriksson, S., Edfeldt, G., Breen, E.J., Khan, A., Mohamedali, A., et al. (2015). A novel multiplexed immunoassay identifies CEA, IL-8 and prolactin as prospective markers for Dukes' stages A-D colorectal cancers. Clinical proteomics 12, 10.

Mahmood, T., and Yang, P.C. (2012). Western blot: technique, theory, and trouble shooting. N Am J Med Sci 4, 429-434.

Marmol, I., Sanchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E., and Rodriguez Yoldi, M.J. (2017). Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. International journal of molecular sciences 18.

Marshall, K.W., Mohr, S., Khettabi, F.E., Nossova, N., Chao, S., Bao, W., Ma, J., Li, X.J., and Liew, C.C. (2010). A blood-based biomarker panel for stratifying current risk for colorectal cancer. International journal of cancer 126, 1177-1186.

Mayeux, R. (2004). Biomarkers: potential uses and limitations. NeuroRx: the journal of the American Society for Experimental NeuroTherapeutics 1, 182-188.

Meacham, C.E., and Morrison, S.J. (2013). Tumour heterogeneity and cancer cell plasticity. Nature 501, 328-337.

Meier, F., Geyer, P.E., Virreira Winter, S., Cox, J., and Mann, M. (2018). BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. Nature methods 15, 440-448.

Meucci, G., Tatarella, M., Vecchi, M., Ranzi, M.L., Biguzzi, E., Beccari, G., Clerici, E., and de Franchis, R. (1997). High prevalence of Helicobacter pylori infection in patients with colonic adenomas and carcinomas. J Clin Gastroenterol 25, 605-607.

Mitchell, P.S., Parkin, R.K., Kroh, E.M., Fritz, B.R., Wyman, S.K., Pogosova-Agadjanyan, E.L., Peterson, A., Noteboom, J., O'Briant, K.C., Allen, A., et al. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. Proceedings of the National Academy of Sciences of the United States of America 105, 10513-10518.

Moorghen, M., Ince, P., Finney, K.J., Sunter, J.P., Appleton, D.R., and Watson, A.J. (1988). A protective effect of sulindac against chemically induced primary colonic tumours in mice. J Pathol 156, 341-347.

Mroczko, B., Groblewska, M., Okulczyk, B., Kedra, B., and Szmitkowski, M. (2010). The diagnostic value of matrix metalloproteinase 9 (MMP-9) and tissue inhibitor of matrix metalloproteinases 1 (TIMP-1) determination in the sera of colorectal adenoma and cancer patients. Int J Colorectal Dis 25, 1177-1184.

Nguyen, M.T., and Weinberg, D.S. (2016). Biomarkers in Colorectal Cancer Screening. J Natl Compr Canc Netw 14, 1033-1040.

Nielsen, H.J., Brunner, N., Frederiksen, C., Lomholt, A.F., King, D., Jorgensen, L.N., Olsen, J., Rahr, H.B., Thygesen, K., Hoyer, U., et al. (2008). Plasma tissue inhibitor of metalloproteinases-1 (TIMP-1): a novel biological marker in the detection of primary colorectal cancer. Protocol outlines of the Danish-Australian endoscopy study group on colorectal cancer detection. Scand J Gastroenterol 43, 242-248.

Ong, S.E., and Mann, M. (2005). Mass spectrometry-based proteomics turns quantitative. Nature chemical biology 1, 252-262.

Pedersen, J.W., Blixt, O., Bennett, E.P., Tarp, M.A., Dar, I., Mandel, U., Poulsen, S.S., Pedersen, A.E., Rasmussen, S., Jess, P., et al. (2011). Seromic profiling of colorectal cancer patients with novel glycopeptide microarray. Int J Cancer 128, 1860-1871.

Peterse, E.F.P., Meester, R.G.S., Siegel, R.L., Chen, J.C., Dwyer, A., Ahnen, D.J., Smith, R.A., Zauber, A.G., and Lansdorp-Vogelaar, I. (2018). The impact of the rising colorectal cancer incidence in young adults on the optimal age to start screening: Microsimulation analysis I to inform the American Cancer Society colorectal cancer screening guideline. Cancer 124, 2964-2973.

Peterson, A.C., Russell, J.D., Bailey, D.J., Westphall, M.S., and Coon, J.J. (2012). Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. Molecular & cellular proteomics: MCP 11, 1475-1488.

Pickup, M., Novitskiy, S., and Moses, H.L. (2013). The roles of TGFbeta in the tumour microenvironment. Nat Rev Cancer 13, 788-799.

Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B., and Aebersold, R. (2009). Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. Cell 138, 795-806.

Pollard, M., and Luckert, P.H. (1980). Indomethacin treatment of rats with dimethylhydrazine-induced intestinal tumours. Cancer Treat Rep 64, 1323-1327.

Potter, J.D., Lindor, Noralane M. (Eds.) (2009). Colorectal Cancer: Epidemiology. In Genetics of Colorectal Cancer, J.D. Potter, Lindor, Noralane M. (Eds.), ed. (Springer).

Prakash, A., Tomazela, D.M., Frewen, B., Maclean, B., Merrihew, G., Peterman, S., and Maccoss, M.J. (2009). Expediting the development of targeted SRM assays: using data from

shotgun proteomics to automate method development. Journal of proteome research 8, 2733-2739.

Rasmussen, L., Ladelund, S., Brunner, N., Jorgensen, L.N., Nielsen, H.J., and Sorensen, L.T. (2013). Tissue inhibitor of metalloproteinases-1 as a biological marker in colorectal cancer: influence of smoking on plasma levels? The International journal of biological markers 28, 226-230.

Rauniyar, N. (2015). Parallel Reaction Monitoring: A Targeted Experiment Performed Using High Resolution and High Mass Accuracy Mass Spectrometry. International journal of molecular sciences 16, 28566-28581.

Reddy, B.S., Maruyama, H., and Kelloff, G. (1987). Dose-related inhibition of colon carcinogenesis by dietary piroxicam, a nonsteroidal antiinflammatory drug, during different stages of rat colon tumour development. Cancer Res 47, 5340-5346.

Rex, D.K., Boland, C.R., Dominitz, J.A., Giardiello, F.M., Johnson, D.A., Kaltenbach, T., Levin, T.R., Lieberman, D., and Robertson, D.J. (2017). Colorectal Cancer Screening: Recommendations for Physicians and Patients from the U.S. Multi-Society Task Force on Colorectal Cancer. The American journal of gastroenterology 112, 1016-1030.

Reya, T., Morrison, S.J., Clarke, M.F., and Weissman, I.L. (2001). Stem cells, cancer, and cancer stem cells. Nature 414, 105-111.

Richter, S. (2008). Faecal DNA screening in colorectal cancer. Can J Gastroenterol 22, 631-633.

Rosenberg, L., Palmer, J.R., Zauber, A.G., Warshauer, M.E., Stolley, P.D., and Shapiro, S. (1991). A hypothesis: nonsteroidal anti-inflammatory drugs reduce the incidence of large-bowel cancer. J Natl Cancer Inst 83, 355-358.

Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., et al. (2004). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Molecular & cellular proteomics: MCP 3, 1154-1169.

Rowan, A.J., Lamlum, H., Ilyas, M., Wheeler, J., Straub, J., Papadopoulou, A., Bicknell, D., Bodmer, W.F., and Tomlinson, I.P. (2000). APC mutations in sporadic colorectal tumours: A mutational "hotspot" and interdependence of the "two hits". Proceedings of the National Academy of Sciences of the United States of America 97, 3352-3357.

Ruder, E.H., Laiyemo, A.O., Graubard, B.I., Hollenbeck, A.R., Schatzkin, A., and Cross, A.J. (2011). Non-steroidal anti-inflammatory drugs and colorectal cancer risk in a large, prospective cohort. The American journal of gastroenterology 106, 1340-1350.

Schee, K., Boye, K., Abrahamsen, T.W., Fodstad, O., and Flatmark, K. (2012). Clinical relevance of microRNA miR-21, miR-31, miR-92a, miR-101, miR-106a and miR-145 in colorectal cancer. BMC cancer 12, 505.

Sethi, M.K., Hancock, W.S., and Fanayan, S. (2016). Identifying N-Glycan Biomarkers in Colorectal Cancer by Mass Spectrometry. Acc Chem Res 49, 2099-2106.

Seyfried, N.T., Gozal, Y.M., Dammer, E.B., Xia, Q., Duong, D.M., Cheng, D., Lah, J.J., Levey, A.I., and Peng, J. (2010). Multiplex SILAC analysis of a cellular TDP-43 proteinopathy model reveals protein inclusions associated with SUMOylation and diverse polyubiquitin chains. Molecular & cellular proteomics : MCP 9, 705-718.

Shah, R., Jones, E., Vidart, V., Kuppen, P.J., Conti, J.A., and Francis, N.K. (2014). Biomarkers for early detection of colorectal cancer and polyps: systematic review. Cancer Epidemiol Biomarkers Prev 23, 1712-1728.

Slatter, M.L., Yakumo, K., Hoffman, M., and Neuhausen, S. (2001). Variants of the VDR gene and risk of colon cancer (United States). Cancer Causes Control 12, 359-364.

Smith, L.M., and Kelleher, N.L. (2013). Proteoform: a single term describing protein complexity. Nature methods 10, 186-187.

Sole, X., Crous-Bou, M., Cordero, D., Olivares, D., Guino, E., Sanz-Pamplona, R., Rodriguez-Moranta, F., Sanjuan, X., de Oca, J., Salazar, R., et al. (2014). Discovery and validation of new potential biomarkers for early detection of colon cancer. PLoS One 9, e106748.

Song, L., Jia, J., Peng, X., Xiao, W., and Li, Y. (2017). The performance of the SEPT9 gene methylation assay and a comparison with other CRC screening tests: A meta-analysis. Scientific reports 7, 3032.

Song, L.L., and Li, Y.M. (2016). Current noninvasive tests for colorectal cancer screening: An overview of colorectal cancer screening tests. World journal of gastrointestinal oncology 8, 793-800.

Stephen B. Edge , C.C.C. (2010) (Verlag: Springer).

Stoffel, E.M., and Kastrinos, F. (2014). Familial colorectal cancer, beyond Lynch syndrome. Clin Gastroenterol Hepatol 12, 1059-1068.

Su, Y.H., Wang, M., Brenner, D.E., Norton, P.A., and Block, T.M. (2008). Detection of mutated K-ras DNA in urine, plasma, and serum of patients with colorectal carcinoma or adenomatous polyps. Annals of the New York Academy of Sciences 1137, 197-206.

Sugimura, T., and Sato, S. (1983). Mutagens-carcinogens in foods. Cancer Res 43, 2415s-2421s.

Suh, O., Mettlin, C., and Petrelli, N.J. (1993). Aspirin use, cancer, and polyps of the large bowel. Cancer 72, 1171-1177.

Surinova, S., Choi, M., Tao, S., Schuffler, P.J., Chang, C.Y., Clough, T., Vyslouzil, K., Khoylou, M., Srovnal, J., Liu, Y., et al. (2015a). Prediction of colorectal cancer diagnosis based on circulating plasma proteins. EMBO molecular medicine 7, 1166-1178.

Surinova, S., Radova, L., Choi, M., Srovnal, J., Brenner, H., Vitek, O., Hajduch, M., and Aebersold, R. (2015b). Non-invasive prognostic protein biomarker signatures associated with colorectal cancer. EMBO molecular medicine 7, 1153-1165.

Symonds, E.L., and Young, G.P. (2015). Blood Tests for Colorectal Cancer Screening in the Standard Risk Population. Current colorectal cancer reports 11, 397-407.

Tan, S.H., Mohamedali, A., Kapur, A., and Baker, M.S. (2013). Ultradepletion of human plasma using chicken antibodies: a proof of concept study. Journal of proteome research 12, 2399-2413.

Tan, S.H., Mohamedali, A., Kapur, A., Lukjanenko, L., and Baker, M.S. (2012). A novel, cost-effective and efficient chicken egg IgY purification procedure. Journal of immunological methods 380, 73-76.

Tanaka, T., Tanaka, M., Tanaka, T., and Ishigamori, R. (2010). Biomarkers for colorectal cancer. Int J Mol Sci 11, 3209-3225.

Tao, S., Haug, U., Kuhn, K., and Brenner, H. (2012). Comparison and combination of blood-based inflammatory markers with faecal occult blood tests for non-invasive colorectal cancer screening. Br J Cancer 106, 1424-1430.

Tipton, J.D., Tran, J.C., Catherman, A.D., Ahlf, D.R., Durbin, K.R., and Kelleher, N.L. (2011). Analysis of intact protein isoforms by mass spectrometry. The Journal of biological chemistry 286, 25451-25458.

Toby, T.K., Fornelli, L., and Kelleher, N.L. (2016). Progress in Top-Down Proteomics and the Analysis of Proteoforms. Annual review of analytical chemistry (Palo Alto, Calif) 9, 499-519.

Todoroki, I., Friedman, G.D., Slattery, M.L., Potter, J.D., and Samowitz, W. (1999). Cholecystectomy and the risk of colon cancer. The American journal of gastroenterology 94, 41-46.

Tu, C., Rudnick, P.A., Martinez, M.Y., Cheek, K.L., Stein, S.E., Slebos, R.J., and Liebler, D.C. (2010). Depletion of abundant plasma proteins and limitations of plasma proteomics. Journal of proteome research 9, 4982-4991.

Umemori, Y., Ohe, Y., Kuribayashi, K., Tsuji, N., Nishidate, T., Kameshima, H., Hirata, K., and Watanabe, N. (2010). Evaluating the utility of N1,N12-diacetylspermine and N1,N8-

diacetylspermidine in urine as tumour markers for breast and colorectal cancers. Clinica chimica acta; international journal of clinical chemistry 411, 1894-1899.

Van Riper, S.K., de Jong, E.P., Carlis, J.V., and Griffin, T.J. (2013). Mass spectrometry-based proteomics: basic principles and emerging technologies and directions. Advances in experimental medicine and biology 990, 1-35.

Vesely, M.D., Kershaw, M.H., Schreiber, R.D., and Smyth, M.J. (2011). Natural innate and adaptive immunity to cancer. Annu Rev Immunol 29, 235-271.

Vijeta Pamudurthy, M.B.a.V.K. (2016). Biomarkers in Colorectal Cancer Screening. J Gastrointest Dig Syst 6, 389.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science (New York, NY) 339, 1546-1558.

Volik, S., Alcaide, M., Morin, R.D., and Collins, C. (2016). Cell-free DNA (cfDNA): Clinical Significance and Utility in Cancer Shaped By Emerging Technologies. Molecular cancer research : MCR 14, 898-908.

Vukobrat-Bijedic, Z., Husic-Selimovic, A., Sofic, A., Bijedic, N., Bjelogrlic, I., Gogov, B., and Mehmedovic, A. (2013). Cancer Antigens (CEA and CA 19-9) as Markers of Advanced Stage of Colorectal Carcinoma. Medical archives (Sarajevo, Bosnia and Herzegovina) 67, 397-401.

Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., Jia, W., Cai, S., and Zhao, L. (2012). Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. Isme j 6, 320-329.

Waszkiewicz, N., Zalewska-Szajda, B., Szajda, S.D., Kepka, A., Waszkiewicz, M., Roszkowska-Jakimiec, W., Wojewodzka-Zelezniakowicz, M., Milewska, A.J., Dadan, J., Szulc, A., et al. (2012). Lysosomal exoglycosidases and cathepsin D in colon adenocarcinoma. Polskie Archiwum Medycyny Wewnetrznej 122, 551-556.

Watany, M.M., Elmashad, N.M., Badawi, R., and Hawash, N. (2018). Serum FBLN1 and STK31 as biomarkers of colorectal cancer and their ability to noninvasively differentiate colorectal cancer from benign polyps. Clin Chim Acta 483, 151-155.

Werner, T., Sweetman, G., Savitski, M.F., Mathieson, T., Bantscheff, M., and Savitski, M.M. (2014). Ion coalescence of neutron encoded TMT 10-plex reporter ions. Analytical chemistry 86, 3594-3601.

WHO (2018). Early detection of cancer.

Wild, N., Andres, H., Rollinger, W., Krause, F., Dilba, P., Tacke, M., and Karl, J. (2010). A combination of serum markers for the early detection of colorectal cancer. Clin Cancer Res 16, 6111-6121.

Wilkins, M.R., Sanchez, J.C., Gooley, A.A., Appel, R.D., Humphery-Smith, I., Hochstrasser, D.F., and Williams, K.L. (1996). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnology & genetic engineering reviews 13, 19-50.

Wolf, A.M.D., Fontham, E.T.H., Church, T.R., Flowers, C.R., Guerra, C.E., LaMonte, S.J., Etzioni, R., McKenna, M.T., Oeffinger, K.C., Shih, Y.T., et al. (2018). Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. CA: a cancer journal for clinicians 68, 250-281.

Wolters, D.A., Washburn, M.P., and Yates, J.R., 3rd (2001). An automated multidimensional protein identification technology for shotgun proteomics. Analytical chemistry 73, 5683-5690.

WRCF (1987). Food, Nutrition and the Prevention of Cancer : A Global Perpective (Washington, DC: American Institute of Cancer Research).

Yanqing, H., Cheng, D., and Ling, X. (2018). Serum CA72-4 as a Biomarker in the Diagnosis of Colorectal Cancer: A Meta-analysis. Open medicine (Warsaw, Poland) 13, 164-171.

Zhang, C., Liu, Y., and Andrews, P.C. (2013). Quantification of histone modifications using (1)(5)N metabolic labeling. Methods (San Diego, Calif) 61, 236-243.

Zhong, L., Liu, J., Hu, Y., Wang, W., Xu, F., Xu, W., Han, J., and Biskup, E. (2017). STK31 as novel biomarker of metastatic potential and tumourigenicity of colorectal cancer. Oncotarget 8, 24354-24361.

Zhu, W., Smith, J.W., and Huang, C.M. (2010). Mass spectrometry-based label-free quantitative proteomics. Journal of biomedicine & biotechnology 2010, 840518.

Thesis Experimental Aims

Aim I: To identify potential early clinical stage colorectal cancer diagnosis using a multi-variate test using MS-based technologies

This study aimed to adopt state-of-the-art proteomic technologies to discover protein biomarkers for the detection of CRC patients at earlier stages (I/II) from 100 EDTA plasma samples. The aim was to visualise and quantify novel lower-abundance proteins, using combinations of commercially available depletion and an in-house ultradepletion systems. The identified candidates were further verified using orthogonal technologies including western blotting, ELISA and machine learning predictive model on synthetic patient data.

Aim II: Verification of a multi-analyte signature assay for early diagnosis using Parallel Reaction Monitoring (PRM) assay

The study then aimed to verify the biomarker candidates identified in plasma discovery studies by targeted peptide measurements to facilitate the development of a robust PRM assay.

Aim III: Development and verification of parallel reaction monitoring assays for CRC epithelial-mesenchymal transition markers uPAR and integrin αvβ6

This study sought to develop a proof-of-concept PRM assay to interrogate plasma samples for expression of two cancer epithelial to mesenchymal transition markers (uPAR and αvβ6) in cell lines and recombinant proteins. This study is ongoing and requires optimisation to verify and measure uPAR and αvβ6 peptide fragments in CRC plasma samples.

Chapter 3

Potential early clinical stage colorectal cancer diagnosis using a proteomics blood test

Abstract

Background: One of the most significant challenges in colorectal cancer (CRC) management is the use of compliant early-stage population-based diagnostic tests as adjuncts to a confirmatory colonoscopy. Despite the near curative nature of early-stage surgical resection, mortality remains unacceptably high; as most patients are diagnosed by faecal haemoglobin followed by colonoscopy occur at latter stages. Additionally, current population-based screens reliant on faecal occult blood tests (FOBT) have low compliance (~40%) and tests suffer low sensitivities. Therefore, blood-based diagnostic tests offer survival benefits from their higher compliance (≥97%), especially if they can match or surpass the sensitivity and specificity of FOBTs. However, discovery of low abundance plasma biomarkers is difficult due to occupancy of a high percentage of proteomic discovery space by relatively few high-abundance proteins.

Methods: A combination of high abundance protein ultradepletion (e.g., MARS-14 and an in-house IgY depletion) strategies, extensive peptide fractionation methods (SCX, SAX, High pH and SEC) and SWATH™-MS were utilised to uncover protein biomarkers from a cohort of 100 plasma samples (i.e., pools of 20 healthy and 20 stages I-IV CRC plasmas). The differentially expressed proteins were analysed using ANOVA and pairwise t-tests ($p<0.05$; fold-change>1.5), and further examined with a neural network classification method using augmented 5,000 patient datasets, *in silico*.

Results: Ultradepletion combined with peptide fractionation allowed for the identification of a total of 513 plasma proteins, 8 of which had not been previously reported in human plasma. SWATH™-MS analysis revealed 37 protein biomarker candidates that exhibited differential expression across CRC stages compared to healthy controls. Of those, 7 candidates (CST3, GPX3, CFD, MRC1, COMP, PON1 and ADAMDEC1) were validated using Western blotting and/or ELISA. The neural network classification narrowed down candidate biomarkers to 5 proteins (SAA2, APCS, APOA4, F2 and AMBP) that had maintained accuracy which could discern early (I/II) from late (III/IV) stage CRC.

Conclusion: MS-based proteomics in combination with ultradepletion strategies have an immense potential of identifying diagnostic protein biosignature.

Potential early clinical stage colorectal cancer diagnosis using a proteomics blood test

Samridhi Sharma[1‡], Seong Beom Ahn[1‡], Abidali Mohamedali[2], Sadia Mahboob[1], William Redmond[1], Dana Pascovici[2], Jemma X. Wu[2], Thiri Zaw[2], Subash Adhikari1, Vineet Vaibhav[1], Edouard C. Nice[3] and Mark S. Baker[1]*

‡ **These authors contributed equally**

[1] Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Macquarie University, NSW, 2109, Australia

[2] Department of Molecular Sciences, Faculty of Science and Engineering, Macquarie University, NSW, 2109, Australia

[3] Department of Biochemistry and Molecular Biology, Faculty of Medicine, Nursing and Health Sciences, Monash University, VIC, 3800, Australia

Keywords: Colorectal cancer, SWATH™-MS, Depletion, Plasma, Biomarkers, Early Stage, Diagnosis, Predictive model

***Corresponding Author:**

Professor Mark S. Baker

Faculty of Medicine and Health Sciences

Level 1, 75 Talavera Road

Macquarie University

NSW, 2109, Australia

Disclosure of Potential Conflicts of Interest:

The authors declare no real or potential conflicts of interest

3.1 Introduction

Global temporal patterns of colorectal cancer (CRC) incidence and mortality are alarming. In 2018, it is estimated that over 1.8 million patients will be diagnosed with CRC, resulting in over 800,000 deaths annually (1). These statistics are expected to increase to ~2.2 million new cases with 1.1 million fatalities by 2030 (2). This trend can partially be explained by the fact that early stages of the disease are especially asymptomatic with the majority of patients diagnosed when tumours have already invaded local lymph nodes (stage III) or metastasised to distant organs (stage IV), leading to survival rates lower than 13% (2, 3). Surgical tumour resection in early stage disease can be both preventive and curative (4) with the 5-year survival rate of early stage I/II CRC patients greater than 90% (5). There is therefore a substantial need to reliably, accurately and consistently diagnose CRC as early as possible.

There are a number of stool-based tests and structural examinations (6, 7) that are in use clinically to aid early CRC detection. In developed countries, stool-based tests like gFOBT (guaiac chemical faecal occult blood tests), FIT (faecal immunochemical tests) and mt-sDNA (multi-target stool DNA tests) are distributed to most-at-risk populations (e.g., those aged 50-74 years) (8). The gFOBT (sensitivity 62-79%; specificity 87%- 96%) and FIT (sensitivity 73-92%; specificity 91-97%) (6) tests rely on the chemical or immunological detection of faecal hemoglobin (Hb) respectively (8). The mt-sDNA test, which has a lower ~90% specificity, (6) identifies multiple molecular biomarkers, such as hypermethylated BMP3/NDRG4, point mutations in KRAS and the beta-actin gene as well as Hb protein (9). However, despite extensive public health education programs worldwide, patient participation/compliance with faecal-based screening tests has rarely (if ever) exceeded 44% (6, 10, 11).

Positive faecal gFOBT/FIT test results (i.e., true or false positives) are referred to more invasive structural tests for confirmation. These structural tests include computed topographic colonography (CTC) and flexible sigmoidoscopy (FS) (6). The efficacy of CTC and FS is restricted by exposure to low-dose radiation and incomplete examination of the proximal colon, respectively (6). As per standard practice of care, all positive non-colonoscopic screening procedures are followed up with a confirmatory colonoscopy.

However, colonoscopy is expensive, invasive, requires unpleasant preparation and causes occasional adverse sedation morbidities as well as unavoidable infrequent mortality from adverse consequences like bowel perforation and sepsis (6). Low compliance and sensitivity

of faecal tests has compelled the investigation of potential blood tests that have a much higher compliance rate (as high as 97% in controlled studies).

Two primary classes of blood-based markers have been developed, namely DNA-based and protein-based. Tests that detect tumour-specific genetic and epigenetically-altered circulating tumour DNA (ctDNA) released from tumour cells are colloquially termed 'liquid biopsy' tests (12). However, there remain some technology barriers to early clinical stage cancer screening using liquid biopsy tests. These include; secretion of negligible levels of ctDNA from small adenomas or early stage tumours meaning large amounts of blood are required, mutational heterogeneity among individual patients (13) and poor association of emerging mutational biomarkers with cancer stages and types, each of which limits use for screening early clinical stage CRC patients (14).

Of protein markers, carcinoembryonic antigen (CEA) was one of the earliest to be used clinically, although it has been subsequently discounted as efficacious for early-stage screening (15). Plasma CEA levels are primarily used to monitor colorectal carcinoma treatment and to identify recurrence after surgical resection, despite having a low 35% sensitivity and 87% specificity (16). Furthermore, CEA is expressed in many other cancers (17, 18) and is not specific to CRC. Multiple other protein markers have been proposed (19), however only a few individuals have shown translational promise. Protein-based blood biomarkers offer significant advantages that make them amenable for the development of an ideal population blood-based CRC screening test. They purport to be accurate, specific, sensitive and inexpensive (11). Furthermore, protein-based tests offer significant advantages in translatability with current technologies and clinical laboratory practices (20). The key, however, remains, to find a molecular protein-based biomarker (or panel) that provides better specificity and sensitivity than gFOBT and FIT, as a pre-colonoscopy screening test.

Blood plasma is a complex body fluid owing to the high dynamic concentration range of proteins found within it. The concentration range of human blood plasma proteins extends 12-13 orders of magnitude (21), with >90% of all plasma protein content covered by a few (10 to 14) highly abundant proteins found above the mg/ml mark. These are primarily haemostatic (e.g., albumin), acute phase response proteins (e.g., serpins), lipid/protein transporters and immunoglobulins (21, 22). The remaining low and medium abundance proteins are found at concentrations ranging from ng/ml down to pg/ml and are often derived from proteins that have leaked or been shed from tissues (including diseased cells/tissues) or that represent interleukins, cytokines or growth factors (21, 23). These low abundance proteins potentially hold critical

information regarding the health and disease status of any individual (24). However, low abundant proteins are masked by more abundant proteins and are difficult to detect in a proteomics discovery experiment. Indeed, the repertoire of often identified disease biomarker candidates from mass spectrometry are usually categorised as general inflammatory response proteins, lipid transporters or coagulation cascade proteins (25-27). In other words, many proteomic biomarker studies unearth proteins of unremarkable biological context, meaning that they code for disease with particularly low specificity (28).

This study aimed to adopt a multilayered plasma proteomic approach to discover protein biomarkers for the detection of CRC patients at earlier stages (I/II) from EDTA plasmas. To visualise and quantify novel lower abundance proteins, a combinations of commercially available depletion (i.e., MARS-14) (29) and an in-house ultradepletion system (30, 31) were used. SWATH™-MS (**S**equential **W**indow **A**cquisition of all **TH**eoretical **M**ass **S**pectra) was employed for deep and reliable exploration of the plasma proteome. These studies were applied to a set of pooled EDTA-plasma samples in order to identify potential candidates for early stage I/II CRC detection. To verify the diagnostic ability of candidate biomarkers, Western blotting and ELISA on pooled and individual samples were performed, where tests were available commercially (experimental procedure summarised in Figure 3.1). Finally, a machine-learning approach to further test the validity of our candidates was utilised. Unsupervised clustering algorithms were used to validate how dissimilar early stage I/II CRC were from healthy subjects. Supervised classifiers on generated data based on the variance found in our individual samples were used, which was then tested on real patient data. This discovery experiment resulted in a novel blood-based multi-analyte biomarker signature panel that requires comprehensive validation to allow population-based detection of stages I and II CRC.

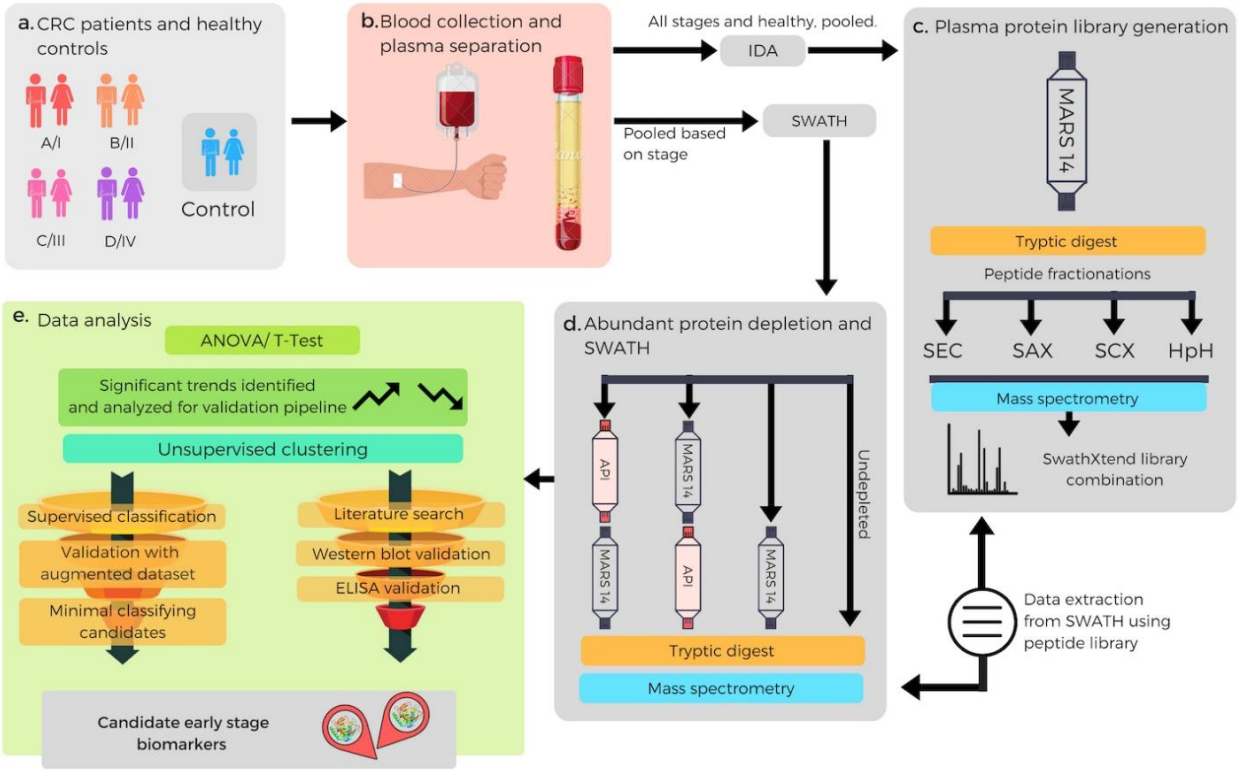


Figure 3.1: *Blood-based multi-analyte proteomic signature discovery workflow: (a) A total of 100 age- and sex-matched EDTA-plasma samples were procured (n=20 per stage I, II, III, IV, and n=20 healthy controls (non-menopausal, non-smoking and no history of any cancers)). (b) Plasma samples were collected as per ethics requirements. To create a plasma reference library, equal volumes of all patients and healthy plasmas were pooled. For the SWATH™-MS experiments, equal volumes of 20 plasma samples were combined to produce pools of each of the 4 CRC stages (I-IV) and healthy controls (c) For library generation, HAPs depleted using MARS-14 column (Agilent) followed by tryptic digestion and peptide fractionation by SAX, SCX, SEC and HpH (independently), followed by IDA-MS analysis. (d) The stage pooled samples were processed through four different experiments (three, where the plasma HAP were depleted and one where it was not). The resulting proteins were digested and subjected to SWATH™-MS. Lists of quantifiable proteins were extracted from the SWATH™-MS dataset using the peptide library generated in (c). (e) Differentially expressed proteins were first identified using ANOVA/t-test (p-value<0.05, fold change cut off ±1.5), resulting in 37 proteins exhibited with differential expression across all CRC stages compared to healthy controls. These 37 proteins were further evaluated by unsupervised clustering method to increase discriminatory power. Differentially expressed proteins were subjected to validation pipeline where they were checked to identify evidence in the literature, followed by experimental validation (ELISA/Western blotting) of a subset that seemed most promising. Concurrently, the*

samples also underwent a supervised classification method which identified potential candidates which were then validated with an augmented dataset (with a SD 10 times the observed variance). This resulted in a subset of 5 candidate proteins that were able to classify the different stages of the disease. SAX: strong anion exchange, SCX: strong cation exchange, SEC: size exclusion chromatography, HpH: high pH reversed phased c18, SWATH™-MS : sequential window acquisition of all theoretical mass spectra, IDA-MS: information-dependent acquisition mass spectrometry, SD: standard deviation, HAPs: high abundant proteins.

3.2 Material and Methods

Ethic statement and sample collection:

This study was performed with approval from the Macquarie University Human Research Ethics Committee (MQ HREC approval #5201200702). The cohort of 100 patient EDTA-plasma samples was procured from the Victorian Cancer Biobank (VCB) in Melbourne, Australia. The experiment assembled 100 individual EDTA-plasma samples, composed of 80 from Dukes' staging system staged CRC (n=20 each for stages A, B, C, and D). These were clinically re-classified as stage I, II, III, and IV CRCs respectively according to the AJCC system. Samples were collected from CRC patients diagnosed with non-malignant/malignant tumours, before they underwent any treatment or surgery for CRC. Plasma was also collected from 20 healthy donors (Victorian Cancer Biobank) that were age- and sex-matched, non-smokers and with no prior history of cancer or other major disease. The demographic and clinical information is summarised in Table S3.5.

Cancer and healthy plasma samples were processed identically throughout the study and supplied for this study by the Victorian Cancer Biobank. The blood samples were collected in 9ml EDTA tubes and centrifuged for 10 mins at room temperature at 1200g. The supernatant liquid (plasma) was transferred to a single tube of 10 ml, centrifuged again at room temperature for 10 minutes at 1800 x g, aliquoted into 15 x 250µl aliquots and then stored at -80˚C (15). The whole process of sample preparation was completed within 2 hours of plasma collection as per the Victorian Cancer Biobank guidelines. TNM staging, 5-year survival and 5-year recurrence data for recruited patients is tabulated in Appendix VI.

Multiple affinity removal system (MARS-14) high abundance plasma protein depletion:

A previous study using the MARS-14 system has shown that depletion columns afford highly repeatable and efficient plasma fractionation with few non-targeted proteins captured (29). The

Agilent MARS-14 high capacity affinity column (4.6×100 mm) was designed to employ anti-human plasma protein monoclonal antibodies to remove the 14 most abundant proteins (human serum albumin, IgG, antitrypsin, IgA, transferrin, haptoglobin, fibrinogen, α2-macroglobulin, α1-acid glycoprotein, IgM, apolipoprotein AI, apolipoprotein AII, complement C3 and transthyretin) from human plasma. Depletion was performed on an Agilent 1260 HPLC system where 40µl EDTA-plasma samples were first diluted 4-fold using buffer A supplied by the manufacturer followed by 0.22µm spin filtering at 4°C in technical triplicates. Eluates plasmas were injected to run on the HPLC, and proteins eluted following the manufacturer's instructions.

In-house abundant protein immuno-depletion (API):

In detail, chicken IgY polyclonal antibodies were raised against 7 dual (SCX followed by SAX including dual flow-through proteins) ion-exchange fractions of human plasma. Purified IgYs were covalently-linked as antigen affinity-purified IgYs to activated hydrazide beads (GE, Uppsala, Sweden) following the manufacturer's instructions and packed into columns as described previously (30, 31). This API (abundant protein immunodepletion) column was subsequently pre-equilibrated at 5ml/min using PBS at pH 7.2. Plasma was injected into the column at 0.1ml/min and washed using 2.5 column volumes of PBS, first at 0.05ml/min for 3min and then at 5ml/min. Bound proteins were subsequently eluted from the API column using 4 column volumes of 0.1M glycine buffer at pH2.5 and a flow rate of 5 ml/min. Neutralisation using glycine 100mM, pH 10 was performed on all bound fractions post-elution for long-term storage at -80°C prior to LC-MS/MS. All samples were buffer exchanged using 3kDa Amicon filtration and total protein quantified using a Micro BCA Protein Assay kit (Thermo Scientific™). API columns were immediately re-equilibrated with 5 column volumes of binding buffer at 5 ml/min for subsequent re-use (30, 31). The samples were prepared as technical triplicates.

Tryptic digestion: Prior to tryptic digestion, protein concentration was measured using a BCA Protein Assay Kit following the manufacturer's protocol (Thermo Fisher Scientific) for both depleted and nondepleted samples. The samples were reduced with 5mM dithiothreitol (DTT) at 60°C for 30min and alkylation with 25mM iodoacetamide (IAA) at room temperature for 30min in the dark. Samples were diluted in 100mM ammonium bicarbonate and digested with sequencing grade porcine trypsin (Promega) at a protease to substrate ratio of 1:30 at 37°C for 16hr. Peptide mixtures were desalted and cleaned with C18 OMIX tips (Agilent) according to the manufacturer's protocol followed by drying by vacuum centrifugation.

Strong Cation Exchange (SCX) peptide fractionation: Tryptic digested peptides (100µg) were fractionated using a poly-sulfonylethyl column A size 200 x 2.1mm, 5µm, 200Å column attached to the 1260 series HPLC (Agilent, Santa Clara, CA, USA). The separation was initiated, at a constant flow rate of 0.3 ml/min, with 100 % buffer A (5 mM KH_2PO_4, pH 2.72, 25% acetonitrile) for 25min. This was followed by a gradual increase in buffer B (5 mM KH_2PO_4, pH 2.72, 350 mM KCl, 25% acetonitrile) concentration from 0 % to 45 % over 70min.

Strong Anion Exchange (SAX) peptide fractionation: Digested peptides (100µg) were fractionated using a UNO^{TM} Q1 column (Bio-Rad, CA, USA) on a 1260 series HPLC (Agilent, Santa Clara, CA, USA). Fractionation was performed at a constant flow rate of 0.5 ml/min with peptides eluted on a linear gradient of buffers A (20mM Tris-HCl, pH 7) for 10min then a linear increase of buffer B (20mM Tris-HCl, pH 7, 1M KCl) to 100% over 60min and held for 10min and finally replaced with buffer C (20mM Tris-HCl, pH 7, 2M KCl) to 100%.

Size Exclusion Chromatography (SEC) peptide fractionation: Peptides (100µg) were fractionated using Tricorn Superdex 75 10/300 GL, 10 × 300-310 mm, 13µm column (Amersham Biosciences) on a 1260 series HPLC (Agilent, Santa Clara, CA, USA). Elution of peptides was performed using a 100mM $NaPO_4$, 250mM NaCl, pH 7 at an isocratic flow rate of 0.5ml/min. Peptides were collected over 80min.

High pH reversed phased C18 (HpH) peptide fractionation: Peptides (100µg) were fractionated using a ZORBAX 300 Extend-C18 2.1x150 mm, 3.5µm column on a 1260 HPLC system (Agilent, Santa Clara, CA, USA). Buffer A (5mM ammonium formate (NH_4COOH)) and B (5mM NH_4COOH, 90% acetonitrile in water) were used for the fractionation at a constant flow rate of 0.3 ml/min.

SWATH™-MS library generation (information-dependent acquisition, IDA): All fractionated peptides obtained from multiple peptide fractionation methods (as descripted above) were used for SWATH™-MS reference library generation (i.e., protein identification). The protein identification was performed on a Sciex TripleTOF 5600 (Sciex, Framingham, MA) coupled with Eksigent Ultra nanoLC system (Eksigent Technologies, Dublin, CA). Peptides were injected onto a reverse phase peptide C18 trap (Bruker peptide Captrap) for pre-concentration and desalted at a flow rate of 10µL per min for 5 min with 0.1% formic acid (v/v) and 2% acetonitrile (v/v). After desalting, the peptide trap was switched in-line with an in-house packed analytical column (150µm × 10cm, solid core Halo C18, 160 Å, 2.7 µm media (Bruker)). Peptides were eluted and separated from the column using the buffer B (99.9%

acetonitrile (v/v), 0.1% formic acid (v/v)) gradient starting from 2% and increasing to 10% for 10min then to 35% over the next 78min at a flow rate of 500nL per minute. After peptide elution, the column was cleaned with 95% buffer B for 10min and equilibrated with 98% buffer A (0.1% formic acid (v/v)) for 20 minutes before next injection. In IDA mode, a TOFMS survey scan was acquired at m/z 350 – 1500 with 0.25 second accumulation time, with the ten most intense precursor ions (2+-5+; counts >150) in the survey scan consecutively isolated for subsequent product ion scans. Dynamic exclusion was used with a window of 20secs. Product ion spectra were accumulated for 50msecs in the mass range m/z 100 – 1500 with rolling collision energy.

IDA data were subjected to database searches by ProteinPilot (V4.2, SCIEX) using the Paragon algorithm (33). *Homo sapiens* database was obtained from SwissProt (20,204 entries, 2015 version). The search parameters were as follows: sample type: identification; cys alkylation: iodoacetamide; digestion: trypsin; instrument: TripleTOF 5600; special factors: none; ID focus: biological modifications; miss-cleavages: one; precursor peptide mass tolerance: ±50ppm; fragment ion mass tolerance: ±0.1Da; peptide length: >7 amino acids. A reverse-decoy database search strategy was used with ProteinPilot, with the calculated protein FDR<1% and a probability cut off at 0.99. The analysis of proteomics experiments conformed to the guidelines provided by Journal of Proteome Research. The link to these guidelines is https://pubsapp.acs.org/paragonplus/submission/jprobs/jprobs_mass_spectrometry_guidelines.pdf

SWATH™-MS: A Sciex TripleTOF 5600 coupled with Eksigent Ultra nanoLC system and identical LC conditions (as described above) were used for SWATH™-MS experiments. Initially, the precursor m/z frequencies from generated IDA data (above) were used to determine the sizes of m/z window. SWATH™-MS variable window acquisition with a set of 60 overlapping windows (1amu for window overlap) was constructed covering the mass range of m/a 399.5 – 1249.5. In SWATH™-MS mode, TOFMS survey scans were acquired (m/z 350-1500, 0.05 sec) then the 60 predefined m/z ranges were sequentially subjected to MS/MS analysis. Product ion spectra were accumulated for 60msecs in the mass range m/z 350-1500 with rolling collision energy optimised for lowed m/z in m/z window +10%.

SWATH™-MS data were extracted using PeakView (v2.1) with the following parameters: top 6 most intense fragments per peptide, fragment tolerance at 75ppm, 10min retention time window, confidence thresholds of 99%, FDR for transitions <1% (based on chromatographic feature after fragment extraction) and exclusion of shared/modified peptides.

Statistical analyses: Peptide quantification was performed using peak areas from extracted ion chromatograms and proteins were quantified using cumulative mean values of the calculated peptide quantities. The extracted data was normalised using total area normalisation, and log-transformed prior to statistical analysis; the data distribution was examined using density plots and boxplots. The overall sample look and consistency of the technical triplicates was examined visually using hierarchical clustering and PCA plots.

Extracted quantitation contained data from pooled samples in technical triplicates, belonging to five categories: CRC stage I-IV and healthy control. Proteins differentially expressed between the five categories were identified based on a one-way ANOVA run separately for each protein, selecting proteins based on an ANOVA p-value criterion (<0.05) and maximum fold change (FC > 1.5). Pairwise t-tests were also carried out, using both a protein level and peptide-level approach. The statistical analysis protocol is embedded in SWATHXtend as described in detail previously (34).

Unsupervised and supervised machine-learning: The differentially expressed protein candidates analysed by one-way ANOVA and pairwise t-test were consolidated in a single dataset from the different depletions, and were further evaluated, first, by being plotted in 3D-space following unsupervised clustering techniques. Dissimilarity matrix were created based on the peak areas of technical replicates for each condition and plotted by using multi-dimensional scaling. The data is represented based on the first dimensions for each CRC stage and healthy. Results from this clustering approach were verified using principal component analysis (PCA). Both methods were done in MATLAB.

Although supervised classification approaches have been used in recent years with proteomics datasets (35, 36), the nature of most proteomics datasets, with a high number of proteins but a small population, make their validity as early predictors of a disease debatable. One way to overcome the limitations of such a dataset is to generate a synthetic dataset based on real participants' information in order to perform classification. Data augmentation is a mainstay for training classification algorithms in the field of machine-vision and medical imaging analysis (37, 38), though not widely used with proteomics data. Here these methods were adapted as further validation of our results. To evaluate the predictive power of the selected panel of candidate protein biomarkers, a synthetic population of patients (1,000 per the 4 CRC stages as well as healthy controls, total=5,000) was created by generating a normal distribution of random number at 10 times the standard deviation (SD) for each protein concentration from our technical replicates. Data augmentation was also performed in MATLAB.

Once the dataset was generated for each group, various classification approaches (including a shallow neural network as well as k-nearest neighbour and decision tree classifiers) were applied, using the MATLAB neural network toolbox and classification app. For the shallow neural network, the network was composed of 10 hidden neurons, with 70% of the data used for training, 15% for validation and 15% for testing. Once the network was trained, it was deployed to test on the dataset comprising our real pooled patient values.

Western blotting: Protein concentration was measured using a BCA Protein Assay Kit following the manufacturer's protocol (Thermo Fisher Scientific). Proteins (25µg/sample) were separated on a 4-12% SDS-PAGE gel and transferred onto nitrocellulose membrane blots using semi-dry blotting system (Bio-Rad) following the manufacturer's protocol. To ensure the equal protein loading in each lane, the blots were stained Ponceau S (Sigma) and imaged on a ChemiDoc™ imaging system (Bio-Rad). Blots were then incubated with primary monoclonal/polyclonal antibodies including CFD (R&D systems, AF1824, 1:2500), GPX3 (R&D systems AF4199, 1:200), CST3 (Abcam ab133495, 1:13000), PON1 (Abcam, ab92466, 1:5000), MRC1 (Abcam ab195193, 1:1000) and COMP (Abcam, ab74524, 1:200), followed by respective HRP-conjugated secondary antibodies. Blots were imaged using a Li-Cor Odyssey Blot imager (LI-COR Biosciences). Quantitation of signal intensity of the bands in Western blots was performed using Image lab software version 5.0 (Bio-Rad) and Image Studio Lite version 5.2 (LI-COR Biosciences).

Enzyme-Linked Immunosorbent Assay (ELISA) verification: Expression level of ADAMDEC1 from pooled and individual plasma (n=100, 20 per stage (I-IV) and 20 heathy control) was measured using MyBioSource ELISA kit (Catalogue #: MBS928931) following the manufacturer's instructions. Optical densities were measured at 450 nm and 570 nm using a PHERAstar® microplate reader (BMG Labtech). Statistical significance of differential expression of the plasma proteins was analysed by one-way ANOVA on Prism software v.7 (graph pad).

3.3 Results

Plasma SWATH™-MS library generated using several protein/peptide fractionation methods

A total of 513 distinct plasma proteins were identified by combined heathy/CRC plasma using HAP depletion and four peptide fractionation methodologies (Figure 3.2a). A total of 361 plasma proteins using HpH fractionation, 295 proteins by SAX, 332 proteins by SEC and 344

by SCX. (Figure 3.2a). The HpH peptide fractionation method identified a most number of proteins with higher stringency MS-based identification criteria (40) (Supp Figure S3.1). Detailed information for peptide/protein identification is shown in Supp Table S3.1 which include (i) list of proteins identified in each fractionation method, (ii) number of unique peptides identified for each protein, (iii) amino acid sequences of each peptides, (iv) number of missed cleavages for each peptide and (v) uniqueness (uniquely mapping non-nested) of each peptide. Although the SCX method provided the highest number of protein identifications, the HpH fractionation method identified more proteins with higher stringency MS-based identification criteria (40), encompassing >2 non-nested, unitypic peptides where each peptide identified should be >9 amino acids in length and PSM, peptide and protein FDRs of <1% (Supp Table S3.1 and Supp Figure S3.1).

To visualise the detectable threshold of plasma proteins in our SWATH™-MS library, a scatter plot analogous to the "Anderson curve" was plotted (21). It exemplifies the high dynamic plasma protein concentration range (Figure 3.2b). Based upon the Plasma Proteome Database, PeptideAtlas and the PubMed literature, reported concentrations for 425 proteins (out of 529 total identified proteins) was identified. These reported concentrations were used to create a scatter plot (Figure 3.2b). It should be noted that, not plot all 3,509 human plasma proteins identified to date at high stringency by the Human (Plasma) Proteome Project (41). It should also be noted that the 427 proteins were uncovered spanned ~10 orders of magnitude in protein concentration. The concentration for the most abundant protein (hman serum albumin; ALB) was found to be ~40.6 mg/ml down to the lowest protein identified at 4.3pg/ml which was found to be multiple EGF-like domains 8 protein (MEGF8), a protein whose function is unclear but may be involved in cell adhesion/attachment (Figure 3.2b, Supp Table S3.2). A significant residual 104/529 human plasma proteins identified in the SWATH™-MS library currently have no reported plasma concentrations, to the best of our knowledge. Interestingly, based on a search against the PeptideAtlas database in May 2019, 8 plasma proteins found in our SWATH™-MS library compilation were reported as plasma proteins for the first time (Supp Table S3.2).

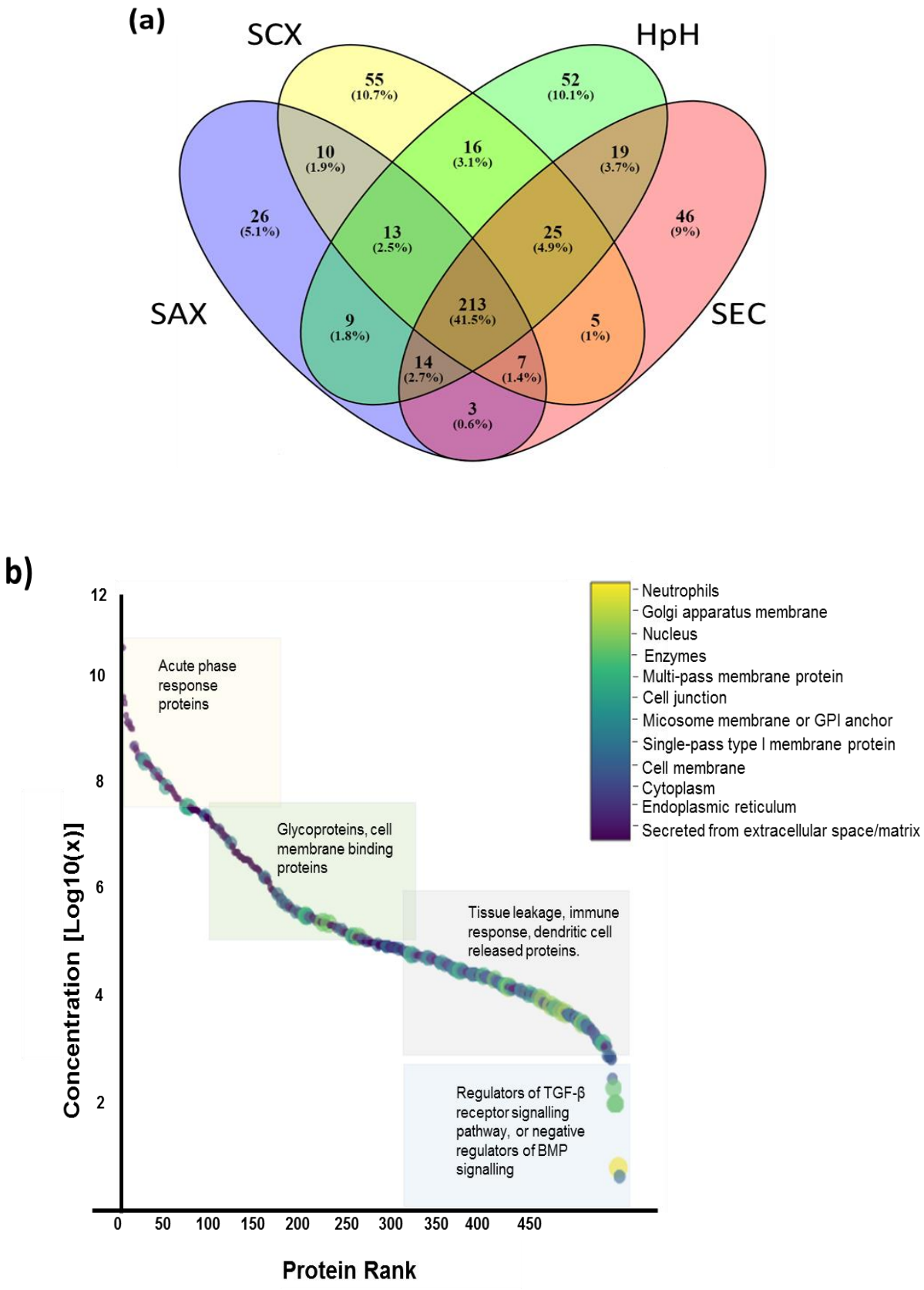


Figure 3.2: SWATH™-MS reference library with functional annotations; *Illustrates; (a) Venn diagram comparing a number of common, unshared and shared proteins identified*

between four peptide fractionation methods used to compile a plasma SWATH™-MS library, with (b) "Anderson curve" superimposed with gene ontology information from plasma proteins identified in the study. The colour code bar shown indicated on the right-hand side of Figure 3.2b corresponds to various gene ontology characteristics applied to data points shown on the concentration curve. ***HpH****: high pH C18 reversed phase separation,* ***SAX****: strong anion exchange,* ***SEC*** *size exclusion chromatography,* ***SCX****: strong cation exchange.*

Functionalities of identified plasma proteins

To visualise the functionalities of proteins found in our plasma SWATH™-MS library, UniProt was employed to annotate; (i) subcellular localisation, (ii) tissue specificity, (iii) gene ontology analyses (biological processes, cellular component, molecular function), and (iv) protein families (Supp Table S3.2, Figure 3.2b). As expected, those proteins found to lie in the high abundance range were mostly classical plasma proteins such as those that are known to be liver-derived or acute phase response proteins, including HAPs like human serum albumin, immunoglobulin (multiple types), fibrinogen, chylomicron proteins, transferrin, haptoglobin, C-reactive protein, clusterin (ApoJ), and complementary factor B. Gene ontology analysis classifies these proteins as involved in biological processes like positive/negative activators of acute phase response, antimicrobial response, blood coagulation or complement activation.

Mid-range proteins, on the other hand, consisted predominantly of peptidases, serpins, S-100 family proteins, glycoproteins, and cell membrane binding proteins like cystatin C, CD59, C1Q, extracellular matrix proteins and superoxide dismutase, amongst others. Some of these plasma proteins were found to have roles in cell-cell signalling, angiogenesis and activation of MAPK activity.

In the low abundance range, cell membrane proteins, extracellular exosome proteins, proteins secreted from the endoplasmic reticulum or lysosome membrane and intracellular secreted proteins were found. Examples included, hyaluronan-binding protein 2, galectin-3-binding protein, phosphatidylinositol-glycan-specific phospholipase D. The lowest discovered plasma proteins found were in the ρg/ml concentration range and included the E3 ubiquitin-protein ligase TRIM33 that is known to be specifically expressed in colon adenomas and adenocarcinomas and is thought to be a regulator of TGF-β receptor signalling pathway (42). A detailed list of the SWATH™-MS library specific peptides, their length, number of peptides per proteins and their unitypicity can be found in Supp Tables S3.1 and S3.2.

Identification of quantifiable plasma proteins in healthy or CRC plasmas using various (ultra)depletion strategies

Having compiled a comprehensive SWATH™-MS reference library, the SWATH™-MS analysis on pooled human healthy and CRC plasma samples was performed. As described, pooled (n=20) human plasmas for each of stages I-IV CRCs and healthy controls were (i) non-depleted, (ii) MARS-14 only depleted, (iii) ultradepleted using MARS-14 followed by API using purified anti-human plasma fraction chicken IgY columns (30, 31) (MARS-14→API), and finally (iv) ultradepleted using API-depletion followed by MARS-14 (API→MARS-14). Each of the nondepleted, depleted and both ultradepleted experiments were run as technical triplicates (refer to Figure 3.1 for an overview of the experimental plan). Compilation of all SWATH™-MS experiments as outlined above, resulted in the identification and quantitation of a total of 444 distinct human plasma proteins from healthy or CRC plasmas (Figure 3.3a). Detailed information of all quantifiable plasma proteins and peptides captured by these non-depletion and depletion strategies are illustrated in Supp Table S3.4.

When non-depleted plasmas were analysed, a total of 315 proteins were identified and quantified that had been deposited prior into the SWATH™-MS library. In agreement with previously published studies (29), use of the Agilent MARS-14 system that removes 14 most highly abundant plasma proteins allowed for the identification of 362 proteins, including an additional 88 plasma proteins not observed in non-depleted plasmas. Equally, non-depleted plasmas contained 41 unique proteins not found after MARS-14 depletion, indicating the distinct possibility of significant co-depletion as an off-target effect of the use of MARS-14 depletion. This observation correlates with previous work illustrating additional proteins are likely bound to targeted MARS-14 proteins and are unexpectedly/inadvertently co-depleted (43).

To comprehensively expose lower abundance proteins differential-expression between healthy and clinically staged CRC plasmas, various ultradepletion approaches were undertaken. Systematic depletion of high-medium abundance proteins performed using MARS-14 followed by API identified 325 proteins. Of these 31 proteins had not been previously observed in non-depleted or MARS-14 depleted plasmas with 29 were not seen by any other method. Reversing the order of ultradepletion (i.e., API depletion followed by MARS-14) identified only 244 proteins, 12 which had not been previously observed in non-depleted or MARS-14 depletion whilst only 10 were newly identified.

In summary, MARS-14 depletion allowed 28 unique proteins to be observed whilst ultradepletion allowed for the visualisation of 41 unique proteins (Figure 3.3a). Collectively,

129 proteins were identified and quantitated additionally (i.e., ~30% of the total 444 plasma proteome subset identified) using all (ultra)depletion strategies employed.

To visualise the protein concentration range of these additional 129 proteins, the red dots were superimposed onto the complete plasma SWATH™-MS library (blue dots) on an "Anderson curve" (Figure 3.3b). This result demonstrates that these additional 129 proteins represented mostly medium-low abundance plasma proteins (e.g., LECT2, ADAMTS13 and PCDH12). These results show that high-medium abundance plasma protein depletion allows for even deeper and more comprehensive (though obvious not complete) proteome coverage.

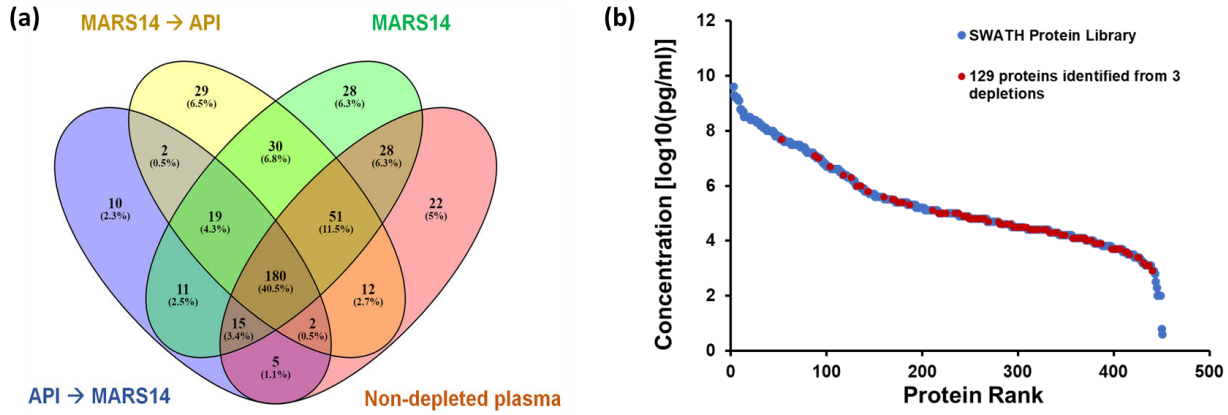


Figure 3.3: Quantifiable plasma proteins found in healthy/CRC plasmas from undepleted and multiple plasma protein depletion strategies. *Venn diagram (a) showing the numbers of unique and common quantifiable proteins following three depletion (MARS-14, API followed by MARS-14 and MARS-14 followed by API) and non-depletion experiments. Protein concentration range (b) of the additional 129 proteins found after high-medium abundance protein depletion on the plasma SWATH™-MS library "Anderson curve".*

Differentially expressed plasma protein biomarkers of early stages I/II CRC:

Prior to statistical analysis, the extracted SWATH™-MS dataset from each depletion and the non-depletion experiment was independently normalised using total area normalisation and data distribution was examined using density plots and boxplots (Supp. Figure S3.2). Furthermore, consistency of sample replication was examined visually using hierarchical clustering and PCA plots (Supp. Figure S3.2).

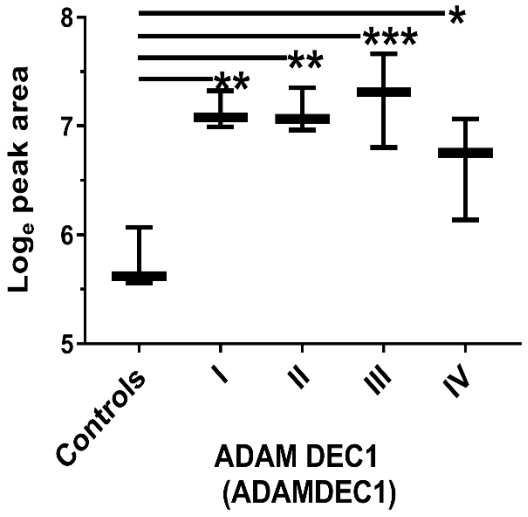
To discover plasma proteins that were differentially expressed between healthy and staged I-IV CRC plasmas, one-way ANOVA and Pairwise t-test at both the protein and peptide levels were employed. All differentially expressed proteins were selected based on a p-value < 0.05

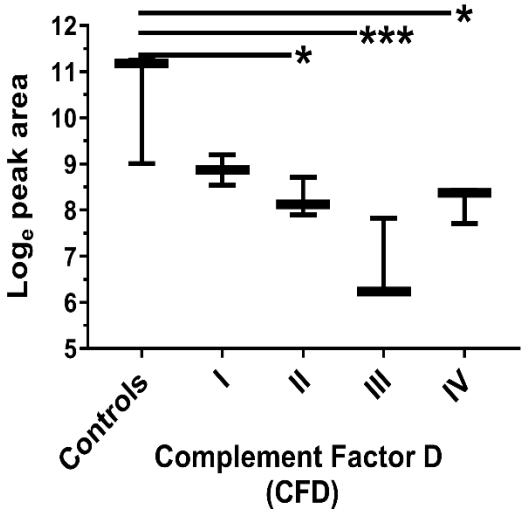
and a fold change ratio cut off of ±1.5. These proteins were further filtered to retain only those candidates that exhibited consistent trends (up or down-regulation) in all stages compared to control, and these results were consolidated from all depletions. This analysis resulted in the identification of a total of 37 protein candidates that exhibited differential (↓↑) expression in all the four (I-IV) CRC stages when compared to healthy controls from a comparison of the non-depleted and three depleted experiments. Detailed information regarding each of these 37 CRC biomarker protein candidates is presented in Supp Table S3.5.

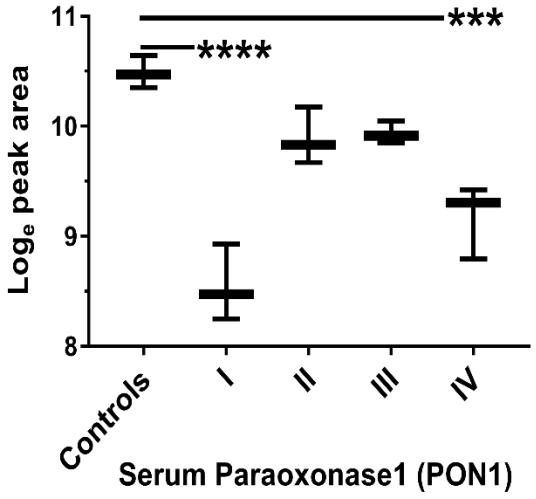
The highest number of differentially expressed proteins were found in the API→MARS-14 ultradepleted healthy against CRC samples, whereas non-depleted samples resulted in the lowest number of differentially expressed proteins. It should be noted that some proteins (e.g., SAA2) were consistently up-regulated in disease CRC plasmas whether the data came from non-depleted or after MARS-14 depletion. Equally, GPX3 was consistently up-regulated in both MARS-14 depleted and MARS-14→API depletion experiments. Additionally, CST3 and CFD were consistently down-regulated in all stages of CRC plasmas using both MARS-14 and API→MARS-14 depletion. Figure 3.4 represents a subset of these data. CRC biomarker candidate proteins were subsequently selected based on biological relevance as well as statistical analysis (e.g., predictive modelling) discussed below.

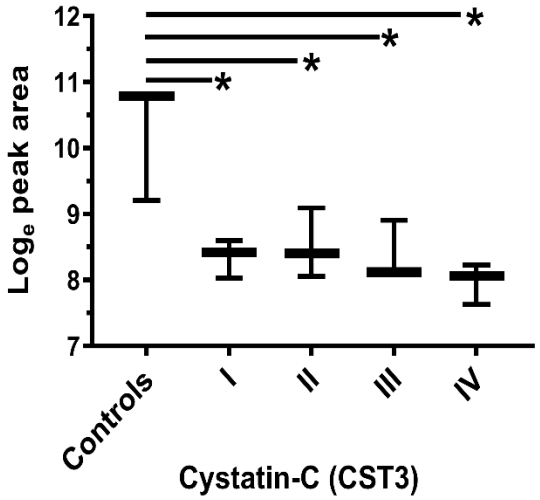
Of the 37 CRC protein biomarker candidates, 31 had reported known concentration whilst the plasma concentration of the remaining 6 proteins had not been reported. These 31 reported proteins were mapped onto the plasma SWATH™-MS library Anderson concentration curve (Figure 3.5), demonstrating that the concentrations of protein candidates were widely represented across a broad plasma protein concentration range.

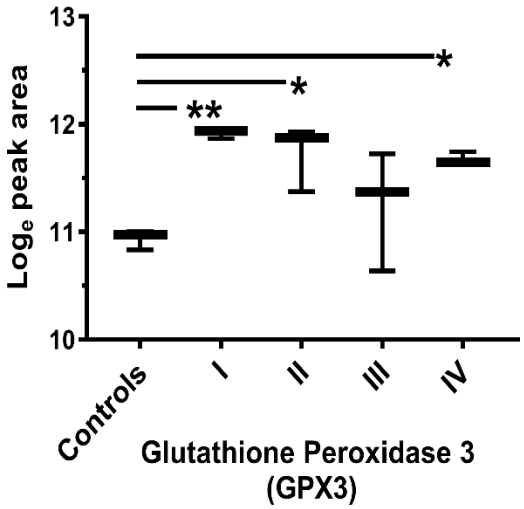
Gene ontology characteristics of the 37 CRC protein biomarker candidates using UniProt and the Human Protein Atlas to determine potential biological relevance was used. Of these, 10 proteins were found to be liver-derived proteins (APOA2, APOC3, F2, APOC2, SERPIN6, PON1, AMBP, SAA1, SAA2, and HGFAC), and *in toto*, all 37 proteins had subcellular attributes associated with the cytosol (APOB, SAA1, HGFAC, S100A8, PFN1, APOA2, F2), exosomes (VASN, COMP), secretory proteins (COMP, ADEC1, SODE, HGFAC, C1QC, ITIH3, CFAD, MASP2, SAA1, SAA2, GPX3, SAMP, AMBP, PON1), or had been shown to be an integral component of cell membranes (VASN). Three candidates were expressed in somatic tissue (MECP2), endothelial cells (ROBO4) or were known to be secreted in response to dendritic cell activation and maturation (ADAMDEC1; Supp Table S3.5).

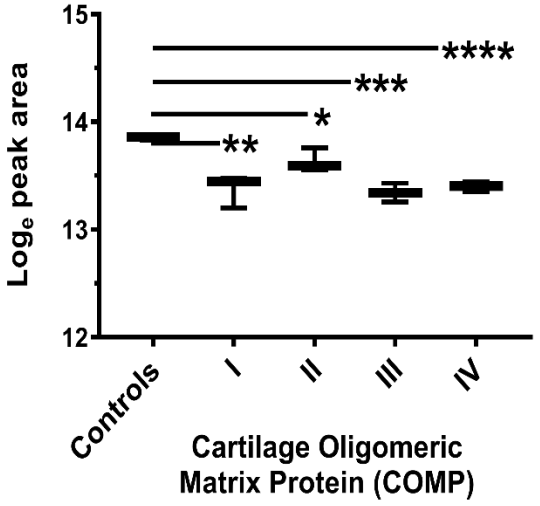

Log$_e$ peak area
8
7
6
5
*
***
**
**
Controls
I
II
III
IV
ADAM DEC1
(ADAMDEC1)


Log$_e$ peak area
12
11
10
9
8
7
6
5
*
***
*
Controls
I
II
III
IV
Complement Factor D
(CFD)


Log$_e$ peak area
11
10
9
8
***
****
Controls
I
II
III
IV
Serum Paraoxonase1 (PON1)


Log$_e$ peak area
12
11
10
9
8
7
*
*
*
Controls
I
II
III
IV
Cystatin-C (CST3)


Log$_e$ peak area
13
12
11
10
*
*
**
Controls
I
II
III
IV
Glutathione Peroxidase 3
(GPX3)


Log$_e$ peak area
15
14
13
12
****
***
*
**
Controls
I
II
III
IV
Cartilage Oligomeric
Matrix Protein (COMP)

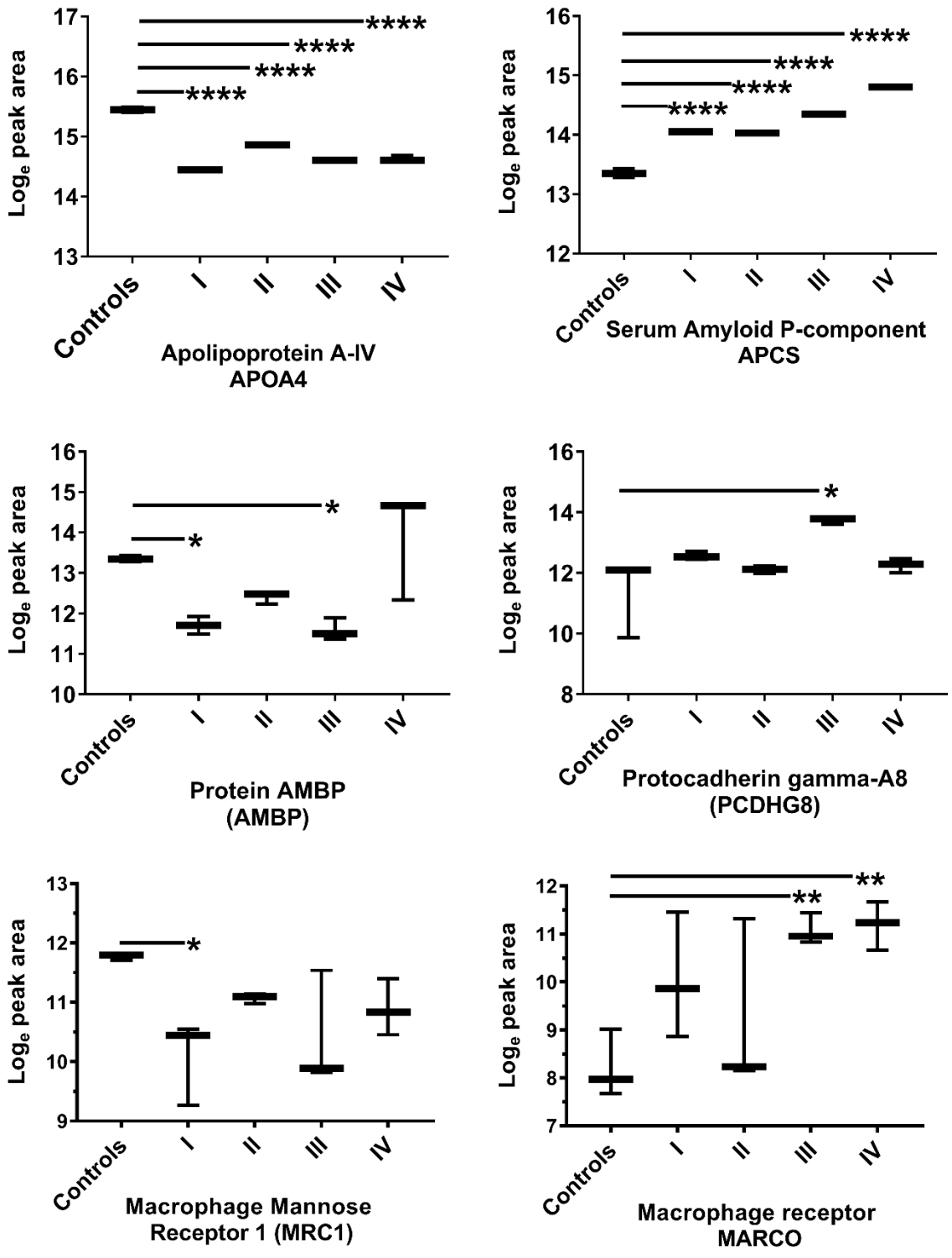


Figure 3.4: Graphical representation of differentially expressed plasma proteins between all CRC stages (I-IV) compared to healthy controls. *Box plots for differentially expressed proteins between healthy control and CRC stages I-IV. * p<0.05, **p<0.005, ***p<0.0005 and ****p<0.0001 calculated using unpaired t-tests.*

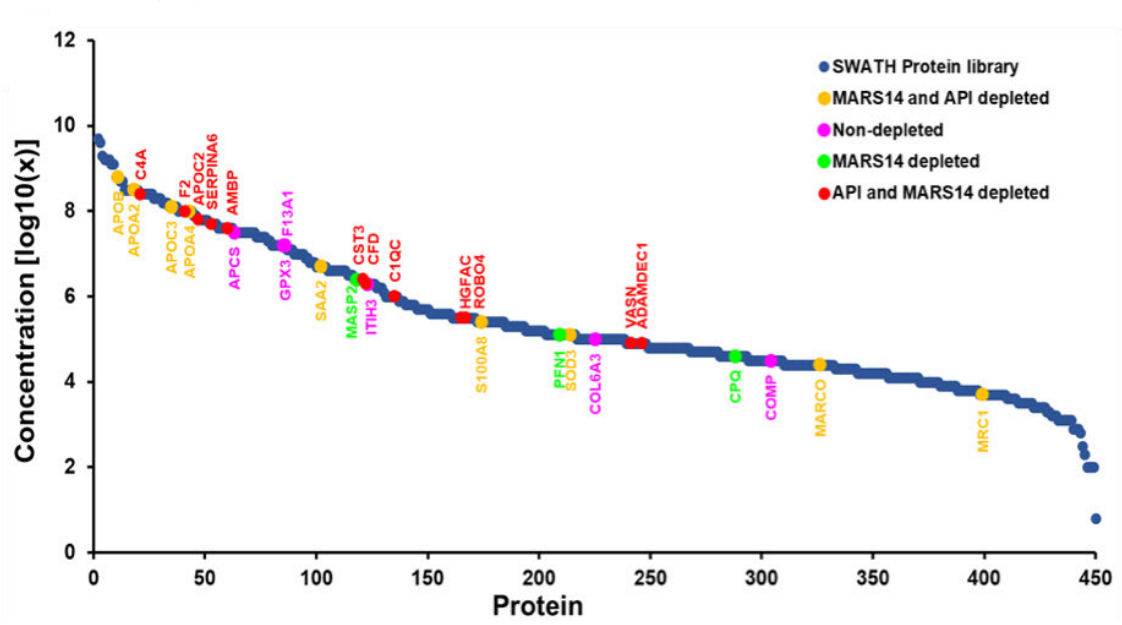


Figure 3.5: Graphical distribution of differentially expressed plasma proteins between all CRC stages (I-IV) compared to healthy controls*: 28 potential candidates identified from four biomarker discovery experiments superimposed on the SWATH™-MS reference library protein concentration curve plotted against protein abundance rank. The color key on the top-right side shows proteins identified from different biomarker discovery experiments.*

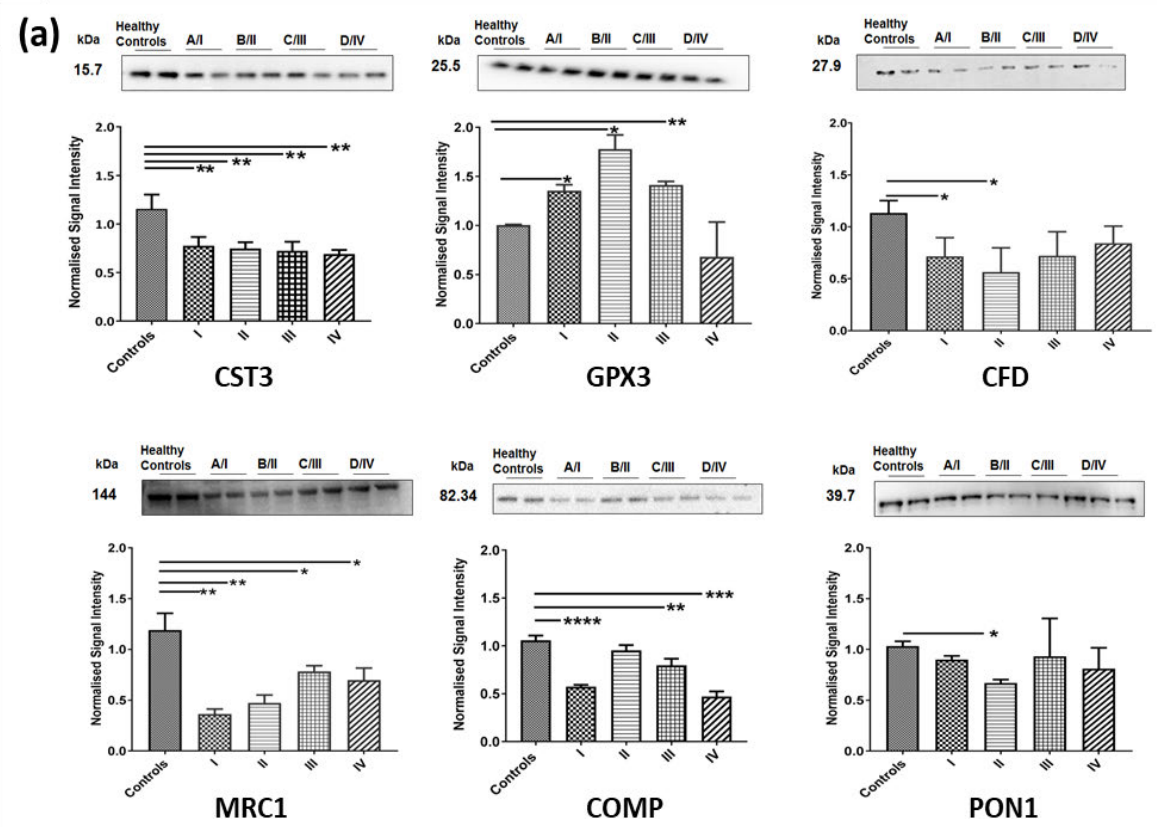
Verification of differentially expressed protein candidates using orthogonal technologies

Selected early stage CRC biomarker candidates were subsequently validated using Western blotting and ELISA. In total, 7 of 37 plasma protein candidates discovered above were available based on previously established biological relevance in cancer, statistical analysis of data and availability of well-established, high-quality antibodies for either Western blotting or ELISA analyses. The expression of 6 proteins (CST3, GPX3, PON1, CFD, COMP and MRC1) was analysed using Western blotting on samples from the pooled healthy and staged (AJCC I-IV) CRC plasma samples (Figure 3.6a) used in the original biomarker discovery study. The expression levels of ADAMDEC1 were measured using a commercially available ELISA kit on the same pooled, as well as the individual (n=100) healthy and staged CRC patient plasma samples (Figures 3.6b & 3.6c).

Consistent with SWATH™-MS results, Western blotting confirmed statistically significant changes in expression levels of CST3, CFD, MRC1, COMP and PON1 were down-regulated in disease plasmas compared to heathy controls. Of these, CST3, MRC1 and COMP levels

were found to be significantly down-regulated in all CRC stages in comparison to healthy, whilst the levels of CFD and PON1 were found to be significantly lower in stage I and/or stage II compared to healthy controls. Equally, GPX3 was shown to be up-regulated in AJCC stages I, II and III compared to heathy plasmas (Figure 3.6a), consistent with SWATH™-MS data for GPX3. Full-length Western blots and Ponceau S Acid Red stained images are shown in Supplementary Figure S3.3. Collectively, expression levels observed in Western blotting for these 6 candidates was consistent with observed SWATH™-MS quantification trends.

ELISA on pooled samples also confirmed that SWATH™-MS expression data for ADAMDEC1, with expression significantly elevated in stage I, II and III CRCs compared to healthy controls (Figure 3.6b). However, when individual patient plasma (n=100) was analysed by ELISA, statistically significant ADAMDEC1 expression level differences ($p < 0.05$) were only found between stage II CRC and healthy control plasma samples (n=20) (Figure 3.6c). ELISA studies on a larger CRC population are in progress to ascertain if ADAMDEC1 SWATH™-MS differences between stage I, II, III, and IV and healthy controls can be substantiated.



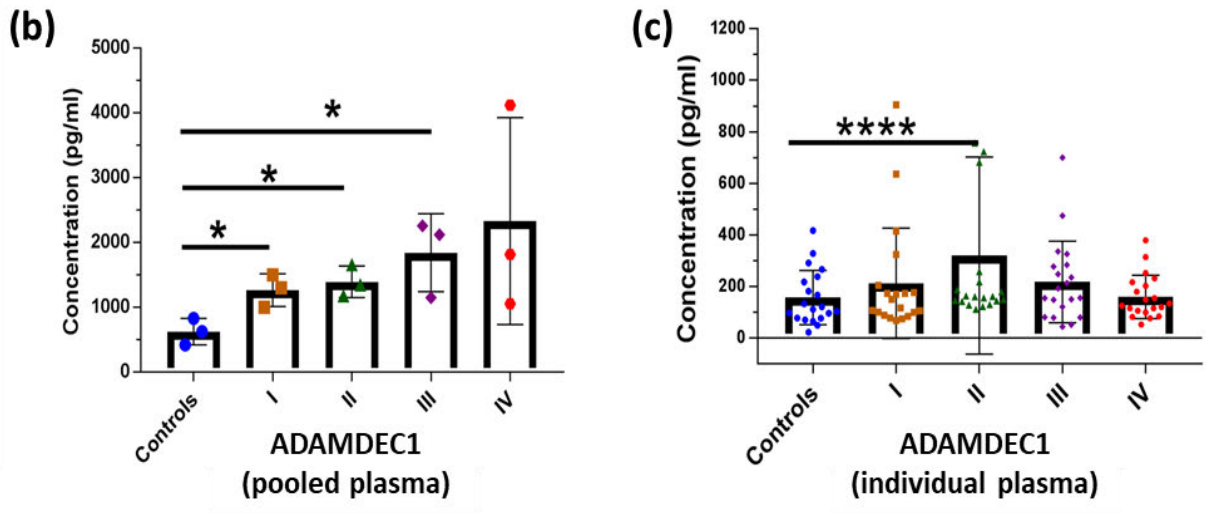


Figure 3.6: *Western blotting and ELISA verification for 7 candidate early-stage CRC plasma protein biomarkers. (a) Panel shows analysis of six biomarker candidates by Western blot and expression level of protein in pooled plasma of all CRC stages (I-IV). (b) ADAMDEC1 ELISA on pooled and (c) individual patients (n=100). The bars indicate the means and SEMs. *p<0.05, **p<0.005, ***p<0.0005 and ****p<0.00005 calculated using unpaired t-test. CST3: Cystatin-C, GPX3: Glutathione peroxidase 3, CFD: Complement factor D, MRC1: Macrophage mannose receptor 1, COMP: Cartilage oligomeric matrix protein, PON1: Serum paraoxonase/arylesterase 1 and ADAMDEC1: ADAM-like decysin 1.*

Neural network-based classification predicts early cancer stage using differentially expressed CRC candidate protein biomarkers

As illustrated above, 37 differentially expressed proteins were identified to discern early stage CRC by SWATH™-MS using pooled plasma samples, rather than individual plasma samples. This approach was used to get stable population values for each stage, but also to limit the enormous cost and time requirement necessary to individually ultradeplete 100 plasma samples in an exploratory study.

An important caveat with the use of exhaustive ultradepletion and peptide fractionation methods is whether candidates identified from pooled SWATH™-MS dataset (technical triplicates of pooled healthy and CRC stages I-IV) are a valid representation of individuals. Extrapolation of pooled data carries inherent risks as the intra- and inter-patient variation of protein candidates is unknown. Being aware of this limitation, a model to test whether each proposed candidate holds statistical power when various noise is added to our pooled data. To overcome this problem, the dataset was synthetically augmented by simulating a large number of hypothetical patients, adding noise far above (up to tenfold) the variance present in our

technical replicates. This data-augmentation made it possible to use state-of-the-art machine-learning based statistical approaches with our dataset to test its stringency.

Before generating synthetic data, the variance of protein concentration from our technical replicates were similar for each stage, which they were (healthy = 33±28%, stage 1 = 36±34%, stage 2 = 42±34%, stage 3 = 45±34% and stage 4 = 31±18%) were verified. A synthetic patient population of a thousand patient per (1000 patients per CRC stage and 1000 healthy subjects) was generated, and application of a conservative variance in protein expression that was 10 times that of the SD of pooled samples in absolute values over a normal distribution around the average response. Of importance, this variance was well above the observed variance of our validated individual concentrations verified by ELISA (Figure 3.6c). This approach gave the possibility to test the widest possible range of protein expression that would expect from a relatively heterogeneous population. At the same time, this approach should prevent overfitting in the training of our algorithm. As can be seen in the dissimilarity matrix per stage, our technical replicates for each CRC stage as well as for the synthetic cohort shows a clear consistency between healthy control and all 4 stages (Figure 3.7a). The distinction between stages also translated well when the data was plotted using the first three dimensions following multi-dimensional scaling, with distances increasing between clusters (healthy and CRC stages) as the disease progresses from an early stage I through to more advanced stage IV.

Subsequently, various supervised classification algorithms to classify each stage separately was trained. Our trained classifier achieved 99.6% correct classification at 10 times the variance for the simulated data used (Figure 3.7b & 3.7c). the deployed algorithm was then verified if it could still properly classify our real dataset which was used to create the synthetic data but completely kept out of the training and achieved 80% correct classification (Figure 3.7d). This is a very encouraging verification of our candidates, and advocates progressing to population cohort studies involving measurement of each of these 37 early stage CRC candidate plasma biomarkers by targeted MRM-based approaches in individual participants to better our predictive model.

The number of proteins necessary were narrowed down to maintain high accuracy. Data mining was performed by examining the dissimilarity distances between proteins rather than in between stages. Five proteins showed clear potential as sufficient to maintain high accuracy, which was further tested. This panel included proteins SAA2, APCS, APOA4, F2 and AMBP. Classification on our synthetic population produced a 94% correct classification from the test dataset (i.e., trained model, Figure 3.7e & 3.7f) and achieved 100% correct classification once deployed on the real pooled samples that were once again kept out of the training of the

algorithm (Figure 3.7g). Importantly, 4 protein candidates (APCS, APOA4, F2 and AMBP) were identified from our in-house ultradepletion experiments (MARS-14→API or API→MARS-14) whilst only 1 candidate (SAA2) was identified from non-ultradepleted experiments. This result clearly indicates the importance of plasma proteomics depth analysis for improved biomarker discovery and shows that a very promising candidates for predicting early occurrence of the pathology.

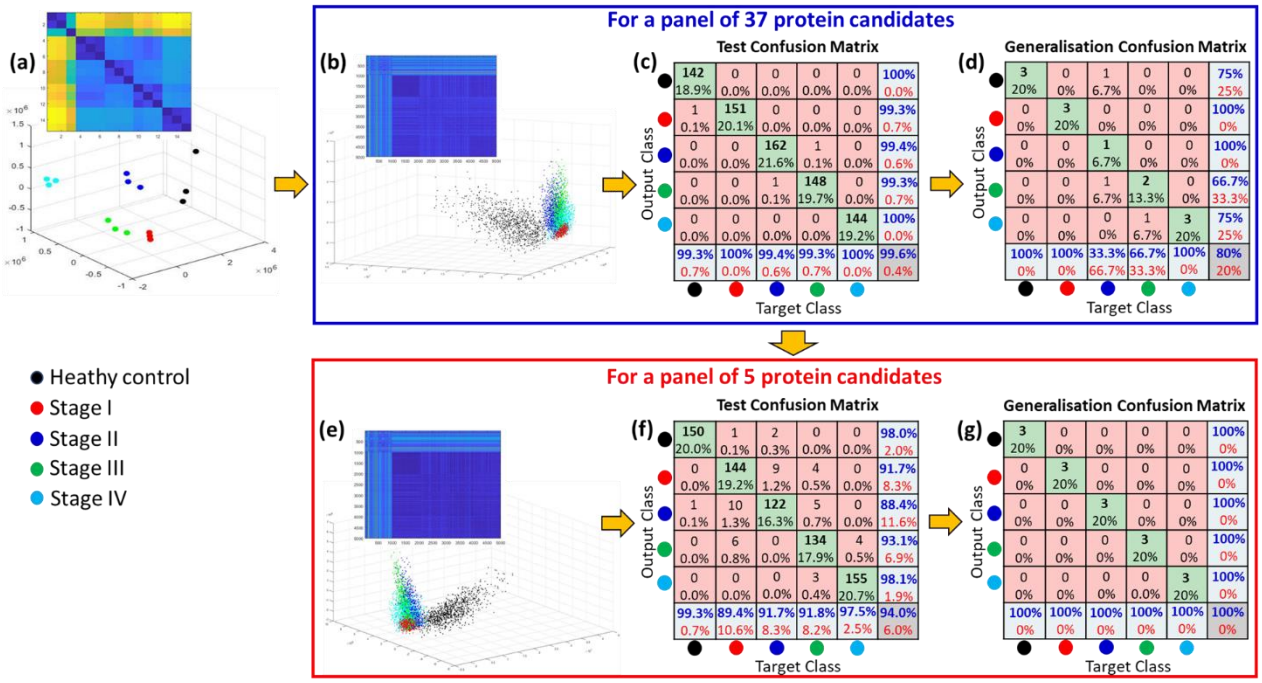


Figure 3.7: A shallow neural network-based classification of synthetic and real datasets with 37 and 5 protein candidates. *(a) The dissimilarity matrix (top left corner) and multi-dimensional scatter (MDS) plot for the triplicates of pooled CRC plasma samples (e.g., healthy control and stages I - IV). (b) The dissimilarity matrix and MDS plot of a synthetic dataset of a panel of 37 protein candidates. A total of 5,000 synthetic patients (1,000 per healthy control and the 4 CRC stages) were created from random numbers falling within a normal distribution of 10 times the standard deviation (SD) of the pooled real CRC plasma samples. (c) Confusion matrix of the synthetic dataset (for 37 protein candidates) for the test phase of the training of the classifier achieved 99.6% success. (d) Confusion matrix for the testing of the classifier on the real dataset kept out of training achieved 80% correct classification. (e) Dissimilarity matrix and MDS plot of the synthetic dataset for a panel of 5 protein candidates (SAA2, APCS, APOA4, F2 and AMB) with a total of 5,000 synthetic patients. (f) Confusion matrix of the synthetic dataset (for 5 protein candidates) for the test phase of the training of the classifier*

achieved 94% success. (g) Confusion matrix for the testing of the classifier on the real dataset kept out of training achieved 100% correct classification.

3.4 Discussion

Early stage diagnosis of CRC has immense actionable curative potential and has been estimated to be able to increase patient survival by >90% (5). Aside from poor compliance (~40%), stool-based testing relies on detection of blood hemoglobin in stool samples, rendering false-positive results from subjects with rectal fissures, hemorrhoids or other ailments where tissue is damaged with consequent bleeding, causing additional burden on health systems due to requisite, unnecessary follow-up colonoscopies (6). In this scenario, blood-based testing would be undisputedly a more reliable, higher compliance (~97%), less invasive and more widely accepted method of screening diagnosis. However, the discovery of reliable biomarkers with high specificity and sensitivity for early stage CRC diagnosis from blood has proven to be challenging.

Comprehensive plasma SWATH™-MS library

The most significant challenge in plasma-based biomarker discovery study is the ability to reliably and accurately measure as many as possible plasma proteins from a single experiment (21). This is complicated by the dominance of many high abundant proteins (HAPs) that mask the identification of more biologically-relevant lower abundance proteins, which may better reflect disease pathophysiology (4). Some antibody-based technologies (e.g., Luminex/Bio-Plex systems (44) have shown some promise, however their high cost has confined discovery to a relatively small number of protein biomarkers. MS-based techniques have made significant recent strides with regards to accuracy and reliability and these, combined with a plethora of analytical techniques (e.g., depletion, ultradepletion, protein/peptide fractionation and IDA) can potentially tackle this challenge.

The use of MARS14 columns to deplete HAPs from plasma samples likely removes some low abundance proteins, unintentionally (43). Nevertheless, this approach is considered as a reliable method for depletion and biomarker discovery (45). Further, extensive fractionations performed on depleted samples are reported to be effective in peptide separation on different tryptic peptides (46-48), and hence resulted in building a comprehensive SWATH™-MS library. Collectively, the multi-fractionation approach covered a broad range of peptide characteristics. As a result, this allowed a total of 513 distinct plasma protein identifications

from combined healthy/CRC plasmas to occur. Moreover, this approach revealed 8 proteins that never previously been identified in plasma. Interestingly, many of these new plasma proteins appeared to be tissue leakage proteins (e.g., ODF3L1, SYN2 and ODF3L1) from organs including brain, testis and ovary, most likely demonstrating these proteins are low to medium abundance in plasma. Furthermore, two proteins of these plasma classified have been previously classified as Human Proteome Project (HPP) 'missing proteins', having a neXtProt protein evidence (PE) level in the PE2-4 range. This illustrates the efficacy of the peptide fractionation method to obtain a plasma snapshot of the human body and by extension of pathophysiology.

Depletion of high abundance proteins has been previously demonstrated to allow identification of lower abundance proteins in human plasma (45). Untargeted proteomic analyses using current LC-MS/MS on MARS-14-depleted plasma do not efficiently reveal many low abundance, disease-specific biomarkers from human plasma (32). The reason for this detection disparity may be due to the particularly steep distribution of protein abundance seen with plasma versus cell proteomes (21). To overcome this problem, an in-house "ultradepletion" method that immunodepletes additional high and medium abundance human plasma proteins (30, 31) has been developed and used here for the first time. However, for quantification purposes, some inconsistencies have been reported (49). To circumvent these issues, a multipronged strategy was applied for reliable protein quantitation. Here either MARS-14 alone, or an ultradepletion strategy with either API or MARS-14 was used in tandem.

These approaches widened the quantifiable plasma proteome by an additional 129 proteins which were predominantly low to medium abundance proteins, demonstrated by on a plasma protein Anderson curve. This study unravelled proteins like MEGF8, CRISP3 and TRIM33 that are known to occur in lower picogram levels in plasma. Of these TRIM33 is known to be a negative regulator of BMP signalling as well as a regulator of TGF-β receptor signaling (42), whist MEGF8 and CRISP-3 are found expressed on extracellular exosomes and are integral component of plasma membranes. These low abundance proteins sit in in lowest section of the Anderson concentration curve belonging to G-protein coupled receptors, Notch family, interleukins, integrin beta chain family members, α and β-transferins, homeobox proteins and zinc finger proteins. Further, proteins like proprotein convertase 9, C-C motif chemokine 16, SPARC-like protein, ADAMT's like protein 4, macrophage receptor, IgG Fc-binding protein, Golgi membrane protein 1 and ADAMDEC1 were mapped for their tissue-specific expression to colon, small intestine, epithelia and lymph nodes. These proteins are known to be involved

in apoptosis, immune response, cell metabolism, cell differentiation and dendritic cell maturation, respectively.

Revealed known potential CRC biomarkers

It was however not surprising to note that the subset of 37 early stage CRC differentially expressed protein biomarkers identified through this study were observed across the entire range of concentrations represented by the Anderson curve. A number of biomarker studies have previously had similar aims to this study, albeit using different samples and analytical techniques. This thesis recapitulated a number of these studies that lends credence to the validity of our approach and suggest that these markers may indeed have significance.

The list of differentially expressed proteins comprised many acute phase response proteins or those involved in the complement cascade. A number of these have been previously reported to be markers of CRC, including serum paraoxonase 1 (PON1), down-regulated in CRC plasma here as well as in other investigations (50). PON1 is a known free radical scavenger possessing antioxidant activities and has been reported to play an important role in CRC carcinogenesis and metastasis (51). Paradoxically, activity of sera PON1 has been demonstrated to be increased in patients with CRC (52), suggesting that a decrease in protein levels may not necessarily be associated with decreased activity, though the authors do propose larger studies needed to be performed to validate their claims.

Plasma is the richest reserve of secretory proteins that potentially reflect abnormal physiology. Unsurprisingly, aberrations in several secretory proteins with relevance to tumour pathophysiology was discovered. The most frequently recurring marker protein was S100A8 (53, 54) found to be elevated in our study. S100A8 is predominantly expressed in myeloid cells and has been identified as a serological marker for CRC in combination with S100A9 (54). Interestingly, Ichikawa *et al*., suggest S100A8/A9 promotes activation of MAPK and NF-kappa B signaling pathways and mediates tumour development. (54, 55).

Another previously established up-regulated marker glutathione peroxidase (GPX3) was found, an extracellular selenoprotein member known to play important roles in oxidative stress-induced apoptosis (56). Barett *et al.,* had previously demonstrated that elevated plasma GPX3 may serve protective roles in inflammation-associated colon carcinogenesis by reducing oxidative DNA damage (32). However, Roman *et al*., reported no significant differences between CRC and healthy control levels of serum GPX3 (57) although they were unable to

validate these findings with orthogonal techniques. In our study, GPX3 was elevated across all CRC stages compared to healthy plasmas. Due to this apparent discrepancy with literature reports, Western blotting was used to validate GPX3 expression which confirmed our SWATH™-MS results.

Apolipoproteins A4 (ApoA-IV) and Apolipoprotein B, both small intestine and duodenum specific proteins also stood out in the data. A recently published study established that aberrant ApoA-IV expression in CRC patients was associated with 8q24 oncogenic SNPs and with diabetes mellitus (DM) with suggestion that this protein may subsequently facilitate CRC development (58). In our study ApoA-IV levels across all CRC stages were found to be significantly down-regulated in comparison to healthy controls consistent with past genomic studies (58). On the other hand, elevated levels of Apo B in serum have previously been associated with CRC risk in a study performed on 28,098 participants, out of which incidence cases were identified in follow-up done from 1991-2012 with a 95% confidence interval (59). This correlated with data from this study where ApoB levels were found to be significantly up-regulated across all CRC stages compared to healthy control plasmas.

A subset of biomarkers emanating from this study have been shown to be expressed in multiple cancer tissue types, including CRC. For example, cystatin C (CST3) is a secretory protein known to be a potent cathepsin B (CTSB) inhibitor (60). It is thought that CTSB participates in remodeling of connective tissues during tumour growth, invasion and metastasis (61). This study found CST3 down-regulated in CRC stages, whereas a number of studies have associated up-regulation of CST3 associated with progression of cancer (62). Several studies have suggested CST3 is not reliable, proposing alternatively that prognostic value lies in disturbances in CTSB/CST3 ratios (50, 60, 63). Nevertheless, data here validated down-regulated levels of CST3 finding significant fold change between all CRC stages and healthy controls. However, subsequent detailed statistical modelling indicated that CST3 did not add particular value in classifying CRC tumour stage. The link between uPAR and CSTB, both being proteases is certainly intriguing and worth investigating further as both are known to be significantly up-regulated and associated with poor outcomes from CRC metastasis (64).

Novel CRC biomarkers

Plasma proteins are largely secreted by liver and tissues through which they circulate (21, 24). In the panel of early CRC stage candidates, it was interesting to observe changes in proteins specifically expressed in colon and associated intestinal mucosal lining tissues. Of such

proteins, one interesting candidate was ADAMDEC1 which is selectively expressed and shed by maturing dendritic cells and macrophages predominantly in the small intestine, caecum and large intestine (65, 66). ADAMDEC1, a disintegrin and metalloprotease, is a particularly unique member of ADAM family in that it lacks a transmembrane domain which allows it to remain soluble (67). It is one of four ADAM's released from thrombin-stimulated platelets and cleaves cell surface pro-epidermal growth factor (pro-EGF) at an arginine residue to generate soluble high-molecular weight EGF (HMW-EGF) (67). HMW-EGF is an effective ligand for EGF receptor members and ultimately triggers the EGF signal transduction pathway (67). A more recent study found ADAMDEC1 up-regulated in normal epithelial cells, specifically after these normal cells had been co-cultured with active mutant RasV12-transformed epithelial cells (68). This study suggested that ADAMDEC1 may be an epithelial intrinsic soluble factor that promotes apical extrusion of RasV12 cells, displaying anti-tumour activity, in a phenomenon called "epithelial defence against cancer" (68). In both studies, increased level of ADAMDEC1 was demonstrated to play a crucial role in tumour division and progression. Here, up-regulated levels of plasma ADAMDEC1 in all CRC stages compared to healthy controls was observed and this trend was confirmed by ELISA performed on both pooled and individual patient (n=20 per CRC stage) plasmas. This study of individual patient plasma samples allowed us to investigate the impact of "pooling" plasma samples in the first place, necessary to complete technical protocols within a reasonable grant timeframe. Although, pooling had advantages in discovery (discussed earlier), extrapolating protein biomarker information to individual patient populations based on that pooled data is counterintuitive. Therefore, ADAMDEC1 was used as a "example" protein to investigate the efficacy of extrapolation of pooled data for the complete list of all 37 candidates. Individual ADAMDEC1 SD values were then used to inform cut-offs for the generation of a machine learning algorithm as discussed.

Another novel finding of this study was identification of a subset of immune system protein biomarkers. Any human body harbouring tumours likely initiates assault on physiological wellbeing. Cells of the immune system continually monitor tissues and provide protection against many types of pathology, including monitoring tumourigenesis (69). Macrophage receptor (MARCO), a scavenger receptor is expressed by suppressive tumour-associated macrophages (TAM) called M2 macrophages. These are known to suppress the immune system favouring tumour growth and promoting metastasis through pro-angiogenesis and tissue remodelling (70). Interestingly, Georgoudaki *et al.*, showed targeting MARCO-expressing TAM's enhance the effect of immune checkpoint therapy in both melanoma and CRC (69). Macrophages are recruited to the tumour via blood circulation or direct immigration to adjacent

tumours from surrounding tissues which might explain the elevated plasma levels of MARCO observed here across all CRC stages. Considerable increases in fold change ratio in later stages (C/III and D/IV) could be the result of immune suppression accelerating clinical tumour growth and metastasis. Another immune regulatory protein**,** macrophage mannose receptor 1 (MRC1) also known as CD206 is an M2 marker and has been found to be co-expressed with MARCO in CRC cell lines by Georgoudaki *et al.*, (69). A study on advanced imaging agents found that MRC1/CD206 a C-type lectin mannose receptor is a major binding receptor for γ-tilmanocept - a compound routinely used for molecular imaging and mapping of sentinel lymph nodes (71). In our study, MRC1 was observed to be down-regulated in all CRC stages compared to healthy controls. In contrast, a previous study found that MRC1 was up-regulated in CRC (72). These data were based on a discovery cohort of only three patients, although the data were validated using ELISA in 96 CRC patients., Importantly, the samples analysed by Fan NJ et. al, were not clinically staged and were a mixed cohort, whilst our study incorporated all four stages of CRC in well-defined cohort as discussed in section3.2. It is also important to note that plasma samples were used in our study while Fan NJ et. al, used serum samples in their study. While it is difficult to speculate on the reasons for this discrepancy in MRC1 expression data, future studies should directly compare serum and plasma levels in all stages of CRC.

Predictive neural network classification reveals a subset of potential biomarkers for early CRC detection

Though ultradepletion of pooled CRC-staged plasmas allowed increased analytical depth and identification of novel low abundance proteins, it can also be a limitation if the overall endgame is to generate tangible, predictive models for high-throughput diagnosis. Machine-learning approaches are becoming more mainstream for proteomics studies (35). These methods are often ill-suited for analysis of limited datasets from demanding, economic and person-hour resource-intensive proteomics studies (e.g., where ultradepletion is performed). In a proof-of-concept experiment, a synthetic patient population to train a classification algorithm was generated and then tested this on real patient samples. An algorithm assuming pooled plasma samples was trained which represented a centroid around which a normal distribution of biomarker concentrations would reside. This hypothetical variance present in human plasma protein concentration needs to be conservative, as high variability even occurs between among twins over time (73). Our supposed variance considered:

1. variance between individuals over time and environmental factors;

2. variance between technologies employed keeping in mind high throughput testing on a population scale is our long-term aim, and
3. variance amongst clinical stages of CRC.

It is important to note that our choice of potential biomarkers was stringent and based upon orthogonal, complementary approaches with consideration of a reasonable biological rationale. With these restrictions in mind, a SD as high as 10 times the SD from the mean for our generated population and maintaining a near perfect classification on disease stages with our 37 candidates was implemented. High classification rates remained with as low as 5 of our proteins of interest. Therefore, this panel of 5 candidates is proposed as highly interesting for potential predictive purposes, and now propose to replace these generated samples with biological ones as a larger patient population dataset (individual targeted protein assays) over time. Of interest regarding the richness of selected biomarkers, progression of CRC from stage I to IV resulted in increased separation distance between stages from healthy to stage IV CRC. This fits very well with a narrative that would be expected as the condition of patients deteriorate, and biological manifestation of cancer increases.

Next steps

A review of PubMed confirms that most biomarker studies do not result in biomarkers entering clinical practice, and this is primarily due to the fact that most candidate markers do not meet stringent specificity and sensitivity criteria (62). Recent publications have promoted the idea of multi-variate biomarker panels (75, 76) as being more efficacious than single markers.

In our study, potential biomarker candidates derived from multiple depletion and peptide fractionation with SWATH™-MS were first selected with unbiased statistical analyses, established biological roles in oncology were further considered for prioritising the candidates. After further verification using immuneassays, the following candidates were proposed as putative early stage CRC markers: ADAMDEC1, MARCO, MRC1, S100A8, ApoAIV, GPX3, COMP, PON1 and CFD. The diagnostic utility of these protein panels needs to be validated by measuring expression in individual healthy and staged CRC patient samples using either or both immunological and targeted mass spectrometry technologies. The first step would be to develop a first-pass parallel reaction monitoring (PRM) assay for absolute and relative quantification of these candidates in clinically staged pooled plasma samples. Once a reproducible and robust assay is developed, it could be used to quantitate and evaluate the specificity and sensitivity of the panel of candidates to differentiate healthy individuals from early stage CRC patients on using both current and a larger, independent validation cohort. In

addition, ELISA-based assays can also be used to quantify potential candidates in individual samples.

This study then applied a machine learning model (predictive neural network classification) to narrow the panel down to 5 proteins that maintained accuracy in identifying early (I/II) stage CRC. This 5-protein marker panel (i.e. SAA2, APCS, APOA4, F2 and AMB) is a subset of the 37 candidates identified from discovery experiments. The predictive power of the 5-protein marker panel will need to be validated as outlined above, but this aspect of my study highlights the power of predictive neural network modelling, which was used for the very first time in MS-based biomarker discovery.

3.5 Conclusions

MS-based proteomics in combination with depletion strategies have the potential to identify multiple protein targets in human plasma. Unfortunately, not many markers identified in the laboratory every reach the clinic and it is essential that putative biomarkers are examined in larger patient cohorts and benchmarked against current screening methods. The 37 candidates identified in this thesis are a statistically filtered list of proteins, with no biological hypothesis underlying their selection. The potential biological role of 9 of these 37 candidates in CRC pathogenesis has been discussed in section 3.4 under "Novel CRC biomarkers." For example, candidates such as MARCO and MRC1 are immune markers, which are known to suppress the immune system and favour tumour growth by promoting metastasis (69,70). Both candidates were found to be differentially expressed in control vs CRC patient plasma samples. Another important hypothesis regards ADAMDEC1 which shows expression limited to the intestinal and colon regions. The up-regulated expression of ADAMDEC1 was demonstrated to play a crucial role in tumour division and progression in multiple studies (67,68). Our data confirmed the increased levels of ADAMDEC1 in CRC patient plasma. Similarly, many of the proteins identified in this study, have previously been identified as standalone biomarkers or as a member of a panel (e.g. PON1, GPX3 and ApoB; see Section 3.4 for more information), but unfortunately, none of these have entered the clinic. In addition, the machine learning model identified a 5-protein marker panel, from the 37 potential candidates, that accurately discriminated early-stage CRC from healthy controls. In conclusion, in our study, the SWATH™-MS analysis revealed 37 protein biomarker candidates and a predictive neural network narrowed down 37 candidate list to 5 proteins that maintained accuracy to discriminate

early stage (I/II) from healthy controls. This panel of 5 protein candidates consist of both upregulated and downregulated proteins. The statistical strength of the panel is based on the combined trends of all 5 biomarkers taken together. This 5-protein marker panel will be the focus of future work and will be tested for specificity and sensitivity in larger patient population datasets.

3.6 Limitations

Plasma is representative of the physiological state of an individual (24). Therefore, biomarkers emerging from plasma may not be specific for the target disease, but rather may reflect different undiagnosed pathologies. For example, elevated lactate dehydrogenase is associated with melanoma, but also liver disease and kidney tissue damage (82,83). For this reason, multi-variate panels of biomarkers are preferred to a single biomarker molecule (77). There are many examples that combination of markers provides superior sensitivity and specificity in biological assays (78,84). For example, Ova1 (a multi-variate test) performs better than CA-125 (a single biomarker) in diagnosing ovarian cancer (84).

There are > 20,000 protein coding genes and 14,500 diseases classified by the ICD code, and it is likely that different diseases will share proteins implicated in their pathogenesis (24). Thus, to identify highly specific, unique plasma biomarkers, it is vital to use appropriate negative controls. This could include patients with early, benign disease or a different disease of the same organ. In this case, an ideal negative control to test the specificity of these biomarkers would have been plasma samples derived from patients with benign polyps or other benign ailments of the GI tract. (78). Thus, although the valuable controls of benign disease were not included, the approach of defining a multi-variate diagnostic biomarker panel may circumvent this limitation (79-81).

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394-424.

2. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut. 2017;66(4):683-91.

3. Gonzalez-Pons M, Cruz-Correa M. Colorectal Cancer Biomarkers: Where Are We Now? Biomed Res Int. 2015; 2015:149014.

4. Makhoul R, Alva S, Wilkins KB. Surveillance and Survivorship after Treatment for Colon Cancer. Clinics in colon and rectal surgery. 2015;28(4):262-70.

5. Brenner H, Jansen L, Ulrich A, Chang-Claude J, Hoffmeister M. Survival of patients with symptom- and screening-detected colorectal cancer. Oncotarget. 2016;7(28):44695-704.

6. Wolf AMD, Fontham ETH, Church TR, Flowers CR, Guerra CE, LaMonte SJ, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. CA: a cancer journal for clinicians. 2018.

7. National Bowel Cancer Screening Program. In: Health Do, editor. Australia: Australian Government; 2016.

8. New National Bowel Cancer Screening Program Test Kit: Department of Health, Australian Government 2018 [Available from: http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/NBCSP-test-kit.

9. Dickinson BT, Kisiel J, Ahlquist DA, Grady WM. Molecular markers for colorectal cancer screening. Gut. 2015;64(9):1485-94.

10. Adler A, Geiger S, Keil A, Bias H, Schatz P, deVos T, et al. Improving compliance to colorectal cancer screening using blood and stool-based tests in patients refusing screening colonoscopy in Germany. BMC gastroenterology. 2014; 14:183.

11. Song LL, Li YM. Current noninvasive tests for colorectal cancer screening: An overview of colorectal cancer screening tests. World J Gastrointest Oncol. 2016;8(11):793-800.

12. Pedersen SK, Symonds EL, Baker RT, Murray DH, McEvoy A, Van Doorn SC, et al. Evaluation of an assay for methylated BCAT1 and IKZF1 in plasma for detection of colorectal neoplasia. BMC Cancer. 2015; 15:654.

13. Cohen JD, Javed AA, Thoburn C, Wong F, Tie J, Gibbs P, et al. Combined circulating tumour DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. Proc Natl Acad Sci U S A. 2017;114(38):10202-7.

14. Lowes LE, Bratman SV, Dittamore R, Done S, Kelley SO, Mai S, et al. Circulating Tumour Cells (CTC) and Cell-Free DNA (cfDNA) Workshop 2016: Scientific Opportunities and Logistics for Cancer Clinical Trial Incorporation. Int J Mol Sci. 2016;17(9).

15. Mahboob S, Ahn SB, Cheruku HR, Cantor D, Rennel E, Fredriksson S, et al. A novel multiplexed immunoassay identifies CEA, IL-8 and prolactin as prospective markers for Dukes' stages A-D colorectal cancers. Clinical proteomics. 2015;12(1):10.

16. Fletcher RH. Carcinoembryonic antigen. Ann Intern Med. 1986;104(1):66-73.

17. Grunnet M, Sorensen JB. Carcinoembryonic antigen (CEA) as tumour marker in lung cancer. Lung cancer (Amsterdam, Netherlands). 2012;76(2):138-43.

18. Tang S, Zhou F, Sun Y, Wei L, Zhu S, Yang R, et al. CEA in breast ductal secretions as a promising biomarker for the diagnosis of breast cancer: a systematic review and meta-analysis. Breast cancer (Tokyo, Japan). 2016;23(6):813-9.

19. Nikolaou S, Qiu S, Fiorentino F, Rasheed S, Tekkis P, Kontovounisios C. Systematic review of blood diagnostic markers in colorectal cancer. Techniques in coloproctology. 2018;22(7):481-98.

20. Nowsheen S, Aziz K, Panayiotidis MI, Georgakilas AG. Molecular markers for cancer prognosis and treatment: have we struck gold? Cancer letters. 2012;327(1-2):142-52.

21. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics. 2002;1(11):845-67.

22. Paulovich AG, Whiteaker JR, Hoofnagle AN, Wang P. The interface between biomarker discovery and clinical validation: The tar pit of the protein biomarker pipeline. Proteomics Clinical applications. 2008;2(10-11):1386-402.

23. Wu TL, Sun YC, Chang PY, Tsao KC, Sun CF, Wu JT. Establishment of ELISA on 384-well microplate for AFP, CEA, CA 19-9, CA 15-3, CA 125, and PSA-ACT: higher sensitivity and lower reagent cost. J Clin Lab Anal. 2003;17(6):241-6.

24. Geyer PE, Holdt LM, Teupser D, Mann M. Revisiting biomarker discovery by plasma proteomics. Mol Syst Biol. 2017;13(9):942.

25. Ma H, Chen G, Guo M. Mass spectrometry-based translational proteomics for biomarker discovery and application in colorectal cancer. Proteomics Clinical applications. 2016;10(4):503-15.

26. Xu W, Hu Y, Li J, He X, Fu Z, Pan T, et al. Study of distinct serum proteomics for the biomarker's discovery in colorectal cancer. Discovery medicine. 2015;20(110):239-53.

27. Yamamoto T, Kudo M, Peng WX, Takata H, Takakura H, Teduka K, et al. Identification of aldolase A as a potential diagnostic biomarker for colorectal cancer based on proteomic analysis using formalin-fixed paraffin-embedded tissue. Tumour biology: the journal of the International Society for Oncodevelopmental Biology and Medicine. 2016;37(10):13595-606.

28. Dudley JT, Butte AJ. Identification of discriminating biomarkers for human disease using integrative network biology. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2009:27-38.

29. Tu C, Rudnick PA, Martinez MY, Cheek KL, Stein SE, Slebos RJ, et al. Depletion of abundant plasma proteins and limitations of plasma proteomics. J Proteome Res. 2010;9(10):4982-91.

30. Tan SH, Mohamedali A, Kapur A, Baker MS. Ultradepletion of human plasma using chicken antibodies: a proof of concept study. J Proteome Res. 2013;12(6):2399-413.

31. Tan SH, Mohamedali A, Kapur A, Lukjanenko L, Baker MS. A novel, cost-effective and efficient chicken egg IgY purification procedure. Journal of immunological methods. 2012;380(1-2):73-6.

32. Wei Ning KW, Kristina E.Hill, Caitlyn W.Barrett, Lori A.Coburn, Raymond F.Burk, Christopher S.Williams. Gpx3 is a Tumour Modifier in Murine Inflammatory Carcinogenesis. Gastroenterology. 2011;140(5).

33. Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, 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.

34. Wu JX, Song X, Pascovici D, Zaw T, Care N, Krisp C, et al. SWATH™-MS Mass Spectrometry Performance Using Extended Peptide MS/MS Assay Libraries. Mol Cell Proteomics. 2016;15(7):2501-14.

35. Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. Omics: a journal of integrative biology. 2013;17(12):595-610.

36. Kelchtermans P, Bittremieux W, De Grave K, Degroeve S, Ramon J, Laukens K, et al. Machine learning applications in proteomics research: how the past can boost the future. Proteomics. 2014;14(4-5):353-66.

37. Wong SC, Gatt A, Stamatescu V, McDonnell MD. Understanding data augmentation for classification: when to warp? In Digital Image Computing: Techniques and Applications (DICTA). 2016; International Conference on:1-6. IEEE.

38. DeVries T, Taylor GW. Dataset Augmentation in Feature Space. ICLR Workshops. 2017;URL: https://arxiv.org/abs/1702.05538v1.

39. Govaert E, Van Steendam K, Willems S, Vossaert L, Dhaenens M, Deforce D. Comparison of fractionation proteomics for local SWATH™-MS library building. Proteomics. 2017;17(15-16).

40. Baker MS, Ahn SB, Mohamedali A, Islam MT, Cantor D, Verhaert PD, et al. Accelerating the search for the missing proteins in the human proteome. Nature communications. 2017; 8:14271.

41. Schwenk JM, Omenn GS, Sun Z, Campbell DS, Baker MS, Overall CM, et al. The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. J Proteome Res. 2017;16(12):4299-310.

42. Massague J. TGFbeta in Cancer. Cell. 2008;134(2):215-30.

43. Pernemalm M, Lehtio J. Mass spectrometry-based plasma proteomics: state of the art and future outlook. Expert review of proteomics. 2014;11(4):431-48.

44. Houser B. Bio-Rad's Bio-Plex(R) suspension array system, xMAP technology overview. Archives of physiology and biochemistry. 2012;118(4):192-6.

45. Yadav AK, Bhardwaj G, Basak T, Kumar D, Ahmad S, Priyadarshini R, et al. A systematic analysis of eluted fraction of plasma post immunoaffinity depletion: implications in biomarker discovery. PloS one. 2011;6(9): e24442.

46. Stein DR, Hu X, McCorrister SJ, Westmacott GR, Plummer FA, Ball TB, et al. High pH reversed-phase chromatography as a superior fractionation scheme compared to off-gel isoelectric focusing for complex proteome analysis. Proteomics. 2013;13(20):2956-66.

47. Liu X, Pohl CA. Comparison of reversed-phase/cation-exchange/anion-exchange trimodal stationary phases and their use in active pharmaceutical ingredient and counterion determinations. Journal of chromatography A. 2012; 1232:190-5.

48. Irvine GB. High-performance size-exclusion chromatography of peptides. Journal of biochemical and biophysical methods. 2003;56(1-3):233-42.

49. Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. Nature. 2008;452(7187):571-9.

50. Surinova S, Choi M, Tao S, Schuffler PJ, Chang CY, Clough T, et al. Prediction of colorectal cancer diagnosis based on circulating plasma proteins. EMBO molecular medicine. 2015;7(9):1166-78.

51. Rosenblat M, Elias A, Volkova N, Aviram M. Monocyte-macrophage membrane possesses free radicals scavenging activity: stimulation by polyphenols or by paraoxonase 1 (PON1). Free radical research. 2013;47(4):257-67.

52. Afsar CU, Gunaldi M, Okuturlar Y, Gedikbasi A, Tiken EE, Kahraman S, et al. Paraoxonase-1 and arylesterase activities in patients with colorectal cancer. International journal of clinical and experimental medicine. 2015;8(11):21599-604.

53. Li S, Xu F, Li H, Zhang J, Zhong A, Huang B, et al. S100A8(+) stroma cells predict a good prognosis and inhibit aggressiveness in colorectal carcinoma. Oncoimmunology. 2017;6(1): e1260213.

54. Kim HJ, Kang HJ, Lee H, Lee ST, Yu MH, Kim H, et al. Identification of S100A8 and S100A9 as serological markers for colorectal cancer. J Proteome Res. 2009;8(3):1368-79.

55. Ichikawa M, Williams R, Wang L, Vogl T, Srikrishna G. S100A8/A9 activate key genes and pathways in colon tumour progression. Molecular cancer research: MCR. 2011;9(2):133-48.

56. Kayanoki Y, Fujii J, Islam KN, Suzuki K, Kawata S, Matsuzawa Y, et al. The protective role of glutathione peroxidase in apoptosis induced by reactive oxygen species. Journal of biochemistry. 1996;119(4):817-22.

57. Marco Roman PJ, Marco Agostini, Giulio Cozzi, Salvatore Pucciarelli, Donato Nitti, Chiara Bedin , Carlo Barbante Serum seleno-proteins status for colorectal cancer screening explored by data mining techniques - a multidisciplinary pilot study. Microchemical Journal. 2012; 105:124-32.

58. Sugimachi K, Yamaguchi R, Eguchi H, Ueda M, Niida A, Sakimura S, et al. 8q24 Polymorphisms and Diabetes Mellitus Regulate Apolipoprotein A-IV in Colorectal Carcinogenesis. Annals of surgical oncology. 2016;23(Suppl 4):546-51.

59. Borgquist S, Butt T, Almgren P, Shiffman D, Stocks T, Orho-Melander M, et al. Apolipoproteins, lipids and risk of cancer. International journal of cancer. 2016;138(11):2648-56.

60. Yan Y, Zhou K, Wang L, Wang F, Chen X, Fan Q. Clinical significance of serum cathepsin B and cystatin C levels and their ratio in the prognosis of patients with esophageal cancer. OncoTargets and therapy. 2017; 10:1947-54.

61. Fraile JM, Ordonez GR, Quiros PM, Astudillo A, Galvan JA, Colomer D, et al. Identification of novel tumour suppressor proteases by degradome profiling of colorectal carcinomas. Oncotarget. 2013;4(11):1931-2.

62. Zhang J, He P, Zhong Q, Li K, Chen D, Lin Q, et al. Increasing Cystatin C and Cathepsin B in Serum of Colorectal Cancer Patients. Clinical laboratory. 2017;63(2):365-71.

63. Wang H, Gao L, Meng C, Yu N, Yang F, Zhang C, et al. Serum Cystatin C Level Is Not a Promising Biomarker for Predicting Clinicopathological Characteristics of Bladder Urothelial Tumours. Biomed Res Int. 2018; 2018:2617439.

64. Herszenyi L, Istvan G, Cardin R, De Paoli M, Plebani M, Tulassay Z, et al. Serum cathepsin B and plasma urokinase-type plasminogen activator levels in gastrointestinal tract cancers. European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP). 2008;17(5):438-45.

65. Lund J, Olsen OH, Sorensen ES, Stennicke HR, Petersen HH, Overgaard MT. ADAMDEC1 is a metzincin metalloprotease with dampened proteolytic activity. The Journal of biological chemistry. 2013;288(29):21367-75.

66. O'Shea NR, Chew TS, Dunne J, Marnane R, Nedjat-Shokouhi B, Smith PJ, et al. Critical Role of the Disintegrin Metalloprotease ADAM-like Decysin-1 [ADAMDEC1] for Intestinal Immunity and Inflammation. Journal of Crohn's & colitis. 2016;10(12):1417-27.

67. Chen R, Jin G, McIntyre TM. The soluble protease ADAMDEC1 released from activated platelets hydrolyzes platelet membrane pro-epidermal growth factor (EGF) to active high-molecular-weight EGF. The Journal of biological chemistry. 2017;292(24):10112-22.

68. Yako Y, Hayashi T, Takeuchi Y, Ishibashi K, Kasai N, Sato N, et al. ADAM-like Decysin-1 (ADAMDEC1) is a positive regulator of Epithelial Defense Against Cancer (EDAC) that promotes apical extrusion of RasV12-transformed cells. Scientific reports. 2018;8(1):9639.

69. Georgoudaki AM, Prokopec KE, Boura VF, Hellqvist E, Sohn S, Ostling J, et al. Reprogramming Tumour-Associated Macrophages by Antibody Targeting Inhibits Cancer Progression and Metastasis. Cell reports. 2016;15(9):2000-11.

70. Lamagna C, Aurrand-Lions M, Imhof BA. Dual role of macrophages in tumour growth and angiogenesis. Journal of leukocyte biology. 2006;80(4):705-13.

71. Azad AK, Rajaram MV, Metz WL, Cope FO, Blue MS, Vera DR, et al. gamma-Tilmanocept, a New Radiopharmaceutical Tracer for Cancer Sentinel Lymph Nodes, Binds to the Mannose Receptor (CD206). Journal of immunology (Baltimore, Md: 1950). 2015;195(5):2019-29.

72. Fan NJ, Chen HM, Song W, Zhang ZY, Zhang MD, Feng LY, et al. Macrophage mannose receptor 1 and S100A9 were identified as serum diagnostic biomarkers for colorectal cancer through a label-free quantitative proteomic analysis. Cancer biomarkers: section A of Disease markers. 2016;16(2):235-43.

73. Liu Y, Buil A, Collins BC, Gillet LC, Blum LC, Cheng LY, et al. Quantitative variability of 342 plasma proteins in a human twin population. Mol Syst Biol. 2015;11(1):786.

74. Das V, Kalita J, Pal M. Predictive and prognostic biomarkers in colorectal cancer: A systematic review of recent advances and challenges. Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie. 2017; 87:8-19.

75. Menon U, Ryan A, Kalsi J, Gentry-Maharaj A, Dawnay A, Habib M, et al. Risk Algorithm Using Serial Biomarker Measurements Doubles the Number of Screen-Detected Cancers Compared With a Single-Threshold Rule in the United Kingdom Collaborative Trial of Ovarian Cancer Screening. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2015;33(18):2062-71.

76. Rho JH, Ladd JJ, Li CI, Potter JD, Zhang Y, Shelley D, et al. Protein and glycomic plasma markers for early detection of adenoma and colon cancer. Gut. 2018;67(3):473-84.

77. Mazzara, Saveria, et al. "CombiROC: an interactive web tool for selecting accurate marker combinations of omics data." *Scientific reports* 7 (2017): 45477.

78. Zhang, Fan, et al. "A neural network approach to multi-biomarker panel discovery by high-throughput plasma proteomics profiling of breast cancer." *BMC proceedings*. Vol. 7. No. 7. BioMed Central, 2013.

79. Surinova, Silvia, et al. "Prediction of colorectal cancer diagnosis based on circulating plasma proteins." *EMBO molecular medicine* 7.9 (2015): 1166-1178.

80. Cristobal, Alba, et al. "Personalized proteome profiles of healthy and tumour human colon organoids reveal both individual diversity and basic features of colorectal cancer." *Cell reports* 18.1 (2017): 263-274.

81. Hao, Jian-Jiang, et al. "Comprehensive Proteomic Characterization of the Human Colorectal Carcinoma Reveals Signature Proteins and Perturbed Pathways." *Scientific reports* 7 (2017): 42436.

82. Weide, B., et al. "Serum markers lactate dehydrogenase and S100B predict independently disease outcome in melanoma patients with distant metastasis." *British journal of cancer*107.3 (2012): 422.

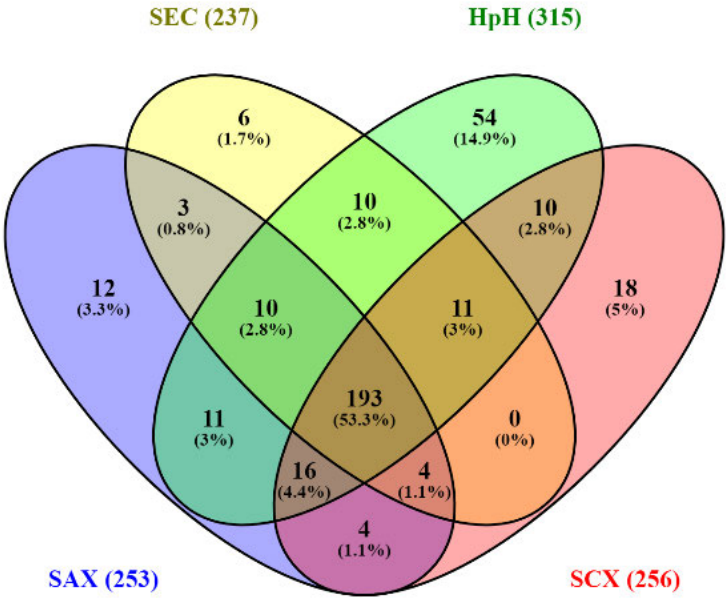
83. Kohn, Mary, and Harry Ross. "Lactate dehydrogenase output of the excised kidney as an index of acute ischaemic renal damage." *Transplantation* 11.5 (1971): 461-464.

84.Nolen, Brian M., and Anna E. Lokshin. "Biomarker testing for ovarian cancer: clinical utility of multiplex assays." *Molecular diagnosis & therapy* 17.3 (2013): 139-146.

Supplementary Information

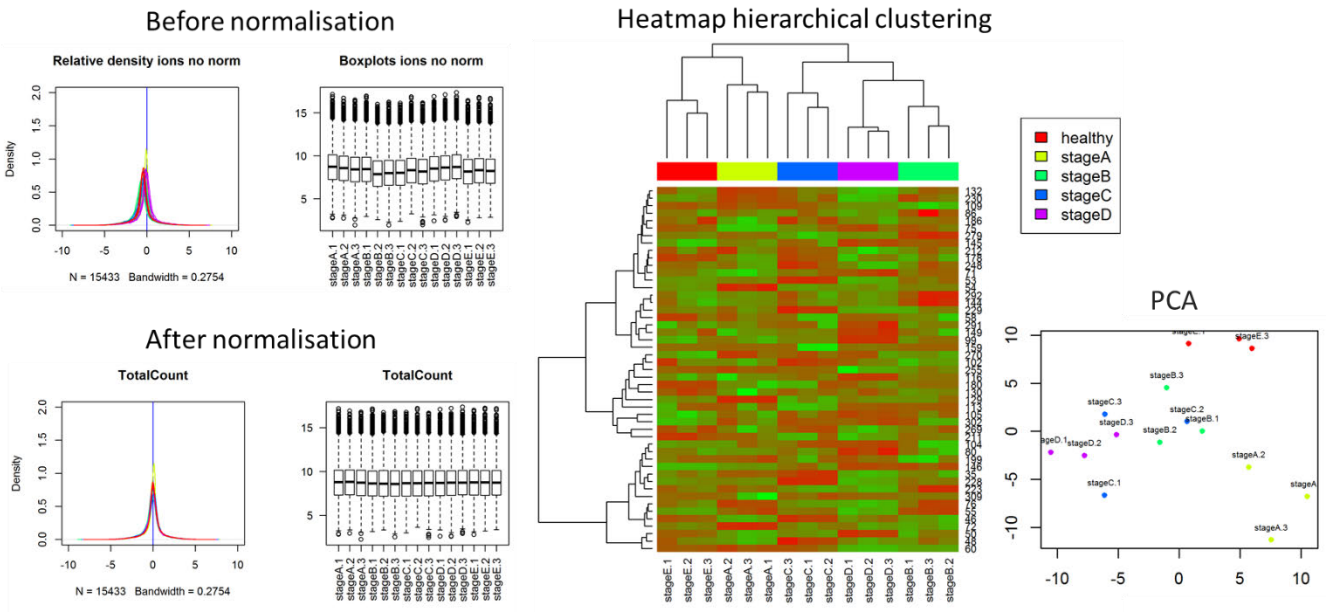
Supplementary Figures S3.1-S3.4

Supplementary Tables S3.1-S3.5

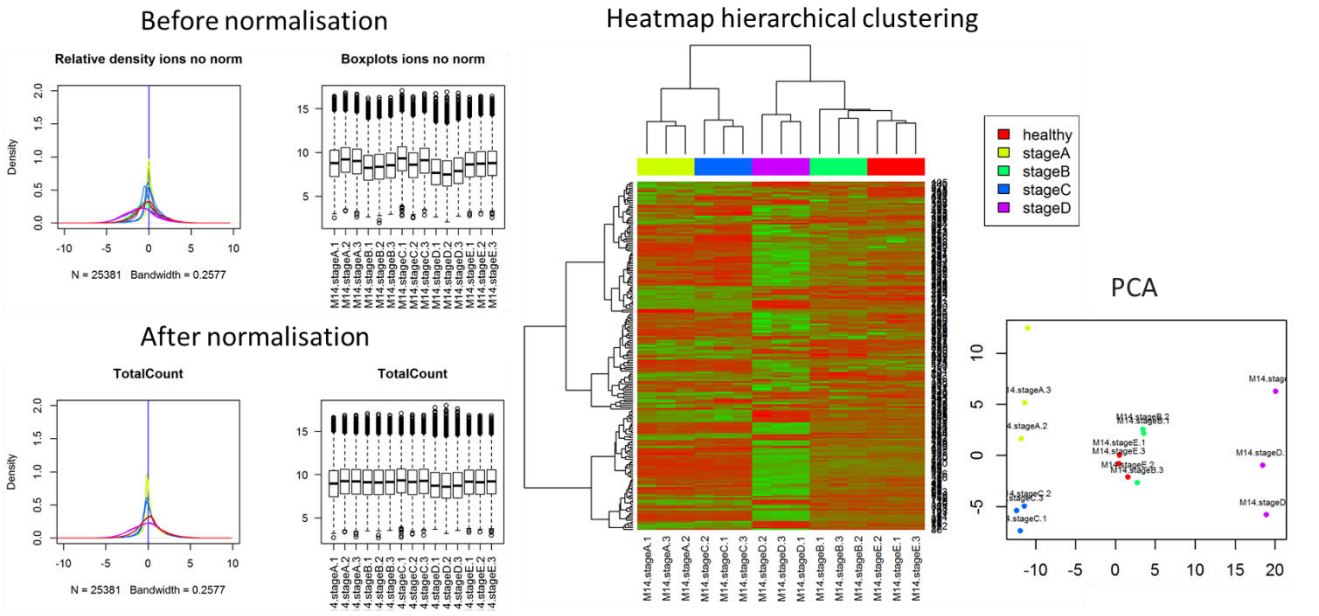


Supplementary Figure S3.1: Venn diagram comparison of number of common, unshared and shared identified proteins (containing ≥2 uniquely mapping non-nested peptides of amino acid length ≥9) between four peptide fractionation methods. **HpH**: High pH C18 reversed phase, **SEC**: Size exclusion chromatography, **SAX**: Strong anion exchange, **SCX**: Strong cation exchange.

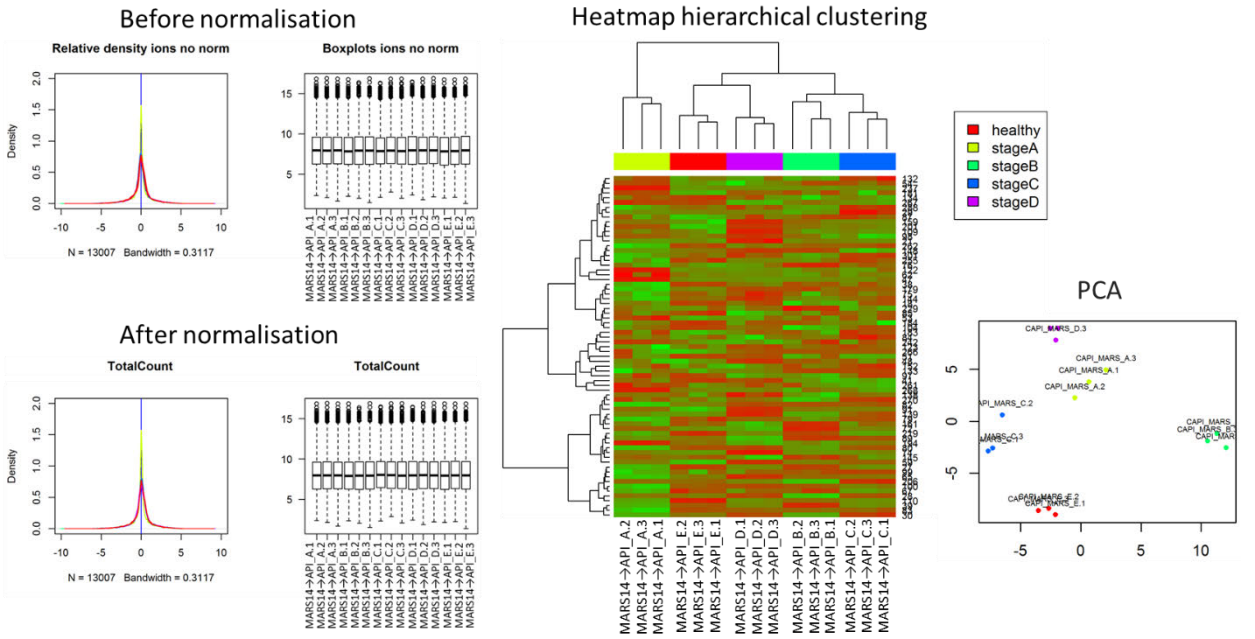
(a) Non-depleted plasma



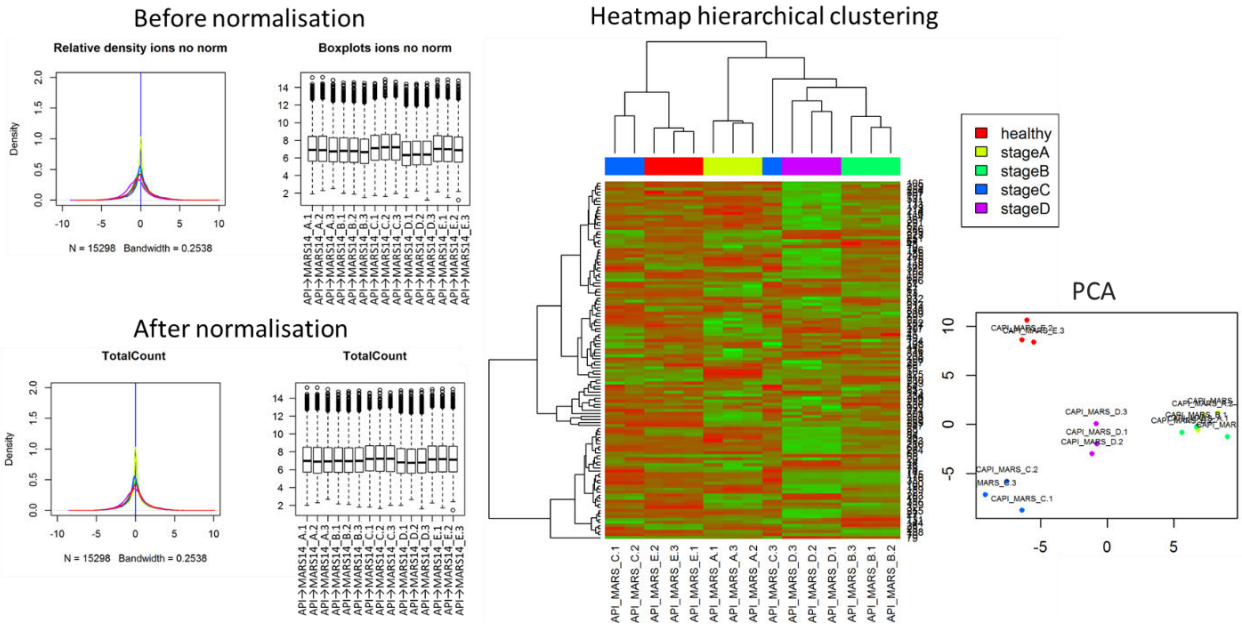
(b) MARS-14 depleted plasma



(c) MARS14→API depleted plasma

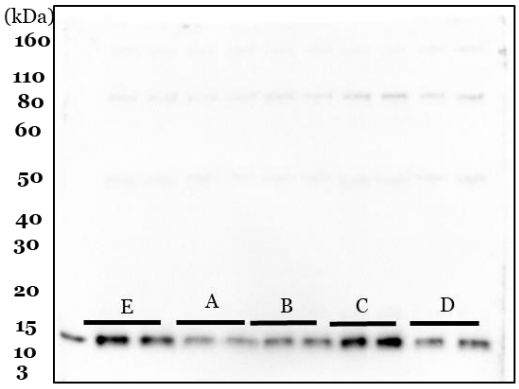


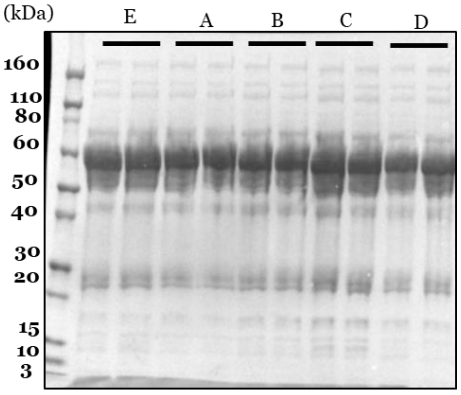
(d) API→MARS14 depleted plasma



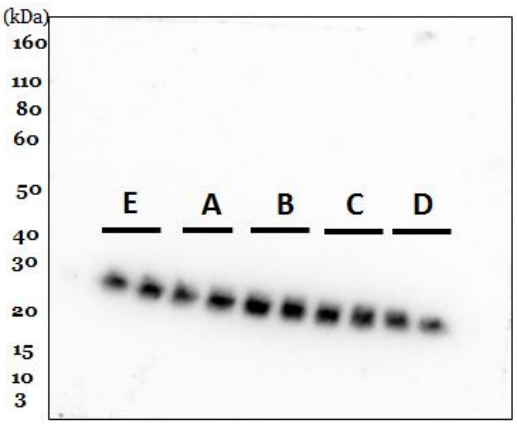
Supplementary Figure S3.2: Extracted SWATH™-MS dataset from **(a)** non-depleted, **(b)** MARS-14 depleted, **(c)** MARS-14→API depleted **(d)** API→MARS-14 depleted experiments were independently normalised using total area normalisation. The data distribution was examined using density plots and boxplots. The consistency of the sample replication was examined visually using heatmap hierarchical clustering and PCA plots. Healthy controls represented as stage E in the images. PCA: Principal component analysis.

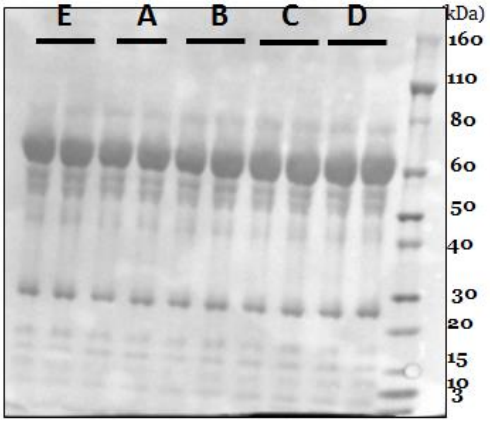
(a) CST3, Molecular weight: 15.7 kDa





(b. i) GPX3, Molecular weight: 25.5 kDa





(b. ii)* GPX3, Molecular weight: 25.5 kDa

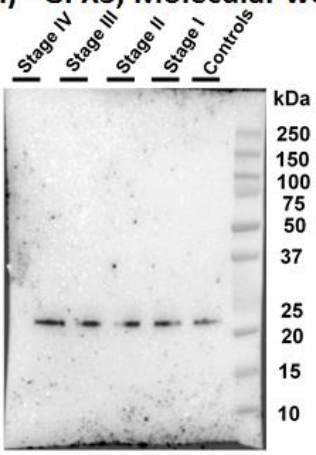
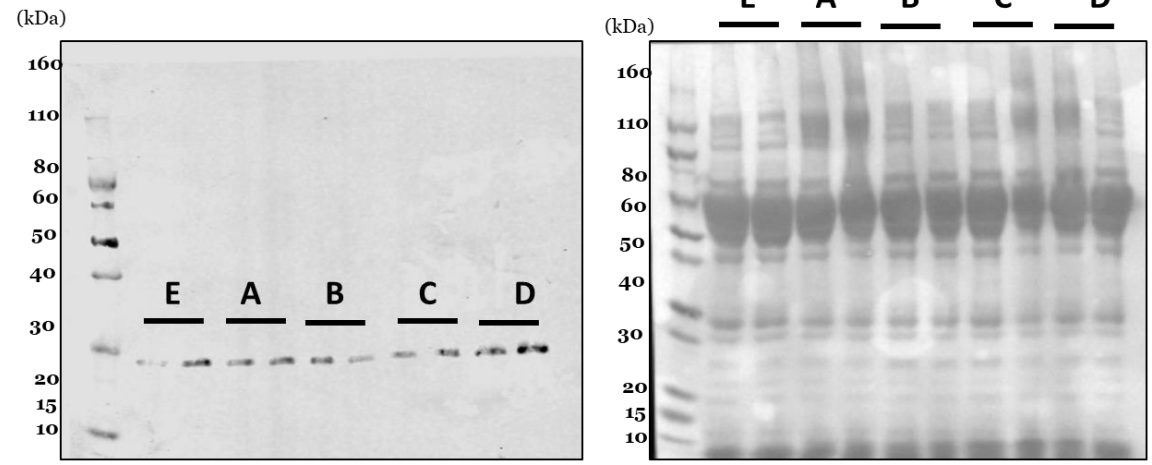
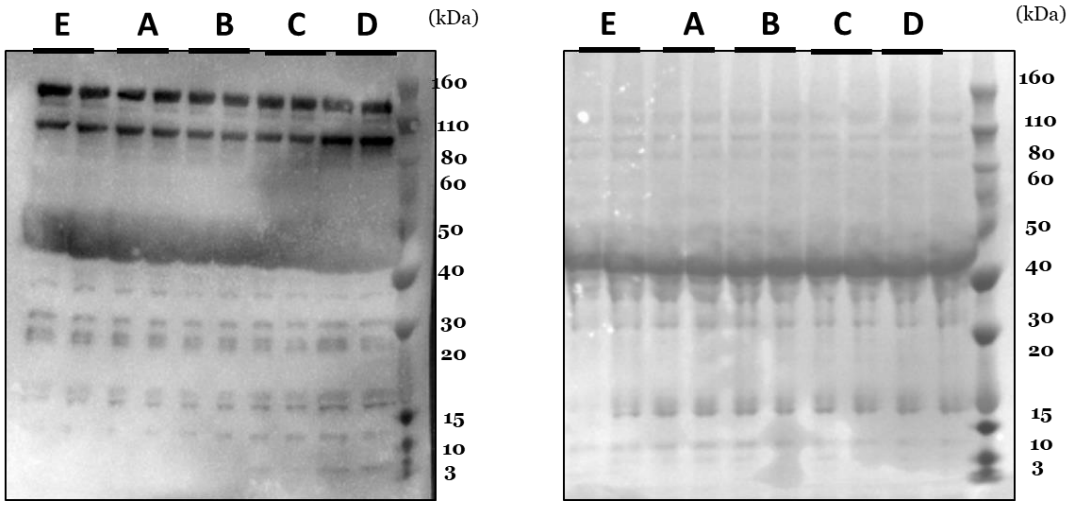


*Figure S3 (b. ii), a replicate gel run for GPX3 western blot is an evidence to support that variation observed in Figure S3 (b. i) is likely an artefact of the gel run.

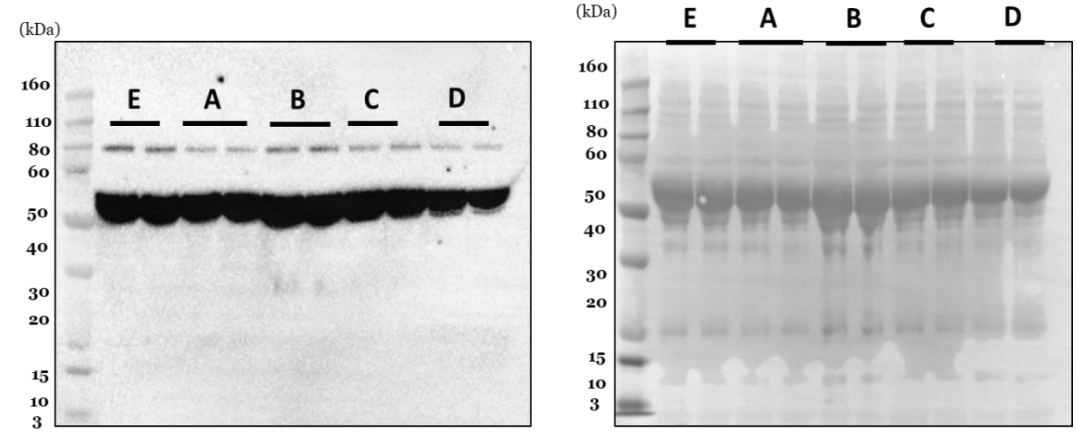
(c) CFD, Molecular weight: 27.9 kDa

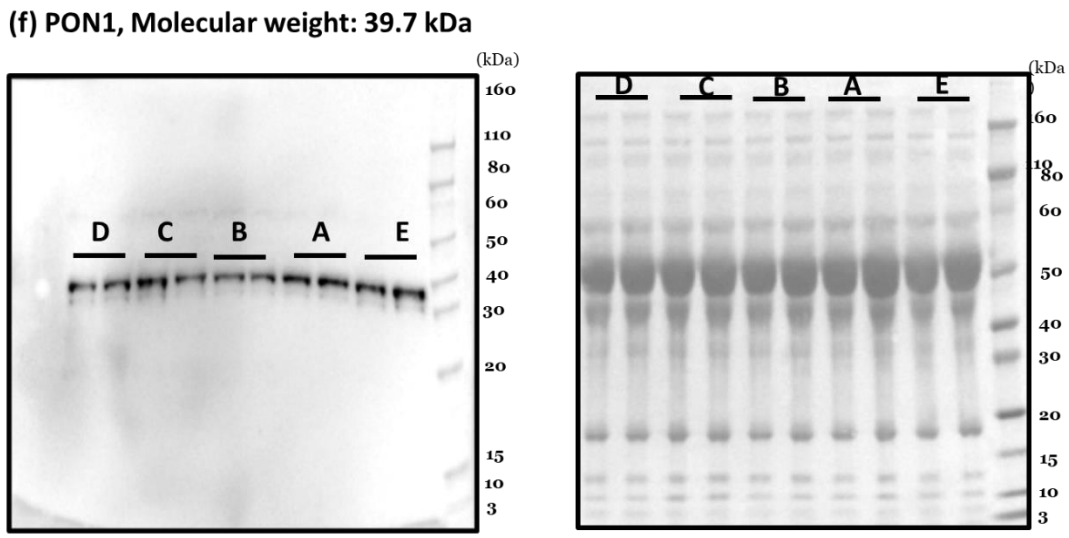


(d) MRC1, Molecular weight: 144 kDa



(e) COMP, Molecular weight: 82.34 kDa





Supplementary Figure S3.3: Western blotting images (left) and Ponceau stained images (right) of **(a)** Complement factor D, CST3, **(b)** Glutathione peroxidase 3, GPX3, **(c)** Complement factor D, CFD, **(d)** Macrophage mannose receptor 1, MRC1, **(e)** Cartilage oligomeric matrix protein, COMP and **(f)** Serum paraoxonase/arylesterase 1, PON1.

Supplementary Tables S3.1-S3.5

- Supplementary Table S3.1: Peptides/proteins identified using different fractionation methods (HpH, SEC, SAX, SCX)

- Supplementary Table S3.2: Gene Ontology (GO) functions of proteins found in our plasma SWATH™-MS library

- Supplementary Table S3.3: 8 proteins identified/observed in plasma for the very first time via proteomics.

- Supplementary Table S3.4: Quantifiable plasma proteins and peptides captured by non-depletion and depletion strategies

- Supplementary Table S3.5: Clinical Details of CRC Patients

The mass spectrometry raw data has been submitted on PRIDE Archive - proteomics data repository (https://www.ebi.ac.uk/pride/archive/) and this information would be provided as soon as it is available.

For the purpose of review of this thesis, these Supplementary Tables S3.1-S3.4 have been uploaded separately as supplementary data.

Table S3.5: Clinical Details of CRC patients				
AJCC staging **(n= 80)**	**I** **(20)**	**II** **(20)**	**III** **(20)**	**IV** **(20)**
Age				
Median ± SD	65 ± 7.2	70 ± 7.9	65 ± 9.0	62 ± 8.0
Sex				
Male	66.7%	53.3%	46.7%	60%
Female	22.2%	46.7%	53.3%	40%
Location of tumour/cancer				
Sigmoid	6 (40%)	6 (40%)	11	8 (53.3%)
Low Rectal	2 (13.3%)	0	1 (6.7%)	1 (6.7%)
Caecal	2 (13.3%)	3 (20%)	3 (20%)	3 (20%)
Ascending colon	2 (13.3%)	2 (13.3%)	0	1 (6.7%)
Transverse colon	2 (13.3%)	4 (26.7%)	0	0
Descending colon	0	0	0	2 (13.3%)
Adenoma	1 (6.7%)	0	0	0
Metastasis/Location				
Lymph Node	0	0	15	0
Liver	0	0	0	8
Gall Bladder/Lung	0	0	0	1
Ovary	0	0	0	1
Other Colonic Regions	0	0	0	5

Chapter 4

Verification of a multi-analyte signature assay for early diagnosis using Parallel Reaction Monitoring (PRM) assay

Abstract

To support the aim of developing a new early stage CRC diagnosis blood test, the combination of mass spectrometry and ultradepletion was utilised to explore plasma at unprecedented depth. With this technical advantage, 9 protein candidates ADAMDEC1, MARCO, MRC1, S100A8, ApoAIV, GPX3, COMP, C1QC and CFD were identified and prioritised. After successful orthogonal verification of 7 of these candidates, I proceeded to develop and accurately measure the 9 candidates in plasma using targeted proteomics parallel reaction monitoring (PRM) assays. Targeted MS-based approaches are widely used to quantitate proteins. More specifically, the parallel PRM-based methods performed using orbitrap-quadrupole instrument provide high-resolution via ion trapping capabilities and have been proven a stronger method than other conventional triple quadrupole instruments used specifically for targeted proteomics (Bourmaud et al., 2016). In this study, first-pass PRM assays were optimised to perform relative quantification of two protein candidates ADAMDEC1 and CFD on pooled plasma. This first-pass assay confirmed the response profile of ADAMDEC1 and CFD as was observed in SWATH™-MS data of Chapter 3. This first-pass assay will further be characterised using CPTAC guidelines and finally will be used to test the efficiency of these candidates in individual patient plasma as part of future biomarker discovery studies with fresh CRC and healthy cohorts. The overarching aim of this study is to evaluate the diagnostic potential of the identified candidates from biomarker discovery experiments in Chapter 3 using a PRM-based assay in plasma samples.

State of Research

This study is ongoing and will require further assay characterisation experiments to be performed in accordance with relevant CPTAC guidelines, in the hope that an assay can be deployed into testing of individual samples and in various laboratories. The details of the needed experiments have been described in detail in Future Experiment section of this Chapter.

4.1 Introduction

The successful development of a multi-analyte biomarker assay comprises of three broad phases, the identification of potential protein candidates, verification and prioritisation of candidates, validation before clinical approval (Paulovich et al., 2008). MS dependent biomarker discovery experiments use combination of depletion, multiple fractionations and liquid chromatography to relatively quantify hundreds of proteins. The identified proteins are hypothesis free and simply based on peptide detection and observed signal intensity of those peptides (Geyer et al., 2017). The identified peptides are carefully prioritised on basis of statistical analysis such as 1% FDR at the peptide level, number of unique peptides observed and fold change differences (±1.5) as was done in our study (chapter 3). Even after stringent statistical analysis there appears to be no shortage of these 'discovered' potential candidate proteins. These biomarkers/potential candidates must be validated on larger patient cohorts to confirm their ability to differentiate between disease stage from healthy controls (Paulovich et al., 2008).

It is not surprising that most of the existing FDA approved blood-tests are dominated by enzyme assays and immune-assays (Chau et al., 2008). Immune-assays such as ELISA and Western blotting (WB) have been considered as standard modalities for detecting protein abundance in biospecimens (Chau et al., 2008). In terms of translation, both these techniques have proved to be powerful standalone methodologies applied to clinics (as tests for HIV, hepatitis, pregnancy, allergen, Lyme disease, syphilis, autoimmune disorders and numerous others) (Hosseini S., 2018). When speaking of protein biomarker research, Western blotting and ELISA can be considered as worthwhile standard orthogonal checks on the validity of primary MS discovery phase data (Chapter 2). It is worthwhile to note that whilst these have been integral to bioscience for over three decades, they have some short comings. Towbin and colleagues, in their landmark paper describing the methodology for Westerns, acknowledge its limitation in accurate quantitation due to suboptimal transfer of proteins from gel to membrane (Towbin et al., 1979).

In retrospect, ELISA and Western blotting can be limited by one or more of the following; (i) linear dynamic range, (ii) limit of detection, (iii) ability to multiplex, and (iv) reproducibility (Aebersold et al., 2013; Towbin et al., 1979). These limitations can be significantly ameliorated

using quantitative assays like single reaction monitoring (SRM) or parallel reaction monitoring assay (PRM) or targeted assay, as discussed in detail in Chapter 2.

The workflow of a targeted assay relies on multiple parameters like retention time, the mass-to-charge ratio of the precursor ion and selected fragment ions of the targeted peptide, and the relative signal intensities of the detected fragment (transition) signals (Ong and Mann, 2005). Using multiplexed data acquisition techniques and software tools, these parameters can be assessed and scored to determine the probability of a targeted peptide being present in the sample (Peterson et al., 2012). The relative presence of hundreds of proteins can be statistically evaluated in a single injection using targeted workflows, reproducibly, making it far superior to ELISA (Michaud et al., 2018; Parker et al., 2014). It is for these reasons that the proteomics community uses targeted assays to validate quantitative bottom-up MS data.

In these targeted assays, multiple pre-selected peptides of candidate proteins are measured in healthy controls and disease plasma with excellent inter and intra laboratory reproducibility (Addona et al., 2009). Moreover, the broad applicability of these assays can be used to measure these proteins in CRC disease patient plasma and distinguish them from healthy controls. Biospecimens are often spiked with heavy labelled target peptides to compare the transitions to enable absolute quantification of a protein in the sample. Such peptides can be purchased commercially (e.g. Australian Peptides, Synpeptide Co., Ltd., OriGene Technologies, Inc.).

However, the mass spectrometer does not equally ionise, separate and detect all peptides.
The PRM methodologies enable the measurement of proteins with concentration as low as 2-10 ng/ml in non-depleted and unfractionated plasma (Rauniyar, 2015), providing the measurement of protein in its innate form (Michaud et al., 2018). The PRM assays can be executed with a simple needle's prick, and requires only microlitres of plasma (Michaud et al., 2018). Targeted MS has been proposed to bridge the gap between discovery and clinical validation (Harlan and Zhang, 2014).

In our previous study (Chapter 3), 9 protein candidates ADAMDEC1, MARCO, MRC1, S100A8, ApoAIV, GPX3, COMP, C1QC and CFD were identified and prioritised based upon statistical data and literature study. Orthogonal technologies such as immune assays evaluated and confirmed the expression of proteins identified from our SWATH™-MS data (Chapter 3). The aim of this study is to develop, and measure target peptides derived from the 9 protein

candidates that were identified from previous SWATH™-MS studies using the same cohort of 100 **individual** patient plasma samples to evaluate diagnostic utility. The study discussed here is a preliminary study. Out of 9 candidates, the peptide targets for CFD and ADAMDEC1 were evaluated in pooled plasma samples as a first-pass assay using a workflow shown in Figure 4.1.

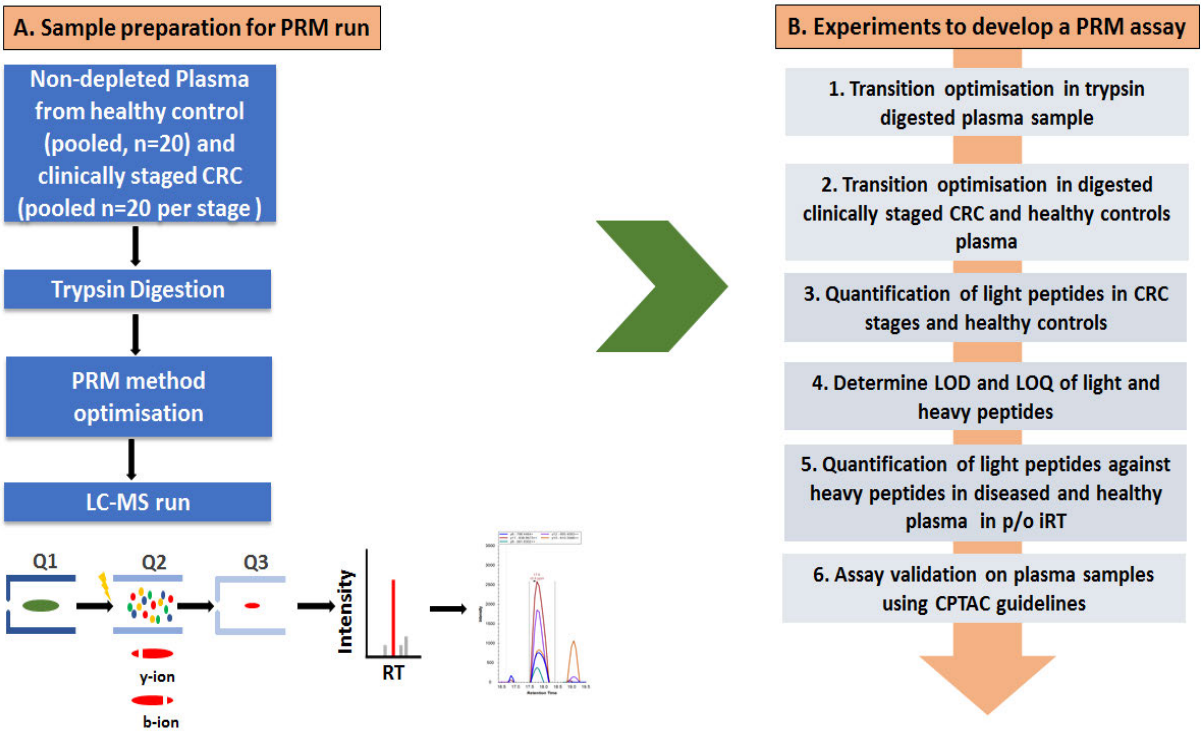


Figure 4.1: The basic workflow to develop a PRM assay *A) Steps for sample preparation before LC-MS runs for targeted PRM. B) Step wise experiments needed to develop a targeted assay after method optimisations*

4.2 Material and Methods

Plasma collection and sample preparation

Plasma samples from 100 individuals, comprising of 80 clinically staged colorectal cancer (I-IV) and 20 healthy controls were procured from the Victorian Cancer Biobank, Melbourne under approved human ethics (HREC Ref #5201600401 and #5201700681). Plasma samples were collected prior to colonoscopy and detailed clinical/pathophysiology reports information of individual, their inclusion/exclusion criteria have been discussed in detail in research article (Mahboob et al., 2015).

PRM assays

Sample preparation for liquid chromatography-parallel reaction monitoring (LC-PRM)

Plasma protein concentration was determined by a bicinchoninic acid (BCA) assay kit (Pierce TM #23225) according to the manufacturer's protocol. Plasma tryptic digests for 100 individual samples and pooled set of plasma (5 samples, n=20 pooled per staged (I-IV) and healthy controls) were prepared by diluting 5µl of non-depleted plasma in ratio of 1:5 with 1X PBS. Denaturation of cysteine bonds in proteins samples was performed using 5mM dithiothreitol for 30 min at 60°C. Reduced plasma samples were alkylated for 30 min at 37°C in dark with 10mM iodoacetamide. Final tryptic digestion was performed using sequencing grade trypsin (Promega, Madison,WI) was added to reduced and alkylated samples in ratio of 20:1 for 16h at 37°C. Tryptic digested samples were desalted using C-18 custom made zip-tips prior to MS-analysis. The sample elution was done using 50% V/V acetonitrile and 0.1% of formic acid. The eluted samples were lyophilised and stored at -80°C. Prior to PRM analysis, lyophilised samples were reconstituted in 0.1% formic acid.

Potential biomarkers identified by SWATH™-MS plasma proteomics

The target protein candidates selected for optimisations used in this study were originally identified from SWATH™-MS biomarker discovery experiment described in detail in previous chapter of this thesis (Sharma *et al*, Chapter 3). In summary, from the biomarker discovery experiments, 37 potential candidates were identified as early stage CRC detection markers. CRC early. Two candidates were selected for PRM assay development i.e. CFD and ADAMDEC1 which were also successfully validated using orthogonal technologies via western blotting and ELISA.

LC-PRM analysis of plasma digests

The plasma digests were tryptic digested samples were subjected to Thermo Scientific™ Q-Exactive™ Hybrid Quadrupole-Orbitrap fitted with a nano-liquid chromatography (LC) system (Thermo). The LC column for peptide separation was pre-packed with michrom Magic C18 (75µm x 15cm, 5µm, 120 A) and was used to separate peptides. A total of 2ug peptide constituted in 0.1% of formic acid (total volume 10µL) were injected into the column in 99% buffer A (0.1% FA) and 1% buffer B (0.1% FA in ACN). The flow rate was kept at 0.3 µl/min with following gradient conditions (50 mins linear gradient from 1% to 65% buffer B, a 2 mins linear gradient from 65% to 85% buffer B, and a final 8 mins gradient from at 85% buffer B). Between each cell lysate sample, one blank was run whereas for recombinant proteins blank

was placed after three runs. An orbitrap resolution was set at 17,500, with a target AGC value of $1e^6$, maximum fill time of 250ms, and an isolation window of 2.0m/z. Collision energy for each pair of positively charged precursor ion and peptide fragments was optimised. The peptides with strongest intensities for four to eight transitions per protein were selected as the detection targets. Some key considerations are important for optimisations to get high quality data for the targets under consideration. These optimisations include mass to charge ratio (m/z) of the peptide, charge of the precursor ions, optimised collision energies and peak intensities of transitions. Further optimisations mainly focused on the number of transitions being monitored every run and its effect on the signal to noise ratio of the targets in addition to changing parameters like dwell time, cycle time and parameters of the ion source.

Heavy synthetic peptides for quantitation

Heavy labelled peptides were custom made for selected targeted peptides that were reproducibly detected with high peak intensity in SWATH™-MS. Peptides are ADAMDEC1: modified NSVASIST**C**DGL [13C6-15N4-R], charge 2+, formula weight 1332.46; CFD: ATLGPAVRPLPWQ [13C6-15N4-R], charge 3+, formula weight 1571.85 manufactured by SynPeptide Co. Ltd., China) with ≥ 98% purity. The amount of heavy labelled peptide to be spiked in the tryptic digest was optimised by running a serial-dilutions from range 0.7-70 pmol/µl (dilution points at 0.7, 3.5,7,17.5,35,70) pmol/µl.

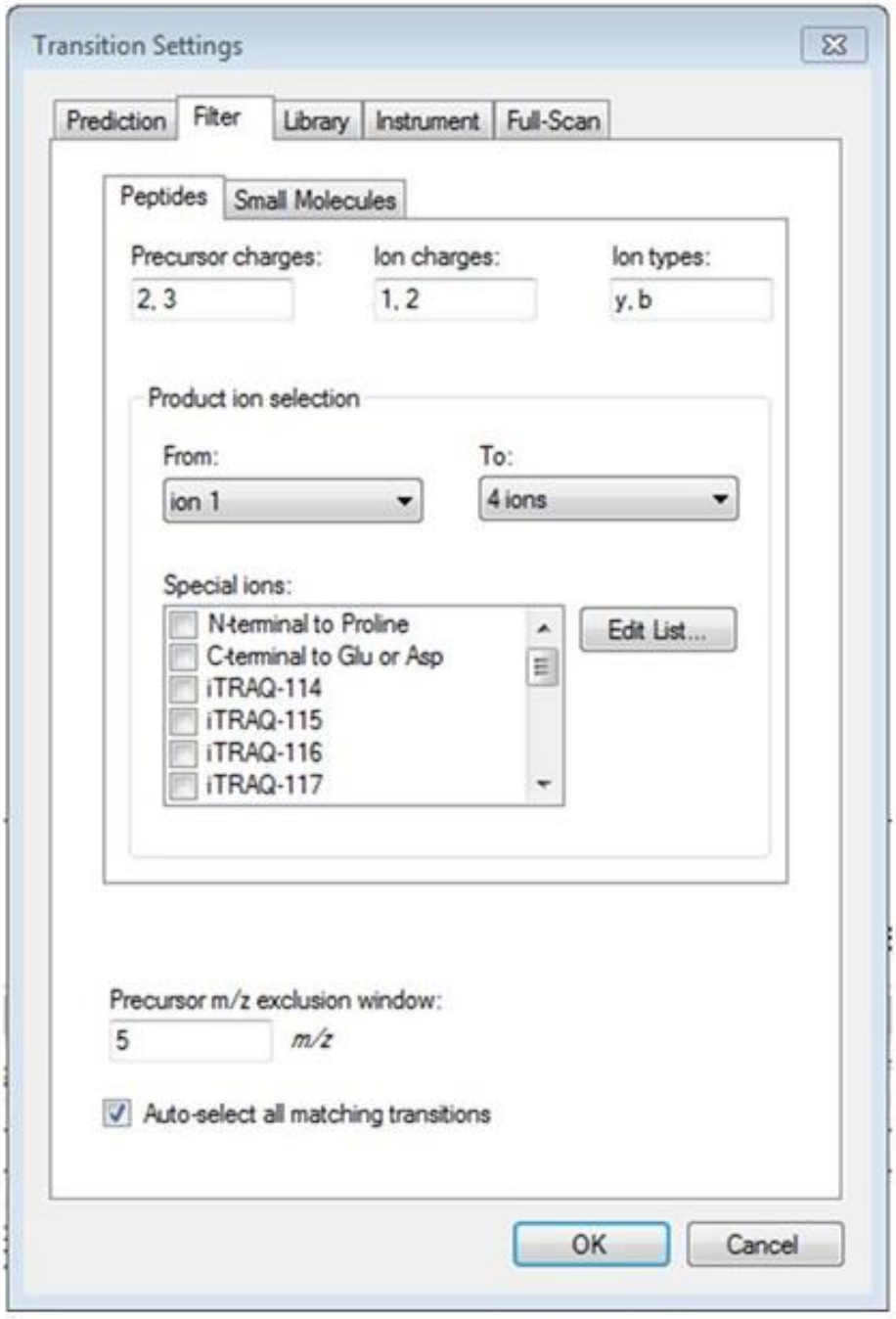
Table 4.1. Peptide sequences and selected transitions for PRM plasma assays of selected candidate biomarkers

Peptide sequence	**Protein**	**UniProt I.D.**	**Heavy Precursor m/z**	**Light Precursor m/z**	**Heavy peptide concentration (pmol/µl)**
NSVASIST**C**DGLR	ADAM-like decysin 1 (ADAMDEC1)	O15204	695.3369++	690.3328++	ND
ATLGPAVRPLPWQR	Complement factor D (CFD)	P00746	524.6397+++	521.3036+++	ND

Data analysis

The quantification of all the two proteins targets were performed on Skyline (version 4.2) (http: //proteome.gs.washington.edu/software/skyline). Skyline is an open-source software that supports several methods of extracting chromatography-based quantitative measurements from the raw data files (Egertson et al., 2015). In brief, the FASTA sequences of two targets was uploaded to the skyline file. In the transition settings, under filter tab, the precursor charges were kept 2, and 3; ion charges as 1, 2, 3 and ion types as y,b, and p (Figure 4.2). The library settings under the transition settings, ion match tolerance was kept at 0.7 m/z. Under the full scan tab, MS/MS filtering was 'targeted', 'QIT' as product mass analyser with a resolution of 0.7 m/z. Import spectral library created in biomarker discovery experiment and enter the cut-off score at 0.95. The spectral library provides an accurate and quick match to experimental detected MS/MS spectra. The normalisations of peak areas, histograms, product dot ion product availability (dotp) and graphs were all made using Skyline and MS stats (http://msstats.org/). In the peptide settings, digestion via trypsin was set with no missed cleavages and 13C (6)15 N (4) C-term Arginine and 13C (6)15 N (1) Leucine as specific isotope modifications. The excel file for spectral library is provided as supplementary file S1. The isolation width of m/z =3 was set up in MS/MS filtering tab at a resolving power of 7500 (at 400 m/z) and orbitrap as a mass analyser. MS/MS peak integration was performed by Skyline software for identification and quantification of the precursors by matching with the spectral library.

A


Transition Settings
Prediction
Filter
Library
Instrument
Full-Scan
Peptides
Small Molecules
Precursor charges:
2, 3
Ion charges:
1, 2
Ion types:
y, b
Product ion selection
From:
ion 1
To:
4 ions
Special ions:
N-terminal to Proline
C-terminal to Glu or Asp
iTRAQ-114
iTRAQ-115
iTRAQ-116
iTRAQ-117
Edit List...
Precursor m/z exclusion window:
5
m/z
Auto-select all matching transitions
OK
Cancel

B

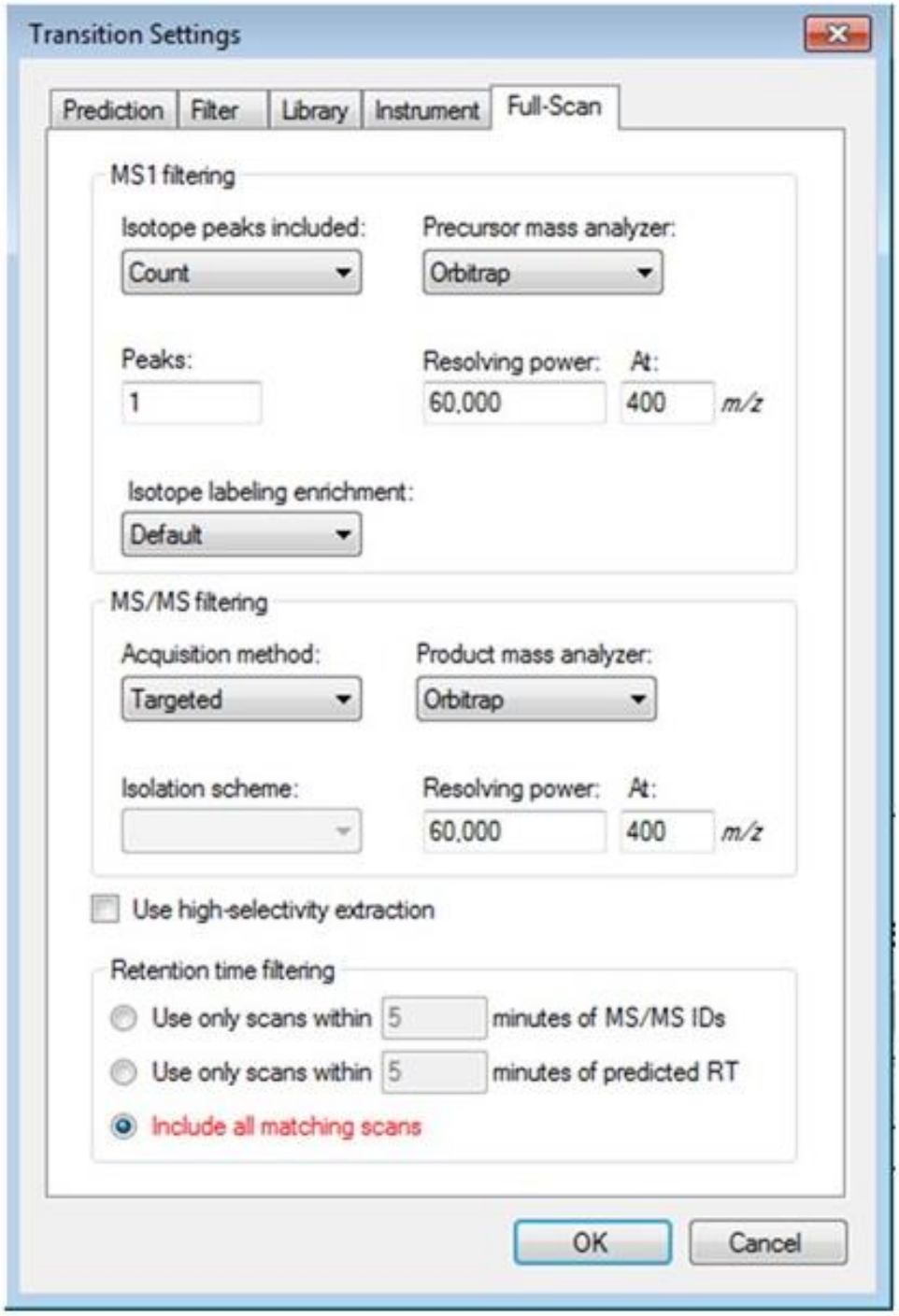


Figure 4.2: Skyline MS/MS filtering options and setting used in the study *A) Filter features for selecting precursor ion charge state and type of product ion. B) Full scan features the isotopes peaks selection, resolution of 60,000 at 400 m/z.*

Statistical analysis

Histograms, bar graphs and analysis between CRC diseased and controls patients were performed using GraphPad Prism version 7.00 for Windows (GraphPad Software, San Diego, USA, www.graphpad.com).

4.3 Results and Discussion

Selection of light and heavy peptides

Unique peptide targets for PRM assay development were selected based on SWATH™-MS biomarker discovery experiments in data-independent mode, which had identified ~450 plasma proteins (Chapter 3). Selected peptides; ADAMDEC1: NSVASIST**C**DGLR, m/z 693.3328, 2+ CFD: ATLGPAVRPLPWQR, m/z 521.3036, 3+ are proteotypic and were uniquely mapped to the ADAMDEC1 and CFD respectively. These were observed reproducibly in multiple runs of biomarker discovery experiments reassuring the 'flyability' of the peptides. Target peptide for ADAMDEC1 contains a modification cite at NSVASIST**C**DGLR and therefore special attention was paid during quantification of this peptide. The peptide charge state of these fragments was 2+ and 3+.

Preliminary target measurement via PRM assay

The main objective of this experiment is to reproducibly detect and relatively quantify endogenous target peptides for proteins ADAMDEC1 and CFD using Thermo Scientific™ Q-Exactive™ Hybrid Quadrupole-Orbitrap. The detailed m/z light precursor mass and charge is shown in Table 4.1. Triplicate aliquots of the trypsin digested CRC stage (I-IV) plasma samples and healthy controls (pooled set) without subjecting to any depletion and enrichment were analysed for relative quantification of the target peptides. Peptide CFD: ATLGPAVRPLPWQR; was reproducibly detected in all 15 runs (triplicate of 5 pooled set of samples, 4 for each CRC stage (I-IV) and 1 for healthy control). Peptide ATLGPAVRPLPWQR is proteotypic and does not contain any cite of PTM modification which makes it an ideal target peptide for developing a PRM assay. The selected daughter ion for quantification are y11: 638.8673++ (rank1), y12: 695.4093++ (rank2), y6: 796.4464+ (rank4), y10: 610.3566++ (rank3) and y9: 561.8302++ (rank 7), Figure 4.2.

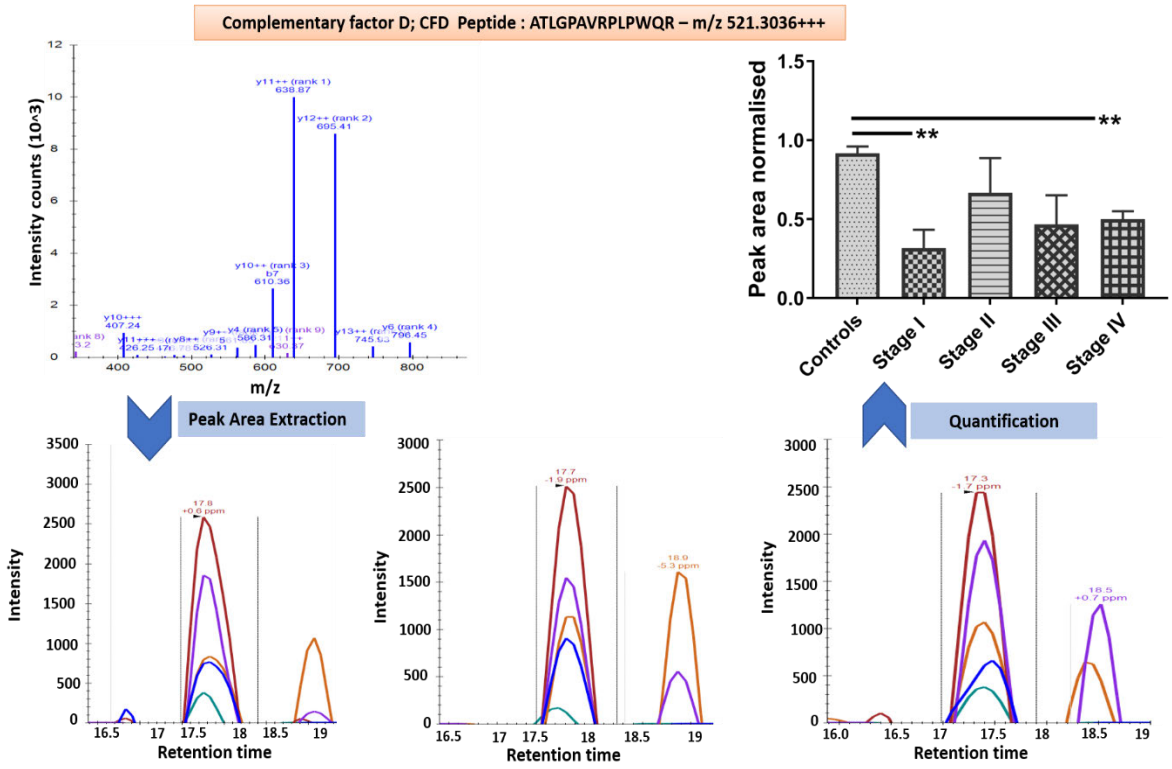


***Figure 4.3*: Skyline MS/MS filtering for CFD.** *A) MS/MS spectra for the targeted peptides highlighting in blue matching fragments for an injected replicate as shown in Skyline for peptide ATLGPAVRPLPWQR with charge 3+ showing the five fragments for light 521.3036+++. B) Chromatograms and peak intensity traces if three technical replications of pool of control patients C) Relative comparison of normalised peak areas between healthy controls and CRC stages (I-IV). ** p < 0.005.*

Peptide NSVASISTCDGLR with charge 2+ showing the six fragments for light 690.3328++. is proteotypic and contains cysteine which is carbamidomethylated after alkylation in trypsin digestion. The selected daughter ion for quantification are y8: 921.4458++ (rank5), y10: 540.2611++ (rank3), y8: 461.2266++ (rank1), y7:404.6845++, b9: 920.4142+(rank 4) and b6: 572.3039+ (rank 2) Figure 4.3. % Coefficient of variance and statistical significance between technical triplicates runs of the trypsin digested plasma samples from CRC stages I-IV for ADAMDEC1 and CFD peptides (Table 4.2). The normalised peak area plotted using graph-pad prism showed relative down expression of the peptide in all CRC stages (I-IV) in comparison to healthy which is in parallel to our SWATH™-MS and Western blot data discussed in the previous chapter (Chapter 3, Figure 3.2). Unfortunately, peptide

NSVASIST**C**DGLR for ADAMDEC1 was not reproducibly identified in all the triplicates, requiring troubleshooting and further optimisation.

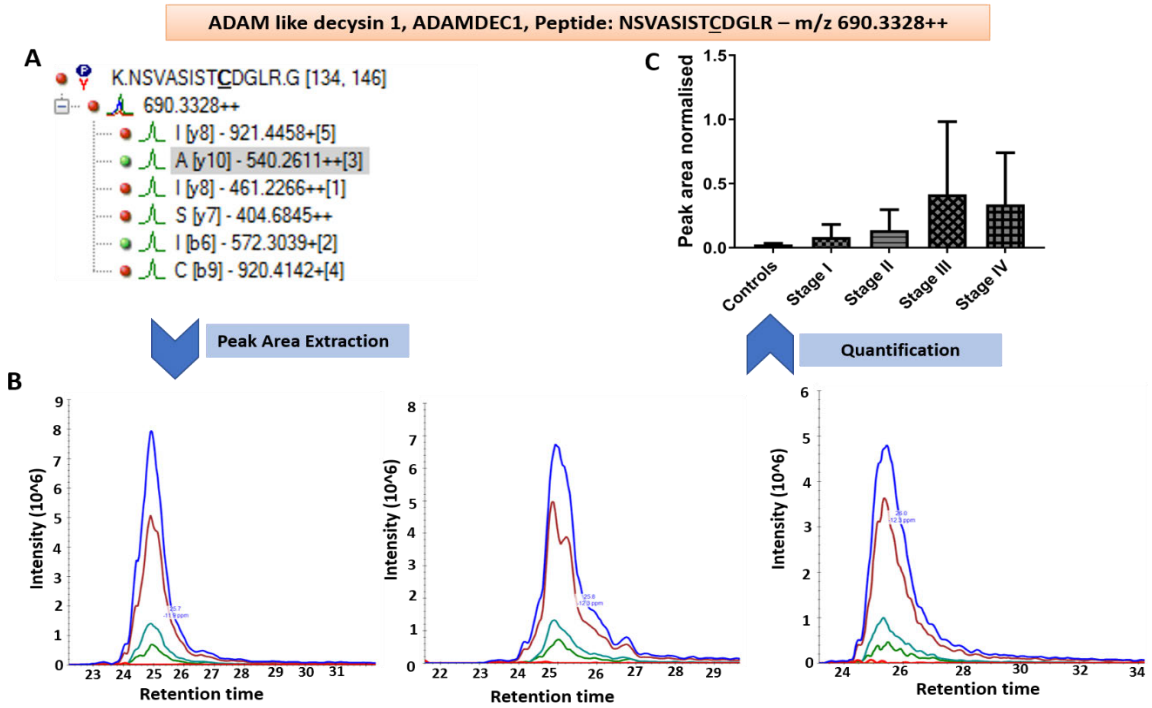


***Figure 4.4**: **Skyline MS/MS filtering for ADAMDEC1.** A) Skyline peptide tree for peptide NSVASISTCDGLR with charge 2+ showing the six fragments for light 690.3328++. B) Chromatograms and peak intensity traces if three technical replications of pool of CRC stage III patients C) Relative comparison of normalised peak areas between healthy controls and CRC stages (I-IV).*

Table 4.2: % Coefficient of variance and statistical significance between technical triplicates runs of the trypsin digested plasma samples from CRC stages I-IV for ADAMDEC1 and CFD peptides. *ns*: non-significant

Peptide	UniProt Accession Number	Protein Name	Total Variability (%CV)	P-Value
NSVASISTCDGLR	O15204	ADAMDEC1	18.40	0.035
ATLGPAVRPLPWQR	P00746	Complement Factor D	15.40	ns

To develop a high-quality PRM assay, it is important to determine of lowest limit of detection (LoD), sensitivity and quantification using heavy-labelled peptides. Clinical Proteomic Tumour Analysis Consortium (CPTAC) provides general guidelines to assist in the

development of any targeted proteomics assay so that it best avoids inter and intra-lab variability (Whiteaker et al., 2016). The data presented in this study is preliminary and requires robust optimisation according to CPTAC guidelines which will be complied with.

4.5 Future experiments

To develop a sensitive and accurate assay, robust analytical characterisation must be performed. The proteomics community has published a common set of guidelines and an assay repository database that jointly form a public repository called the Clinical Proteomic Tumour Analysis Consortium (CPTAC) Assay Portal (http://assays.cancer.gov/). CPTAC provides guidelines for standard operating procedures, protocols, and assay characterisation data associated with targeted mass spectrometry-based assays. Out of 7 candidates confirmed using Western blot and ELISA in Chapter 3, only CST3 has an available PRM assay in CPTAC library portal. The remaining protein candidates require the development of PRM assays for measurement in plasma.

Following the standard operating procedures, there are still multiple experiments that are needed for developing a well-established assay (Whiteaker, J. R. et. al., 2014). These experiments include:

1) Dilution of the synthetic peptide/recombinant protein/internal standard into the trypsin digested plasma. This is essential to establish the stoichiometric levels between the endogenous peptides and internal standards by determining the limit of detection and quantification. It also determines the extent of interference due to the presence of the sample matrix.

2) Evaluation of reproducibility: A minimum of three replicates of peptides spiked at three concentrations (i.e. low, medium, and high) to determine the reproducibility of the experiments at multiple concentrations.

3) Evaluation of the variation of peptide detection in multiple biological samples to see the selectivity.

4) Measuring the stability of peptides across various freeze and thaw cycles.

5) Reproducible quantification of the peptides in a relevant matrix (in our case CRC plasma)

Once a characterized assay is ready, it can be used to measure the levels of any target protein in individual patient samples across CRC stages. The measurements can also be evaluated using negative controls such as GI or other tumours to evaluate the specificity of the peptide in

specifically discriminating **CRC** patients from healthy controls. The success of these experiments will decide the ultimate clinical utility of all candidate biomarkers.

4.4 Critical Appraisal and Challenges

PRM-based methods are widely used to verify the results from discovery-based proteomic workflows. The relative expression of ADAMDEC1 and CFD peptides obtained from plasma samples of CRC patients was verified using PRM. The targeted proteomics results were consistent with Western blotting data for CFD, and with the ELISA data for ADAMDEC1. The CFD peptide was observed to be down-regulated in CRC stages (I-IV) in comparison to healthy controls. The ADAMDEC1 peptide was found to be up-regulated in CRC stages (I-IV) in comparison to healthy controls.

Nevertheless, targeted proteomics still lacks the sensitivity to detect low abundant peptides with precision in plasma samples. This was particularly evident in my attempts to detect expression levels of ADAMDEC1 in healthy plasma, where the reported concentration of plasma ADAMDEC1 can be as high as 76 ng/ml. The consistent identification of an ADAMDEC peptide proved to be challenging as it was observed in only one of three healthy control sample runs with a reasonable intensity of $5 * 10^6$ whereas the other two replicates showed transition intensity of 800-1000. Careful analysis showed that the ADAMDEC1 peptide observed in the first run may have been bleed-through from a previous sample, as a recombinant protein was run before initiating this experiment (Figure S4.1 E (i)). However, only further method development/experimentation can help to determine whether this was an instance of sample contamination.

On the other hand, CFD has a protein concentration of 8 ug/ml based on PeptideAtlas data. The reproducible detection of CFD peptide successfully quantified CFD in CRC stage I-IV and healthy control plasma samples (Figure S4.2). It is also important to note that the methods described in this study determine quantification of a single proteotypic peptide. It is important to validate targeted-MS data using $\geq$2 proteotypic peptides (Sioud, M. 2007; Oswald et. al., 2015). In addition, batch-to-batch digestion efficiency varies between one plasma digest to another. Therefore, calculation of variance between different tryptic digests, and between freeze thaw cycles will be an important step in evaluating future results. Furthermore, PRM experiments were performed using Quadrupole-Orbitrap, whereas a triple quadrupole mass

spectrometer linked with a nano-ESI source and a nano-LC system may have increased analytical sensitivity (Sioud, M. 2007; Rauniyar, N. 2015).

Concluding remarks

Out of the list of 37 candidates, 9 candidates were selected based on their biological roles in cancer and verification results using either western blot or ELISA immunoassays (Figure 3.6, Chapter 3). From these 9, four protein targets PON1, CST3, CFD and ADAMDEC1 were further processed to develop a PRM assay.

Out of these four candidates, only ADAMDEC1 and CFD were successfully completed as shown in Figures 4.3 and 4.4. The remaining two candidates PON1 and CST3, did not show reliable identification at one retention time in consecutive runs. The error between retention times was greater than 5 minutes and hence further method optimisation is required to validation and reliably capture the transitions from PON1 and CST3 in plasma samples. In summary, this study provides preliminary data overviewing the trends of identified potential targets ADAMDEC1 and CFD. Absolute quantification of target proteins can now be obtained by spiking heavy peptide into plasma samples after determination of lowest limit of quantification of both heavy and light peptide. This study lays a foundation for the future development of a potentially translatable assay for CFD for early CRC diagnosis.

References

Addona, T.A., Abbatiello, S.E., Schilling, B., Skates, S.J., Mani, D.R., Bunk, D.M., Spiegelman, C.H., Zimmerman, L.J., Ham, A.J., Keshishian, H., *et al.* (2009). Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. Nature biotechnology *27*, 633-641.

Aebersold, R., Burlingame, A.L., and Bradshaw, R.A. (2013). Western blots versus selected reaction monitoring assays: time to turn the tables? Molecular & cellular proteomics : MCP *12*, 2381-2382.

Bourmaud, A., Gallien, S., and Domon, B. (2016). Parallel reaction monitoring using quadrupole-Orbitrap mass spectrometer: Principle and applications. Proteomics *16*, 2146-2159.

Chau, C.H., Rixe, O., McLeod, H., and Figg, W.D. (2008). Validation of analytic methods for biomarkers used in drug development. Clinical cancer research : an official journal of the American Association for Cancer Research *14*, 5967-5976.

Egertson, J.D., MacLean, B., Johnson, R., Xuan, Y., and MacCoss, M.J. (2015). Multiplexed peptide analysis using data-independent acquisition and Skyline. Nature protocols *10*, 887-903.

Geyer, P.E., Holdt, L.M., Teupser, D., and Mann, M. (2017). Revisiting biomarker discovery by plasma proteomics. Molecular systems biology *13*, 942.

Harlan, R., and Zhang, H. (2014). Targeted proteomics: a bridge between discovery and validation. Expert review of proteomics *11*, 657-661.

Hosseini S., V.-V.P., Rito-Palomares M., Martinez-Chapa S.O. (2018). General Overviews on Applications of ELISA. In: Enzyme-linked Immunosorbent Assay (ELISA) (Springer, Singapore).

Mahboob, S., Ahn, S.B., Cheruku, H.R., Cantor, D., Rennel, E., Fredriksson, S., Edfeldt, G., Breen, E.J., Khan, A., Mohamedali, A., *et al.* (2015). A novel multiplexed immunoassay identifies CEA, IL-8 and prolactin as prospective markers for Dukes' stages A-D colorectal cancers. Clinical proteomics *12*, 10.

Michaud, S.A., Sinclair, N.J., Petrosova, H., Palmer, A.L., Pistawka, A.J., Zhang, S., Hardie, D.B., Mohammed, Y., Eshghi, A., Richard, V.R., *et al.* (2018). Molecular phenotyping of laboratory mouse strains using 500 multiple reaction monitoring mass spectrometry plasma assays. Communications biology *1*, 78.

Ong, S.E., and Mann, M. (2005). Mass spectrometry-based proteomics turns quantitative. Nature chemical biology *1*, 252-262.

Oswald, S., Gröer, C., Drozdzik, M. et al. AAPS J (2013) 15: 1128. https://doi.org/10.1208/s12248-013-9521-3

Parker, C.E., Domanski, D., Percy, A.J., Chambers, A.G., Camenzind, A.G., Smith, D.S., and Borchers, C.H. (2014). Mass spectrometry in high-throughput clinical biomarker assays: multiple reaction monitoring. Topics in current chemistry *336*, 117-137.

Paulovich, A.G., Whiteaker, J.R., Hoofnagle, A.N., and Wang, P. (2008). The interface between biomarker discovery and clinical validation: The tar pit of the protein biomarker pipeline. Proteomics Clinical applications *2*, 1386-1402.

Peterson, A.C., Russell, J.D., Bailey, D.J., Westphall, M.S., and Coon, J.J. (2012). Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. Molecular & cellular proteomics : MCP *11*, 1475-1488.

Rauniyar, N. (2015). Parallel Reaction Monitoring: A Targeted Experiment Performed Using High Resolution and High Mass Accuracy Mass Spectrometry. International journal of molecular sciences *16*, 28566-28581.

Rauniyar, N. (2015). Parallel reaction monitoring: a targeted experiment performed using high resolution and high mass accuracy mass spectrometry. International journal of molecular sciences, 16(12), 28566-28581.

Sioud M. (2007) Main Approaches to Target Discovery and Validation. In: Sioud M. (eds) Target Discovery and Validation Reviews and Protocols. Methods in Molecular Biology™, vol 360. Humana Press

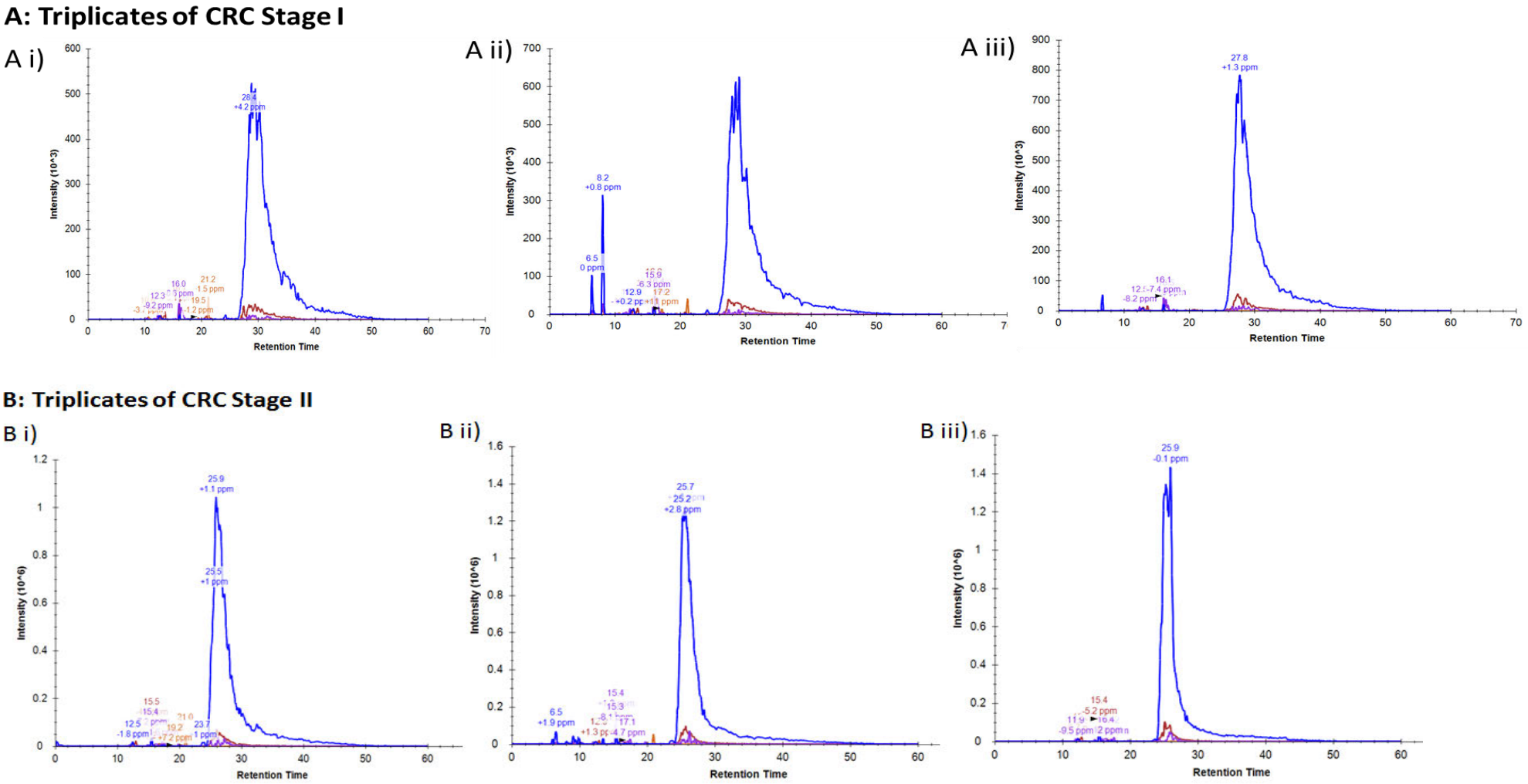
Towbin, H., Staehelin, T., and Gordon, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. Proceedings of the National Academy of Sciences of the United States of America *76*, 4350-4354.

Whiteaker, J.R., Halusa, G.N., Hoofnagle, A.N., Sharma, V., MacLean, B., Yan, P., Wrobel, J.A., Kennedy, J., Mani, D.R., Zimmerman, L.J., *et al.* (2016). Using the CPTAC Assay Portal to Identify and Implement Highly Characterized Targeted Proteomics Assays. Methods in molecular biology (Clifton, NJ) *1410*, 223-236.

Whiteaker, J. R., Halusa, G. N., Hoofnagle, A. N., Sharma, V., MacLean, B., Yan, P., ... & Meyer, M. R. (2014). CPTAC Assay Portal: a repository of targeted proteomic assays. Nature methods, 11(7), 703.

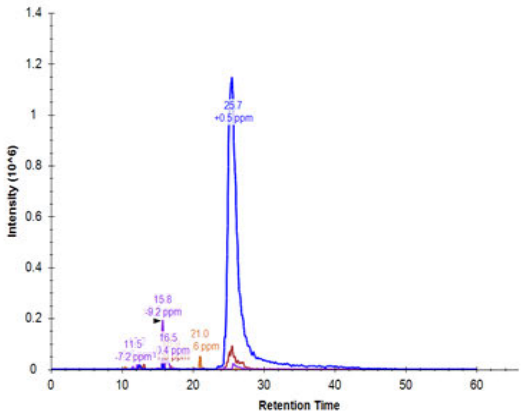
Supplementary Information

Figure S4.1 Transitions identified from peptide NSVASISTCDGLR (ADAMDEC1) in technical triplicates of CRC stages I-IV (pooled plasma) and healthy controls

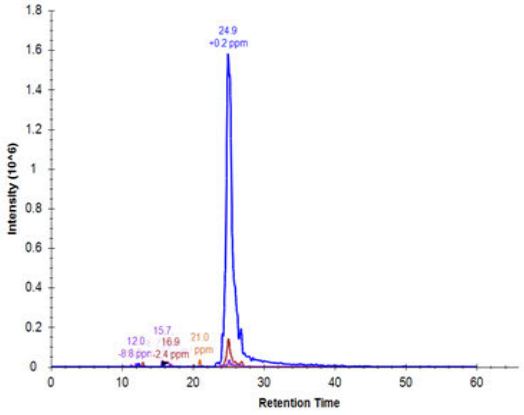


C: Triplicates of CRC Stage III

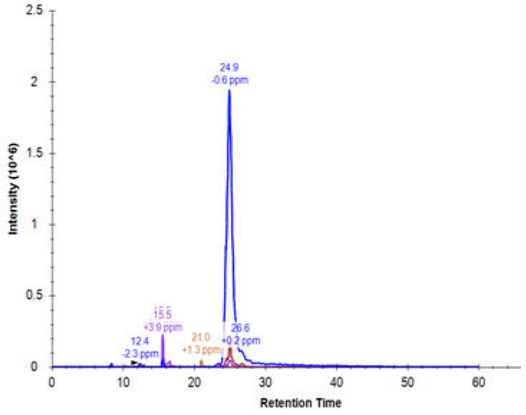
C i)



C ii)



C iii)

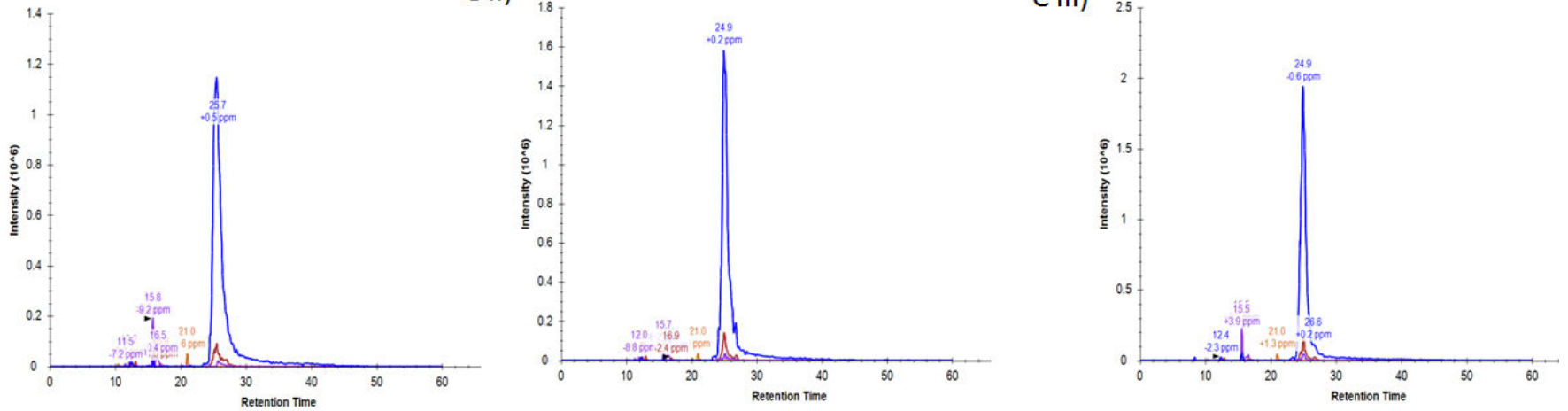


C: Triplicates of CRC Stage III

C i)

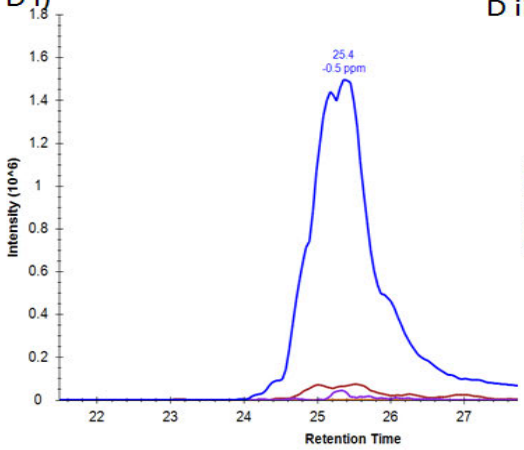
C ii)

C iii)

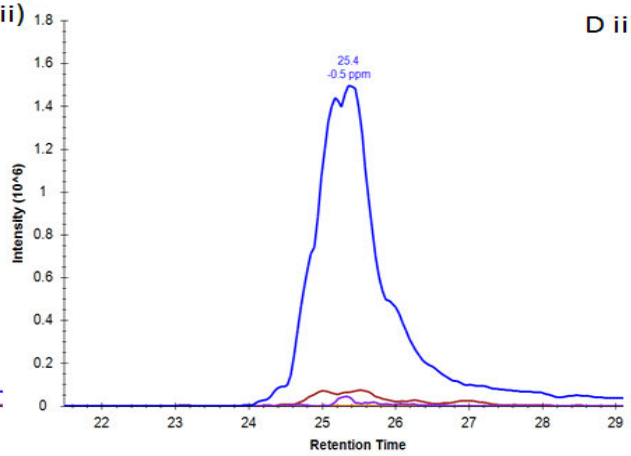


D: Triplicates of CRC Stage IV

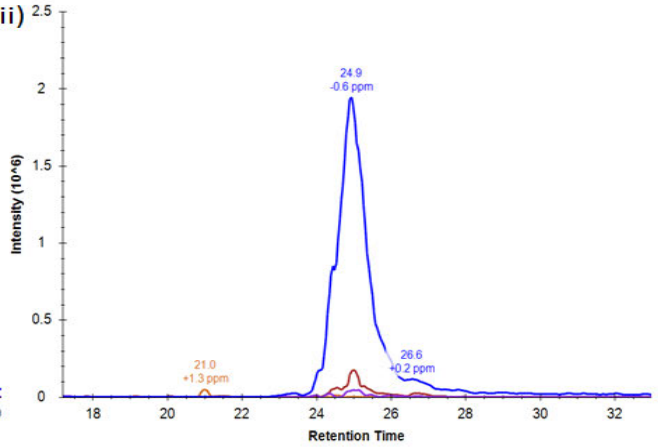
D i)



D ii)

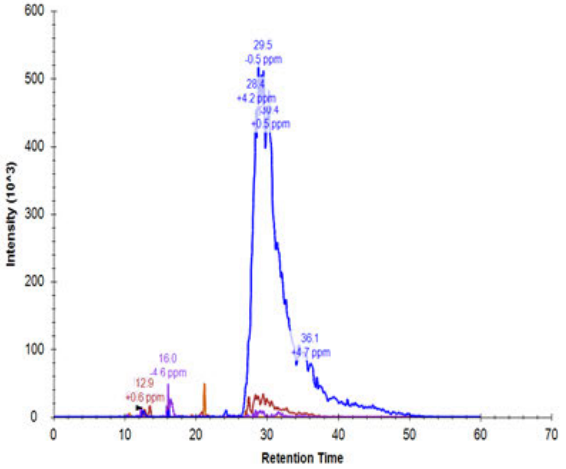


D iii)

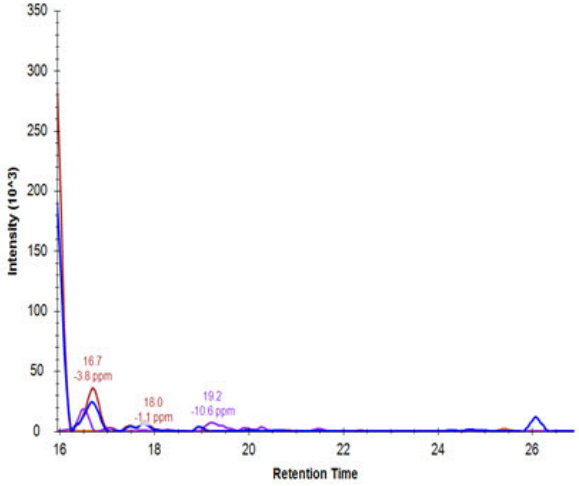


E: Triplicates of Healthy Controls

E i)



E ii)



E iii)

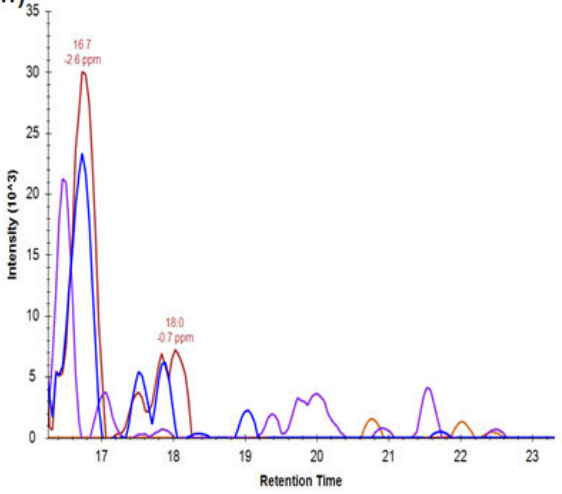
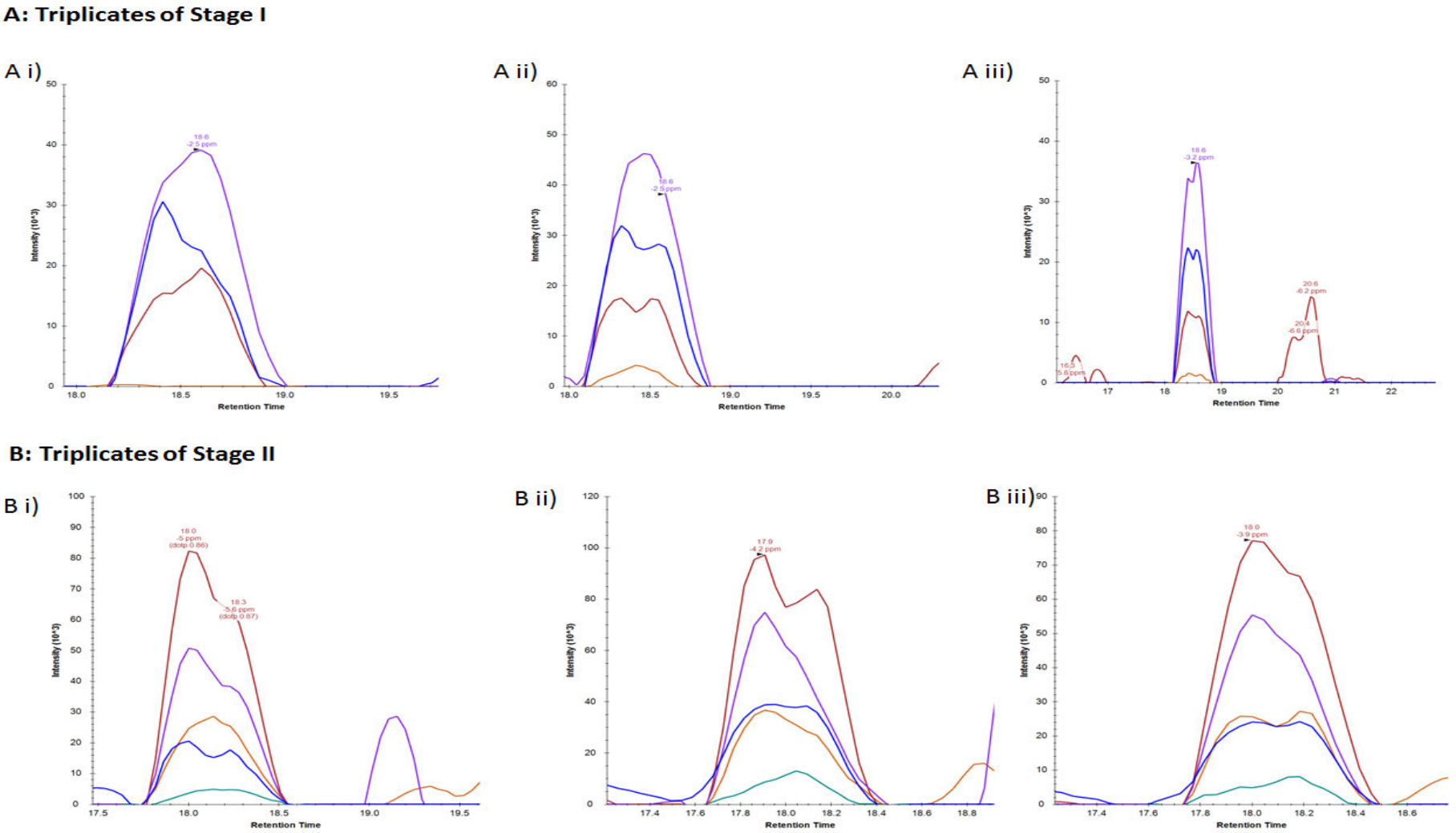


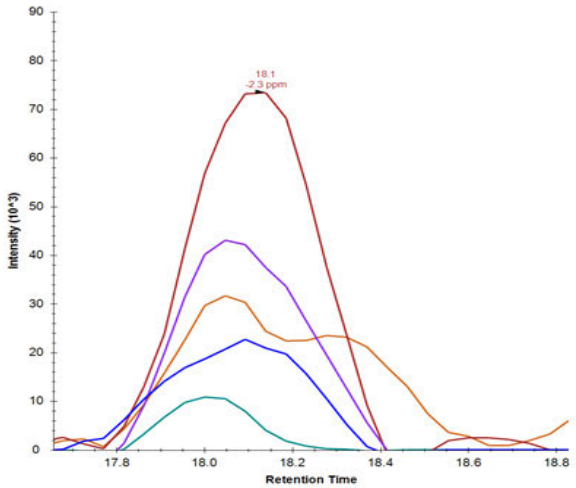
Table S4.1 ADAMDEC1 Normalised Peak Area Intensity				
Normalized Area	**Precursor Result**	**Precursor Results Summary**	**Protein Name**	**Replicate**
1.74E-02	571582	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	A1
1.96E-01	2895062	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	A2
3.84E-02	701774	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	A3
4.34E-02	504112	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	B1
3.21E-01	15801390	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	B2
4.79E-02	352682	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	B3
1.01E-01	26125030	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	C1
1.07E+00	1252529	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	C2
8.07E-02	1251567	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	C3
5.22E-02	742869	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	D1
6.23E-01	3320430	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	D2
3.58E-02	1008936	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	E1
1.96E-02	3350372	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	E2
2.24E-02	128802	RT: 15.56+/-3.89 Area: 4143367+/-7485407	sp|O15204|ADEC1_HUMAN	E3

Figure S4.2 Transitions identified from peptide ATLGPAVRPLPWQR (CFD) in technical triplicates of CRC stages I-IV (pooled plasma) and healthy controls

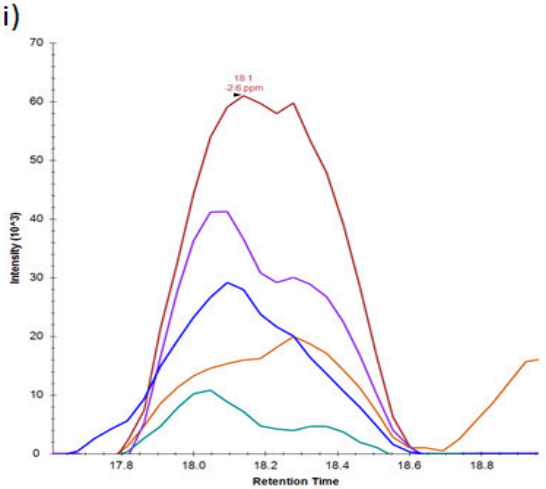


C: Triplicates of Stage III

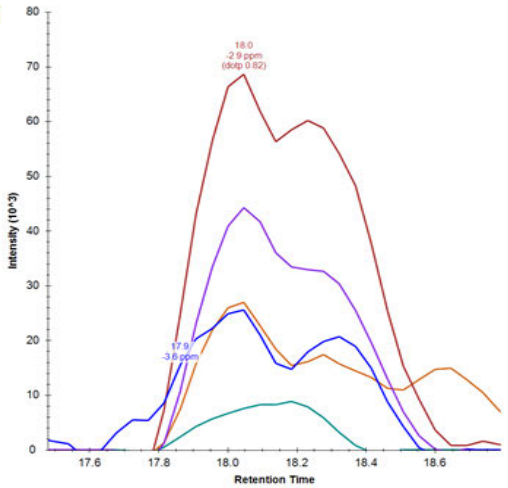
C i)



C ii)

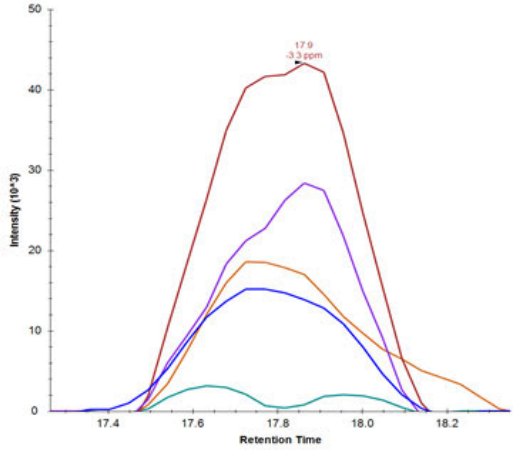


C i

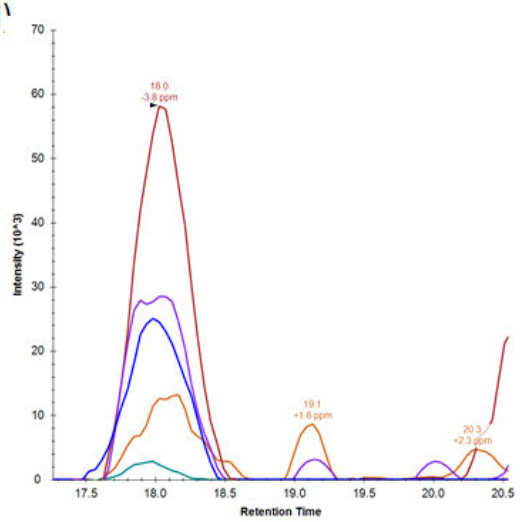


D: Triplicates of Stage IV

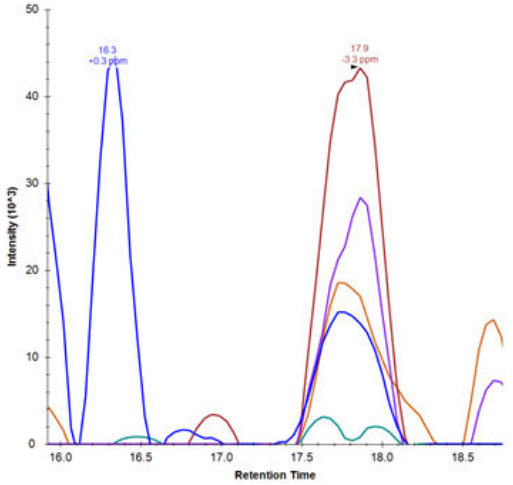
D i)



D ii)

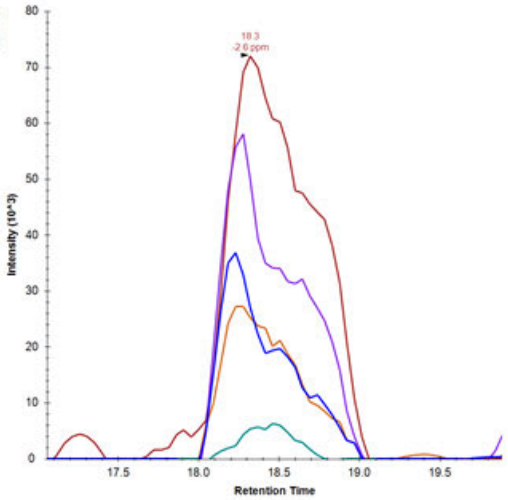


D iii)

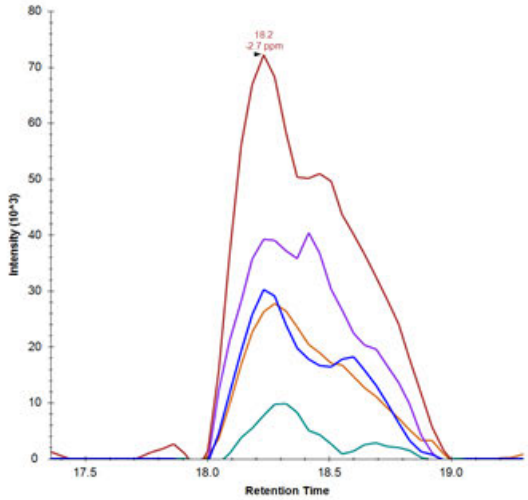


E: Triplicates of Healthy Controls

E i)



E ii)



E iii)

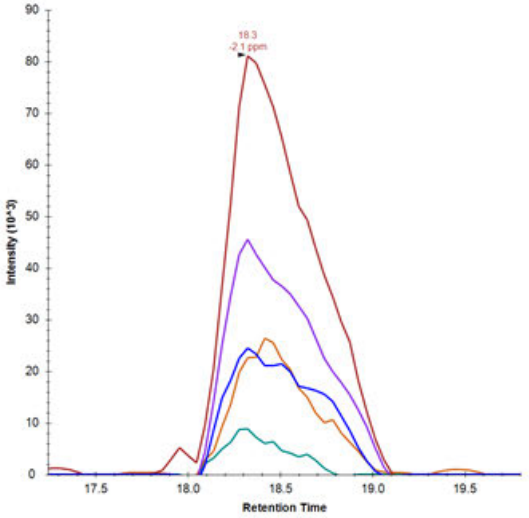


Table S4.2 Complement Factor D Normalised Peak Area Intensity					
Area Normalized	**Transition Results Summary**	**Precursor Result**	**Precursor Results Summary**	**Protein Name**	**Replicates**
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	2409645	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	A1
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	2201094	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	A2
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	1877391	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	A3
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	9537801	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	B1
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	3049100	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	B2
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	1364333	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	B3
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	2282177	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	C1
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	4710353	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	C2
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	4200453	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	C3
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	4713250	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	D1
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	2519008	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	D2
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	3274652	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	D3
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	8543592	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	E1
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	13981250	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	E2
1.4219%+/-1.566%	RT: 19.14+/-1.7 Area: 113346+/-148586	8192211	RT: 18.9+/-2 Area: 4857087+/-3608898	sp|P00746|CFAD_HUMAN	E3

Chapter 5

μPAR and αvβ6 as metastatic marker of colorectal cancer quantified using parallel reaction monitoring

Abstract

The Baker research team based at Macquarie University has focused on developing strategies to identify novel diagnostic/prognostic markers for colorectal cancer (CRC) metastasis, whilst simultaneously elucidating the molecular interactions responsible for the disease. These efforts incorporate design and use of earlier stage plasma diagnostics, better tissue prognostic biomarkers and improved targeted therapeutic options to improve CRC patient survival.

The objective of this chapter was to explore spatiotemporal aspects of increased expression of circulating uPAR and integrin αvβ6 potentially 'shed' from the cancer cell surface into during the development of CRC.

The role of uPAR as a stage II prognostic marker in CRC tissues has been well established by our group, and it has been demonstrated that increased expression of uPAR determines the likelihood of CRC recurrence and overall survival in stage II CRC patients (Ahn et al., 2015). Similarly, αvβ6 expression has been shown to be elevated at early stages of CRC tumours, but not expressed in normal epithelium (Ahn S B et al., 2014, Bandyopadhyay A et al., 2009)). In addition, the NIH Clinical Trials database indicates at least 7 breast and prostate cancer clinical trials are targeting either uPAR or suPAR. It is possible that these proteins are expressed at extremely low levels and remained undetectable in undepleted plasma. This chapter describes work to identify and quantify uPAR and αvβ6 peptides from ultradepleted plasma using a targeted PRM-based assay approach.

Contributors

This work was completed in collaboration with Ms Sachini Fonseka, Dr Mathew McKay, Dr Seong Beom Ahn and Professor Mark Baker at Macquarie University, Australia. Some of the collaborative work presented in this chapter has been included in a Master of Research thesis published in 2017 by Ms Fonseka as some experiments were performed by her independently

whilst others were undertaken collaboratively. The details of contributions made by both authors is outlined below. I acknowledge Ms Fonseka as a significant contributor to data generation and analysis. This work was funded by an iMQRES scholarship #2015158 and a Sydney Vital Research Scholarship #50468/00.

Addendum:

A detailed summary of contributions made by both first authors Ms Samridhi Sharma and Ms Sachini Fonseka is detailed below in Figure 5A and Table 5A. These two displays show schematic progress across the project and itemise contributions made by candidates in the completion of milestones.

SS: Ms Samridhi Sharma
SF: Ms Sachini Fonseka

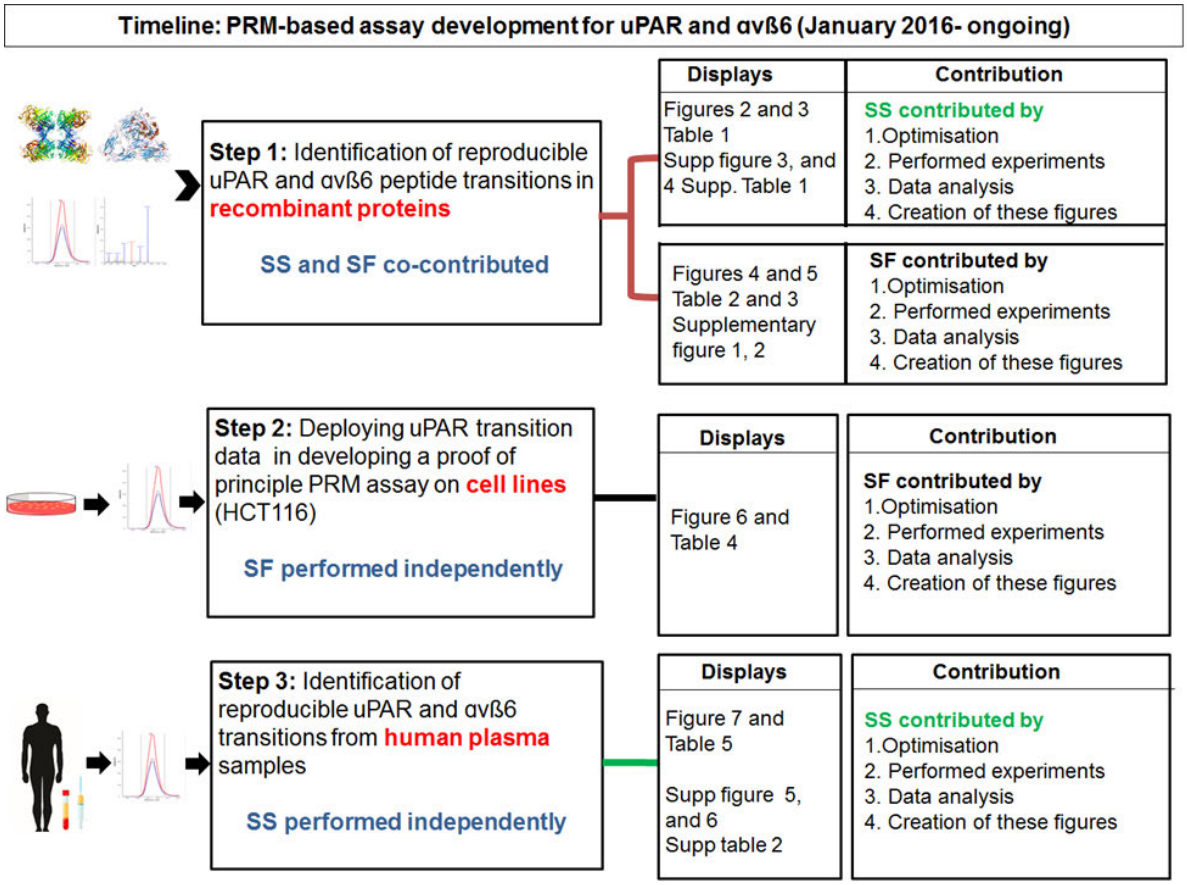


Figure 5A: *Schematic of the workflow involved in developing a PRM assay, pictographic timeline and contributions from joint first authors through major milestones of the project.*

Table 5A: Data contributed by joint first authors				
Display name	**Display type**	**Role Definition**	**Contribution**	
			Sharma S	**Foneska S**
Figure 5.1	Schematic representation of mechanism	NA	__	__
Figure 5.2	Data	This figure is based on experiments performed by SS. Results from this experiment were run by SS. Repeats of these experiments were performed by SF and data are provided as Supplementary Figure S5.1.	Performed the experiments, analysed the data and prepared the figure	__
Table 5.1	Data	Table 5.1 is derived from SS experiments reported in Figures 5.1 and 5.2. The data generated by SF are added as Table S5.2 to avoid the redundancy in the thesis chapter	Performed the experiments, analysed the data and made the figure	__
Figure 5.3	Data	This figure is based on experiments performed by SS. Results from this experiment were run by SS. A set of experiments were performed by SF and are provided as Figure S5.2.	Performed the experiments, analysed the data and made the figure	__
Figure 5.4, 5.5 and Table 5.2, 5.3	Data	This figure is based on experiments performed by SF. Results from this experiment were run by SF.	__	Performed the experiments, analysed the data and made the figure
Figure 5.6 and Table 5.4	Data	Experiments based on whole cell lysates of **HCT116 cell line** done by **SF exclusively**	__	Performed the experiments, analysed the data and made the figure
Figure 5.7, Figure 5.8 and Table 5.5	Data	uPAR and αvβ6 transitions detected in trypsin digested crude human plasma done by **SS exclusively.**	Performed the experiments, analysed the data and made the figure	__

Development of parallel reaction monitoring (PRM) assays for the μPAR and αvβ6 colorectal cancer biomarkers (in preparation)

Samridhi Sharma[1‡], Sachini Fonseka[2‡], Seong Beom Ahn[1], David Cantor[3], Edouard C. Nice [4] and Mark S. Baker[1]*

[1] Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Macquarie University, NSW, 2109, Australia

[2] School of Chemistry and Molecular Biosciences, Faculty of Science, Institute of Molecular Sciences, The University of Queensland, Brisbane Queensland, 4072 Australia

[3]Australian Proteome Analysis Facility (APAF), Department of Molecular Sciences, Macquarie University, NSW, 2109, Australia

[4]Department of Biochemistry and Molecular Biology, Monash University, VIC, 3800, Australia

‡ These authors contributed equally and should be considered equal first authors

Keywords: Colorectal Cancer, HCT116, EMT biomarkers, Prognosis, Diagnosis

***Corresponding Author**

Professor Mark S. Baker

Department of Biomedical Sciences

Faculty of Medicine and Health Sciences

75T, Talavera Road, Macquarie University

NSW, 2109, Australia

Disclosure of Potential Conflicts of Interest

The authors declare no real, potential or perceived conflicts of interest

5.1 Introduction

Colorectal cancer (CRC) is the third leading cause of cancer-related morbidity and mortality affecting both male and female populations (Siegel et al., 2014). The survival rate of patients is greater than 90% when diagnosed at an early stage (AJCC, stage I/II), as the tumour is confined within the bowel wall and is usually curable by surgery (Haggar and Boushey, 2009). Unfortunately, due to the asymptomatic nature of early stage disease, tumours remain undiagnosed and are primarily detected when they have spread (metastasised) to local lymph nodes and distant organs (late stages, AJCC III/IV). Metastasis to distant organs reduces patient survival to lower than 13% (Siegel et al., 2014) and is responsible for 90% of cancer-related deaths (Seyfried and Huysentruyt, 2013). Metastasis is described as a process in which cancer cells leave the primary tumour and move to surrounding tissues and distant organs where they form new lesions (Seyfried and Huysentruyt, 2013).

The sequential metastatic process can be distilled into distinct pathophysiological stages; emergence of the primary tumour growth, angiogenesis to meet increased metabolic demand, epithelial-to-mesenchymal transition (EMT), invasion, intravasation, survival in circulation, extravasation and dormancy or secondary tumour growth (Duffy et al., 2008; Eccles and Welch, 2007; Steeg, 2006). It is to be noted that the epithelial-to-mesenchymal transition (EMT) marks the initiation of metastasis in epithelial cancers, including CRC. During this transition, the epithelial cancer cell changes its interactions with basement membranes and undergoes biochemical changes to phenotypically transform into mesenchymal cells. This change provides additional motility to cells, essential for metastasis, invasiveness, resistance to apoptosis and increased turnover of extracellular matrix components (ECM) that assist in the completion of the EMT (Lamouille et al., 2014).

Several distinct molecular processes collectively initiate an EMT-like activation of transcription factors, cell matrix remodelling and cytoskeletal proteins, ECM-degrading enzymes and variations in the expression of specific cell-surface proteins (Kalluri and Weinberg, 2009; Lamouille et al., 2014). Some specific proteins such as integrins, proteolytic enzymes, growth factors and their respective downstream signalling are often regarded as lynchpins in this process and can be used as biomarkers of EMT progression (Kalluri and Weinberg, 2009).

The expression of two proteins in focus are namely integrin αvβ6 (Bengs et al., 2013; Cantor et al., 2015; Niu et al., 2002) and the urokinase-type plasminogen activator receptor (uPAR) (Brabletz et al., 2005; Eden et al., 2011; Lester et al., 2007) have been proposed as EMT markers. The serine protease urokinase plasminogen activator (uPA) plays a pivotal role in epithelial tissue remodelling as well as cancer development and progression by interacting with its specific GPI-linked membrane-bound receptor uPAR (Ploug et al., 1991; Uszynski et al., 2004). The interaction between uPA and uPAR activates downstream cell-surface plasminogen by proteolytic cleavage which starts plasmin-mediated pericellular ECM proteolysis, facilitating cell migration and invasion (Uszynski et al., 2004).

Apart from its role as a specific receptor for uPA, uPAR is overexpressed in both neoplastic and tumour-stromal invasive microenvironments (Uszynski et al., 2004). uPAR plays a critical role in cancer progression through its interaction with structural ECM proteins such as integrins and vitronectin, in addition to functioning as a regulator of angiogenesis (Zhang et al., 2003). Many epithelial cancers, including CRC, have are found to exhibit uPAR expression (Lester et al., 2007). The expression of uPAR in CRC tumour cells is characteristically limited to the invasive front of tumour islands, which facilitates its ability to increase cell motility or cancer metastasis (Pyke et al., 1994). A close correlation has been observed between high uPAR expression and poor patient prognosis, especially, during the transition to an invasive carcinoma (Suzuki et al., 1998). Furthermore, Ahn *et al.*, established that high epithelial cell uPAR expression differentiates poorer survival for stage II rectal cancer patients (Ahn et al., 2015).

Another class of lynchpin protein participating in EMT transition are the integrins. They represent a major class of ubiquitous transmembrane α/β heterodimer glycoproteins (Hamidi and Ivaska, 2018). Different integrin α/β heterodimers interact with the ECM through specific, high-affinity ligands. Some integrin heterodimers have been found to participate in ECM disruption to promote the EMT (Figure 5.1). Some, such as αvβ6 that is a high-affinity interactor with Transforming growth factor beta 1, are specifically expressed by epithelial cancer cells or cells undergoing large-scale tissue remodelling. As a result, it is not surprising that αvβ6 is poorly expressed by normal epithelial tissues (Hamidi and Ivaska, 2018). Unequivocal data supports the contention that integrin αvβ6 and uPAR interact directly and that both are involved in cancer proliferation, adhesion, migration, invasion and the EMT - all hallmarks of metastasis (Cantor et al., 2015). Saldanha et al., 2007 first identified interaction between uPAR and αvβ6 using a human ovarian epithelial cancer cell line using forward and

reverse co-immunoprecipitation (Saldanha et al., 2007). This interaction was confirmed by *in silico* structural modelling and peptide array analysis and that identified the exact binding sites which were validated by proximity ligation assays (Ahn et al., 2014) (Sowmya et al., 2014). Furthermore, uPAR and αvβ6 are co-expressed at invasive regions of CRC tumours, both are reported to delineate clinical stage and both are independent negative survival prognostic factors (Bengs et al., 2013) (Boonstra et al., 2011).

High expression of β6 (a subunit of the αvβ6 heterodimer) has been associated with poor patient survival in a study performed on 500 metastatic CRC patients (Bandyopadhyay and Raghavan, 2009). In another study by Scharl *et al.*, higher serum levels of β6 were observed in CRC patient plasma with 100% metastasis and poor patient survival (Bengs et al., 2013). In conclusion, multiple lines of evidence describe the association of elevated uPAR and β6 levels with cancer metastasis.

Current determination of both αvβ6 and uPAR are based on ELISA and immunohistochemistry (IHC) studies, which rely upon the sensitivity and specificity of commercially available antibodies to produce high-quality assays. Often, antibodies are poorly characterized, leading to cross-reactivity with other antigens present in biological samples, thereby providing non-robust and irreproducible results which can lead to the drawing of ambiguous conclusions (Lin et al., 2013). Therefore, there is a need to develop alternative, sensitive and specific assays for absolute protein quantification. In this context, lLiquid-chromatography (LC)-based parallel reaction monitoring (PRM) assays are robust and reproducible. PRM's can be multiplexed and deployed for use to test different biospecimens after optimisations (Picotti et al., 2010).

This study has an overarching aim to evaluate both uPAR and αvβ6 as potential EMT and CRC biomarkers in CRC plasma. Plasma as a biospecimen is ideal owing to its ease of accessibility and reflection of systemic physiology/pathology. However, the high dynamic concentration range of plasma makes it a particularly challenging biospecimen, as 90% of plasma is covered by high-abundant plasma proteins in mg/ml, whilst low abundance proteins or tissue-shed proteins like uPAR and αvβ6 have concentrations in the order of ng/ml. This makes detectability of low abundant proteins challenging. Therfore, before initiating our plasma PRM assay, a pilot study to optimise LC-PRM parameters was performed to determine expression of uPAR in the HCT116 cell line. A suitably developed PRM assay would then be deployed on human plasma samples to design a fit-for-purpose biomarker assay.

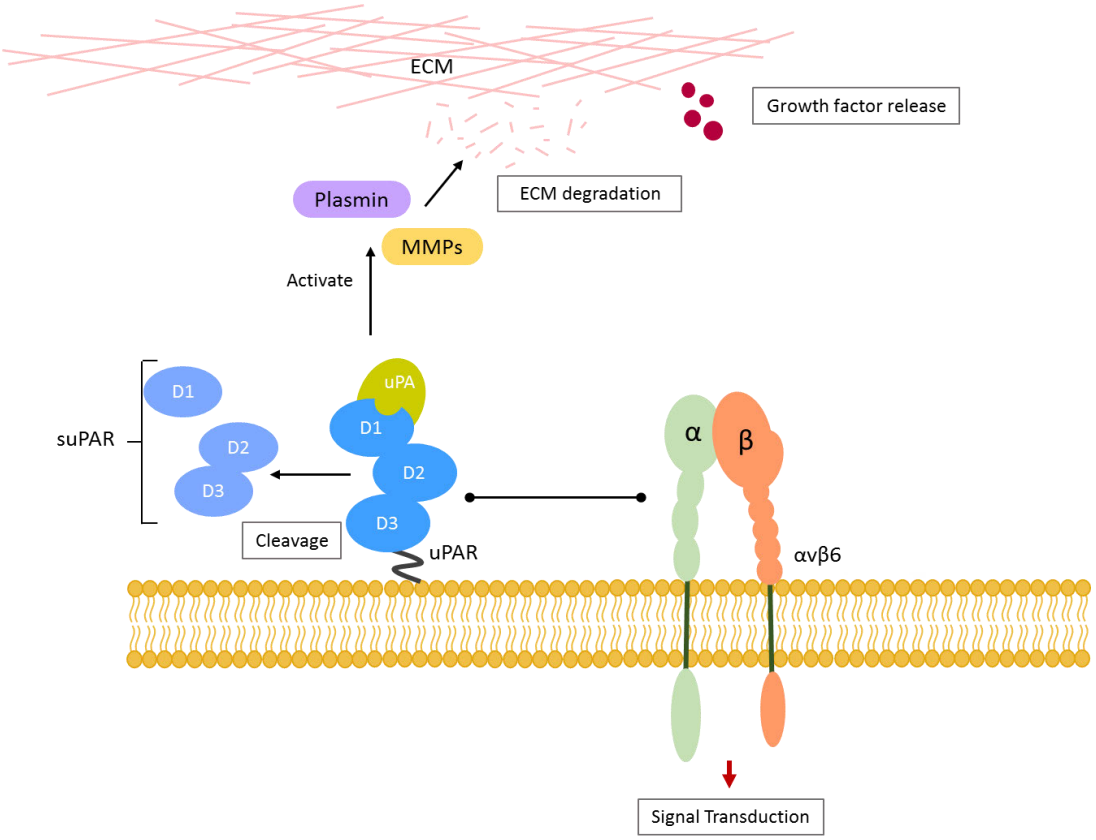


Figure 5.1: The schematic diagram of the role of uPAR and αvβ6 in the epithelial to mesenchymal transition at the cell surface. *At the extracellular plasma membrane, the zymogen protease pro-uPA binds to the GPI-anchored uPAR, where it is cleaved to release active twin-chain uPA as part of the plasminogen activation cascade, resulting in the activation of plasminogen to plasmin. Plasmin triggers the release and activation of other matrix metalloproteases (MMPs) from stromal cells to assist in degrading the ECM basement membrane, culminating in intravasation. αvβ6 expression directly relates with MMP secretion and co-interacts with the uPAR for intracellular transduction. Adapted from (Smith and Marshall, 2010). Image is adapted from (Cantor et al., 2015).*

5.2 Materials and Methods

All reagents, organic solvents and mass spectrometric-related chemicals employed in this study were obtained from Sigma Aldrich and were of MS grade with 99.9% purity, unless stated otherwise.

Recombinant proteins

Recombinant αvβ6 and uPAR proteins were used to generate and evaluate "flyable" peptides and transitions for successful PRM assays. Recombinant human αvβ6, composed covalently of αv at 110.5kDa and β6 at 68.6kDa in a 1:1 ratio, and uPAR protein standards were obtained from R&D, Minneapolis, USA (R&D catalogue numbers #3817-AV-050 and #807-UK/CF-100, respectively).

CRC cell lysate preparation and determination of total protein

This study employed two sub-clones of HCT116 cells (ATCC® CCL-247™) that were produced by the Baker/Wang/Doe research team at JCSMR, ANU and were gifted to our current research group by our colleague Dr Yao Wang, St George Hospital, NSW. Wild type (WT) HTC116 is a cell line derived from a stage B CRC tumour with known endogenous uPAR expression (Ahmed et al., 2003). The HCT116 anti-sense cell line (AS) has stable transfection of a uPAR-siRNA construct, resulting in an approximate 35-40% decrease in uPAR expression, relative to both the mock-transfected and/or wild type HCT116 cell line as determined by Western blotting (data not shown). The stable decrease in uPAR expression in HCT116 AS cell lines has been independently confirmed to be ~27% (Liu et al., 2014). In brief, cells were cultured to 90% confluency in 150cm tissue culture dishes in Dulbecco's Modified Eagle media, supplemented with 400µg/ml Hygromycin-B at 37^{o}C with 5% CO_2. Cells were washed in cold phosphate-buffered saline prior to lysis in a 100mM triethylammonium bicarbonate (TEAB) buffer containing 0.1% sodium deoxycholate. Protein concentration in the cell lysates was determined using bicinchoninic acid (BCA) assay kit (Pierce^{TM} #23225) following the manufacturer's protocol.

Plasma collection and sample preparation

Plasma samples from 100 individuals, comprising of 80 clinically staged colorectal cancer (I-IV; 20 in each) and 20 healthy controls were procured from Victorian Cancer Biobank, Melbourne under approved human ethics (HREC Ref #5201600401 and #5201700681). Plasma samples were collected prior to colonoscopy and detailed clinical/pathophysiology reports information of individual, their inclusion/exclusion criteria have been discussed in detail in research article (Mahboob et al., 2015) and section 3.2, Chapter 3 of this thesis.

Tryptic digestion of recombinant proteins, cell lysates and human plasma samples:

All recombinant protein samples, cell lysates and human plasma samples were reduced, alkylated, tryptically digested and desalted as outlined in material and methods prior to analysis by mass spectrometry

Liquid Chromatography (LC)-Parallel Reaction Monitoring (PRM)-MS analysis

Trypsin digested lysate samples were analysed on a Thermo Scientific™ Q-Exactive™ Hybrid Quadrupole-Orbitrap fitted with a nano-liquid chromatography (LC) system (Thermo). The LC column for peptide separation was pre-packed with Michrom Magic C18 (75µm x 15cm, 5µm, 120 A) to separate peptides. A total of 2ug peptide was reconstituted in 0.1% formic acid (total volume 10µL) before being injected onto the column in 99% buffer A (0.1% FA) and 1% buffer B (0.1% FA in ACN). The flow rate was kept at 0.3µL/min throughout the gradient conditions (50mins linear gradient from 1% to 65% buffer B, a 2 mins linear gradient from 65-85% buffer B, and a final 8mins gradient from 85% buffer B). A blank injection was performed between each lysate sample to avoid potential carry-over. A blank injection was performed after every third injection of purified recombinant protein.

In PRM mode, a predefined precursor m/z mass for target peptides and inclusion list was generated. The peptide transition list was downloaded from SRMAtlas (http://www.srmatlas.org/), where peptide length is >8 amino acids in length, no reported post-translational modification and charge state of +2 and +3 included in the inclusion list and scheduled for PRM runs.

All peptide transitions were checked for proteotypicity using BLASTP (UniProt) and the unicity checker in NextProt[1] and confirmed to be unique for either αvβ6 or uPAR. In instrument settings, Orbitrap resolution was set at 17,500 with a target AGC value of $1e^6$, maximum fill time of 250ms, and an isolation window of 2.0m/z. Normalised collision energy was kept at 30 after multiple optimisations.

Data analysis

All raw data files from recombinant proteins, cell lysates and plasma upon subjecting to PRM assay were imported and processed in Skyline (v. 4.2) (MacLean et al., 2010), a freely available tool for PRM analysis.

First, the FASTA sequence of both proteins was uploaded to the software, along with basic information for data quality checking in order to extract the chromatograms from raw data files. This information included, setting up the peptide, protein and transition parameters such as digestion was specified as trypsin, with carboxyamidomethylation modifications enabled. In transition settings (under filter tab) precursor charges were kept 2, and 3; ion charges as 1, 2, 3 and ion types as y, b and p. Under the full scan tab, MS/MS filtering was 'targeted', product mass analyser as 'QIT' with a resolution of 0.7m/z. Peak integration and data generation and normalisation was performed automatically by Skyline based on default integration parameters. Regression model to calculate the sensitivity of the PRM assay and an unpaired t-test determine the significance (p-value) of uPAR expression between recombinant and cell lysate samples were performed using GraphPad Prism (version Prism 7).

5.3 Results

Proteotypic peptide selection and determination of optimal ion intensity peak for uPAR and β6 PRMs

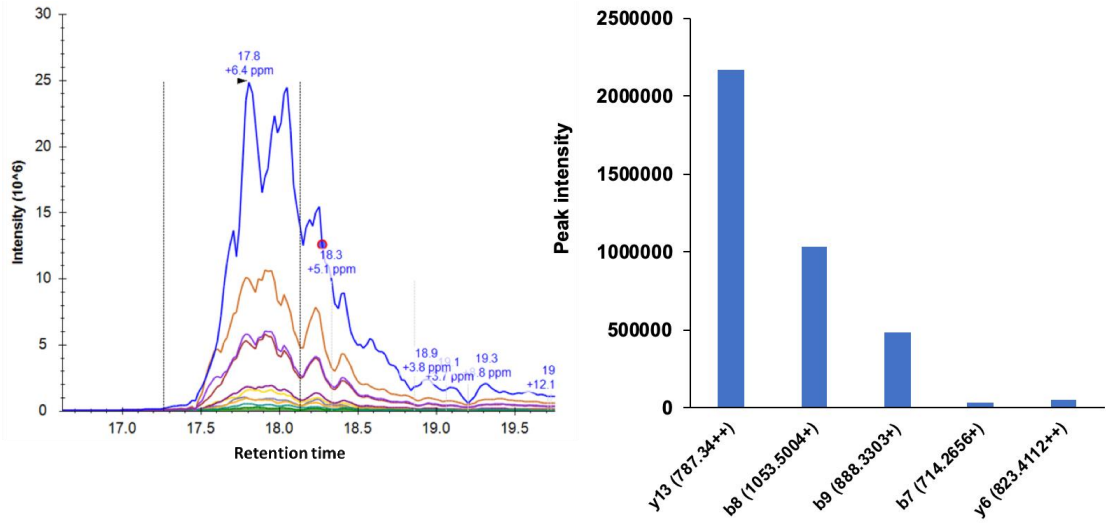
The main objective of this study was to select proteotypic peptides that allowed unambiguous identification and quantitation of the cancer-associated cell-surface proteins uPAR and β6. This step was key for the successful development of PRM assays.

Our aim was to select at least five peptide transitions for each of the two proteins. From 100ng of digested recombinant protein, PRM analysis identified multiple (13) peptides from uPAR and 10 from β6, respectively. Representative peaks for uPAR and β6 peptides are shown in Figures 5.2 and 5.3 with transition data outlined in Table 5.1. The average peak area of each daughter ion was compared for each peptide to determine highest intensity ion for each peptide and is shown as mean ± standard error (SEM) as part of routine data curation and where selected peptides were required to be observed in every replicate for continued investigation.

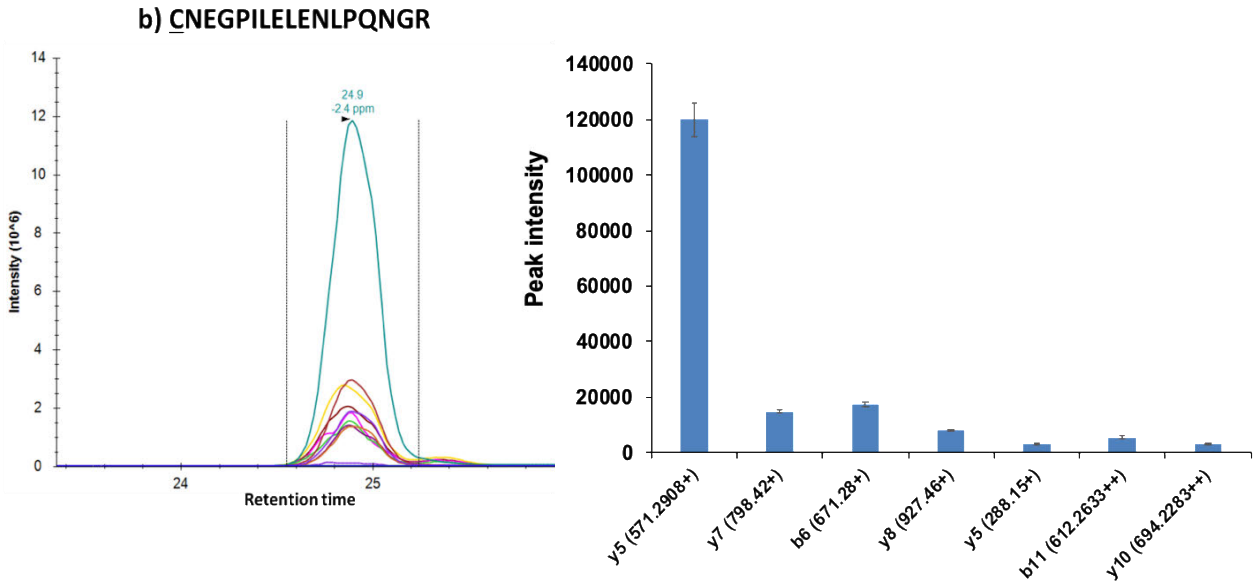
uPAR peptides: From the 13 peptides (downloaded from SRM Atlas, Table S5.1), scheduled for parallel reaction monitoring, four peptides demonstrated reproducible transitions with top-ranking peak intensity for daughter ions. The selection of top-ranked daughter ions of respective peptides was based on the p-value significance of peak intensities of all the transitions detected. The summary data and statistical evaluation for daughter ions ranked from highest to lowest intensity are collated in Table 5.1a. Four uPAR transitions with significantly higher intensity compared to other uPAR peptides were for the peptides

GNSTHGCSSEETFLIDCR (690.63), CNEGPILELENLPQNGR (651.66), SGCNHPDLDVQYR (520.90) and LWEEGEELELVEK (801.90) (Figures 5.2A, 5.2B, 5.2C, and 5.2D, respectively). The reproducible identification of these four peptides was confirmed in a separate validation experiment (Supp Figure S5.1 and Table S5.2).

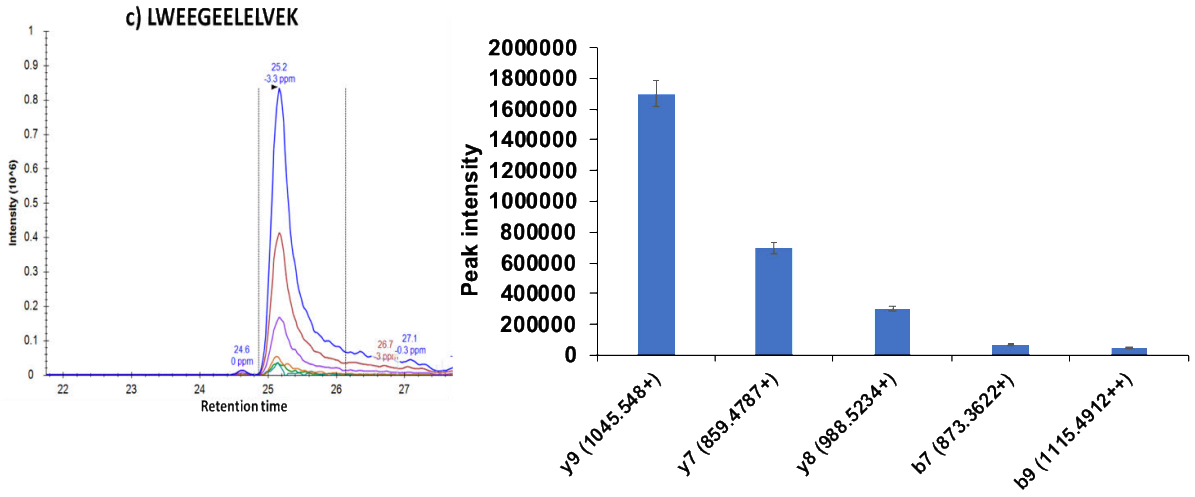
A) Peptide: SGCNHPDLDVQR; m/z = 520.9, charge = 3+



B) Peptide: CNEGPILELENLPQNGR; m/z = 651.66, charge = 3+



C) Peptide: LWEEGEELELVEK; m/z = 801.9, charge = 2+



D) Peptide: GNSTHGCSSEETFLIDCR; m/z = 690.63, charge = 3+

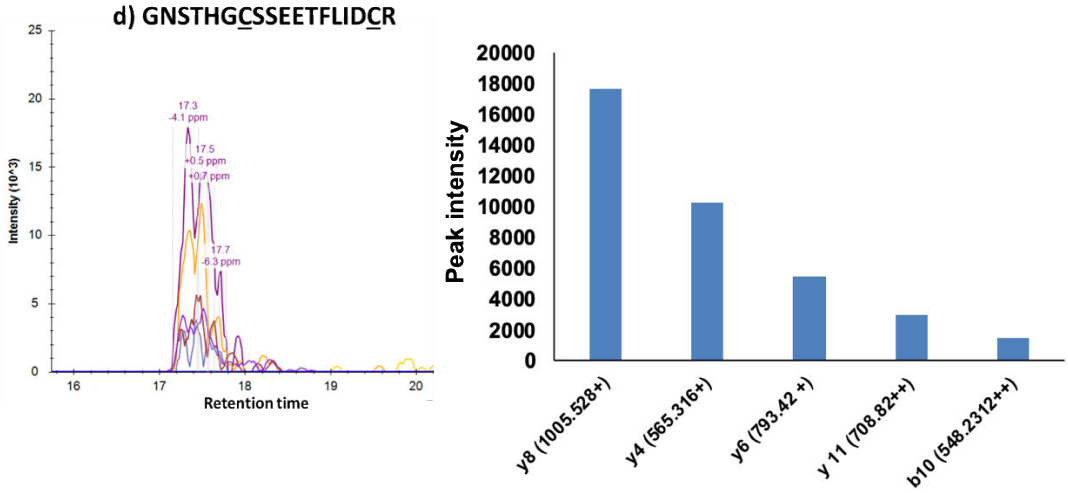
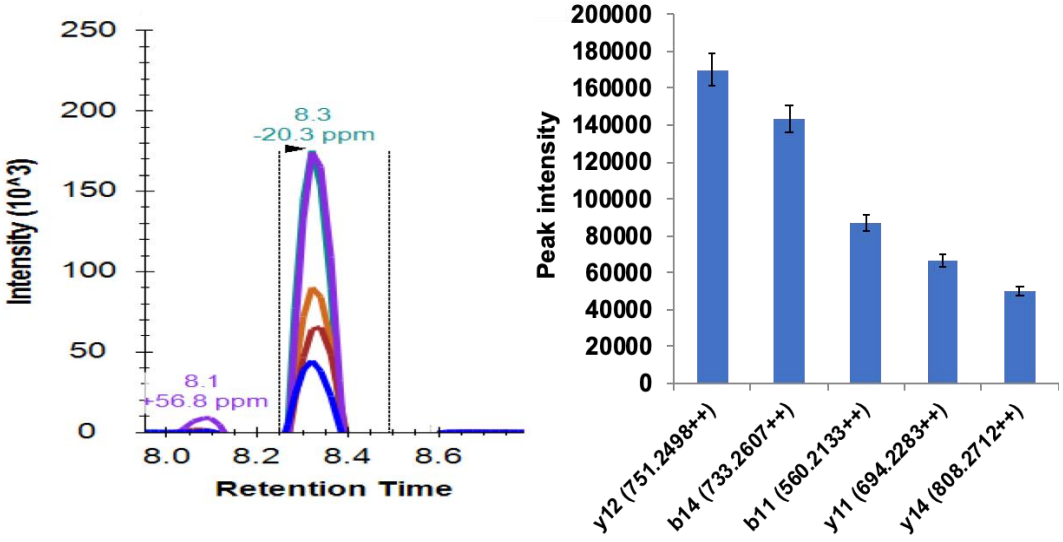


Figure 5.2: *Representation of uPAR peptide peak intensity and detected transitions: (A-D) The peak area of all reproducibly detected daughter ions from each peptide was compared in triplicates. Mean and SEM is shown in histograms below. Histograms without error bars were not run in triplicates*

Table 5.1: uPAR and αvβ6 transitions detected in recombinant proteins in triplicates					
a) uPAR					
Peptide sequence	**m/z**	**Charge**	**Daughter ions observed**	**Maximum Peak Intensity**	**Top rank ion**
GNSTHGCSSEETFLIDCR	690.63	3	5	3000	b13+
CNEGPILELENLPQNGR	651.66	3	7	$8*10^4$	y5+
SGCNHPDLDVQYR	520.9	3	3	$1*10^4$	y4+
LWEEGEELELVEK	801.9	2	2	2	-
b) αvβ6					
GLLCGGNGDCDCGECVCR	686.923	3	5	$1.7*10^5$	y12+
SCIECHLSAAGQAR	520.573	3	-	$8*10^4$	y1+

β6 peptides: From the 10 peptides procured from SRMAtlas (Table S5.1), two peptides showed reproducible transitions with top ranking peak intensity of daughter ions. Again, the selection of top-ranked daughter ions of respective peptides was based on the p-value significance of peak intensities of all the transitions detected. The summary data and statistical evaluation is outlined in Table 5.1b. The β6 peptides SCIECHLSAAGQAR (520.57) and GLLCGGNGDCDCGECVCR (686.92) had statistically significant top ranked ions (Figure 5.3a and 5.3b). The reproducible identification of these two peptides was confirmed in a separate validation experiment (Supp Figure S5.1 and Table 5.2).

A) Peptide: GLLCGGNGDCDCGECVCR; m/z = 686.9233, charge = 3+



B) Peptide: SCIECHLSAAGQAR; m/z = 520.5733, charge = 3+

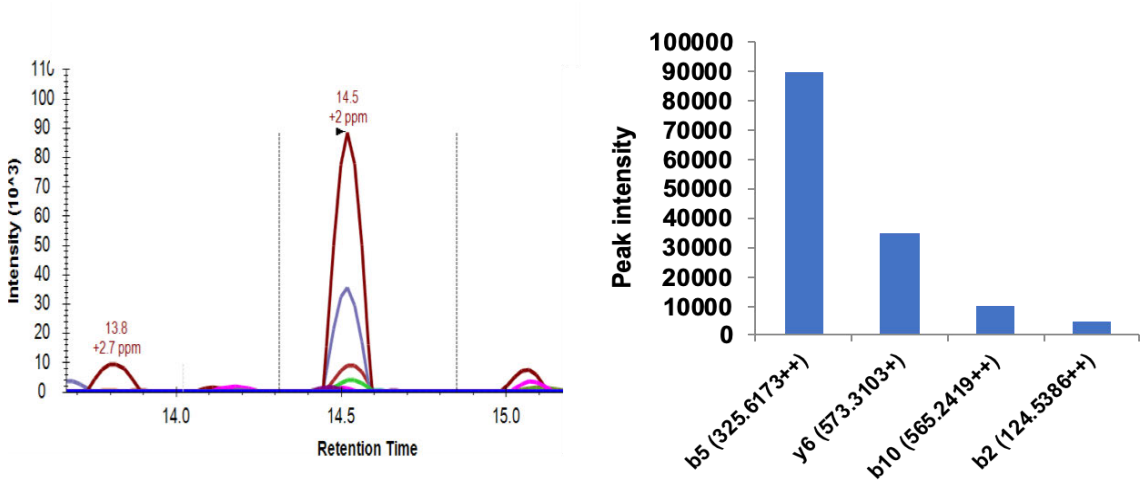


Figure 5.3: *Representation of αvβ6 peptide peak intensity and detected transitions: (A-B) The peak area of all reproducibly detected daughter ions from each peptide was compared in triplicates. Mean and SEM is shown in histograms below.*

uPAR transition limit of detection (LoD)

The LoD was determined using the linear range of recombinant uPAR transitions by measuring relative responses of an increasing concentration of digested recombinant uPAR (5, 2.5, 1 and 0.1µg/mL). Samples were run in triplicate and data analysed by linear regression to assess accuracy of correlation between concentration and peak intensity. Reproducibly observed transitions with highest peak ion are shown as mean ± SEM (Figure 5.4). The lowest concentration point that determines a confident peak was defined as the lowest limit of detection. The linearity was evaluated by the correlation coefficient (R^2) of the standard curve of multiple transitions observed. Using R^2 to assess the best fit of data to a straight-line curve assumes that the data have a constant variance over the whole range of concentrations. Respective R^2 and LoD of each transition are listed in Table 5.2. In terms of cartesian coordinates, the average correlation coefficient calculated between peptide concentration and peak area 0.95. GNSTHGCSSEETFLIDCR uPAR peptide, detected with high confidence in previous experiments, was not confidently detected in sensitivity experiment.

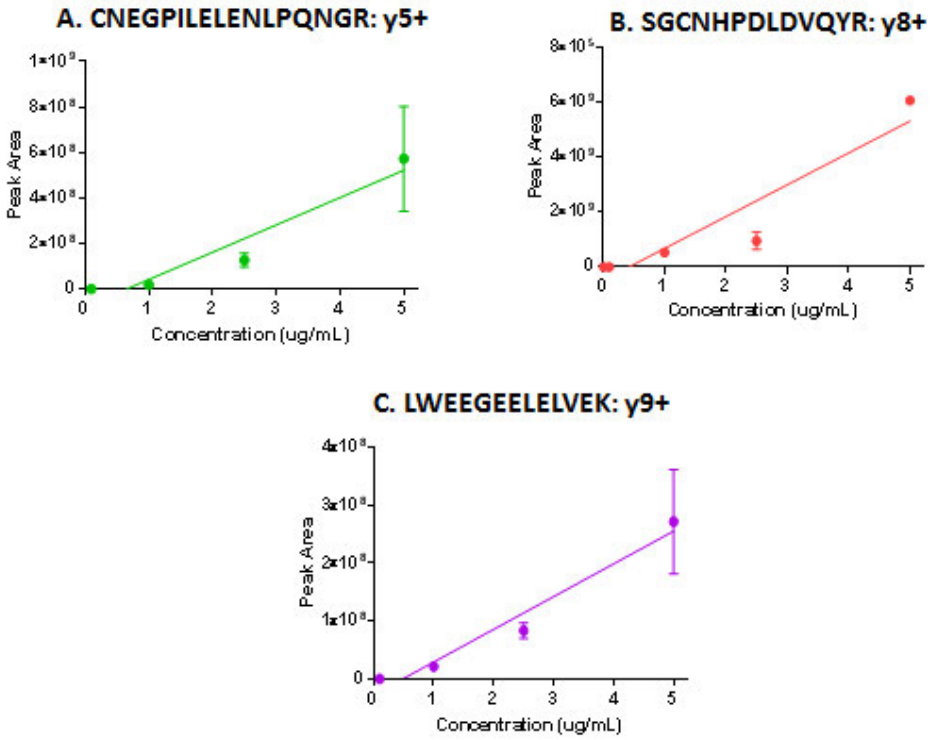


Figure 5.4: Determination of linearity and limit of detection for high intensity uPAR transitions. *A best-fit regression model assesed linearity and the last confident peak taken as the LoD. The transitions are shown as peptide sequence and top intensity daughter ion. Experiments to generate this data were performed by Ms Sachini Fonseka and were previously presented in her M.Res thesis. This figure has been reproduced and adapted from (Fonseka) after seeking permission.*

Table 5.2: Correlation coefficient of two linear models of uPAR transitions response curve and its limit of detection (LoD).

Transitions	**Correlation coefficient**		**LOD (µg/mL)**
	Best fit	**(x=0, y=0)**	
CNEGPILELENLPQNGR: y5+	0.92	0.8746	0.1
SGCNHPDLDVQYR: y8+	0.9067	0.8881	0.01
LWEEGEELELVEK: y9+	0.962	0.9346	0.1

Experiments to generate this data were performed by Ms Sachini Fonseka and were previously presented in her M.Res thesis. This table has been reproduced and adapted from (Fonseka) after seeking permission.

β6 transitions limit of detection (LoD)

This experiment remains at a preliminary stage, as samples were only performed in singlicate in the concentration range of 5, 2.5, 1.25, 0.08 and 0.04µg/mL. The analysis and presentation of data for observed β6 transitions was the same as described with the exception discussed below. The peptide GLLCGGNGDCDCGECVCR (686.92) detected in the previous experiment was not confidently detected. Conversely, peptides SCIECHLSAAGQAR, LGFGSFVEKPVSPFVK, GCQLNFIENPVSQVEILK, and HILPLTNDAER were detected. The correlation coefficient (R^2) and LoD of each transition are listed in Table 5.3. The average correlation coefficient between peak area and peptide concentration was 0.93 with a best-fit regression fit (Figure 5.5). Additional repeats are required for confident analysis.

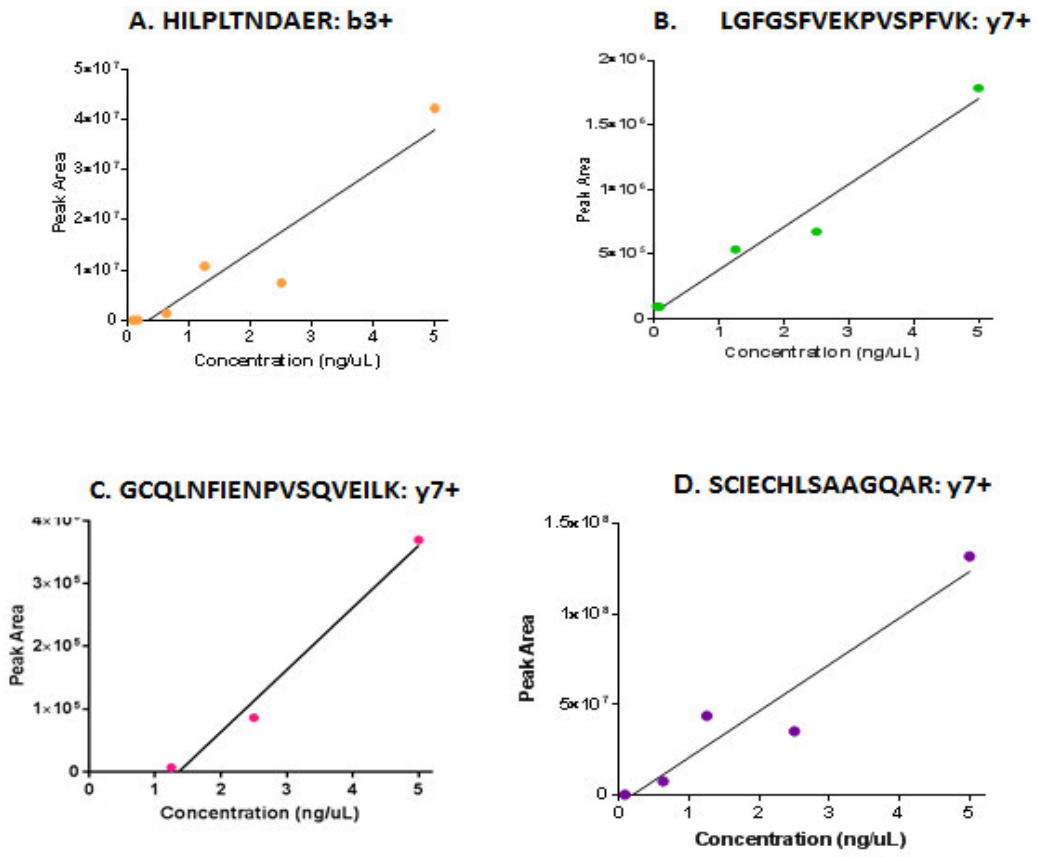


Figure 5.5: Determination of linearity and limit of detection for high intensity identified transitions of β6. *A best-fit regression model assesed linearity and the last confident peak taken as the limit of detection (LoD). The transitions are shown as peptide sequence and top intensity daughter ion. Experiments to generate this data were performed by Ms Sachini Fonseka and were previously presented in her M.Res thesis. This figure has been reproduced and adapted from (Fonseka) after seeking permission.*

Table 5.3: The correlation coefficient of two linear models of the β6 transitions standard curve and its limit of detection (LoD).

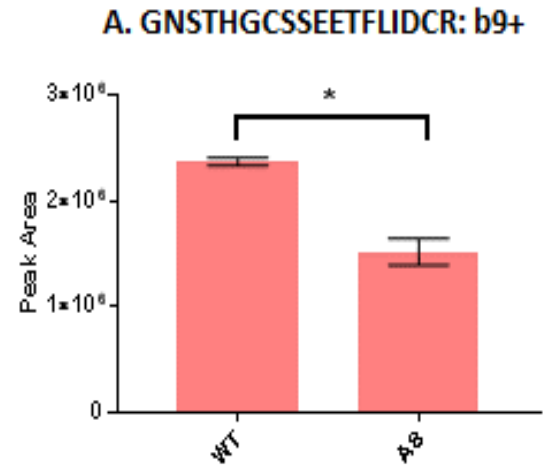
Transitions	**Correlation coefficient**		**LOD (µg/mL)**
	Best fit	**(x=0, y=0)**	
HILPLTNDAER: b3+	0.8937	0.8759	0.08
LGFGSFVEKPVSPFVK: y7+	0.9719	0.9682	0.04
GCQLNFIENPVSQVEILK: y7+	0.9168	0.8977	1.25
SCIECHLSAAGQAR: y7+	0.9144	0.9096	0.08

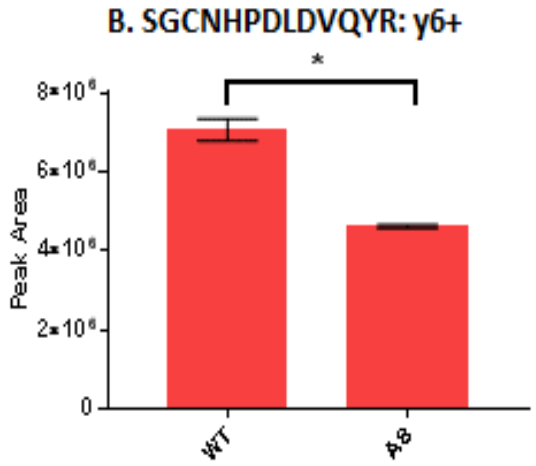
Experiments to generate this data were performed by Ms Sachini Fonseka and were previously presented in her M.Res thesis. This figure has been reproduced and adapted from (Fonseka) after seeking permission.

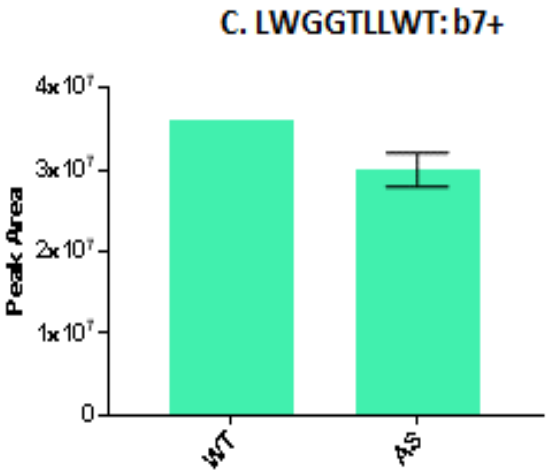
uPAR PRM-MS assay validation in HCT116 whole cell lysates

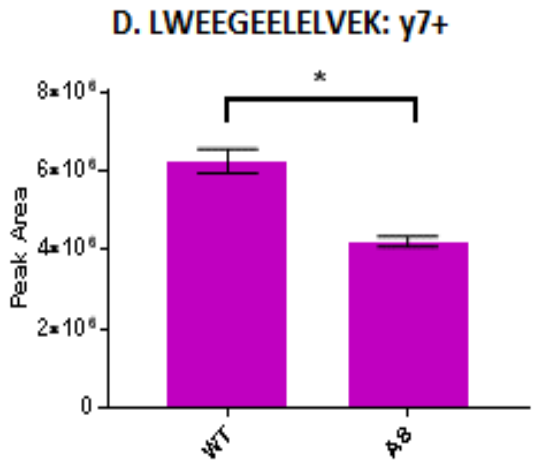
The aim of this experiment was to test applicability of the uPAR PRM assay using the HCT116 cell line. Furthermore, the sensitivity was tested using the HCT116 AS cell line which has a reported 35-45% knock-down of uPAR protein (Liu *et al.*, 2014). Ten transitions and six uPAR peptides were detected from the inclusion list (Figure 5.6, Table 5.4). The sequence location of detected peptides within the mature uPAR protein map to each of the three domains of uPAR - one from each of domain D1 and the so-called domain linker region D2D3 and four from domain D3. This data clearly indicated the presence of many peptides and hence full-length uPAR protein in HCT116 whole cell lysates. The six peptides detected were consistently observed including from analysis of recombinant protein samples in both experiments. In contrast to the multiple daughter ions per peptide observed in the recombinant analysis in the lysate, all peptides except CNEGPILELENLPQNGR and SGAAPQPGPAHLSLTITLLMTAR were detected by a single daughter ion (Figure 5.6E, 5.6F,5.6G,5.6H,5.6I, and 5.6J), Table 5.4. All transitions showed the expected reduced intensity in the AS lysate compared to WT, with a reduction range of 13% to 58%. Transition CNEGPILELENLPQNGR: b6+ had the highest reduction of 58% whereas SGAAPQPGPAHLSLTITLLMTAR: b8+ had the lowest of 13% (Table 5.4). On average, the transitions have a 33.8% decreased intensity in the AS cell line

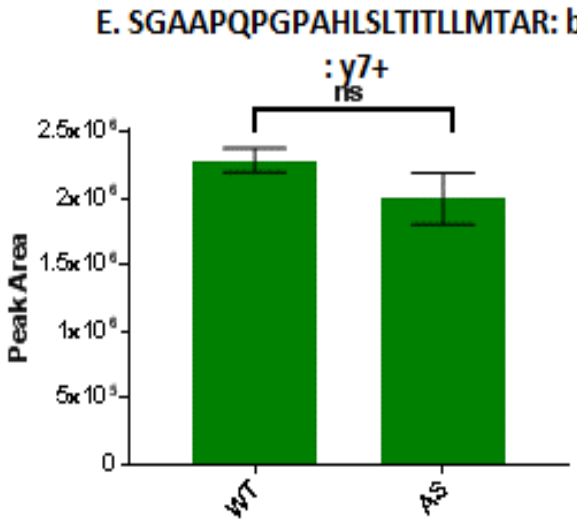
when compared to WT. This decrease in intensity was significant for 5 out of the 10 transitions (Table 5.4).

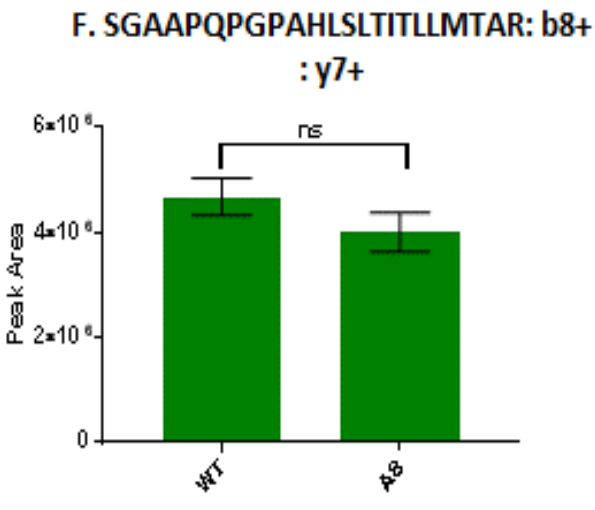












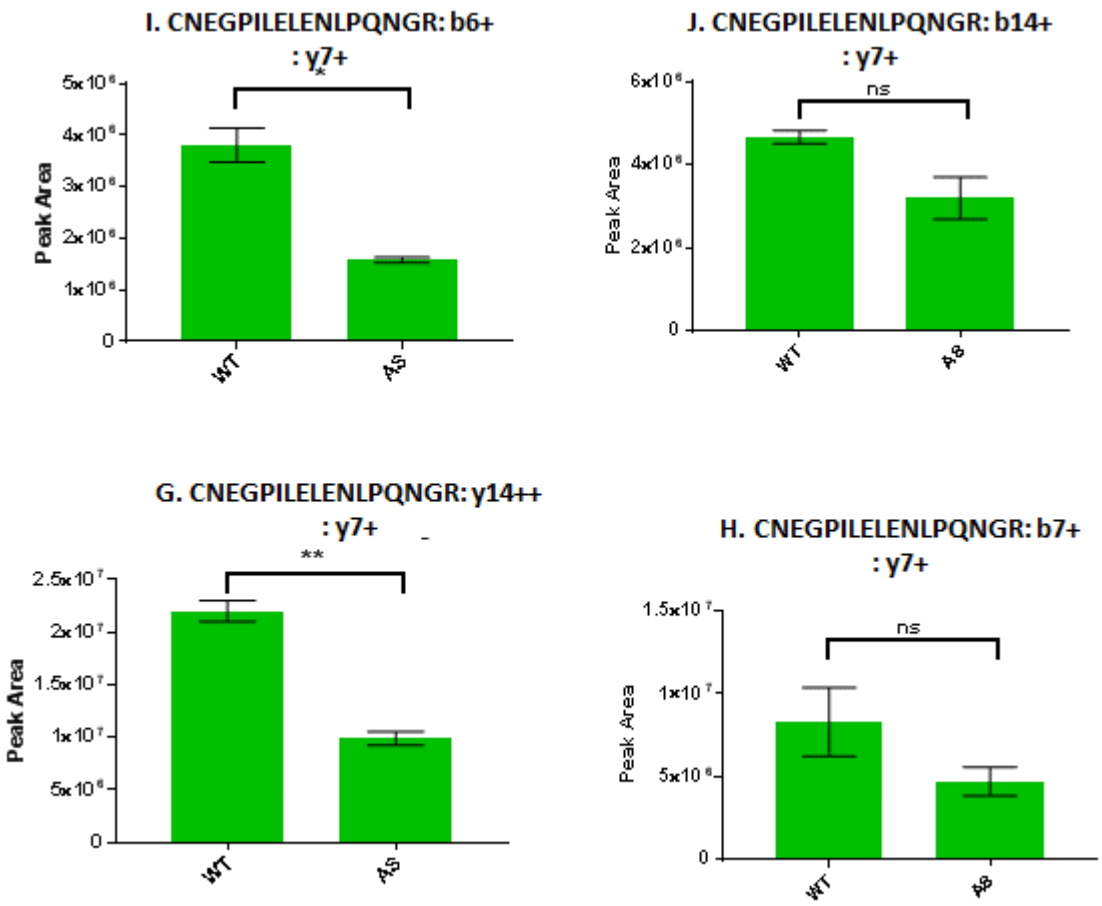


Figure 5.6: uPAR transitions detected in HCT116 WT and AS whole-cell lysates. *An unpaired t-test determined significance reduced expression of uPAR peptides between WT and AS cell lines. Experiments to generate this data were performed by Ms Sachini Fonseka and were previously presented in her M.Res thesis. This figure has been reproduced and adapted from (Fonseka) after seeking permission.*

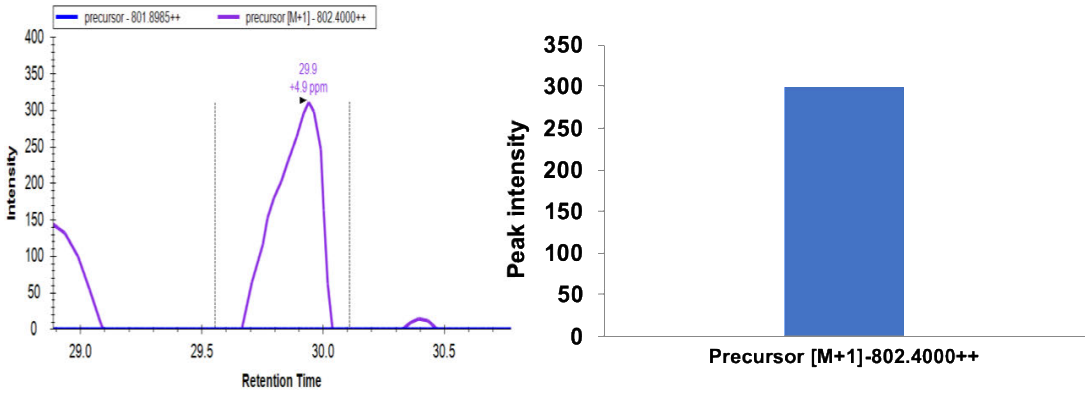
Table 5.4: uPAR transitions detected in HTCC WT and AS lysate					
Transitions		**WT**	**AS**	**% decrease**	**Comparison**
GNSTHGCSSEETFLIDCR	b9+	$2.4x10^6 \pm 3.7x10^4$	$1.5x10^6 \pm 1.2x10^5$	37	*p=0.02**
CNEGPILELENLPQNGR	y14++	$2.2\ x10^7 \pm 9.8\ x10^5$	$9.9\ x10^6 \pm 6.5\ x10^5$	55	*p=0.009***
CNEGPILELENLPQNGR	b7++	$8.3\ x10^6 \pm 2.1\ x10^6$	$4.7\ x10^6 \pm 8.6\ x10^5$	43	*p=0.25*
CNEGPILELENLPQNGR	b14	$4.6\ x10^6 \pm 1.6\ x10^5$	$3.2\ x10^6 \pm 5.1\ x10^5$	31	*p=0.11*
CNEGPILELENLPQNGR	b6+	$3.8\ x10^6 \pm 3.3\ x10^5$	$1.6\ x10^6 \pm 5.2\ x10^4$	58	*p=0.02**
SGCNHPDLDVQYR	y6+	$7.1\ x10^6 \pm 2.8\ x10^5$	$4.6\ x10^6 \pm 4.3\ x10^4$	34	*p=0.01**
LWGGTLLWT	b7+	$3.6\ x10^7$	$3.0\ x10^7 \pm 3.0\ x10^6$	17	*N/A*
LWEEGEELELVEK	y7+	$6.3\ x10^6 \pm 3.0\ x10^5$	$4.2\ x10^6 \pm 1.4\ x10^5$	32	*p=0.03**
SGAAPQPGPAHLSLTITLLMTAR	b8+	$4.7\ x10^6 \pm 3.5x10^5$	$4.0\ x10^6 \pm 3.7\ x10^5$	18	*p=0.32*
SGAAPQPGPAHLSLTITLLMTAR	b4+	$2.3\ x10^6 \pm 9.0x10^4$	$2.0\ x10^6 \pm 1.9\ x10^5$	13	*p=0.30*
Experiments to generate this data were performed by Ms Sachini Fonseka and were previously presented in her M.Res thesis. This table has been reproduced and adapted from (Fonseka) after seeking permission.					

Proteotypic peptide identification for quantitation of uPAR and αvβ6 in human plasma samples

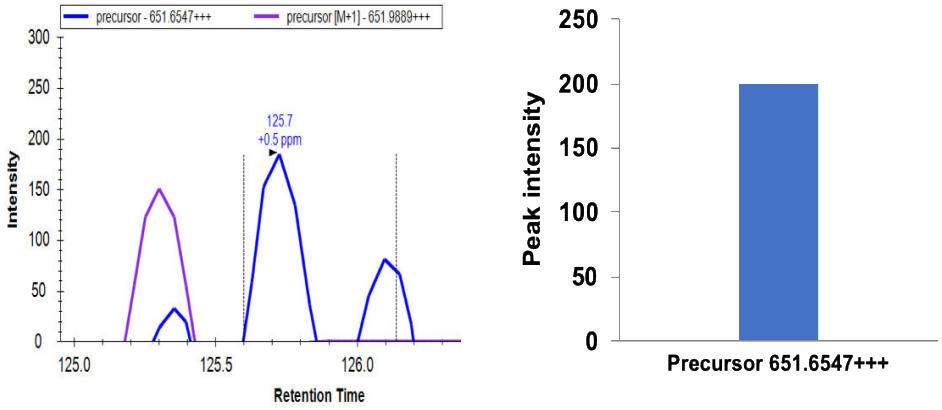
The detection of uPAR and αvβ6 peptides from human plasma samples was performed using 2ug and 5ug of plasma samples. 2ug of undepleted plasma did not show any data in spectral graphs and therefore, the amount of protein was increased from 2ug to 5ug.

uPAR: 5ug of plasma sample yielded precursor ion peaks for five peptides, four of which matched with the peptides identified from recombinant proteins (Figure 5.2). Peptide, LWEEGEELELVEK; m/z = 801.8985, charge = 2+ was reproducibly identified from recombinant uPAR digest (Figure 5.2 and Supp Figure 5.1) with 6 reported transitions and a maximum peak intensity of $6x10^8$, whereas in 5ug human digested plasma, this same peptide was identified with peak intensity of 300 and with one precursor ion peak. Peptide CNEGPILELENLPQNGR, m/z = 651.66, charge = 3+ was detected with multiple transitions and high peak intensity in recombinant proteins (Figure 5.2 and Supp Figure 5.1), but a single peak of peptide was observed in plasma with an intensity of 200 and not a single transition was identified (Table 5.5 and Supp Figure 5.5). Further, peptides GNSTHGCSSEETFLIDCR; m/z = 690.6266; charge = 4 and peptide SGCNHPDLDVQYR; m/z = 520.9002; Charge = 3+ were identified with one precursor ion peak. It is important to note that no transition was observed for these two peptides from 5ug of plasma. Another peptide VEECALGQDLCR, m/z = 725.3267 and charge = 2 + was identified with maximum peak intensity of $3x10^4$ (Table 5.5 and Figure 5.7). However, no clear transition peaks were observed for this peptide.

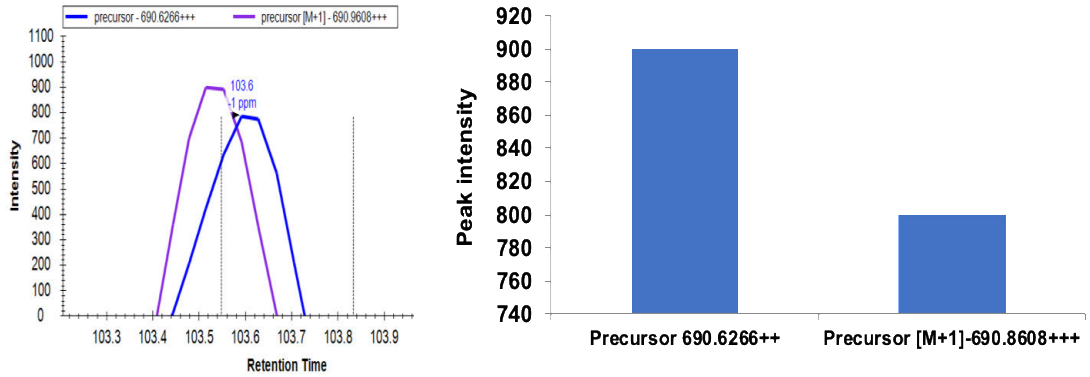
A) Peptide: LWEEGEELELVEK; m/z = 801.8985, charge = 2+



B) Peptide: CNEGPILELENLPQNGR, m/z = 651.66, charge = 3+



C) Peptide: GNSTHGCSSEETFLIDCR; m/z = 690.6266; charge = 4+



D) Peptide: VEECALGQDLCR; m/z = 725.3267; charge = 2+

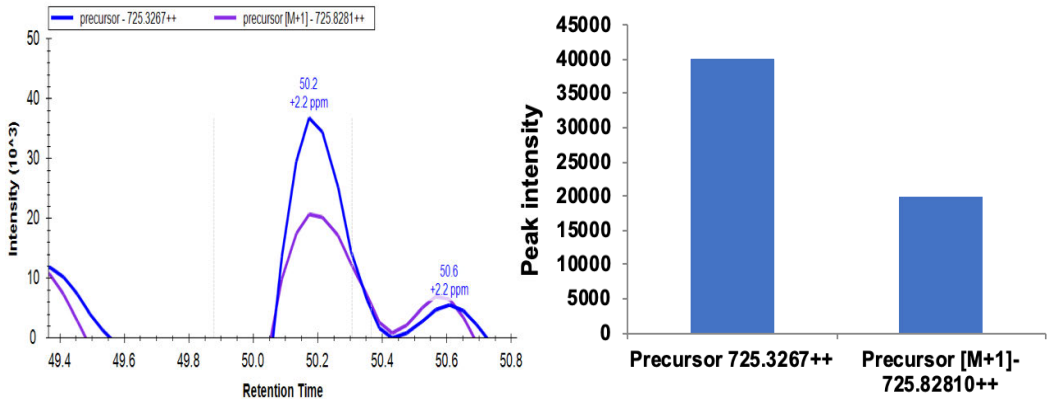
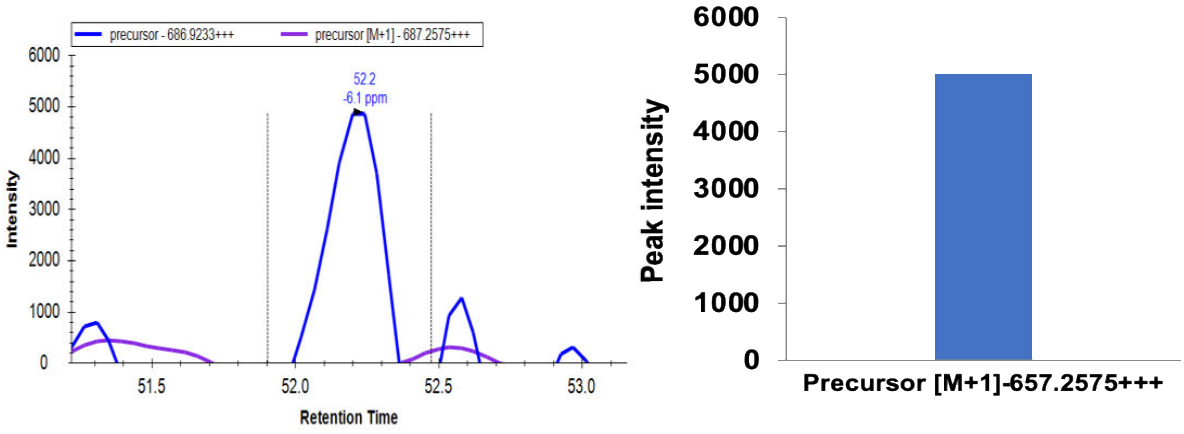


Figure 5.7: *uPAR peptide peak intensity and detected transitions from 5µg of trypsin digested non-depleted plasma: (A-D) The peak area of all reproducibly detected daughter ions from each peptide was compared in triplicates. Mean is shown in histograms.*

β6: The precuror ion of peptide GLLCGGNGDCDCGECVCR, m/z = 686.92, charge = 3+ was identified with a peak intensity of 5000 and only one daughter ion (Table 5.5 and Supp Figure 5.3). In comparison to recombinant protein data, this peptide was reproducibly detected from αvβ6 recombinant proteins with 2 transitions and maximum peak intensity of $1x10^4$. Peptide SCIECHLSAAGQAR, m/z = 520.5733, charge = 3+ only shows a straight line that doesn't qualify as a valid peak. Peptide WQTGTNPLYR, m/z = 309.6596, charge = 4+ was identified with maximum peak intensity of 450 (Table 5.5 and Figure 5.8). The data for both uPAR and αvβ6 peptides has been tabulated in Table 5.5.

A. Peptide 1: GLLCGGNGDCDCGECVCR, m/z = 686.92, charge = 3+



B. Peptide 2: WQTGTNPLYR, m/z = 309.6596, charge = 4+

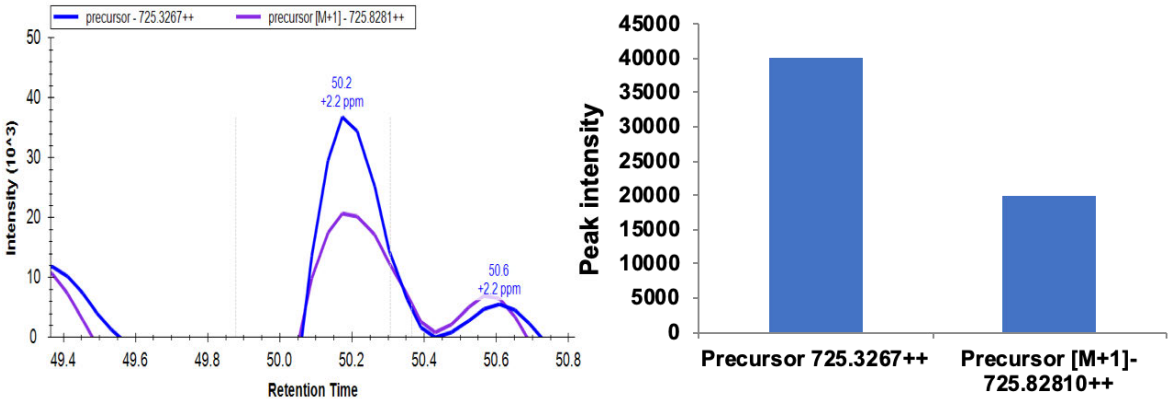


Figure 5.8: *αvβ6 peptide peak intensity and detected transitions from 5μg of trypsin digested non-depleted plasma: (A-B) The peak area of all reproducibly detected daughter ions from each peptide was compared in triplicates. Mean is shown in histograms.*

Table 5.5: uPAR and αvβ6 transitions detected in trypsin digested crude human plasma					
uPAR					
Peptide sequence	**m/z**	**Charge**	**Daughter ions observed**	**Maximum Peak Intensity**	**Top rank ion**
GNSTHGCSSEETFLIDCR	690.63	3	-	800	-
CNEGPILELENLPQNGR	651.66	3	-	200	-
SGCNHPDLDVQYR	520.9	3	-	-	-
LWEEGEELELVEK	801.9	2	-	300	-
VEECALGQDLCR	725.3267	2	-	$4*10^4$	-
αvβ6					
GLLCGGNGDCDCGECVCR	682.92	3	-	5000	-
SCIECHLSAAGQAR	520.5733	3	-	-	-
WQTGTNPLYR	309.6596	4	-	450	-

5.4 Discussion

uPAR and αvβ6 are established EMT drivers and have been suggested to orchestrate cancer metastasis (Cantor et al., 2015; Lester et al., 2007). Several studies have associated elevated levels of uPAR and β6 in plasma independently with cancer progression (Bengs et al., 2013; Boonstra et al., 2011).

In this study, the aim was to design/develop robust LC-PRM assays as alternatives to immunoassays. These could be used to accurately monitor levels of uPAR and β6 as markers of the metastatic CRC phenotype and identify the EMT. A pilot study to verify and quantify uPAR and β6 from HCT116 cell lysates was performed.

The first step was to reproducibly identify highest product ions, number of transitions and highest peak intensities from a complete digest of recombinant proteins. Multiple runs of the proteins are expected to produce equimolar peptides in multiple runs following no other variations is expected to produce equimolar peptide. However, differences in peak intensity (or relative abundance) were observed, which was likely due to the peptide performance variability in the MS (Keshishian et al., 2007). Thus, peak intensity was used as a measure of transition performance. Low-intensity transitions at a high concentration (10µg/ml in the transition detection experiment) are more likely to be undetectable in a plasma sample with much lower (~ng/mL) target protein (Keshishian et al., 2007; Liu et al., 2013).

Peptides that produce high ion-current responses and high abundance fragment ions are likely to have the best detection sensitivity (Fusaro et al., 2009). Therefore, all ions detected from each peptide were compared using an unpaired t-test analysis to determine the highest intensity transition for each peptide. uPAR peptides from the recombinant protein, YLECISCGSSDMSCER (651.92), NQSYMVR (449.22) LGDAFSMNHIDVSCCTK (652.29) and NSSDIVQIAPQSLILK (863.98) (Table S5.1) were never detected in scheduled PRM. This was expected considering that peptides can have varying mass spectrometry detectability due to varying ionisation efficiencies (Keshishian et al., 2007). Additionally, poor chromatography, solubility problems, matrix interference and failure of recovery from digest can all effect the peptides detection (Jaffe et al., 2008).

Further, the most important factor to be considered while designing a PRM-based assay is selection of a target peptide (Liebler and Zimmerman, 2013). The peptides when subjected to PRM enter the Q1 quadrupole that filters peptides of interest based on m/z. The mass inclusion list used for scheduled-PRMs contains m/z of signature proteotypic peptides and can be developed by in-house DIA experiments or identification from data repositories such as MRMaid (Mead et al., 2009) and SRMAtlas (Kusebauch et al., 2014) or *in silico* computational methods such as OpenMS/TOPP (Nahnsen and Kohlbacher, 2012).

In this study, SRMAtlas, an extension of PeptideAtlas, was used to generate inclusion lists for uPAR and β6. After performing experiments, selected peptide transitions were evaluated on basis of mass spectrometric characteristics such as length of peptides, number of amino acids (aa), hydrophobicity, proteotypicity, number of observations and known single nucleotide polymorphisms (SNPs) (Uchida et al., 2013). All reproducible peptides observed were between 10-16 aa in length and observations from multiple experiments has been reported by PeptideAtlas. Recommendations for peptide evaluation includes a hydrophobicity check as water-soluble peptides are preferred for optimal LC performance (Keshishian et al., 2007), therefore, peptide hydrophobicity was calculated and confirmed to be less than 40% using an online tool Peptide 2.0. The transitions selected for uPAR and β6 fulfil the aforementioned criteria of detectable signature proteotypic peptides for PRM, though, their detectability with plasma matrix as a background is yet to be determined. As stated above, it is important to select peptides that have previously been observed in other MS/MS experiments, as this confirms their 'detectability/flyability' and utility for successful PRM assay development. However, most β6 and uPAR data in SRMAtlas is derived from tissue and cell line studies, suggesting

that these two proteins are understudied in plasma or may have limited detection due to their plasma low abundance.

In PeptideAtlas, two (2) integrin β6 and eleven (11) uPAR peptides were previously observed by the diagnostic company Roche in plasma proteome studies and through other unpublished data. The transition optimisation cannot solely rely on information derived from SRMAtlas or PeptideAtlas as endogenous peptides show different properties in different matrices. The best strategy would be to include peptides generated from SWATH™-MS or DIA or label-free MS studies, but that will also require further transition optimisation to develop a comprehensive PRM assay.

The quality of a 'fit for purpose' diagnostic assay is based on sensitivity, peptide stability and requirement for standardisation. In order to characterise the analytical performance of a PRM assay, it is important to establish, via a systematic process, that endogenous peptides are efficiently performing in clinical settings (Shrivastava et al., 2011). This is done via determination of LoQ and LoD, which measure the lowest concentration of endogenous peptide that can be measured by mass spectrometry. LoD is an important step in discriminating between the presence and absence of a peptide from the biological sample while LoQ can reliably measure low level of peptides for PRM assay. The recombinant proteins used in this study assessed the fragmentation pattern of the peptide and the transition stability (variables determined through LoD/LoQ).

It is important to note that the chromatographs and fragmentation pattern of endogenous peptides is identical to that of a peptide derived from recombinant protein. This is a part of method evaluation as establishing these parameters can increase the robustness and the statistical confidence of an experiment (Mani et al., 2012). It accounts for the variability in various concentration ratios which might be encountered while working with endogenous samples. Further, in a step towards developing a reproducible PRM assay, it is a common strategy to use isotopically labelled peptides as internal standards. This technique is widely known as stable isotope dilution (SIS) (Ozcan et al., 2017). The stable isotope peptides correspond to endogenous peptide of interest and are synthetically produced with a heavy arginine or lysine. The endogenous sample is spiked with stable isotope peptides in a 1:1 ratio. Again, owing to similar chromatographic, ionisation, and fragmentation patterns with endogenous peptides, spiking the samples helps in absolute and relative quantification of endogenous peptides by measuring LoD and LoQ (Ozcan et al., 2017). However, it is important

to note that both peptide selection and transitions are affected by the biological matrix, and the experiment. Therefore, each study needs to be optimised to the highest sensitivity and specificity based upon biological samples and specific instrument settings (Ozcan et al., 2017). In summary, determination of LoD and LoQ tests using recombinant protein and use of isotope peptides as internal standards is imperative for method development especially for studies where PRM assays are to be deployed into large scale clinical samples (Ozcan et al., 2017).

The sensitivity or LoD of the transitions was determined by assessing dynamic range and the linearity of uPAR and β6 transitions. Many peptides from uPAR or β6, detected in the previous experiment, were not detected in the LoD experiment. This may be attributed to peptide degradation within the prepared protein digest stock. Here, %CV calculation for multiple PRM runs at various time points and between freeze/thaw cycles according to CPTAC guidelines can assist development of a robust and reproducible assay. In this study, the sensitivity of identified peptides was evaluated by best-fit linear regression. This strategy was employed because of the broad concentration range used for uPAR and β6, and because MS standard curves are known to lose linearity at very high concentrations (Tang et al., 2004; Tang and Kebarle, 1993). The best fit regression was accurate at modelling higher concentrations but accurate modelling at lower concentrations is yet to be determined.

Hence, more experiments are required to complete determination of sensitivity and specificity of uPAR and β6 transitions using plasma as the sample matrix/background. Specifically (i) robust and reproducible determination of uPAR and β6 peptides, at more than 6 time points and between freeze/thaw sessions, (ii) the uPAR and β6 transitions with the lowest LoD should be re-identified concentrating on robust transitions observed in previous experiments, and (iii) determination of specificity and sensitivity of both proteins with plasma as a background and should be analysed for matrix interference. Upon successful completion of this, heavy labelled peptides should be employed in a series of future quantification experiments.

The proof-of-concept for uPAR PRM assay validation in CRC cell lysate

The aim of this experiment was to perform a proof-of-concept study to determine 1) uPAR recovery from whole lysates, 2) identify transitions using cell lysates as a matrix background and 3) test the sensitivity of the PRM assay to detect and verify decreased uPAR expression from HCT116 AS cell lines and mock wild type HCT116 (known to express unaffected uPAR levels) (Ahmed et al., 2003).

In total, ten (10) uPAR transitions and six (6) uPAR peptides were identified from whole cell lysate of both HCT116 WT and the HCT116 AS cell lines. The six peptides observed, originated from all three D1-D3 domains of uPAR, indicating the presence of full-length uPAR in our cell lysates and that HCT116 whole cell lysate preparation sufficiently recovered membrane GPI-anchored uPAR.

Further, it is important to comment on the transitions observed between endogenous (cell lysate) and recombinant uPAR. It was evident that the recombinant protein uPAR could be detected with multiple transitions whilst single daughter ion transitions were observed from cell lysate PRMs. This further corroborates variability in peptide fragmentation or ion detection in the presence of a complex background matrix (Barnes et al., 2011). Despite the variability observed in detection of daughter ions, similar and high peptide intensities were observed between cell lysates and recombinant uPAR. To further explain this, peptide CNEGPILELENLPQNGR was seen with highest intensity from the recombinant protein analysis and the same peptide had the 2nd highest intensity from cell lysates. Similarity in the intensity of recombinant and endogenous peptides indicates consistent ionization and behavior in MS, validating the use of recombinant proteins for assay optimization

It should be noted that not all peptides detected from HCT116 cell lysates were detected in the recombinant uPAR PRM analysis. This was not unexpected considering a lysate extraction is liable to protein loss, or degradation and biological variability (Feist and Hummon, 2015). Therefore, undetected uPAR peptides may either have been degraded, lost or not recovered after lysis. Additionally, HCT116-derived uPAR could have endogenous posttranslational modifications (PTMs) that cannot be observed in the recombinant protein (Hoofnagle and Wener, 2009). Fortunately, MRM/PRM-MS has the advantage that the known PTMs and isoforms can be targeted by extending the selection criterion for peptides (Liu et al., 2013). Therefore, if necessary, known uPAR glycosylation variants can be included in the PRM assay design (Ploug et al., 1991).

The PRM assay developed to detect uPAR differential expression between WT and AS HCT116 cells is adequately sensitive. The observed peptide intensity data suggested what we know from experience to be an ~33% decrease in uPAR expression in HCT116 AS cells as this closely parallels published protein knock-down data of ~27% (Ahmed et al., 2003; Liu et al., 2014). In addition, the use of the HCT116 CRC cell line confirms the applicability of the endogenous uPAR test to quantitatively measure this protein.

Measurement of uPAR and αvβ6 peptides in plasma samples

The aim of this study was to evaluate the capability of PRM assay to detect uPAR and αvβ6 from human plasma sample. Considering the complex nature of plasma, and interference caused by abundant proteins in MS studies, the first basic criteria adopted was to reproducibly and precisely detect uPAR and αvβ6 peptide transitions from non-depleted plasma.

In a PRM experiment, several single reaction monitoring transitions are monitored in a single assay. The reliability of the transitions and metrics of reproducibility are calculated using product ions of the peptides (i.e. uPAR and αvβ6) by assessing the reproducibility of signal peak intensity. The average of signal peak intensity for the m/z peaks observed for product ions is calculated across experiments including the %CV (coefficient of variance) and with standard deviation values. These statistics indicate the reproducibility of the fragment ions for a given precursor and also suggest number of PRM experimental runs needed to observe a reliable result for the transitions.

While peptides from endogenous proteins (uPAR and αvβ6) were reproducibly observed in the non-depleted human plasma samples, the transitions fragmentation pattern did not show good peak intensities. In this study, the uPAR peptide VEECALGQDLCR was observed with a peak intensity of 0.3 X 102 with no reproducible transitions and product ions. The detection of low intensity precursor ion peaks suggested the presence of circulating fragments of uPAR and αvβ6 peptides in plasma. Further optimisation is therefore required and may include changes to sample preparation and using plasma digests after MARS14 depletion. Since uPAR and αvβ6 are CRC stage II prognostic and metastatic markers, respectively (Ahn S B et al., 2014 and Ahn S B et al., 2015), using later stages of CRC (i.e. stage II and III) plasma digests could also allow reliable detection of these peptides, as both protein increase in expression with CRC progression.

5.5 Next Steps

In conclusion, while preliminary data using recombinant uPAR and αvβ6 proteins provide promising starting points to developing a PRM assay, translation of this assay to cell lysates and plasma have not met the robustness to reach clinical assay standards. While certain peptides and transitions can be reproducibility detected, the primary concern continues to be that of

detectable peak intensities in plasma. In addition to optimising sample processing and the PRM protocol itself, a possible alternative would be to explore and mimic an accurate model of an uPAR expressing patient tumour.

Clinical translation of any biomarker requires massive population screening, with success hinging upon achieving the best sensitivity and specificity with accurate determination of protein concentration. A PRM assay, once fully developed and optimised, stands to offer immediate implementation and cost-efficient verification of numerous markers and high throughput in larger cohorts (Anderson and Hunter, 2006; Gallien et al., 2011).

Further clinical evaluation of these markers may help in diagnosing CRC patients before progression to later stage disease or prior to initiation of the metastatic process and result in considerably improved patient survival.

References

Ahmed, N., Oliva, K., Wang, Y., Quinn, M., and Rice, G. (2003). Proteomic profiling of proteins associated with urokinase plasminogen activator receptor in a colon cancer cell line using an antisense approach. PROTEOMICS *3*, 288-298.

Ahn, S.B., Chan, C., Dent, O.F., Mohamedali, A., Kwun, S.Y., Clarke, C., Fletcher, J., Chapuis, P.H., Nice, E.C., and Baker, M.S. (2015). Epithelial and stromal cell urokinase plasminogen activator receptor expression differentially correlates with survival in rectal cancer stages B and C patients. PLoS One *10*, e0117786.

Ahn, S.B., Mohamedali, A., Anand, S., Cheruku, H.R., Birch, D., Sowmya, G., Cantor, D., Ranganathan, S., Inglis, D.W., Frank, R., *et al.* (2014). Characterization of the interaction between heterodimeric alphavbeta6 integrin and urokinase plasminogen activator receptor (uPAR) using functional proteomics. J Proteome Res *13*, 5956-5964.

Bandyopadhyay, A., and Raghavan, S. (2009). Defining the role of integrin alphavbeta6 in cancer. Curr Drug Targets *10*, 645-652.

Bengs, S., Spalinger, M.R., Lang, S., Boehmer, L.V., Weber, A., Vavricka, S.R., Frei, P., Knuth, A., Fried, M., Rogler, G., *et al.* (2013). Integrin αVβ6 (ITGB6) Is a Novel Serum Tumour Marker in Colorectal Cancer (CRC) Patients and Is Associated With Invasion and Metastasis of CRC. In American Gastroenterological Association.

Boonstra, M.C., Verspaget, H.W., Ganesh, S., Kubben, F.J., Vahrmeijer, A.L., van de Velde, C.J., Kuppen, P.J., Quax, P.H., and Sier, C.F. (2011). Clinical applications of the urokinase receptor (uPAR) for cancer patients. Curr Pharm Des *17*, 1890-1910.

Brabletz, T., Hlubek, F., Spaderna, S., Schmalhofer, O., Hiendlmeyer, E., Jung, A., and Kirchner, T. (2005). Invasion and metastasis in colorectal cancer: epithelial-mesenchymal transition, mesenchymal-epithelial transition, stem cells and beta-catenin. Cells Tissues Organs *179*, 56-65.

Cantor, D.I., Cheruku, H.R., Nice, E.C., and Baker, M.S. (2015). Integrin alphavbeta6 sets the stage for colorectal cancer metastasis. Cancer Metastasis Rev *34*, 715-734.

Duffy, M.J., McGowan, P.M., and Gallagher, W.M. (2008). Cancer invasion and metastasis: changing views. J Pathol *214*, 283-293.

Eccles, S.A., and Welch, D.R. (2007). Metastasis: recent discoveries and novel treatment strategies. Lancet *369*, 1742-1757.

Eden, G., Archinti, M., Furlan, F., Murphy, R., and Degryse, B. (2011). The urokinase receptor interactome. Curr Pharm Des *17*, 1874-1889.

Feist, P., and Hummon, A.B. (2015). Proteomic Challenges: Sample Preparation Techniques for Microgram-Quantity Protein Analysis from Biological Samples. International Journal of Molecular Sciences *16*, 3537-3563.

Fonseka, S.

Fusaro, V.A., Mani, D.R., Mesirov, J.P., and Carr, S.A. (2009). Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nature biotechnology *27*, 190-198.

Haggar, F.A., and Boushey, R.P. (2009). Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. Clin Colon Rectal Surg *22*, 191-197.

Hamidi, H., and Ivaska, J. (2018). Every step of the way: integrins in cancer progression and metastasis. Nat Rev Cancer *18*, 533-548.

Hoofnagle, A.N., and Wener, M.H. (2009). The Fundamental Flaws of Immunoassays and Potential Solutions Using Tandem Mass Spectrometry. Journal of immunological methods *347*, 3-11.

Jaffe, J.D., Keshishian, H., Chang, B., Addona, T.A., Gillette, M.A., and Carr, S.A. (2008). Accurate Inclusion Mass Screening: A Bridge from Unbiased Discovery to Targeted Assay Development for Biomarker Verification. Molecular & Cellular Proteomics : MCP *7*, 1952-1962.

Kalluri, R., and Weinberg, R.A. (2009). The basics of epithelial-mesenchymal transition. J Clin Invest *119*, 1420-1428.

Keshishian, H., Addona, T., Burgess, M., Kuhn, E., and Carr, S.A. (2007). Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. Mol Cell Proteomics *6*, 2212-2229.

Kusebauch, U., Deutsch, E.W., Campbell, D.S., Sun, Z., Farrah, T., and Moritz, R.L. (2014). Using PeptideAtlas, SRMAtlas and PASSEL – Comprehensive Resources for discovery and targeted proteomics. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *46*, 13.25.11-13.25.28.

Lamouille, S., Xu, J., and Derynck, R. (2014). Molecular mechanisms of epithelial-mesenchymal transition. Nat Rev Mol Cell Biol *15*, 178-196.

Lester, R.D., Jo, M., Montel, V., Takimoto, S., and Gonias, S.L. (2007). uPAR induces epithelial-mesenchymal transition in hypoxic breast cancer cells. J Cell Biol *178*, 425-436.

Liebler, D.C., and Zimmerman, L.J. (2013). Targeted Quantitation of Proteins by Mass Spectrometry. Biochemistry *52*, 3797-3806.

Lin, D., Alborn, W.E., Slebos, R.J., and Liebler, D.C. (2013). Comparison of protein immunoprecipitation-multiple reaction monitoring with ELISA for assay of biomarker candidates in plasma. J Proteome Res *12*, 5996-6003.

Liu, X., Jin, Z., O'Brien, R., Bathon, J., Dietz, H.C., Grote, E., and Van Eyk, J.E. (2013). Constrained Selected Reaction Monitoring: Quantification of selected post-translational modifications and protein isoforms. Methods (San Diego, Calif) *61*, 304-312.

Liu, X., Qiu, F., Liu, Z., Lan, Y., Wang, K., Zhou, P.-K., Wang, Y., and Hua, Z.-C. (2014). Urokinase-type plasminogen activator receptor regulates apoptotic sensitivity of colon cancer HCT116 cell line to TRAIL via JNK-p53 pathway. Apoptosis *19*, 1532-1544.

MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern, R., Tabb, D.L., Liebler, D.C., and MacCoss, M.J. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics *26*, 966-968.

Mahboob, S., Ahn, S.B., Cheruku, H.R., Cantor, D., Rennel, E., Fredriksson, S., Edfeldt, G., Breen, E.J., Khan, A., Mohamedali, A., et al. (2015). A novel multiplexed immunoassay identifies CEA, IL-8 and prolactin as prospective markers for Dukes' stages A-D colorectal cancers. Clinical proteomics 12, 10.

Mead, J.A., Bianco, L., Ottone, V., Barton, C., Kay, R.G., Lilley, K.S., Bond, N.J., and Bessant, C. (2009). MRMaid, the web-based tool for designing multiple reaction monitoring (MRM) transitions. Mol Cell Proteomics *8*, 696-705.

Nahnsen, S., and Kohlbacher, O. (2012). In silico design of targeted SRM-based experiments. BMC Bioinformatics *13*, S8.

Niu, J., Dorahy, D.J., Gu, X., Scott, R.J., Draganic, B., Ahmed, N., and Agrez, M.V. (2002). Integrin expression in colon cancer cells is regulated by the cytoplasmic domain of the β6 integrin subunit. International Journal of Cancer *99*, 529-537.

Ozcan, Sureyya, et al. "Towards reproducible MRM based biomarker discovery using dried blood spots." Scientific reports 7 (2017): 45178.

Picotti, P., Rinner, O., Stallmach, R., Dautel, F., Farrah, T., Domon, B., Wenschuh, H., and Aebersold, R. (2010). High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. Nat Methods *7*, 43-46.

Ploug, M., Ronne, E., Behrendt, N., Jensen, A.L., Blasi, F., and Dano, K. (1991). Cellular receptor for urokinase plasminogen activator. Carboxyl-terminal processing and membrane anchoring by glycosyl-phosphatidylinositol. J Biol Chem *266*, 1926-1933.

Pyke, C., Ralfkiaer, E., Ronne, E., Hoyer-Hansen, G., Kirkeby, L., and Dano, K. (1994). Immunohistochemical detection of the receptor for urokinase plasminogen activator in human colon cancer. Histopathology *24*, 131-138.

Saldanha, R.G., Molloy, M.P., Bdeir, K., Cines, D.B., Song, X., Uitto, P.M., Weinreb, P.H., Violette, S.M., and Baker, M.S. (2007). Proteomic identification of lynchpin urokinase plasminogen activator receptor protein interactions associated with epithelial cancer malignancy. J Proteome Res *6*, 1016-1028.

Seyfried, T.N., and Huysentruyt, L.C. (2013). On the origin of cancer metastasis. Crit Rev Oncog *18*, 43-73.

Siegel, R., Desantis, C., and Jemal, A. (2014). Colorectal cancer statistics, 2014. CA Cancer J Clin *64*, 104-117.

Shrivastava, Alankar, and Vipin B. Gupta. "Methods for the determination of limit of detection and limit of quantitation of the analytical methods." Chronicles of young scientists 2.1 (2011): 21.

Smith, H.W., and Marshall, C.J. (2010). Regulation of cell signalling by uPAR. Nat Rev Mol Cell Biol *11*, 23-36.

Sowmya, G., Khan, J.M., Anand, S., Ahn, S.B., Baker, M.S., and Ranganathan, S. (2014). A site for direct integrin alphavbeta6.uPAR interaction from structural modelling and docking. J Struct Biol *185*, 327-335.

Steeg, P.S. (2006). Tumour metastasis: mechanistic insights and clinical challenges. Nat Med *12*, 895-904.

Tang, K., Page, J.S., and Smith, R.D. (2004). Charge Competition and the Linear Dynamic Range of Detection in Electrospray Ionization Mass Spectrometry. Journal of the American Society for Mass Spectrometry *15*, 1416-1423.

Tang, L., and Kebarle, P. (1993). Dependence of ion intensity in electrospray mass spectrometry on the concentration of the analytes in the electrosprayed solution. Analytical Chemistry *65*, 3654-3668.

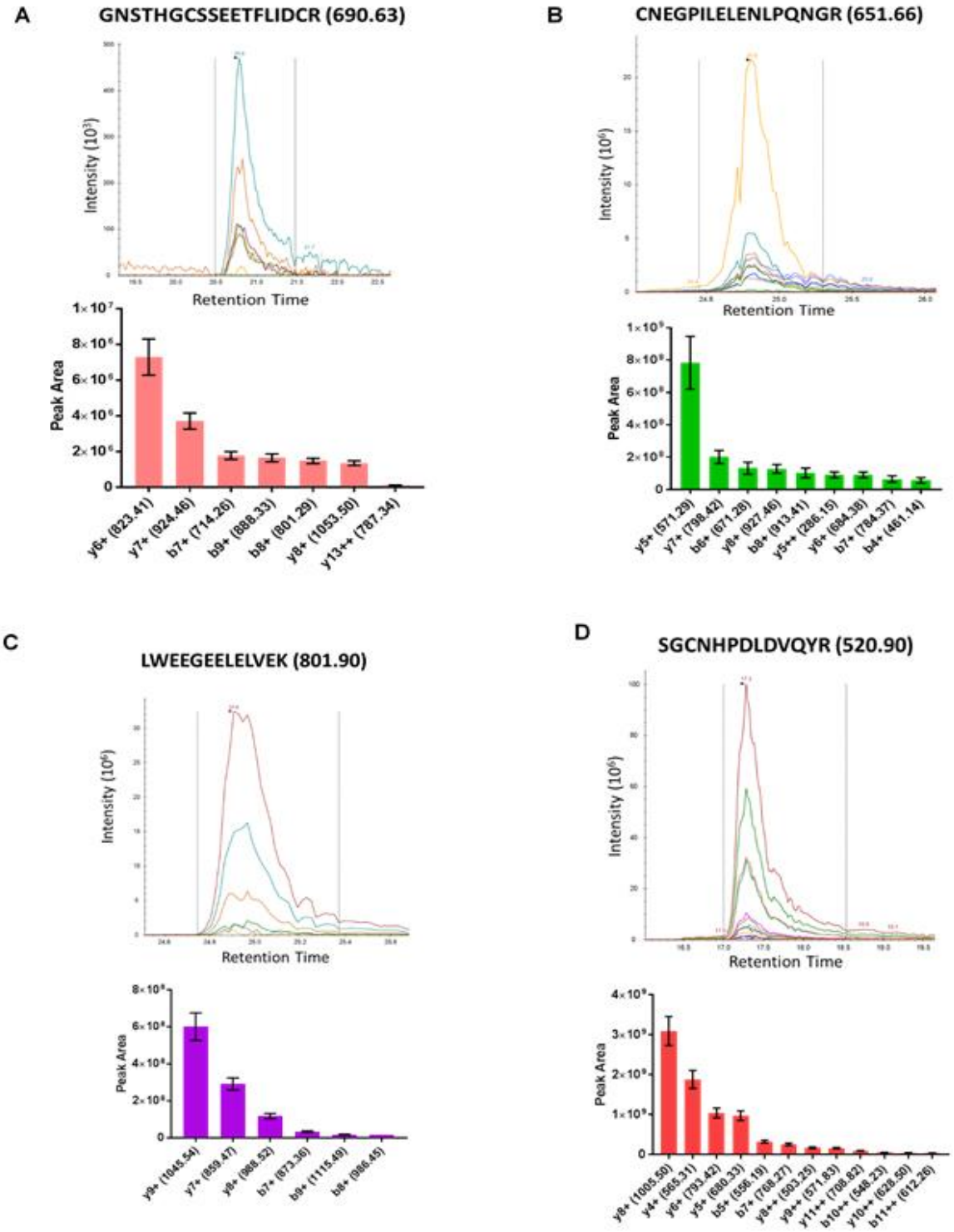
Uchida, Y., Tachikawa, M., Obuchi, W., Hoshi, Y., Tomioka, Y., Ohtsuki, S., and Terasaki, T. (2013). A study protocol for quantitative targeted absolute proteomics (QTAP) by LC-MS/MS: application for inter-strain differences in protein expression levels of transporters, receptors, claudin-5, and marker proteins at the blood–brain barrier in ddY, FVB, and C57BL/6J mice. Fluids and Barriers of the CNS *10*, 21-21.

Uszynski, M., Perlik, M., Uszynski, W., and Zekanowska, E. (2004). Urokinase plasminogen activator (uPA) and its receptor (uPAR) in gestational tissues; Measurements and clinical implications. Eur J Obstet Gynecol Reprod Biol *114*, 54-58.

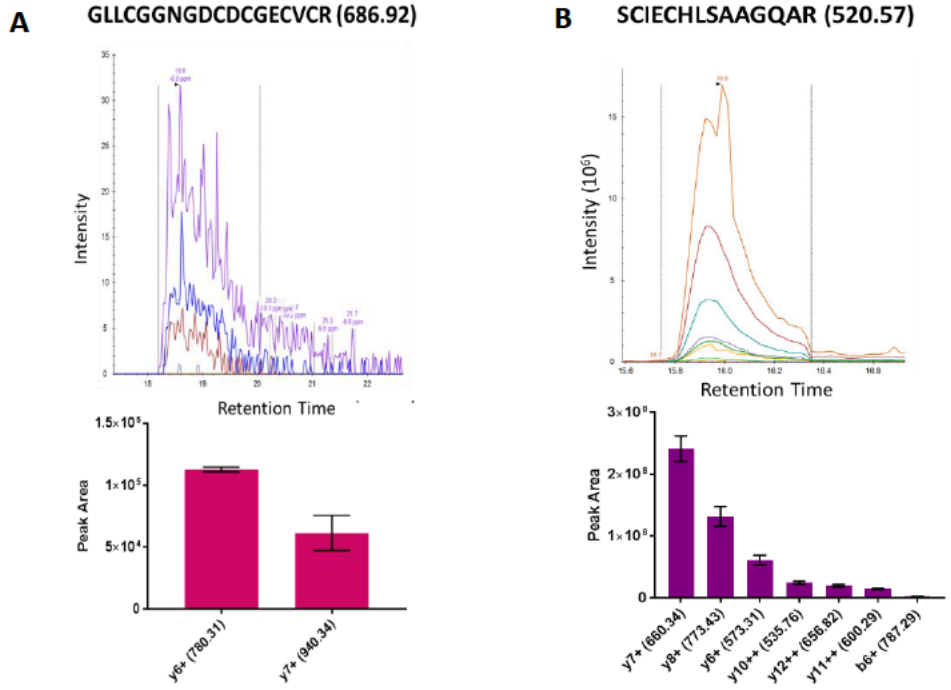
Zhang, G., Kim, H., Cai, X., Lopez-Guisa, J.M., Carmeliet, P., and Eddy, A.A. (2003). Urokinase receptor modulates cellular and angiogenic responses in obstructive nephropathy. J Am Soc Nephrol *14*, 1234-1253.

Supplementary Information

Supp Figure S5.1: Representation of uPAR peptide peak intensity and detected transitions: (A-D) The peak area of all reproducibly detected daughter ions from each peptide was compared in triplicates. Mean and SEM is shown in histograms below. The experiment for the data in this figure was performed by Ms Sachini Fonseka and is presented in her M.Res. thesis. This figure has been reproduced and adapted from (Fonseka) after seeking due permissions.

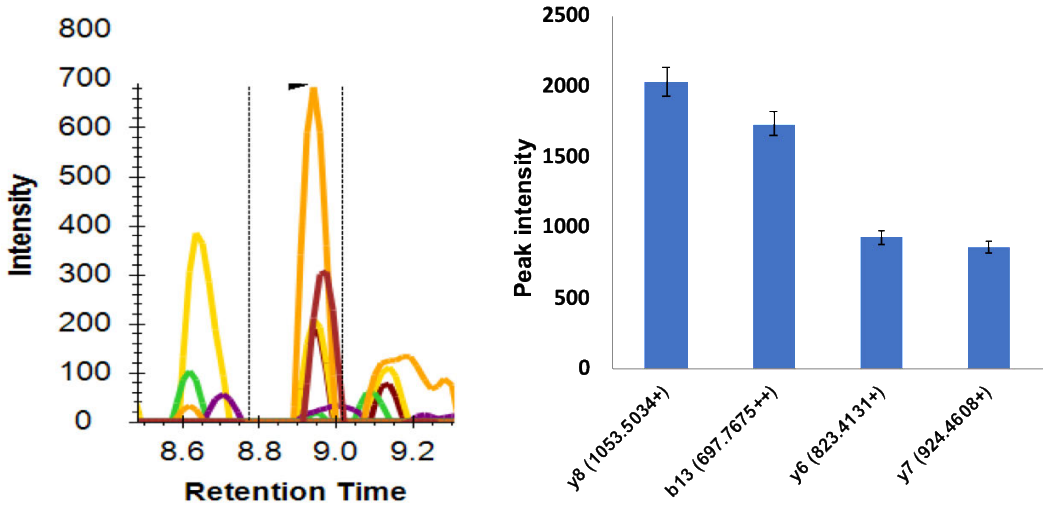


Supp Figure S5.2: αvβ6 peptide peak intensity and detected transitions: (A-B) The peak area of all reproducibly detected daughter ions from each peptide was compared in triplicates. Mean and SEM is shown in histograms. The experiment for the data in this figure was performed by Ms Sachini Fonseka and is presented in her M.Res thesis. This figure has been reproduced and adapted from (Fonseka) after seeking due permissions.

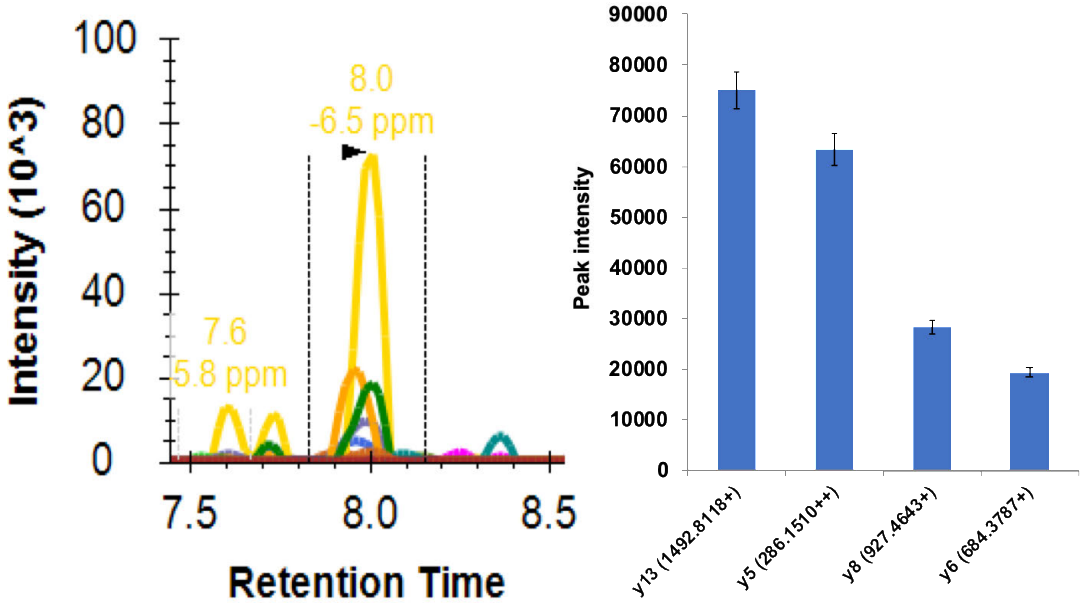


Supp Figure S5.3: Validation of reproducible uPAR peptide transitions from recombinant proteins and between freeze/thaw cycles. The peak area of all reproducibly detected daughter ions from each peptide was compared in triplicates (A-D). Mean and SEM is shown in histograms.

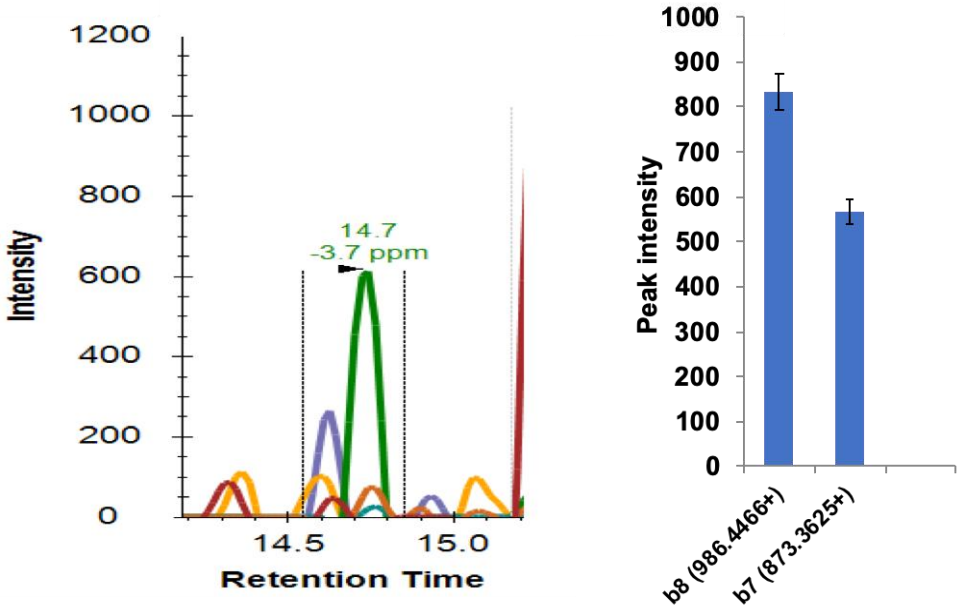
A) Peptide: GNSTHGCSSEETFLIDCR; m/z = 690.63, charge = 3+



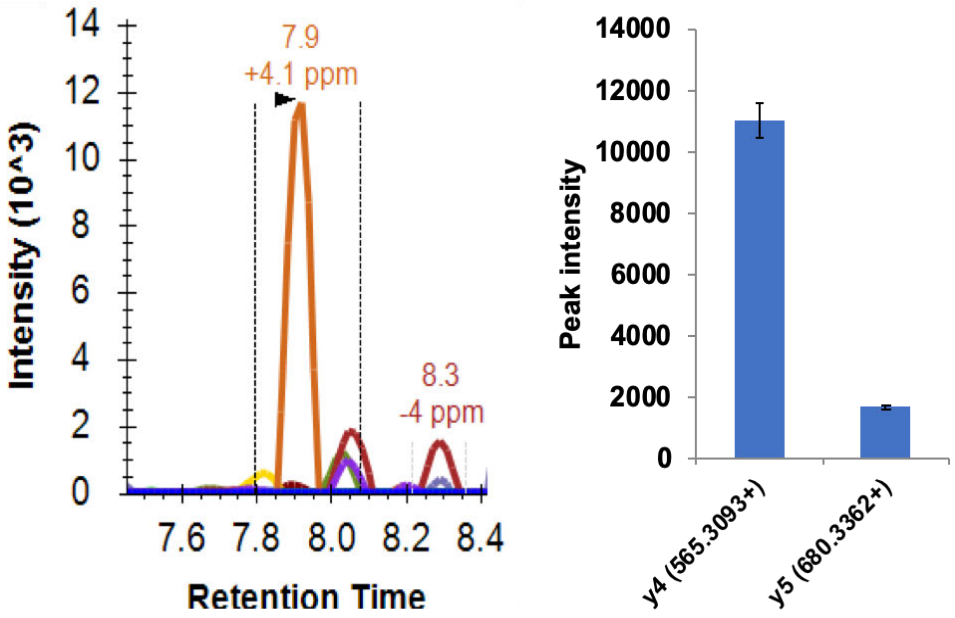
B) Peptide: CNEGPILELENLPQNGR; m/z = 651.66, charge = 3+



C) Peptide: LWEEGEELELVEK; m/z = 801.9, charge = 2+



D) Peptide: SGCNHPDLDVQYR; m/z = 520.9002, charge = 3+



Supplementary Tables

Supp Table S5.1: Peptide mass inclusion list for identifying reproducible transitions

Table S5.1. Peptide mass inclusion list for PRM MS analysis		
uPAR		
Peptide sequence	**m/z**	**Charge**
GNSTHGCSSEETFLIDCR	690.63	3
GPMNQCLVATGTHEPK	870.41	2
CNEGPILELENLPQNGR	651.66	3
YLECISCGSSDMSCER	651.92	3
YLECISCGSSDMSCER	977.38	2
SGCNHPDLDVQYR	520.9	3
LWGGTLLWT	523.79	2
NQSYMVR	449.22	2
LWEEGEELELVEK	801.9	2
SGAAPQPGPAHLSLTITLLMTAR	768.42	3
ITSLTEVVCGLDLCNQGNSGR	1147.05	2
SPEEQCLDVVTH	707.32	2
LGDAFSMNHIDVSCCTK	652.29	3
αvβ6		
Peptide sequence	**m/z**	**Charge**
HILPLTNDAER	426.9	3
LGFGSFVEKPVSPFVK	579.99	3
NSSDIVQIAPQSLILK	576	3
NSSDIVQIAPQSLILK	863.98	2
GCQLNFIENPVSQVEILK	696.7	3
VGDTASFSVTVNIPHCER	663.66	3
NEYSMSTVLEYPTIGQLIDK	767.72	3
SCIECHLSAAGQAR	520.57	3
WQTGTNPLYR	618.31	2
GLLCGGNGDCDCGECVCR	686.92	3

Table S5.2. a) **uPAR** transitions detected in recombinant protein sample triplicates						
Peptide sequence	***m/z***	**Charge**	**# daughter ions obs.**	**Maximum Peak Intensity**	**Top rank ion**	**Top ranking ion significance**
GNSTHGCSSEETFLIDCR	690.63	+3	7	$7.3x10^6$	y6+	*p*=0.03*
CNEGPILELENLPQNGR	651.66	+3	9	$7.8x10^8$	y5+	*p*=0.03*
SGCNHPDLDVQYR	520.9	+3	12	$3.1x10^9$	y8+	*p*=0.04*
LWEEGEELELVEK	801.9	+2	6	$6.0x10^8$	Y9+	*p*=0.02*
b) αvβ6 transitions detected in recombinant protein sample triplicates						
SCIECHLSAAGQAR	520.57	+3	7	$3.0x10^6$	y7+	*p*=0.03*
GLLCGGNGDCDCGECVCR	686.92	+3	2	$1.0x10^4$	y6+	*p*=0.03*
The data shown in this table is derived from experiments conducted by Ms Sachini Fonseka and Ms Samridhi Sharma. The table is reproduced and modified to fit in the manuscript (Fonseka)						

Chapter 6

General Discussion, Conclusion and Future directions

6.1 General discussion

Broadly, this thesis is an attempt to address and add novel additions to the two branches of HPP a) C-HPP and b) B/D-HPP. Chapter 1 of this thesis focuses on the analysis undertaken to demonstrate that the rate of progress of the HPP in finding PE1 proteins needs to be accelerated in order to meet proposed HPP timeline/schedules. However, the rate of identification of missing proteins seems to have slowed down in past few years as discussed in publication I. The goal of which is to ultimately hasten the progress of identification of whole human proteome and identify all the missing proteins. To do so, there are few avenues that can be explored:

- Capture and account for all credible scientific data for PE2-4 missing proteins including genomic and antibody-based evidences.
- Understanding the characteristics of the proteins in PE2-4 categories to elucidate the reasons for them falling through the cracks from current data analysis via mass spectrometry.

Another strategy to expedite the above-mentioned point is to group missing proteins with respect to their descriptions (as mentioned in neXtProt). Currently, neXtProt is only available source that exclusively classifies and provides missing proteins information. Hence, neXtProt descriptors can be used to group proteins together to study the differences in the composition and sequences of proteins. This exercise will allow understanding of why certain proteins cannot be identified using mass spectrometry due to placement of arginine and lysine in transmembrane domains for an example. This is covered in Introduction chapter, publication II of this thesis. Similar study can be performed on taste receptors and other G-protein coupled receptor due to a common transmembrane domain in all these proteins.

The above-mentioned strategies, and complementary MissingProteinPedia as information engine can improve the rate of progress of the HPP and assembling information in one single platform towards completion of C-HPP.

The second branch of HPP the biology and disease driven HPP was strategised to associate the dysregulation of human proteins with biology and disease condition of a human, which in our

case was colorectal cancer (CRC). This particular aspect was extensively explored in chapter 3, 4 and 5 of this thesis. Chapter 3 focused on identification of early stage markers for CRC to improve patient survival using proteomics-based technologies. The past years has seen technological progress of MS-based proteomics in aspects of plasma-based biomarker discovery and the application of these was well exploited in initial biomarker discovery. Specifically, design a workflow that involves intense sample preparation and multi-stage fractionation to quantify the plasma proteome via SWATH™-MS. This study was conducted on pooled plasma samples as a strategy adopted to identify differentially expressed proteins candidates to be verified later by orthogonal technology and targeted proteomics. But, if the workflow used in this thesis is to be deployed in the individual plasma samples, there needs to develop a fast and automated workflow for initial biomarker discovery. This need for an automated system to identify maximum number of plasma proteins, is a goal yet to be achieved in MS technology in future. The large initial number of plasma proteome profiles from hundreds of individual samples has maximum likelihood of revealing a statistically significant pattern that might differentiate healthy controls from early stage CRC. However, the aim is to develop a robust method that can find in-depth of plasma proteome without the depletion of plasma sample. This objective could require a lot of technological advances not only in proteomics-based techniques but also more sensitive and high throughput mass spectrometers. This aim is yet to be accomplished in terms of technology.

Colorectal cancer (CRC), a malignant neoplasm of colon, rectum and appendix, remains to be the third most common cause of cancer-related morbidity and mortality (IARC Global Cancer Observatory web site (http://gco.iarc.fr/)). The overall patient survival is directly related to the time of diagnosis of CRC. It is estimated that more than one -third of patients are often diagnosed when CRC has already metastasised to distant organs. At this stage, CRC is rarely curative via surgery and often leads to poor patient survival rate. On contrary, if CRC is detected at an early stage, it is curable after surgical resection and dramatically improves patient survival. Several screening modalities have been implemented for early CRC screening which includes stool-based tests like gFOBT and FIT, later to be confirmed by gold standard colonoscopy, if positive (Wolf et al., 2018). Despite of multiple annual stool-screening programs in place to detect CRC early, the patient compliance for stool-based tests is approximately 40% (Wolf et al., 2018). In addition, the reliance of these tests on blood haemoglobin in stool samples renders false-positive results and exposes them to unnecessary

invasive colonoscopic procedure. The discovery of reliable biomarkers with high sensitivity and sensitivity for early stage diagnosis is a logical choice.

The primary aim of this thesis was to identify and evaluate novel biomarkers with diagnostic ability that can differentiate early CRC stages from healthy controls. While achieving the primary goal, this thesis focuses on the challenges of plasma as an ideal, though a challenging source of biomarkers. To overcome this challenge of identifying similar repertoire of plasma proteins, the combinatory power of the immunoaffinity depletion columns (MARS14 and in-house IgY depletion columns) was utilised for in-depth analysis plasma proteome. Using these techniques, high and mid abundant plasma proteins that mask the low abundant proteins were depleted. Further, to accurately measure and reproducibly quantify plasma proteins, the current state-of-art proteomics technology SWATH™-MS was used. The plasma library generation is a pre-requisite for a SWATH™-MS experiment. This was achieved using strength of multiple fractionation methods (SCX, SAX, SEC and HpH). The resulting library identified proteins in concentration range of 50 ug/ul-2ng/ml, which was represented on a protein concentration curve, or "Anderson Curve". The final biomarker discovery experiments identified 37 potential candidates that could distinguish CRC early stages from healthy controls. These proteins were observed across the entire concentration range of the library Anderson curve showing the broad functional coverage of plasma protein across the concentration range. Of these 37 proteins, 10 proteins were found to be liver-derived proteins (APOA2, APOC3, F2, APOC2, SERPIN6, PON1, AMBP, SAA1, SAA2, and HGFAC), and *in toto*, all 37 proteins had subcellular attributes associated with the cytosol (APOB, SAA1, HGFAC, S100A8, PFN1, APOA2, F2), exosomes (VASN, COMP), secretory proteins (COMP, ADEC1, SODE, HGFAC, C1QC, ITIH3, CFAD, MASP2, SAA1, SAA2, GPX3, SAMP, AMBP, PON1), or had been shown to be an integral component of cell membranes (VASN). Three candidates were expressed in somatic tissue (MECP2), endothelial cells (ROBO4) or were known to be secreted in response to dendritic cell activation and maturation (ADAMDEC1). Many of observed proteins were identified as CRC biomarkers in more than one proteomics study encouraging the confidence in the data.

Further, the biomarker discovery study done in this thesis was performed on pooled patient data. Though, the biomarkers must be measured in individual samples in future studies, for discovery/identification of potential candidates, pooling is advantageous. It not only reduces the sample numbers but also reduces the sample volume required. More so, a pooling strategy

reduces the cost of depletion and fractionations which are must for sample preparation in plasma-based proteomics as discussed in Introduction section 2.8. Further, plasma ultradepletion with SWATH™-MS is extremely time consuming as was demonstrated by our 3-year preliminary study. Our study used deep ultradepletion discovery on pooled CRC staged plasmas followed by scalable high-throughput technologies. The discovered candidates are amenable to quantitative multiplexing via (MRM/ELISA) for further validation and absolute quantitation of novel candidate biomarkers.

Mass spectrometry is an analytical technique and results from it leads to generation of large datasets. It is always accompanied with bioinformatic approaches to perform data analysis, for fold-change determination, normalisation, gene-ontology classification, to map the pathways or to determine the protein and protein interactors. Realising the need of machine learning approaches in handling large proteomics data sets (Swan. et al., 2013). Machine learning has now taken a big leap into determining the predictive candidates using predictive neural networks (Zhang, Fan, et al.,2013). The predictive neural networks emulate the brain functioning. Just like the connections of a neuron and the network is activated-based on the signal received. Each neuron is a variant of linear classifiers. Multiple neurons can be included in the different neurons and layers resulting in formation of complex network laying a foundation for non-linear classifiers. This application has now bled into proteomics and was employed in this thesis (chapter 3) to identify protein candidates from large proteomics data sets. In a proof-of-concept experiment, a classification algorithm was trained to identify potential candidates from real patients and synthetic patient populations was generated to train panel of 5 protein candidates to discriminate early stage CRC from healthy. The profile of 37 candidates (including these 5) managed to yield perfect cancer stage classification, not only for the presence of cancer, but also in differentiating the actual clinical stage (I-IV) of disease. Therefore, our study proposes this 5-protein marker panel of candidates as highly interesting for potential predictive purposes, and now propose to replace these generated samples with biological ones as a larger patient population dataset.

Further, the few of potential candidates of 37 identified candidates were validated using orthogonal immune assays confirming and backing up the mass spectrometry data (from chapter 3). The results obtained in this study encourage further evaluation of these biomarkers on individual datasets.

The fourth chapter of this thesis focused on employing the strength of targeted proteomics to accurately measure identified potential candidates from SWATH™-MS biomarker discovery study in individual patient samples. The experiments are still in preliminary stages with the first-pass assay for CFD in place, which relatively quantifies and confirms the SWATH™-MS data encouraging to do furthermore validations on individual population cohorts.

6.2 Applications and Limitations

Circulating plasma is one of the most useful biofluids for the study of human pathology and biomarker discovery. This is because, plasma contains a repertoire of proteins that are dynamically modulated based on the pathology of an individual (Geyer et al., 2018). In the case of diseases like cancer, plasma often contains proteins derived from tissue and tumour leakage that may otherwise not be present in plasma. Using proteomics, CRC staged (I-IV) patients were compared to "healthy" controls to identify CRC-specific plasma protein markers. One of the limitations of this study was the use of pooled plasma samples for the biomarker discovery phase. A total of 20 patients for each stage of the four CRC stages (I-IV) and 20 healthy controls were employed in this study. To identify unique CRC stage-specific plasma biomarkers, the 20 plasma samples from each stage were pooled prior to SWATH™-MS analysis. While pooling specimens is not the best practice for developing biomarkers (Geyer et al., 2017), the goal was to verify the SWATH™-MS data using individual specimens. This two-step approach relied on the fact that SWATH™-MS is a powerful and sensitive protein detection method (Doerr, 2014). The verification experiments performed on 7 proteins to date on pooled samples, confirm the SWATH™-MS quantitation. However, many proteins remain to be validated, and additional work is required in order to determine the value and accuracy of performing discovery proteomics in pooled rather than individual patient samples. Another critical limitation of using this approach is that proteomic data could not be analysed based on patients' clinical outcomes such as recurrence-free or overall survival due to pooling of plasma samples.

Another important consideration and potential limitation of SWATH™-MS is that it involves an intensive workflow, which initially involves developing a robust peptide library followed by subjecting samples to mass spectrometry analysis. Nevertheless, SWATH™-MS was selected due to its sensitivity and the increased likelihood of detecting low abundance circulating CRC biomarkers. In DDA experiments, where the most abundant precursor ions are

selected for MS/MS analyses, low abundant proteins may be missed. Although beyond the scope of this thesis, it would have been interesting to compare the data derived from DDA and SWATH™-MS experiments.

Another aspect which could be worth exploring is tissue-based proteomics. It offers the most accurate insight of bioprocesses and pathogenic pathways contributing to disease progression. These candidates are, therefore, also the most sought-after targets for therapeutics. The biomarkers obtained from such an analysis are also most easily translated into clinical assays like immunohistochemistry assays, following biopsy. However, from a clinical perspective, translational utility of tissue biomarkers necessitates highly invasive procedures (Kalinina et. al., 2011). Therefore, blood-based biomarkers were the most attractive choice for CRC screening. Blood is a rich source of tissue leakage proteins, immunoglobulins and exosomes. It would be interesting to map the proteomic profile of CRC plasma against matched CRC tissues, across stages to correlate trends. This exercise could potentially yield a subset of markers released in blood from tumour tissues representative of CRC stage progression. The comparison of circulating and tissue derived markers was beyond the scope of this thesis.
Proteomic technologies have been extensively used in the identification of disease biomarkers in the circulation and yet most of these putative biomarkers have not been independently validated and never reach the clinic. A major reason for this lack of biomarker confirmation is the lack of a consistent, structured pipeline to validate putative candidates and evaluate their diagnostic utility. This is not difficult to achieve but does require meticulous characterisation of the discriminating power each protein, and careful consideration when generating multi-protein panels. These validation steps are discussed in Chapter 6, Future Directions.

Machine-learning approaches are becoming more mainstream for proteomics studies. Particularly use of artificial neural network that emulates the human brain network has been used in breast cancer early detection imaging studies (Zhang, Fan, et al. 2013). This technology was used to determine a panel of candidates that distinguish early stages of CRC from healthy on a synthetic data set. However, machine learning techniques often suffer from data overfitting which is inherent to all machine learning techniques, and it is hard to completely disprove. This could happen especially while using technical triplicate data, which usually has a low standard error. Another cause for data overfitting could be model learning the details and noise of the data set and using it as a concept which could negatively impact the ability of the model to

generalise the results. There are some avenues adopted in our study that ensures that overfitting is minimised:

i). In training algorithms, we trained several biomarkers with a different response profile independently. The results suggested that no single marker cluster in hyperspace, whereas a panel of biomarkers shows a well-defined clustering.

ii). For supervised learning, the training dataset algorithm was kept separate from the testing data. In our case, we trained our model on a very noisy synthetic patient dataset with 10-fold standard deviation. 10-fold standard deviation was derived from the standard deviation values obtained from ADAMDEC1 concentration observed in individual patient data (n=100). The trained model was then tested on the real data from mass spectrometry experiments and hence the results observed prove that the algorithm generalises to the new dataset.

iii). Another avenue that was used in our study is that, a small percentage of the training dataset (15% in our case) was kept for the internal validation of our algorithm to prevent over-fitting (70% training, 15% validation and 15% testing). There was no difference in any of the scores for all three components.

iv). For supervised learning, the training dataset algorithm was kept separate from the testing data. In our case, we train our model on a very noisy synthetic patient dataset with 10-fold standard deviation. 10-fold standard deviation was derived from the standard deviation values obtained from ADAMDEC1 concentration observed in individual patient data (n=100). The trained model was then tested on the real data from mass spectrometry experiments and hence the results observed prove that the algorithm generalises to the new dataset. The next step would be to get more patient data in order to further test the generalisation power of our trained model, and gradually, to replace the synthetic data for the training phase as well.

The advancements in field of plasma biomarker discovery using proteomics has been extensively exploited to identify, thousands of diagnostic, prognostic or therapeutic markers. Unfortunately, many of these falls through the cracks, failing to reach the clinic. One of the major reasons is a lack of a structured pipeline to pursue the identified candidate and evaluate its diagnostic ability in preliminary stages. This is not difficult to achieve but does require meticulous characterisation of the discriminating power of these proteins, as a panel. These steps are discussed below in the future directions.

6.3 Future directions

There are many challenges to develop a highly sensitive, accurate and clinically relevant biomarkers for disease diagnosis, prognosis and treatment prediction.

In this study, some interesting circulating plasma protein candidates for early stage CRC detection were identified. Many of these have an established biological role in tumour development and progression. In order to extend this work, and validate the diagnostic value of these markers, a series of potential additional experiments are outlined below:

1. Development of a multiplex quantitative MRM assay for prioritised potential candidates - ADAMDEC1, MARCO, MRC1, S100A8, ApoAIV, GPX3, COMP, C1QC and CFD, and the 5-protein marker panel (SAA2, APCS, APOA4, F2 and AMB) identified from predictive neural network classification.
2. Benchmark new protein panels against current CRC screening modalities. This includes establishing normal reference ranges for new biomarkers in healthy controls, comparing sensitivity and specificity of the new biomarker against existing screening modalities such as stool-based test (gFOBT/FIT).
3. Measuring uPAR and αvβ6 peptide fragments in CRC plasma samples using MRM for accurate quantification.
4. Estimating cost/benefit effectiveness of any new test in the Australian **N**ational **B**owel **C**ancer **S**creening **P**rogram (NBCSP). The long-term effectiveness of a blood based multi-variate biomarker test should be determined through a modelled economic evaluation. The patient compliance to such a test would be evaluated based on findings of a discrete choice experiment (DCE). This includes evaluating patient compliance to a new test and measuring cost effectiveness per early stage CRC case identified (i.e., adenomas and stage I/II), cost per life years gained and cost per Quality Adjusted Life Year gained (Lansdorp-Vogelaar et al., 2009; Siebert et al., 2012).

Recent Technical Advances in Biomarker Discovery

The last two decades have seen advancements in proteomics-based workflows for biomarker discovery; and mass spectrometry has been at the epicenter of proteomics and biomarker research (Aebersold and Mann, 2003). The technical advances in mass spectrometry in terms of method developments have been instrumental in providing insights to understand the pathophysiology of several diseases. These tools, thereby, behold the potential to unveil protein markers that function as lynchpins in CRC progression. SWATH™-MS is an emerging mass

spectrometry approach that allows comprehensive proteome profiling (Röst et al., 2014). The use of this SWATH™-MS for early detection of CRC has helped to identify several novel putative CRC biomarkers along with four previously known CRC markers (Chapter 3). A major goal ahead is to ascertain their diagnostic value in clinics and understand their role in CRC pathobiology.

Integrative research has been the hallmark of life-science research in the past decade. This is especially true for the multi-omics research. A landmark endeavour projecting the advancement of proteomics is personalised integrative personal omics profiling (iPOP). iPOP is a breakthrough in improved disease diagnosis, risk assessment and monitoring response to treatment. It involves examination of patient blood/plasma/serum samples while accounting for environment (including diet, exercise, etc.), medical history, and clinical data and performs the omics analysis (including genetic variations, proteomics, RNA and DNA information), matching data to the chromosomes (Li-Pook-Than et al., 2013). The data is integrated at multiple time points. It is then compared over the period and further integrated with DNA variants and pharmacogenome to assess disease risk. The integrative profiling system is underway to determine personalised health care measures in addition to providing a better understanding of biochemical mechanisms in disease manifestations. This approach could particularly bridge the missing link between DNA sequence mutation, their variants and their protein products. This approach also circumvents the challenges posed by tumour heterogeneity in accurate diagnosis and treatment of malignancies. As an extension of the work presented in this thesis, the iPOP model could be implemented to study individual patient cohorts and profile CRC samples pre and post tumour resection. This approach may also bring to light any predominant mutation in a given population that might be associated with these malignancies (Bedard et al.,2013).

One part of this thesis focused largely on in-depth measurement of the entire plasma proteome for plasma library generation. A new alternative and more comprehensive approach to identify novel peptides and CRC associated proteins is via proteogenomics. In a proteogenomic approach, genomic and transciptomics data is used to generate customised proteins sequences to interpret proteomics data (Nesvizhskii and Alexey, 2014). Recently, a proteogenomic study of human colon cancer performed on 110 CRC patients produced a catalog of colon cancer associated proteins and phosphosites including known and putative new markers and drug targets (Vasaikar et al., 2019). It contained genomically inferred targets including copy number

alterations, driver mutations and derived neoantigens but also yielded novel findings. It will be interesting to compare the individual patient data with the catalogue presented in this study. The re-analysis of SWATH™-MS proteomics can also help in integration of new proteomics data with existing genomics data. This effort can demonstrate the ability of proteogenomic to reveal new insights into the colon cancer biology. The success and potential of proteogenomic approach inspired national cancer institute initiated International Cancer Proteogenomic Consortium (ICPC) in year 2016 (Rodriguez et. al., 2018). This consortium also unites the experts across the cancer researchers towards a common goal of eradicating cancer.

The proteomics community has also come together to form a collection of common guidelines and an assay repository database that jointly form a public repository called the Clinical Proteomic Tumour Analysis Consortium (CPTAC) Assay Portal (http://assays.cancer.gov/) as described extensively in Chapter 4. The goal of the CPTAC portal is to disseminate assays to the scientific community, at large. CPTAC provides guidelines that pertain to standard operating procedure, protocols and assay characterisation data associated with targeted mass spectrometry-based assays. An effort that would significantly increase the clinical significance of the work in this thesis, would be, to develop an assay that aligns with these guidelines. A foundation is set towards this endeavour using PRM-based workflows as defined in Chapter 4.

The aforementioned advances are currently representative of the constantly evolving field of proteomics that can identify and measure novel peptides accurately. In the context to CRC diagnosis, these technologies could be instrumental in filling the void of standard iterative methods that can build a connection between discovery, identification and validation of relevant CRC early stage detection markers.

6.4 Clinical significance of the multi-variate protein biomarker assays.

The 5-year survival of CRC is ~85% only if diagnosed early (i.e., AJCC stages I/II), when tumour is localised within the mucosa or submucosa (Figure 2.6). Once the tumour infiltrates the lymph nodes, the survival rates drop to ~62% and is reduces to ~10% by the time liver metastasis occurs (Figure 2.6). This study identified circulating proteins, many shown previously to be associated with metastatic CRC, that differentiated early stage CRC patients from healthy controls. Although, validation of these putative biomarkers in fresh and individual cohorts is necessary, there is substantial value in identifying sensitive and selective blood-based

early diagnostic markers. These liquid biopsies can decrease unnecessary procedures, diminish associated co-morbidities and prevent mortalities as discussed in section 2.4.2.2, as well as lowering the economic burden of a preventable disease. The clinical significance of a multi-variate blood-based biomarkers of early stage CRC would be extremely high, as diagnosis by blood-based testing has patient compliance of more than 90% in comparison to existing faecal-based tests, with compliance rates of <40% (Wolf et al., 2018).

References

Aebersold, Ruedi, and Matthias Mann. "Mass spectrometry-based proteomics." Nature 422.6928 (2003): 198.

Bedard, Philippe L., et al. "Tumour heterogeneity in the clinic." Nature 501.7467 (2013): 355.

Doerr, Allison. "DIA: mass spectrometry." Nature methods 12.1 (2014): 35.

Drucker, Elisabeth, and Kurt Krapfenbauer. "Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine." EPMA journal4.1 (2013): 7.

Geyer PE, Holdt LM, Teupser D, Mann M. Revisiting biomarker discovery by plasma proteomics. Mol Syst Biol. 2017;13(9):942.

Geyer, Philipp Emanuel, et al. "Plasma proteome profiling to detect and avoid sample-related biases in biomarker studies." bioRxiv (2018): 478305.

IARC Global Cancer Observatory web site (http://gco.iarc.fr/)

Kalinina, Juliya, et al. "Proteomics of gliomas: initial biomarker discovery and evolution of technology." Neuro-oncology 13.9 (2011): 926-942.

Lansdorp-Vogelaar I, van Ballegooijen M, et al. Effect of rising chemotherapy costs on the cost savings of CRC screening. J. National Cancer Institute. 2009;101(20):1412-2

Li-Pook-Than, Jennifer, and Michael Snyder. "iPOP goes the world: integrated personalised Omics profiling and the road toward improved health care." Chemistry & biology 20.5 (2013): 660-666.

Liu, Wentao, et al. "Proteomic identification of serum biomarkers for gastric cancer using multi-dimensional liquid chromatography and 2D differential gel electrophoresis." Clinica chimica acta 413.13-14 (2012): 1098-1106.

Nesvizhskii, Alexey I. "Proteogenomics: concepts, applications and computational strategies." Nature methods 11.11 (2014): 1114.

Rodriguez, Henry, and Stephen R. Pennington. "Revolutionizing precision oncology through collaborative proteogenomics and data sharing." Cell 173.3 (2018): 535-539.

Röst, H.L. et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. 32, 219–223 (2014).

Siebert U, Alagoz O, et al. State-transition Modeling: Report of ISPOR-SMDM Modeling Good Research Practices Task Force-3. J. Int. Soc. Pharmaeco & Out. Res. 2012;15(6):812-20.

Shussman, Noam, and Steven D. Wexner. "Colorectal polyps and polyposis syndromes." Gastroenterology report 2.1 (2014): 1-15.

Spratt, John S., and Lauren V. Ackerman. "Pathologic significance of polyps of the rectum and colon." Diseases of the Colon & Rectum 3.4 (1960): 330-335.

Swan, Anna Louise, et al. "Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology." *Omics: a journal of integrative biology*17.12 (2013): 595-610.

Vasaikar, Suhas, et al. "Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities." Cell177.4 (2019): 1035-1049.

Whiteaker, Jeffrey R., et al. "CPTAC Assay Portal: a repository of targeted proteomic assays." Nature methods 11.7 (2014): 703.

Wolf AMD, Fontham ETH, Church TR, Flowers CR, Guerra CE, LaMonte SJ, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. CA: a cancer journal for clinicians. 2018.

Zhang, Bing, et al. "Proteogenomic characterization of human colon and rectal cancer." Nature 513.7518 (2014): 382.

Zhang, Fan, et al. "A neural network approach to multi-biomarker panel discovery by high-throughput plasma proteomics profiling of breast cancer." BMC proceedings. Vol. 7. No. 7. BioMed Central, 2013.

Appendix

Appendix I: Approved Human Research Ethics Letter.

Appendix II: Permission/License to publish manuscript in print and electronic format for Publication 1 from *Nature Communications*.

Appendix III: Permission/License to publish manuscript in print and electronic format for Publication 2 from Journal of Proteome Research.

Appendix IV: Off-site Research program initiated ICPC Pilots International Student Training under Cancer Moonshot Project at Fred Hutch Cancer Research Centre.

Appendix V: High-resolution images of Publication 1.

Appendix VI: Age, sex TNM staging, 5-year survival and 5-year recurrence data for recruited patients and healthy controls (n=100).

Appendix I: Approved Human Research Ethics Letter

Office of the Deputy Vice-Chancellor
(Research)

Research Office
Research Hub, Building C5C East
Macquarie University
NSW 2109 Australia
T: +61 (2) 9850 4459
http://www.research.mq.edu.au/
ABN 90 952 801 237

MACQUARIE University
SYDNEY·AUSTRALIA

20 September 2017

Dear Professor Baker

Reference No: 5201700681

Title: *Early-Stage Blood-Based Proteomic Diagnosis of Colorectal Cancer*

Thank you for submitting the above application for ethical and scientific review. Your application was considered by the Macquarie University Human Research Ethics Committee (HREC (Medical Sciences)).

I am pleased to advise that ethical and scientific approval has been granted for this project to be conducted at:

- Macquarie University

This research meets the requirements set out in the *National Statement on Ethical Conduct in Human Research* (2007 – Updated May 2015) (the *National Statement*).

Standard Conditions of Approval:

1. Continuing compliance with the requirements of the *National Statement,* which is available at the following website:

http://www.nhmrc.gov.au/book/national-statement-ethical-conduct-human-research

2. This approval is valid for five (5) years, subject to the submission of annual reports. Please submit your reports on the anniversary of the approval for this protocol.

3. All adverse events, including events which might affect the continued ethical and scientific acceptability of the project, must be reported to the HREC within 72 hours.

4. Proposed changes to the protocol and associated documents must be submitted to the Committee for approval before implementation.

It is the responsibility of the Chief investigator to retain a copy of all documentation related to this project and to forward a copy of this approval letter to all personnel listed on the project.

Should you have any queries regarding your project, please contact the Ethics Secretariat on 9850 4194 or by email ethics.secretariat@mq.edu.au

The HREC (Medical Sciences) Terms of Reference and Standard Operating Procedures are available from the Research Office website at:

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics

The HREC (Medical Sciences) wishes you every success in your research.

Yours sincerely

Professor Tony Eyers
Chair, Macquarie University Human Research Ethics Committee (Medical Sciences)

This HREC is constituted and operates in accordance with the National Health and Medical Research Council's (NHMRC) *National Statement on Ethical Conduct in Human Research* (2007) and the *CPMP/ICH Note for Guidance on Good Clinical Practice*.

Details of this approval are as follows:

Approval Date: 20 September 2017

The following documentation has been reviewed and approved by the HREC (Medical Sciences):

Documents reviewed	Version no.	Date
Correspondence responding to the issues raised by the HREC (Medical Sciences)		Received 10 Sep 2017
Macquarie University Ethics Application Form	1.1*	11 Sep 2017
Project Proposal	1.1*	11 Sep 2017
MQ Participant Information and Consent Form (PICF)	1.1*	11 Sep 2017

Documents Noted	Version no.	Date
Amendment approval to obtain additional control human plasma samples – Guillemin Project (5201600401)	N/A	8-9-2017

***If the document has no version date listed one will be created for you. Please ensure the footer of these documents are updated to include this version date to ensure ongoing version control.**

Appendix II: Permission/License to publish manuscript in print and electronic format for Publication 1 from *Nature Communications*.

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Dec 20, 2018

This Agreement between Ms. Samridhi Sharma ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4493290261696
License date	Dec 20, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Cancer and Metastasis Reviews
Licensed Content Title	Integrin αvβ6 sets the stage for colorectal cancer metastasis
Licensed Content Author	D. I. Cantor, H. R. Cheruku, E. C. Nice et al
Licensed Content Date	Jan 1, 2015
Licensed Content Volume	34
Licensed Content Issue	4
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	no
Title	Early stage diagnosis of colorectal cancer
Institution name	macquarie university
Expected presentation date	Dec 2018
Portions	Figure 2
Requestor Location	Ms. Samridhi Sharma Level 1, 75 Talavera Road Macquarie Park Sydney, NSW 2109 Australia Attn: Ms. Samridhi Sharma
Billing Type	Invoice
Billing Address	Ms. Samridhi Sharma Level 1, 75 Talavera Road Macquarie Park Sydney, Australia 2109 Attn: Ms. Samridhi Sharma
Total	0.00 AUD
Terms and Conditions	

Springer Nature Terms and Conditions for RightsLink Permissions

Springer Nature Customer Service Centre GmbH (the Licensor) hereby grants you a non-exclusive, world-wide licence to reproduce the material and for the purpose and requirements specified in the attached copy of your order form, and for no other use, subject to the conditions below:

1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

   If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Where **print only** permission has been granted for a fee, separate permission must be obtained for any additional electronic re-use.

3. Permission granted **free of charge** for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

4. A licence for 'post on a website' is valid for 12 months from the licence date. This licence does not cover use of full text articles on websites.

5. Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity, NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

6. Permission granted for books and journals is granted for the lifetime of the first edition and does not apply to second and subsequent editions (except where the first edition permission was granted free of charge or for signatories to the STM Permissions Guidelines http://www.stm-assoc.org/copyright-legal-affairs/permissions/permissions-guidelines/), and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence.

7. Rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

8. The Licensor's permission must be acknowledged next to the licensed material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

9. Use of the material for incidental promotional use, minor editing privileges (this does not include cropping, adapting, omitting material or any other changes that affect the meaning, intention or moral rights of the author) and copies for the disabled are permitted under this licence.

10. Minor adaptations of single figures (changes of format, colour and style) do not require the Licensor's approval. However, the adaptation should be credited as shown in Appendix below.

Appendix — Acknowledgements:

For Journal Content:
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g.

Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

For Adaptations/Translations:
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

For Book content:
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g. Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)] [**COPYRIGHT**] (year of publication)

Other Conditions:

Version 1.1

Appendix III: Permission/License to publish manuscript in print and electronic format for Publication 2 from Journal of Proteome Research.

Copyright Clearance Center RightsLink®

Home | Create Account | Help

ACS Publications Most Trusted. Most Cited. Most Read.

Title: Ultradepletion of Human Plasma using Chicken Antibodies: A Proof of Concept Study
Author: Sock-Hwee Tan, Abidali Mohamedali, Amit Kapur, et al
Publication: Journal of Proteome Research
Publisher: American Chemical Society
Date: Jun 1, 2013

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

If credit is given to another source for the material you requested, permission must be obtained from that source.

BACK | CLOSE WINDOW

Appendix IV: Off-site research program initiated ICPC Pilots International Student Training under Cancer Moonshot Project at Fred Hutch Cancer Research Centre.

NIH NATIONAL CANCER INSTITUTE
Office of Cancer Clinical Proteomics Research (http://proteomics.cancer.gov/)
CONTACT US (/CONTACT-US) | SIGN UP FOR UPDATES (/SIGN-EMAIL-UPDATES)

Search... SEARCH

Center for Strategic Scientific Initiatives

(/) ABOUT PROTEOMICS PROGRAMS

ICPC Pilots International Student Training, Paving a Path for Tomorrow's Cancer Researchers

Wednesday, March 21, 2018

The Office of Cancer Clinical Proteomics Research at the National Cancer Institute, part of the United States National Institutes of Health, is spearheading the preparation and training of the proteogenomic research workforce on an international scale.

Samridhi Sharma in Dr. Mark Baker's laboratory at Macquarie University. Credit: Samridhi Sharma

The NCI and Macquarie University in Australia, among other research agencies and institutes, are members of the International Cancer Proteogenome Consortium (https://proteomics.cancer.gov/programs/international-cancer-proteogenome-consortium). Launched in 2016 and aligning efforts with the U.S. Cancer Moonshot Initiative, NCI and Macquarie University recently piloted an international student exchange in the field of proteogenomics—the comprehensive study of cancer patient tissue, proteins and genes to better understand cancer development, progression, diagnosis, therapeutic response, and potential novel therapeutic targets.

Memoranda of Understanding are foundational to ICPC. The MOU between NCI and Macquarie University, as well as other participating institutions, support training and the exchange of scientists with expertise in proteogenomic data. ICPC aims to improve patient outcomes by disseminating and sharing proteomics, genomics, and

improve patient outcomes by disseminating and sharing proteomics, genomics, and imaging data publicly with the international cancer research community.

Samridhi Sharma, a second-year doctoral student with Mark Baker, Ph.D. (https://www.mq.edu.au/about_us/faculties_and_departments/faculty_of_medicine_and_h at Macquarie University, leveraged her postgraduate research fund to attend the 2017 Human Proteome Organization (https://hupo.org/) World Congress in Dublin, Ireland. She then traveled to the U.S. to complete four-weeks of proteogenomic training under Amanda Paulovich, M.D., Ph.D. (https://research.fhcrc.org/paulovich/en.html), a Clinical Proteomic Tumor Analysis Consortium Investigator and Member at the Fred Hutchinson Cancer Research Center in Seattle, Washington.

"I wanted to work in proteomics because it is a very powerful tool for potentially finding diagnostic plasma markers of early stage cancer," Sharma said. "Proteomics is a strong technique that adds a high level of value to genomics, cell-based and animal model studies about particular biology or diseases, helping us filter out important protein candidates and moving forward with validation studies."

Sharma is working to identify proteins in the blood that can act as diagnostic or prognostic markers for Stage I/II colorectal cancer. But before she can take the potential markers she has found to the clinic, Sharma had to learn how to develop laboratory methods to quantify the identified plasma proteins.

"In Dr. Paulovich's lab, I learned how to do a technique called immuno-multiple reaction monitoring, or immuno-MRM, to pull down protein and peptides from plasma samples and then run [the samples] on a mass spectrometer," Sharma said. "The objective of this is to validate our current identifiers from a prior SWATH study—the newest type of mass spectrometry—so that we can measure [the proteins] in multiple clinical samples."

While studying the genome has led to new therapies, most of the current targeted cancer therapies (https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies/targeted-therapies-fact-sheet) pinpoint proteins. Dr. Paulovich and her research team have lead the development of the immuno-MRM technology platform to address the lack of reliable assays needed to measure human proteins, filling in some of the missing biology that has not been identified by genomically sequencing tumors alone.

"It made sense initially to focus on the genome. Then it became clear that a lot of patients seemed to not respond to drugs that they should respond to or they would respond initially and then develop resistance," Dr. Paulovich said. "Our thought was if we looked at proteins, which are the targets of many new therapies, perhaps we could do a better job of predicting patient responses and understand the mechanisms through which the drugs work."

According to Dr. Paulovich, the immuno-MRM technology platform is robust, precise, and can simultaneously measure multiple proteins in samples across a single run, a drastic improvement over the traditional platform. Immuno-MRM also uses an internal standard that can be implemented in the assay and used across laboratories, providing a standardized tool for quantifying human proteins with the end goal of improving patient diagnosis and treatment.

"Immuno-MRM is a technological capability that will help the translation of protein biomarkers into the clinical realm," Dr. Paulovich said.

Dr. Baker, Professor of Proteomics and Biochemistry at Macquarie's Faculty of Medicine and Health Sciences, says that the most important contribution to ICPC from Sharma's work will be the knowledge to help us better understand the enemy that we know cancer to be.

"I hope that Samridhi's work will show that we can see changes in the blood of patients affected with colorectal cancer at the very earliest stages when the cancer is no bigger than a pea," Dr. Baker said. "These protein changes in the blood may reflect that the immune system recognizes the cancer as foreign in some way."

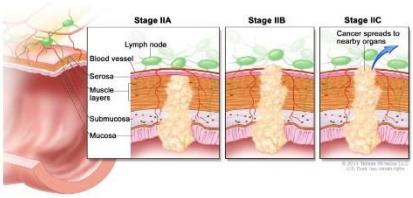
As part of the international collaboration, the proteomics data that Sharma produces through her training will be contributed to the CPTAC Data Portal (https://proteomics.cancer.gov/data-portal), which makes proteomics data publicly available for use by cancer researchers and physicians worldwide.

"We have agreed to participate in a data exchange common that aligns with the U.S. Cancer Moonshot Initiative, the NCI and CPTAC," Dr. Baker said.

The data sets contributed by Dr. Baker and his research team fulfill a call-to-action (https://obamawhitehouse.archives.gov/the-press-office/2016/07/16/fact-sheet-victoria-comprehensive-cancer-center-vice-president-biden) in cancer research that represents the global diversity of people and commonly diagnosed cancers in their unique populations.

Dr. Baker's expertise and contribution since working in the U.S. has made Macquarie University a center of proteomics and technology excellence, a hub for first-class proteomics research into colorectal cancer, and an obvious choice for collaboration.

Dr. Baker's research group focuses on two simple aims that target colorectal cancer. The first aim is to find blood-based, early stage biomarkers that when combined with a conventional treatment like surgery can be used to cure cancer. The second aim focuses on developing drugs that reverse the biology of metastasis (https://www.cancer.gov/about-cancer/understanding/what-is-cancer), a component of late stage cancer.



Stage II colorectal cancer. In Stage IIA (left), cancer spreads through the muscle layer of the colon/rectum wall to the serosa. In Stage IIB (middle), cancer spreads through the serosa but not to nearby organs. Stage IIC (right), cancer spread through serosa to nearby organs. Credit: NCI Visuals

"We will detect a lot of early cancers that can be excised and save people's lives. We want to get it before it recurs," Dr. Baker said. "If [colorectal cancer] does recur, is it possible to slow it down and make it dormant for a little bit longer so we can prolong lives? To do that, we need to know the genes and proteins involved in cancer spread and how they work."

According to the World Health Organization (http://globocan.iarc.fr/Pages/fact_sheets_population.aspx), colorectal cancer is the third most common cancer resulting in 1.36 million new cases and 694,000 deaths worldwide in 2012.

Australia has the highest incidence (http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx?cancer=colorectal) of colorectal cancer worldwide. While colon cancer can be a treatable and often curable disease, fewer than 40% of cancer cases are detected early enough according to Bowel Cancer Australia (https://www.bowelcanceraustralia.org/bowel-cancer-facts).

In 2015, Dr. Baker collaborated with Concord Repatriation General Hospital researchers to draw data from a 20-year colorectal cancer surveillance study. The teams discovered uPAR (http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117786#sec002) (urokinase plasminogen activator receptor), a potential biomarker that may allow clinicians to better predict whether a Stage II (https://www.cancer.gov/types/colorectal/patient/colon-treatment-pdq#section/_112) colorectal tumor is likely to recur or whether surgery has been successful.

According to Dr. Baker, Stage IIB patients have cancer that is thought to be confined within the bowel wall, not spreading to nearby organs or affecting the lymph nodes. Surgery is a treatment option for Stage II patients, however, 25% of patients on average still recur, leaving patients with a much lower survival rate. If uPAR is expressed on the cancer cells in the primary tumor, it may predict recurrence in Stage II patients.

"Clinicians in the past haven't known how to judge if a Stage II patient is high risk, going to recur, or if surgery will be curative," Dr. Baker said. "This could be a real personalized approach to this population of rectal cancer patients, affecting approximately 200,000 people on the planet every year."

By using uPAR expression as a potential survival indicator in this subset of cancer patients, Dr. Baker hopes to arm clinicians with a simple test to further advise patients about the likelihood of cancer recurrence.

ICPC is fostering international collaborations like those between Drs. Baker and Paulovich laboratories, arming the international cancer research community with the knowledge necessary to better fight the terrible enemy that the world understands to be cancer.

"We're starting to see new therapies, better detection of cancer earlier, and improvements to existing interventions based on the fact of knowledge," Dr. Baker said. "The more knowledge you have about your enemy, the better off you are at fighting it."

Appendix V: High-resolution images of Publication 1.

High Resolution images for Section 1.4 (From Baker, Mark S., et al. "Accelerating the search for the missing proteins in the human proteome." *Nature communications* 8 (2017): 14271.)

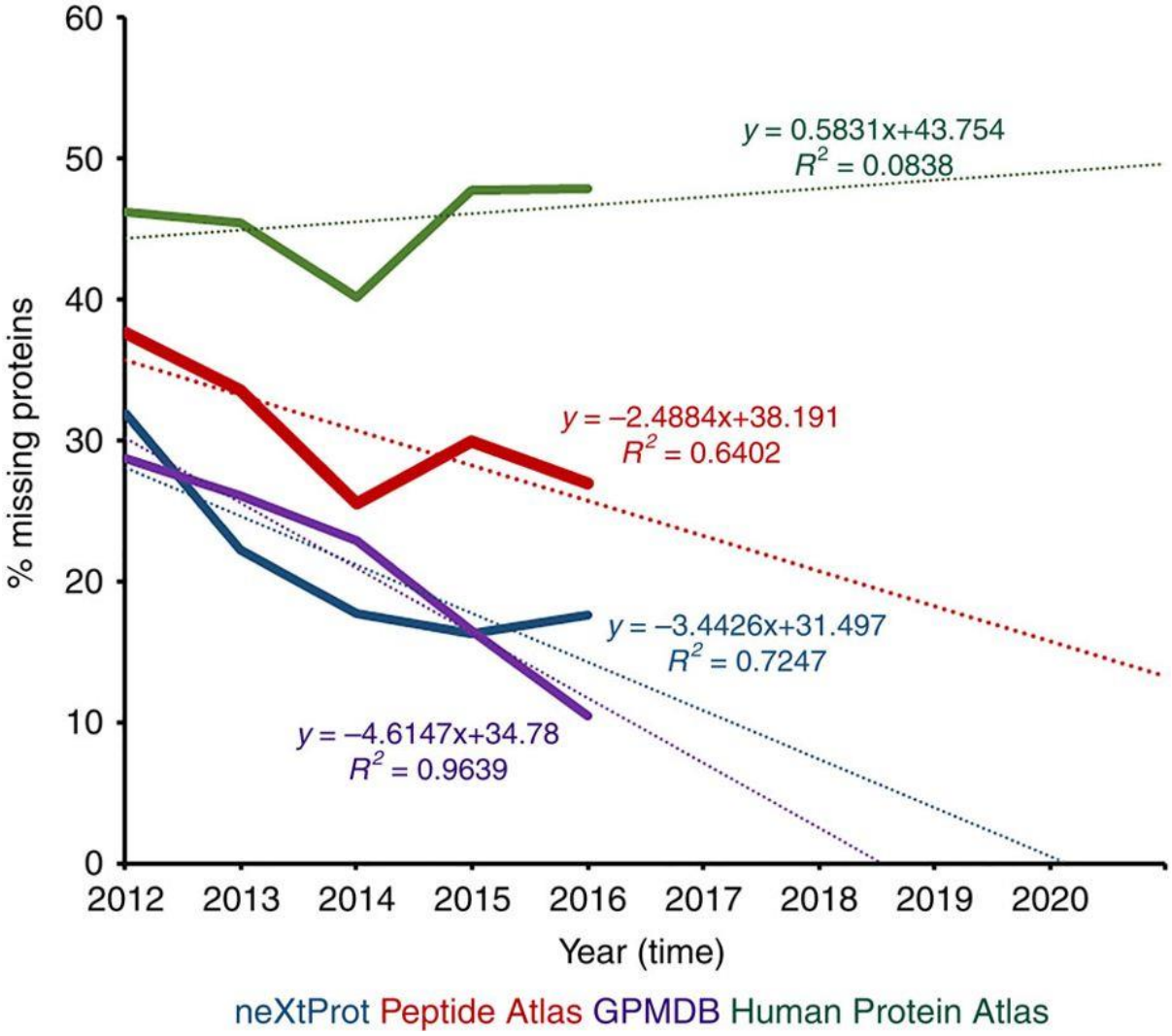


Figure 1: *Extrapolation of linear best-fit rate equations demonstrates the rate at which various HPP input databases and GPMDB are currently finding PE2-4 proteins.*

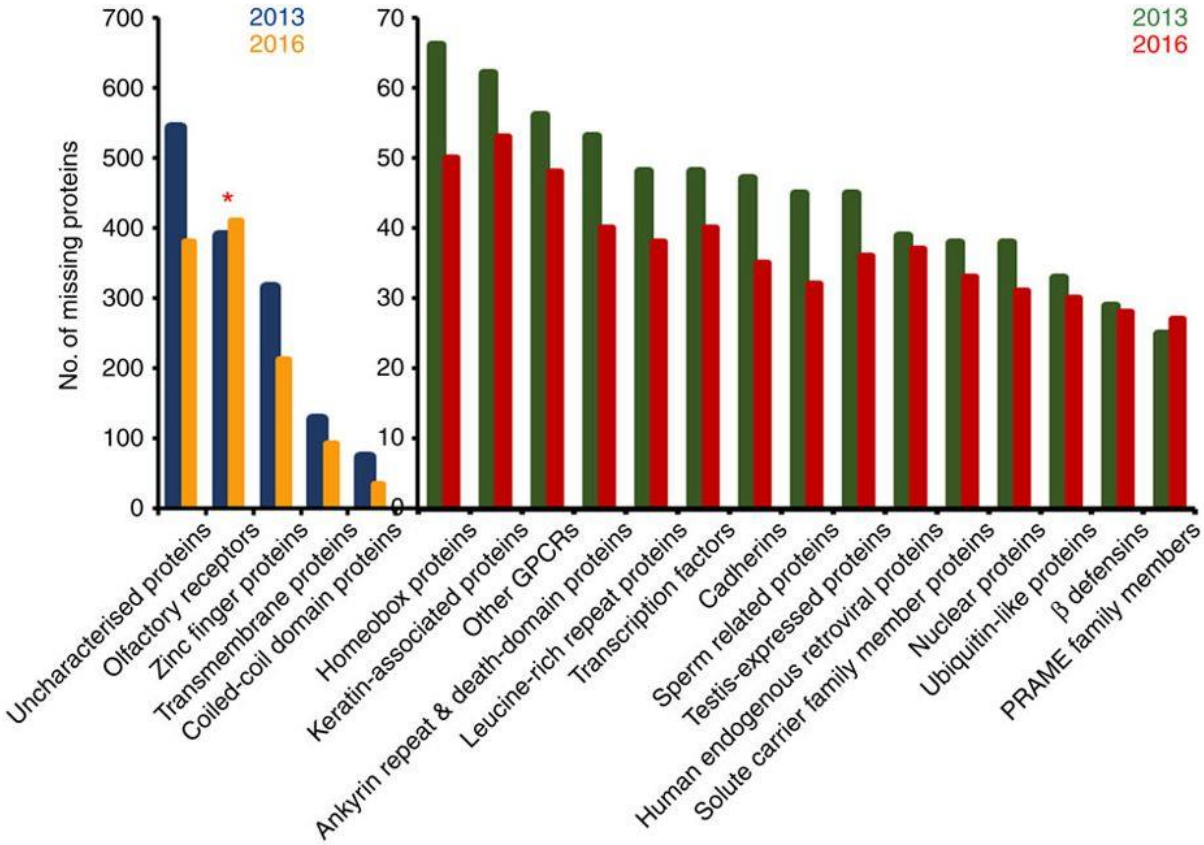


Figure 2: *Top 20 missing protein families to determine protein families enriched in the February 2016 neXtProt PE2-4 report list.*

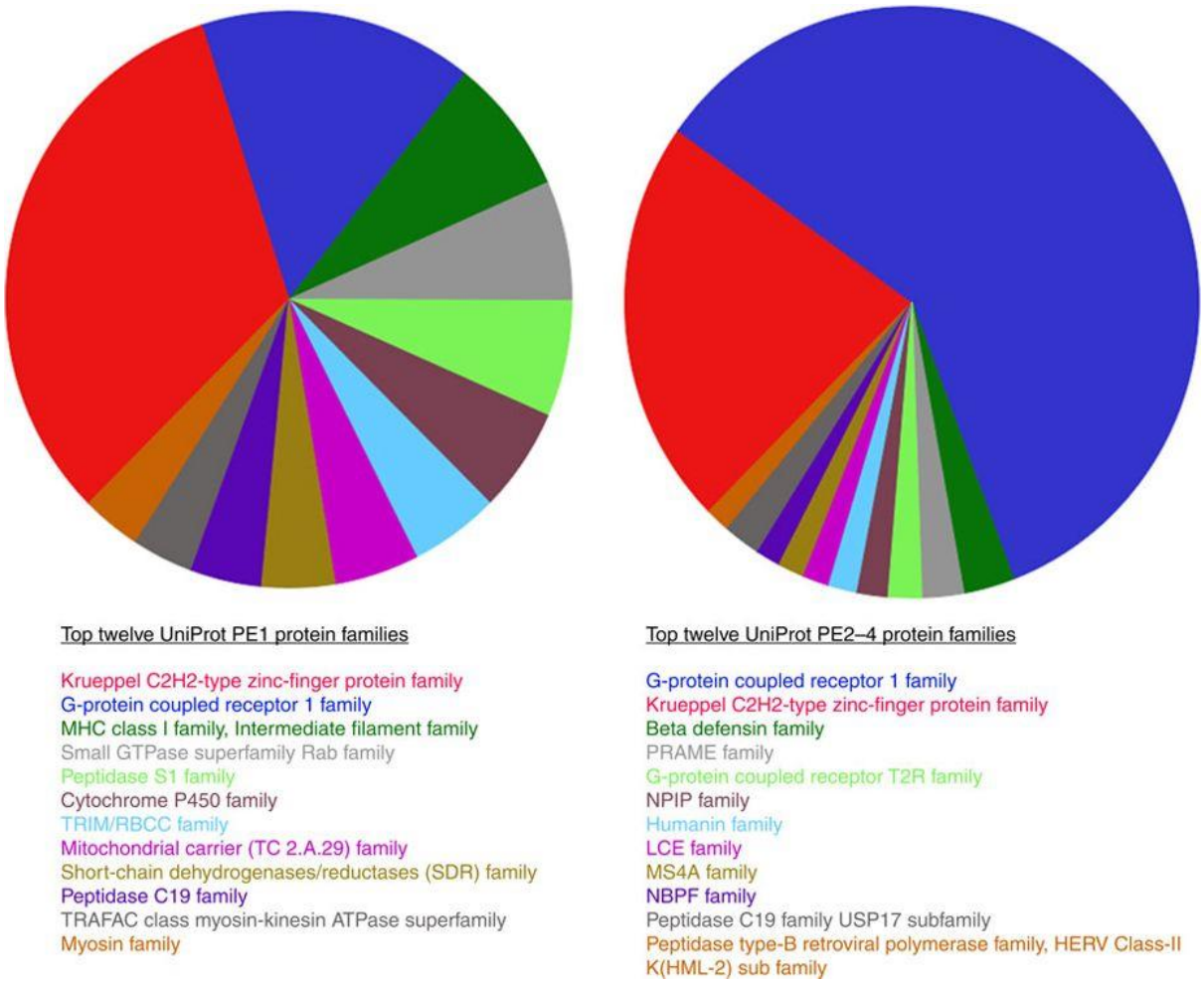


Figure 3: *Most prolific PE1 and 12 PE2-4 UniProt protein families represented in the HPP neXtProt February 2016 release.*

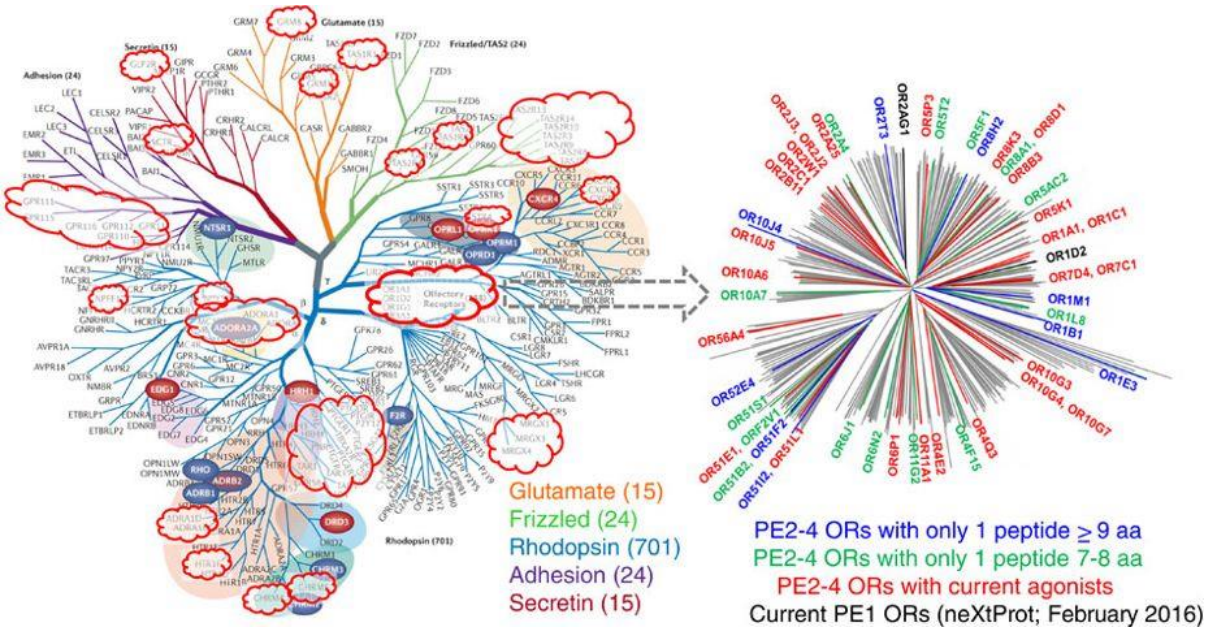


Figure 4: *Phylogenetic analysis of PE distribution across GPCRs and olfactory receptors.*

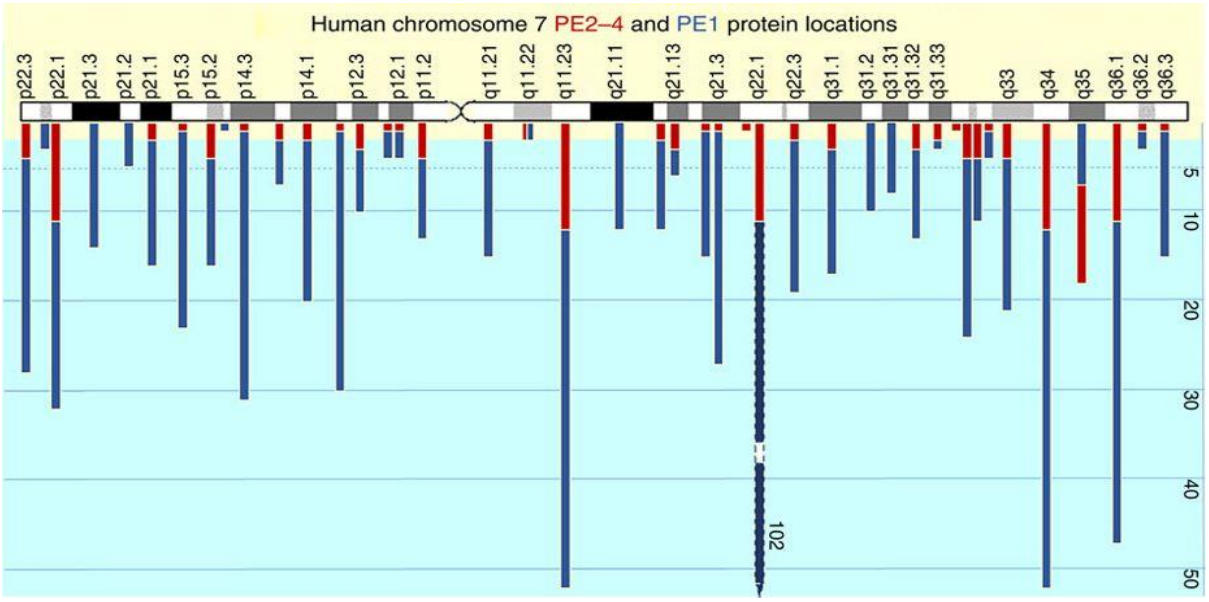


Figure 5: *Positional mapping of the PE1 (757) and PE2-4 (139) proteins along human Chromosome 7.*

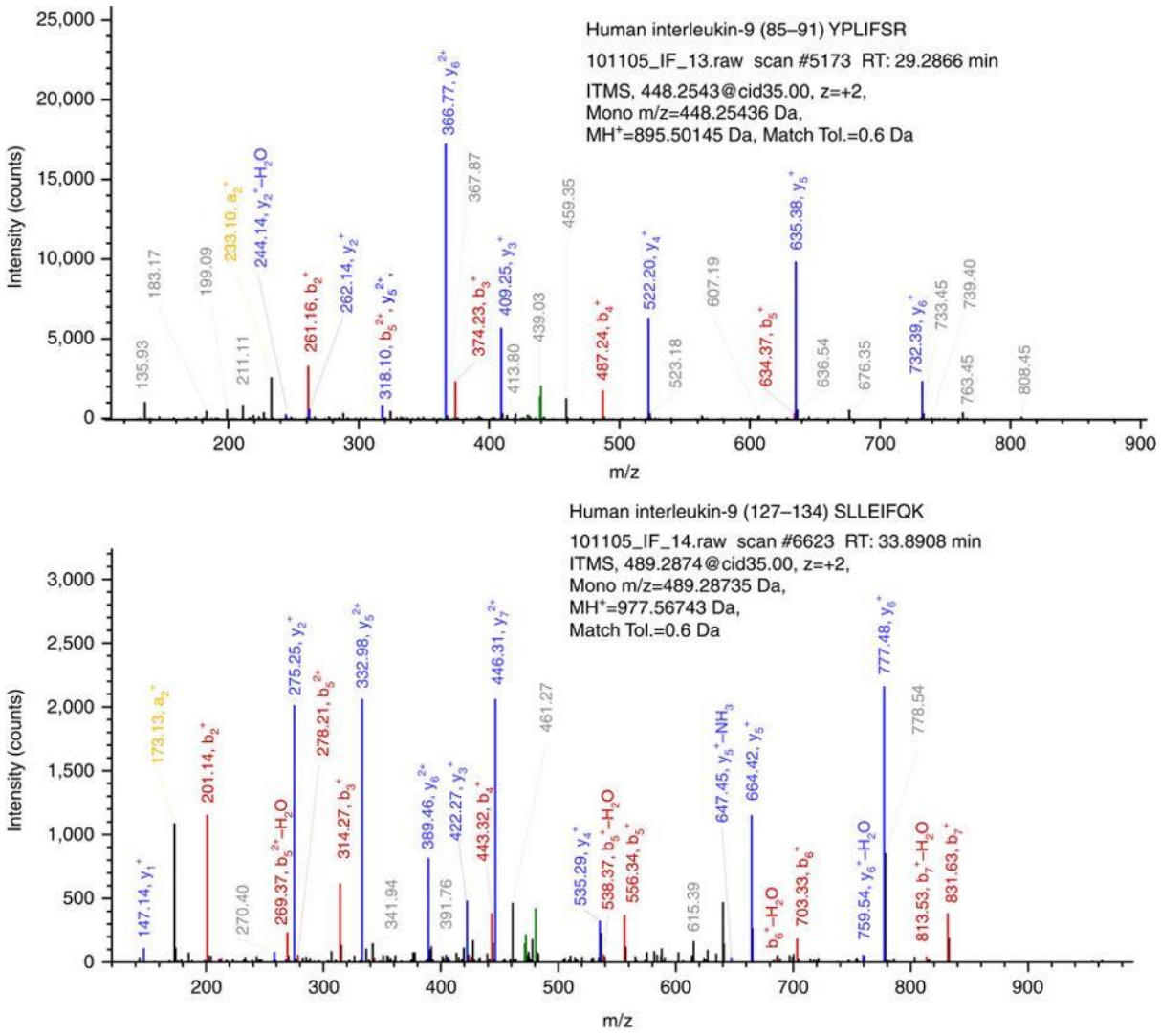


Figure 6: *Fragmentation spectra of two IL-9 proteotypic peptides detected in the secretome of activated T-cells.*

Appendix VI: Age, sex TNM staging, 5-year survival and 5-year recurrence data for recruited patients and healthy controls (n=100).

Age, sex TNM staging, 5-year survival and 5-year recurrence data for recruited patients and healthy controls (n=100). E: Healthy controls, A: CRC Stage I patients, B: CRC stage II patients, C: CRC stage III patients, D: CRC stage IV patients

IDs	Description	Sex	age	overall stage	TNM	5-yr Overall survival	5-yr Recurrence
E1	06AH087	M	69	E		Alive	No
E2	06AH276	F	58	E		Alive	No
E3	06AH301	M	73	E		Alive	No
E4	06AH360	F	66	E		Alive	No
E5	06AH367	F	65	E		Alive	No
E6	07AH130	F	50	E		Alive	No
E7	07AH253	M	55	E		Alive	No
E8	09AH109	F	53	E		Alive	No
E9	09AH296	F	54	E		Alive	No
E10	09AH695	M	65	E		Alive	No
E11	09AH710	M	68	E		Alive	No
E12	09AH727	M	69	E		Alive	No
E13	09AH794	F	57	E		Alive	No
E14	09AH795	M	79	E		Alive	No
E15	10AH033	M	57	E		Alive	No
E16	10AH484	M	75	E		Alive	No
E17	13AH0086	M	56	E		Alive	No
E18	13AH0087	F	62	E		Alive	No
E19	12MH0815	F	61	E		Alive	No
E20	12MH0063	F	65	E		Alive	No
A1	07AH359	M	56	A	T2, N0, MX.	Alive	No
A2	08AH228	F	70	A	pT1,N0,MX	Alive	No
A3	09AH457	F	63	A	T2 N0 MX	Alive	No
A4	09AH671	M	58	A	T2 N0 MX	Alive	Yes
A5	09NH111	F	64	A	T1, N0, MX, R0	Alive	No
A6	09NH115	M	74	A	T2, N0, MX, R0	Alive	No

A7	09NH135	F	71	A	pT2, L1, V0, N0, Mx	Alive	No
A8	09NH209	M	66	A	pT1, V1, L1, N0, MX	Alive	No
A9	09NH231	M	78	A	pT1, N0, Mx	Alive	No
A10	10AH547	M	68	A	pT2, N0, MX	Alive	No
A11	10AH703	M	62	A	T1 N0 Mx	Alive	No
A12	10AH752	F	55	A	pT1, N0	Alive	No
A13	10NH154	M	54	A	pT1,N0,Mx	Alive	Yes
A14	11AH0281	M	66	A	pT1 N0	Alive	No
A15	11AH0558	M	57	A	pT2 N0 MX	Alive	No
A16	05RMH238	F	79	A		Alive	No
A17	08WH253	F	72	A	pT1 N0 MX V0 R0	Alive	No
A18	11WH0063	F	51	A	pT1 pN0 MX R0 V0 G2)	Alive	No
A19	12WH0182	F	62	A	T2 N0 MX V0 R0	Alive	No
A20	12WH0194	F	56	A		Alive	No
B1	05AH277	F	74	B	Pt3n0	Alive	No
B2	05AH425	M	51	B	T3,N0	Alive	No
B3	05AH452	M	76	B	T3, N0, Mx	Alive	No
B4	06AH023	M	75	B	T3, N0, MX	Alive	No
B5	06AH287	F	66	B	T3, N0, MX	Alive	No
B6	06AH309	M	74	B	T3,No,Mx	Dead	No
B7	06AH352	F	52	B	T3, N0, MX	Alive	No
B8	07AH500	M	66	B	T3N0MX	Alive	No
B9	07AH613	F	73	B	pT3,N0,MX	Alive	No
B10	08AH702	M	68	B	pT3, N0	Alive	No
B11	08AH726	F	70	B	T4, N0, Mx	Alive	Yes
B12	10NH059	F	78	B	pT3b, N0, Mx	Dead	Yes
B13	10NH137	M	72	B	pT3b, V0, L1, N0, Mx	Alive	No
B14	10NH191	F	70	B		Alive	No
B15	10NH261	F	68	B	pT3 N0	Alive	No

B16	11AH0208	F	76	B	pT3 N0	Alive	No
B17	11AH0227	M	64	B	pT3 NO	Alive	No
B18	11AH0341	F	78	B	pT3 N0	Dead	Yes
B19	11NH0038	M	64	B	pT3, N0	Dead	Yes
B20	11NH0060	M	71	B	pT3 N0	Alive	No
C1	05AH355	F	73	C	T3, N1, MX.	Dead	Yes
C2	06AH097	M	66	C	T3N1	Alive	No
C3	06AH319	F	56	C	T3, N2, MX	Alive	No
C4	07AH070	F	76	C		Alive	No
C5	07AH233	M	52	C	T4, N2, MX	Dead	Yes
C6	07AH351	M	65	C	pT1, N1, MX	Alive	No
C7	07AH373	M	63	C	TNM T4, N2, MX	Dead	Yes
C8	07AH572	M	68	C	pT4, N1, M0	Alive	No
C9	08NH066	F	54	C	pT3b, N1, L0, V0, Mx	Alive	No
C10	09NH062	F	72	C	pT3a, N1, Mx	Alive	No
C11	09NH217	M	51	C	pT2, V1, L1, N1, MX	Dead	Yes
C12	09NH219	M	56	C	pT3A, N2, MX	Alive	No
C13	10NH004	F	69	C	pT3,N2,MX	Dead	Yes
C14	10NH009	F	58	C	T3B, N2, Mx	Dead	Yes
C15	10NH024	F	79	C	pT3d, V1, L0, N2, Mx	Dead	Yes
C16	10NH052	F	58	C	T3 N1 MX	Alive	No
C17	11NH0126	F	58	C	pT4a, N1	Alive	No
C18	11NH0210	M	67	C	pT3, N2a	Alive	No
C19	11NH0250	M	55	C	pT3 N1b	Alive	No
C20	12NH0036	F	57	C	pT3, N2b	Alive	No
D1	07AH401	M	79	D	T3 N1 M1	Alive	-
D2	10AH277	F	69	D	pT4 N2 M1	Dead	-
D3	11AH0102	M	61	D	pT3, N2a, M1	Dead	-
D4	11AH0490	F	71	D	-	Dead	-
D5	12AH0446	M	74	D	pT4a, N2b, M1a	Dead	-
D6	13AH0040	F	55	D	pT4a, N2b, M1a	Dead	-

D7	05WH102	M	53	D	T4 N2 M1	Dead	-
D8	05RMH117	F	59	D		Dead	-
D9	05WH131	M	63	D	T4 N2 M1	Dead	-
D10	06WH132	M	73	D	T3,N2,M1	Alive	-
D11	06WH176	M	62	D	T4, N2, M1	Dead	-
D12	07RMH006	M	62	D		Dead	-
D13	07WH211	F	61	D	T4,N1,M1,V0,R 2	Dead	-
D14	07WH218	F	75	D	T3 N2 M1 V1 RX	Dead	-
D15	07RMH580	M	56	D	T3 N0 M1	Dead	-
D16	08WH075	F	67	D	T4, N2, M0	Alive	-
D17	08RMH268	F	62	D		Alive	-
D18	08RMH506	F	78	D	T3 N1 M1 V2	Dead	-
D19	08SH655	F	63	D	Pt3n1m1	Alive	-
D20	10SH613	F		D	T4a, N2b	Alive	-