Application of informatics and quantitative proteomics to identify missing proteins and proteins involved in colorectal cancer metastasis

By

Subash Adhikari

Principal Supervisor: Professor Mark S. Baker

Co-Supervisor: Dr Seong Beom (Charlie) 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 2020

Statement of Originality

I hereby declare that the work presented in this thesis entitled "**Application of informatics and quantitative proteomics to identify missing proteins and proteins involved in colorectal cancer metastasis**" has not been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Subash Adhikari
44722656

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

31/01/2020

Acknowledgements

I would first and foremost like to thank my principal supervisor, Prof. Mark S. Baker, for presenting me an opportunity to be a part of the project. I am grateful to his continuous support, expert guidance and encouragement. Undertaking this thesis would not have been possible without a generous iMQRES scholarship offered by Macquarie University. I am grateful to my co-supervisor Dr Seong Beom (Charlie) Ahn and adjunct supervisor Prof. Eduard Nice for their continued support throughout my PhD candidature. I would like to sincerely thank Prof. Gilles Guillemin for being my acting supervisor during the later stages of my PhD.

I would like to especially thank Dr Abidali Mohamedali for assistance with every aspect of my PhD along with other past and current members of the Baker research team including Dr Susan Fanayan, Dr David I. Cantor, Dr Harish Cheruku, Dr Vineet Vaibhav, Dr Samridhi Sharma and Ms Sachini Fonseka for their cooperation and support. I would also like to thank Prof. Shoba Ranganathan for her support. Likewise, I would like to thank Prof. Janne Lehtiö for allowing me to revisit his lab at Science for Life Laboratory and Karolinska institutet, sincere thanks to Lehtiö lab team members including Dr Ghazaleh Assadi, Dr Rui Branca, Dr Henrik Johansson, Dr Helena Bäckvall and Dr Georgios Mermelekas.

I would like to acknowledge mass spectrometry access provided by APAF and HiRIEF-mass spectrometry access by Science for life laboratory Clinical Proteomics Mass Spectrometry facility, Karolinska Institutet/Karolinska University Hospital/ Science for Life Laboratory.

I am grateful to Sydney vital research scholar award, EMBL-Australia travel grant, Skipper Jacobs travel grant, HUPO travel awards during 2018 and 2019 and Macquarie

University Biomedical Sciences departmental travel funds; which contributed towards my research and travel expenditures.

I am thankful to Ms Laura Newey, Ms Viviana Bong and Ms Anne Clark for assistance on candidature related issues. Similarly, I would like to thank our lab operations team Ms Louise Marr, Ms Tamara Leo, Ms Lucy Lu and Mr Mitchel Borton for facilitating the smooth operation of my experiments.

Finally, I would like to express my gratitude to my parents, sisters and my beloved wife for providing support and encouragement throughout the years of my PhD candidature.

Subash Adhikari
January 2020, Sydney

Contribution Statement

The thesis contains manuscripts that have been published or are prepared for submission. Contributions leading to the generation of these manuscripts are listed below.

Manuscript I: *In Silico* Peptide Repertoire of Human Olfactory Receptor Proteomes on High-Stringency Mass Spectrometry

J Proteome Res. 2019 May 22. doi: 10.1021/acs.jproteome.8b00494

Subash Adhikari, Samridhi Sharma, Seong Beom, Ahn, Mark S. Baker

S.A performed R-based *in silico* analysis and generated an interactive HTML file that allows querying for multiple peptide annotations. M.S.B conceived the idea, S.S contributed on the update of the neXtProt missing proteins. **S.A**., S.B.A., and M.S.B. prepared the manuscript.

Manuscript II: Tryptic Peptidomic Landscape of All Human Multi-Transmembrane Domain Proteins

Subash Adhikari, Seong Beom Ahn, Mark S. Baker

M.S.B conceived the idea and **S.A performed all analysis** with inputs from M.S.B. **S.A**, S.B.A and M.S.B prepared and edited the manuscript.

Manuscript III: Quantification of Proteins from Proteomic Analysis

Zainab Noor, **Subash Adhikari,** Shoba Ranganathan, and Abidali Mohamedali

Encyclopedia of Bioinformatics and Computational Biology, 871-890, 2019

https://doi.org/10.1016/b978-0-12-809633-8.20677-8

S.A contributed to the label-free techniques. S.R and A.M conceived the idea, Z.N prepared Skyline workflow and analysis. Z.N, **S.A**, A.M and S.R prepared the manuscript.

Manuscript IV: uPAR interactome analysis to identify uPAR-ligand binding sites and their interaction confidence levels

Subash Adhikari, Seong Beom Ahn, Abidali Mohamedali, Mark S. Baker

S.A performed the interactome retrieval and analysis, sequence alignment and cystoscope based network creation with inputs from M.S.B and S.B.A. M.S.B conceived the idea, S.A, S.B.A, A.M and M.S.B prepared the manuscript.

Manuscript V: uPAR based interference peptides inhibit multiple metastatic phenotypes in CRC cell model expressing the integrin β6

Subash Adhikari, David Cantor, Seong Beom Ahn, Abidali Mohamedali, Janne Lehtiö, Edouard C. Nice and Mark S. Baker

M.S.B conceived the idea. M.S.B, E.N, S.B.A and J.L designed the experiments. **S.A** performed the proteomics experiments while S.B.A and D.C performed the background functional cell assays. **S.A** performed proteomics analysis with inputs from S.B.A, A.M and M.S.B. All authors contributed in preparing and revising the manuscript. HiRIEF-LC-MS analysis was performed by the Clinical Proteomics Mass Spectrometry facility, Karolinska Institutet/ Karolinska University Hospital/ Science for Life Laboratory.

Supplementary Manuscript I: Proteomics Reveals Cell-Surface Urokinase Plasminogen Activator Receptor (uPAR) Expression Impacts Most Hallmarks of Cancer Proteomics, DOI: 10.1002/pmic.201900026

Seong Beom Ahn, Abidali Mohamedali, Dana Pascovici, **Subash Adhikari**, Samridhi Sharma, Edouard C Nice and Mark S. Baker.

S.A performed the high pH fractionation, MS data submission to ProteomeXchange via MassIVE repository and generated the interactive HTML volcano plot based on JSON. M.S.B obtained the vectors for the construction of HCT116^{uPARAS} cell line and initiated the study. M.S.B, S.B.A, A.M and E.C.N designed the experiments. S.B.A, A.M, **S.A**, S.S **performed the experiment**, S.B.A, A.M, D.P, M.S.B, **S.A** prepared manuscript figures and tables. S.B.A and A.M conceived Ingenuity pathway analysis against the hallmarks of cancer. All authors contributed to writing and reviewing all versions of the manuscript.

Supplementary Manuscript II: (not a part of this thesis)

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

Clinical Proteomics, DOI: 10.1186/s12014-019-9255-z

Seong Beom Ahn, Samridhi Sharma, Abidali Mohamedali, Sadia Mahboob, William J. Redmond, Dana Pascovici, Jemma X. Wu, Thiri Zaw, **Subash Adhikari**, Vineet Vaibhav, Edouard C. Nice and Mark S. Baker

MSB, SBA, ECN designed experiments. S.B.A, S.S, A.M, S.M, T.Z, **S.A,** V.V performed experiments. W.R, D.P, J.W performed the statistical analysis. S.B.A, S.S, A.M, W.R, D.P, J.W, **S.A** prepared figures and tables. All authors contributed to the writing/reviewing of each manuscript version. All authors read and approved the final manuscript.

International presentations

Oral Presentations

1. Suppression of colorectal cancer proliferation and invasion by antagonising uPAR·αvβ6 interaction

Subash Adhikari, David Cantor, Seong Beom Ahn, Abidali Mohamedali, Mark S. Baker

Poster and Oral presentation, HUPO2017

2. Identifiable Human Olfactory Receptor Proteome Using High-Stringency Mass Spectrometry.

Subash Adhikari, Samridhi Sharma, Seong Beom Ahn, MarkS. Baker

Oral Presentation, HUPO2018

3. Suppressing proliferation, invasion and non-canonical MAPK signaling by antagonizing the cancer cell surface-restricted uPAR·αvβ6 protein interaction (#22)

Mark S. S Baker, **Subash Adhikari**, Seong Beom Ahn, Abidali Mohamedali, David Cantor

Oral Presentation, 23rd Annual Lorne Proteomics Symposium 2018

http://proteomics-2018.p.asnevents.com.au/days/2018-02-02/abstract/50205

Poster Presentations

1. Interference peptides (iPEPs) inhibits metastatic phenotypes in CRC cell model

Subash Adhikari, David Cantor, Seong Beom Ahn, Abidali Mohamedali, Mark S. Baker

Poster presentation, EMBL-EBI, Proteomics Informatics course, July 2018

2. Detection of Plasma Colorectal Cancer Prognostic Biomarkers

Seong Beom Ahn, Abidali Mohamedali, Dana Pascovici, Jemma Wu, **Subash Adhikari**, Samridhi Sharma, Thiri Zaw, Edouard C. Nice, Mark S. Baker

Poster presentation, HUPO2017

3. Bioinformatics analysis of DIA based mass spectrometry data to quantify the protein expression

Zainab Noor, **Subash Adhikari**, Abidali Mohamedali, Mark S. Baker, Shoba Ranganathan

Poster presentation, 23rd Annual Lorne Proteomics Symposium 2018

http://proteomics-2018.p.asnevents.com.au/days/2018-02-02/abstract/49721

4. Proteomics confirms lower cancer cell-surface uPAR superimposed on KRAS mutation carrying cells can negate many of the hallmarks of cancer

Seong Beom Ahn, Abidali Mohamedali, Dana Pascovici, **Subash Adhikari**, Mark Baker

Poster presentation, HUPO2018

5. Decreased cancer cell-surface urokinase plasminogen activator receptor (uPAR) negates several hallmarks of cancer

Seong Beom Ahn, Abidali Mohamedali, Dana Pascovici, **Subash Adhikari**, Edouard Nice, Mark Baker

COSA's 45th Annual Scientific Meeting, 2018

Grants

1. uPAR interference peptide (iPEP) antagonists of avß6·uPAR interaction as colorectal cancer (CRC) metastasis therapy leads

Sydney Vital Research Scholar Award

Adhikari Subash, Ahn Seong Beom, Fanayan Susan, Baker Mark

https://researchers.mq.edu.au/en/projects/upar-interference-peptide-ipep-antagonists-of-av%C3%9F6upar-interactio

2. Travel Grant, Skipper Jacobs Charitable Trust, July 2018

3. Travel Grant, EMBL-Australia, May 2018

4. Travel Grant, HUPO2018, September 2018

5. Bursary, Proteomics Informatics course, EMBL-EBI, July 2018

List of abbreviations

AGC	Automatic Gain Control
AJCC	American Joint Committee on Cancer
AKT	Protein kinase B
AOC	Area under the curve
APAF	Australian proteome analysis facility
APC	Adenomatous polyposis coli
API	Application programming interface
ATF	Amino terminal fragment
CEA	Carcinoembryonic antigen
CHAT	Cancer hallmarks analysis tool
CID	Collision induced dissociation
CIN	Chromosome instability
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CRC	Colorectal cancer
CT	Chromosome translocations
DAVID	The Database for Annotation, Visualization and Integrated Discovery
DDA	Data dependent acquisition
DIA	Data independent acquisition
EBI	European bioinformatics institute
ECD	Electron capture dissociation
ECM	Extracellular matrix
EGFR	Epidermal growth factor receptor
EMBL	European molecular biology laboratory
EMT	Epithelial mesenchymal transition
ERK	Extracellular-signal-regulated kinase
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FAK	Focal adhesion kinase
FDR	False discovery rate
FIT	Fecal immunochemical test
FMLP	N-formyl-L-methionyl-L-leucyl-phenylalanine
FOBT	Faecal occult blood test
FPR	False positive rate
GO	Gene ontology
GPCR	G-protein-coupled receptors
GPI	Glycosylphosphatidylinositol
GRAVY	grand average of hydropathy
GSH	Glutathione
HCD	Higher-energy collisional dissociation
HDI	Human development index
HGP	Human genome project
HIF	Hypoxia-Inducible Factor

HK	High-molecular-weight kininogen
HPA	Human protein atlas
HPLC	High-Performance Liquid Chromatography
HPP	Human proteome project
HTML	Hypertext Markup Language
HUPO	Human proteome organization
IARC	International Agency for Research on Cancer
ICPC	International Cancer Proteogenome Consortium
IEF	Isoelectric focusing (IEF)
IPA	Ingenuity Pathway Analysis
IPG	Immobilized pH gradient
JAK	Janus kinase
JNK	c-Jun N-terminal kinases
JSON	JavaScript Object Notation
KB	Knowledgebase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KKS	Kallikrein-kinin system
LCMS	Liquid chromatography and mass spectrometry
LDL	Low-density lipoprotein
LIT	Linear ion trap
LOD	Limit of detection
LRP	Low density lipoprotein receptor-related protein
LTQ	Linear triple quadrupole
MAPK	Mitogen-activated protein kinases
MMP	Matrix metallopeptidases
MP	Missing proteins
NSAF	Normalized spectral abundance factor
PAI	Plasminogen activator inhibitor
PAICQIC	Proteomics Standard Initiative Common QUery InterfaCe
PAS	Plasminogen activation system
PE	Protein existence
PET	Positron-emission tomography
PKC	Protein kinase C
PPI	Protein -protein interaction
PRIDE	PRIDE PRoteomics IDEntifications
PSI	Proteomics standards initiative
PSM	Peptide spectrum match
PTM	Post translational modification
QE	Q-Exactive mass spectrometer
RTK	Receptor tyrosine kinase
SILAC	Stable Isotope Labeling by/with Amino acids in Cell culture
SRM	Selected reaction monitoring
STAT	Signal transducer and activator of transcription

SWATH	Sequential window acquisition of all theoretical fragment ion spectra
TDA	Target decoy approach
TMD	Transmembrane Domain
TMT	Tandem mass tag
TNM	Tumor node metastasis
TOF	Time of flight
VEGF	Vascular endothelial growth factor
VN	Vitronectin
XIC	Extracted Ion Chromatogram

Table of contents

Thesis Executive Summary

Accurate management of any human disease requires a thorough understanding of the molecular underpinnings (i.e., genome + epigenome + transcriptome + proteome + peptidome + protein post-translational modifications + metabolome + microbiome) driving the biology of that specific disease [1].

Given the intricate and expanding roles played by proteins in human health and disease, these have been studied extensively to uncover disease mechanisms, define diagnostic, prognostic and theranostic markers and identify novel therapeutic targets [2].

Over the past decade, high throughput mass spectrometry (MS)-based proteomics with subsequent bioinformatic analysis have emerged as one major technological driving force in our attempts to expand human proteomics so that it has a noticeable impact on medicine, human health and the life sciences alike [3,4].

The Human Proteome Project (HPP) provides a framework for communal proteomics research. It specifically adds value to the task of *'knowing thyself'* in strictly molecular terms by mapping the ~20,000 proteins encoded by the human genome [5,6]. It aims to do so as a corollary to the human genome using measurements at the highest possible accuracy and stringency [7].

The HPP states that it has three initial primary aims, namely to; "

i. *complete the protein 'parts list' of Homo sapiens by identifying and characterizing at least one protein product and as many posttranslational modifications, single amino acid polymorphisms and splice variant isoforms as possible for each protein-coding gene;*
ii. *transform proteomics so it becomes complementary to genomics across clinical, biomedical and life sciences through technological advances*

iii. *create knowledgebases for the identification, quantitation and characterization of the functionally networked human proteome."* [8]

This thesis contributes to two major elements of the HPP (**C-HPP**, Chromosome-Centric-HPP and **B/D-HPP**, Biology/Disease-HPP) through the use of informatics and proteomics approaches. Expanding previous research efforts by our HPP team at Macquarie University, ***Chapter 1*** aims to advance community-centric resources to accelerate the identification of missing proteins (MPs). This chapter provides a plausible explanation for the observed paucity in identifications of certain missing protein family groups that have failed to be identified by MS over the last decade - namely the olfactory receptors (ORs). Analysis of OR hydrophobicity, topological distribution, tryptic cleavage site and frequency, and ability to predict *in silico* uniquely-mapping, non-nested tryptic peptides of a communally-required length (9 amino acids or longer) indicated that multiple ORs are unable to generate peptides as per requirements set by the HPP to be called protein existence 1 (PE1 for short) proteins. These ORs may not be MS-identifiable unless they qualify for relaxed stringency criteria or other proteases are used to generate suitable peptides from which we can infer protein identification. A similar observation was made when this analysis was expanded for all multi-transmembrane domain (TMD)-containing human membrane proteins coded for by the human genome.

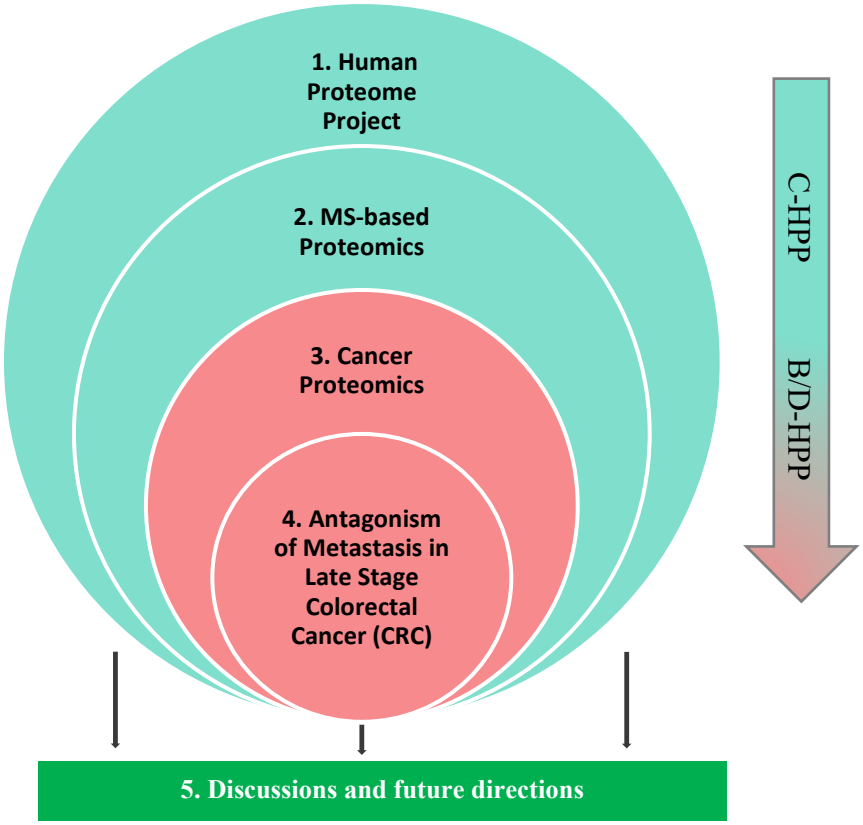


Figure 1: Thesis Table of Contents Graphic: This thesis aims to expand our knowledge of the HPP by contributing to some of its central aims, namely; (i) identification and characterization of previously uncharacterized proteins (Chapter 1), and (ii) advancing the diagnosis and treatment of human disease (medicine) thorough a better utilisation of MS-based proteomics research (Chapter 2-4). The final chapter 5 provides a summary and the potential future research endeavours based on this thesis.

Chapter 2 presents details on the mass spectrometry instrumentation and its application in quantitative proteomics research, the chapter contains an exemplar label-free quantitative proteomics analysis workflow on Skyline [9].

In **Chapter 3**, state of the art proteomics methodologies was employed to understand possible means to image, target and treat human disease. In particular, the second part of my thesis contributes to the Cancer-HPP efforts (part of the Biology/Disease HPP) specifically in colorectal cancer (CRC). In that section, I outline studies directed against

a candidate biomarker of aggressive, invasive, metastatic CRC, namely urokinase plasminogen activator receptor (uPAR). uPAR has been reported by many teams to be elevated during tissue remodelling, inflammation and in many human epithelial cancers. uPAR mediates extracellular matrix (ECM) proteolysis and cellular-ECM interactions, which are facilitated by cross-talk between uPAR and multiple ligands (e.g., integrins). Systematic analysis of these ligands was performed to generate a uPAR interactome based on publicly available protein-protein interaction resources. Clustering of uPAR binary interactions was performed based on the type of method used for interaction confirmation, to produce a score which represents interaction confidence.

Finally, the interaction of cancer cell surface proteins like uPAR with specific integrins (in this case with αvβ6) have shown to produce a complex referred to here as uPAR-αvβ6. This has been discovered in our group and is now known to express in a percentage of CRC cases to be involved in driving late-stage CRC metastasis.

Rationally-designed interference peptides (iPEPs) derived from uPAR surfaces were explored as potential antagonists of uPAR-αvβ6 in **chapter 4**. Binding of uPAR iPEPs to αvβ6 inhibited many hallmarks of cancer involved in mediating the effects of this interaction. For example, various uPAR sequence-derived iPEPs bound to the external surface of the SW480 cells that are known to express the uPAR-αvβ6 and this resulted in a switch from canonical transforming growth factor-beta (TGFβ) signaling to non-canonical mitogen-activated protein kinases (MAPK) signaling. iPEPs treatment decreased the proliferation, migration and invasion on cells expressing the interaction when compared to the controls. High-Throughput proteomics analysis was performed to identify processes associated with these iPEPs induced biologies by a combination of a tandem mass tag (TMT) [10], high-resolution isoelectric focusing (HiRIEF) [11] and liquid chromatography with tandem mass spectrometry (LC-MSMS). The study led to

the identification of proteins involved in embryogenesis, wound healing, cell differentiation and migration. iPEPs inhibit many hallmarks of CRC and represent a potential therapeutic lead that may be explored for uses in the future treatment of advanced CRC.

Finally, **chapter 5** summarizes the thesis and provides potential future endeavours of this thesis.

References

1. Petricoin, E. F., Zoon, K. C., Kohn, E. C., Barrett, J. C. & Liotta, L. A. Clinical proteomics: Translating benchside promise into bedside reality. *Nature Reviews Drug Discovery* (2002). doi:10.1038/nrd891
2. Ong, S.-E. & Mann, M. Mass spectrometry–based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).
3. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
4. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
5. Hanash, S. & Celis, J. E. The Human Proteome Organization: a mission to advance proteome knowledge. *Mol. Cell. Proteomics* **1**, 413–4 (2002).
6. Legrain, P. *et al.* The Human Proteome Project: Current State and Future Direction. *Mol. Cell. Proteomics* **10**, M111.009993 (2011).
7. Deutsch, E. W. *et al.* Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* **15**, 3961–3970 (2016).
8. Baker, M. S. *et al.* Accelerating the search for the missing proteins in the human proteome. *Nat. Commun.* **8**, 14271 (2017).

9. Pino, L. K. *et al.* The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrom. Rev.* (2017). doi:10.1002/mas.21540

10. Andrew Thompson, † *et al.* Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. (2003). doi:10.1021/AC0262560

11. Branca, R. M. M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* **11**, 59–62 (2014).

Chapter 1: Human proteome project

1.1 Overview

The Human Proteome Organisation (HUPO) has coordinated a global collaborative scientific project called the Human Proteome Project (HPP), since its official launch at the HUPO2010 Congress in Sydney on September 23rd, 2010 [1]. The HPP has made significant contributions in setting community guidelines, metrics and structures to facilitate collaborative human proteomics research and to support the identification and characterization of what has been colloquially called “missing proteins” (MPs) over the past decade [2,3].

Collaboration between various international groups has resulted in the definition of many proteomics standards and guidelines that ensure reliable, accurate, high-stringency data analysis, interpretation and reporting [3]. Characterization of the complete human proteome is the main goal of HUPO’s HPP, and this would be instrumental in coming to a better, more complete understanding of both normal human biology and human disease.

Progress on identifying and characterizing the complete human proteome is manifested through a neXtProt protein existence (PE) status. Proteins characterized with "definitive" evidence their existence are assigned PE1 status, whereas protein lacking the evidence or containing partial evidence are categorized into PE2-4 proteins, referred to MPs and finally uncharacterized proteins are categorized into PE5 status [4]. MS-based PE1 assignment requires a protein to generate a minimum of two experimentally derived uniquely mapping non-nested peptides of minimum nine amino acid length [2].

This chapter describes various HUPO’s HPP initiatives and provides an explanation for paucity in the MS-based PE1 assignment for certain family of proteins over the last

decade, i.e., membrane proteins. The ability of these proteins to generate theoretical peptides qualifying MS-based PE1 status was exemplified by a seven-transmembrane domain (TMD) containing olfactory receptors (ORs) and subsequently expanded to all multi-TMD containing membrane proteins. Lack of any MS-based PE1 assignment over the last decade makes ORs a suitable candidate for such analysis to identify the reasons behind their paucity in MS evidence.

In silico analysis of all ORs domains allowed me to hypothesize that the low frequency of tryptic proteolytic resides, i.e., arginine (R) and lysine (K) on TMD regions, potentially restricts tryptic activity and limits soluble peptide yield during experimental MS analysis. This suggests that the experimental peptide yield from a conventional MS approach should be similar to *in silico* peptide sub-population derived from the non-TMD regions. Our analysis suggested that most of the ORs can generate peptides as per the HPP stringency requirement from their soluble (non-TMD) domains, with a few exceptions even when TMDs are considered. Similarly, upon expansion of the analysis to all TMD containing human proteins, additional proteins that could not generate peptides qualifying HPP guidelines upon whole sequence tryptic digest were identified.

1.2 Introduction

The Human Genome Project (HGP) presented the potency of omics approach in regard to improving our understanding of human biology through a genomics perspective [5]. Multi-omics approaches are now utilized to complement genomics to decode the information flow between genome and phenotype [6]. Mass spectrometry is one the widely adopted omics platform in systems levels study of the expressed phenotype by quantifying the proteome level changes [7]. Identification of a critical mutational landscape, associated protein expression levels and post-translational modifications

have the potential to aid in the establishment of the novel actionable targets and define advanced disease management strategies [8]. Recent advancements in MS technology with; (i) increased sensitivity and specificity of the instruments, and (ii) cost-effective methodologies have stimulated multiple frontiers of medical research [9,10].

Proteomics is the large scale simultaneous identification and/or quantitation and/or determination of post-translational modification of proteins. It has been widely adopted to evaluate protein level systemic changes in a range of human biology and disease applications [11]. Advancement in MS instrumentation such as the introduction of electrospray ionization (ESI)[12] and Orbitrap™ [13] have enabled proteomics to transform into a globally adopted high-throughput application, routinely utilized in identification and quantification of thousands of proteins in tandem [14]. Global efforts in the identification and characterization of the human proteins were institutionalized with the establishment of the Human Proteome Organization (HUPO).

1.2.1 C-HPP and B/D-HPP

During 2008-2010, the HUPO HPP initiated two parallel streams, namely the Chromosome-Centric HPP (C-HPP) and the Biology/Disease HPP (B/D-HPP). These provided a matrix-approach maximising the identification and quantitation of genome-coded protein data from both genomics-focussed approaches (where each gene was effectively given equal prominence) for the C-HPP mirrored against a more practical approach focussed on those proteins involved in specific biologies/diseases under the B/D-HPP. It was argued that such bi-streamed approach would ensure all proteins coded by the genome were provided equal opportunity to be investigated as part of the HPP [15]. The C-HPP was launched in 2012 and is involved in mapping and characterization of proteins encoded by human chromosomes 1-22, X and Y and include mitochondrially-coded genes.

B/D-HPP aims at studying the molecular mechanisms and biological processes associated with the human disease through state-of-the-art proteomics applications. The B/D-HPP provides a foundation for community-wide collaboration between proteomics-based biology and disease research groups [16]. B/D-HPP has generated a highly specific protein cohort that plays a vital role in human disease and biology, maintained by PeptideAtlas. (https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/proteinListSelector) [17].

Similarly, other HPP initiatives include CompMS (Computational Mass Spectrometry), Early Career Researcher (ECR), Human Antibody, iMOP (Initiative on MultiOrganism Proteomes) and Proteome Analyzer (https://hupo.org/hupo-sponsored-initiatives). Along with these initiatives, the Cancer (Biden) Moonshot and NCI's Clinical Proteomic Tumor Analysis Consortium (CPTAC), International Cancer Proteogenomics Consortium (ICPC) are the crucial independent initiative that shares close collaborative ties with HUPO. Collectively, with HUPO's Cancer-HPP initiative under B/D-HPP, these programs aim to accelerate cancer research to aid early cancer diagnosis and develop effective cancer treatment strategies for patients. (https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative). CPTAC and the ICPC aim to accelerate cancer research through integrating proteomics, post-translational modification, genomics, epigenomics and transcriptomics data (i.e., proteogenomics; https://proteomics.cancer.gov/programs/cptac) to similar ends [18, 19].

1.2.2 HPP pillars

The HPP is a HUPO-coordinated global scientific initiative that aims at the characterization of the complete human proteome, which will expand the understanding of human biology at the molecular level, and aid in the characterization of novel diagnostic, prognostic, therapeutic and preventive measures for multiple diseases [20]. HPP constitutes four pillars i) mass spectrometry, ii) knowledgebase (KB), iii) antibody and iv) recently introduced pathology pillars, established based on their importance on a large scale proteomics analysis (https://hupo.org/About-the-HPP) [20].

The HPP KB pillar compiles and curates protein annotations. neXtProt (www.nextprot.org) [21] is a primary KB component that provides an update on the PE status, derived from the annual communal reanalysis of selected publicly-available MS datasets by PeptideAtlas (www.peptideatlas.org) [17]. UniProtKB database (www.uniprot.org) [22] is one the widely used and comprehensive KB for protein annotation. Data-intensive nature of the MS application has made the establishment of data repositories unavoidable, leading to the establishment of the ProteomeXchange Consortium (http://www.proteomexchange.org/). The consortium encourages open data policies between significant proteomics repositories and provides a framework for communal standards for data submission and dissemination [23]. Members of this consortium constitute: PRIDE PRoteomics IDEntifications (PRIDE) database (https://www.ebi.ac.uk/pride/archive/) [24], PeptideAtlas (http://www.peptideatlas.org/), [17] MassIVE (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp), jPOST (https://jpostdb.org/), iProX (https://www.iprox.org/) [25] and Panorama Public (https://panoramaweb.org) [26]. HUPO Proteomics standards initiative (PSI) (http://www.psidev.info/), defines the standards for proteomics and interactomics data representation standards. PSI has released a set of controlled vocabularies for data

annotation standardization [27]. Data formats for different MS-related data is being regularly updated, facilitating data comparison and exchange [28].

The HPP MS Pillar comprises the analytical methods and technology platforms that allow high-throughput identification and quantification of peptides and proteins. Verified reference spectra of human proteome derived peptides facilitate proteomics studies. These peptides provide a basis for the development and validation of targeted studies such as Selection Reaction Monitoring (SRM).

The HPP affinity/antibody pillar aims to facilitate the adoption of antibody-based data to map the human proteome in cells, tissue, health and disease and to accelerate human "omics" research. The initiative primarily reports through the Human Protein Atlas (HPA) and seeks to index validated antibodies and generate a protein and transcriptome expression and localization atlas across a broad range of healthy and diseased human tissues (https://hupo.org/Human-Antibody-Initiative). Ongoing findings of this initiative will be available through the publicly available antibody resource database (www.antibodypedia.org). The human protein atlas initiative (www.proteinatlas.org) houses annotated tissue profiling between healthy and cancer tissues, stratified based on the high-resolution tissue imaging [29].

Introduction of the Pathology Pillar during the HUPO2018 world congress highlights the role played by disease pathophysiology in understanding potential uses of proteomics-based biomarker discovery in diagnostic settings. The pillar will collaborate with the identification of unmet clinical needs and formulate guidelines for reproducible and valid clinical assays. The pillar aims to provide linkages between the acquisition of clinical samples and integration of associated proteomics data for clinical applications (https://hupo.org/page-1757231). A diagrammatic representation of the initial structure

of the HPP, including the two original streams and three pillars is summarized in Figure 1.2.

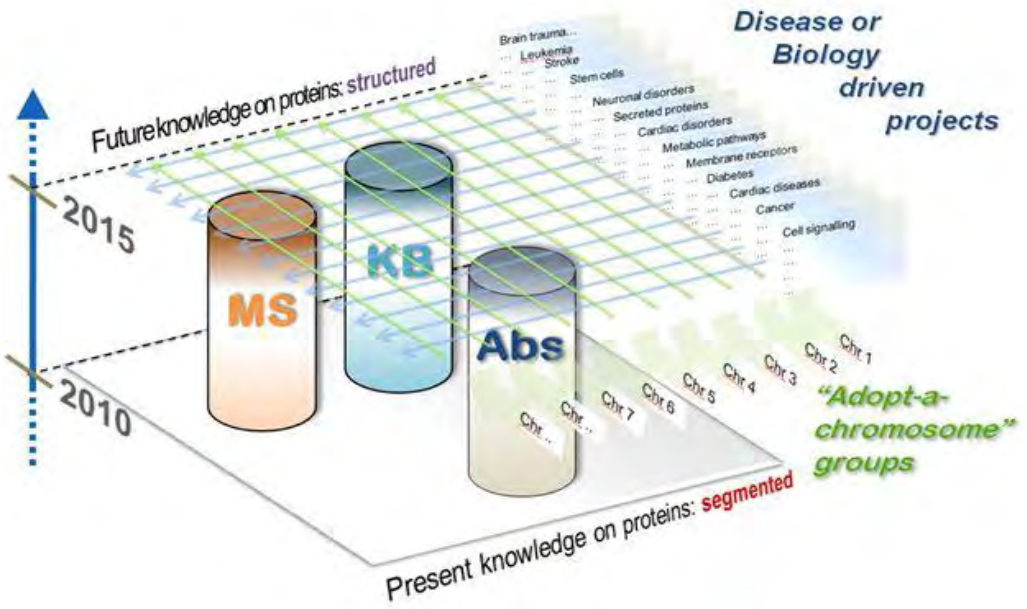
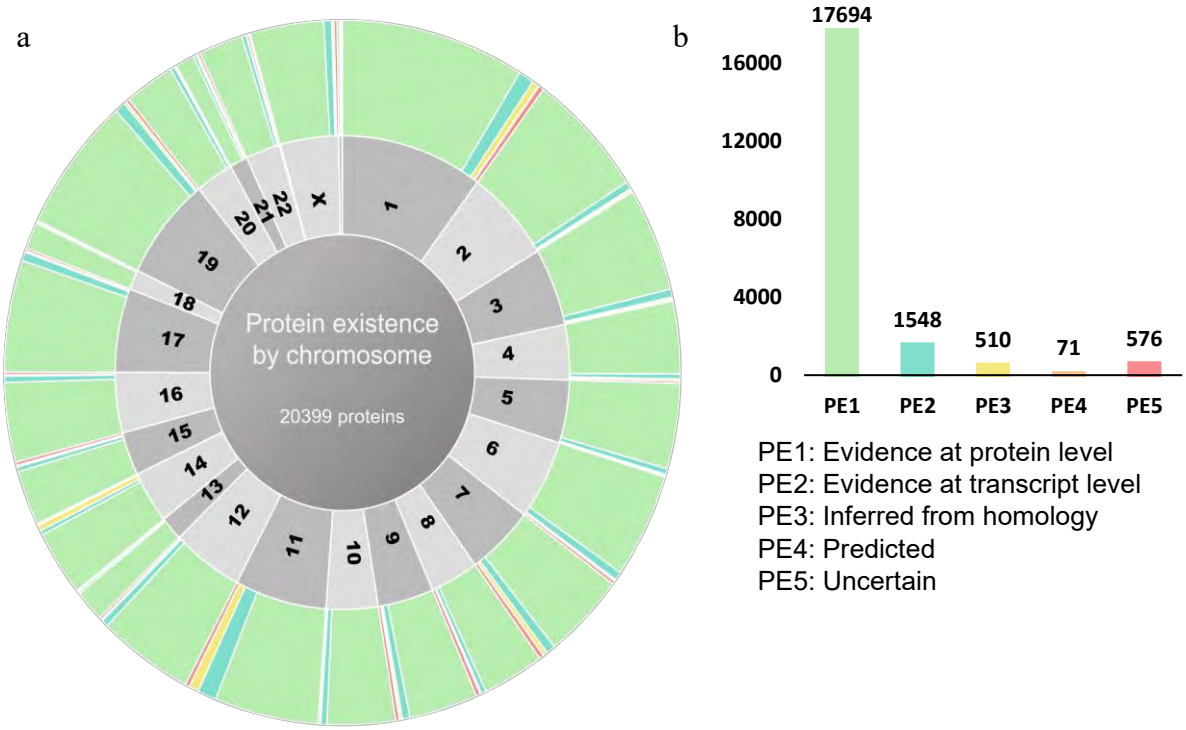


Figure 1.2: HPP initiated complementary C-HPP and B/D-HPP project based on four different HPP pillars. Advancement in MS-based technology combined with HPP collaborative efforts led to the characterization of multiple missing proteins over the last decade [30]. Reproduced from https://hupo.org/About-the-HPP.

1.2.3 Missing proteins



HPP categorizes all human proteins into five distinct (protein existence; PE) groups based on the evidence that supports an argument for their existence, referred to as PE status with five categories (PE1, PE2, PE3, PE4, PE5). Proteins with high-stringency MS or neXtProt-qualified antibody-based evidence are assigned PE1 status. PE2-5

Figure 1.3: Current (v2.22.2) neXtProt PE status for the human proteome. Through neXtProt, the HPP has progressively assigned PE1 status to multiple proteins since 2002. However, challenges remain in the identification and characterization of certain protein groups such as membrane proteins. a) Distribution of human proteins across different chromosome and PE status shows significant number of human proteins are PE1s. b) Number of human proteins across PE status, current neXtProt build contains 2129 MPs (PE2-4).

proteins refer to proteins with evidence-based on transcript level only (PE2), homology only (PE3), prediction only (PE4) or those that are uncertain/dubious genes/proteins (PE5) [21]. The "missing proteins" (MPs) originally referred to all PE2-5 proteins, but the definition was recently modified to only refer to ΣPE2-4 human proteins. Current HPP release lists 2129 proteins as MPs out of total 20,399 proteins (https://www.nextprot.org/about/protein-existence). Lack of protein level MS or antibody evidence exists due to; low abundance, nil or low-level expression, incorrect genome annotation, incompatible structure for MS studies etc. Advancement in proteomics techniques combined with collaborative HPP efforts has resulted in a gradual decrease in the number of missing proteins from 6,568 during the inception of C-HPP in 2012 to the current 2,109 PE2-4 MPs [30].

neXtProt maintains and updates PE status for all human proteins at regular interval and this data is available at the web page link https://www.nextprot.org/about/protein-existence [4]. HPP has also set guidelines for the assignment of PE1 status. These guidelines require; (i) protein containing MS-based experimental 2+ uniquely mapping non-nested peptides that are nine or more amino acid (aa) in length, (ii) Protein existence validated based on antibody evidence and (iii) PE2-4 proteins with GOLD binary interaction data in neXtProt [4]. neXtProt provides an update on PE status of human proteins through an annual HPP date and the 2018/2019 release lists 2,129 PE2-4 MPs proteins out of a denominator total of 20,399 proteins derived from all currently known protein-coding genes.

Table 1.1: Summary of current (v2.22.2) neXtProt release statistics. neXtProt provides an annual update on PE1-5 and MP identification. Currently, Chromosome (Chr) 1 (the largest) and Chr 11 contain the highest number and percentage of PE2-4 MPs.

Chromosome	*PE1*	*PE2*	*PE3*	*PE4*	*PE5*	*Total*
1	1796	150	66	8	49	2,069
2	1178	71	18	1	17	1,285
3	969	76	19	3	15	1,082
4	678	46	21	2	21	768
5	796	55	9	3	13	876
6	983	79	12	6	30	1,110
7	808	89	41	5	42	985
8	602	45	14	2	39	702
9	671	71	21	12	35	810
10	665	60	7	1	17	750
11	1014	184	95	1	39	1,333
12	932	64	14	0	23	1,033
13	297	20	4	1	12	334
14	613	40	56	4	14	727
15	513	43	17	1	30	604
16	745	58	10	2	24	839
17	1049	82	13	3	22	1,169
18	250	13	1	2	10	276
19	1271	104	25	2	32	1,434
20	488	43	3	5	12	551
21	186	28	18	1	23	256

22	432	39	5	1	21	498
Mitochondrial	15	0	0	0	0	15
X	720	76	18	5	29	848
Y	26	11	3	0	7	47
Unknown	2	2	0	0	0	4
Total	**PE1**	**PE2**	**PE3**	**PE4**	**PE5**	**Total**
	17,694	1,548	510	71	576	20,399
	86.7%	7.6%	2.5%	0.4%	2.8%	100%

The neXtProt uniqueness checker (https://www.nextprot.org/tools/peptide-uniqueness-checker) provides information on the identification of uniquely-mapping non-nested peptides [31]. Multiple tools are available for protein and peptide annotation based on the SPARQL search/query engine through neXtProt REST API (https://api.nextprot.org/).

The HPP has made a significant achievement over the last decade in assigning high portion (86.7%) of proteins from the protein-coding genome to PE1 status. For examples, multiple PE2-4 MPs of protein families/groups like the coiled-coiled domain proteins, homeobox proteins and keratin-associated proteins have been assigned PE1 status. However, there remain significant challenges for several PE2-4 protein families/groups. MS identification of multi- transmembrane domain-containing proteins largely remains problematic. These include notable examples like the GPCR olfactory receptor (OR) and taste receptor families, primarily due to sample scarcity and technical challenges associated with the identification of hydrophobic membrane proteins [32].

HPP data interpretation guidelines v2.1 define requirements for MS-based identification of any PE2-4 MPs. These guidelines mandate PE1 assignment should be based on the

presence of a minimum of two uniquely-mapping non-nested experimentally-derived peptides that are nine or more amino acid in length [2]. Exceptions to these rules can now be considered on a case-by-case basis, provided there is plausible scientific evidence that a given protein is not able to generate the stringency required for peptides as per HPP guidelines [2]. Exception criteria can be expected to streamline characterization of certain protein groups/families, such ORs where no MS evidence exists, as ~1% of the ORs have been classified as a PE1 based on non-MS evidence [32]. OR1D4 and OR2AG1 PE1 assignment were derived from a convincing biochemical, genetic or haplotype data, whereas PE1 assignment of OR1D2 and OR2J3 was based on protein interaction data [32].

1.3 Paucity in the identification of multi-transmembrane domain (TMD)-containing membrane proteins

ORs and taste receptors are the only protein groups where zero progress was made since 2013. ORs continues to be the significant PE2-4 missing protein family, representing a significant 19% of all PE2-4 proteins (400 out 2,129) [3]. Noteworthy progress has been made in the characterization of keratin-associated, leucine-rich repeats, sperm and testes related proteins and zinc finger proteins [32]. Current neXtProt build (application release: v2.22.2) lists 4 out of 404 ORs as PE1s, solely based on non-MS evidence, i.e., biochemical, genomic or protein-protein interaction data [32].

1.3.1 Paucity in the Identification of ORs

Manuscript 1 Context: ORs play a significant role in olfaction, mediated by chemical interaction with a multitude of odorant ligands including volatiles, metals, nutrients, neurotransmitters, photons and protons [33–35]. ORs signal in many ectopic physiologies with escalating chemosensory roles, independent of nasal epithelial tissues with known expression in brain, colon, liver, breast, thyroid, testes, etc [34]. Context-dependent OR responses provide humans with cognizance on the spatiotemporal environment, far beyond the smell sensation [36,37].

Hydrophobic interface and limited frequency of tryptic sites on the TMDs limit their ability to generate peptides with ease when compared to non-TMD regions. Based on this evidence, we proposed that experimental peptide of a TMD containing proteins should be comparable to *in silico* peptide yield from the non-TMD regions. *In silico* analysis was performed to examine the ability of ORs to generate the peptides qualifying the current HPP MS PE1 assignment guidelines. Our analysis indicated that 58% of ORs could generate *in silico* peptides qualifying the current HPP PE1 guidelines from concatenated OR soluble domains, which increases up to 94 % upon relaxing the stringency requirements. Similarly, 98% of ORs were able to generate PE1 qualifying peptides upon complete OR sequence digestion.

Journal of proteome research

Cite This: J. Proteome Res. XXXX, XXX, XXX−XXX

Perspective

pubs.acs.org/jpr

In Silico Peptide Repertoire of Human Olfactory Receptor Proteomes on High-Stringency Mass Spectrometry

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

[†]Department of Biomedical Sciences, Faculty of Medicine & Health Sciences, Macquarie University, Sydney, New South Wales 2109, Australia

Supporting Information

ABSTRACT: Human olfactory receptors (ORs) are seven-pass transmembrane G-protein coupled receptors (GPCR) involved in smell perception and many other signaling pathways. They are primarily expressed in the olfactory epithelium and ectopically expressed in several other organs and tissues. neXtProt contains 4 PE1 (protein existence 1, evidenced at the protein level) ORs, determined on the basis of either protein interaction data (i.e., OR1D4 and OR2AG1) or convincing genetic, haplotype, or biochemical data (i.e., OR1D2 and 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 (iv) limited tissue availability. OR transmembrane domains (TMDs) inherently limit MS identification because the hydrophobic nature restricts the access of trypsin to potential cleavage sites. Equally, the extremely low frequency or lack of accessible arginine and lysine residues in TMDs renders trypsin cleavage ineffective. Here, we demonstrate an analytical approach specifically focused on the hydrophilic (trypsin-accessible) domains of ORs [i.e., with all transmembrane segments and anchored peptides excluded). We predicted the ability of OR soluble (hydrophilic) domains to yield 2 or more >9 amino acids (aa) length unique mapping (unique to a protein only), non-nested (peptides with varying length at the N or C terminal but containing the same core sequence), leucine/isoleucine (I/L) switch examined (I and L have same mass and cannot be distinguished by MS) tryptic peptides. Our analysis showed that ~58% of the human OR proteome could potentially generate tryptic peptides that satisfy current the Human Proteome Project data interpretation guidelines (version 2.1) when no missed cleavages are allowed and increases to ~78% when one missed cleavage is allowed. The utilization of current biological data (adjuvant genomics, expression profile, transcriptomics, epigenome silencing data, etc.) and the adoption of a non-conventional proteomics approach (e.g., Confetti multiprotease digestion, CNBr cleavage of TMDs, and more-extreme chromatographic and 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*

Downloaded via MACQUARIE UNIV on August 22, 2019 at 22:00:30 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

Human olfactory receptors (ORs) are a family of 404 (per the UniProt release from March 24th, 2019) G-protein coupled receptors (GPCR) signaling proteins containing 7 transmembrane domains (TMD).[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 signaling in response to a range of ligands, including protons, photons, low-molecular-weight (<30 kD) hormones, metals, nutrients, volatiles, and neurotransmitters.[5−7] Elucidating OR function is progressing, coincident with advances in structural and physiological analysis, signaling 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 the brain, breast, colon, liver, lung, testes, thyroid, etc., usually with fragments per kilobase per million mapped fragments (FKPM) values 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 endodomain loops between these TMDs, respectively. Each of these domains varies in length, sequence, and arginine/lysine (R/K) composition that makes them heterogeneously susceptible to tryptic digestion.[1] For example, while hydro-

Special Issue: Human Proteome Project 2019

Received: June 22, 2018
Published: May 3, 2019

ACS Publications © XXXX American Chemical Society

DOI: 10.1021/acs.jproteome.8b00494
J. Proteome Res. XXXX, XXX, XXX−XXX

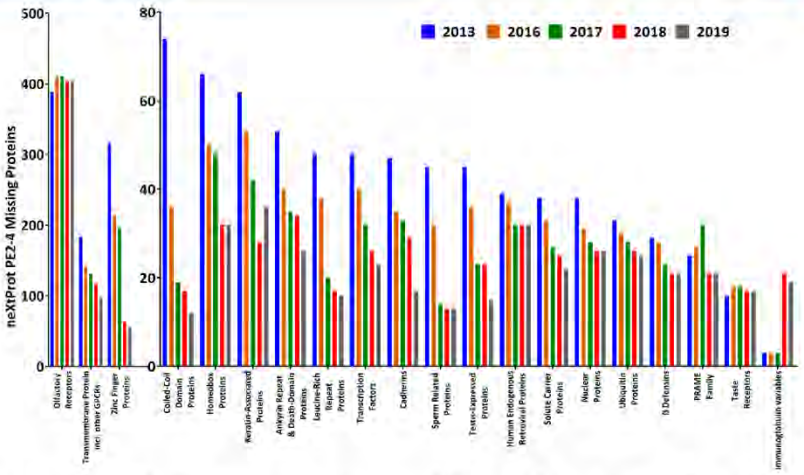


Figure 1. Most-abundant 20 neXtProt PE2−4 missing protein descriptors. neXtProt protein data sets were captured from neXtProt chromosomal download reports as previously described. PE2−4 proteins were sorted by neXtProt protein descriptions under the term "descriptions".

philic 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 or solubilized using more-extreme sample-preparation strategies[13] and the utilization of heated chromatography.[14]

In our previous study, we concatenated all 122 717 "stranded" OR peptide spectra (≥7 amino acids, aa, in length) available 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 false discovery rate (FDR), inclusion of shorter nonproteotypic observed peptides (≤8 aa) and numerous erroneous or ambiguous identifications.[18] Our study (see the Supporting Information) concluded that at very best, there was patchy or unconvincing mass spectrometry (MS) evidence for ∼6% (i.e., 23) of the 404 ORs.[15] We *definitively* concluded that no human OR currently met the high-stringency MS criteria set by the Human Proteome Project (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:

- they have no or low transcription;
- there are few OR-expressing cells;
- they have limited tissue or sample availability;
- they have restricted spatiotemporal- or differentiation-dependent expression;
- gene expression is inactivated in olfactory sensory neurons for all but a single OR; and
- there is a lack of availability or solubility of trypsin-accessible sites in many ORs.

Here, we have analyzed the ability of ORs to generate peptides from trypsin accessible domains exclusively. Because it is extremely unlikely that membrane protein TMDs contribute to MS data collations unless specifically enriched, we undertook in silico digestion of both the full-length ORs and concatenations of the exposed hydrophilic OR domains (free N- and C-termini strands plus 3 each of the ecto-domain and endodomain loops). We illustrate that ∼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 data set release for the years 2013 and 2014−2019 (ftp://ftp.nextprot.org/pub/current_release/). The chromosomal reports (1−22; X, Y, and MT) were downloade,d 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. The 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 and probable status. The cluster of uncharacterized proteins was not included in the figure owing to the lack of their functional annotation (Figure 1).

Retrieval of Current ORs from UniProt

Human Swiss-Prot entries with the term "Olfactory receptor" were retrieved from UniProt (March 24th, 2019 release). These entries were filtered to exclude (i) non-olfactory

DOI: 10.1021/acs.jproteomc.8b00494
J. Proteome Res. XXXX, XXX, XXX−XXX

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 the extracellular hydrophilic interface to the mid-TMD hydrophobic location was normalized. The ratio of R-to-K residue location from the 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, while values of 0.2–0.8 reflected positioning within hydrophobic TMDs (Figure 2).

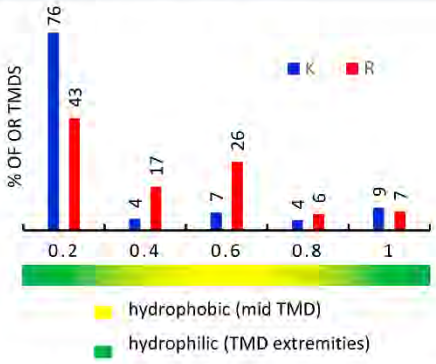


Figure 2. Relative location of R/K residues in TMDs of 278 human OR subset that contains R or K residues or both. We analyzed the relative location of OR TMD R and K residues found in the 278 of a total 404 human ORs that contain these residues. Among those 278 proteins, only 57 contain ≥2 R or K residues. The majority of K (~85%) and R (~50%) residues are found proximal to hydrophobic–hydrophilic TMD interfaces. Most R and 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.

Identification of ORs Capable of Generating Peptides Qualifying for PE1 Status

UniProt definitions were used for topological classification of all OR ecto- and endodomain loop and strand regions (March 24th, 2019 release). 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 version 2.21.0).[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 analyzed (Figure 3). 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 that 58% (235) of 404 ORs could generate tryptic peptides qualifying the current PE1 HPP data interpretation guidelines[19] with no missed cleavages allowed. The workflow was repeated, allowing one tryptic

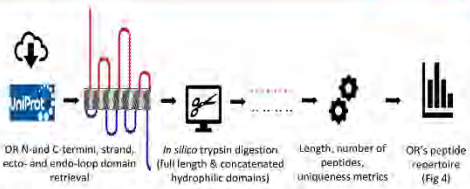


Figure 3. Summary of methods used for identifying ORs at different metrics stringency. UniProt-derived soluble OR domains were trypsin digested in silico with no missed cleavages allowed. These peptides were checked for uniqueness using neXtProt's peptide uniqueness checker. Uniquely mapping non-nested peptides from the digest were matched to respective ORs to count their ability to generate peptides at different stringencies.

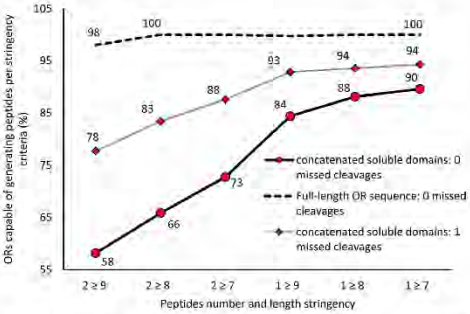


Figure 4. ORs capable of generating peptide as per the current HPP guidelines for missing proteins and lower stringency. Uniquely mapping non-nested tryptic peptides (only) produced by in silico digestion (non-missed-cleavage) of full-length OR sequences (inclusive of TMDs) and concatenated soluble domains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

missed cleavage for the remaining 42% ORs, leading to an inclusion of an 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 (see the Supporting Information) was prepared using the DT R package.[23]

RESULTS

Figure 1 contains past and current neXtProt PE data sets 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 out of 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 1 also demonstrates that membrane protein identification at high-stringency remains problematic, including across neXtProt

DOI: 10.1021/acs.jproteome.8b00494
J. Proteome Res. XXXX, XXX, XXX−XXX

protein groups covering the olfactory receptors, other transmembrane non-GPCR transmembrane proteins, non-OR GPCRs, taste receptors, and solute carrier proteins.

The only protein group in which *absolutely zero* progress has been made since 2013 are the ORs. Figure 1 also shows that the 2019 neXtProt release demonstrates that ORs continue to be the major PE2−4 missing protein family, representing a massive ~19% of all PE2−4 proteins (400 out 2129). 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 and accessibility of R/K residues in domains is crucial for determining susceptibility to cleavage. The limited MS-based detection of TMD peptides has been previously established, owing to hydrophobicity[25] and poor MS signal.[26] Unusual liquid chromatography (LC) and MS conditions are required to comprehensively analyze peptides generated from membrane protein TMDs.[27,28] We undertook an initial structural analysis of the OR proteome to determine where all OR TMD R/K residues resided.

Our UniProt-based analysis (Figure 2) 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 2 illustrates that occasionally R/K residues (278/404 ORs with 1 reside and 57/404 ORs with ≥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. Recognizing 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 not correlate with experimental peptide yield from membrane proteins containing multiple TMDs (e.g., 7 TMDs in all ORs and GPCRs) because peptide cleavage and release is restricted. This is because the soluble ecto- and endodomain loops between adjacent hydrophobic membrane-embedded TMDs need to be cleaved at a minimum of two locations before tryptic peptides can be released. When a single tryptic cleavage is present, both nascent ends of these loops remain completely anchored through the adjacent 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 tryptic peptides that can contribute to the PE1 assignment.

We argue that 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). Analysis of TMD-containing peptides warrants the more generalized adoption of nonconventional, extreme sample preparation, peptide cleavage (e.g., CNBr), and heated-column LC methodologies to allow the digestion and release of measurable peptides.[13,14]

To explain the paucity of MS-based OR identifications (0 of 404) in the 2019 neXtProt HPP, we sought to find 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 and (ii) trypsin-accessible soluble domains from the 404 ORs (i.e., N-termini, C- termini, and ecto- and endodomain loops) allowing no missed cleavages and (iii) OR-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 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. A total of ~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 the 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 the consideration of TMD-containing peptides, which are not easily detected. In contrast, ~58% of ORs trypsin-accessible domains (N-termini strands, C-termini strands, and ecto- and endodomain loops) qualify for the HPP PE1 guidelines, and the number extends to ~78% upon including missed cleavages. By default, this presents the fact that ~22% of human ORs may not meet PE1 requirements based on their soluble domains. Should the 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.

The 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 cleavages are 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 amino acids in length. In contrast, OR5MA generates three non-uniquely mapping non-nested tryptic peptides (yellow), one from the ecto-domain (MLSPNHTIVTEFILLGLTDDPVLEK), and two from the endo-domain (YVAICSPLHYSSR and DVILAIQQMIR). Additionally, OR5MA contains three "hanging" peptides (light blue) (NVTPNMLHNFLSEQK,

DOI: 10.1021/acs.jproteomc.8b00494
J. Proteome Res. XXXX, XXX, XXX−XXX

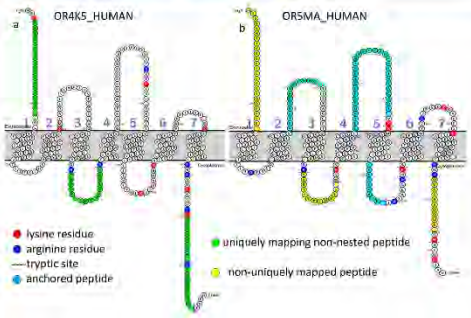


Figure 5. Topological distribution of exemplar OR (OR4K5 and OR5MA) N-termini strands, C-termini strands, intra-TMD loops, TMDs, R/Ks, residue number, and trypsin cleavage sites. ORs contain 7 TMDs with 3 each of ecto- and endodomains (loops) and 1 N-terminal and 1 C-terminal strands. The figure illustrates 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. However, Q6IEU7 generates 3 non-uniquely mapped peptides (yellow) and 3 strands that remain anchored to the plasma membrane (light blue).

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, determined solely based on tryptic peptides.

We identified 8 ORs that could not generate peptides as per PE1 assignment criteria upon whole-sequence digestion. This observation directed us to probe if any evidence of experimental uniquely mapping non-nested OR peptide exists. We accessed PeptideAtlas (http://www.peptideatlas.org/)[29] and neXtProt (https://www.nextprot.org/)[21] in search of publicly available experimental OR peptide evidence. The current PeptideAtlas build does not contain any evidence for experimental OR peptide spectra;[29] the HPP does not permit the use of synthetic peptides as evidence for PE1 assignment. The current neXtProt release has assigned a PE1 status to four ORs with non-MS-based evidence i.e., either with 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]

DISCUSSION

There has been no progress (i.e., zero identification) in MS-based ORs identification since the 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 guidelines. Apart from inherent technical challenges attributed by limited abundance,[33] ORs topology plays a significant role in the paucity of OR identification.[28] The restriction of trypsin proteolytic activity on membrane-embedded R/K residues, if any, and the requirement of a minimum of two cleavage sites within any domain loops result in low peptide yield. Peptides generated by a potential PE1 candidate OR requires LC−MS compatibility and "flyability" to be identified on the MS platform,[34] further limiting identification. We have provided interactive HTML files (supplementary file, requires file type conversion to HTML) containing none or one missed cleaved tryptic peptides generated from ORs endo- and ecto- domain with associated annotations in the Supporting Information. These files 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 cleavages 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 the years, the assignment of any ORs as PE1 proteins based solely on the current HPP guidelines seems far-fetched. These ORs require adjuvant genomics, expression profiles, 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.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00494.

A file describing the supplementary HTML files (PDF)
HTML files containing a list of theoretical none (file S1) or one missed cleaved (file S2) tryptic peptides from concatenated soluble domains (file can be accessed by common browsers upon changing the file extension to HTML) (TXT)
(TXT

AUTHOR INFORMATION

Corresponding Author

*E-mail: mark.baker@mq.edu.au.

ORCID

Subash Adhikari: 0000-0001-5945-7804
Seong Beom Ahn: 0000-0001-5907-3544
Mark S. Baker: 0000-0001-5858-4035

Author Contributions

M.S.B. conceived of the idea, S.A. performed in silico analysis, and S.S. contributed to the neXtProt missing proteins update. S.A., S.B.A., and M.S.B. prepared the manuscript. All authors read and approved the final manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Eric Deutsch and Lydie Lane for the confirmation of current OR peptides information from PeptideAtlas and neXtProt, respectively. S.A. and S.S. acknowledge iMQRES funding from Macquarie University, S.B.A. received funding

DOI: 10.1021/acs.jproteomc.8b00494
J. Proteome Res. XXXX, XXX, XXX−XXX

from the Cancer Institute NSW grant no. 15/ECF/1-38), and M.S.B. received funding from NHMRC project grant APP1010303.

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. *Nutr. Rev.* **2004**, *62*, S184–S188.
(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., Des. Sel.* **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), No. 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**, *45*, 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, Biomembr.* **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.; Alu, 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. https://github.com/sgibb/cleaver/ (accessed March 25, 2019).
(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*, 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 "DataTables". https://CRAN.R-project.org/package=DT (accessed March 25, 2019).
(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), No. 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*, D655–D658.
(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*, 3047–3056.
(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.

DOI: 10.1021/acs.jproteome.8b00494
J. Proteome Res. XXXX, XXX, XXX–XXX

(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 Bioinf.* **2007**, *8*, S23.

DOI: 10.1021/acs.jproteome.8b00494
J. Proteome Res. XXXX, XXX, XXX−XXX

1.3.2 Paucity in the identification of TMD containing membrane proteins

Manuscript 2 Context: Membrane proteins mediate a multitude of cellular processes. These proteins facilitate the transfer of biomolecule across the plasma membrane, initiate a cellular response in the extracellular matrix (ECM) through interaction with multiple ligands and provide physical separation of the intracellular components against the extracellular environments. Extracellular accessibility and their role in multiple human diseases make them particularly important in clinical settings, constituting over 45% of current pharmaceutical targets [38,39]. However, membrane hydrophobicity and low abundance present a technical challenge in the extraction of membrane proteins and recovery of corresponding peptides [40]. The requirement of more than one cleavage site on the domain loops (non-TMD) further limits soluble peptide yield [32].

Under-representation of the TMD regions in traditional bottom-up MS analysis highlights the hurdles in identification and quantification of TMD containing proteins [41]. Owing to the difficulty in the generation of soluble peptides from these TMDs and requirement of minimum two cleavage site on non-TMD domains, we proposed experimental peptide yield from membrane proteins in traditional MS analysis is comparable to the *in silico* peptide yield form the concatenated soluble domains.

Our analysis presented that 12040/15515 (77%) of human TMDs lack R/K sites, whereas 2812 (18%) TMDs contained only one R or K sites. This represents a maximum of 7 % human TMDs can theoretically generate tryptic peptides, explaining the under-representation of peptides from the TMDs in the traditional bottom-up approach.13409/18085 (75%) of corresponding soluble domains on the TMD containing proteins contained more than one R or K sites, making them capable of generating

soluble tryptic peptides. 3674/3878 (95%) TMD containing proteins can generate peptides from their soluble domains which can qualify current HPP PE1 guidelines.

The following work has been prepared as a manuscript; therefore, it contains its own citations.

Tryptic Peptidomic Landscape of All Human Multi-Transmembrane Domain Proteins

Subash Adhikari†, Seong Beom Ahn† and Mark S. Baker†*

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

Abstract

Membrane proteins represent a significant percentage of FDA approved clinical targets against common therapeutic interventions. Difficulty in the extraction of membrane proteins from the lipid bilayer results in the restricted mass spectrometry (MS)-based identification and characterization of these proteins. The phenomenon is more pronounced on the multi-transmembrane domain (TMD) containing membrane proteins. TMD regions of membrane proteins are under-represented in traditional bottom-up MS datasets, primarily due to the membrane hydrophobicity and inhibition of tryptic activity in such hydrophobic environment. Limited tryptic sites, i.e., arginine (R) and lysine (K) residues within the TMD domains, the requirement of ≥ 2 accessible proteolytic sites in ecto- or endo- domain loops for peptide release, limited sample availability and the requirement of enrichment, further lessen the likelihood of identification of these membrane proteins by MS.

In contrast, soluble domains (i.e., ecto-, endo-, N- and C- terminal strands) are analytically suited to generate tryptic peptides with ease compared to TMD regions, due to the ease in the solubilization and tryptic digestion. The TMD-containing proteome constitutes a considerable number of neXtProt (version 2.22.8) PE2-4 “missing proteins” (MPs).

16085/20967 (77%) of total human TMD are devoid of tryptic cleavage sites and 4008/20967 (19%) of TMD constitute a single tryptic proteolytic site. Hence these domains fail to generate soluble tryptic peptides due to the lack of sufficient proteolytic sites. Accounting for the lack of tryptic peptide generation from TMD regions, experimental peptidome yield is liable to be originated from the non-TMD regions. Based on this observation, a systematic analysis was performed on ≥ 2 TMD-containing membrane proteins to examine what percentage of membrane proteins can theoretically generate peptides qualifying the current Human Proteome Project (HPP) guidelines for MS-based identification of MPs. *In silico* digestion of concatenated soluble domains derived from ≥2 TMD-containing proteins was performed to explore tryptic peptide yield from these soluble domains at different MS stringencies. The analysis demonstrated that 2479/2814 (88%) of multi TMD containing proteins can generate ≥2 non-nested uniquely-mapping peptides of ≥9 amino acids (AAs) in length, qualifying the HPP guidelines for the identification of MPs by MS. Relaxation of stringency requirements to ≥1 non-nested uniquely-mapping peptides of ≥7 AA length increased the qualifying proteins to 2744/2814 (98% of the ≥2 TMD-containing human membrane proteome). Utilization of specialized membrane protein enrichment methods, sample extraction, multiple proteolytic fractionations and the use of MS techniques explicitly aimed at multi-TMDs may allow better proteome coverage and aid in the identification of these membrane proteins.

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

Introduction

The cell membrane is composed of membrane proteins, phospholipid bilayer and glycoproteins. The cellular membrane generates a physical separation between individual intracellular components of a cell, and between the cells and extracellular environment [1]. Membrane proteins are the constituents of membrane pores, channels, pumps and transporters, which allow the exchange of bioactive molecules between the cell and its extracellular environment. Membrane proteins also act as receptors (e.g., growth factor receptors, protease receptors, integrins) to a range of ligands that originate from the external environment, surrounding extracellular matrix and/or other cells, which enable the initiation of multiple signaling pathways within the cell [2, 3]. Due to the importance of their cellular functions and extracellular accessibility, it is not surprising that >70% of all FDA-approved targeted anti-cancer drugs currently in clinical trials target membrane proteins, according to the current My Cancer Genome (MCG) database [4].

Even though ~30% of all human genes code for membrane proteins, the membrane proteome has remained relatively poorly investigated [5]. This issue is even more acute in the multi-transmembrane domain(TMD) containing integral membrane proteins such as G-protein coupled receptors (GPCR) signalling proteins [6]. The potential role of TMD in restriction of MS-based identification is reflected by the fact that multi-TMD containing proteins form a substantial proportion of the Human Proteome Project (HPP) protein existence (PE) 2-4 missing proteins (MP) [7, 8], suggesting that very little evidence is available for their existence supported by the high MS stringency data. 668 multi-TMD containing proteins are classified as MPs on current neXtProt release [7]. Multi-TMD containing GPCRs; human olfactory receptors (ORs) and taste receptors (TRs) are two major protein groups lacking high stringency MS data to support their existence [9].

Regrettably, not a single OR qualifies for MS-based PE1 assignment guidelines set by the HPP [10]. Paucity in the detection of multi TMD-containing proteins is primarily due to the hydrophobic nature of membrane impeding solubilization and trypsin activity, compounded with limited tryptic digestion sites, the low copy number of genes/proteins and the lack of sample availability [11, 6].

Considering the inherent limitations of TMD-containing protein identifications by MS, we recently examined how many OR proteins can generate peptides from their soluble domains that would meet the high-stringency MS criteria set by the HPP data-interpretation guidelines v2.1 [6]. HPP guidelines require proteins to generate two or more experimentally derived uniquely-mapping non-nested peptides that are greater than or equal to nine amino acids (AA) in length. ORs are seven TMD containing proteins, by inference they contain one each of N- and C- terminal strands, and three each of ecto- and endo-domain loops between these 7 TMD [12]. The study performed *in-silico* tryptic digestion specifically on concatenated soluble domains of ORs (i.e., excluding all TMDs and "anchored" peptides that result from being constrained by one TMD after only a single tryptic cleavage within any loop domain). We were able to demonstrate that 58% of the human olfactory receptor proteome could potentially generate tryptic peptides from their concatenated domains satisfying the current HPP guidelines [6]. This suggests one clear physicochemical reason why they remain a refractory protein family overrepresented in all PE2-4 MPs.

In this study, we extended our OR analysis to all human TMD-containing proteins. We undertook *in silico* tryptic digestion of both the i) full-length TMD-containing integral membrane proteins and ii) concatenations of the exposed free N- and C-termini strands, ecto- and endo- domain loops. We then analysed how many (or percentage) membrane proteins could generate peptides exclusively from their soluble domains that would allow

them to qualify for the MS-based HPP PE1 assignment requirements. This comprehensive study of all TMD-containing proteins will provide useful insights for the development of more inclusive criteria for PE assignment, would allow researchers to more carefully target membrane proteins using more advanced MS techniques and finally reveal the potential for alternative methods of detection of MPs.

Methods

Identification of TMD containing proteins,

Domain topologies and their corresponding sequence of multi-TMD-containing proteins were obtained from UniProt SPARQL endpoint (https://sparql.uniprot.org/) (accessed 26th June 2019) [13]. Classification of these proteins into different PE status was performed based on neXtProt (v2.22.8) [14].

Distribution of R/K sites

The number of R/K sites within the TMD and hydrophobicity of topological domains were calculated to evaluate the likelihood of tryptic digestion. Grand Average of Hydropathy (GRAVY) based hydrophobicity was calculated with Sequence Manipulation Suite [15].

Generation of in silico peptidome of multi-TMD-containing membrane proteins

Multi-TMD containing proteins were tryptic digested *in silico* with protein digestion simulator (https://github.com/PNNL-Comp-Mass-Spec/Protein-Digestion-Simulator). The complete sequence and concatenated soluble domains were digested independently to examine the TMD induced peptidome differences. The uniqueness of peptides derived from these two sets of digestion was substantiated with a custom R-script querying neXtProt API (https://api.nextprot.org/). Standalone web-based peptide uniqueness checker tool (https://www.nextprot.org/tools/peptide-uniqueness-checker) is available that allows querying for up to 1000 peptides in a single batch [16]. The number

of proteins capable of generating peptides qualifying the HPP guidelines (two or more uniquely mapping non-nested peptides of nine or more AA length) and lower stringency (one uniquely mapping non-nested peptides of seven or more AA length) was calculated. “Anchored peptides” generated from the domain loops containing single tryptic cleavage site were discarded during the analysis.

Results

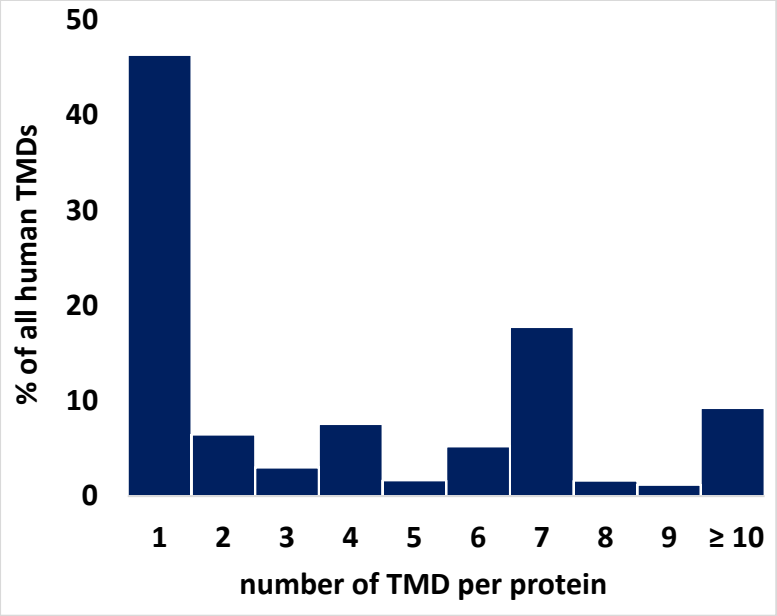


Figure 1: Number of TMD present in all human TMD containing proteins. Significant SP entries (2425/20431, 12%) contain a single TMD, followed by seven TMD containing GPCRs (930/20431, 5%). TMD containing proteins are the sizable portion of neXtProt MPs (880/2129, 41%). Presence of domains loops in multi-TMD domains containing proteins requires a minimum of two or more tryptic cleavage sites to release the peptides during an experimental MS analysis, which further decreases the likelihood of peptide generation from a multi-TMD domain-containing protein.

5239 (of total 20431) human Swiss-Prot (SP) entries contain annotation for TMDs, among which 2425 entries (12% of SP entries and 46% of TMD-containing proteins) contain single TMD and 2814 SP entries contain multi-TMDs. The second most prevalent domains were 7 TMDs-containing proteins (929 SP entries, 18% of TMD containing proteins) which includes 404 ORs, TRs and other GPCR family of proteins (figure 1). ORs and TRs are the largest protein groups which have failed to be identified by MS over the last decades. neXtProt has assigned PE1status for 4 ORs solely based-on non-MS evidence [14]. 4268/5239 (81%) TMD-containing membrane proteins have been classified as PE1, a substantial proportion of TMD-containing proteins are classified as PE2-4 proteins (800/2129, 38% of all PE2-4 proteins).

Additionally, 83/5239 TMD-containing proteins are classified as PE5 proteins (figure 2). GRAVY analysis of TMD-containing proteins reflects the hydrophobic nature of the TMDs compared to the non-TMDs domains, as shown in figure 3a. Hydrophobicity is known to inhibit the trypsin activities, leading to the suppression of soluble peptide yield during conventional MS experiments [6]. Non-conventional MS approaches are utilized to extract peptides from TMD regions to overcome the pH and solubility restriction [17, 18]. Sequence analysis of TMDs revealed that these sequences are conserved in human with a median and average AA length of 21 and 21.18, respectively (figure 3b). The TMD length is alike across multiple organisms and is thought to provide the evidence for a common origin of eukaryotic cell membranes. The complexity of signalling events through the membrane tends to define the thickness of the membrane, thereby modulating the TMD length [19].

We further analysed the likelihood of tryptic digestion of TMDs and soluble domains, solely based on the presence of tryptic cleavage sites. TMD region exhibited a diminished frequency of R and K residues compared to the soluble domains when

analysing 20967 TMDs and 26124 soluble domains from 5239 TMD containing proteins. 16085/20967 (77 %) TMDs lack any R or K residues, whereas 4008/20967 (19 %) TMDs contain a single R or K residue (figure 3c).

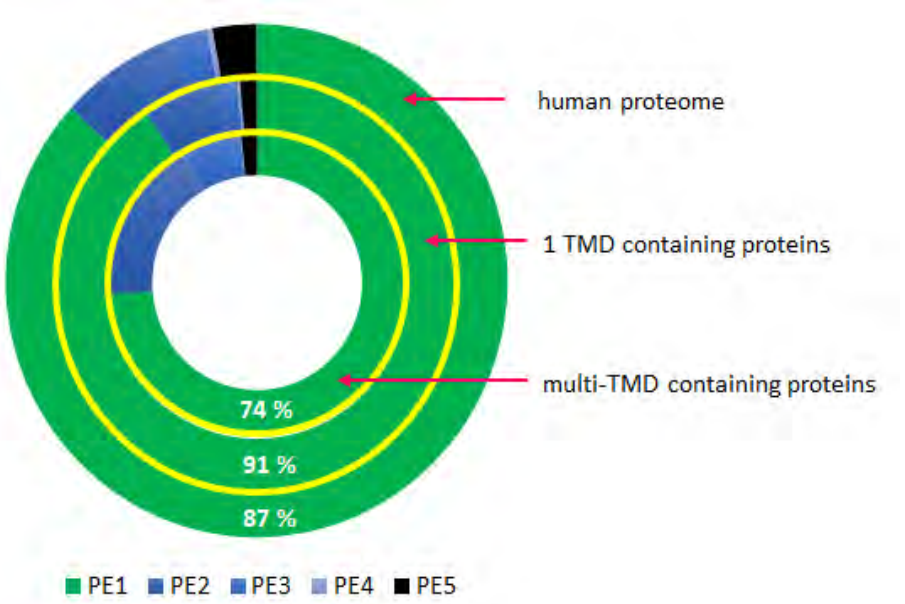


Figure 2: Comparison of PE status between complete human proteome and TMD containing proteins. 880/2119 (42 %) current missing proteins are TMD-containing proteins. Most (4268/5239, 82%) of the TMD-containing proteins are PE1. TMD domain-containing protein includes TRs and ORs, where no considerable progress has been made over the last decade. Multi-TMD containing proteins has the highest fraction of MPs (688/2812), followed by complete human proteome (2129/20399) and one TMD containing proteins (192/2429).

Considering the requirement of two cleavage sites to solubilize the peptides, most (96 %) TMDs are not capable of generating tryptic peptides. Limited tryptic activity on the remaining 874 (4%) TMDs in the hydrophobic interface further lessen the prospect for the generation of tryptic peptides. In comparison, 2395/26124 (9%) soluble domains were devoid of the R or K residues, whereas 3920/26124 (15%) soluble domains contained a single R or K residue. 19809/26124 (75%) soluble domains contained more than one tryptic cleavage sites. A total of 20355 (78%) soluble domains, including 546

N- and C- terminal domains with single R/K residues can theoretically generate tryptic peptides during conventional MS analysis (figure 3c). Location and frequency of R/K sites are the principal factors that define the tryptic peptidome repertoire. Release of soluble peptides from a domain loop requires the presence of two or more tryptic cleavage sites. Presence of a single cleavage sites leads the generation of “anchored peptides” that remain attached to the membrane. Similarly, the number of TMDs has a discernible effect on peptide yield upon tryptic digest. Human TMDs constitute a fraction of complete sequence (average/mean length of ~21 AA) on a single TMD-containing protein and the remaining sequence region can potentially generate the tryptic peptides during conventional MS analysis. Moreover, proteins with single TMD are not restricted to the requirement of multi cleavage sites in their domains, which is mandatory on soluble domains of multi-TMD containing proteins. Lack of insufficient R/K residues in most 20093/20967 (96%) TMDs for tryptic peptide generation lead us to analyse what percentage of multi-TMD containing proteins can generate tryptic peptides that qualify current HPP PE1 guidelines from their soluble domains.

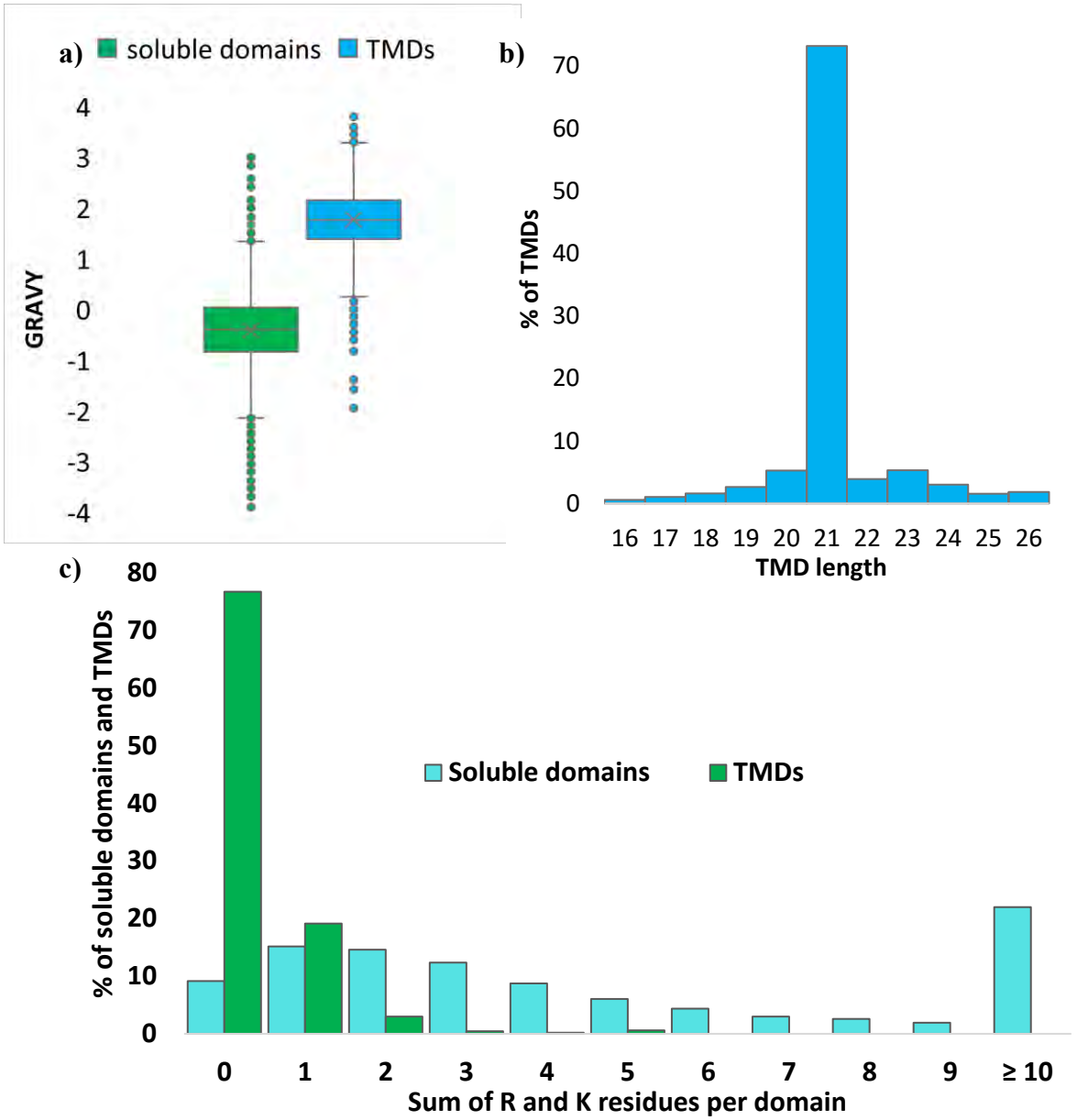


Figure 3: a) Hydrophobicity measurement of the concatenated soluble domains and TMDs. TMD regions are hydrophobic compared to the soluble domains based on hydrophobicity measurement by GRAVY index. TMDs are situated on the hydrophobic lipid environment which restricts tryptic digestion and subsequent peptide generation. b) TMDs length is conserved across species. Human TMDs were found to be of ~21 AA length. Conserved TMD length across species is taken as evidence for a common origin of eukaryote membranes. c) Distribution of R and K residues along the soluble domains and TMDs. Most 16085/20967 (77%) of the TMDs were devoid of R/K residues, whereas 4008/20967 (19%) TMDs contain single R/K residues. Sparse distribution of

R/K residues decreases the possibility of tryptic peptides generation. Buried R/K residues are known to modulate transmembrane helix orientation towards the aqueous interface, changes in ionization state of these residues regulate membrane function [20]. Soluble domains were found to contain a higher frequency of R/K residues. Our analysis indicated that 20355 (78%) soluble domains could theoretically generate tryptic peptides, based on the presence of more than one tryptic cleavage sites on the domain loops and a single tryptic cleavage sites on the N- and C- terminal strands.

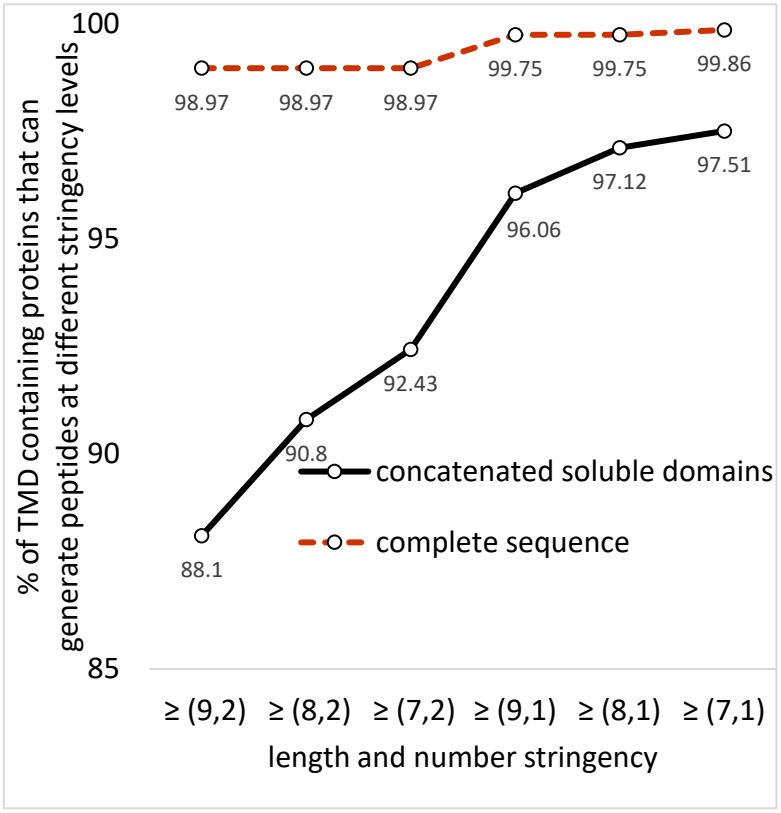


Figure 4: Multi-TMD containing proteins that can generate peptides qualifying the current MS-based PE1 guidelines set by the HPP. The ability of the multi-TMDs containing proteins to generate *in silico* uniquely mapping non-nested tryptic peptides from their soluble domains (excluding TMDs) and full-length protein sequence (including TMDs) at different MS stringencies was employed to obtain the peptidome differences at each stringency levels. 2479/2814 (88%) TMD-containing proteins could generate peptides qualifying HPP PE1 assignment guidelines.

In silico analysis of 2814 human proteins containing multi-TMDs provided the measurement of the proteins that can generate peptides qualifying HPP PE1 criteria for

MPs from their soluble domains. Our analysis indicated that 2479/2814 (88%) multi-TMD containing proteins could generate peptides qualifying the current HPP PE1 criteria. The gradual relaxation of the stringency requirements to a single uniquely mapping non-nested peptide of seven AA in length led to the identification of 2744/2814 (97.5%) proteins that can theoretically generate peptides at that stringency level. *In silico* digestion of complete sequence of multi-TMD containing proteins resulted in the identification of 2785/2814 (99%) multi-TMD-containing proteins that can generate peptides qualifying HPP PE1 guidelines, lowering the stringency requirement increases the number of proteins to 2810/2814.

Discussion

HPP has been successful in the identification and characterization of multiple proteins over the last decade. However, assignment of PE1 status on multiple proteins groups including membrane proteins remains problematic due to intrinsic technical challenge presented by i) sample limitations that require membrane enrichments, ii) trouble in proteolytic cleavage to solubilize the hydrophobic domains, iii) requirement of multiple cleavage sites on domains loops to release a peptide and iv) analysis of resulting hydrophobic peptides that do not fragment well on MS. Promotion of MS data sharing through MS data repositories has increased the availability of MS data. Communal reanalysis of these data, like those performed annually by PeptideAtlas [21], is bound to characterize additional MPs in days to come. We anticipate the release of updated HPP data interpretation guidelines from current v2.1, which might change the number of proteins that can generate peptide as per the PE1 assignment criteria. This analysis was performed based on UniProt annotations for TMDs and may differ from other TMD database or TMD predictions tools.

Supporting information

Supplementary file S1.1: *In silico* tryptic peptidome derived from the concatenated soluble domains of multi-TMD containing human proteins.

Author information

Corresponding Author

*E-mail: mark.baker@mq.edu.au.

ORCID

Subash Adhikari: 0000-0001-5945-7804

Seong Beom Ahn: 0000-0001-5907-3544

Mark S. Baker: 0000-0001-5858-4035

Author Contributions

M.S.B conceived the research and S.A performed all analyses with inputs from M.S.B. S.A, S.B.A and M.S.B prepared and edited this manuscript.

Acknowledgements

MSB and SBA thank NHMRC Project Grant APP1010303 and Cancer Institute NSW Grant #15/ECF/1-38 respectively. SA acknowledges the iMQRES scholarship provided by Macquarie University, Sydney, Australia.

References

(1) Vit, O.; Petrak, J. Integral Membrane Proteins in Proteomics. How to Break Open the Black Box? *Journal of Proteomics*. 2017.

(2) Gault, J.; Donlan, J. A. C.; Liko, I.; Hopper, J. T. S.; Gupta, K.; Housden, N. G.; Struwe, W. B.; Marty, M. T.; Mize, T.; Bechara, C.; et al. High-Resolution Mass Spectrometry of Small Molecules Bound to Membrane Proteins. *Nat. Methods* **2016**, *13* (4), 333–336.

(3) Bausch-Fluck, D.; Hofmann, A.; Bock, T.; Frei, A. P.; Cerciello, F.; Jacobs, A.; Moest, H.; Omasits, U.; Gundry, R. L.; Yoon, C.; et al. A Mass Spectrometric-Derived Cell Surface Protein Atlas. *PLoS One* **2015**, *10* (4), e0121314.

(4) Abramson, R. Overview of Targeted Therapies for Cancer. My Cancer Genomehttps://www.mycancergenome.org/content/molecular-medicine/overview-of-targeted-therapies-for-cancer/ (Updated May 25). https://www.mycancergenome.org/content/page/overview-of-targeted-therapies-for-cancer/ (accessed Jul 4, 2019).

(5) Tan, S.; Tan, H. T.; Chung, M. C. M. Membrane Proteins and Membrane Proteomics. *Proteomics* **2008**, *8* (19), 3924–3932.

(6) Adhikari, S.; Sharma, S.; Ahn, S. B.; Baker, M. S. In Silico Peptide Repertoire of Human Olfactory Receptor Proteome on High-Stringency Mass Spectrometry. *J. Proteome Res.* **2019**, acs.jproteome.8b00494.

(7) 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. Progress on Identifying and Characterizing the Human Proteome: 2018 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **2018**, *17* (12), 4031–4041.

(8) 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. Accelerating the Search for the Missing Proteins in the Human Proteome. *Nat. Commun.* **2017**, *8* (1), 14271.

(9) Hauser, A. S.; Attwood, M. M.; Rask-Andersen, M.; Schiöth, H. B.; Gloriam, D. E. Trends in GPCR Drug Discovery: New Agents, Targets and Indications. *Nat. Rev. Drug Discov.* **2017**, *16* (12), 829–842.

(10) Gaudet, P.; Michel, P.-A.; Zahn-Zabal, M.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Teixeira, D.; et al. The NeXtProt Knowledgebase on Human Proteins: Current Status. *Nucleic Acids Res.* **2015**, *43* (D1), D764–D770.

(11) Josic, D.; Clifton, J. G. Mammalian Plasma Membrane Proteomics. *Proteomics* **2007**, *7* (16), 3010–3029.

(12) Hilger, D.; Masureel, M.; Kobilka, B. K. Structure and Dynamics of GPCR Signaling Complexes. *Nat. Struct. Mol. Biol.* **2018**, *25* (1), 4–12.

(13) Bateman, A.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Alpi, E.; Antunes, R.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; et al. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45* (D1), D158–D169.

(14) Gaudet, P.; Michel, P.-A.; Zahn-Zabal, M.; Britan, A.; Cusin, I.; Domagalski, M.; Duek, P. D.; Gateau, A.; Gleizes, A.; Hinard, V.; et al. The NeXtProt Knowledgebase on Human Proteins: 2017 Update. *Nucleic Acids Res.* **2017**, *45* (D1), D177–D182.

(15) Stothard, P. The Sequence Manipulation Suite: JavaScript Programs for Analyzing and Formatting Protein and DNA Sequences. *Biotechniques* **2000**, *28* (6), 1102–1104.

(16) Schaeffer, M.; Gateau, A.; Teixeira, D.; Michel, P.-A.; Zahn-Zabal, M.; Lane, L.

The NeXtProt Peptide Uniqueness Checker: A Tool for the Proteomics Community. *Bioinformatics* **2017**, *33* (21), 3471–3472.

(17) 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), 3956–3964.

(18) Farias, S. E.; Kline, K. G.; Klepacki, J.; Wu, C. C. Quantitative Improvements in Peptide Recovery at Elevated Chromatographic Temperatures from Microcapillary Liquid Chromatography-Mass Spectrometry Analyses of Brain Using Selected Reaction Monitoring. *Anal. Chem.* **2010**, *82* (9), 3435–3440.

(19) Singh, S.; Mittal, A. Transmembrane Domain Lengths Serve as Signatures of Organismal Complexity and Viral Transport Mechanisms. *Sci. Rep.* **2016**, *6* (1), 22352.

(20) Gleason, N. J.; Vostrikov, V. V; Greathouse, D. V; Koeppe, R. E. Buried Lysine, but Not Arginine, Titrates and Alters Transmembrane Helix Tilt. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (5), 1692–1695.

(21) Deutsch, E. W. The PeptideAtlas Project; Humana Press, 2010; pp 285–296.

1.4 Conclusions

HPP has provided an international platform for coordinated proteomics research, leading to the identification and characterization of multiple MPs. Progressive assignment of PE1 status to various proteins was observed during the last decade [30]. neXtProt lists 2129 proteins as missing, representing 10% of all human proteins. Identification of some specific proteins groups has been insignificant, particularly membrane proteins. Limited expression levels, finite R/K sites, restricted tryptic activity on hydrophobic environment and requirement of more than one R/K sites restricts soluble peptides generation that is identifiable on a conventional MS approach [32]. Non-conventional methods such as; use of heated columns and adaptation of confetti approach have the potential to identify previously undetected peptide regions and aid in the subsequent characterization of MPs [42,43]. As per the HPP data interpretation guidelines, MS-based PE1 assignment requires the protein to generate two or more experimental uniquely mapping non-nested peptides that are nine or more aa in length. The upcoming update of HPP data interpretation guidelines (v 2.3, to be announced) will allow the use of nested peptides provided the combined length is 11 aa or more. However, not all proteins can generate MS identifiable peptides at the stringency criteria set by the HPP. HPP has a provision of “exclusions criteria” for such proteins and these proteins can be assigned PE1 status on a case by case basis with relaxed stringency requirements. Inclusion of antibody-based evidence and neXtProt GOLD binary interaction data have allowed assignment of 1096 (~5 %) proteins as PE1s, that do not have any MS evidence.

1.5 References

1. Hancock, W., Omenn, G., Legrain, P. & Paik, Y. K. Editorial: Proteomics, human proteome project, and chromosomes. *Journal of Proteome Research* **10**, 210 (2011).
2. Deutsch, E. W. *et al.* Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* **15**, 3961–3970 (2016).
3. Omenn, G. S. *et al.* Progress on Identifying and Characterizing the Human Proteome: 2018 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* **17**, 4031–4041 (2018).
4. Gaudet, P. *et al.* The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* **45**, D177–D182 (2017).
5. Green, E. D., Watson, J. D. & Collins, F. S. Human Genome Project: Twenty-five years of big biology. *Nature* **526**, 29–31 (2015).
6. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).
7. Doerr, A. Mass spectrometry–based targeted proteomics. *Nat. Methods* **10**, 23–23 (2013).
8. Wang, P. I. & Marcotte, E. M. It's the machine that matters: Predicting gene function and phenotype from protein networks. *J. Proteomics* **73**, 2277–2289 (2010).
9. Timms, J. F., Hale, O. J. & Cramer, R. Advances in mass spectrometry-based cancer research and analysis: from cancer proteomics to clinical diagnostics. *Expert Rev. Proteomics* **13**, 593–607 (2016).
10. Crutchfield, C. A., Thomas, S. N., Sokoll, L. J. & Chan, D. W. Advances in mass spectrometry-based clinical biomarker discovery. *Clin. Proteomics* **13**, 1 (2016).

11. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
12. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* (1989). doi:10.1126/science.2675315
13. Scigelova, M. & Makarov, A. Orbitrap Mass Analyzer – Overview and Applications in Proteomics. *Proteomics* **6**, 16–21 (2006).
14. Branca, R. M. M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* **11**, 59–62 (2014).
15. Paik, Y.-K., Omenn, G. S., Hancock, W. S., Lane, L. & Overall, C. M. Advances in the Chromosome-Centric Human Proteome Project: looking to the future. *Expert Rev. Proteomics* **14**, 1059–1071 (2017).
16. Aebersold, R. *et al.* The Biology/Disease-driven Human Proteome Project (B/D-HPP): Enabling Protein Research for the Life Sciences Community. *J. Proteome Res.* **12**, 23–27 (2013).
17. Deutsch, E. W. The PeptideAtlas Project. in 285–296 (Humana Press, 2010). doi:10.1007/978-1-60761-444-9_19
18. Whiteaker, J. R. *et al.* CPTAC Assay Portal: a repository of targeted proteomic assays. *Nat. Methods* **11**, 703–704 (2014).
19. Singer, D. S., Jacks, T. & Jaffee, E. A U.S. "Cancer Moonshot" to accelerate cancer research. *Science (80-.).* **353**, 1105–1106 (2016).
20. Legrain, P. *et al.* The Human Proteome Project: Current State and Future Direction. *Mol. Cell. Proteomics* **10**, M111.009993 (2011).
21. Gaudet, P. *et al.* The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Res.* **43**, D764–D770 (2015).

22. Bateman, A. *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).

23. Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* **45**, D1100–D1106 (2017).

24. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456 (2016).

25. Ma, J. *et al.* iProX: an integrated proteome resource. *Nucleic Acids Res.* **47**, D1211–D1217 (2019).

26. Sharma, V. *et al.* Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline. *Mol. Cell. Proteomics* **17**, 1239–1244 (2018).

27. Vizcaíno, J. A., Perkins, S., Jones, A. R. & Deutsch, E. W. Chapter 11. Data Formats of the Proteomics Standards Initiative. in 229–258 (2016). doi:10.1039/9781782626732-00229

28. Sivade, M. *et al.* Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. *BMC Bioinformatics* **19**, 134 (2018).

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

30. Omenn, G. S. *et al.* Progress on Identifying and Characterizing the Human Proteome: 2018-2019 Metrics from the HUPO Human Proteome Project. *J. Proteome Res.* acs.jproteome.9b00434 (2019). doi:10.1021/acs.jproteome.9b00434

31. Schaeffer, M. *et al.* The neXtProt peptide uniqueness checker: a tool for the proteomics community. *Bioinformatics* **33**, 3471–3472 (2017).

32. Adhikari, S., Sharma, S., Ahn, S. B. & Baker, M. S. In Silico Peptide Repertoire

of Human Olfactory Receptor Proteome on High-Stringency Mass Spectrometry. *J. Proteome Res.* acs.jproteome.8b00494 (2019). doi:10.1021/acs.jproteome.8b00494

33. Aisenberg, W. H. *et al.* Defining an olfactory receptor function in airway smooth muscle cells. *Sci. Rep.* (2016). doi:10.1038/srep38231
34. Flegel, C., Manteniotis, S., Osthold, S., Hatt, H. & Gisselmann, G. Expression Profile of Ectopic Olfactory Receptors Determined by Deep Sequencing. *PLoS One* (2013). doi:10.1371/journal.pone.0055368
35. Horowitz, L. F. *et al.* Olfactory receptor patterning in a higher primate. *J. Neurosci.* (2014). doi:10.1523/JNEUROSCI.1779-14.2014
36. Fleischer, J., Breer, H. & Strotmann, J. Mammalian olfactory receptors. *Front. Cell. Neurosci.* (2009). doi:10.3389/neuro.03.009.2009
37. Kang, N. & Koo, J. Olfactory receptors in non-chemosensory tissues. *BMB Reports* (2012). doi:10.5483/BMBRep.2012.45.11.232
38. Han, D. K., Eng, J., Zhou, H. & Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **19**, 946–951 (2001).
39. Rucevic, M., Hixson, D. & Josic, D. Mammalian plasma membrane proteins as potential biomarkers and drug targets. *Electrophoresis* **32**, 1549–1564 (2011).
40. Lu, B., McClatchy, D. B., Kim, J. Y. & Yates, J. R. Strategies for shotgun identification of integral membrane proteins by tandem mass spectrometry. *Proteomics* **8**, 3947–3955 (2008).
41. Kar, U. K., Simonian, M. & Whitelegge, J. P. Integral membrane proteins: bottom-up, top-down and structural proteomics. *Expert Rev. Proteomics* **14**, 715–723 (2017).

42. Blackler, A. R., Speers, A. E. & Wu, C. C. Chromatographic benefits of elevated temperature for the proteomic analysis of membrane proteins. *Proteomics* **8**, 3956–64 (2008).

43. Farias, S. E., Kline, K. G., Klepacki, J. & Wu, C. C. Quantitative improvements in peptide recovery at elevated chromatographic temperatures from microcapillary liquid chromatography-mass spectrometry analyses of brain using selected reaction monitoring. *Anal. Chem.* **82**, 3435–40 (2010).

Chapter 2: Mass Spectrometry based Proteomics

2.1 Overview

A proteome can be defined as the assembly of all proteins temporarily and spatially encoded by a genome under any given circumstances [1]. Protein synthesis is a highly regulated process in the human body, and in any given proteome the expression levels of proteins are tuned to match the demand of the signalling events that govern biological and functional characteristics in those cells and tissues. Alterations in protein expression levels and functions are associated with a range of human diseases [2]. The proteomics analysis aims at high-throughput identification, characterization and quantitation of all proteins in a proteome, including associated modifications, protein complexes and interactions between these proteins [3,4]. High-throughput proteome analysis allows one to decipher disease-related physiology, pathophysiology and effects of the treatment-induced signalling change [5]. In recent years, bottom-up shotgun mass spectrometry (MS) has evolved as the preferred proteome analysis tool and is often utilised to better understand disease processes, as well as being used in the development of diagnostics and prognostic markers [6]. Availability of cost-efficient genomic sequencing technologies and databases has further aided in the adoption of the technique as most of the proteomics applications rely heavily upon the pre-existence of a highly-curated genome sequence [7].

Despite current MS instruments reaching an impressive level of detection (LOD), down often to levels approaching one attomole of a non-complex sample [8], 2,129 out of 20,399 human proteins of the human proteome do not contain sufficient MS evidence.

These proteins in the human proteome are colloquially referred to as the “missing proteins” [9,10]. The outcome of a proteomics analysis is heavily dependent on efficiency and method used for sample preparation, adopted workflow, choice of MS instrument and what types of informatic stringency is applied to any data.

2.2 MS-based proteomics

Proteins modulate the expressed phenotype through dynamic crosstalk with a range of biomolecules. The proteome is characterized by substantial variations all of which influence the biology [11], including the;

i. intra- and extracellular protein abundance levels,
ii. cell type/s and subcellular compartment/s,
iii. temporal expression changes,
iv. stimuli-dependent expression signature, and
v. presence/abundance of various different protein isoforms (proteoforms), post-translational modifications (PTMs) and splice variants (SVs).

The presence of these variables makes precise analysis of any comprehensive proteome particularly challenging. MS has provided a platform to determine an unbiased map of abundance, interactions, sequence modifications, PTM, splice variants and subcellular localization of thousands of proteins in tandem [12].

In detail, a typical MS-based proteomics analysis involves;

i) protein extraction and preparation,
ii) chromatographic separation and MS analysis,
iii) identification and quantification, and
iv) exploration of the biological relevance of the MS data (see Fig. 2.1).

Availability of multiple types of mass analysers has made an investigation of human proteome feasible [13,14]. Two fundamental MS acquisition strategies exist - namely, discovery and targeted proteomics. Discovery-based proteomics aims to identify proteomics perturbations at the global level where peptide ion fragmentation is selected automatically, mainly based on signal intensities. On the other hand, targeted

proteomics deals with the identification of proteins of interest using a predefined peptide ion fragmentation model [15].

The MS workflow is further classified into top-down or bottom-up approaches. Bottom-up proteomics (also termed “shotgun proteomics”) involves proteolytic digestion of proteins by proteases (e.g. trypsin, lys-c, chymotrypsin) into peptides and their subsequent identification by LCMS analysis. The identification is based on matching the experimental spectra with the theoretical spectra, commonly referred to as a peptide-spectrum match (PSM). On the other hand, a top-down approach involves an analysis of intact proteins by MS, allowing enhanced precision regarding the identification of the primary structure of proteoforms and where data is obtained without the prior use of various proteolytic agents [16].

2.3 MS workflow

2.3.1. Sample processing

Choice of the proteomics sample preparation influences overall MS sensitivity and accuracy [17]. Sample preparation methods are tailored for optimal protein extraction and optimal processing from a range of biological samples. These involve a multi-step process starting with extraction and solubilization of proteins followed by denaturation, reduction, alkylation and finally proteolytic digestion [17]. Trypsin is most widely utilized for protein digestion in bottom-up approaches, due to its high specificity for C-terminal proteolytic cleavage next to arginines (R) and lysines (K) [18]. Introduction of the multi-protease digestion (i.e., colloquially termed *Confetti*), such as by a combination of Lys-c and trypsin digestion [19] has claimed to produce more comprehensive peptide coverage of a proteome when compared to single enzyme digestion [20]. Sample loss is

unavoidable during sample preparation, giving rise to different methods that allow completion of sample preparation in a single reactor cell, minimizing sample loss [17,21].

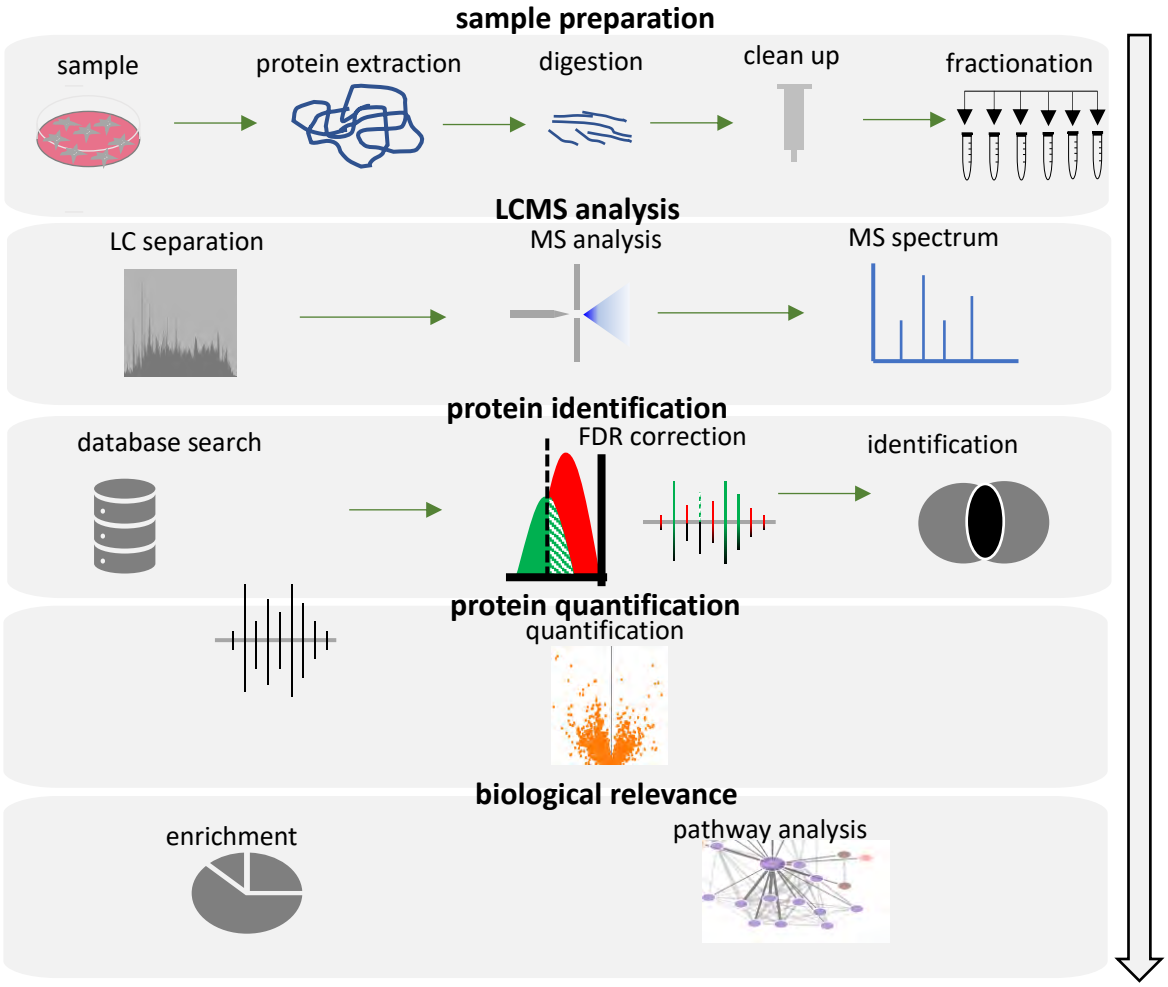


Figure 2.1: MS-based bottom-up proteomics analysis workflow. Proteins from a biological source are digested into peptides and introduced to LC-MS/MS analyses. Experimental peptide spectra from MS analysis are matched to their respective theoretical spectra for peptide spectrum matching (identification), which acts as a surrogate for protein inference and protein quantification. The differences observed in proteome number and individual protein abundances provide a measure of systemic proteome changes.

Many known protein biomarkers (e.g., CEA, PSA, cytokines, growth factors, interleukins) are present in low concentrations [22]. Analytical challenges introduced by an observed broad dynamic range in protein concentration can be addressed by numerous fractionation techniques, that allow the separation of protein or peptide subpopulations based on their physio-chemical characteristics (e.g., hydrophobicity, charge at a given pH, mass and globular shape). During traditional bottom-up MS approaches, highly abundant ion species mask low abundance ion species. Protein/peptide fractionation facilitates an increase in the proteome coverage by decreasing sample complexity. It is particularly crucial in samples with high dynamic range, such human plasma that has a concentration range of ≥12 log orders of magnitude [23]. Various fractionation methods such as high pH fractionation, isoelectric focussing, strong cation exchange are available, that aid to decrease sample complexity [24]. In addition, the presence of specific contaminants like detergents, salts and polymers lead to suppression in fragmentation efficiency and peak intensity during MS [25]. Sample clean-up procedures followed by the protein digestion allows removal of residual salts and detergents that may inhibit ionization efficiency during MS. Sample loss during sample preparation and the inability of MS instruments to characterize entire proteome results in the identification of high abundant proteins (Figure 2.2).

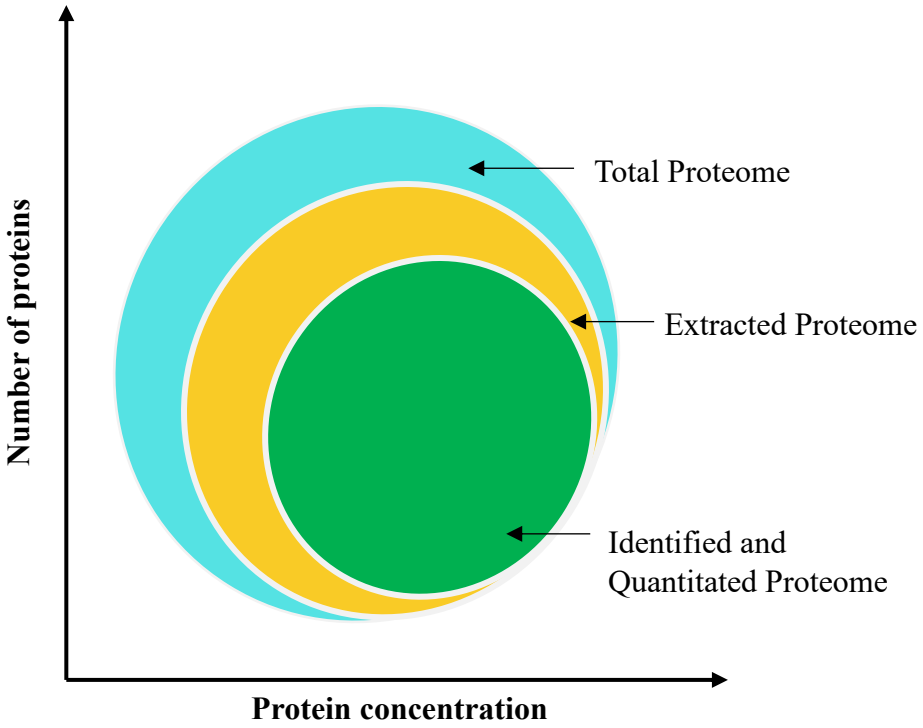


Figure 2.2: Representation of a quantifiable proteome fraction during MS analysis. Proteome is characterized by the large variation in protein concentration that spans across multiple orders of magnitude. Protein loss during sample preparation and the masking of low abundant proteins by higher abundant proteins are the major challenges during MS analysis, leading to the decrease in proteome coverage. These effects are further pronounced in MS analysis of samples with high dynamic range of concentration (e.g., in identification and quantification of low abundant proteins from plasma), during analysis of low abundant proteins (e.g., Olfactory receptors that has limited sample availability and expression levels) and analysis of low abundant protein modification (e.g; estimation of tyrosine modifications). Specific protein depletion, enrichment and fractionation of protein/peptide techniques are available to enhance the proteome coverage through reduction of sample complexity.

2.3.2 MS instrumentation

In proteomics, all MS instruments measure the mass-to-charge ratio (m/z) of a parent or fragmented “daughter” peptide species, which act as a surrogate for protein identification from complex protein mixtures in bottom-up approaches [14]. Peptide identity is derived from the mass and the fragmentation pattern, whereas parent ion intensity provides a basis for quantitation. MS instruments operate in two modes to obtain this spectrum-specific information. Mass and intensity profiles of co-eluting peptides are obtained during MS mode, whilst a fragmentation pattern is recorded upon fragmentation of the peptides along their peptide bonds during MS/MS mode [12]. The speed and sensitivity of modern MS instruments allow fragmentation and the quantitation of intensity for thousands of ion species within a fraction of a second [12].

Multiple fragmentation methods exist that are applicable to different MS workflows. Most common of these fragmentation techniques are;

i) collision-induced dissociation (CID),

ii) electron capture dissociation (ECD),

iii) electron transfer dissociation (ETD) and

iv) higher-energy collisional dissociation (HCD) [12].

The linear ion trap (LIT)-based CID technique fragments peptides upon collision with a low-pressure inert gas, generating a series of b- and y- ions. LIT increases the internal energy of peptides through excitation with an electric field, leading to peptide-bond CID cleavage. This CID fragmentation technique generates obscure spectra during neutral loss. Identification of such obscure spectra requires additional fragmentation modes on any MS workflow [26].

HCD fragmentation, on the other hand, is performed in a collision cell where ions are transported to a c-trap for high-resolution analysis in the Orbitrap™ [27]. Although analysis of HCD spectra in an Orbitrap™ instrument produces high-quality spectra, spectral acquisition times are increased due to the requirement of Fourier transform-based detection, compared to electron-multiplier-based detection in CID. Routine implementation of HCD-based fragmentation has been made possible with the development of efficient HCD collision cells with enhanced performance [28].

During ECD and ETD fragmentation, the electron transfer from a “radical anion to a protonated peptide” [29] increases the internal energy, thereby neutralizing one of the positive charges [30,31]. This induces fragmentation of peptide along a peptide-bond, results in c- and z- series ions. PTMs that are labile by CID are retained in ETD mode. Hence this fragmentation is preferred for PTM analysis, [29] using quadrupole detectors [32].

MS fragmentation can be achieved in a positive or negative ion mode. During positive ion mode, the analyte is fragmented at low pH to allow the formation of positive ions whereas, analysis is carried out at higher pH to deprotonate ion species during negative ion mode [33]. Peptide sequences can be deduced by concatenating the increasing size of fragments derived from the N-terminus (b-ions) towards C-terminus (y-ions), based on successive amino acid-specific mass differences of fragments [34].

A typical MS instrument contains an ion source, mass analyser and detector. Modern MS instruments are coupled to an additional online separation component that provides a continuous separation dimension for enhanced peptide separation efficiencies, such as an HPLC LC system [35] or ion mobility component [36]. Shotgun proteomics primarily utilizes linear ion trap (LIT)-Orbitrap™ and Quadrupole-Time of flight (TOF) configurations for

selection, fragmentation and detection of ion species [12]. On Quadrupole-TOF configurations, the Quadrupole mass filter allows transmission of entire ion species in MS mode or selected transmission of ion species around a precursor mass range in MS/MS mode. TOF analyses the fragmented ions generated in a collision cell. Quadrupole-TOF instrument achieves separation when peptides pass through the first Quadrupole or TOF, where only selected mass-range ion species maintain a stable trajectory to reach the detector, usually referred to as peptide separation “in space”. In contrast, ion trap instruments fragment peptides by application of an electrostatic field, where only selected peptide fragments within a certain mass-range maintain stability within a trap and this is referred to as peptide separation “in-time” [37]. In an Orbitrap™ mass analyser, a mass spectrum is generated from a frequency of oscillation of peptide ions around a “ *central spindle-shaped electrode*” using Fourier transformation [38,39]. The ion species axial oscillation frequency is directly proportional to the square root of m/z. High precision determination of the oscillation frequency leads to the accurate measurement of m/z in an Orbitrap™ instrument [40].

Some MS instruments house combinations of low-resolution LIT and high-resolution Orbitrap™. These are commercially available as LTQ-Orbitrap™ and have been widely adopted in proteomics applications. In MS mode, the LIT collects the peptide population and transfers to an intermediate c-trap that injects ions into Orbitrap™ for analysis at high-resolution. During MS/MS mode, LIT performs precursor isolation, fragmentation and mass scan at low resolution. LTQ-Orbitrap™ allows a combination of high-resolution MS and low-resolution MS/MS scan in tandem [41,42], as exemplified in the Thermo Orbitrap™ Velos and Thermo QExactive™ instruments. The label-based proteomics analysis in

Chapter 4 of this thesis was performed in such QExactive™ instrument. The recent introduction of tri-hybrid architecture, commercially available as the Thermo Orbitrap Fusion™ Tribrid™ with integrated Quadrupole-Orbitrap™-LIT analysers has enabled MS to reach the unprecedented resolution of 500,000 at 200 m/z. Also, the Thermo Orbitrap Fusion™ Tribrid™ instrument is available with an optional ETD source [43]. The architecture is flexible enough to perform complex MS acquisition in the presence of 3 mass analysers, where simultaneous peptide fragmentation in one mass analyser and detection at other two mass analysers at different accuracy and resolution is possible [44]. The tri-hybrid architecture is compatible with multiple fragmentation strategies, as described above, at any MS^n level [45].

2.3.4 Protein identification

The assembly of identified peptides into assumed proteins has been previously referred to as protein inference. This is a significant component of any MS-based proteomics workflow [46]. MS analysis is a data-intense application, usually generating thousands of MS/MS spectra. Manual interpretation of such a large number of spectra is no longer feasible and requires a dedicated computational pipeline for automated PSM assignment [47]. Computational algorithms for protein inference are required to solve a “many-to-many” relationship between peptide and proteins due to the possibility of the generation of same peptide sequence from different proteins (i.e., degenerate peptides) upon proteolytic digestion [48]. Peptides are the surrogate in the process of protein inference, but only a selected few peptides provide the capability of being able to distinguish corresponding proteins. Recently, for example, neXtProt has introduced the definition of “peptide

uniqueness" to allow the segregation of peptides that can definitively aid in high-stringency protein identification from peptide sequences [49]. The so-called neXtProt peptide uniqueness checker tool performs a series of analysis to interpret the uniqueness of any peptide [50]. Usually, peptides of longer amino acid (AA) length incorporate a higher probability of being specific and unique to any given protein. Previous studies have shown that peptides with seven or more AAs in length are more informative during protein inference determinations [48].

The Human Proteome Project (HPP) has defined in their HPP MS Data Interpretation Guidelines v2.1 for new PE1 identification of previous PE2-4 missing proteins that the protein identification should be based upon the identification of two or more uniquely-mapping, non-nested peptides of nine or more AAs in length with matching synthetic peptide fragmentation patterns. These are considered to be one of the highest stringency requirements that definitively confirm the presence/identification of a corresponding protein by MS [51].

Multiple database search algorithms exist that perform PSM assignment, adopting either probabilistic or heuristic approaches [52]. Mascot [53], SEQUEST [54], X!Tandem [55] and Andromeda [56] are the most widely used database search algorithms for PSM assignment. Protein inference with PSM alone is ambiguous due to the existence of degenerate peptides and what has been colloquially called 'one-hit-wonder' peptides [57]. Hence, validation of PSM assignment is mandatory during protein inference protocols [58]. Figure 2.3 contains a method for one of such peptide validation technique.

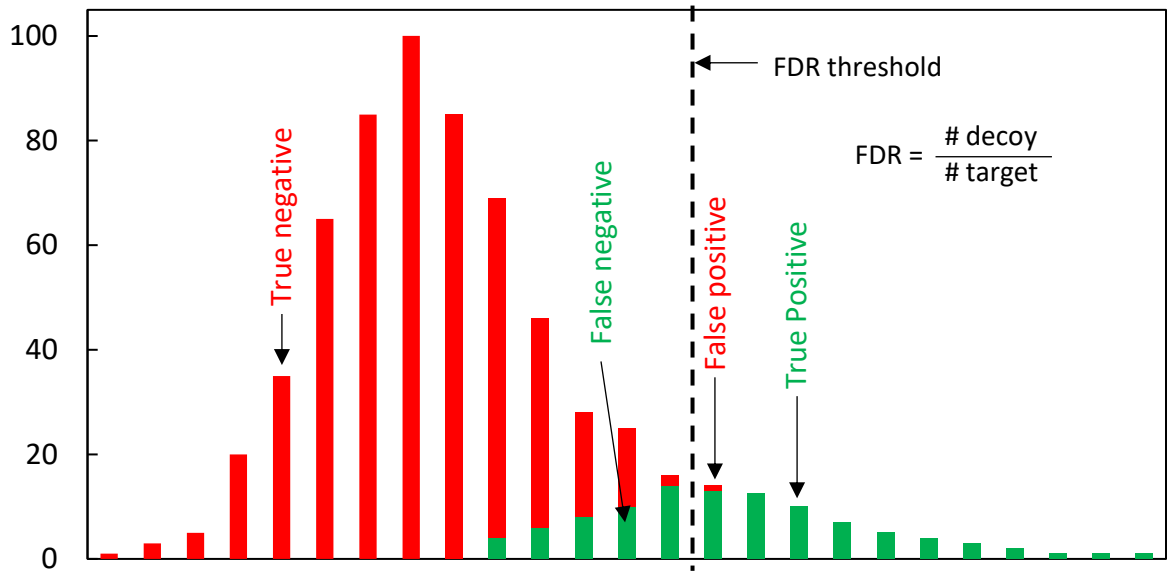


Figure 2.3: Estimation of the reliability of the peptide identification through the FDR approach. During a target-decoy approach, the ratio of the decoy match against a target match provides a measure for FDR. Decoy PSMs correlates to the incorrect PSM assignments.

Target-decoy approaches (TDA) are widely used in proteomics for FDR estimation [59]. These assume the likelihood of the occurrence of a random match is similar between the target database and corresponding decoy databases (i.e., reversed, shuffled or randomized). So, the decoy database match provides a measure of the expected random match to the target database. FDR estimation for peptides is derived from the number of targets that match the decoy and target databases respectively. Counts of target and decoy match above a certain threshold can be used to control FDR levels (figure 2.3) at both the protein or peptide level [52].

Heterogeneity of proteomics data presents a considerable challenge in confidently assigning a PSM from the fragment ion spectra obtained from an MS [60]. The complexity

of FDR assignment increases with any increase in the number of peptide spectra or the database size and multiplicative tendencies of FDRs should be taken into consideration when combining multiple datasets [60]. FDR estimation based on TDA primarily focuses on peptide level and cannot ensure quality-control at the protein level. MAYU presented the capability to quantify the definition of FDR on protein levels based on the assembly of PSM at certain FDR threshold [60]. In contrast to the TDA approach that applies a single FDR threshold to a dataset, p-value based confidence scoring reports the p-value for individual PSMs. The method fits the score distribution with a parametric model to define the p-value and utilize spectra properties to increase the accuracy of confidence scoring [58].

2.3.4 Protein quantification

MS-based protein quantitation is widely used to determine changes in protein abundance differences between diverse cellular states and/or after perturbation. Quantitation methods rely on the quantification of any MS feature that provides a measure of protein abundances, such as intensity, area-under-curve (AOC) and spectral count [61]. Quantitative proteomics can broadly be classified into;

i) label and label-free methods, and

ii) relative and absolute quantification [18].

Label-based quantification is achieved via incorporation of heavy isotopologues into a peptide or a protein through chemical or metabolic processes [62]. MS instruments can differentiate the isotopic variant induced by labelling, hence allowing the combination of multiple samples in any single MS run.

Metabolic labelling is achieved by supplementation of a heavy isotope during the growth phase of a cell or organism, thereby achieving selective labelling only on newly synthesized proteins [63]. Chemical labelling incorporates isotopomeric tags onto protein or peptides, thereby allowing multiplexing of different samples in a single MS run [3]. Chemical labelling with multiple isotope variants in a single experiment is routinely utilized in regular proteomics analysis for quantitative analysis. Examples of most widely adopted techniques include; iTRAQ [64], TMT [65], dimethyl labelling [66] and SILAC [67].

2.2.5.1 Manuscript 3 Context: Label-free quantification.

Label-free quantification adopts

i) comparative spectral counting that compares the number of acquired fragment ion spectra of a peptide from different experiments [68] and

ii) quantification based on chromatogram intensity or AOC measurement. Correlation between fragment ion spectra and the peptide concentration provides a measure of quantification during spectral counting based quantification [69].

Modified spectral counting approaches that take peptide properties such as peptide length and physicochemical properties of a peptide (e.g., normalized spectral abundance factor (NSAF) [70] and absolute protein expression (APEX) [71] have been shown to increase the accuracy of spectral counting-based quantification. Extracted ion chromatogram (XIC) as a function of monoisotopic mass and retention time provides a measure of AOC. AOC correlates linearly to protein abundance, providing a feasible quantification strategy.

MS data requires post-processing upon data acquisition to achieve accurate quantification to overcome run-to-run MS variations. Combinations of retention time alignment, feature detection, normalization of MS intensities and definition of intensity threshold and noise allow controls over the quality of quantitative analysis [72]. Integration of known concentration of a reference peptide during MS analysis allows quantification of endogenous peptides based on the principle of relative quantification, providing a basis for absolute quantification. Peptide level quantification can then be extrapolated to the protein level [73].

The manuscript is attached after the reference section of this chapter.

2.3.5 Interpretation of biological relevance of MS data.

MS analysis results in multivariate quantitative proteomics data that contains identification and quantification information for many thousands of proteins. Exploration of the biological data associated with these proteins requires a multi-step informatics approach [74]. In this regards, HUPO's Proteomics Standards Initiative (PSI) defines the MS data format required to enable cross-platform data-compatibility [75]. Multiple tools that assist in the process of biological data interpretation are available, which require the proteins IDs and data to be in a specific format. For example, UniProt accession is a widely accepted protein identifier format and UniProt ID conversion tools allow interconversion between protein identifiers [76].

Ontology enrichment is one of the preliminary analysis performed on the MS data. This evaluates the biological function, molecular process and subcellular localisation of any identified/quantified proteins. The Gene Ontology (GO) consortium provides different ontology identifier based on controlled vocabularies [77]. Ranking of these ontologies

provides an overview of the proteome complement identified from any MS analysis [78]. Ranking of identified proteins and their ontologies based on the likelihood of occurrence on a specific pathway and superimposition of identified proteins on predefined pathways or network provides a premise for pathway analysis. Pathway analysis provides a capability to map the signalling events and predict upstream and downstream regulators [79]. PANTHER [78] and DAVID [80] are well adopted GO analysis tools. STRING [81], IntAct [82] and BioGRID [83] provide protein-protein interaction (PPI) information, expansion of which generates the interactome map. The Proteomics Standard Initiative Common QUery InterfaCe (PSICQUIC) allows simultaneous PPI query across multiple database, through the web-based interface or query language [84]. Reactome is the largest open-source web-based pathway analysis tool [85], based on KEGG pathways [86]. Cytoscape allows visualization of the pathway maps and standalone protein-protein interactions [87]. UniProt maintains a collection of protein sequence and annotations sourcing from multiple databases, including PPI resources mentioned above and is the largest knowledgebases of its kind [76].

2.4 Conclusion

The dynamic range of individual protein content on a complex proteome range between multiple orders of magnitude, significantly higher than the dynamic range of identification achievable on modern MS instrument [88]. This has translated to the neXtProt confirmation of only 16,598 PE1 proteins of a total of 20,399 human genome coding proteins by MS approach, as per the current neXtProt v2.22.8 release statistics [9]. Increasing the proteome coverage requires minimization of biological and technical variabilities [88]. Studies have shown that intra-laboratory proteomics of the same sample produces

different results [89,90] suggesting a need for communal proteomics analysis standards. The outcome of any MS analysis should require validation with some orthogonal approach, particularly for results associated with low-resolution and modified peptides [34]. Reliability and reproducibility are paramount when using high-throughput technologies (like MS-based proteomics). The recent focus on MS data sharing through repositories should increase the transparency of MS analysis (see section 1 for MS data repositories). HUPO has been advocating for common data standards [75] and data interpretation guidelines [51]. PeptideAtlas performs an annual reanalysis of selected MS data with high MS data stringency [91].

2.5 References

1. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
2. Labbadia, J. & Morimoto, R. I. The Biology of Proteostasis in Aging and Disease. *Annu. Rev. Biochem.* **84**, 435–464 (2015).
3. Ong, S.-E. & Mann, M. Mass spectrometry–based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).
4. Gisbert, J. P. & Chaparro, M. Clinical Usefulness of Proteomics in Inflammatory Bowel Disease: A Comprehensive Review. *J. Crohn's Colitis* **13**, 374–384 (2019).
5. Schubert, O. T., Röst, H. L., Collins, B. C., Rosenberger, G. & Aebersold, R. Quantitative proteomics: challenges and opportunities in basic and applied research. *Nat. Protoc.* **12**, 1289–1294 (2017).
6. Hanash, S. Disease proteomics. *Nature* **422**, 226–232 (2003).

7. Shruthi, B. S., Vinodhkumar, P. & Selvamani. Proteomics: A new perspective for cancer. *Adv. Biomed. Res.* **5**, 67 (2016).

8. Huguet R, Blank M, Soltero N, Sharma S, Z. Low Attomole Limit of Quantification on an Orbitrap Fusion Lumos Tribrid Mass Spectrometer. in *ASMS* (2015).

9. Gaudet, P. *et al.* The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* **45**, D177–D182 (2017).

10. Baker, M. S. *et al.* Accelerating the search for the missing proteins in the human proteome. *Nat. Commun.* **8**, 14271 (2017).

11. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).

12. Choudhary, C. & Mann, M. Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* **11**, 427–439 (2010).

13. Paik, Y.-K., Omenn, G. S., Hancock, W. S., Lane, L. & Overall, C. M. Advances in the Chromosome-Centric Human Proteome Project: looking to the future. *Expert Rev. Proteomics* **14**, 1059–1071 (2017).

14. Matthiesen, R. & Bunkenborg, J. Introduction to Mass Spectrometry-Based Proteomics. in *Methods in molecular biology (Clifton, N.J.)* **1007**, 1–45 (Humana Press, Totowa, NJ, 2013).

15. Doerr, A. Mass spectrometry–based targeted proteomics. *Nat. Methods* **10**, 23–23 (2013).

16. Savaryn, J. P., Catherman, A. D., Thomas, P. M., Abecassis, M. M. & Kelleher, N. L. The emergence of top-down proteomics in clinical research. *Genome Med.* **5**, 53 (2013).

17. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).

18. Ong, S.-E., Foster, L. J. & Mann, M. Mass spectrometric-based approaches in quantitative proteomics. *Methods* **29**, 124–130 (2003).

19. Glatter, T. *et al.* Large-Scale Quantitative Assessment of Different In-Solution Protein Digestion Protocols Reveals Superior Cleavage Efficiency of Tandem Lys-C/Trypsin Proteolysis over Trypsin Digestion. *J. Proteome Res.* **11**, 5145–5156 (2012).

20. Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics. *J. Proteome Res.* **9**, 1323–1329 (2010).

21. Chen, W. *et al.* 3D-SISPROT: A simple and integrated spintip-based protein digestion and three-dimensional peptide fractionation technology for deep proteome profiling. *J. Chromatogr. A* **1498**, 207–214 (2017).

22. Bodzon-Kulakowska, A. *et al.* Methods for samples preparation in proteomic research. *J. Chromatogr. B* **849**, 1–31 (2007).

23. Percy, A. J. *et al.* Clinical translation of MS-based, quantitative plasma proteomics: status, challenges, requirements, and potential. *Expert Rev. Proteomics* **13**, 673–684 (2016).

24. Mostovenko, E. *et al.* Comparison of peptide and protein fractionation methods in proteomics. *EuPA Open Proteomics* **1**, 30–37 (2013).

25. Annesley, T. M. Ion suppression in mass spectrometry. *Clin. Chem.* **49**, 1041–4

(2003).

26. Melanie J. Schroeder, †, Jeffrey Shabanowitz, †, Jae C. Schwartz, §, Donald F. Hunt, †,‡ and & Joshua J. Coon*, †. A Neutral Loss Activation Method for Improved Phosphopeptide Sequence Analysis by Quadrupole Ion Trap Mass Spectrometry. (2004). doi:10.1021/AC0497104
27. Olsen, J. V *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**, 709–712 (2007).
28. Pekar Second, T. *et al.* Dual-Pressure Linear Ion Trap Mass Spectrometer Improving the Analysis of Complex Protein Mixtures. *Anal. Chem.* **81**, 7757–7765 (2009).
29. Mikesh, L. M. *et al.* The utility of ETD mass spectrometry in proteomic analysis. *Biochim. Biophys. Acta - Proteins Proteomics* **1764**, 1811–1822 (2006).
30. Zubarev, R. A. *et al.* Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal. Chem.* **72**, 563–73 (2000).
31. Syka, J. E. P. *et al.* Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J. Proteome Res.* **3**, 621–6
32. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci.* **101**, 9528–9533 (2004).
33. Yuan, M., Breitkopf, S. B., Yang, X. & Asara, J. M. A positive/negative ion–switching, targeted mass spectrometry–based metabolomics platform for bodily fluids, cells, and fresh and fixed tissue. *Nat. Protoc.* **7**, 872–881 (2012).

34. Walther, T. C. & Mann, M. Mass spectrometry-based proteomics in cell biology. *J. Cell Biol.* **190**, 491–500 (2010).

35. Snyder, L., Kirkland, J. & Dolan, J. *Introduction to modern liquid chromatography*. (2011).

36. Kanu, A. B., Dwivedi, P., Tam, M., Matz, L. & Hill, H. H. Ion mobility-mass spectrometry. *J. Mass Spectrom.* **43**, 1–22 (2008).

37. Michalski, A. *et al.* Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Mol. Cell. Proteomics* **10**, M111.011015 (2011).

38. Hardman, M., Makarov, A. A., and, M. H. & Makarov*, A. A. Interfacing the Orbitrap Mass Analyzer to an Electrospray Ion Source. *Anal. Chem.* **75**, 1699–1705 (2003).

39. Scigelova, M. & Makarov, A. Orbitrap Mass Analyzer – Overview and Applications in Proteomics. *Proteomics* **6**, 16–21 (2006).

40. Makarov*, A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. (2000). doi:10.1021/AC991131P

41. Alexander Makarov, * *et al.* Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer. (2006). doi:10.1021/AC0518811

42. Olsen, J. V. *et al.* A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed. *Mol. Cell. Proteomics* **8**, 2759–2769 (2009).

43. Williamson, J. C. *et al.* High-performance hybrid Orbitrap mass spectrometers for quantitative proteome analysis: Observations and implications. *Proteomics* (2016). doi:10.1002/pmic.201400545

44. Senko, M. W. *et al.* Novel Parallelized Quadrupole/Linear Ion Trap/Orbitrap Tribrid Mass Spectrometer Improving Proteome Coverage and Peptide Identification Rates. *Anal. Chem.* **85**, 11710–11714 (2013).

45. Eliuk, S. & Makarov, A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu. Rev. Anal. Chem.* **8**, 61–80 (2015).

46. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4**, 787–797 (2007).

47. Nesvizhskii, A. I. & Aebersold, R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov. Today* **9**, 173–181 (2004).

48. Sinitcyn, P., Rudolph, J. D. & Cox, J. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annu. Rev. Biomed. Data Sci.* **1**, 207–234 (2018).

49. Gaudet, P. *et al.* The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Res.* **43**, D764–D770 (2015).

50. Schaeffer, M. *et al.* The neXtProt peptide uniqueness checker: a tool for the proteomics community. *Bioinformatics* **33**, 3471–3472 (2017).

51. Deutsch, E. W. *et al.* Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* **15**, 3961–3970 (2016).

52. Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Mol. Cell. Proteomics* **14**, 2394–2404 (2015).

53. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-based

protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).

54. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).

55. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).

56. Cox, J. *et al.* Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

57. Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–40 (2005).

58. Huang, T., Wang, J., Yu, W. & He, Z. Protein inference: a review. *Brief. Bioinform.* **13**, 586–614 (2012).

59. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).

60. Reiter, L. *et al.* Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. *Mol. Cell. Proteomics* **8**, 2405–2417 (2009).

61. Griffin, N. M. *et al.* Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.* **28**, 83–89 (2010).

62. Rauniyar, N. & Yates, J. R. Isobaric Labeling-Based Relative Quantification in Shotgun Proteomics. *J. Proteome Res.* **13**, 5293–5309 (2014).

63. Heck, A. J. & Krijgsveld, J. Mass spectrometry-based quantitative proteomics. *Expert Rev. Proteomics* **1**, 317–326 (2004).

64. Wiese, S., Reidegeld, K. A., Meyer, H. E. & Warscheid, B. Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics* **7**, 340–350 (2007).

65. Andrew Thompson, † *et al.* Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. (2003). doi:10.1021/AC0262560

66. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* **4**, 484–494 (2009).

67. Ong, S.-E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–86 (2002).

68. Hongbin Liu, †,§,‖, Rovshan G. Sadygov, †,§ and & John R. Yates, I. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. (2004). doi:10.1021/AC0498563

69. Geromanos, S. J. *et al.* The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics* **9**, 1683–1695 (2009).

70. Florens, L. *et al.* Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40**, 303–311 (2006).

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

72. Megger, D. A., Bracht, T., Meyer, H. E. & Sitek, B. Label-free quantification in clinical proteomics. *Biochim. Biophys. Acta - Proteins Proteomics* **1834**, 1581–1590 (2013).

73. Brönstrup, M. Absolute quantification strategies in proteomics based on mass spectrometry. *Expert Rev. Proteomics* **1**, 503–512 (2004).

74. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).

75. Vizcaíno, J. A., Perkins, S., Jones, A. R. & Deutsch, E. W. Chapter 11. Data Formats of the Proteomics Standards Initiative. in 229–258 (2016). doi:10.1039/9781782626732-00229

76. Bateman, A. *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).

77. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

78. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).

79. García-Campos, M. A., Espinal-Enríquez, J. & Hernández-Lemus, E. Pathway Analysis: State of the Art. *Front. Physiol.* **6**, 383 (2015).

80. Huang, D. *et al.* The DAVID Gene Functional Classification Tool: a novel

biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8**, R183 (2007).

81. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
82. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* **40**, D841–D846 (2012).
83. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529–D541 (2019).
84. del-Toro, N. *et al.* A new reference implementation of the PSICQUIC web service. *Nucleic Acids Res.* **41**, W601–W606 (2013).
85. Hermjakob, H. Reactome - Pathway Context and Visualisation for Omics Data. *Biophys. J.* **116**, 329a (2019).
86. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).
87. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
88. Nilsson, T. *et al.* Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* **7**, 681–685 (2010).
89. Addona, T. A. *et al.* Multi-site assessment of the precision and reproducibility of multiple reaction monitoring–based measurements of proteins in plasma. *Nat. Biotechnol.* **27**, 633–641 (2009).

90. Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8**, 291 (2017).

91. Deutsch, E. W. The PeptideAtlas Project. in 285–296 (Humana Press, 2010). doi:10.1007/978-1-60761-444-9_19

Author's personal copy

Quantification of Proteins From Proteomic Analysis

Zainab Noor, Subash Adhikari, Shoba Ranganathan, and Abidali Mohamedali, Macquarie University, Sydney, NSW, Australia

© 2019 Elsevier Inc. All rights reserved.

Protein Quantification – Introduction

All living biological entities rely on an intricate balance of metabolic activity, bio-molecules and non-organic molecules to maintain life. Inherent to this balance is the ability to respond rapidly to changes from within or without. Although the breadth of this response can be vast from activating enzymes to opening channels to allow ions to leave or enter, its temporal nature can be varied leading to permanent changes due to chronic signals or transient changes due to acute signals and every variation in between. One of the most fundamental changes that occur in cells is the variation in the abundance of proteins (Vogel and Marcotte, 2012) in response to external or internal stimuli. Indeed, it is well understood that although transient and rapid alterations in cells is often due to modification of the activity of proteins, sustained stimuli lead to sizeable changes in the proteome (the protein complement) of an organism due to changes in transcriptional and translational activity; therefore, protein abundance. This fact has been the source of understanding how biological systems respond to change which has led to the development of markers of change that have revolutionized medicine. An illustration is the discovery of Early pregnancy factor (EPF) as an indicator of early pregnancy (Fan and Zheng, 1997) has revolutionized early maternity practice.

Although it has been known that multiple proteins change expression in relation to stimuli (or treatment), most studies were able to measure and quantitate only a small number of proteins at a time using immunological methodologies (such as, ELISA and Western blotting). Upon the advent of mass spectrometry, researchers began to investigate how this new method could not only identify but also quantitate multiple proteins. The rapid expansion of the technologies and the development of novel label-based quantification and methodologies of label-free quantification using spectral counting led to significant advances in the field. By early 2007, it was possible to quantitate thousands of proteins in a single experiment. These methodologies had significant drawbacks especially around experiment design, technical variability, reproducibility and reliability (Bantscheff *et al.*, 2007). A more recent and powerful technique, loosely based on targeted quantification (multiple reaction monitoring), using a Data Independent Acquisition (DIA) mode on a mass spectrometer has made quantification far more reliable and robust. **Fig. 1** summarises the most common methodologies used to quantitate proteins. This book article will discuss in brief detail some of the most common proteomic methodologies of protein quantification, with a particular emphasis on DIA quantification. Using an example, the article will guide a reader through setting up their own data analysis workflow and to apply statistical and experimental validation techniques to accurately quantify thousands of proteins from a biological sample.

Label-Based Quantification

The basic principle of label-based protein quantification technique is labelling or tagging of peptides/proteins with stable heavy isotopes (13C, 15N, 18O and 2H) and comparing the unlabelled 'light' form of the sample with labelled 'heavy' variant in mass spectrometry. In a given mass spectrum, a mass shift of 3–4 Da occurs in a heavy isoform where the ratio of peak intensities of both isoforms depict the ratio of abundances (Lindemann *et al.*, 2017). This method has been employed in studying modified proteins, membrane proteins, quantifying disturbances or perturbations in a system due to changes in proteins and effects of drugs or inhibitory compounds on protein expression. Label-based quantification is limited by the number of labels that can be applied at the same time (multiplexing) such that up to 12 samples can currently be studied simultaneously (King *et al.*, 2017).

Label-Free Quantification

Label free quantitative proteomics is an extensively used semi-quantitative approach capable of multiplexing multiple MS runs into a single experiment. This approach has wide application in high throughput applications with large individual variation and large sample size, such as in clinical proteomics. Label free quantification in such case can yield more analytical depth and higher dynamic range for quantification (Distler *et al.*, 2016).

Label free quantification is based on comparison of precursor ion intensities between MS runs, measured in terms of peak area or peak intensity after feature alignment according to retention time, m/z, charge states etc. A reproducible LC-MS system is an absolute requirement for efficient feature alignment (Norbeck *et al.*, 2005). Label free quantification requires resource intensive computational capability for feature detection and alignment (Mueller *et al.*, 2008), but unlike label-based quantification, the only limit to number of samples that can be analyzed is computational.

Regardless of platform and method adopted, label free quantification relies on extracted ion chromatogram (XIC) intensity-based quantitation or spectral counting-based quantitation. It has been observed that ion concentration correlates with ESI signal intensity; that is, higher peptide concentration in sample yields higher area under integrated chromatogram of a MS spectra. Relative area under curve/chromatogram (peak area) or peak intensity between samples can be used for relative quantitation of peptides/proteins across samples (Neilson *et al.*, 2011).

Author's personal copy

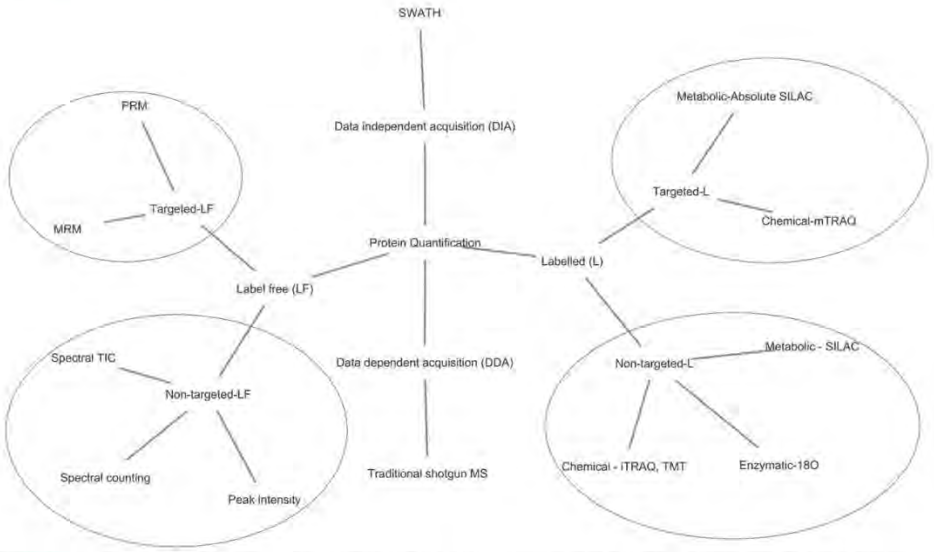


Fig. 1 Multiple approaches exist for MS based protein quantification. Labelling quantifies protein on basis of difference in signal intensities between heavy and light labelled peptides from different samples combined into a single run. Label free quantification is performed by measuring relative signal intensities between multiple sample runs. Quantifications are accomplished on DDA and DIA mode, depending on targeted and comprehensive data analysis.

Spectral counting-based quantification relies on assumption that number of tandem mass spectra are proportional to abundance of a protein, that more spectra are observed with increasing peptide concentration in the sample (Washburn *et al.*, 2001). Normalization factors relating to length and physiological properties of a peptide and its ability to be fragmented are taken into consideration (Ahrne *et al.*, 2013). One challenge with combining XIC-based quantitation with DDA-based identification is finding a balance between MS (quantitation) and MS^2 (identification) cycling. Devoting more scanning time to peptide identification decreases the number of data points available for MS quantitation, which limits the resolution, and hence accuracy of XIC-based quantitation.

Relative Quantification

Relative quantification is the measure of change in protein expression patterns as a function of biological function, expressed in terms of ratios between different biological states. Relative quantification is measured by either of labelled-based or label-free quantification, depending upon scope of the experiment. A number of labelled-based relative quantification strategies have been developed which are characterized by their mode of action. They are broadly classified into three categories (i) metabolic labelling, biological incorporation of isotopes or labels in a cell culture; (ii) chemical labelling, integration of heavy isotopes into the proteins using isotopic reagents or isobaric tags and (iii) enzymatic labelling, labelling with heavy isotope of oxygen during hydrolysis of protein through trypsin or other proteases. Similarly, various label-free approach could be adopted for relative quantification.

Absolute Quantification

Absolute quantification is the process of measuring the exact amount of a protein in a proteome. Proteins of interest (sample proteins) are spiked with known concentrations of labelled standard peptides bearing identical sequence and physicochemical properties as sample peptides. Quantification is performed by comparing the intensities of the sample and standard peptides/proteins, where unknown quantities of sample proteins can be deduced by means of known concentrations of standard proteins. Several peptide or protein-based approaches have been designed for absolute quantification including absolute SILAC, protein standard absolute quantification (PSAQ), full-length expressed stable isotope-labelled quantification (FLEXIQuant) and quantification Concatemers (QconCAT). Similarly, in the label-free context, MRM assays use the same approach described to accurately quantify selected peptide transitions and hence infer absolute protein abundance.

Author's personal copy

Quantification Through Data Independent Acquisition (DIA)

During the last decade, most of the MS-based proteomics studies were based on a discovery-based shotgun proteomics method, run in data-dependent acquisition (DDA) mode. Although, this helps to maximize amount of information obtained from an experiment, it has certain inherent limitations such as scan speed (Wenner and Lynn, 2004), selection of ion for further fragmentation, poor reproducibility, narrow dynamic range etc. (Law and Lim, 2013). Due to biases towards certain most intense fragments, DDA method has limitation of not being able to detect low abundant ion fragments. Targeted MS-based proteomics was developed to overcome these limitations of DDA mode. Targeted MS approach on Data independent mode (DIA) is dependent on a known reference spectrum, where identification could be performed from selected few fragment ion, rather than all possible fragment in DDA mode (Gillette and Carr, 2013). Success of data independent quantification relies on proper selection of distinct precursor ion-transitions (Tang *et al.*, 2014). Hence, DDA mode is employed to filter top performing precursor-ions as inclusion list for DIA mode, so the MS instrument could select those specific fragments to aid in identification of low abundant ions/spectra.

Sequential Window Acquisition of all THeoretical fragment ion spectra coupled to Tandem Mass Spectrometry (SWATH-MS) is a novel identification/quantification methodology that operates in DIA mode in a label-free context. SWATH-MS, currently takes into account the peptides of mass/charge (m/z) range from 400 to 1200 and generates fragment ion spectra of all the precursor peptides falling within the window width of 25 Da in each cycle. A prerequisite assay library of peptides/proteins of interest, created in shotgun or discovery mode is required to identify and translate the detailed fragment ion spectral data which is produced in DIA mode and extracted through targeted data extraction approach (Vidova and Spacil, 2017). While performing the data analysis, a number of parameters are usually taken into consideration to perform efficient and high-confidence data analysis. These parameters are, precursor and product ion mass/charge values, retention times, generated spectra and relative intensities. Scores are calculated against peak groups and statistical validation is performed (see **Fig. 2**).

In this article, label-free relative quantification analysis using computational tools have been discussed and demonstrated on DIA data. Quantification data setup and pre-requisite steps to extract the useful information and measurements from experimental data have been illustrated in detail. Moreover, statistical analysis has been carried out on quantified experimental data to analyze proteome expression.

Data Analysis Software

The data independent acquisition strategy has led to the development of a number of specialized software and packages which implement high-speed accurate DIA analysis, from chromatogram extraction to quantification. Currently, in addition to the generalized data analysis packages, mass spectrometry vendors also encompass built-in software responsible for not only

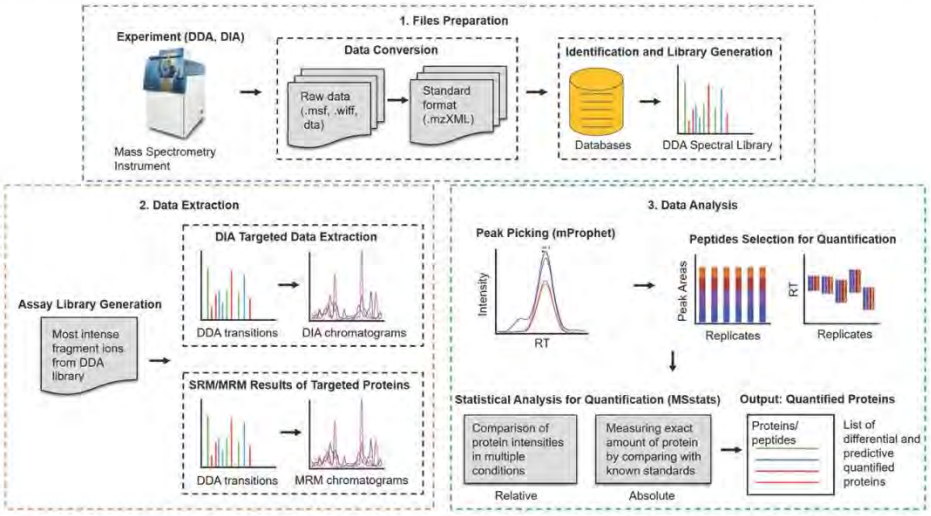


Fig. 2 Workflow of mass spectrometry based protein quantification. Data generated either in DIA or DDA (for the library) mode from a TOF or Orbitrap instrument are first converted to suitable standards followed by library generation from DDA data. Peaks are then extracted and picked for quantification. Finally, statistical analysis is performed to identify a list of differentially quantified proteins.

Author's personal copy

Table 1 Commonly used DIA (Data Independent Acquisition) data analysis software packages

Software	*Enterprise*	*Specifications*	*Versions*	*Download link*
PeakView	AB Sciex	• Commercial • Windows-based • Bio Tool Kit	V 1.1 (2010–2011) V 2.2 (2014)	https://sciex.com/products/software/peakview-software
OpenSWATH (Rost *et al.*, 2014)	OpenMS	• Free – Open Source • Windows, Linux, Mac-based • Compatible with AB Sciex, Thermo and Water DIA data	V 2.2.0 (2017) V 2.1.0 (2016)	https://github.com/OpenMS/OpenMS/releases
Skyline (Pino *et al.*, 2017)	MacCoss lab	• Free • Windows-based • Panorama repository	12 versions V 3.3 (2017)	https://skyline.ms/project/home/software/Skyline/begin.view
Spectronaut (Bruderer *et al.*, 2017)	Biognosys	• Commercial • Windows-based • Quantification of 1000 proteins from a single run	5 versions V 10.0 – Orion (latest)	https://biognosys.com/shop/spectronaut
DIA-Umpire (Tsou *et al.*, 2015)	Alexey Nesvizhskii lab	• Free – Open Source • Untargeted, library-free peptide identification from DIA data	V 1.4 (2015) V 2.0 (2016) V 2.1 (2016)	http://diaumpire.sourceforge.net/

monitoring the MS experiment but also performing all the downstream data analysis. Some of the widely used software for DIA data analysis are presented in **Table 1**.

Sample Dataset

The sample data was collected for the detection and quantification studies of yeast (*Saccharomyces cerevisiae*) proteome (Selevsek *et al.*, 2015). In this study, changes in yeast proteome expression have been analyzed in the presence of osmotic stress, progressively at different times. As stated above, DIA data analysis requires initial DDA runs to generate the library of assays. It is preferred to run the shotgun injections on the same instrument type as the final DIA data, however, the library assays can be transferred to different instrument platforms if similar fragmentation and chromatography techniques are adopted. In the current example, shotgun and DIA runs were performed on a 5600 TripleTof mass spectrometer (ABSciex) with NanoLC-2Dplus HPLC system. The instrument setup, parameters, collision energy and window settings are available in the study (Selevsek *et al.*, 2015). Additionally, all the data files were submitted to the ProteomeXchange repository (PRIDE) (ID: PXD001010) (see "Relevant Websites section"). These data files include DDA and DIA .wiff files which can be accessed and utilized for performing the quantitative analysis.

Computational Tools

For this article, to perform the identification and quantification of yeast proteome using DIA strategy, a number of tools were employed. For instance, DDA data was identified using TPP (Deutsch *et al.*, 2015). The identified spectra were converted into spectral libraries using SpectraST (Lam *et al.*, 2007) and Skyline. Targeted data extraction was performed in Skyline (Pino *et al.*, 2017). Peak scoring and statistical analysis were executed in Skyline using mProphet algorithm (Reiter *et al.*, 2011). Lastly, protein level relative quantification and significance analysis were carried out using MSstats functionalities (Choi *et al.*, 2014), supported in Skyline. Skyline software (see **Fig. 3**) can be freely downloaded, with mProphet and MSstats modules incorporated in Skyline.

Step1: Spectral Identification and Library Generation

The DDA-based library was generated by running shotgun experiments, comprising of 46 MS injections. These include samples from BY4741 strain and at different time points in response to osmotic shock. Details of these 46 injections are provided in the study (Selevsek *et al.*, 2015). The 46 data files are available online in wiff format with the following names:

1. *'nselevse_L120203_006.wiff' to 'nselevse_L120203_029.wiff'* (24 files);
2. *'nselevse_L120327_001.wiff' to 'nselevse_L120327_018.wiff'* (18 files);
3. *'igillet_L120122_001.wiff' to 'igillet_L120122_003.wiff'* (3 files), and
4. *'igillet_L120124_003.wiff'* (1 file).

The generated spectra were identified using TPP tools against Saccharomyces Genome Database (SGD) (Cherry *et al.*, 2012). The search results were statistically sorted according to probability at 1% FDR. The identified spectra from all the samples were then compiled to generate *'iProphet_Combined.pepXML'* file and used to build a consensus spectral library.

Author's personal copy

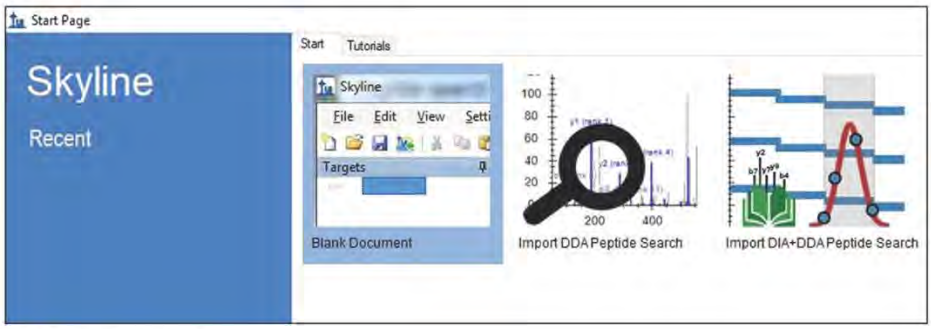


Fig. 3 Skyline software start-up page.

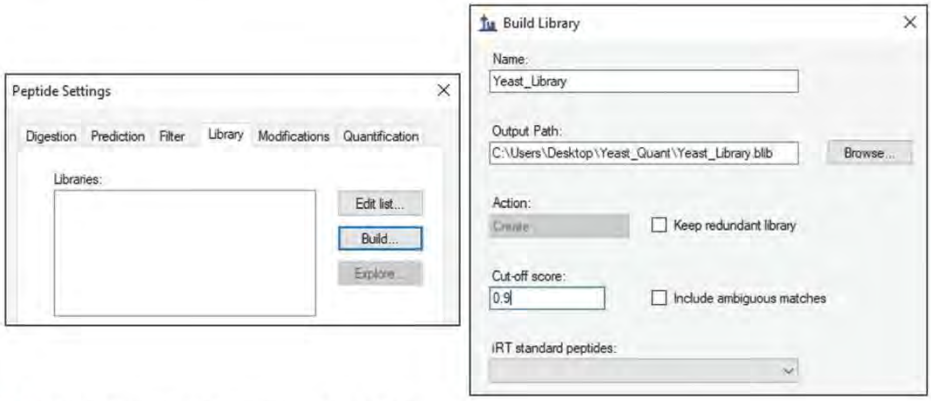


Fig. 4 Library settings in Skyline for generation of spectral library.

Building a spectral library requires the following files:

- pepXML file (search engine result)
- mzXML files (spectral data)

The pepXML file is available at ProteomeXchange repository (ID: PXD001010) whereas mzXML spectra files can be acquired by converting 46 DDA wiff files to XML format using Msconvert tool (Kessner *et al.*, 2008). These converted mzXML data files are necessary to extract the chromatograms from the spectral library during DIA targeted data extraction.

Spectral library using Skyline

Skyline has a 'Build Library' module in 'Peptide Settings' to generate the spectral library (see Fig. 4). It requires library name as *'Yeast_Library.blib'*, output path and PeptideProphet cut-off score as *'0.9'* (false discovery rate of below 1%). Retention times, mass to charge values and relative intensity values for each spectrum are also stored in the spectral library, which can be visualized in the Skyline (see Fig. 5).

Step 2: Fragment Ion Library or Assay Library

Fragment ion library or assay library is a subset of the spectral library. It contains the most intense particular precursor and product ions (MS/MS transitions) from the spectral library or a targeted proteins list. These assays contribute in DIA targeted data extraction of peptides and proteins of interest. These characteristics have to be determined prior to importing the DIA data into the Skyline. For this purpose, Skyline provides various peptide and transition settings with 'Filter', 'Library' and 'Modifications' modules. 'Modification' allows one to include and exclude different types of structural and isotopic modifications. In addition to the default options, new

Author's personal copy

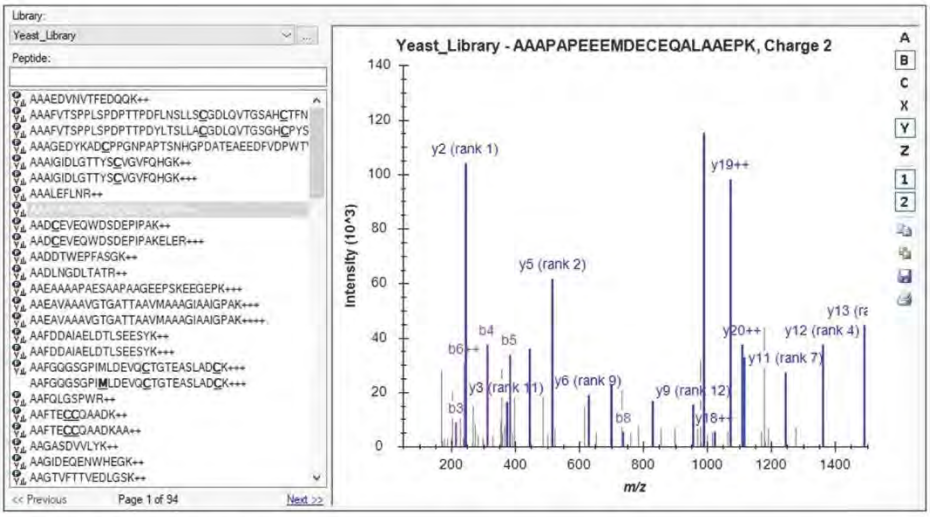


Fig. 5 Spectral library visualization in Skyline.

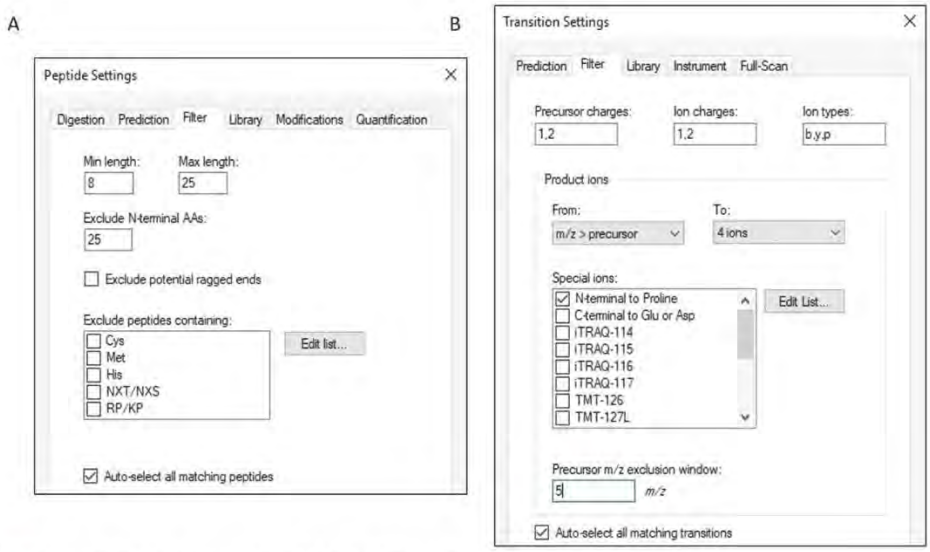


Fig. 6 Filter settings for transition selection in Skyline. (A) Peptide settings. (B) Transition settings.

modifications can also be added to the Skyline. In 'Filter' module, peptides length, precursor/product ion charges and types, and a number of ions can be determined (see Fig. 6). Subsequently, through 'Library' settings, ion match tolerance values (match between DIA and DDA data) and library to be added in the analysis can be set. It also provides options for selecting most intense transitions from the library, which also fulfils the 'Filter' settings (see Fig. 7). For the current data set, following criteria were set (see Figs. 6 and 7):

- Peptide length=min 8 and max 25
- Structural modifications=no oxidation on methionine

Author's personal copy

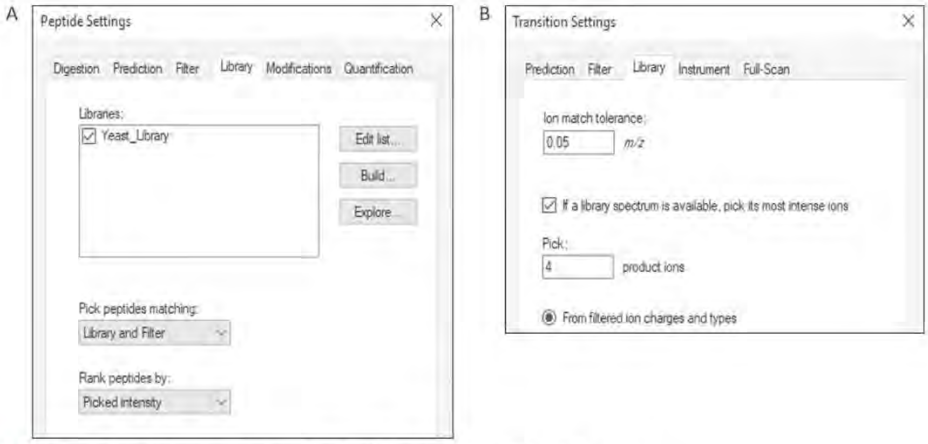


Fig. 7 Library settings for transition selection in Skyline. (A) Peptide settings. (B) Transition settings.

- Ion Intensities = highly intense in spectral library
- Ion Charges (precursor and product ions) = single (1) and double (2)
- Ion types = b, y and p (precursor) ions
- Precursor *m/z* exclusion window = 5 *m/z*
- Match tolerance = 0.05 *m/z*
- Number of selected transitions = top four

Using 'precursor m/z exclusion window' option, spectra in a mass window around precursor *m/z* value can be excluded to minimize the noise and extract the refined transitions. In the 'Library' and 'Filter' settings, further transition options are also given such as, 'auto-select all matching transitions', 'pick most intense transitions from library spectrum' and 'pick peptides matching both the library and filter settings'. After setting up the environment for transitions selection, a list of targeted proteins is loaded into the Skyline to generate the fragment or assay library.

Step 3: Targeted Proteins

In contrast to MRM data analysis, where proteins of interest are already determined prior to performing the experiment, in DIA, the list of targeted proteins or peptides are loaded after the experiment has been performed. This allows the extraction and analysis of DIA data that is only related to proteins of interest from the whole data independent run, which usually analyses all the peptides within the range of 400–1200 *m/z* (depending upon the instrument settings). The targeted proteins/peptides list can be provided in different ways in Skyline, such as, selected externally and imported into the skyline using 'Import Fasta' module, and/or adding all those proteins/peptides which are available in the spectral library. It can also be added by simply inserting the proteins and peptide sequences in 'Targets' module. In this study, peptides which were identified and stored in the spectral library, and external protein data i.e. *'Yeast.fasta'* were added to the target list. The proteins present in the fasta file also undergo *in silico* tryptic digestion in Skyline. While the protein data is provided, all those transitions which matched the transition settings (in step 2) were added to the target list along with their abundances and retention time values, as shown in Fig. 8.

Step 4: DIA Data Extraction Environment

Before importing and extracting the DIA data from experimental runs, Skyline needs to set up the DIA environment. In this, all the instrument and isolation settings for MS1 and MS/MS filtering are configured in 'Transition Settings' under 'Full-Scan' module. The full-scan features include:

- Precursor mass analyzer
- Product mass analyzer
- Resolving powers of analyzers

Author's personal copy

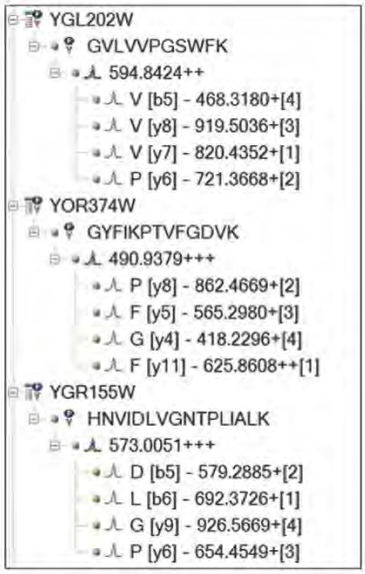


Fig. 8 Visualization of targeted transitions list in Skyline.

- Acquisition method
- Isolation Scheme (pattern of precursor isolation windows in DIA acquisition)
- Retention time filtering

A DIA isolation scheme defines the pattern for each acquisition window (m/z range) through the entire mass range by the specific instrument. Skyline provides default isolation schemes with a variety of patterns, for example, SWATH (15 m/z), SWATH (25 m/z) etc. A customized pattern can also be designed and added according to the particular instrument used for data acquisition experiment. To set up a user-specific isolation scheme, following parameters are required:

- Mass range = 400 to 1200 m/z
- Window (isolation) width = 25 m/z
- Number of windows = 32
- Total cycle time = 3.3 s
- Accumulation time (time spent on single transition) = 10 ms

For the current data set, these values are filled up according to the 5600 TripleTof mass spectrometer used for yeast proteome quantification, provided in the study (Selevsek *et al.*, 2015). After specifying the settings, Skyline displays the isolation windows in tabular and graphical form.

Retention time filtering defines the time range over which each chromatogram is extracted from DIA data against a single targeted transition. The chromatogram can be extracted over the entire gradient, however, it may result in the noisy and distorted peaks with a minimum understanding of correct chromatogram peak. For precise data extraction and identification, retention time value for each peptide in the experiment is required. Skyline provides couple of options limiting the RT range, such as, (i) using RT values of peptides found in the spectral library, (through DDA runs), (ii) predicting these values using SSRCalc hydrophobicity algorithm (Spicer *et al.*, 2007) and (iii) calculating RT values using standard RT peptides through empirical measurements (also called as normalized RTs or iRTs). Details of the SSRCalc and normalized RTs is available in the Skyline tutorial.

Subsequently, assigning the values to the other parameters in 'Full-Scan' module, set up the DIA and DDA environment in Skyline for data extraction and analysis (see **Fig. 9**).

- Precursor mass analyzer = TOF
- Product mass analyzer = TOF
- Resolving powers of analyzers = 30,000
- Acquisition method = DIA

Author's personal copy

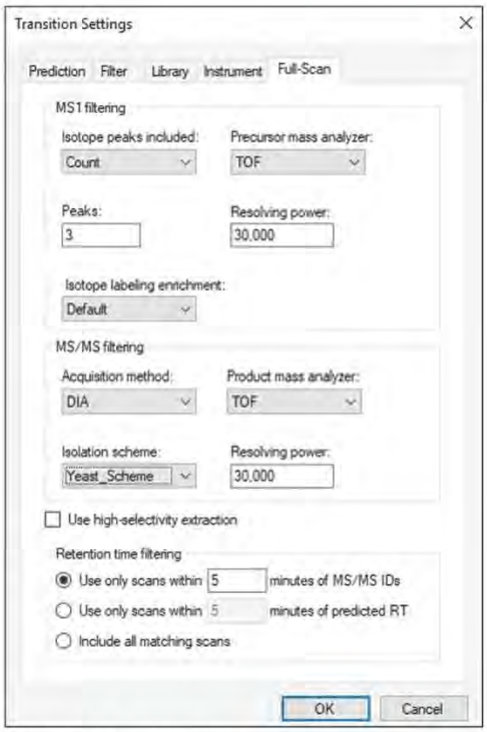


Fig. 9 MS1 and MS/MS filtering settings for DIA data extraction in Skyline.

- Isolation Scheme (pattern of precursor isolation windows in DIA acquisition)=Yeast_Scheme (user-defined)
- Retention time filtering=scans within 5 min of MS/MS IDs

This retention time filtering value will extract the chromatograms from DIA spectra within 5 min range of the peptide-spectrum match for each targeted peptide from the DDA run in a spectral library.

Step 5: DIA Targeted Data Extraction

Following setting up the environment in Skyline, data files from SWATH-MS runs for yeast biological replicates can be imported. In this study, changes in the *S. cerevisiae* proteome were quantified over 6-time points i.e. 0 min (T0), 15 min (T1), 30 min (T2), 60 min (T3), 90 min (T4) and 120 min (T5), as a result of osmolarity stress. Samples were collected, digested and analyzed using SWATH-MS in triplicates at each time point. The 18 data files are available on ProteomeXchange database (see "Relevant Websites section") in wiff format with following names:

1. *'L120412_001_SW.wiff' to 'L120412_003_SW.wiff'* (T0) (3 files);
2. *'L120412_004_SW.wiff' to 'L120412_006_SW.wiff'* (T1) (3 files);
3. *'L120412_007_SW.wiff' to 'L120412_009_SW.wiff'* (T2) (3 files);
4. *'L120412_010_SW.wiff' to 'L120412_012_SW.wiff'* (T3) (3 files);
5. *'L120412_013_SW.wiff' to 'L120412_015_SW.wiff'* (T4) (3 files); and
6. *'L120412_016_SW.wiff' to 'L120412_018_SW.wiff'* (T5) (3 files).

These files can be imported directly into Skyline (compatible with different file formats) using 'Import Results' option in 'File' menu. 'Import Results' provides multiple preferences including (i) add single-injection replicates in files, (ii) add multi-injection replicates in directories, (iii) add one new replicate for different types of results (see Fig. 10). As the SWATH runs were added to the Skyline for analysis, it extracted the chromatograms against peptides available in the 'Targets' list while taking into account the retention time filtering parameter. This is also called as 'Targeted Data Extraction'.

Author's personal copy

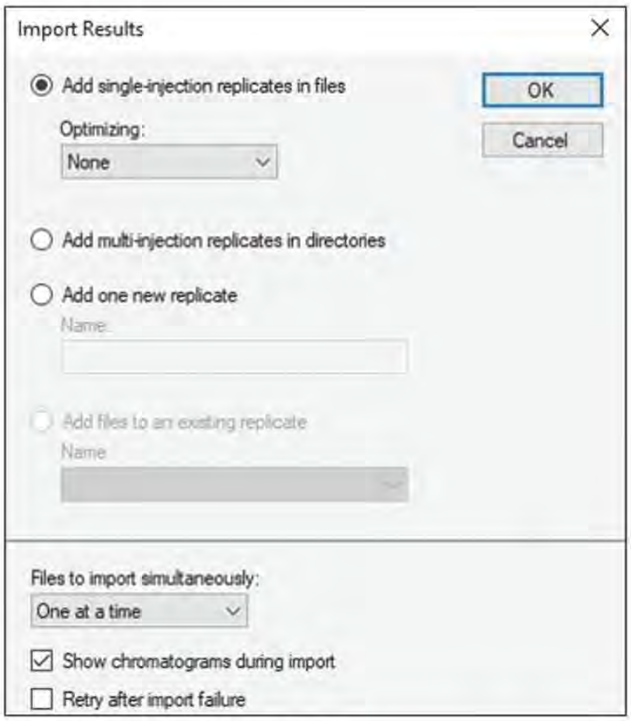


Fig. 10 Import results module for importing mass spectrometry results in Skyline.

Peak picking and scoring

The extracted chromatograms then undergo the process of identifying the correct peak for each targeted peptide, which correlates best with the spectrum in the spectral library. Although this is done automatically, peaks can also be picked manually. For each match, matching scores are allocated, also called as 'dot products'. In order to select the best peak, in addition to its conventional heuristic method, Skyline has also incorporated mProphet algorithm (Reiter *et al.*, 2011) scoring strategy for measuring the similarity between chromatogram and library peaks to choose the best match. The mProphet algorithm takes into account various parameters for scoring purposes, such as:

- Co-elution
- Mass accuracy
- Peak shape
- Ion intensity
- Relative product ion abundance correlation
- Predicted retention time

Additionally, Skyline is equipped with the facility of training and customizing the mProphet peak picking and scoring model for peak determination. This can be executed by generating decoy peptides and calculating FDRs for peaks selection. Details of manual peak-spectrum matching are available in further readings.

Step 6: DIA Data Exploration and Analysis

After importing the DIA data and performing peak selection, chromatograms for 18 biological replicates can be visualized, simultaneously. This can be done by arranging result files systematically into three groups in tabular form using 'Arrange Graphs' module of the Skyline in 'View' menu. Each group indicates the data for all six time points in each replicate. Chromatograms for the first replicate at time T0 to T5 are displayed in Fig. 11. In each graph, best peaks are displayed within the black dotted lines

Author's personal copy

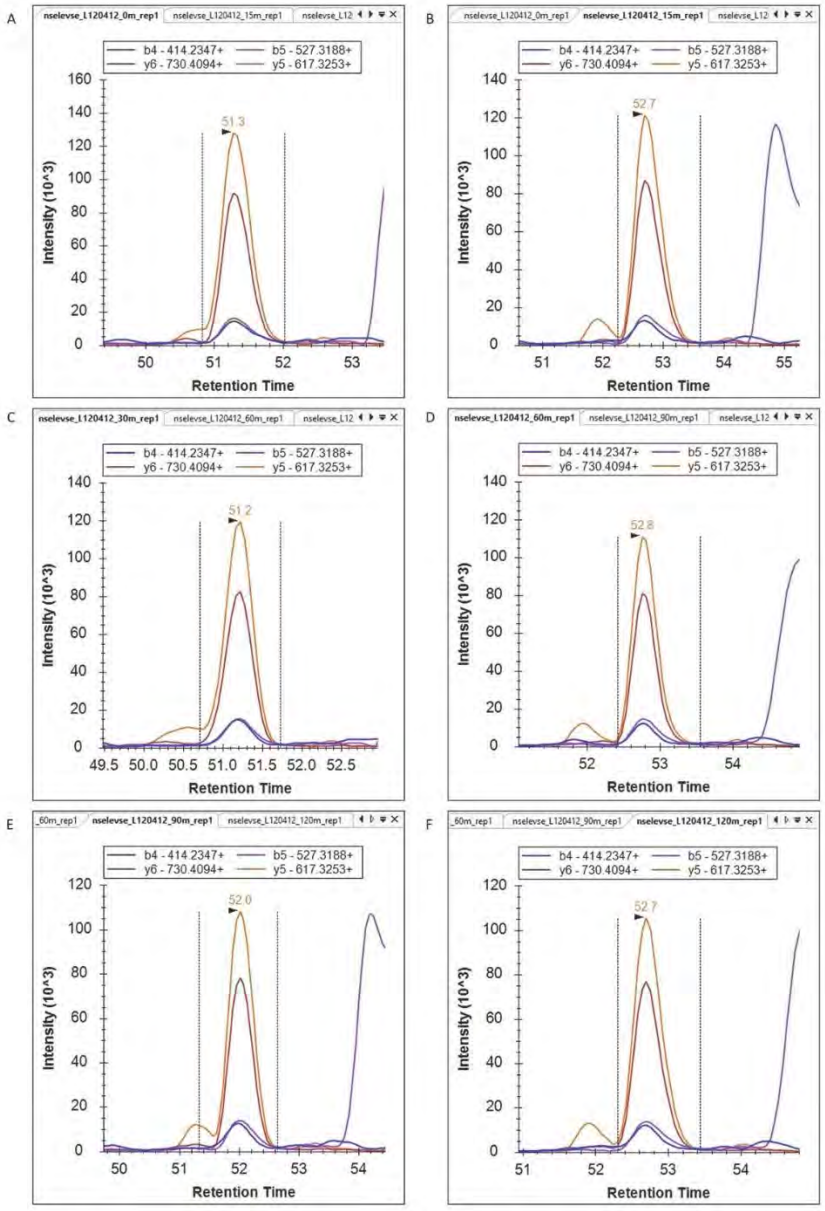


Fig. 11 Chromatograms of yeast peptide from DIA data in replicate 1 at six time points. (A) 0 min (T0). (B) 15 min (T1). (C) 30 min (T2). (D) 60 min (T3). (E) 90 min (T4). (F) 120 min (T5).

Author's personal copy

which indicates the peak boundaries. Visualization and graphical attributes of chromatograms can be customized using a variety of options in 'View' menu, such as:

- Retention times (all or best peak) = Best peak
- Peptide ID times (none or matching with library) = Matching
- Transitions (all, precursor or products) = All
- Transform (interpolated or Savitsky-Golay smoothing) = Savitzky-Golay Smoothing
- Auto-zoom X-axis = Best peak
- Auto-scale Y-axis

Peptides selection for quantification

From the given list of targeted proteins, a number of peptides or peaks are matched and identified (from DIA and DDA data). For the downstream quantification measurements, those peptides should be incorporated in a way that shows consistency in their peak areas and retention times in multiple replicates. So, in order to perform this refinement, peak intensities and RTs of all the peptides can be compared and inspected using Skyline. In 'View' menu, 'Replicate Comparison' option for 'Retention Times' and 'Peak Areas' can be used to analyze the intensity and time comparisons in all 18 DIA replicate runs for yeast proteome at six time points.

In retention time replicate view (see Fig. 12(A)), following graph parameters were set:

- Value (all or retention time) = All
- Transitions (all, precursor or product ions) = All
- Order (acquired time or document) = Document

The RT of each fragment ion is represented in separate color and collectively each set of bars indicates the RT pattern of the whole peptide, over 18 runs. The start and end retention time points depict the elution times of the peptide in triplicates. The same height of all the groups shows that the peptide elutes in approximately equal time at all six time points.

In peak intensity replicate view (see Fig. 12(B)), following graph parameters were set:

- Normalized to (total, maximize or none) = None
- Transitions (all, precursor or product ions) = All
- Order (acquired time or document) = Document

The intensity of each fragment ion is illustrated in different colors and collectively each bar indicates the peak intensity of the whole peptide, over 18 runs. From the graph, it can be demonstrated that intensity values get lower after 15 min of exposure to osmotic stress in replicate 1 and 3, whereas, a slight increase in intensity is observed at 90 min and 120 min of exposure. However, in replicate 2, the intensity values during first 60 min are approximately the same and get decreased at 90 and 120 min. Moreover, relative transition intensity of peptides in all 18 runs can be visualized by normalizing the values to 'Total' in 'Peak Area' option in the replicate graph.

Consequently, similar peak intensity and retention time pattern among all the replicates validate the reproducibility of the MS instrument and experiment. It also depicts that the Skyline has identified and picked the correct peak for that particular peptide. Similarly, all the peptides can be inspected through these graphs individually from the target list, to select for quantification purposes.

If the intensity and retention time data for a peptide is not consistent among replicates (see Fig. 13), it can be improved by manual selection of correct peaks. It can be observed in Fig. 14 that for this particular peptide, unlike replicate 3, the correct peak is not selected in replicate 1 and replicate 2. This can be resolved by moving the peak boundaries via dragging the black dotted lines

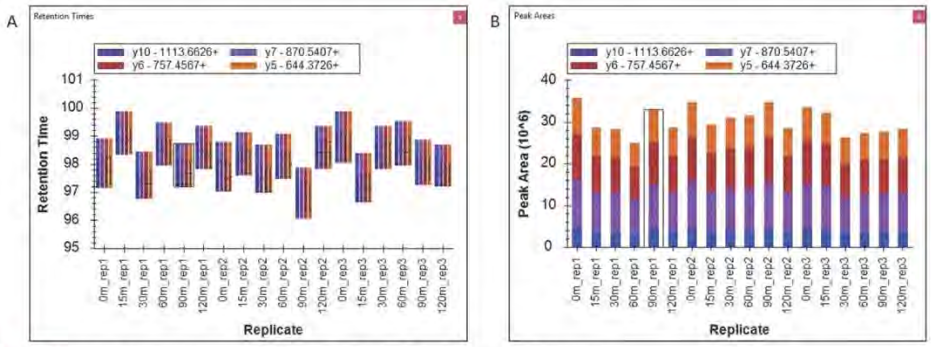


Fig. 12 Consistent replicate comparison graphs in Skyline. (A) Retention times (B) Peak areas.

Author's personal copy

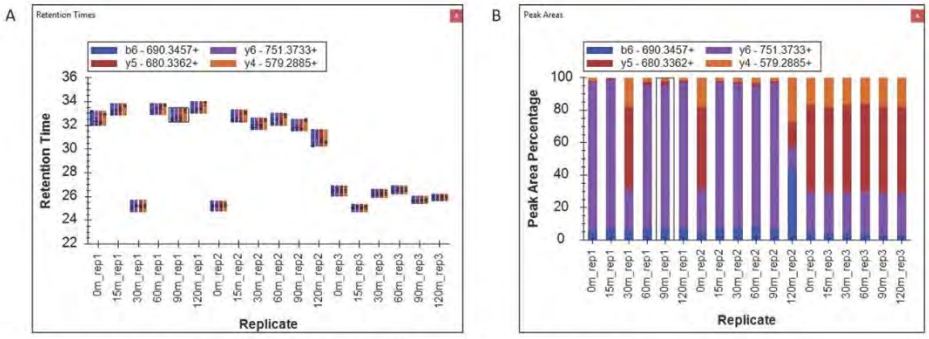


Fig. 13 Inconsistent replicate comparison graphs in Skyline. (A) Retention times (B) Peak areas.

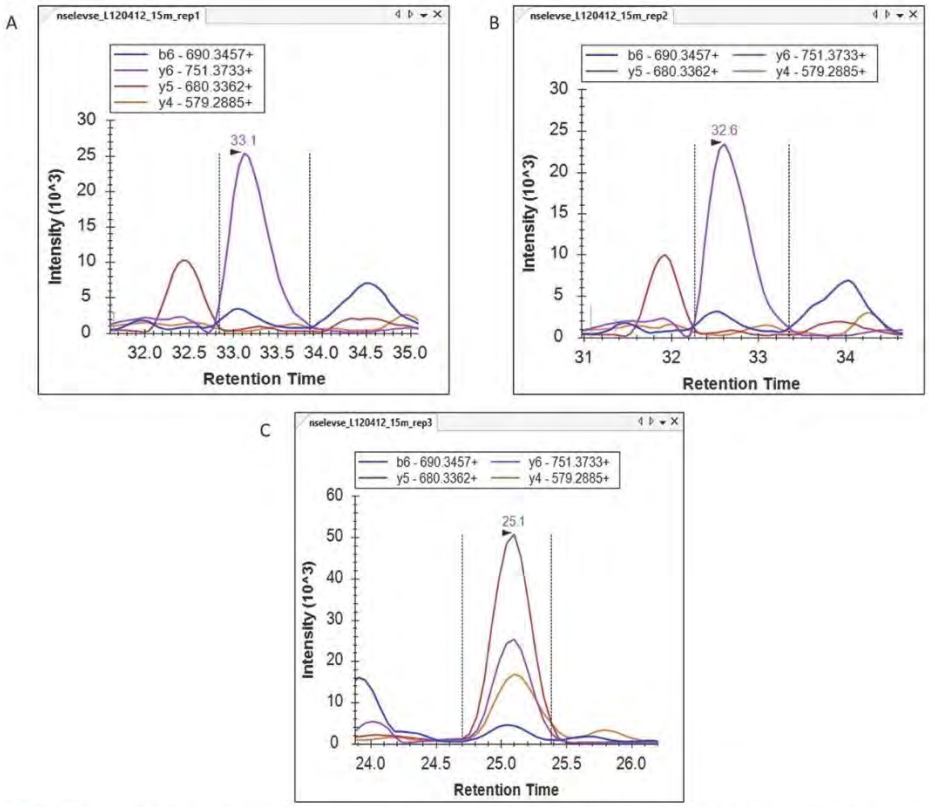


Fig. 14 Chromatograms of selected peaks at 15 min in Skyline. Incorrect peaks are selected in (A) Replicate 1 and (B) Replicate 2. (C) Peak in Replicate 3 is correctly identified.

Author's personal copy

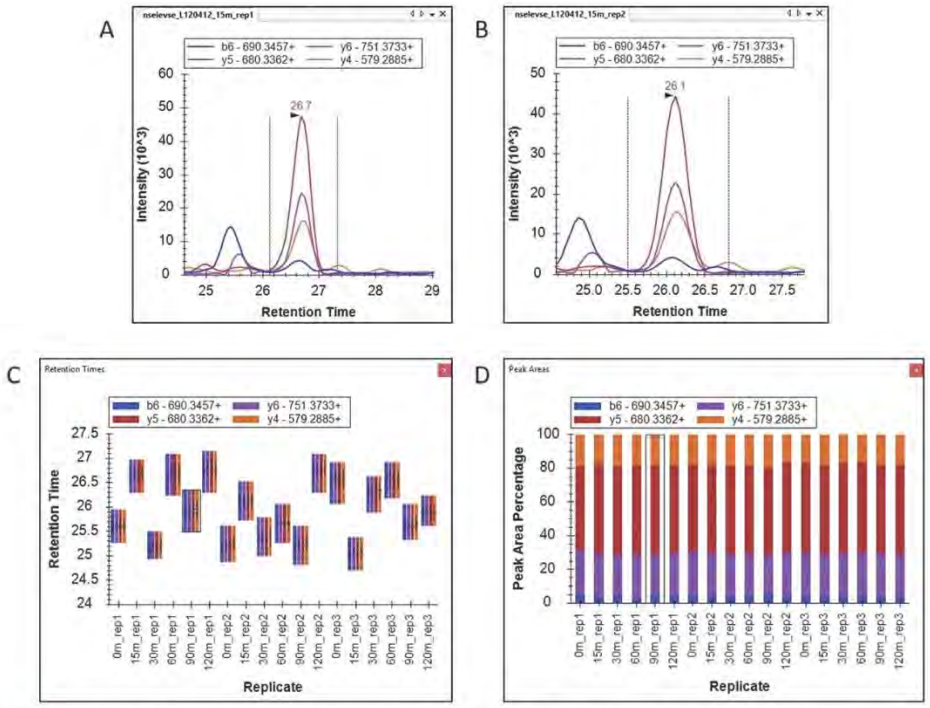


Fig. 15 Chromatograms of correctly selected peaks for specific peptide at 15 min. (A) Replicate 1. (B) Replicate 2. (C) Retention times consistency. (D) Peak areas consistency.

around the peaks. These corrections can now be observed in peak area and retention time graphs as well (see Fig. 15). If the data does not show consistency even after adjustments, the insignificant peptide can be deleted from the study.

Step 7: Statistical Analysis for Quantification

The aim of this study was to quantify the abundance of yeast proteome in response to the osmotic stress condition and identify those proteins which show a significant change over the course of the time period of two hours. Skyline provides the facility of performing this statistical analysis by annotating the replicate data with some additional classes or conditions. In this case study, conditions are the different time points, along which we have to observe the changes. Data annotation can be performed using 'Define Annotation' module in the 'Document Settings' in Skyline. Condition values are named as 'a0m', 'b15m', 'c30m', 'd60m', 'e90m' and 'f120m' and applied onto 'Replicates' (see Fig. 16). Select 'Condition' and 'BioReplicate' in 'Document Settings' to incorporate these annotations into the current Skyline document to analyze yeast data.

Intensity and time variation among technical replicates

Using replicate comparison function of Skyline, variation among the intensities and retention times of the quantotypic peptides in different conditions can be analyzed. In peak areas and retention time views, following graph parameters were set:

- Transitions (all, precursor, products or total) = Total
- Group by (replicate or condition) = Condition
- Normalized to (maximum, total or none) = Maximize
- Select CV values

Graphs shown in Fig. 17 illustrate the variation among different conditions or time points for this particular peptide belonging to protein *'YGR240C'*, in all replicates. All the peptides can be analyzed by iterating over the entire set of targeted proteins. It can be deduced from the graph (see Fig. 17) that intensity of this specific peptide greatly varies among all replicates at 90 min after the exposure to osmotic stress environment and the least variation is observed at 0 min and 120 min.

Author's personal copy

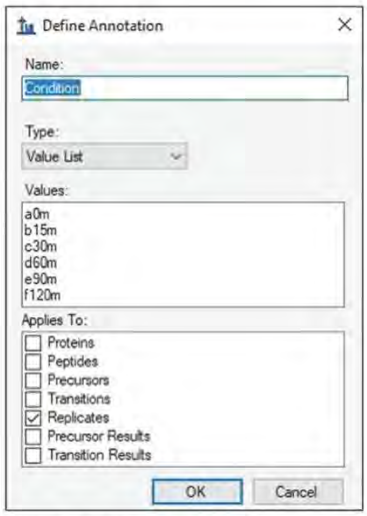


Fig. 16 Annotation module for defining new annotations for replicate data in Skyline.

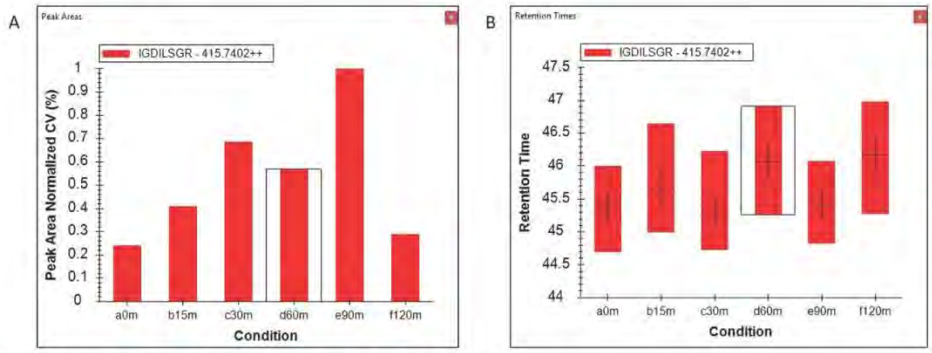


Fig. 17 Variation among replicates at six time points. (A) Intensity. (B) Time.

Average peptide abundance among conditions

To study the average peptide abundance or difference in peptide expression among different conditions, following graph parameters were set in peak areas replicate comparison view:

- Transitions (all, precursor, products or total)=Total
- Group by (replicate or condition)=Condition
- Normalized to (maximum, total or none)=Total
- Deselect CV values

The graph in Fig. 18 presents the average mean of peptide abundances in triplicates among all conditions. The red bars in the graph represent the mean average values and black lines on the bars represent the standard deviation, which measures the amount of dispersion in the intensity values. A high standard deviation at 90 min depicts the high variation in data values at this time in all replicates.

Overlapping mean values for this specific peptide at six time points, shown in the Figs. 17 and 18, do not contribute greatly to the differential abundance analysis, and therefore, demonstrate that protein *'YGR240C'* exhibits no significant response to osmotic stress. Contrary to this, peptide response from protein *'YMR169C'*, shown in Fig. 19, is highly predictive and presents a supposition

Author's personal copy

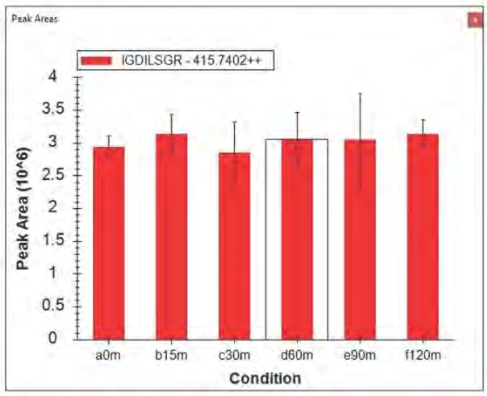


Fig. 18 Average peptide abundances among replicates at six time points.

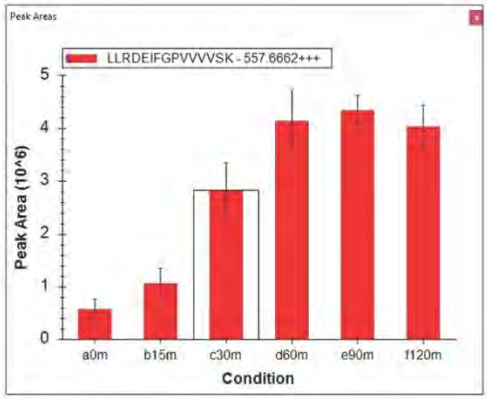


Fig. 19 Significant change in abundances during six time points after osmotic shock.

that this protein may have been upregulated under the influence of osmotic stress. However, this hypothesis requires the similar behaviour from more than one peptide to validate the up or down-regulation of the corresponding protein.

Group comparison among conditions

Differential abundance analysis between time points can also be performed through pairwise group comparison investigation i.e. measure the change in abundance between time points:

- 0 min (T0) to 15 min (T1)
- 0 min (T0) to 30 min (T2)
- 0 min (T0) to 60 min (T3)
- 0 min (T0) to 90 min (T4)
- 0 min (T0) to 120 min (T1)

In Skyline, 'Group Comparison' module in 'Document Settings' provides the facility of adding as many comparisons as required. In 'Edit Group Comparison' form, for the first comparison, following parameters were set:

- Name=T0 v. T1
- Control group annotation=Condition
- Control group value=a0m

Author's personal copy

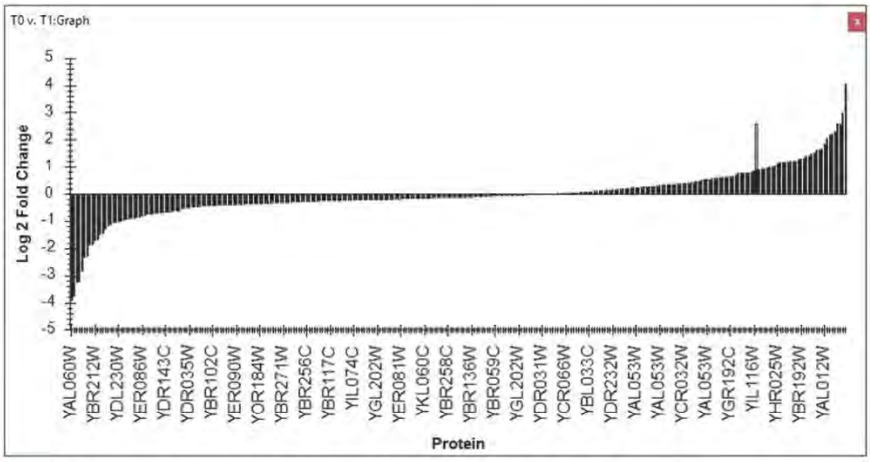


Fig. 20 Log2 fold change in abundances of proteins during the time period between 0 min (T0) and 15 min (T1).

- Value to compare against=b15m
- Identity annotation=BioReplicate
- Normalization method=Equalize medians
- Confidence level=95%
- Scope=Peptide or Protein for peptide or protein level analysis, respectively

To inspect the abundances of proteins within the time frame of first 15 min, the graph can be visualized in 'View' and 'Group Comparison' menu of Skyline. The graph shown in **Fig. 20** illustrates the overall fold change, either increase or decrease or no effect, in yeast proteins. By employing the same method on all other comparison groups, protein expressions during different time periods can be analyzed.

Data Visualization

Different quantification strategies provide a measure of change in proteome between different biological conditions/samples. Elucidation of biological relevancy of these quantifications requires additional analysis. Perseus is widely adopted for analysis post-database search, for visualization and statistical analysis. There are various tools available for pathway enrichment and analysis, Gene ontology (GO) enrichment could be performed by Panther (Mi *et al.*, 2016), protein-protein interaction information could be obtained from STRING (Szklarczyk *et al.*, 2015), IntAct (Orchard *et al.*, 2014), Human Protein Reference Database (HPRD) (Keshava Prasad *et al.*, 2009) and Biological General Repository for Interaction Datasets (BioGRID) (Stark *et al.*, 2011) amongst others.

Perseus (Tyanova *et al.*, 2016) is a freely available software that performs shotgun proteomics analysis of biological relevance from processed raw files. In addition, can perform various visualization, clustering, principal component analysis (PCA) and statistical tests on quantitative and time series data, details of which are available on Perseus documentation.

Conclusion

Mass spectrometry based quantitative proteomics has rapidly changed the landscape of proteomics research spurred on by rapid technological advances and more importantly huge strides ahead in computational analysis and informatics. It has to be noted though, that the major limitation of quantitative studies is the thresholds and stringencies set by the users. Although there is yet to be standards developed for quantitative proteomics (Mohamedali *et al.*, 2017), it behoves researchers and users to be wary of and thoroughly statistically analyze all mass spectrometry data keeping strict stringency to eliminate false positives. The example used in this article illustrates that the procedure for the analysis for protein quantification by proteomics is relatively straightforward with freely available tools. It has to be emphasised though, that users have a good understanding of the theoretical underpinnings of Mass Spectrometry and informatics to ensure that the results obtained from their studies are robust and reproducible.

Furthermore, upon the identification of differentially expressed proteins between samples after a rigorous experimental and analytical approach, orthogonal techniques to validate potential candidates using targeted techniques such as MRM's and ELISA are almost a compulsory part of the reliability of MS quantification (Vidova and Spacil, 2017). The use therefore, of large scale quantitative proteomics using a label-free DIA approach has been a mainstay of the discovery part of proteomic studies and indeed

Author's personal copy

has proven to be a formidable, robust and reliable method of quantifying large numbers of proteins in complex biological samples. Certain limitations such as the size and nature of libraries, quality of sample preparation, reproducibility on varied instruments, quality of data, statistical threshold guidelines etc. all still mean that quantitative proteomics has a way to go to achieve the robustness and reliability of quantitative transcriptomics.

See also: Clinical Proteomics. Genome-Wide Scanning of Gene Expression. Identification of Proteins from Proteomic Analysis. Natural Language Processing Approaches in Bioinformatics. Sequence Analysis

References

Ahrne, E., Molzahn, L., Glatter, T., Schmidt, A., 2013. Critical assessment of proteome-wide label-free absolute abundance estimation strategies. Proteomics 13 (17), 2567–2578.
Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B., 2007. Quantitative mass spectrometry in proteomics: A critical review. Anal. Bioanal. Chem. 389 (4), 1017–1031.
Bruderer, R., Sondermann, J., Tsou, C.C., *et al.*, 2017. New targeted approaches for the quantification of data-independent acquisition mass spectrometry. Proteomics 17 (9),
Cherry, J.M., Hong, E.L., Amundsen, C., *et al.*, 2012. Saccharomyces genome database: The genomics resource of budding yeast. Nucleic Acids Res. 40 (D1), D700–D705.
Choi, M., Chang, C.Y., Clough, T., *et al.*, 2014. MSstats: An R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics 30 (17), 2524–2526.
Deutsch, E.W., Mendoza, L., Shteynberg, D., *et al.*, 2015. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. Proteomics Clin. Appl. 9 (7–8), 745–754.
Distler, U., Kuharev, J., Navarro, P., Tenzer, S., 2016. Label-free quantification in ion mobility-enhanced data-independent acquisition proteomics. Nat. Protoc. 11 (4), 795–812.
Fan, X.G., Zheng, Z.Q., 1997. A study of early pregnancy factor activity in preimplantation. Am. J. Reprod. Immunol. 37 (5), 359–364.
Gillette, M.A., Carr, S.A., 2013. Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. Nat. Methods 10 (1), 28–34.
Keshava Prasad, T.S., Goel, R., Kandasamy, K., *et al.*, 2009. Human protein reference database – 2009 update. Nucleic Acids Res. 37 (Database issue), D767–D772.
Kessner, D., Chambers, M., Burke, R., Agusand, D., Mallick, P., 2008. ProteoWizard: Open source software for rapid proteomics tools development. Bioinformatics 24 (21), 2534–2536.
King, C.D., Dudenhoeffer, J.D., Gu, L., Evans, A.R., Robinson, R.A.S., 2017. Enhanced sample multiplexing of tissues using combined precursor isotopic labeling and isobaric tagging (cPILOT). J. Vis. Exp. 123).
Lam, H., Deutsch, E.W., Eddes, J.S., *et al.*, 2007. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 7 (5), 655–667.
Law, K.P., Lim, Y.P., 2013. Recent advances in mass spectrometry: Data independent analysis and hyper reaction monitoring. Expert Rev. Proteomics 10 (6), 551–566.
Lindemann, C., Thomanek, N., Hundt, F., *et al.*, 2017. Strategies in relative and absolute quantitative mass spectrometry based proteomics. Biol. Chem. 398 (5–6), 687–699.
Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T., Thomas, P.D., 2016. PANTHER version 10: Expanded protein families and functions, and analysis tools. Nucleic Acids Res. 44 (D1), D336–D342.
Mohamedali, A., Ahn, S.B., Sreenivasan, V.K.A., Ranganathan, S., Baker, M.S., 2017. Human Prestin: A candidate PE1 protein lacking stringent mass spectrometric evidence? J. Proteome Res. 16 (12), 4531–4535.
Mueller, L.N., Brusniak, M.Y., Mani, D.R., Aebersold, R., 2008. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J. Proteome Res. 7 (1), 51–61.
Neilson, K.A., Ali, N.A., Muralidharan, S., *et al.*, 2011. Less label, more free: Approaches in label-free quantitative mass spectrometry. Proteomics 11 (4), 535–553.
Norbeck, A.D., Monroe, M.E., Adkins, J.N., *et al.*, 2005. The utility of accurate mass and LC elution time information in the analysis of complex proteomes. J. Am. Soc. Mass Spectrom. 16 (8), 1239–1249.
Orchard, S., Ammari, M., Aranda, B., *et al.*, 2014. The MIntAct project – IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 42 (Database issue), D358–D363.
Pino, L.K., Searle, B.C., Bollinger, J.G., *et al.*, 2017. The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. Mass Spectrom. Rev.
Reiter, L., Rinner, O., Picotti, P., *et al.*, 2011. mProphet: Automated data processing and statistical validation for large-scale SRM experiments. Nat. Methods 8 (5), 430–435.
Rost, H.L., Rosenberger, G., Navarro, P., *et al.*, 2014. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. 32 (3), 219–223.
Selevsek, N., Chang, C.Y., Gillet, L.C., *et al.*, 2015. Reproducible and consistent quantification of the Saccharomyces cerevisiae proteome by SWATH-mass spectrometry. Mol. Cell. Proteomics 14 (3), 739–749.
Spicer, V., Yamchuk, A., Cortens, J., *et al.*, 2007. Sequence-specific retention calculator. A family of peptide retention time prediction algorithms in reversed-phase HPLC: Applicability to various chromatographic conditions and columns. Anal. Chem. 79 (22), 8762–8768.
Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., *et al.*, 2011. The BioGRID interaction database: 2011 update. Nucleic Acids Res. 39 (Database issue), D698–D704.
Szklarczyk, D., Franceschini, A., Wyder, S., *et al.*, 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43 (Database issue), D447–D452.
Tang, H., Fang, H., Yin, E., *et al.*, 2014. Multiplexed parallel reaction monitoring targeting histone modifications on the QExactive mass spectrometer. Anal. Chem. 86 (11), 5526–5534.
Tsou, C.C., Avtonomov, D., Larsen, B., *et al.*, 2015. DIA-Umpire: Comprehensive computational framework for data-independent acquisition proteomics. Nat. Methods 12 (3), 258–264.
Tyanova, S., Temu, T., Sinitcyn, P., *et al.*, 2016. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods 13 (9), 731–740.
Vidova, V., Spacil, Z., 2017. A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. Anal. Chim. Acta 964, 7–23.
Vogel, C., Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. 13 (4), 227–232.
Washburn, M.P., Wolters, D., Yates 3rd, J.R., 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. 19 (3), 242–247.
Wenner, B.R., Lynn, B.C., 2004. Factors that affect ion trap data-dependent MS/MS in proteomics. J. Am. Soc. Mass Spectrom. 15 (2), 150–157.

Further Readings

Anand, S., Samuel, M., Ang, C.S., Keerthikumar, S., Mathivanan, S., 2017. Label-based and label-free strategies for protein quantitation. Proteome Bioinform. 31–43.
Anderle, M., Roy, S., Lin, H., Becker, C., Joho, K., 2004. Quantifying reproducibility for differential proteomics: Noise analysis for protein liquid chromatography-mass spectrometry of human serum. Bioinformatics 20 (18), 3575–3582.

Author's personal copy

Bantscheff, M., Lemeer, S., Savitski, M.M., Kuster, B., 2012. Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. Anal. Bioanal. Chem. 404 (4), 939–965.
Chen, Y., Wang, F., Xu, F., Yang, T., 2016. Mass spectrometry-based protein quantification. In: Mirzaei, H., Carrasco, M. (Eds.), Modern Proteomics–Sample Preparation, Analysis and Practical Applications. Springer International Publishing, pp. 255–279.
Clough, T., Thaminy, S., Ragg, S., Aebersold, R., Vitek, O., 2012. Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. BMC Bioinform. 13 (16), S6.
Karpievitch, Y., Stanley, J., Taverner, T., *et al.*, 2009. A statistical framework for protein quantitation in bottom-up MS-based proteomics. Bioinformatics 25 (16), 2028–2034.
Lill, J., 2003. Proteomic tools for quantitation by mass spectrometry. Mass Spectrom. Rev. 22 (3), 182–194.
Pan, S., Aebersold, R., Chen, R., *et al.*, 2008. Mass spectrometry based targeted protein quantification: Methods and applications. J. Proteome Res. 8 (2), 787–797.
Wang, M., You, J., Bemis, K.G., Tegeler, T.J., Brown, D.P., 2008. Label-free mass spectrometry-based protein quantification technologies in proteomic analysis. Brief. Funct. Genom. Proteom. 7 (5), 329–339.
Wilm, M., 2009. Quantitative proteomics in biological research. Proteomics 9 (20), 4590–4605.

Relevant Websites

http://www.coxdocs.org/doku.php?id=maxquant:start
MaxQuant.
http://www.openms.de/tutorials/
OpenSWATH.
http://www.coxdocs.org/doku.php?id=perseus:start
Perseus.
http://www.ebi.ac.uk/pride/archive/projects/PXD001010
PRIDE Archive - PXD001010.
https://skyline.ms/wiki/home/software/Skyline/page.view?name=tutorials
Skyline.

Biographical Sketch

Zainab is a PhD student in the department of Chemistry and Biomolecular sciences (Bioinformatics group) at Macquarie University. She has been working in the area of protein structure and function analysis during her bachelors and masters' studies. Previously, she worked on identifying novel potent drug-like compounds against histone deacetylases proteins by employing *in silico* drug designing techniques. Currently, she is involved in programming-based mass spectrometry data analysis of data generated through data-dependent acquisition (DDA) and data-independent acquisition (DIA) approaches related to colorectal cancer. More specifically, she is dealing with the customization of DDA based spectral libraries to be used in DIA data analysis for biomarker discovery.

Subash is a PhD student in Cancer proteomics. His main interest is on mass spectrometry based quantitative proteomics. He is currently working on peptide antagonist against specific membrane protein-protein interaction that are known to drive epithelial cancer metastasis.

Author's personal copy

Shoba Ranganathan holds a Chair in Bioinformatics at Macquarie University since 2004. She has held research and academic positions in India, USA, Singapore and Australia as well as a consultancy in industry. Shoba's research addresses several key areas of bioinformatics to understand biological systems using computational approaches. Her group has achieved both experience and expertise in different aspects of computational biology, ranging from metabolites and small molecules to biochemical networks, pathway analysis and computational systems biology. She has authored as well as edited several books as well as contributed several articles to Springer's Encyclopedia of Systems Biology. She is currently the Editor-in-Chief of Elsevier's Encyclopedia of Bioinformatics and Computational Biology as well as the Bioinformatics Section Editor for Elsevier's Reference Module in Life Sciences.

Abidali is a Lecturer at Macquarie University in the Faculty of Science and Engineering. He completed his PhD in 2010 studying, using proteomics, the effects of single mutations in mouse model on brain development in the autism spectrum disorder, Rett syndrome. He then returned to Macquarie University undertaking a post-doc to study the biology of metastasis in colorectal cancer using quantitative proteomics and to determine novel therapeutic targets and develop novel technologies to examine the plasma proteome.

Chapter 3: Role of uPAR in Colorectal Cancer

3.1 Overview

Significant cancer-related deaths associated with substantial financial and social burden have made the development of cancer management strategies a research priority. Establishment of novel prognostic and diagnostic biomarkers that abrogate cancer metastatic progression and relapse is a prominent research challenge.

Most current prognostic markers are only appropriate in late-stage cancer settings when treatment options for aggressive late-stage metastatic cancers are not as effective, as the tumour cells have already metastasised to distal organs. Late-stage cancer metastasis can be the primary cause of cancer-related deaths [1]. However, at early cancer stages (e.g., American Joint Committee on Cancer (AJCC) stages I/II), a combination of surgical resection with chemo-, radio-, targeted- and immune-therapies are mostly curative as the tumours are confined within its site of origin (i.e. histologically with no metastasis).

Our research team at Macquarie University has an interest in the study of the biologies that drive CRC metastasis, especially with regards to contributions made by different proteases, growth factors and integrins in CRC pathobiology.

Our team have made an extensive contribution over the last decade in deciphering the role of the urokinase plasminogen activator receptor (uPAR) and uPAR interactome components in CRC metastasis; including uPAR, urokinase plasminogen activator (uPA), and most recently integrin αvβ6 and transforming growth factor β1 (TGFβ1).

This chapter specifically explores the role/s of uPAR in CRC metastatic progression. uPAR is a pleiotropic cell surface receptor protein known to be overexpressed during inflammation, wound healing and in many epithelial cancers (carcinomas). uPAR can only actuate signalling cascades through interaction with a reported range of multiple lateral partners, primarily because its glycophosphatidylinositol (GPI) membrane anchor lacks both transmembrane and intracellular protein domains [2].

uPAR initiates several hallmarks of cancer (HoC) by activation of extracellular proteolytic cascade modulated through the plasminogen activation system (PAS) or non-proteolytic mechanisms after interactions with membrane proteins like a variety of integrins and extracellular proteins like vitronectin (Vn) [2].

To identify uPAR ligands, their respective binding sites on uPAR and the biologies driven by these interactions, I performed a systematic analysis of human uPAR interactome based on the published literature. Analysis of the binding site location to identify competitive ligands was also performed. Classification of the known uPAR ligands based on the method utilized to decipher the interaction provided different levels of confidence regarding reported uPAR interactors. The literature identifies 111 potential binary uPAR interactions, among which 12 interactions contained specific data-binding site location that could be superimposed on the uPAR structure. Integrin signalling was one of the most significant pathways that emerge upon this type of enrichment analysis of possible uPAR interactomes.

In a collaborative study with other members of our team, I further investigated the systematic effects of decreased uPAR expression levels (↓ by ~43%) on the CRC uPAR-

driven metastasome with the aid of comprehensive membrane proteome analysis of a stable anti-sense uPAR CRC cell model (modified cell line called HCT116^{ASuPAR}) [3]. Pathway analysis with Ingenuity pathway analysis (IPA) [4] and a recently released Cancer Hallmarks Analytics Tool (CHAT) [5] showed that the decreases in the level of the uPAR protein suppressed many metastasis-related pathways illuminated through analysis of several HoCs.

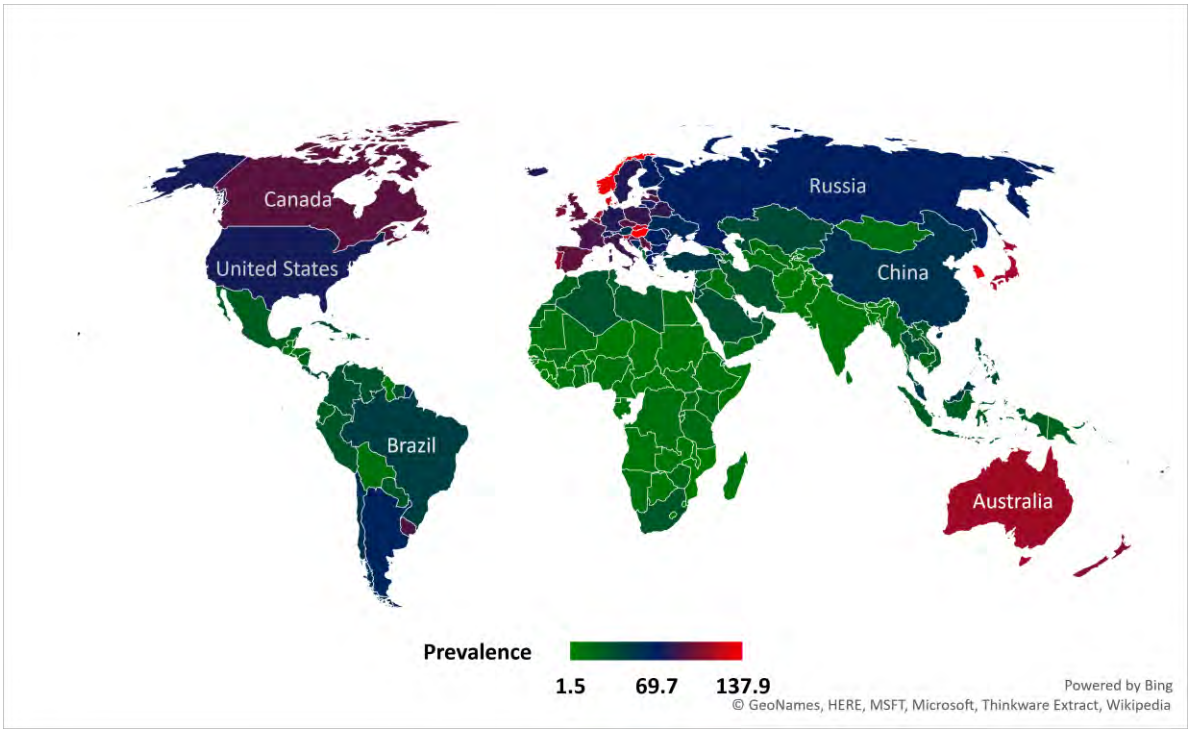


Figure 3.1: CRC Age-standardized prevalence rate during 2018. Hungary had the highest prevalence rate at 137.9 per 100,000, followed by Norway at 132 per 100,000. Australian prevalence rate stands at 111.9 per 100,000. CRC incidence and prevalence pattern aligns with the HDI, the role of meal selection and sedentary lifestyle has been linked to the CRC incidence. CRC mortality rates have decreased over the last decade. However, CRC remains the most significant contributor to cancer-related deaths due to the lack of early-stage diagnostic and late-stage prognostic markers. The data was obtained from GLOBOCAN 2018 [6], IARC (http://gco.iarc.fr/today), World Health

Organization. Permanent link for the data is available at https://tinyurl.com/CRC-prevalence-2018.

3.2 CRC Incidence, Prevalence and Mortality

Globally, 1.8 million new cases of colorectal cancer (CRC) were reported in 2018. Of these, 75% of patients are predicted to die due to the disease within five years. CRC results in the second-highest cancer-related deaths in Australian men and women. Australians have an elevated risk of being diagnosed with CRC, which currently stands at 10% in men and 6% in women respectively by the age of 85 [7]. An age-standardized rate for the same year for the Australian population stands at 36.9 per 100,000 of population [6] (Fig 3.1).

High CRC prevalence is observed in selected European countries, Korea, Japan, Australia and Canada, as demonstrated in Figure 1. CRC prevalence pattern aligns with the human development index (HDI). Hungary has the highest CRC prevalence rate at 137.9 per 100,000 population, followed by Norway at 132 per 100,000 population. CRC prevalence is lowest in Gambia and Guinea at 1.5 and 2.6 per 100,000 population respectively. Australian CRC prevalence rate stands at 111.9 for the year 2018. CRC incidence and mortality rates coincide with prevalence rates (Fig 3.2).

Australia has the highest age-matched CRC incidence and mortality rates. Average incidence and mortality rate between male and female population during 2018 was 36.9 and 11.2 per 100,000 population respectively. Data (Fig 3.2) shows that men have higher CRC incidence and mortality compared to women. The Australian CRC mortality rate has decreased over the last decade, aided by the rise in bowel screening and improved CRC management strategies. The Australian government has initiated a National Bowel

Cancer Screening Program (NBCSP) that provides a complimentary bowel screening for people without CRC symptoms starting at the age of 50 through to age 75. The screening was based on the FOBT test (now on FIT) followed up by confirmatory or exclusion colonoscopy on any replicated positive FOBT test. By the year 2020, the program aims to screen eligible Australians aged 50-74 every two years, which could potentially save 500 Australian lives annually [8].

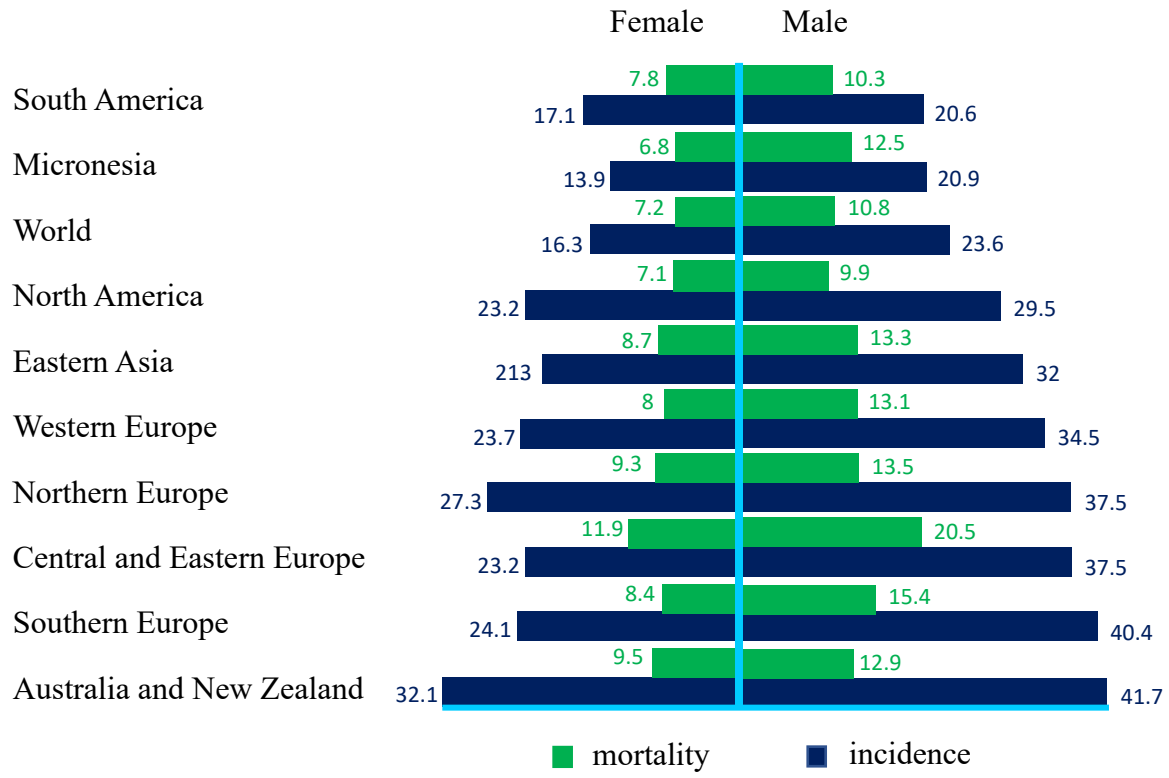


Figure 3.2: Worldwide age-matched CRC mortality and incidence rates per 100,000 population during 2018. This data was obtained from GLOBOCAN 2018 [6], IARC (http://gco.iarc.fr/today), World Health Organization. Permanent link for the data is available at: https://tinyurl.com/CRC-incidence-and-mortality.

3.2 CRC staging

Stratification of cancer into distinct clinical AJCC stages (or other older staging systems e.g., Dukes') facilitates the efficient delivery of cancer management strategies. Accurate staging is crucial for the implementation of efficient treatment options. Definition of the universal staging system streamlines the worldwide efforts for the development of novel prognostic and diagnostic CRC markers. American Joint Committee on Cancer (AJCC) classification is the latest iteration of the CRC clinical and pathological staging. AJCC (also referred to as TNM based staging) considers several parameters for CRC stage classification, including;

i) size and levels of tumor spread (T),

ii) presence of the tumors in the local lymph nodes (N), and

iii) the severity of distal metastasis on the distal organ (M) [9].

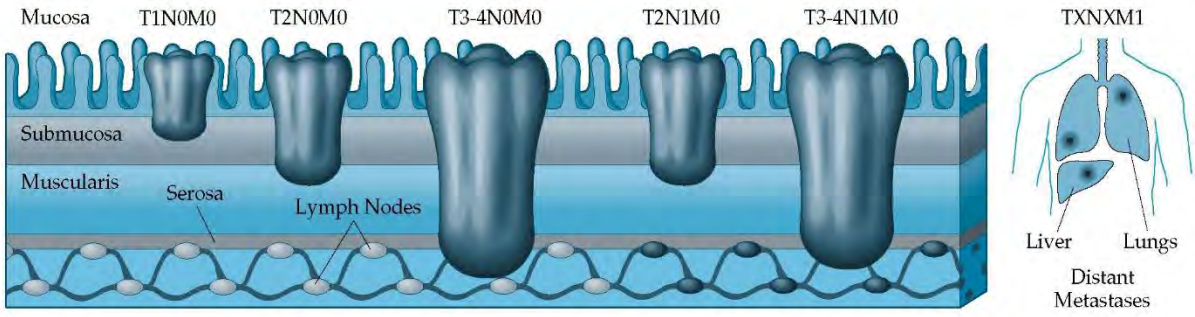
The AJCC system replaced the previous Dukes' staging CRC system. Most tissue biobanks previously have relied on Dukes' staging. Dukes' classified CRC into four stages (A, B, C and D). The Dukes' staging system stratified CRC tumors that were;

i. confined tumor within mucosal or sub-mucosal walls,

ii. tumor penetration through the mucosa to muscularis propria,

iii. tumor spread into the lymph node, and

iv. distal metastasis [10].



TNM Stage	*Description*
T1N0M0	Infiltration no deeper than submucosa
T2N0M0	Infiltration of muscularis; no penetration through colonic wall; no lymph node involvement
T3-4N0M0	Extension through colonic wall; no lymph node involvement
T2N1M0	Infiltration of muscularis; no penetration through colonic wall; lymph node involvement
T3-4N1M0	Extension through colonic wall; lymph node involvement
TXNXM1	Distant metastases

Figure 3.3: AJCC-TNM CRC staging: CRC is classified into distinct stages according to the severity of the CRC metastasis based on T-N-M system. The staging accounts for tumor size (T), diffusion to the lymph nodes (N) and distal metastasis (M) during staging. The staging system provides a premise for treatment strategies, survival rate and a measure of the tumor spread from its site of origin. TNM classification is crucial in deciding the appropriate treatment options. Figure reproduced from https://tinyurl.com/TNMstaging-SA .

3.3 CRC screening

Late-stage diagnosis is the primary factor in CRC-related deaths. Surgical resection, in combination with adjuvant therapies, is curative for most early-stage patients, with a 5-year survival rate above 90%. However, the 5-year survival rate falls to near 10% when diagnosed at later stages [11]. CRC diagnostic approach primarily employs biochemical tests and visual examination. The faecal occult blood test (FOBT) was one of the first biochemical diagnostic tests for CRC. It measures traces of blood haemoglobin (Hb) in a patient stool sample. The test utilizes the detection of the peroxidase activity between the Hb-derived heme group and guaiac chemical groups. However, the test may report false-positive results in the presence of non-CRC related upper gastrointestinal (GI) tract bleeding, recta/anal anatomical lesions (e.g., haemorrhoids) and/or the presence of dietary peroxidases and antioxidants [12]. The sensitivity of the FOBT test has been found to range between 12.9% and 79.4%, whereas the specificity ranges between 87% and 98% [13]. The immunoassay-based fecal immunochemical test (FIT) has enhanced screening specificity, compared to FOBT primarily due to the reliance on an antibody that is highly specific to Hb. FIT can detect bleeding from the lower GI tracts associated with CRC and non-neoplastic lesions [14]. The FIT test is now more commonly utilized in CRC screening than older FOBT tests [13]. Sensitivity and specificity of FIT are reported to be 79% and 86% respectively, based on 113,360 patients, including 437 patients with colonoscopy confirmed CRC [15,13]. The recent introduction of reformulated multitarget stool DNA tests (MT-sDNA) that measures multiple CRC mutations has been reported to possess a sensitivity range between 53% and 73 %. The MT-sDNA technique can achieve higher sensitivity than FIT but suffers from lower specificity [16].

Diagnostic visual examination of CRC involves the identification of polyps and CRC lesions. Colonoscopy is the technique of choice for positive CRC diagnosis for patients positive on a FIT test. The sensitivity of the procedure in the hands of a trained medical professional is up to 96 % on lesions less than 6 mm and the sensitivity increases to 98% for lesions of size greater than 1 cm [17]. Colonoscopy fails to identify lesions located on or around the sharp turns such as anatomical hepatic and sigmoid flexure and flat adenomas. Despite the excellent specificity, unease to patients before (bowel cleaning), during (sedation, potential bleeding associated with bowel rupture) and after (recovery taking some days) the procedure may enhance CRC co-morbidity [17]. When colonoscopy is not feasible, computerized tomographic colonography (CTC) is recommended for diagnostic purposes. However, the procedure has limited specificity of up to 75% when compared to colonoscopy [18]. Surgical resection of adenomas is the most efficient method for CRC management. However, patients with advanced aggressive carcinoma require a combination of surgical resection and systemic therapy to enhance survival [19]. Combination of colonoscopy and surgical resection can be curative during initial stages I/II of CRC.

3.4 CRC metastasome

Genetics plays a key role in CRC inception and progression. The biologies behind cancer progression can be summarized into a common framework termed as HoCs, introduced and recently modified by Hanahan and Weinberg [20,21]. The ten HoCs are;

i. sustaining proliferative signalling,
ii. evading growth suppressors,
iii. avoiding immune destruction,
iv. enabling replicative immortality,
v. tumour-promoting inflammation,
vi. activating invasion and metastasis,
vii. inducing angiogenesis,
viii. genome stability and mutation,
ix. resisting cell death and
x. deregulating cellular energetics

These HoCs recapitulate characteristic phenotypes involved in cancer progression and are based upon a combination of patho-phenotypic and genomics perspectives. HoCs provides a rational framework for understanding the various biologies behind the progression of neoplastic disease, including CRC [21]. Multiple mutations enable CRC metastasis to gain and retain cellular proliferative ability (e.g., RAF, RAS, MYC), elude growth suppression (e.g., RB, TGFβ, TP53), oppose apoptosis (e.g., Bcl-2 family, TP53), enable perpetual replication (e.g., telomerase, TERT), enhanced angiogenesis (e.g., TSP, VEGF) and metastasis followed by the migration and invasion (e.g., β-catenin, MMPs, uPAR, BCF4) [20,21].

Patterns of somatic mutation involved in the transformation and progression of hyperplasia to metastatic carcinoma have already established, few of those include mutations in adenomatous polyposis coli (APC), RAS-family genes and TP53 [22]. CRC can also be associated with pathogenic germline variants in CRC oncogenes. Germline mutation in DNA mismatch repair genes is evident in non-polyposis and APC [23]. Identification of key germline and somatic mutations have proven helpful in improving our understanding of carcinogenesis and will allow us to better design and develop diagnostic, therapeutic and preventive strategies against CRC [22].

Sequential events leading to the transformation of the colon mucosa towards adenocarcinoma have been well documented [24] and are linked to alterations in key genes like epidermal growth factor receptor (EGFR), p53, mitogen-activated protein kinases (MAPK), TGFβ and WNT signalling pathways [25]. Many of these pathways are responsible for maintaining colon homeostasis and/or are involved in colon epithelial cell oncogenic transformation [26].

Around 60-80% of CRC patients express EGFR, where overexpression usually reflects poor prognosis along with aggressive histological and clinical phenotypes [27]. Hyperactivation of the WNT pathway is believed to be a key oncogenic driver in most CRC [28]. Mutations in transcriptional regulator β-catenin or APC are thought to initiate WNT signalling. In healthy cells, β-catenin is controlled by a protein complex containing APC, disruption of this complex by the WNT receptors leads to the downstream activation of WNT target genes [24].

MAPK is a serine/threonine kinase involved in regulating cell proliferation through the activation of protooncogenes and growth factors. Extracellular signal-regulated kinase

(ERK) cascade is involved in maintaining homeostasis, crucial multiple growth factors and oncogenes promote growth and differentiation through this pathway. Three major MAPK subfamilies mediate MAPK signalling; i) extracellularly signalling kinases (ERK and RAS/Raf1/MEK/ERK), ii) c-Jun N-terminal kinase (JNK) and iii) MAPK14 [29]. Ras/Raf/MEK/ERK regulates cell survival and invasion and is usually deregulated in about 30% of the CRC. Activation of this signal cascade enables cells to acquire the ability to differentiate and migrate [30]. MEK followed by ERK activation is induced by protein kinase C (PKC) or Ras triggering the ERK pathway [29].

TGFβ signalling mediates multiple cellular processes including cell adhesion, differentiation, proliferation and migration [31,32]. Mutational inactivation of the TGFβ signalling is observed during CRC progression through inactivation of TGFβ receptors (*TGF*β*R1* and *TGF*β*R2*) or inactivation of SMAD regulators (SMAD*4, SMAD2* and *SMAD3)*. Reversal of the TGFβ activity inhibits cell proliferation and tumorigenicity, indicating its tumor suppressor roles [32]. TGFβ signalling also acts as a negative regulator of the metastasis through conditioning stromal cell [33]. However, elevated levels of TGFβ in serum is associated with poor patient outcome in clinical settings. TGFβ signalling in CRC is 'good early and bad late'. In epithelial cancers including CRC that allow TGFβ expression, it induces epithelial-mesenchymal transition (EMT) and signalling changes to switch genes that are responsible for invasion and migration [33].

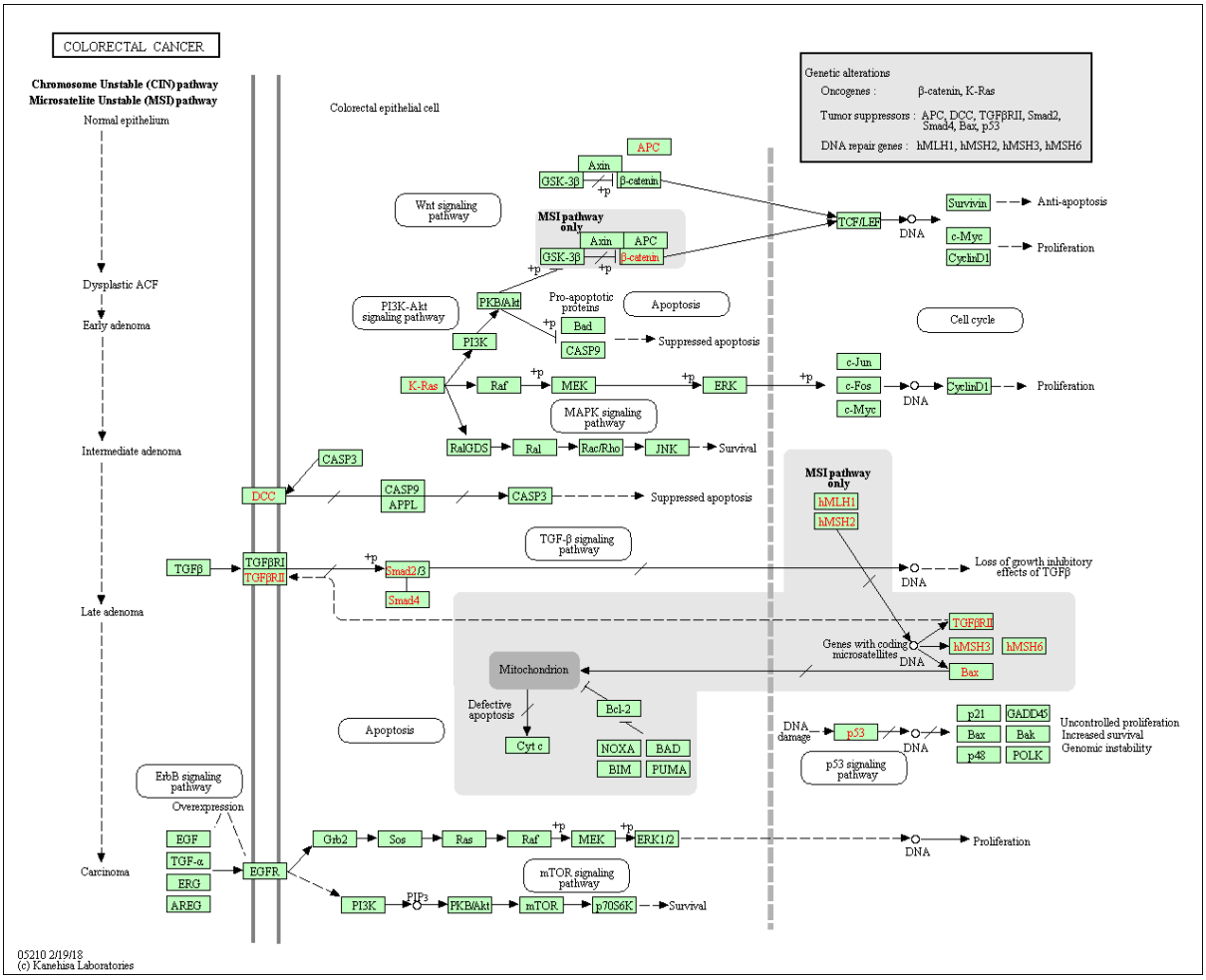


Figure 3.4. CRC gene signalling network: CRC is characterized by multiple gene mutations in CRC oncogenes and tumor suppressor genes. Multiple genomic instabilities are reported in CRC metastasis, primarily driven by microsatellite Instability (MSI), chromosome instability (CIN) and chromosome translocations (CT). MSI and CIN mutations result in alterations of WNT, EGFR, TGFβ, prostaglandins, epithelial cadherin signalling which facilitates the CRC progression [34–36]. Figure reproduced from KEGG [37] colorectal cancer gene signalling network (hsa05210) weblink https://www.genome.jp/kegg/pathway/hsa/hsa05210.html.

Of many biologies driving CRC metastasis, the role of proteolytic systems in general, including the plasminogen activation system (PAS) proteolytic cascade is implicated in driving CRC metastasis [38,39]. uPAR is a multidomain protein anchored to the cell membrane via a glycosylphosphatidylinositol (GPI) anchor and plays a central role in the PAS dependant proteolytic cascade and is considered a “systems organizer” with roles in activation plasminogen, growth factors and pro-metalloproteases (pro-MMP). These have all been implicated in epithelial cell proliferation, invasion and migration [40]. Elevated uPAR expression levels are associated with a poor prognosis in most epithelial cancer [41]. Involvement and overexpression of uPAR in late-stage metastasis have attracted research interest, particularly in the development of uPAR-based diagnostic imaging and prognostic targets for late-stage CRC treatment. uPAR is a potential target in late-stage metastasis of many epithelial cancers including CRC [42,43].

3.5 Role of uPAR in CRC metastasis

3.5.1 uPAR interactome analysis to identify uPAR-ligand binding sites and their interaction confidence levels: Manuscript 4

uPAR is a cell surface protein which functions as a receptor for serine protease uPA. Binding of twin chain uPA to uPAR in the proximity of abundant circulating or cell-bound plasminogen allows activation of plasminogen to plasmin, which in turn mediates extracellular matrix (ECM) degradation. uPAR expression is elevated during inflammation and wound healing [40]. Epithelial carcinomas are characterized by overexpression of uPAR, and it has been implicated in cancer proliferation, invasion, adhesion, migration and subsequent metastasis.

uPAR can modulate a range of intracellular signalling upon interaction with multiple lateral partners. CRC mutational landscape significantly overlaps with the downstream uPAR signalling [2].

Identification of uPAR ligands and specific binding sites on uPAR allows exploration of novel therapeutic avenues for antagonism of signaling derived from various uPAR interactome/s. Antagonization of the specific uPAR-ligand interaction or alteration of expression levels of the key molecules involved in metastatic signalling has the potential to abrogate metastasis.

We have previously demonstrated that transfection of the CRC cell model SW480 cells with vector inducing integrin αvβ6 increased cellular proliferation and invasion, compared to cells containing an empty vector [44]. SW480 cells lack endogenous αvβ6 expression. Similarly, decreasing expression levels of uPAR on CRC cell models has been shown to inhibit multiple cancer metastatic phenotypes (Section 3.5.2: Supplemental manuscript 1).

Understanding uPAR signalling and elucidating roles in the CRC metastasis requires a thorough investigation of uPAR interactomes. To identify uPAR ligands and their respective binding sites on the uPAR, I performed a systematic text-mining study, to extract uPAR protein-protein interaction data from publicly available resources. Analysis of uPAR interactomes provided some information on uPAR binding “hotspots” involved in uPAR downstream signalling. Confidence in any mentioned uPAR-ligand interaction was measured based on the methods utilized to deduce interactions. Interactions identified by high stringent methods, such as crystallography were provided higher scores compared to low stringent methods such as genomics. The analysis generated a comprehensive

uPAR interactome with 111 potential binary interactions, among which specific residue binding sites for 12 interactions were mapped to the level of uPAR crystal structure surfaces.

Citations on the manuscript are formatted as a part of the manuscript.

3.5.2 Review: The uPAR Interactome: Identification of uPAR-ligand binding sites and analysis of interactome confidence levels

Subash Adhikari[1], David Cantor[2], Seong Beom Ahn[1], Abidali Mohamedali[3], Edouard C. Nice[4] and Mark S. Baker[1*]

[1]Department of Biomedical Sciences, Faculty of Medicine and Health Science, [2]Australian Proteome Analysis Facility, [3]Department of Molecular Sciences, Faculty of Science and Engineering, Macquarie University, Sydney, Australia. [4]Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia.

* Corresponding Author: Mark S. Baker, Level 1, 75 Talavera Road, Macquarie University, 2109, Australia, +61 2 9850 8211, mark.baker@mq.edu.au

Abstract

Urokinase plasminogen activation receptor (uPAR) is a pleiotropic glycosylphosphatidylinositol (GPI)-anchored, cell surface receptor capable of binding inactive single-chain urokinase plasminogen activator (sc-uPA) or active twin-chain (tc-uPA). It drives several hallmarks of cancer (HoC) that were originally identified by cancer researchers Douglas Hanahan and Robert Weinberg in cell and recently revised by the same authors. These changes in HoC biologies mediated by uPAR are achieved through;

i) activating the plasminogen activation system (PAS) and causing substrate proteolysis (e.g., extracellular matrix (ECM) proteins, zymogen growth factors) through either plasmin and additional downstream plasmin-activatable proteases (e.g., matrix metalloproteases; MMPs), and/or

ii) non-proteolytic process (e.g., signal transduction) upon interaction with various lateral partners, including integrins and vitronectin (Vn).

The level of uPAR expression is enhanced during inflammation, tissue remodelling and in epithelial carcinoma. Abnormal spatiotemporal uPAR expression has shown to indicate poor prognosis in cancer.

uPAR lacks transmembrane or intracellular domains and thus requires interaction with alternative lateral plasma membrane components (i.e., other proteins) to exert effects on downstream signalling. The ability of the uPAR to interact with major membrane protein families, (e.g., integrins, G-protein-coupled receptors (GPCRs) and receptor tyrosine kinases (RTKs)) give rise to an extensive uPAR interactome, that is involved across multiple biological pathways/systems. Protein-protein interaction (PPi) analysis of the uPAR interactome based on the open-source PPI database revealed 111 potential binary interaction partners. Re-classification of this set of proposed interactions based upon the stringency of particular technologies utilized for uPAR interaction inference was undertaken in order to score likely confidence in any particular uPAR-protein-ligand interaction. Among 111 potential binary interactions, the amino acid primary sequence residue responsible for uPAR binding was identifiable for 12 ligands - which allowed mapping of binding “hotspots” on the uPAR structure.

Introduction

Most cancer patients with aggressive carcinomas face a terminal illness driven by distal metastasis. This leads to reduced survival rates when compared to earlier benign stages of the disease. Most late-stage patients succumb to their disease due to the lack of late-stage clinical treatment options [1].

Initiation of primary tumor motility from their site of origin to distal metastasis is a multi-step process, referred to as the "invasion-metastasis cascade" [1]. These sequences of events can be characterized by the tumor cells acquiring metastatic phenotypes such as; an ability to proliferate, invade and eventually migrate to distal locations [2,3]. Neoplastic epithelial cells degrade the extracellular matrix (ECM) as the epithelial-mesenchymal transition (EMT) drives cells towards increased motility and invasiveness. EMT signalling is triggered and regulated by the heterotypic signal originating from the tumor stroma [4]. EMT degradation is a representative process in embryonic development, fibrosis, wound healing, regeneration, and in multiple cancer phenotypes [5]. Loss of ECM integrity upon cleavage of cell-cell contact, change in

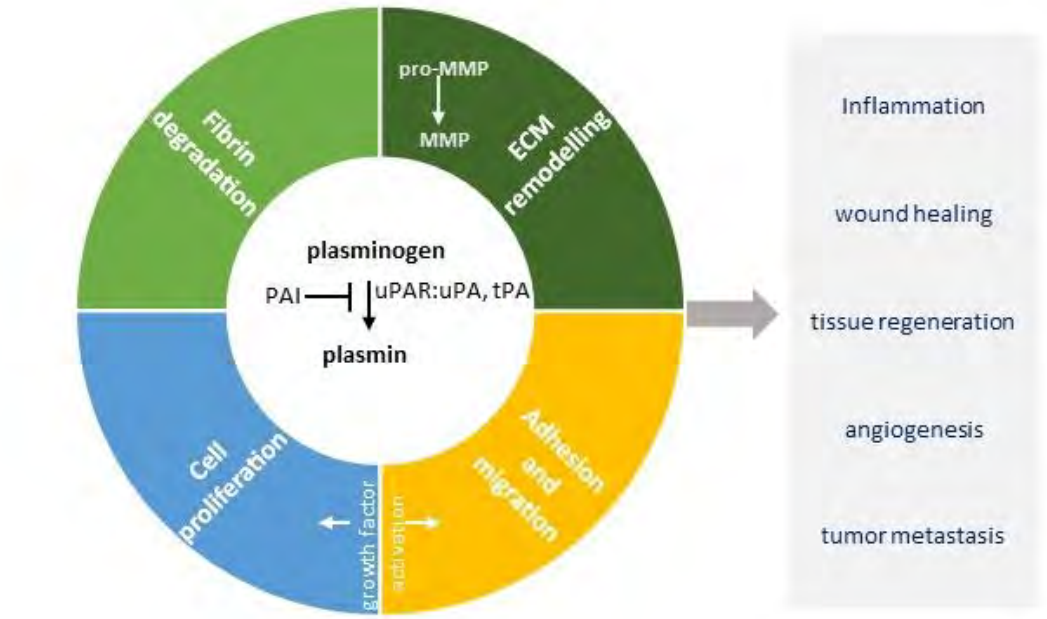


Figure 1: Components of the PAS. Binding of the uPA-uPAR induces plasmin activation along the cell membrane which activates matrix metalloproteases (MMPs) and induce the release of growth factors. Plasmin induces tissue proteolysis via fibrin activation and ECM degradation. PAS-dependent cell proliferation, invasion, adhesion and migration is modulated by the interaction of uPAR with its multiple ligands, whereas plasmin activation is inhibited upon binding of plasminogen activator inhibitor 1 (PAI-1) or PAI-2 to the active tc-uPA-uPAR complex. Regulation of the PAS is modulated by internalization and redistribution of the uPAR-uPA-PAI-1 complex by low-density lipoprotein (LDL) receptor-related protein (LRP). Collectively, PAS system modulates ECM proteolysis, cellular proliferation and migration, tumor signalling, invasion and metastasis. Apical-basal polarity and reorganisation of actin, intermediate filaments and tubulin cytoskeleton network imparts migratory and invasive behaviour in tumor cells [6]. Tumor stromal components; fibroblasts, chemokines, growth factors and host tissue lymphoid cells

mediate EMT through interleukins, transforming growth factor-beta (TGFβ) and Wnt signalling. Distal metastasis is facilitated by circulating tumor cells invading into the vasculature [4].

The urokinase plasminogen activator (uPA)-driven PAS is one of the primary factors driving the EMT during cancer metastasis [7,8]. In detail, PAS is composed of the serine protease urokinase plasminogen uPA, uPA receptor (uPAR), the plasminogen activator inhibitor serine protease inhibitors (SERPINS) and PAIs. PAS converts plasminogen to plasmin and aids in subsequent activation of various MMPs and growth factors. The process is mediated by tPA and uPA. tPA and uPA are endogenously inhibited by the PAI-1 and PAI-2, whereas the SERPINS α2-antiplasmin and α2 -macroglobulin inhibit plasmin activation [9–11].

uPAR

uPAR is a 335 amino acid (AA) long cell membrane receptor [12] encoded on chromosome 19q13 [13]. It is constituted by three extracellular domains named as D1, D2 and D3 and where D2-D3 is interconnected by a linker region [12]. The right-handed orientation of a three-fingered glove-shaped uPAR generates a globular receptor with an opening between the D1 and D3 domains creating a central cavity suitable for lateral interactions with other ligands [14]. The C- terminus of uPAR D3 anchors uPAR to the plasma membrane by a GPI moiety, which determines the uPAR localisation and its conformation in the cell plasma membrane [15].

uPAR GPI anchor can be cleaved by phosphatidylinositol-specific phospholipase D, plasmin and cathepsin, releasing soluble forms of uPAR (suPAR) into blood plasma [15].

The presence of suPAR in plasma allows for non-invasive detection of uPAR from the blood, where levels can act as a surrogates activation of the PAS system [16]. Given the intricate role of uPAR in PAS, the study of uPAR signalling and interruption of processes associated with the PAS proteolytic cascade may provide novel avenues for cancer therapeutic intervention [17].

The lack of transmembrane and cytoplasmic domains necessitates uPAR to establish interaction with lateral partners such as integrins and Vn for the process that do not depend upon uPA-induced PAS activation. The uPAR crystal structure reveals that it is capable of simultaneous interaction with the uPA ATF domain along with other ligands capable of facilitating downstream signalling [14].

uPAR-based drug targets

Due to the intricate role of uPAR in modulating the multiple HoC, uPAR interactome can be potentially targeted to achieve anti-metastatic effects [8,18]. Modulating the levels of the uPAR with the aid of RNA antisense technologies has been employed to quantitate the systematic effects of uPAR on invasion and metastasis. Rreduced uPAR levels on colon cancer cell line models negate multiple phenotypic characteristics associated with cancer metastasis [19,20]. Likewise, a decrease in uPAR expression using antisense transfection resulted in decreased pulmonary metastasis in mouse models and inhibition of tumorigenesis in human glioblastoma cell lines [22-21]. Similarly, RNA interference (RNAi) based on short interfering RNAs (siRNA) resulted in a reduction of pre-established prostate tumors in mice with no evidence of secondary tumors [22].

The establishment of uPAR as “systems organizer” has led to the development of anti-uPAR agents targeting the uPAR interactome. Many developed anti-uPAR agents are based on small molecules, antibodies and/or peptides [23] that antagonize a particular uPAR-ligand interaction. Monoclonal antibodies (mABs) targeting uPAR have been characterized. For example, the ATN-658 antibody is targeted against integrin αM and it can inhibit migratory and invasive behaviour of tumor cells *in vitro* and demonstrates strong anti-tumour activity in solid tumor xenografts. ATN-658 inhibits many uPAR-derived signals involved in regulating proliferation and metastasis [24]. Utilization of small molecules targeting the uPAR interactome has proved equally popular. The inhibition of extracellular-signal-regulated kinase (ERK) activity upon antagonization of uPAR-α5β1 was achieved with 2-(Pyridin-2-ylamino)- quinolin-8-ol and 2,2'-(methylimino) di (8-quinolinol), a molecule generated computationally using a molecular docking strategy [25]. Equally, small molecule antagonism of the uPA-uPAR interaction using IPR-456 inhibits both tumour cell migration and invasion [26].

Apart from roles as a prognostic agent, overexpression of uPAR at cancer invasive fronts allows it to be a diagnostic imaging target – where it can be employed to identify the location, severity and aggressiveness of tumor lesions [27]. For example, multiple positron emission tomography (PET) targets have been developed against uPAR with either Cu^{64} [28], Ga^{68} [29] and F^{18} [30] probes conjugated to uPAR-specific peptides.

Method

A meta-analysis of uPAR interactomes

The range of uPAR PPIs was queried with the search term “Q03405” in BioGRID [31], IntAct [32] and STRING [33] databases through the PSICQIC platform [34]. This search retrieved 111 unique interactions. Interactions were then annotated based upon the type/quality/stringency of the protein-protein interaction evidence and were used to assign a score for uPAR interaction confidence. Interactions identified from detailed structural studies (e.g., those using x-ray crystallography and/or NMR) were ranked higher compared to lower confidence methods (e.g., genomics and/or co-cellular or tissue localization studies). The details of all methods and scoring criteria are presented in Supplementary file S 3.1, whereas binding site data for 12 interactors are present in Supplementary file S 3.2. uPAR sequence alignment was performed against identified uPAR binding site regions.

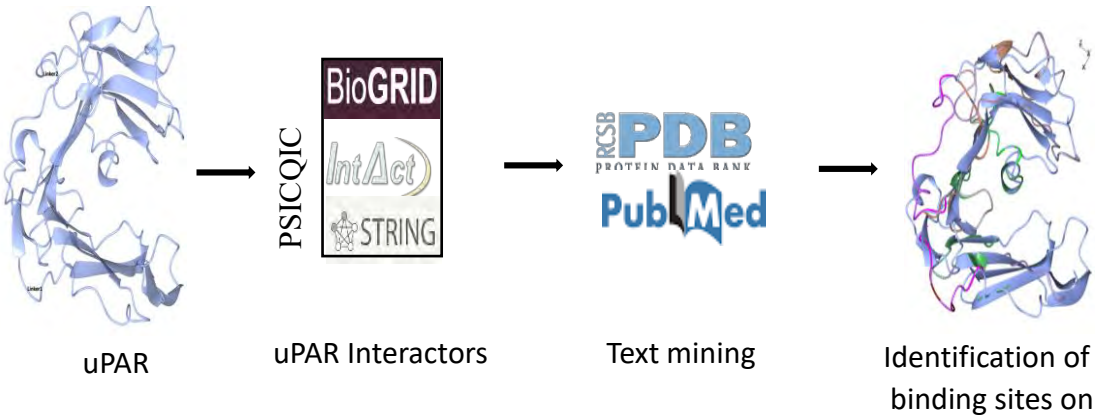


Figure 2: Method adopted for identification of uPAR interactors and their respective binding sites on uPAR. uPAR interactors were extracted and consolidated from open-source BioGrid, IntAct and STRING PPI database via the PSICQUIC platform. Text mining was performed to identify the binding sites of uPAR ligands on uPAR.

Results

uPAR binding site locations

Binding site location analysis of uPAR interaction partners indicates potential binding site hotspots along the uPAR sequence. uPA was identified to be the most significant binding partner followed by Vn based on the number of uPAR residues involved in binding to these respective molecules (fig 3.3). uPA and Vn were found to share a common uPAR binding site, indicating potentially competitive inhibition [35]. The uPA-uPAR interaction is known to induce conformational changes in uPAR, leading to the exposure of the SRSRY region responsible for other G-protein-coupled receptors (GPCR) interaction [36]. The SRSRY motif is chemotactically active and has been widely studied in terms of how it affects/induces other uPAR biologies. Vn binding on the SRSRY motif inhibits uPAR-induced GPCR signalling and as a consequence suppresses uPA-induced cell migration [37].

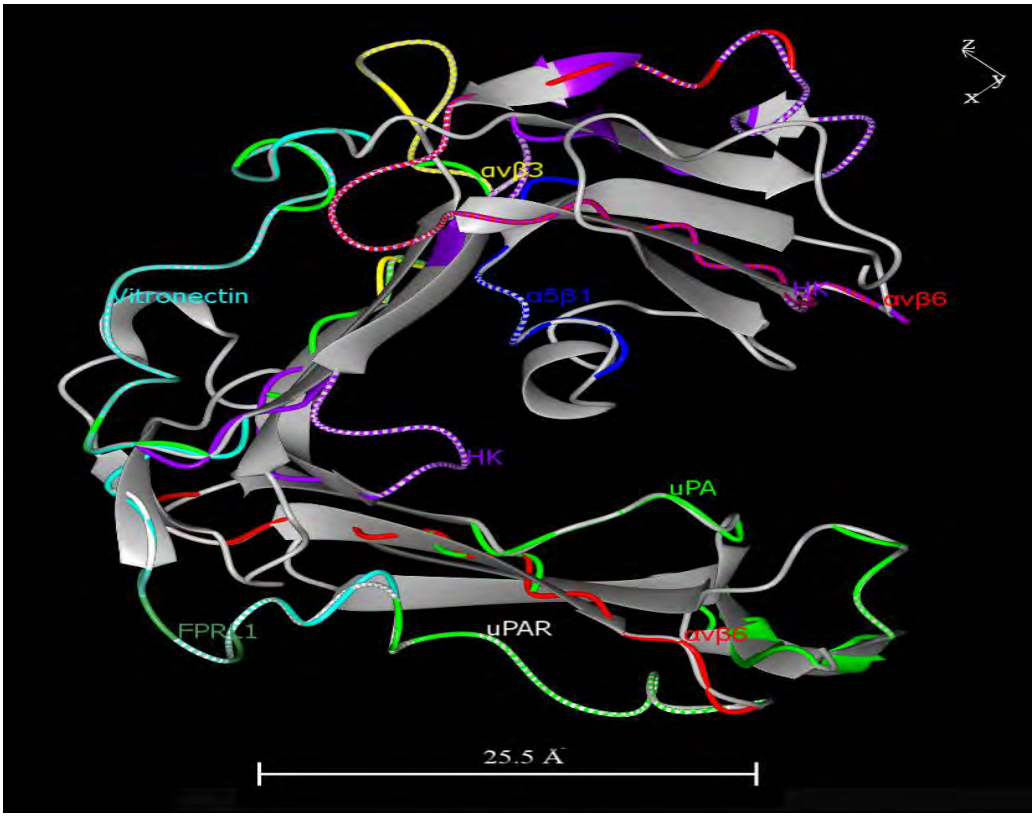


Figure 3: Binding site location of multiple uPAR ligands on uPAR. Sequence alignment of the binding site location on uPAR allows identification of the ligand “hotspots” on the uPAR. Multiple ligands were found to compete with the uPAR chemotactic sequence SRSRY. uPA was identified as the most significant uPAR interaction partner, covering most of the uPAR sequence region.

uPAR ligands

uPAR interacts with multiple cell surface partners [20,38]. In the absence of transmembrane and intracellular domains, uPAR promotes signalling activity via interaction with co-receptor integrins, RTK, GPCRs and multiple effector molecules like Akt, focal adhesion kinase (FAK) and Src [39,40].

uPAR can promote tumorigenesis independent of PAS-dependent proteolysis. uPAR suppression is associated with a decrease in phosphorylation levels of extracellular-signal-regulated kinase (ERK) 1/2, FAK, c-Jun N- terminal kinases (JNK), and p38 mitogen-activated protein kinases (MAPK), each inhibiting cell migration [41]. uPAR downregulation suppresses Notch-1-associated gene expression levels. Notch-1 regulates cancer metastasis through cross-talk with ERK, NF-κB and phosphoinositide 3-kinase (PI3K)-AKT-mTOR pathways. Equally, uPAR promotes the EMT via activation of ERK, PI3K-AKt, Rac1 and Src pathways [42].

ECM proteins, integrins and their ligands like vinculin and FAK are involved in establishing ECM focal contacts. These focal contacts contain multi-protein complexes that form a physical link between the ECM and intracellular actin tubules [43]. The presence of the uPAR on focal contacts corroborates a role, particularly for uPAR-integrin interactions. uPA enhances cytoskeletal induced cell motility changes upon activation of Rac, a Rho family GTPase [44]. Additionally, uPAR has been found to co-immunoprecipitate with actin, α-actinin, Janus kinase (JAK), signal transducers and activators of transcription (STAT), protein kinase C (PKC) and vinculin. In addition, uPAR has been shown to promote cell migration via stimulation of tyrosine kinase 2 (TK2) with subsequent PI3K activation [20].

Low-density lipoprotein (LDL) receptor-related protein (LRP) regulates tumor cell migration via uPA-PAI-1 dependent internalization of uPAR. Similarly, mannose-6-phosphate (M6P) binds to the uPAR D2-D3 fragment that initiates lysosomal-dependent degradation. Subsequently, uPAR recycling allows;

i) the salvation of uPA:uPAR:PAI-1 complexes;

ii) lowering of uPAR availability on the cell surface inhibiting uPAR-dependant proteolysis,

iii) redistribution of principally unused unoccupied uPAR to the cell surface, and

iv) dissociation of various uPAR-ligand interactions, inhibiting associated downstream signalling events [13].

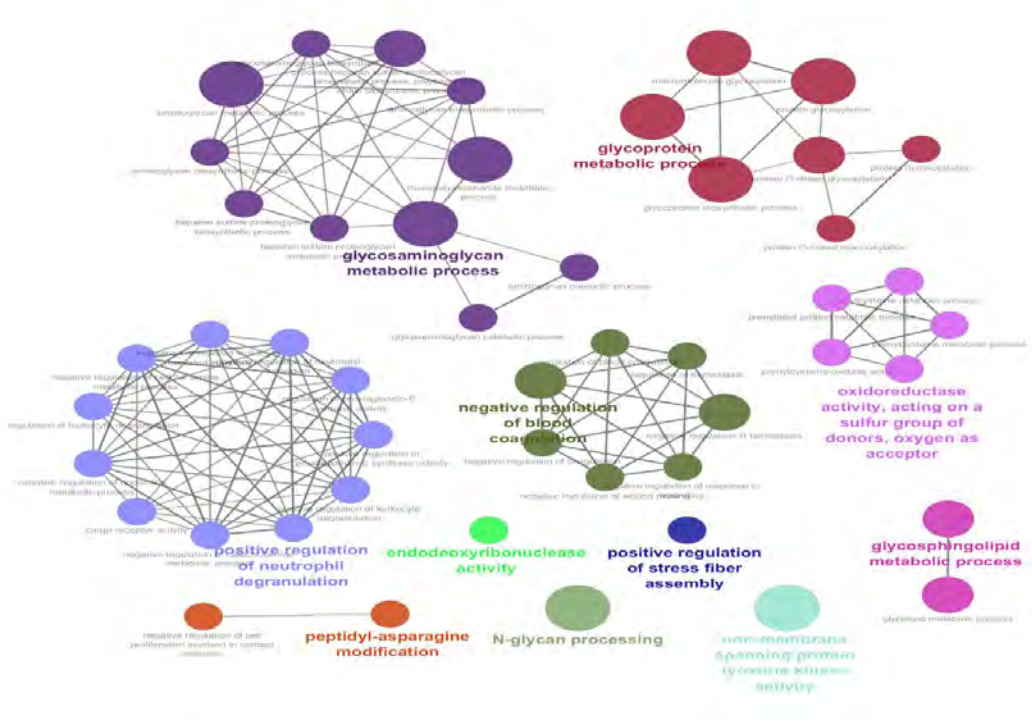


Figure 4: Functional enrichment of the uPAR interactors. uPAR interactors are involved in RTK signalling, multiple metabolic processes, neutrophil degranulation and regulation of proliferation, these processes have known role in inflammation and metastatic cancer progression.

Soluble ligands

uPA

uPA is the most significant uPAR ligand. It is responsible for regulating proteolysis and metastasis [45]. uPA and its tissue type homolog (tPA) activate plasminogen to the serine protease plasmin, and mediate degradation of fibrin and ECM components causing extracellular proteolysis and cell migration [11]. uPAR binds the zymogen, pro-single-chain form of uPA which is converted into active twin-chain uPA at which point it mediates the proteolytic cascade [41].

Plasmin degrades fibrin and other ECM protein components directly or through activation of MMPs [46]. uPA and uPAR interaction initiate PAS and subsequent ECM degradation, and release of growth factors [47]. Components of the PAS (i.e., uPA, uPAR and inhibitor PAI-1) have been implicated in an array of tissue homeostasis processes such as extracellular proteolysis through activation of cell surface plasminogen and enhanced adhesion and signalling via interaction with matrix protein vitronectin [48].

The uPA-uPAR system is inhibited by a series of SERPIN family proteins, including PAIs. PAI-1 induces uPAR recycling through cleavage and internalization of uPA-uPAR in the presence of LRP. LRP binds with both uPAR and epitopes of the relaxed serpin conformation of-PAI-1 [49]. The uPAR induced proteolytic cascade is inhibited upon the formation of PAI1-uPA-uPAR complexes, which are eventually internalized for uPAR recycling and re-distribution to the plasma membrane [20].

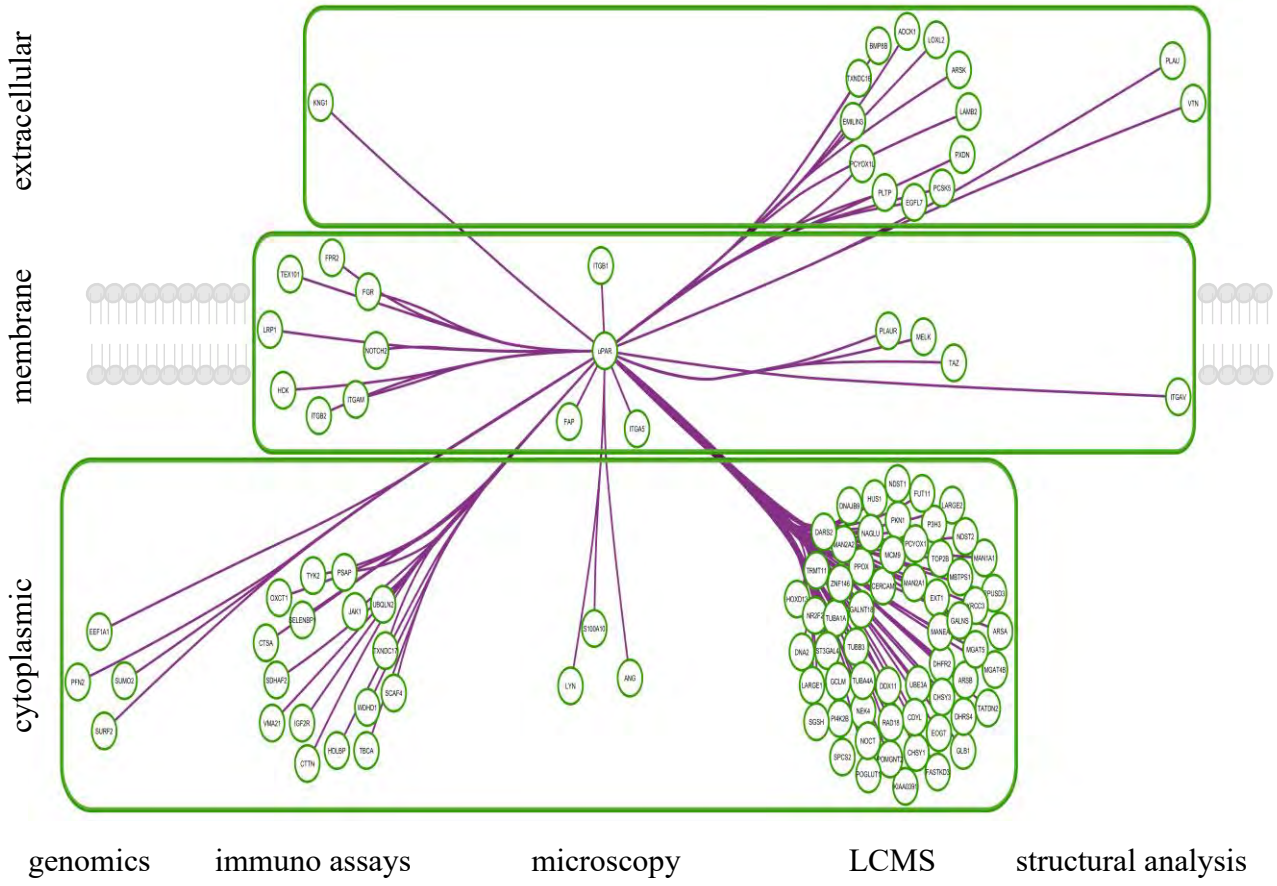


Figure 5a: Classification of uPAR ligands by the subcellular location and uPAR interaction confidence levels. uPAR interactors were derived from BioGRID, IntAct and STRING via PSICQIC platform and clustered into groups based on their subcellular location and methods utilized to infer the interaction. Mass spectrometry (MS) analysis contributed to the identification of most of the uPAR ligands. Most of the identified uPAR ligands were in the cytoplasmic region.

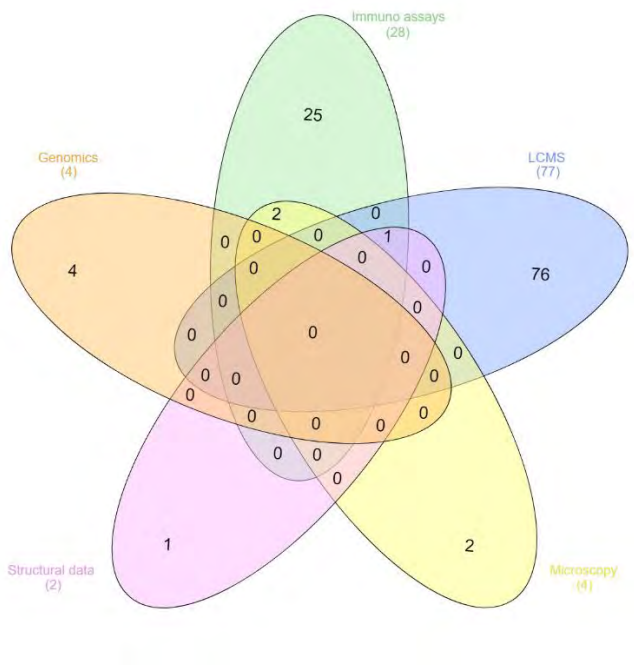


Figure 5b: Overlap of proteins identified across multiple methods. Most of the identified proteins were unique to the method utilized. Mass spectrometry contributed to the identification of the highest number of proteins.

Vitronectin (Vn)

Binding of the Vn SMB domain to uPAR residues has been elucidated because of the availability of a uPAR crystal structure [50]. The direct interaction between uPAR and Vn promotes cell adhesion and tumor growth [51]. Tumor cells that have acquired uPAR induced proliferative phenotypes spread towards the ECM protein Vn and activate integrin-dependent signalling, leading to cytoskeletal reorganization and cell shape changes [49]. uPAR mediated Vn signalling regulates intracellular signalling via MAPK in multiple cell lines [52].

uPAR simultaneously binds with uPA and Vn, phosphorylation of the Vn by the protein kinase casein kinase-2 (CK2) regulates the uPA-induced migration and adhesion to Vn [49]. Binding of VN to PAI-1 inhibits cellular motility, mediated by the antagonization of the Vn-integrins interaction [50]. PAI-1 is thought to modulate the integrin-uPA-uPAR interaction since the affinity of PAI1 for Vn decreases by ~100 fold when it binds with

proteases (e.g., uPA), thereby allowing the interaction between Vn, uPAR and integrins. Additionally, Vn plays a significant role in uPAR-dependent differentiation of monocytes to macrophages, where the adhesion induced by uPAR-Vn binding is essential for subsequent cellular differentiation. PAI-1 suppresses both Vn-dependent adhesion and differentiation processes, whereas uPA acts as an inducer [20].

High molecular weight kininogen

High molecular weight kininogen (HK) is a plasma protein that signals during inflammation and fibrinolysis process [53]. HK is a central molecule of the kallikrein-kinin system (KKS) [54]. The KKS is composed of serine proteases that are primarily involved in the production of kinin peptides (i.e., Lys-bradykinin (kallidin) and bradykinin) [55]. KKS has a role in physiological processes such as inflammation, coagulation, vasodilation, fibrinolysis and control of blood pressure [56,57].

Six HK domains form a complex with prekallikrein in plasma [58]. Plasma kallikrein cleaves HK upon activation of the factor XII (FXII) generating a cleaved HK (HKc) and releasing pro-inflammatory bradykinin [59]. Bradykinin release results in the generation of twin chain HK (HKa) which induces apoptosis upon interaction with uPAR, HKa competitively inhibits the binding of uPAR with integrins αvβ3 and αvβ5. HKa shares an LRP binding site on uPAR, suggesting that it may have a role in the modulation of uPAR internalization and redistribution. HKa signals downstream via inhibition of uPA-dependent ERK phosphorylation [60], stimulation of EGFR signalling and enhancing VEGF expression [61].

uPAR, along with cytokeratin 1 and globular C1q-receptor (gC1qR, p33) are the endothelial cell receptors for factor XII (FXII) [62]. FXII stimulation drives cyclic adenosine

monophosphate (cAMP) accumulation and phosphorylation of ERK and AKT, leading to the enhanced cellular proliferation [63]. Integrin β1 and αMβ2 modulate the interaction of FXII with uPAR [64].

Lateral partners

In the absence of transmembrane and intracellular domains, uPAR promotes its signalling via its coreceptors integrins, RTK and GPCRs [38]. Integrins are the most widely studied and potentially the most significant uPAR binding partners [65].

Integrins

Integrins are thought to be the most significant uPAR coreceptors. uPAR is localized within integrin-containing adhesomes and found to co-precipitate with integrins and molecules involved in integrin signalling (e.g., FAK and Src) [20]. uPAR-integrin interactions have been implicated in wound healing, inflammation and across multiple epithelial cancers [66].

Integrins are heterodimer transmembrane receptors composed of non-covalently linked α- and β- glycoprotein subunits. Eighteen different α subunits and eight different β subunits are known to be expressed in metazoans exclusively. Up to 24 distinct combinations of these subunits have been discovered thus far [67]. They collectively serve as receptors for ECM components, such as collagens, fibronectins and laminins and do so with widely different affinities [65].

Integrins serve as transmembrane linkers facilitating interaction between cytoskeleton and ECM, where the β subunit tail binds with various intracellular anchor proteins through its c-terminal domain. These anchor proteins can either bind directly to actin filaments or other anchor proteins, thus indirectly linking integrins with cellular actin filaments. Linking of integrins and actin filaments lead to the formation of focal adhesion that can be followed by integrin clustering [65]. Integrin clustering influences proteins involved in cytoskeletal organization. Many lines of evidence point to the role of integrins in uPA- and uPAR-induced cellular migration. Interestingly, uPA-uPAR complexes are found in cellular focal contacts in the presence of integrins, but these are much more diffuse over the cell-surface in the absence of integrins [49]. Integrins laterally interact with receptors other than uPAR, forming complexes which in turn recruit additional signalling ligands via these complexes [68].

Integrins enable uPAR to modulate multiple signalling pathways by providing signal specificity, as each uPAR-integrin partnership signals through different pathways. For example, uPAR-β1 is primarily involved in ERK and FAK activation, whereas the uPAR-β3 interaction is involved in Rac activation [20]. Similarly, integrins are known to interact with multiple partners. For example, αMβ2-uPAR promotes uPA-dependent adhesion and migration upon complexation with fibronectin, α5β1-fibronectin interaction induces FAK phosphorylation and subsequent activation of Ras-ERK signalling. Integrin α3β1-uPAR interaction is known to enhance Src signalling induced by subsequent binding of the integrin with laminin. Equally, the integrin α3β1-uPAR interaction is responsible for switching on ERK signalling upon activation of EGFR and FAK [13]. Co-precipitation of uPAR with EGFR suggests a role for uPAR in EGFR-integrin α5β1 physiology. EGFR-

dependent ERK signalling enhances cell proliferation, driven by the integrin α5β1-uPAR interaction, where proliferation is inhibited by EGFR kinase inhibitors or by downregulation of uPAR expression [13].

Our team has made significant contributions to the roles of uPAR and the integrin αvβ6 in human cancer. Elevated expression of integrin αvβ6 in epithelial carcinomas is generally accepted as a marker of poor prognosis [69] due to its role in the enhancement of tumor invasion, EMT and distal metastasis [70].

Roles for uPAR and integrin αvβ6 in cancer has been established and strong evidence exists regarding direct binding of uPAR with αvβ6. This evidence is primarily based on co-immunoprecipitation, proximity ligation assay, peptide array and *in silico* structural modelling data [71]. We have previously established that downregulation of integrin αvβ6 inhibits tumor growth and MAPK activity *in vivo* [72]. Integrin αvβ6 was identified as a component of the uPAR interactome during pull-down studies using human ovarian cancer cell lines and inhibition of uPAR and αvβ6 equally suppressed ERK phosphorylation. The data showed the uPAR signals through the formation of the multi-protein complex [73]. Similarly, enhanced levels of integrin β6 in human colon cancer cell models results in increased proliferation, migration and invasion, compared to cells lacking the β6 subunit [74].

G-protein-coupled receptors

GPCRs are one of the most significant cell surface protein families. Each GPCR structure contains seven transmembrane-spanning domains (TMDs). GPCRs transduce a wide array of extracellular stimuli driven by ligands like amines, hormones, peptides, photons,

nutrients, ions and odorants. GPCR signalling is involved in many diverse physiological processes, including the chemosensory perception of smell and taste, neurotransmission, cellular metabolism and embryogenesis [75,76]. Their involvement in multiple physiological functions and the presence of accessible membrane druggable sites have made these proteins a significant subset of approved drug targets [77]. GPCR expression in proliferating cells has a role in tissue modelling, embryogenesis and inflammation [78–80]. For example, the fMet-Leu-Phe (fMLP) receptor family of proteins are involved in the PAS-induced proteolytic cascade, they are composed of the formyl peptide receptor (FPR) and its homologues FPR1 and FPR2. FPR is involved in chemotaxis [81], whereas FPRL1 activation upon binding to the suPAR SRSRY sequence induces cellular migration. FMLP-dependent cell proliferation and migration are driven by uPAR expression, whereas cell migration dependent on uPA requires the expression of FPRL1 [82]. FMLP receptors bind to uPAR through the D2-D3 linker sequence SRSRY to induce cell proliferation, migration and adhesion. However, the sequence can independently antagonise uPAR-Vn interactions and subsequent uPAR-dependent adhesion [82].

Mannose 6 Phosphate receptor

Mannose 6 phosphate receptor (M6PR) is a single TMD-containing glycoprotein that is ubiquitously expressed in human tissues [83,84]. M6PR is a multifunctional receptor with diverse ligands and is considered a tumor suppressor in multiple human malignancies. M6PR-dependent ECM remodelling is driven by uPAR-induced latent TGFβ1 and plasmin activation [85]. Enhanced expression of the M6PR leads to the proteolytic cleavage of uPAR, releasing its D2 and D3 fragments. These uPAR fragments bind to multiple uPAR

ligands generating downstream signalling. M6PR inhibits cell proliferation and invasion via a type of regulation of integrin αv and cleavage of uPAR, resulting in the dissolution of uPAR binding sites for uPA, VN and integrins [86].

Conclusion

uPAR regulates multiple cancer metastatic signatures, both uPA and uPAR have been suggested as potential diagnostic, prognostic and theranostic biomarkers in the breast, colon, rectal, pancreas, lung, kidney, prostate, ovary and liver cancers [87,88,47]. However, the predictive capacity and prognostic value of these markers have not yet been translated into clinical practice. Targeting components of the PAS restrain metastatic progression and prolong lifespan in animal models. However, prioritisation of the diverse role of the PAS and the uPAR interactome present a challenge. Alternate proteases can potentially compensate for many functions of the PAS rendering this target obsolete, one of such proteases is MMP which have similar pathological and physiological roles in cancer.

Characterization of quantitative differences in the expression of PAS components, (uPA, uPAR and PAI-1) in tumor tissues and blood plasma has potential diagnostic value in cancer. Quantitation may aid in stratification of disease from healthy patients or may allow for the stratification of patients at various clinical stages of cancer.

Notably, antagonization of the uPAR interactome has shown to suppress the cancer cell migration and induce apoptosis. Selective inhibition of the uPAR interactome with i) small molecules [89,90], ii) peptides [71,91] iii) antibodies [92,93] and iv) gene silencing, [94,95] has a potential to curb late-stage metastatic phenotypes. Strong preclinical data on inhibition of

the PAS driven metastasome with targeted approach against uPAR interactome is bound to define new diagnostic and prognostic strategy against cancer in clinical settings.

References

1. Lambert, A. W., Pattabiraman, D. R. & Weinberg, R. A. *Emerging Biological Principles of Metastasis*. *Cell* **168**, 670–691 (Cell Press, 2017).
2. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
3. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
4. Diepenbruck, M. & Christofori, G. Epithelial–mesenchymal transition (EMT) and metastasis: yes, no, maybe? *Curr. Opin. Cell Biol.* **43**, 7–13 (2016).
5. Blobe, G. C., Schiemann, W. P. & Lodish, H. F. Role of transforming growth factor β in human disease. *N. Engl. J. Med.* **342**, 1350–1358 (2000).
6. Nieto, M. A., Huang, R. Y. Y. J., Jackson, R. A. A. & Thiery, J. P. P. EMT: 2016. *Cell* (2016). doi:10.1016/j.cell.2016.06.028
7. Duffy, M. The Urokinase Plasminogen Activator System: Role in Malignancy. *Curr. Pharm. Des.* **10**, 39–49 (2004).
8. Su, S.-C., Lin, C.-W., Yang, W.-E., Fan, W.-L. & Yang, S.-F. The urokinase-type plasminogen activator (uPA) system as a biomarker and therapeutic target in human malignancies. *Expert Opin. Ther. Targets* **20**, 551–566 (2016).
9. Dass, K., Ahmad, A., Azmi, A. S., Sarkar, S. H. & Sarkar, F. H. Evolving role of uPA/uPAR system in human cancers. *Cancer Treatment Reviews* (2008). doi:10.1016/j.ctrv.2007.10.005

10. Kwaan, H. C., Mazar, A. P. & McMahon, B. J. The apparent uPA/PAI-1 paradox in cancer: More than meets the eye. *Semin. Thromb. Hemost.* (2013). doi:10.1055/s-0033-1338127

11. Mekkawy, A. H., Pourgholami, M. H. & Morris, D. L. Involvement of Urokinase-Type Plasminogen Activator System in Cancer: An Overview. *Med. Res. Rev.* (2014). doi:10.1002/med.21308

12. Xu, X. *et al.* Crystal structure of the urokinase receptor in a ligand-free form. *J. Mol. Biol.* (2012). doi:10.1016/j.jmb.2011.12.058

13. Noh, H., Hong, S. & Huang, S. Role of urokinase receptor in tumor progression and development. *Theranostics* **3**, 487–95 (2013).

14. Huai, Q. *et al.* Structure of human urokinase plasminogen activator in complex with its receptor. *Science* **311**, 656–659 (2006).

15. Santibanez, J. F. Transforming Growth Factor-Beta and Urokinase-Type Plasminogen Activator: Dangerous Partners in Tumorigenesis—Implications in Skin Cancer. *ISRN Dermatol.* (2013). doi:10.1155/2013/597927

16. De Bock, C. E. & Wang, Y. Clinical Significance of Urokinase-Type Plasminogen Activator Receptor (uPAR) Expression in Cancer. *Medicinal Research Reviews* (2004). doi:10.1002/med.10054

17. Ma, Y.-Y. & Tao, H.-Q. Role of Urokinase Plasminogen Activator Receptor in Gastric Cancer: A Potential Therapeutic Target. *Cancer Biother. Radiopharm.* **27**, 285–290 (2012).

18. C. Boonstra, M. *et al.* Clinical Applications of the Urokinase Receptor (uPAR) for Cancer Patients. *Curr. Pharm. Des.* **17**, 1890–1910 (2011).

19. Dergilev, K. V, Stepanova, V. V, Beloglazova, I. B., Tsokolayev, Z. I. & Parfenova, E. V. Multifaced Roles of the Urokinase System in the Regulation of Stem Cell Niches. *Acta Naturae* **10**, 19–32 (2018).

20. Smith, H. W. & Marshall, C. J. *Regulation of cell signalling by uPAR*. *Nature Reviews Molecular Cell Biology* **11**, 23–36 (Nature Publishing Group, 2010).

21. Lakka, S. S. *et al.* Adenovirus-mediated antisense urokinase-type plasminogen activator receptor gene transfer reduces tumor cell invasion and metastasis in non-small cell lung cancer cell lines. *Clin. Cancer Res.* **7**, 1087–93 (2001).

22. Pulukuri, S. M. K. *et al.* RNA interference-directed knockdown of urokinase plasminogen activator and urokinase plasminogen activator receptor inhibits prostate cancer cell invasion, survival, and tumorigenicity in vivo. *J. Biol. Chem.* (2005). doi:10.1074/jbc.M503111200

23. Montuori, N. *et al.* Urokinase type plasminogen activator receptor (uPAR) as a new therapeutic target in cancer. *Transl. Med. @ UniSa* (2016).

24. Xu, X. *et al.* Identification of a new epitope in uPAR as a target for the cancer therapeutic monoclonal antibody ATN-658, a structural homolog of the uPAR binding integrin CD11b (αM). *PLoS One* (2014). doi:10.1371/journal.pone.0085349

25. Chaurasia, P., Mezei, M., Zhou, M.-M. & Ossowski, L. Computer aided identification of small molecules disrupting uPAR/alpha5beta1--integrin interaction: a new paradigm for metastasis prevention. *PLoS One* **4**, e4617 (2009).

26. De Souza, M., Matthews, H., Lee, J. A., Ranson, M. & Kelso, M. J. Small

molecule antagonists of the urokinase (uPA): urokinase receptor (uPAR) interaction with high reported potencies show only weak effects in cell-based competition assays employing the native uPAR ligand. *Bioorg. Med. Chem.* **19**, 2549–56 (2011).

27. Persson, M. & Kjaer, A. Urokinase-type plasminogen activator receptor (uPAR) as a promising new imaging target: Potential clinical applications. *Clinical Physiology and Functional Imaging* (2013). doi:10.1111/cpf.12037

28. Persson, M. *et al.* First-in-human uPAR PET: Imaging of Cancer Aggressiveness. *Theranostics* **5**, 1303–16 (2015).

29. Persson, M., Madsen, J., Østergaard, S., Ploug, M. & Kjaer, A. 68Ga-labeling and in vivo evaluation of a uPAR binding DOTA- and NODAGA-conjugated peptide for PET imaging of invasive cancers. *Nucl. Med. Biol.* (2012). doi:10.1016/j.nucmedbio.2011.10.011

30. Persson, M., Liu, H., Madsen, J., Cheng, Z. & Kjaer, A. First 18F-labeled ligand for PET imaging of uPAR: In vivo studies in human prostate cancer xenografts. *Nucl. Med. Biol.* (2013). doi:10.1016/j.nucmedbio.2013.03.001

31. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529–D541 (2019).

32. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* **40**, D841–D846 (2012).

33. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

34. del-Toro, N. *et al.* A new reference implementation of the PSICQUIC web service. *Nucleic Acids Res.* **41**, W601–W606 (2013).

35. Llinas, P. *et al.* Crystal structure of the human urokinase plasminogen activator receptor bound to an antagonist peptide. *EMBO J.* **24**, 1655–63 (2005).

36. Montuori, N. *et al.* uPAR regulates pericellular proteolysis through a mechanism involving integrins and fMLF-receptors. *Thromb. Haemost.* (2013). doi:10.1160/TH12-08-0546

37. Gårdsvoll, H. *et al.* Characterization of the functional epitope on the urokinase receptor. Complete alanine scanning mutagenesis supplemented by chemical cross-linking. *J. Biol. Chem.* **281**, 19260–72 (2006).

38. Eden, G., Archinti, M., Furlan, F., Murphy, R. & Degryse, B. The urokinase receptor interactome. *Curr. Pharm. Des.* **17**, 1874–89 (2011).

39. Kyuno, D. *et al.* Claudin7-dependent exosome-promoted reprogramming of nonmetastasizing tumor cells. *Int. J. Cancer* (2019). doi:10.1002/ijc.32312

40. Eden, G. *et al.* D2A sequence of the urokinase receptor induces cell growth through αvβ3 integrin and EGFR. *Cell. Mol. Life Sci.* (2018). doi:10.1007/s00018-017-2718-3

41. Gondi, C., Kandhukuri, N., Dinh, D., Gujrati, M. & Rao, J. Down-regulation of uPAR and uPA activates caspase-mediated apoptosis and inhibits the PI3K/AKT pathway. *Int. J. Oncol.* (2007). doi:10.3892/ijo.31.1.19

42. Wang, Z., Li, Y., Banerjee, S. & Sarkar, F. H. Exploitation of the Notch signaling pathway as a novel target for cancer therapy. *Anticancer Res.* **28**, 3621–30

43. Leube, R. E., Moch, M. & Windoffer, R. Intermediate filaments and the regulation

of focal adhesion. *Current Opinion in Cell Biology* (2015). doi:10.1016/j.ceb.2014.09.011

44. Kjøller, L. & Hall, A. Rac mediates cytoskeletal rearrangements and increased cell motility induced by urokinase-type plasminogen activator receptor binding to vitronectin. *J. Cell Biol.* (2001). doi:10.1083/jcb.152.6.1145
45. Herkenne, S. Involvement of PAI-1/uPA/uPAR interactome in angiogenesis. (2013).
46. Pepper, M. S. Role of the Matrix Metalloproteinase and Plasminogen Activator–Plasmin Systems in Angiogenesis. *Arterioscler. Thromb. Vasc. Biol.* **21**, 1104–1117 (2001).
47. Mahmood, N., Mihalcioiu, C. & Rabbani, S. A. Multifaceted Role of the Urokinase-Type Plasminogen Activator (uPA) and Its Receptor (uPAR): Diagnostic, Prognostic, and Therapeutic Applications. *Front. Oncol.* **8**, 24 (2018).
48. De Lorenzi, V. *et al.* Urokinase links plasminogen activation and cell adhesion by cleavage of the RGD motif in vitronectin. *EMBO Rep.* **17**, 982–998 (2016).
49. Blasi, F. & Carmeliet, P. uPAR: A versatile signalling orchestrator. *Nature Reviews Molecular Cell Biology* (2002). doi:10.1038/nrm977
50. Huai, Q. *et al.* Crystal structures of two human vitronectin, urokinase and urokinase receptor complexes. *Nat. Struct. Mol. Biol.* (2008). doi:10.1038/nsmb.1404
51. Ferraris, G. M. S. *et al.* The interaction between uPAR and vitronectin triggers ligand-independent adhesion signalling by integrins. *EMBO J.* (2014). doi:10.15252/embj.201387611

52. Pirazzoli, V., Ferraris, G. M. S. & Sidenius, N. Direct evidence of the importance of vitronectin and its interaction with the urokinase receptor in tumor growth. *Blood* (2013). doi:10.1182/blood-2012-08-451187

53. Liu, Y. *et al.* Cleaved high-molecular-weight kininogen and its domain 5 inhibit migration and invasion of human prostate cancer cells through the epidermal growth factor receptor pathway. *Oncogene* (2009). doi:10.1038/onc.2009.132

54. Shukla, M. *et al.* Regulation of the Tumor Microenvironment By High Molecular Weight Kininogen. *Blood* **128**, (2016).

55. Wu, Y. Contact pathway of coagulation and inflammation. *Thrombosis Journal* (2015). doi:10.1186/s12959-015-0048-y

56. Pathak, M., Wong, S. S., Dreveny, I. & Emsley, J. Structure of plasma and tissue kallikreins. *Thromb. Haemost.* (2013). doi:10.1160/TH12-11-0840

57. Bryant, J. & Shariat-Madar, Z. Human Plasma Kallikrein-Kinin System: Physiological and Biochemical Parameters. *Cardiovasc. Hematol. Agents Med. Chem.* (2009). doi:10.2174/187152509789105444

58. Schmaier, A. H. & McCrae, K. R. The plasma kallikrein-kinin system: Its evolution from contact activation. *Journal of Thrombosis and Haemostasis* (2007). doi:10.1111/j.1538-7836.2007.02770.x

59. Yamamoto-Imoto, H. *et al.* A novel detection method of cleaved plasma high-molecular-weight kininogen reveals its correlation with Alzheimer's pathology and cognitive impairment. *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.* (2018). doi:10.1016/j.dadm.2018.06.008

60. Liu, Y., Cao, D. J., Sainz, I. M., Guo, Y.-L. & Colman, R. W. The inhibitory effect

of HKa in endothelial cell tube formation is mediated by disrupting the uPA-uPAR complex and inhibiting its signaling and internalization. *Am. J. Physiol. Physiol.* (2008). doi:10.1152/ajpcell.00569.2007

61. Xu, J. *et al.* Overexpression of the Kininogen-1 inhibits proliferation and induces apoptosis of glioma cells. *J. Exp. Clin. Cancer Res.* (2018). doi:10.1186/s13046-018-0833-0

62. Mahdi, F., Madar, Z. S., Figueroa, C. D. & Schmaier, A. H. Factor XII interacts with the multiprotein assembly of urokinase plasminogen activator receptor, gC1qR, and cytokeratin 1 on endothelial cell membranes. *Blood* (2002). doi:10.1182/blood.V99.10.3585

63. LaRusch, G. A. *et al.* Factor XII stimulates ERK1/2 and Akt through uPAR, integrins, and the EGFR to initiate angiogenesis. *Blood* (2010). doi:10.1182/blood-2009-08-236430

64. Göbel, K. *et al.* Blood coagulation factor XII drives adaptive immunity during neuroinflammation via CD87-mediated modulation of dendritic cells. *Nat. Commun.* (2016). doi:10.1038/ncomms11626

65. Danen, E. H. J. Integrins: An Overview of Structural and Functional Aspects. (2013).

66. Cantor, D. I., Cheruku, H. R., Nice, E. C. & Baker, M. S. Integrin αvβ6 sets the stage for colorectal cancer metastasis. *Cancer Metastasis Rev.* (2015). doi:10.1007/s10555-015-9591-z

67. Hynes, R. O. Integrins: Bidirectional, allosteric signaling machines. *Cell* **110**, 673–687 (2002).

68. Tang, C. H. & Wei, Y. The urokinase receptor and integrins in cancer progression. *Cell. Mol. Life Sci.* **65**, 1916–1932 (2008).

69. Niu, J. & Li, Z. *The roles of integrin αvβ6 in cancer*. *Cancer Letters* **403**, 128–137 (2017).

70. Brown, E. J. Integrin-associated proteins. *Current Opinion in Cell Biology* (2002). doi:10.1016/S0955-0674(02)00360-5

71. Ahn, S. B. *et al.* Characterization of the interaction between heterodimeric αvβ6 integrin and urokinase plasminogen activator receptor (uPAR) using functional proteomics. *J. Proteome Res.* (2014). doi:10.1021/pr500849x

72. Ahmed, N. *et al.* Direct integrin αvβ6-ERK binding: Implications for tumour growth. *Oncogene* (2002). doi:10.1038/sj/onc/1205286

73. Saldanha, R. G. *et al.* Proteomic identification of lynchpin urokinase plasminogen activator receptor protein interactions associated with epithelial cancer malignancy. *J. Proteome Res.* (2007). doi:10.1021/pr060518n

74. Cantor, D., Slapetova, I., Kan, A., McQuade, L. R. & Baker, M. S. Overexpression of αvβ6 integrin alters the colorectal cancer cell proteome in favor of elevated proliferation and a switching in cellular adhesion that increases invasion. *J. Proteome Res.* (2013). doi:10.1021/pr301099f

75. Adhikari, S., Sharma, S., Ahn, S. B. & Baker, M. S. In Silico Peptide Repertoire of Human Olfactory Receptor Proteome on High-Stringency Mass Spectrometry. *J. Proteome Res.* acs.jproteome.8b00494 (2019). doi:10.1021/acs.jproteome.8b00494

76. Nieto Gutierrez, A. & McDonald, P. H. GPCRs: Emerging anti-cancer drug

targets. *Cellular Signalling* (2018). doi:10.1016/j.cellsig.2017.09.005

77. Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B. & Gloriam, D. E. Trends in GPCR drug discovery: New agents, targets and indications. *Nat. Rev. Drug Discov.* **16**, 829–842 (2017).

78. Dorsam, R. T. & Gutkind, J. S. G-protein-coupled receptors and cancer. *Nature Reviews Cancer* (2007). doi:10.1038/nrc2069

79. O'Hayre, M., Degese, M. S. & Gutkind, J. S. Novel insights into G protein and G protein-coupled receptor signaling in cancer. *Current Opinion in Cell Biology* (2014). doi:10.1016/j.ceb.2014.01.005

80. Lynch, J. R. & Wang, J. Y. G protein-coupled receptor signaling in stem cells and cancer. *International Journal of Molecular Sciences* (2016). doi:10.3390/ijms17050707

81. Montuori, N., Carriero, M. V., Salzano, S., Rossi, G. & Ragno, P. The cleavage of the urokinase receptor regulates its multiple functions. *J. Biol. Chem.* (2002). doi:10.1074/jbc.M207494200

82. Montuori, N. & Ragno, P. Multiple activities of a multifaceted receptor: roles of cleaved and soluble uPAR. *Front. Biosci.* **14**, 2494–503 (2009).

83. El-Shewy, H. M. & Luttrell, L. M. Chapter 24 Insulin-Like Growth Factor-2/Mannose-6 Phosphate Receptors. *Vitamins and Hormones* (2009). doi:10.1016/S0083-6729(08)00624-9

84. Martin-Kleiner, I. & Gall Troselj, K. Mannose-6-phosphate/insulin-like growth factor 2 receptor (M6P/IGF2R) in carcinogenesis. *Cancer Letters* (2010). doi:10.1016/j.canlet.2009.06.036

85. Chang, M. H. *et al.* IGF-II/mannose 6-phosphate receptor activation induces metalloproteinase-9 matrix activity and increases plasminogen activator expression in H9c2 cardiomyoblast cells. *J. Mol. Endocrinol.* (2008). doi:10.1677/JME-08-0051

86. Schiller, H. B., Szekeres, A., Binder, B. R., Stockinger, H. & Leksa, V. Mannose 6-Phosphate/Insulin-like Growth Factor 2 Receptor Limits Cell Invasion by Controlling αVβ3 Integrin Expression and Proteolytic Processing of Urokinase-type Plasminogen Activator Receptor. *Mol. Biol. Cell* (2009). doi:10.1091/mbc.e08-06-0569

87. McMahon, B. J. & Kwaan, H. C. Components of the plasminogen-plasmin system as biologic markers for cancer. in *Advances in Experimental Medicine and Biology* (2015). doi:10.1007/978-94-017-7215-0_10

88. Rasmussen, L. J. H. *et al.* Inflammatory biomarkers and cancer: CRP and suPAR as markers of incident cancer in patients with serious nonspecific symptoms and signs of cancer. *Int. J. Cancer* (2017). doi:10.1002/ijc.30732

89. Mani, T. *et al.* Small-molecule inhibition of the uPAR·uPA interaction: Synthesis, biochemical, cellular, in vivo pharmacokinetics and efficacy studies in breast cancer metastasis. *Bioorganic Med. Chem.* (2013). doi:10.1016/j.bmc.2012.12.047

90. Liu, D., Zhou, D., Wang, B., Knabe, W. E. & Meroueh, S. O. A New Class of Orthosteric uPAR·uPA Small-Molecule Antagonists Are Allosteric Inhibitors of the uPAR·Vitronectin Interaction. *ACS Chem. Biol.* **10**, 1521–1534 (2015).

91. Yamada, Y., Kanayama, S., Ito, F., Kurita, N. & Kobayashi, H. A novel peptide

blocking cancer cell invasion by structure-based drug design. *Biomed. Reports* (2017). doi:10.3892/br.2017.957

92. K. Lund, I., Illemann, M., Thurison, T., J. Christensen, I. & Hoyer-Hansen, G. uPAR as Anti-Cancer Target: Evaluation of Biomarker Potential, Histological Localization, and Antibody-Based Therapy. *Curr. Drug Targets* (2011). doi:10.2174/138945011797635902

93. Duriseti, S. *et al.* Antagonistic anti-urokinase plasminogen activator receptor (uPAR) antibodies significantly inhibit uPAR-mediated cellular signaling and migration. *J. Biol. Chem.* (2010). doi:10.1074/jbc.M109.077677

94. Kunigal, S., Lakka, S. S., Gondi, C. S., Estes, N. & Rao, J. S. RNAi-mediated downregulation of urokinase plasminogen activator receptor and matrix metalloprotease-9 in human breast cancer cells results in decreased tumor invasion, angiogenesis and growth. *Int. J. Cancer* (2007). doi:10.1002/ijc.22962

95. Raghu, H., Gondi, C. S., Dinh, D. H., Gujrati, M. & Rao, J. S. Specific knockdown of uPA/uPAR attenuates invasion in glioblastoma cells and xenografts by inhibition of cleavage and trafficking of Notch -1 receptor. *Mol. Cancer* (2011). doi:10.1186/1476-4598-10-130

3.5.2 Proteomics reveals cell-surface urokinase plasminogen activator receptor (uPAR) levels impact most hallmarks of cancer: Supplementary Manuscript 1

I also investigated the role of uPAR in CRC metastasis through a systemic proteomics analysis on CRC cell model that allowed decreased uPAR expression (by ~43%). Reduction in uPAR levels leads to suppression many proteins associated with the metastasis-related components of the uPAR interactome (e.g., caveolin, EGFR, integrin β4, vitronectin) as validated by Ingenuity pathway (IPA) analysis and use of the new Cancer Hallmarks Analytics Tool (CHAT) [5]. The study, for the first time, demonstrated that reduction in uPAR expression negates many HoCs when these are superimposed on a particularly common CRC mutational background. Comprehensive proteome depth was achieved by a combination of membrane enrichment combined with peptide-fractionation strategies and these allowed me to decipher proteome changes not previously seen in other CRC models.

Accepted Article

Proteomics reveals cell-surface urokinase plasminogen activator receptor (uPAR) expression impacts most hallmarks of cancer

Seong Beom Ahn[1+], Abidali Mohamedali[2+], Dana Pascovici[3], Subash Adhikari[1], Samridhi Sharma[1], Edouard C. Nice[4], and Mark S. Baker[1*]

[1]Department of Biomedical Sciences, Faculty of Medicine and Health Science, [2]Department of Molecular Sciences, Faculty of Science and Engineering, [3]Australian Proteome Analysis Facility, Macquarie University, Sydney, Australia. [4]Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia.

* Corresponding Author: Mark S. Baker, Level 1, 75 Talavera Road, Macquarie University, 2109, Australia, +61 2 9850 8211, mark.baker@mq.edu.au

[+] These authors contributed equally to this work

Abbreviations: CHAT, cancer hallmarks analytics tool; CRC, colorectal cancer; CW, carbonate wash; ECM, extracellular matrix; HoCs, hallmarks of cancer; IPA, Ingenuity Pathway Analysis; PRM, parallel

Received: 27/02/2019; Revised: 25/07/2019; Accepted: 09/08/2019

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201900026.

This article is protected by copyright. All rights reserved.

reaction monitoring; TX-114, Triton X-114 phase partitioning; uPA, urokinase plasminogen activator; uPAR, urokinase plasminogen activator receptor; WCE, whole cell extraction;

Keywords: uPAR, Hallmarks of Cancer, Colorectal Cancer, Membrane Protein Extraction

Accepted Article

Abstract

While metastasis is the primary cause of colorectal cancer (CRC) mortality, the molecular mechanisms underpinning it remains elusive. Metastasis is propagated through driver oncogene/suppressor gene mutations, accompanied by passenger mutations and underlying genomic instability. To understand cancer biology, a unifying framework called the *hallmarks of cancer* (HoCs) had been developed, which organizes cell biological alterations under ten key hallmarks. Underlying these HoCs, genome instability generates mutational diversity that is amplified by inflammation. Recognizing how critical cancer cell-surface proteins influence these HoCs has been proposed to accelerate precision medicine therapeutic development. Here, we asked if a moderate decrease in expression (↓43%) of HCT116 cell line urokinase plasminogen activator receptor (uPAR) negates HoCs driven by its *KRAS* and *PIK3CA* mutational background found in CRC. Comprehensive proteomics (whole cell lysis with two membrane protein enrichments) coupled with Ingenuity Pathway Analysis (IPA) demonstrated that uPAR negates essential pathways across the HoC spectrum, particularly those associated with metastasis, resisting cell death and sustaining proliferation, and parallels Cancer Hallmarks Analytics Tool analysis. Decreasing uPAR predominantly alters metastasis-related and uPAR-interactome protein expression (e.g., EGFR, caveolin, vitronectin, integrin β4). Collectively, we demonstrate that uPAR is a lynchpin protein capable of regulating several HoC pathways in a classical CRC mutational background.

This article is protected by copyright. All rights reserved.

Accepted Article

Significance of the study

Lowering expression of a lynchpin protein uPAR negates many hallmarks of cancer by altering primary cancer signaling pathways and the uPAR interactome, supporting the hypothesis that lowering uPAR expression may be a reasonable therapeutic strategy to abrogate metastasis.

Introduction

Genomics, epigenomics, transcriptomics, metabolomics, proteomics and bioinformatics have facilitated advances in our understanding of cancer. Despite this, the molecular mechanisms underpinning cancer metastasis remain mostly intangible.[1] Metastasis is thought to be a multifactorial process involving extracellular matrix (ECM) degradation, epithelial-to-mesenchymal transition (EMT), increased motility, escape from immune surveillance and ability of released cancer cells to re-establish themselves elsewhere in vital organs, effectively shutting these down and eventually leading to mortality.[1] A significant framework used to understand biologies central to cancer has been called the 'Hallmarks of Cancer' (HoCs). This framework, published in 2000[2] has been updated in 2011.[3] Overall, a prevalent concept is that cancer is fundamentally a series of concatenated genomic diseases, sustained by driver mutations of critical oncogenes and tumor suppressor genes[4] that eventually lead to metastasis. Each step in the evolution to cancer metastasis has been explained from a genomic perspective, and a comprehensive list of known alterations is located in COSMIC (Catalogue of Somatic Mutations in Cancer).[5] Many mutations lead to altered proliferation (e.g., RAS, MYC, RAF), evasion of growth suppression (e.g., RB, TP53, TGF-β),

This article is protected by copyright. All rights reserved.

Accepted Article

resistance to apoptosis (e.g., Bcl-2 family, TP53), replicative immortality (e.g., telomerase, TERT) angiogenesis (e.g., VEGF, TSP), and/or the EMT (e.g., β-catenin, BCF4).[2, 3]

Similar genomic approaches have focused on CRC, which accounts for >8% of all cancer deaths worldwide.[6] In CRC, it is well-established that transformation of normal epithelium to adenomas to carcinomas involves combinations of mutational phenotypes superimposed on genomic abnormalities, including chromosomal instability (CIN), microsatellite instability (MSI) and CpG island methylator phenotypes (CIMP).[7] These instabilities are acquired in the presence of an abnormal number of chromosomes (i.e., aneuploid), deficiencies in DNA mismatch-repair and aberrant DNA methylation,[7] leading to increased oncogene and/or tumor suppressor mutation rates (e.g., *KRAS, BRAF, APC, PTEN, SMAD4, PIK3CA, AKT1, TP53*).[8] These mutations lead to activation of RAS-RAF-MEK-ERK (known as MAPK/ERK) signaling pathway, ultimately playing crucial roles in regulating cell proliferation, growth, survival, invasion and metastasis.[9]

Given the prevailing view, HoCs are principally driven by gene mutation we enquired whether protein expression changes derived as part of a prominent CRC genomic mutation had significant effects. Many studies have suggested altering the expression of key "lynchpin" cancer-related proteins may affect proliferation and metastasis. We previously demonstrated that transfecting a subclone of SW480 with a vector inducing αvβ6 integrin overexpression elevated proliferation and invasion compared to "empty" mock vector controls[10] and we have subsequently proposed these biological changes likely occurred because of the interaction of αvβ6 with cell-surface protein urokinase plasminogen activator receptor (uPAR)[10, 11].

Many proteins are involved or interact with the plasminogen activation system, including matrix metalloproteases (MMPs), plasminogen activator inhibitors 1/2 (PAI-1/2), urokinase plasminogen activator (uPA), uPAR and cell-surface integrins.[12] Most play roles in proteolytic ECM degradation and cancer metastasis.[12] There is a significant suite of evidence illustrating that up- or down-

This article is protected by copyright. All rights reserved.

Accepted Article

regulating specific proteins can alter phenotypic outcomes in cancer. Here, we ask if a modest change in expression of one "lynchpin" protein (uPAR) by ~43%↓ in the HCT116 cell line that is known to harbor common CRC *KRAS* and *PIK3CA* mutations, can modify HoCs in a manner analogous to driver mutational change, as partly suggested by our previous study.[13] There, we demonstrated that major cancer-related signaling pathway changes (e.g., β-catenin, c-myc or p53) were dysregulated by uPAR down-regulation.[13] However, that study relied on less-sensitive proteomics tools like 2DE with MALDI-MS and had limited membrane proteome coverage. In this current study, we address these limitations and perform a comprehensive proteome analysis using a well-studied CRC model.

uPAR is a 55-65 kDa glycosyl-phosphatidylinositol (GPI)-linked protein found on many cell type surfaces. Bound twin chain uPA activates plasminogen to plasmin which can directly degrade several ECM glycoproteins, or indirectly do so by activation of zymogen MMPs that degrade many human collagens.[12] Under physiological conditions, uPAR expression is low, with overexpression observed during wound healing, inflammation, trophoblast implantation, neural crest migration and cancer[12, 14] In epithelial carcinomas, uPAR is highly-overexpressed with normal physiological functions hijacked resulting in inflammation, angiogenesis, cancer cell proliferation, adhesion, invasion, migration, colonization and metastasis.[12] Currently, various proteoforms of uPAR[15] are candidate tissue or plasma diagnostic/prognostic biomarkers for clinical stage, within stage prognosis or overall patient survival.[14, 16]

uPAR's biological functions in cancer are masked by a complex interactome involving multiple partners. These interactions fit three major categories. Firstly, biologically/structurally well-established and characterized primary ligands for uPAR, like substrate uPA and vitronectin (VN).[12, 17] A second group comprises interacting partners that although not as well-established have substantial evidence supporting direct interaction. These include kininogen, integrin dimers ($\alpha_v\beta_{1\text{-}6}$),

This article is protected by copyright. All rights reserved.

Accepted Article

caveolin, EGFR and PDGFR.[12, 17] The last group has substantial indirect interaction evidence but may equally represent downstream interactions through intermediaries. These include chemotactic receptors, α-enolase and PAI-1/2.[12, 17] Because of these three distinct levels of interaction, uPAR may modulate a range of intracellular signaling including but not restricted to MAPK/ERK, SMADS and other tyrosine and serine kinases.[12, 17] In human cancers, the mutational landscape overlaps significantly with identified downstream targets of uPAR, such that the combination of genomic mutations with proteomic changes could allow cells to reach a tipping point - favouring metastasis. The dynamic "interactome" in cancer cells responds to, and results in, differences in signaling that affect many downstream biological phenomena.[17] To understand the different functions of uPAR, and how it fits within a CRC mutational landscape, necessitates a unifying conceptual framework as adopted here.

We investigated the global and plasma membrane proteomes that result from a moderate reduction in uPAR expression (~43%↓ in HCT116$^{\Delta SuPAR}$ compared to wild-type HCT116^{WT}). HCT116 cells were derived from Dukes' stage D CRC and established in 1981, [18] and contain CRC-related DNA instabilities (*MSI, CIMP*) and gene mutations (*KRAS, PIK3CA*).[7, 19] Here, the effects of decreasing expression of uPAR superimposed upon that 'classic' CRC mutational background is investigated, allowing a "bird's eye" view of molecular mechanisms that are disrupted. We utilized two membrane protein extraction methodologies with whole cell extraction to optimize membrane proteome coverage. For the first time, using IPA [20] and the broad HoC conceptual framework [3] analyzable through Cancer Hallmarks Analytics Tool (CHAT)[21], we overcome bottlenecks in analyzing large proteomics datasets, specifically within a cancer biology framework. Tying comprehensive proteomics to the HoC framework yield insights into the mechanisms underpinning uPAR biology, potentially uncovering interactions that could explain some enigmatic roles in cancer metastasis and that allowing exploration of targets for metastasis therapeutic intervention.

This article is protected by copyright. All rights reserved.

Accepted Article

Materials and methods

Our experimental workflow used in this study is summarized in Figure 1.

Cell Culture: The human colon carcinoma cell line HCT116^{WT} (HCT116 wild-type) (ATCC) and construction of the stable anti-sense uPAR HCT116^{ASuPAR} has been described previously.[22] HCT116^{WT} cells were maintained in Dulbecco's Modified Eagle's Medium (Sigma-Aldrich) in 10% FBS at 37°C and 5% CO_2, whilst media for HCT116^{ASuPAR} supplemented 400µg/mL hygromycin B. Cells were cultured in 150mm dishes until ~80% confluence, washed twice with PBS and lysed with appropriate lysis buffers.

Whole Cell Protein Extraction: Cells (~1x10^7), in triplicate, were lysed with whole cell lysis buffer (100mM TEAB (Tetraethylammonium bicarbonate), 1% sodium deoxycholate) using a probe sonicator (Branson Sonifier 450; 10 bursts at 40% amplitude, output 2 setting, repeated 3x). Samples were then heated at 95°C for 5min before storage at -80°C for mass spectrometry (MS) analysis.

Triton X-114 Phase Partitioning Membrane Protein Extraction:[23] Cells (~1x10^7) were lysed with lysis buffer (50mM Tris-HCl, 100mM NaCl, 1mM EDTA, pH 7.4 containing protease inhibitor cocktail: cOmplete™) using a probe sonicator as described above, followed by low speed (1,000g) centrifugation at 4°C for 15min to remove nuclei and cell debris but not plasma membrane components. Supernatants were collected and mixed with binding buffer (20mM Tris-HCl, 100mM NaCl) and ultracentrifuged (120,000g) at 4°C for 1hr to enrich membrane components. Pellets were washed with 0.1M Na_2CO_3 (pH 11) and resuspended with binding buffer containing 1%(v/v) TritonX-114. Resuspended samples were heated at 37°C for 20min and phase partitioned by centrifugation at 300g for 5min. The upper aqueous phase was removed, and the detergent phase mixed with ice-cold acetone and incubated at -20°C overnight to precipitate proteins. Precipitated proteins were pelleted by centrifugation at 4°C and 5,000g for 30min and pellets re-solubilized in 100mM NH_4HCO_3 upon being at -80°C.

This article is protected by copyright. All rights reserved.

Accepted Article

Sodium Carbonate Membrane Protein Extraction: Cells (~$1x10^7$) were lysed with HEPES buffer (20mM HEPES, 150mM NaCl, 10mM NaF, 1mM Na-EDTA, 1mM Na-EGTA, pH 7.5 and protease inhibitor cocktail) with probe sonication, followed by low-speed centrifugation as above. Supernatants were collected and mixed with 0.1M Na_2CO_3 and incubated for 1hr at 4°C on a rocking platform. The mixtures were ultracentrifuged at 4°C and 120,000g for 1hr. The membrane enriched pellets were solubilized with 100mM NH_4HCO_3 and stored at -80°C.

Western Immunoblotting/Densitometry: Protein concentration was measured using a BCA Protein Assay Kit following the manufacturer's protocol (Thermo Fisher Scientific). Proteins were separated on a 4-12% SDS-PAGE gel and transferred onto PVDF membranes with immunoblotting performed using SNAP i.d. system (Merck Millipore) following the manufacturer's protocol. The blots were incubated with selected primary antibodies (anti-uPAR R4 (1:1000; ab82220; Abcam), anti-integrin β4 (1:1000; M126; ab2942; Abcam), anti-EGFR (1:1000; 11E19; Sigma-Aldrich), anti-β actin; (1:10000; AC-15; ab6276; Abcam)) followed by HRP-conjugated anti-mouse (R&D systems) secondary (1:10000). Blots were imaged using a Luminescent Image Analyzer LAS-3000 (Fujifilm Australia), and band densities from immunoblots were analyzed using ImageJ software.[24] One-tailed independent t-tests assuming equal variances were performed on all data.

Immunofluorescence Confocal Microscopy: Cells were fixed with 2% paraformaldehyde, blocked with 1% BSA in PBS, stained with 2.5μg/mL anti-uPAR R4 MAb and detected using a goat anti-mouse Alexa Fluor-488 antibody. Cells were washed with PBS between steps. Mouse IgG1 (MAB002; R&D Systems) was used as an isotype control with identical methods as above. Alexa Fluor® 488 fluorescence was visualized using an Olympus FLUOVIEW 300 laser scanning confocal microscope equipped with a 60X, 1.42 numerical aperture oil (refractive index, n = 1.518) immersion objective. Horizontal optical sections (1024×1024 pixels) were taken through the middle of representative cells using Kalman averaging.

Protein Digestion and High-pH Peptide Fractionation: Lysates (25μg protein) were reduced with 5mM DTT for 30min at 60°C and then alkylated with 14mM iodoacetamide for 30min at room

This article is protected by copyright. All rights reserved.

Accepted Article

temperature in the dark. Proteins were digested using trypsin at a ratio of 1:30 at 37°C overnight and the digested peptides were fractionated with sequential elution by increasing ACN concentration (3%, 6%, 9%, 15%, 80% (v/v) in 5mM NH_4COOH) using a C18 (Empore 2215) stage-tip fractionation method.[25]

Mass Spectrometry: Protein identification was performed on an ABSciex TripleTOF 5600 coupled to an Eksigent Ultra nanoLC system (Eksigent) on a C18 Halo 2.7µm 160Å ES-C18, 150µm x 10cm analytical column. Samples injected onto Halo 2cm peptide trap column for pre-concentration and desalting with 0.1% formic acid (FA), 2% ACN at 5µL/min. Peptides eluted from the column using a linear 80min gradient from H_2O:ACN (98:2; + 0.1% FA) to H_2O:ACN (2:98; + 0.1% FA) at a constant flow 0.6µL/min. The LC eluent was subjected to positive ion nanoflow electrospray MS analysis in data-dependent acquisition mode (DDA). TOF-MS survey scan was acquired (m/z 350-1500, 0.25s), with ten largest multiply charged ions (counts >150) in the survey scan sequentially subjected to MS/MS analysis. MS/MS spectra were accumulated for 100msecs (m/z 100-1500) with rolling collision energy. The data has been deposited to ProteomeXchange Consortium via the PRIDE partner repository[26] (PXD010219).

Data Processing and Statistics: Raw data files were searched with Mascot (v.2.4.1) against the SwissProt human database (updated 05/2017 containing 554,515 sequences). The peptide FDR was calculated with searches against a corresponding reverse database. All searches were carried out with a value of ≤1% FDR, precursor peptide mass tolerance of ±50ppm, fragment ion mass tolerance of ±0.1Da, with one missed cleavage allowed and default peptide length of ≥7 amino acids. Fixed modifications were cysteine carbamidomethylation and oxidation. UniProt Gene Ontology (GO) annotation was used to assess the relative % of plasma membrane proteins obtained from each proteome extraction method.

Data files were analyzed with the label-free R data analysis tool Scrappy,[27] also used for subsequent statistical analysis. For each protein extraction method, proteins were retained only if reproducibly

This article is protected by copyright. All rights reserved.

Accepted Article

present, present in all three triplicates and with a total spectral count of ≥5 in at least one of the cell lines. Protein abundance was determined following past Normalized Spectral Abundance Factors (NSAF) protocols, where the protein spectral count was normalized to account for protein length and total sample amount then log-transformed for statistical analysis.[28] Data quality and distribution were examined using boxplots and density plots for the complete dataset and quantile-quantile plots for each pairwise comparison.

Differentially-expressed proteins were identified for each proteome extraction method used. Unpaired t-tests were run for each protein, using p-value thresholds of ≤0.05 and fold change limits of ≥1.5 up or down. Ingenuity Pathway Analysis (IPA v01-06 from Qiagen) was used to categories proteins into canonical pathways and biological interaction networks based on the proteins passing the thresholds above, using default IPA parameters for the core analysis option.[20] For HoC analysis, first, the resulting subsets of differentially-expressed proteins from each extraction method were combined, and IPA was performed to categorize all proteins into relevant diseases and biological functions. Then, the dataset from IPA was manually inspected, and each protein superimposed onto the ten established HoC[3], based-on any protein's biological characteristics and function/s. Finally, HoC-focused categorized protein groups were re-analyzed by IPA to identify which specific signaling pathways were activated or inhibited under each cancer hallmark. The outcomes from HoC analysis were then compared to a recently-released novel CHAT PubMed text-mining engine (http://chat.lionproject.net/[21]) with the following settings; Query: uPAR, Metric: count, Chart: bar, Hallmarks: top.

Parallel Reaction Monitoring (PRM) Analysis: PRM was performed on a Thermo Scientific™ Q-Exactive™ Hybrid Quadrupole-Orbitrap using a 15cm x 75μm I.D analytical column packed with 5μm C18 Michrom magic beads. Peptides were separated at a flow rate of 0.3μl/min with a 50min linear gradient from 1% to 65% buffer B (0.1% FA in 90% ACN), 2min gradient (65%-85%),

This article is protected by copyright. All rights reserved.

Accepted Article

followed by 8min at 85%. Recombinant uPAR digest (100ng) was subjected to Q-Exactive analysis in DDA mode. The peak area of all daughter ions detected for each peptide was compared using an unpaired t-test to determine the top intensity transition for each peptide. Identified top intensity transitions (i.e., uPAR peptides) from DDA mode were captured as an inclusion list and used to confirm the differential uPAR expression between the $HCT116^{WT}$ and $HCT116^{ASuPAR}$ whole cell lysates (injected 2µg each) in PRM mode (resolving power 17,500 at 200m/z with 2.0m/z isolation window at normalized collision energy 30%). Unpaired t-tests used to compare relative changes in uPAR expression between $HCT116^{WT}$ and $HCT116^{ASuPAR}$ based on the area under the curve for identified uPAR peptide transitions.

Results

$HCT116^{WT}$ and $HCT116^{ASuPAR}$ uPAR Expression

The use of stable antisense technologies to study the function of genes is widespread over the past two decades in the scientific literature.[29, 30] Before undertaking this study, we investigated a multitude of gene editing technologies and discovered that evidence remained about the efficiency and off-target effects for even newer and potentially clinically-applicable technologies like CRISPR/Cas9.[31] The $HCT116^{ASuPAR}$ cell line we developed some years ago now has proven to be exceptionally reliably, stable, viable and had been used in a multitude of other studies that inform and complement this current study.[22, 30, 32, 33] As our study required stable, but only partial downregulation of uPAR, we determined that no other methodology would be able to deliver this reliably and consistently. Hence, we chose to use the well-characterized cell lines produced by stable antisense methodology some years ago.

The original $HCT116^{ASuPAR}$ cell line was constructed in our JCSMR labs at Australian National University almost two decades ago, and the level of decrease in uPAR expression in this particular

This article is protected by copyright. All rights reserved.

Accepted Article

line was previously shown to be ~50% compared to wild-type HCT116^{WT}.[22, 33] Given the long-term use and storage of both HCT116 subclones and the fact that more accurate quantitative comprehensive MS proteomics MS assays are now becoming available, we re-measured the uPAR expression levels in both the wild-type and antisense HCT116 subclones. We utilized three different methods to validate uPAR expression – namely, Western blotting, immunofluorescence and a series of MS-based uPAR peptide PRM assays (Fig 2). As previously found on multiple occasions, semi-quantitative Western blotting and densitometric analysis or immunofluorescence consistently indicated uPAR expression was ~43% lower in the stable antisense clone HCT116^{ASuPAR} compared to the wild-type HCT116^{WT} (Fig 2a-c).

MS-based PRM assays provided a more accurate estimation of absolute quantitative uPAR level suppression in both subclones. In detail, a total of five unitypic uPAR peptides (i.e., uniquely-expressed across all published human gene variants as determined by the neXtProt peptide unicity checker[34]) were quantified by PRM analysis (Fig 2d and Fig S2). Collectively, PRM assays confirmed all prior and current Western blotting data to verify uPAR in the HCT116^{ASuPAR} subclone had decreased by 43±12% compared to HCT116^{WT} (Fig 2d and Fig S2). Additionally, immunofluorescence confirmed not only the reduction in uPAR but also importantly that the protein was expressed commensurately (primarily in the membrane). The IgG1 isotype control showed no non-specific binding. This data also confirmed the antisense construct had remained stable and that the antisense construct for uPAR continues to be expressed in the subclone. Taken together, these data guarantee biological pathway and HoC changes observed between HCT116^{WT} and HCT116^{ASuPAR} as being the result of a reproducible and stable yet only a moderate reduction in the expression of cell-surface uPAR.

Furthermore, from a traditional shotgun-MS analytical perspective, no uPAR peptides were detected in HCT116^{ASuPAR} subclone whereas four unique uPAR peptides were identified in HCT116^{WT} using the CW method whereas only two using the Tx114 method (Fig 2e) while none were seen in the whole cell lysate. The expression of uPAR in comparison to other proteins in the cell (such as

This article is protected by copyright. All rights reserved.

Accepted Article

housekeeping proteins) is very low. For this reason, the WCE method was unable to detect the same number of peptides by MS likely due to uPAR being masked by other more abundant proteins in the cell. A similar phenomenon is often observed in blood plasma studies.[35] This result reinforces that uPAR represents a particularly low-abundance membrane protein and that membrane enrichment is required to observe it in proteomics experiments.

To ensure only the highest quality PRM assays possible were used, quantitation of the PRM uPAR transitions met the highest stringency MS-based identification criteria, identical to those required for Human Proteome Project neXtProt PE1 assignment to "missing proteins". Proteins must be identified with ≥2 non-nested, unitypic (uniquely-mapping) peptides where each peptide identified must be ≥9 amino acids in length and, peptide and protein FDRs must be <1%.[36]

Protein Extraction from Whole Cells and Enriched Plasma Membranes

As the depth of membrane proteome coverage is dependent on the protein extraction methodologies used, we used three different extraction methods. These were (i) a whole cell extraction (WCE) and two membrane-enrichment methods, namely (ii) Triton X-114 phase partitioning (TX-114) and (iii) carbonate wash (CW). We aimed to maximize cell-surface plasma membrane protein identification for three reasons. Firstly, low abundant membrane proteins are difficult to isolate due to their hydrophobicity, and they are often masked by highly abundant intracellular proteins. Second, plasma membrane proteins (e.g., growth factor receptors, protease receptors, integrins) play noteworthy roles in crucial HoC - notably cell proliferation, invasion and metastasis. Finally, uPAR and its interacting partners have collectively been identified to be enriched and associated with the plasma membrane.

As revealed (Fig 3 and Table S1), a total of 5,495 discrete proteins were identified across all extraction methods using data from both cell lines ($HCT116^{WT}$ and $HCT116^{\Delta SuPAR}$). As expected, the highest number of proteins identified from WCE (4,123), followed by TX-114 (3,966) and finally CW (3,614). Upon UniProt GO annotation, CW identified the highest number and most comprehensive coverage of membrane proteins (860 membrane proteins out of a total of 3,614 total; 23.8%; Fig 3a).

This article is protected by copyright. All rights reserved.

Accepted Article

A proportional Venn diagram visualizes the number of common and unique proteins obtained from the three extraction methods (Fig 3b). A total of 2,348 common proteins were identified, with WCE identifying the highest number of unique proteins (i.e., 843). Interestingly, although CW identified fewer unique overall proteins (422), it did result in the unique identification of the highest number of plasma membrane proteins (128). Additionally, 1,370 proteins (368+580+422) were observed from ≥1 membrane enrichment method as compared to WCE alone (Fig 3b), supporting our tenet that multiple extraction strategies allow us to "dig deeper" into the HCT116 membrane proteome. WCE also identified many plasma membrane proteins (112) not detected by either of the other membrane-enrichment methods, indicating value-add. A full list of protein identifications from each extraction method is illustrated in Table S1.

HCT116^{WT} vs HCT116^{ASuPAR} Differentially-Expressed Proteins

To ensure accurate protein quantification, only proteins present in all triplicates from any extraction method were further considered. A volcano plot (Fig 4) mapped all quantified proteins across all extraction methods, illustrating differentially expressed proteins (FC ≥1.5, p-value ≤0.05 and ≤1% FDR at the peptide level). Additionally, an interactive volcano plot was generated (Fig S3) for comprehensive visualization of specific protein expression from the different extraction methods. From WCE, a total of 2,464 proteins were quantifiable across all replicates. Of these 150 proteins (6%) were differentially-expressed when uPAR was suppressed (101 proteins↑ and 49 proteins↓). Using TX-114 and CW, 2,302 and 2,045 proteins were reproducibly quantified respectively, with 300 (13%; 221↑, 79↓) and 250 (11.7%; 149↑, 91↓) significantly dysregulated after uPAR suppression (Fig 4). Of all differentially-expressed proteins identified from plasma membrane 55 out of total 150 (37%) were from WCE, 105 of 300 (35%) from TX-114 and 74 of 250 (30%) from CW. This confirms a significant proportion of the plasma membrane proteome (>30%) is found to change after a moderate reduction in uPAR - irrespective of the protein extraction employed.

This article is protected by copyright. All rights reserved.

Accepted Article

Our data identified a total of 77 proteins (52↑ and 15↓) that were up- or down-regulated in the HCT116^{WT} subclone compared to HCT116^{ASuPAR}, from either two or all extraction methods (Table 1 and Table S2). Of those, 36 specific proteins (30↑ and 6↓) were identified as plasma membrane proteins. Remarkably, major uPAR-interacting partners like EGFR, integrin β4, caveolin and VN and other indirect interacting partners were found to be consistently up-regulated in HCT116^{WT} compared to HCT116^{ASuPAR}. To validate our MS results, the expression levels of selected proteins (e.g., EGFR and integrin β4) were examined by western blotting analysis. Both EGFR and integrin β4 expressions were significantly higher in HCT116^{WT} compared to HCT116^{ASuPAR} (Fig S4). Moreover, we found a total of 13 different proteoforms of histocompatibility antigen (HLA) complex class I that were up-regulated in HCT116^{WT} cells (Table 1). The significance of these findings and how they fit in the broader context of cancer biology are discussed below.

Signaling Pathways Regulating Hallmarks of Cancer

MS-based proteome experiment generates enormous datasets. Quantitative proteomics allows a comprehensive view of biological pathways, processes and interactions disturbed/altered in human cancer. However, generating large datasets poses challenges, especially with regards to deciphering overall biochemical impact and drawing judicious biological conclusions.

As illustrated above, we demonstrated changes in individual protein expression against a subset of dysregulated proteins based solely on function. However, it was challenging to intelligently interpret the outcome of simultaneous change in hundreds of proteins, especially when data were derived from 3 different extraction methods. To overcome this challenge, we combined data from all extractions, increasing the comprehensiveness of the data while ensuring no data was lost. To then contextualize pathway changes to human cancer, we categorized all differentially-expressed proteins based on each protein's biological characteristics and function and then superimposed these functions onto ten known HoCs.[3] As anticipated, several proteins exhibiting multiple "moonlight" functions that could

This article is protected by copyright. All rights reserved.

Accepted Article

be categorized across several HoCs. IPA[20] was then used for each categorized group (Fig 5) to identify which signaling pathways were activated or inhibited for each of the ten distinct HoC.

Eight of the ten revised HoCs were found to contain some change associated with them. Of these particular HoCs, namely, resisting cell death, activating invasion/metastasis and sustaining proliferation, were connected to the highest number of differentially-expressed proteins (Fig 5, Table S3). Fewer proteins were associated with each of the remaining five HoCs, namely deregulating cellular energetics, evading growth suppression, avoiding immune destruction, tumor promoting inflammation and inducing angiogenesis. The final two HoCs (enabling replicative immortality and genome instability/mutation) were unable to be analyzed in depth because of an insufficient number of proteins being associated with these HoCs.

Resisting Cell Death: Data shows that the PI3K/AKT pathways were more highly activated while apoptosis was more inhibited in $HCT116^{WT}$ (higher uPAR) compared to $HCT116^{ASuPAR}$ (Fig 5). As PI3K/AKT signaling regulates numerous biological processes including cell growth, differentiation, survival, proliferation and migration,[9] it is not surprising that apoptosis was simultaneously suppressed. This observation was supported by a concurrent protein network study (Fig S5a) demonstrating that the apoptosis-promoting transcription regulator chromatin remodelling complex proteins SMARCC1, SMARCE1, SAMARCA, TRIM28 and NPS1 were repressed in $HCT116^{WT}$. Another critical observation was that the activation of HGF (hepatocyte growth factor, scatter factor) signaling, which is a multi-functional cytokine known to play a central role in growth, angiogenesis, tumorigenesis, motility, invasion and tissue regeneration by activating tyrosine kinase signaling cascade after binding to the proto-oncogene c-Met receptor.[37] In particular, interaction with heparan sulphate allows HGF to complex with c-Met so that it transduces intracellular signals leading to cell division and migration.[37]

Sustaining Proliferative Signaling: Activation of ERK/MAPK and calcium signaling (Fig 5) in our data was consistent with a previous study[38] showing $HCT116^{WT}$ proliferates at a higher rate than $HCT116^{ASuPAR}$. High uPAR expression was demonstrated to significantly activate intrinsic

This article is protected by copyright. All rights reserved.

Accepted Article

proliferative signaling pathways by up-regulating several key integrins (Fig S5b). Additionally, the inhibition of the STAT3 pathway supports previous works finding that STAT3 acts as a tumor suppressor.[39] Importantly, both Rho family GTPases and CDC42 signaling pathways were also activated in HCT116^{WT} relative to HCT116$^{\Delta SuPAR}$. The family of GTPase proteins includes Cdc42, Rho and Rac which are activated by cytokines/integrins and regulate a wide range of biological processes, like mitogenesis, apoptosis and tumorigenesis.[40] It has been previously suggested that direct interaction of uPAR/α3β1/laminin-5 is involved in the activation of Rho family GTPase Cdc42 in oral cancer[41], a finding recapitulated in this study.

Activating Invasion and Metastasis: Activation of GNRH (gonadotropin-releasing hormone) signaling was observed in HCT116^{WT} (Fig 5) relative to HCT116$^{\Delta SuPAR}$. Although expressions of GNRH and its receptor have been shown in many human epithelial cancers as a part of an autocrine regulatory system,[42], this is the first demonstration of any relationship with uPAR. HGF and Cdc42 signaling pathways were shown to be activated, illustrating the multiplicity of pathways active across different HoCs. Interestingly, unlike other up-regulated Rho members, RHOA was down-regulated in HCT116^{WT} (Fig S5c). This observation partially explains a previous study,[38] suggesting that similar down-regulation was necessary to permit ERK-dependent signaling via uPAR, which in turn activates Rac, driving cellular motility. Previous studies in our laboratory have supported this hypothesis, and we have been able to demonstrate a critical role for ERK signaling in CRC invasion/metastasis via uPAR specifically through its interaction with the epithelial-restricted and wound-healing-related integrin dimer αvβ6.[10, 11]

Other HoCs: The expected activation of integrin and ephrin receptor signaling in evading growth suppressors supports known phenotypic observations associated with metastasis (Fig 5). Similarly, although a role for nitric oxide signaling in inducing angiogenesis has been reported previously,[43] the observation of nitric oxide-related pathways in CRC in relation to uPAR has mostly been related to migration and proliferation rather than angiogenesis.[44]

This article is protected by copyright. All rights reserved.

Accepted Article

Metastasizing cancer cells are adept at shunting energy resources to fuel growth and movement. Although described by Hanahan and Weinberg as an "emerging" HoC,[3] there is considerable evidence supporting the deregulation of cellular energetics in metastasis.[45] One interesting observation made here under the HoC of deregulation of cellular energetics was the activation of calcium signaling in $HCT116^{WT}$. Calcium signaling not only has implications in cellular energetics but other aspects of signaling as well.

Collectively, this data verified the framework and showed activation of many anticipated pathways and HoCs whilst providing many novel and biologically relevant means to visualize cancer proteomics data.

Discussion

Whereas many genomic and biochemical pathways have been implicated in "driving" cancer metastasis, no specific gene/s or chromosomal abnormality appear to be primarily responsible. The processes of invasion and subsequent metastasis likely requires coordinated dysregulation of many genes[46] capable of dramatically altering the expressed proteome of both transformed epithelial tumors and resident normal host tissues in a manner that effects many of the ten hallmarks of cancer (HoCs).[2, 3] Identification of proteogenomic (i.e., both genomic and proteomic) drivers for HoCs is critical to improve our understanding of the biology complexity of human cancers. Equally, the discovery of new therapeutic modalities for the treatment of metastasis should be based upon changes in cancer (proteogenomic) biologies, not biased only by the exploration of cancer genomic/epigenomic alterations.

This article is protected by copyright. All rights reserved.

Accepted Article

Our comprehensive proteome extraction and bioinformatic analysis revealed those HoCs most influenced by a moderate decrease in cell-surface uPAR. The data demonstrate reversal of many central signaling pathways involved in HoCs by a moderate reduction (43%) in uPAR, notably superimposed upon a 'classical' CRC DNA instability and gene mutational background (i.e., *MSI*, *CIMP*, *KRAS* and *PIK3CA*) and recapitulated in the HCT116 cell line.[7, 19]

Membrane-enrichment enhances comprehensive human proteome coverage

The recent My Cancer Genome (MCG) database (May/2019) indicates that ≥70% of all FDA-approved targeted anti-cancer drugs currently in clinical trials target plasma membrane proteins.[47] This observation is likely because they are ; (i) more easily targeted by molecules unable to penetrate the cytoplasm/nucleus, (ii) capable of playing central roles in cancer signaling, (ii) central to cancer metastasis, and (iv) able to drive many HoCs.

Plasma membrane proteins, however, are often difficult to study mainly due to difficulties in extracting, digesting and analyzing them at high MS stringency such as those required by the Human Proteome Project.[48] Proteins with transmembrane domains represent a substantial proportion of what is colloquially termed the "missing proteins" (i.e., those without high stringency peptide MS evidence for their protein existence) classified as PE2-4 proteins according to neXtProt release 2/2018.[48] In recent years, several proteomic methods have been introduced to allow an analysis of enriched membrane protein preparations specifically. These include; (i) plasma membrane phase partitioning by centrifugation with/without hydrophilic polyethylene oxide chain non-ionic surfactant detergents (i.e., TritonX-114 or sucrose) and/or (ii) high pH carbonate washing, that strips many loosely attached cytoplasmic proteins.

This study demonstrated that employing multiple extraction methods under a single experimental design does indeed allow more comprehensive human proteome coverage (i.e., higher % of both

This article is protected by copyright. All rights reserved.

Accepted Article

cytoplasmic and plasma membrane proteomes). Comparable comprehensive coverage to that achieved using genomics is essential for understanding the overall effects of perturbing particular variables/biologies (here moderate uPAR suppression) and generating a more holistic understanding of how these the human proteome influences key HoCs.

Signaling pathways regulating HoCs are reversed by uPAR suppression

HoCs provide a comprehensive, logical framework from which to explain many complex biologies underpinning cancer and metastasis.[2, 3] These are based on genotypic to phenotypic observations, forecasting how normal cells acquire certain signaling pathway alterations resulting from transformation. Most data used to elucidate the HoCs have been genomic and/or phenomic in nature and reinforces the suggestion that cancer is principally a genomic/mutational disease. The cell line HCT116 used here possesses a series of CRC-related DNA instabilities (*MSI*, *CIMP*) and mutations (*KRAS, PIK3CA*).[7, 19] Our study demonstrates that a moderate reduction (↓43%) in the expression of a lynchpin protein uPAR can reverse many central signaling pathways involved in 8/10 central HoCs (Fig 5), potentially with an impact similar to that resulting from expression of driver mutations.

Data presented here substantiates that the major HoCs affected by uPAR involve resisting cell death, invasion/metastasis, and sustaining proliferation. Results closely parallel uPAR cancer biologies revealed by analysis using the recent novel CHAT PubMed text-mining engine.[21] CHAT analysis suggests uPAR is primarily involved in invasion/metastasis, with a strong secondary association in sustaining proliferation (including related apoptosis biologies; Fig S6).

Interestingly, higher uPAR expression in $HCT116^{WT}$ cells closely parallels elevated expression of well-established metastasis and uPAR-interacting proteins (e.g., EGFR, caveolin, VN and integrin β4 (Table 1)). It is well-known that the complexes of uPA-uPAR-VN trigger ECM proteolysis through the

This article is protected by copyright. All rights reserved.

Accepted Article

activation of plasmin and downstream MMP's.[12] The complex also activates intracellular ERK/MAPK signaling meditated through its interaction with αvβ1 or αvβ4 integrins, as demonstrated by recruitment of caveolin.[12, 49] Furthermore, EGFR mediates the uPAR/integrins/fibronectin induced signaling pathways that control tumor cell growth, proliferation, and resistance to apoptosis.[49] Collectively, recognizing how the expression of this single critical interacting nodal hub protein has an impact on HoCs has the potential to accelerate the discovery of novel protein therapeutic targets, and perhaps allow for a clearer understanding of the biologies underlying metastatic disease.

Elevated uPAR leads to evasion of cancer cell death by inhibiting apoptosis

Apoptosis induction is one of the most common events observed in human cancers. It appears to be equally so when uPAR is knocked-down.[50, 51] A regulatory circuitry involving apoptosis, necrosis and autophagy can be triggered by intrinsic intracellular signals such as DNA damage and oxidative stress, whilst extrinsic pathways are triggered by ligand binding to receptors, like tumor necrosis factor receptor (TNFR) family members. Thus, not surprisingly and consistent with previous reports, apoptosis signaling was inhibited in more highly uPAR expressing HCT116^{WT} relative to HCT116^{ASuPAR} (Fig 5).[50, 51] Similarly, the seminal cellular survival mTOR pathway was lower in HCT116^{WT} (Fig 5), as observed previously.[52] Another notable observation was the dysregulation of proteasome function, a multi-catalytic enzyme complex recognized as playing a significant role in tumorigenesis[53] and whose inhibition leads to increased apoptosis.[54] Our data provides intriguing evidence that uPAR may affect the ubiquitin-proteasome system, although further work is required to confirm this connection.

uPAR promotes cancer cell proliferation and activates invasion/metastasis

Our observations recapitulate elevated uPAR as influencing many well-known cancer pathways, like elevated ERK/MAPK and suppressed STAT3 and PTEN (phosphatase and tensin homolog) signaling.

This article is protected by copyright. All rights reserved.

Accepted Article

Inhibition of STAT3 signaling (Fig 5) supports other information demonstrating STAT3 as a tumor suppressor.[39] Additionally, uPAR and/or MMP9 knock-down in a medulloblastoma cell allows cells to revert to higher apoptosis rates with reduced proliferation, both of which were found to be driven by EGFR/STAT3 signaling.[51] Similarly, PTEN signaling was lower in HCT116^{WT} cells (Fig 5). PTEN is known to function as a tumor suppressor by negatively regulating Akt pathways,[55] while an earlier study demonstrated PTEN down-regulation by uPAR occurs in an integrin-dependent manner.[56] How PTEN works in metastasis is not yet clear, but our study may provide a better understanding of potential roles.

Sustaining proliferative signaling is intimately linked to the invasion/metastasis activation with uPAR but not necessarily through uPA.[57] Proliferative signalling is partially thought to be regulated by the Rho family of GTPases, critical members of the RhoGDI pathway.[40] RhoGDI signaling consists of several molecular switches that regulate cell proliferation and membrane trafficking. This regulation appears to be protein complex and context-dependent, especially in that precise RhoGDI function across different cancer subtypes can be distinctly unique.[40] Our study provides the first evidence for a connection between uPAR and RhoGDI pathway, suggesting inactivation allows Rho GTPases signaling.

uPAR promotes immune evasion via HLA class I histocompatibility antigen complex

Avoidance of immune destruction is a fundamental survival tactic used by cancer cells to evade our omnipresent immune system. Our data indicated a total of 13 different HLA class I isoforms were associated with higher uPAR expression in HCT116^{WT} (Table 1). Many of these isoforms were associated with immune-related signaling pathways, including neuroinflammation, dendritic cell maturation and NFAT and B lymphocyte PI3K signaling (Fig 5 and Table S3). HLA class I is one of three major types of human MHC (major histocompatibility complex) class I cell surface receptors that

This article is protected by copyright. All rights reserved.

Accepted Article

represent an important cancer immune escape mechanism, and that has been shown to play a critical role in immune recognition of "non-self" peptides or aberrantly expressed proteins.[58] Up-regulation of HLA (Fig S5f) suggests that HCT116^{WT} cells are likely to present foreign neoepitopes antigens on their surface. However, the concurrent activation of signaling to recruit immunosuppressive macrophages, CD4+, CD25+ regulatory T cells[58] (by activating PKC signaling in T lymphocytes) may, in the case of tumors, skew the response towards a more immunosuppressive phenotype. Together, our study exposes possible involvement of new immune-related signaling pathways regulated in CRC by elevated uPAR.

Conclusion

This study demonstrates, for the first time, that reduction in expression of a "lynchpin" protein (uPAR) superimposed on a common CRC mutational background can negate many HoCs. Similar global biochemical and phenotypic HoC changes have previously been thought to be principally "driven' by CRC driver mutations. Our findings were possible because of the extensive proteome coverage afforded by enrichment of membrane proteins in addition to whole cell lysates, with a comprehensive bioinformatics approach and a focus on their role in changing HoCs, in a single experiment.

Despite broad literature demonstrating the importance of uPAR in human cancers, effective direct uPAR-targeted and/or uPAR interaction-targeted MAbs, peptides and small molecule tools for optimal diagnosis, imaging and therapy remain to be effectively developed, exploited and translated to the clinic. In this regard, the NIH Clinical Trials database indicates at least seven breast and prostate cancer clinical trials currently underway targeting either uPAR or suPAR. uPAR (and its

This article is protected by copyright. All rights reserved.

interactome partners) are re-emerging as promising targets capable of distinguishing invasive tumors in metastasis.

Acknowledgements: The authors acknowledge and thank by Dr Alison Kan, Sachini Fonseka, Sean Barton, Adam Bentamy, Iveta Slapetova and Ilze Simpson for contributions to initial aspects of this study. We thank Professor EKO Kruithof for the supply of uPAR cDNA in 1990, Dr Yao Wang for production and supply of stably-transfected HCT116^{ASuPAR} cell line, to the Cancer Institute NSW for an ECR fellowship 15/ECF/1-38 (SBA), the NHMRC for project grant #1010303, Cancer Council NSW RG19-04, RG10-04 & RG08-16 (MSB & ECN) and the iMQRES funding from Macquarie University (SA, SS). This study is a collaboration with HUPO's Cancer-HPP initiative and the International Cancer Proteogenomics Consortium.

Author Contributions: MSB initiated this study, obtained uPAR vectors from EKOK and devised construction of HCT116^{uPARAS} line. MSB, SBA, AM, ECN designed experiments. SBA, AM, SA, SS performed experiments. DP performed statistical analysis. SBA, AM conceived Ingenuity data analysis against hallmarks of cancer. SBA, AM, DP, MSB, SA prepared figures and tables. All authors contributed to writing/reviewing of each manuscript version.

Supporting information: Supporting Figures S1-S6 and Supporting Tables S1-S3 are available online. Mass spectrometry data is available through the ProteomeXchange consortium via the PRIDE partner repository with the dataset identifier **PXD010219**.

This article is protected by copyright. All rights reserved.

Accepted Article

Conflict of Interest Statement: The authors declare no conflicts of interest.

Accepted Article

References

[1] X. X. Jie, X. Y. Zhang, C. J. Xu, Oncotarget 2017, 8, 81558.

[2] D. Hanahan, R. A. Weinberg, Cell 2000, 100, 57.

[3] D. Hanahan, R. A. Weinberg, Cell 2011, 144, 646.

[4] S. Valastyan, R. A. Weinberg, Cell 2011, 147, 275.

[5] S. A. Forbes, D. Beare, N. Bindal, S. Bamford, S. Ward, C. G. Cole, M. Jia, C. Kok, H. Boutselakis, T. De, Z. Sondka, L. Ponting, R. Stefancsik, B. Harsha, J. Tate, E. Dawson, S. Thompson, H. Jubb, P. J. Campbell, Current protocols in human genetics 2016, 91, 10.11.1.

[6] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Gut 2017, 66, 683.

[7] D. Ahmed, P. W. Eide, I. A. Eilertsen, S. A. Danielsen, M. Eknaes, M. Hektoen, G. E. Lind, R. A. Lothe, Oncogenesis 2013, 2, e71.

[8] T. Armaghany, J. D. Wilson, Q. Chu, G. Mills, Gastrointestinal Cancer Research : GCR 2012, 5, 19.

[9] A. De Luca, M. R. Maiello, A. D'Alessio, M. Pergameno, N. Normanno, Expert opinion on therapeutic targets 2012, 16 Suppl 2, S17.

[10] D. Cantor, I. Slapetova, A. Kan, L. R. McQuade, M. S. Baker, Journal of proteome research 2013, 12, 2477.

[11] S. B. Ahn, A. Mohamedali, S. Anand, H. R. Cheruku, D. Birch, G. Sowmya, D. Cantor, S. Ranganathan, D. W. Inglis, R. Frank, M. Agrez, E. C. Nice, M. S. Baker, Journal of proteome research 2014, 13, 5956.

[12] H. W. Smith, C. J. Marshall, Nature reviews. Molecular cell biology 2010, 11, 23.

[13] R. G. Saldanha, N. Xu, M. P. Molloy, D. A. Veal, M. S. Baker, Journal of proteome research 2008, 7, 4792.

[14] M. C. Boonstra, H. W. Verspaget, S. Ganesh, F. J. Kubben, A. L. Vahrmeijer, C. J. van de Velde, P. J. Kuppen, P. H. Quax, C. F. Sier, Curr Pharm Des 2011, 17, 1890.

This article is protected by copyright. All rights reserved.

Accepted Article

[15] R. Aebersold, J. N. Agar, I. J. Amster, M. S. Baker, C. R. Bertozzi, E. S. Boja, C. E. Costello, B. F. Cravatt, C. Fenselau, B. A. Garcia, Y. Ge, J. Gunawardena, R. C. Hendrickson, P. J. Hergenrother, C. G. Huber, A. R. Ivanov, O. N. Jensen, M. C. Jewett, N. L. Kelleher, L. L. Kiessling, N. J. Krogan, M. R. Larsen, J. A. Loo, R. R. Ogorzalek Loo, E. Lundberg, M. J. MacCoss, P. Mallick, V. K. Mootha, M. Mrksich, T. W. Muir, S. M. Patrie, J. J. Pesavento, S. J. Pitteri, H. Rodriguez, A. Saghatelian, W. Sandoval, H. Schluter, S. Sechi, S. A. Slavoff, L. M. Smith, M. P. Snyder, P. M. Thomas, M. Uhlen, J. E. Van Eyk, M. Vidal, D. R. Walt, F. M. White, E. R. Williams, T. Wohlschlager, V. H. Wysocki, N. A. Yates, N. L. Young, B. Zhang, Nature chemical biology 2018, 14, 206.

[16] S. B. Ahn, C. Chan, O. F. Dent, A. Mohamedali, S. Y. Kwun, C. Clarke, J. Fletcher, P. H. Chapuis, E. C. Nice, M. S. Baker, PloS one 2015, 10, e0117786; K. L. Liu, J. H. Fan, J. Wu, Clinical laboratory 2017, 63, 871.

[17] G. Eden, M. Archinti, F. Furlan, R. Murphy, B. Degryse, Current pharmaceutical design 2011, 17, 1874.

[18] M. G. Brattain, D. E. Brattain, W. D. Fine, F. M. Khaled, M. E. Marks, P. M. Kimball, L. A. Arcolano, B. H. Danbury, Oncodevelopmental biology and medicine : the journal of the International Society for Oncodevelopmental Biology and Medicine 1981, 2, 355.

[19] K. Kleivi, M. R. Teixeira, M. Eknaes, C. B. Diep, K. S. Jakobsen, R. Hamelin, R. A. Lothe, Cancer genetics and cytogenetics 2004, 155, 119.

[20] A. Kramer, J. Green, J. Pollard, Jr., S. Tugendreich, Bioinformatics (Oxford, England) 2014, 30, 523.

[21] S. Baker, I. Ali, I. Silins, S. Pyysalo, Y. Guo, J. Hogberg, U. Stenius, A. Korhonen, Bioinformatics (Oxford, England) 2017, 33, 3973.

[22] Y. Wang, X. Liang, S. Wu, G. A. Murrell, W. F. Doe, International journal of cancer 2001, 92, 257.

[23] A. Lee, D. Kolarich, P. A. Haynes, P. H. Jensen, M. S. Baker, N. H. Packer, Journal of proteome research 2009, 8, 770.

[24] M. D. Abràmoff, P. J. Magalhães, S. J. Ram, Biophotonics International 2004, 11, 36.

[25] S. Adhikari, L. Chen, P. Huang, R. Tian, Methods in molecular biology (Clifton, N.J.) 2017, 1662, 45.

[26] J. A. Vizcaino, A. Csordas, N. del-Toro, J. A. Dianes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, Q. W. Xu, R. Wang, H. Hermjakob, Nucleic acids research 2016, 44, D447.

[27] K. A. Neilson, T. Keighley, D. Pascovici, B. Cooke, P. A. Haynes, Methods in molecular biology (Clifton, N.J.) 2013, 1002, 205.

This article is protected by copyright. All rights reserved.

Accepted Article

[28] B. Zybailov, A. L. Mosley, M. E. Sardiu, M. K. Coleman, L. Florens, M. P. Washburn, Journal of proteome research 2006, 5, 2339.

[29] Y. Kojima, N. Otsuki, M. Kubo, J. Kitamoto, E. Takata, H. Saito, K. Kosaka, N. Morishita, N. Uehara, T. Shirakawa, K. I. Nibu, Cancer gene therapy 2018, 25, 274; Z. Wang, Cancer and Noncoding RNAs 2018, 203.

[30] Y. Lin, N. Peng, H. Zhuang, D. Zhang, Y. Wang, Z. C. Hua, BMC cancer 2014, 14, 639.

[31] M. Kosicki, K. Tomberg, A. Bradley, Nature biotechnology 2018, 36, 765; S. W. Cho, S. Kim, Y. Kim, J. Kweon, H. S. Kim, S. Bae, J. S. Kim, Genome research 2014, 24, 132; X. H. Zhang, L. Y. Tee, X. G. Wang, Q. S. Huang, S. H. Yang, Molecular therapy. Nucleic acids 2015, 4, e264.

[32] J. L. Yang, D. Seetoo, Y. Wang, M. Ranson, C. R. Berney, J. M. Ham, P. J. Russell, P. J. Crowe, International journal of cancer 2000, 89, 431.

[33] N. Ahmed, K. Oliva, Y. Wang, M. Quinn, G. Rice, British journal of cancer 2003, 89, 374.

[34] M. Schaeffer, A. Gateau, D. Teixeira, P. A. Michel, M. Zahn-Zabal, L. Lane, Bioinformatics (Oxford, England) 2017, 33, 3471.

[35] S. H. Tan, A. Mohamedali, A. Kapur, M. S. Baker, Journal of proteome research 2013, 12, 2399.

[36] M. S. Baker, S. B. Ahn, A. Mohamedali, M. T. Islam, D. Cantor, P. D. Verhaert, S. Fanayan, S. Sharma, E. C. Nice, M. Connor, S. Ranganathan, Nature communications 2017, 8, 14271.

[37] T. Nakamura, S. Mizuno, Proceedings of the Japan Academy. Series B, Physical and biological sciences 2010, 86, 588.

[38] E. Vial, E. Sahai, C. J. Marshall, Cancer cell 2003, 4, 67.

[39] M. Musteanu, L. Blaas, M. Mair, M. Schlederer, M. Bilban, S. Tauber, H. Esterbauer, M. Mueller, E. Casanova, L. Kenner, V. Poli, R. Eferl, Gastroenterology 2010, 138, 1003.

[40] A. Dovas, J. R. Couchman, The Biochemical journal 2005, 390, 1.

[41] Z. Shi, Y. Liu, J. J. Johnson, M. S. Stack, Molecular and cellular biochemistry 2011, 357, 151.

[42] P. Limonta, M. Montagnani Marelli, S. Mai, M. Motta, L. Martini, R. M. Moretti, Endocrine reviews 2012, 33, 784.

[43] S. K. Choudhari, M. Chaudhary, S. Bagde, A. R. Gadbail, V. Joshi, World journal of surgical oncology 2013, 11, 118.

[44] T. Zhuang, B. Chelluboina, S. Ponnala, K. K. Velpula, A. A. Rehman, C. Chetty, E. Zakharian, J. S. Rao, K. K. Veeravalli, BMC cancer 2013, 13, 590.

This article is protected by copyright. All rights reserved.

Accepted Article

[45] F. A. Urra, F. Munoz, A. Lovy, C. Cardenas, Frontiers in oncology 2017, 7, 118.

[46] A. P. Mazar, Clinical cancer research : an official journal of the American Association for Cancer Research 2008, 14, 5649.

[47]

[48] G. S. Omenn, L. Lane, C. M. Overall, F. J. Corrales, J. M. Schwenk, Y. K. Paik, J. E. Van Eyk, S. Liu, M. Snyder, M. S. Baker, E. W. Deutsch, Journal of proteome research 2018.

[49] D. Liu, J. Aguirre Ghiso, Y. Estrada, L. Ossowski, Cancer cell 2002, 1, 445.

[50] C. S. Gondi, N. Kandhukuri, D. H. Dinh, M. Gujrati, J. S. Rao, International journal of oncology 2007, 31, 19.

[51] R. R. Kotipatruni, A. K. Nalla, S. Asuthkar, C. S. Gondi, D. H. Dinh, J. S. Rao, PloS one 2012, 7, e44798.

[52] R. T. Abraham, J. J. Gibbons, Clinical cancer research : an official journal of the American Association for Cancer Research 2007, 13, 3109.

[53] Y. Tu, C. Chen, J. Pan, J. Xu, Z. G. Zhou, C. Y. Wang, International journal of clinical and experimental pathology 2012, 5, 726.

[54] S. Frankland-Searby, S. R. Bhaumik, Biochimica et biophysica acta 2012, 1825, 64.

[55] Z. Chen, L. C. Trotman, D. Shaffer, H. K. Lin, Z. A. Dotan, M. Niki, J. A. Koutcher, H. I. Scher, T. Ludwig, W. Gerald, C. Cordon-Cardo, P. P. Pandolfi, Nature 2005, 436, 725.

[56] M. Unseld, A. Chilla, C. Pausz, R. Mawas, J. Breuss, C. Zielinski, G. Schabbauer, G. W. Prager, Thrombosis and haemostasis 2015, 114, 379.

[57] M. Jo, S. Takimoto, V. Montel, S. L. Gonias, The American journal of pathology 2009, 175, 190.

[58] B. D. Tait, Human immunology 2000, 61, 158.

This article is protected by copyright. All rights reserved.

Accepted Article

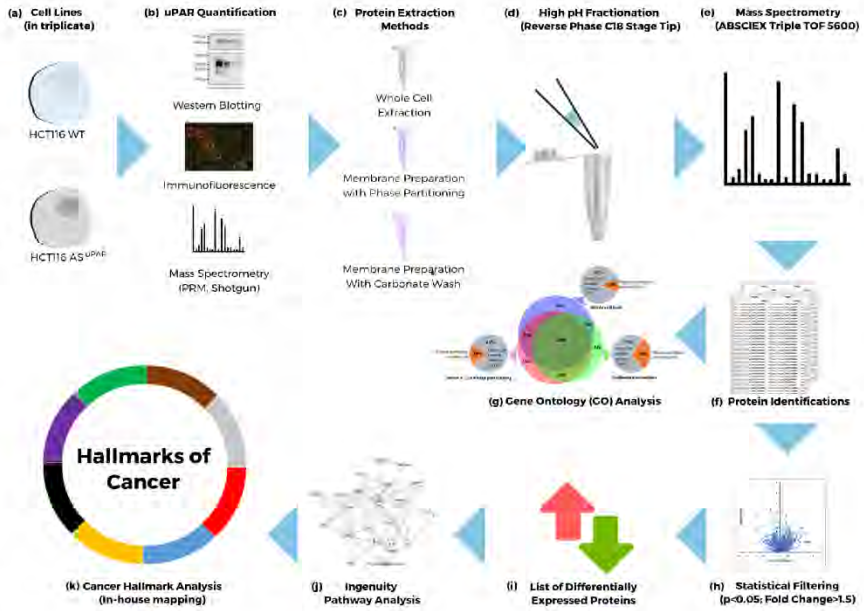
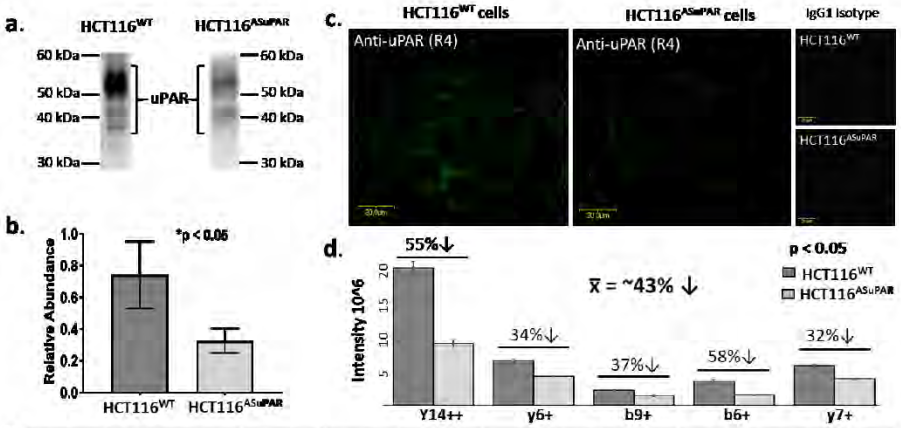


Figure 1. Experimental work flow. (**a**) The human colon carcinoma cell line, HCT116^{WT} and HCT116^{ASuPAR} (reduced uPAR expression by ~43%) cells were grown in triplicate. (**b**) The uPAR expression level was validated by Western blotting, immunofluorescence, shotgun proteomics and targeted proteomics (parallel reaction monitoring-PRM) in HCT116^{WT} and HCT116^{ASuPAR} cells. (**c**) Protein extracts were prepared by three different extraction methods including a whole cell extraction method and two membrane preparation with triton X-114 phase partitioning and carbonate wash. (**d**) Extracted proteins were tryptic digested into peptides and the peptides were fractionated into 5 fractions (3%, 6%, 9%, 15% and 80% acetonitrile) using a C18 (Empore 2215) stage-tip fractionation method. (**e**) Fractionated peptides were subjected to ABSciex Triple TOF 5600 mass spectrometry for protein identifications and quantifications. (**f**) Listed all identified proteins from different extraction methods. (**g**) UniProt Gene Ontology Annotation was performed to assess the percentage of plasma membrane proteins obtained for each extraction method. (**h**) Statistical analysis was performed using unpaired t-tests for each protein with p-value thresholds of <0.05 and fold change limits of >1.5 up or down. (**i**) Combined all differentially expressed protein from all methods. (**j**) Protein pathway and network analyses were performed using Qiagen's Ingenuity Pathway Analysis (IPA) Software. (**k**) Signaling pathways regulation each cancer hallmarks were

This article is protected by copyright. All rights reserved.

analyzed using IPA results. **HCT116^{WT}**: HCT116 wild-type cells, **HCT116^{ASuPAR}**: stable anti-sense uPAR HCT116 cells.



e.

Identified uPAR sequences	Amino acid location	Amino acid length	Peptide Unicity* Check	HCT116^{WT} Method			HCT116^{ASuPAR} Method		
				WCE	TX-114	CW	WCE	TX-114	CW
TNGDCRVEECALGQDLCR	30-47	18	√	-	√	√	-	-	-
ITSLTEVVCGLDLCNQGNSGR	85-105	21	√	-	-	√	-	-	-
CNEGPILELENLPQNGR	198-214	17	√	-	√	√	-	-	-
SGCNHPDLDVQYR	291-303	13	√	-	-	√	-	-	-

Figure 2. Quantification of uPAR expression in HCT116^{WT} and HCT116^{ASuPAR} using various methods. (**a**) Western blot images of uPAR expression in both cell lines. The images were cropped from two different blots. Identical methods and same exposure time were applied for preparation of both images. Refer to Figure S1a-d for full-length Western blot images (i.e., uncropped images); (**b**) Densitometry analysis of Western blot: uPAR expression reduced by ~45% in HCT116^{ASuPAR}. Refer to Figure S1e for the full-length Western blot image used for densitometry analysis; (**c**) Immunofluorescence staining images for both cell lines demonstrating a reduction of uPAR expression in the membrane (the IgG1 isotype control showed no non-specific binding); (**d**) PRM-MS analysis: 5 uPAR peptides were quantified. Average uPAR peptides reduction in HCT116^{ASuPAR} was ~43% (refer to Figure S2 for detailed information including quantified uPAR sequences); (**e**) MS peptides identification: Unitypic peptides of ≥9 amino acid were detected in HCT116^{WT} only. More

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

peptides were detected in the CW method (4 peptides) than in Tx114 (2 peptides). **HCT116^{WT}**: HCT116 wild-type cells, **HCT116^{ASuPAR}**: stable anti-sense uPAR HCT116 cells, **PRM-MS**: parallel reaction monitoring-mass spectrometry, **WCE**: whole cell extraction, **TX-114**: Triton X-114 phase partitioning, **CW**: carbonate wash.

a.

	Protein extraction methods			
	Whole Cell Extraction	**Triton X114 Phase Partitioning**	**Carbonate Wash**	**Total protein identifications**
Total # of proteins identified	4,125	3,966	3,614	5,495
Plasma membrane proteins	732 (17.8%)	838 (21.1%)	860 (23.8%)	1,126 (20.5%)

b.

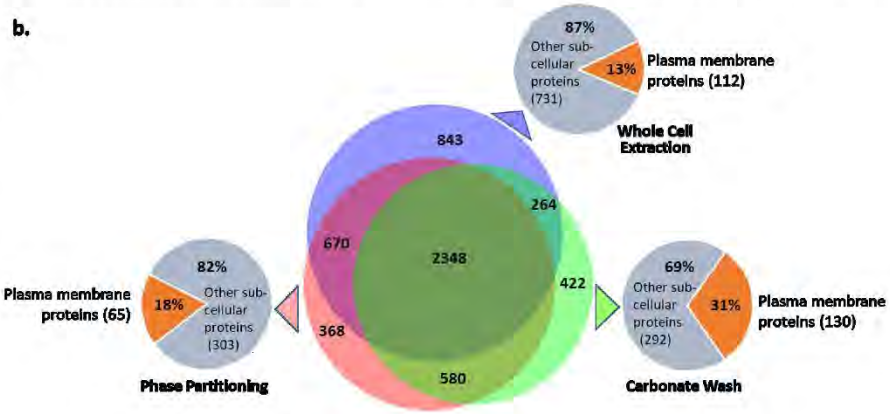


Figure 3. Gene Ontology annotation of identified protein. (**a**) number of identified total and plasma membrane proteins from different protein extraction methods, a total of 5495 proteins were unique (**b**) a proportional Venn diagram to visualize the number of unique and replicate proteins from different extraction methods. See Table S1 for a full list of all identified proteins from each protein extraction method.

This article is protected by copyright. All rights reserved.

Accepted Article

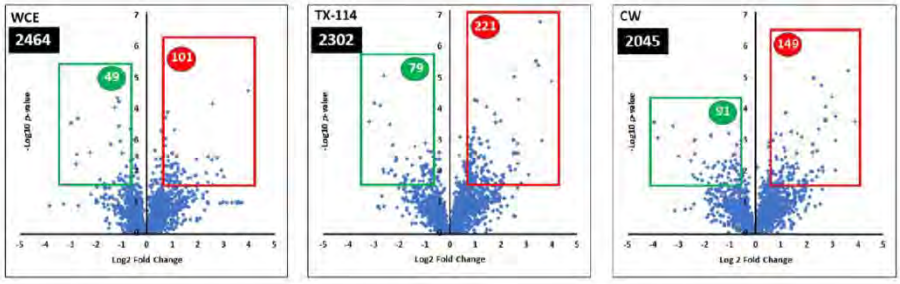


Figure 4. Volcano plots representing differentially-expressed proteins identified by each extraction method. Red boxes: significantly up-regulated proteins (fold change > 1.5; *p*-value < 0.05) in $HCT116^{WT}$ compared to $HCT116^{ASuPAR}$. **Green boxes**: significantly down-regulated proteins (fold change > 1.5; *p*-value < 0.05) in $HCT116^{WT}$ compared to $HCT116^{ASuPAR}$. **Black boxes**: total identified/quantitated proteins. See Figure S3 for interactive volcano plots for visualisation of expression levels for each protein across all extraction methods. **WCE**: whole cell extraction, **TX-114**: Triton X-114 phase partitioning, **CW**: carbonate wash.

This article is protected by copyright. All rights reserved.

Accepted Article

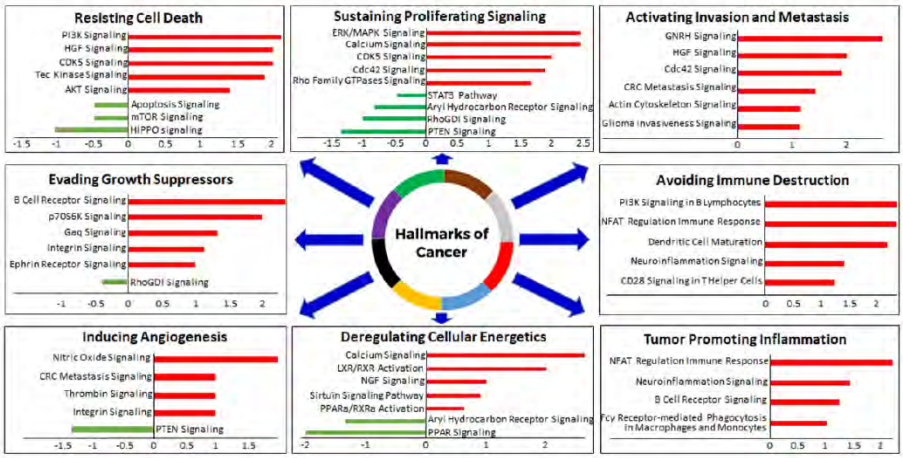


Figure 5: Alterations in signaling pathways involved in the hallmarks of cancer caused by increased CRC cellular uPAR expression. Proteomic comparison of human colon carcinoma cell lines, $HCT116^{WT}$ and $HCT116^{ASuPAR}$ (anti-sense uPAR leads to decreased expression by ~43%), was performed using the combination of three (3) extraction methodologies and quantitative proteomics followed by Ingenuity Pathways Analysis. **X-axis**: z-score; **Red bars**: activation; **Green bars**: inhibition. Detailed information of individual proteins dysregulated is available in Table S3. The signal pathway analyses were performed through the use of IPA (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis).

This article is protected by copyright. All rights reserved.

Table 1 Differentially-expressed proteins observed across ≥ two (2) independent extraction methods.

Gene Names*	WCL	TX-114	CW	Gene Names*	WCL	TX-114	CW	Gene Names*	WCL	TX-114	CW
1A02	**2.14 ↑**	**1.65 ↑**	-	COPG1	-	**2.70 ↑**	**1.89 ↑**	LONM	-	**2.70 ↑**	**2.73 ↑**
1B07	**2.21 ↑**	**1.69 ↑**	-	DD19A	-	**2.53 ↑**	**6.79 ↑**	MAP1B	**7.74 ↓**	**6.55 ↓**	-
1B13	**2.14 ↑**	**1.76 ↑**	-	DDX42	**1.81 ↓**	-	**2.25**	MVP	**2.38 ↑**	**2.43 ↑**	-
1B15	**1.95 ↑**	**1.76 ↑**	-	DHCR7	**6.06 ↑**	-	**1.52 ↑**	NQO1	-	**5.42 ↑**	**2.05 ↑**
1B18	**5.26 ↑**	**1.73 ↑**	**1.62 ↑**	DREB	**1.53 ↓**	**2.77 ↓**	-	P5CR1	**1.54 ↓**	-	**1.78 ↓**
1B44	**8.25 ↑**	**1.77 ↑**	-	EAA1	-	**2.62 ↓**	**4.13 ↓**	PALLD	**3.92 ↑**	**5.73 ↑**	**4.84 ↑**
1B48	**2.49 ↑**	**1.57 ↑**	-	EF2	-	**2.30 ↑**	**2.22 ↑**	PLD3	-	**1.91 ↑**	**2.39 ↑**
1B54	-	**1.79 ↑**	**1.54 ↑**	EFHD2	**1.74 ↓**	**3.45 ↓**	-	PSMD1	-	**1.93 ↑**	**1.87 ↑**
1C01	**2.77↑**	-	**1.61 ↑**	EGFR	**5.41 ↑**	**2.91 ↑**	-	RBM15	**1.79 ↓**	**1.55 ↓**	-
1C06	**16.13 ↑**	**1.69 ↑**	-	ENPL	-	**2.07 ↑**	**1.68 ↑**	RHG18	**2.18 ↑**	-	**2.45 ↑**

Received: 27/02/2019; Revised: 25/07/2019; Accepted: 09/08/2019

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201900026.

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

1C07	**6.92 ↑**	**1.66 ↑**	-	ERO1A	**1.64 ↑**	**3.52 ↑**	-	RPA34	**3.02 ↓**	**3.30 ↓**	-
1C17	**2.11 ↑**	-	**1.51 ↑**	ETFA	-	**2.01 ↑**	**3.36 ↑**	S38A2	**6.90 ↑**	**3.40 ↑**	**2.75 ↑**
5NTD	-	**16.12 ↑**	**4.56 ↑**	FUBP1	**1.85 ↓**	-	**3.47 ↓**	S39A7	-	**1.90 ↑**	**3.22 ↑**
ACACA	-	**1.79 ↑**	**2.04 ↑**	GARS	-	**3.47 ↑**	**2.19 ↑**	SPF27	**1.94 ↓**	**3.76 ↓**	-
ACADV	-	**1.62 ↑**	**5.37 ↑**	GBF1	-	**4.01 ↑**	**9.00 ↑**	SPIT2	**4.09 ↑**	**7.12 ↑**	-
ACLY	-	**4.23 ↑**	**1.72 ↑**	GDF15	-	**6.49 ↑**	**3.09 ↑**	SQSTM	-	**12.50 ↑**	**14.92 ↑**
AHNK	**1.67 ↓**	-	**2.57 ↓**	H2AW	**2.02 ↓**	**1.78 ↓**	**3.56 ↓**	STING	-	**5.29 ↓**	**5.58 ↓**
AL1A3	-	**10.56 ↑**	**12.67 ↑**	HBA	**7.89 ↑**	**8.69 ↑**	**6.23 ↑**	SUSD2	**6.11 ↑**	-	**2.68 ↑**
AOFA	-	**2.56 ↑**	**1.88 ↑**	HBE	-	**2.74 ↑**	**4.04 ↑**	TFR1	**1.72 ↑**	**1.56 ↑**	-
AOFB	**2.37 ↑**	**1.68 ↑**	-	HLAG	**4.01 ↑**	**1.67 ↑**	-	THOC4	-	**2.60 ↓**	**1.90 ↓**
APT	-	**2.83 ↑**	**2.77 ↑**	HMGB1	-	**2.04 ↓**	**4.72 ↓**	TM41B	**3.77 ↑**	-	**1.66 ↑**
B2MG	**3.95 ↑**	-	**3.78 ↑**	IDH3A	-	**2.56 ↑**	**4.01 ↑**	TMM59	-	**6.47 ↑**	**3.52 ↑**
CALR	-	**2.83 ↑**	**2.85 ↑**	IMDH2	-	**2.34 ↑**	**2.94 ↑**	TMX2	**3.21 ↑**	**2.26 ↑**	-
CAV3	-	**4.12 ↑**	**5.43 ↑**	ITB4	**2.18 ↑**	**1.58 ↑**	-	UBA1	-	**3.50 ↑**	**1.89 ↑**
COF1	-	**2.58 ↑**	**1.87 ↑**	ITIH2	-	**6.47 ↑**	**6.14 ↑**	VTNC	**2.24 ↑**	**3.54 ↑**	-
COPB	-	**2.05 ↑**	**2.85 ↑**	KAP0	-	**1.83 ↑**	**2.22 ↑**				

*Gene names listed as alphabet order from left top to right bottom. **WCL**: Whole Cell Lysis, **TX-114**: Triton X114 Phase Partitioning, **CW**: Carbonate Wash. Numbers with up and down narrows represent fold changes. ↑: up-regulated in $HCT116^{WT}$ compared to $HCT116^{ASuPAR}$ and ↓: down-regulated in $HCT116^{WT}$ compared to $HCT116^{ASuPAR}$. See Table S2 for detailed information including full protein names and list of all differentially expressed proteins in each method.

Accepted Article

Received: 27/02/2019; Revised: 25/07/2019; Accepted: 09/08/2019

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201900026.

This article is protected by copyright. All rights reserved.

3.6 References

1. Seyfried, T. N. & Huysentruyt, L. C. On the origin of cancer metastasis. *Crit. Rev. Oncog.* **18**, 43–73 (2013).
2. Eden, G., Archinti, M., Furlan, F., Murphy, R. & Degryse, B. The urokinase receptor interactome. *Curr. Pharm. Des.* **17**, 1874–89 (2011).
3. Wang, Y., Liang, X., Wu, S., Murrell, G. A. C. & Doe, W. F. Inhibition of colon cancer metastasis by a 3′ - end antisense urokinase receptor mRNA in a nude mouse model. *Int. J. Cancer* (2001). doi:10.1002/1097-0215(200102)9999:9999<::AID-IJC1178>3.0.CO;2-6
4. Krämer, A., Green, J., Pollard, J. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btt703
5. Baker, S. *et al.* Cancer Hallmarks Analytics Tool (CHAT): a text mining approach to organize and evaluate scientific literature on cancer. *Bioinformatics* **33**, 3973–3981 (2017).
6. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
7. Bowel cancer - Cancer Council Australia. Available at: https://www.cancer.org.au/about-cancer/types-of-cancer/bowel-cancer/. (Accessed: 13th July 2019)
8. Health, A. G. D. of. Bowel Screening.
9. Edge, S. B. & Compton, C. C. The American Joint Committee on Cancer: the 7th

Edition of the AJCC Cancer Staging Manual and the Future of TNM. *Ann. Surg. Oncol.* **17**, 1471–1474 (2010).

10. Bergström, R., Glimelius, B. & Påh1man, L. The association of preoperative serum tumour markers with Dukes' stage and survival in colorectal cancer. *Br. J. Cancer* (1995). doi:10.1038/bjc.1995.211
11. Edwards, B. K. *et al.* Annual Report to the Nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. *Cancer* (2014). doi:10.1002/cncr.28509
12. Young, G. P. *et al.* Advances in Fecal Occult Blood Tests: The FIT Revolution. *Dig. Dis. Sci.* **60**, 609–622 (2015).
13. Song, L.-L. & Li, Y.-M. Current noninvasive tests for colorectal cancer screening: An overview of colorectal cancer screening tests. *World J. Gastrointest. Oncol.* **8**, 793–800 (2016).
14. Wolf, A. M. D. *et al.* Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. *CA. Cancer J. Clin.* **68**, 250–281 (2018).
15. Lee, J. K., Liles, E. G., Bent, S., Levin, T. R. & Corley, D. A. Accuracy of Fecal Immunochemical Tests for Colorectal Cancer. *Ann. Intern. Med.* **160**, 171–181 (2014).
16. Imperiale, T. F. *et al.* Multitarget Stool DNA Testing for Colorectal-Cancer Screening. *N. Engl. J. Med.* **370**, 1287–1297 (2014).
17. Rex, D. K. *et al.* Colorectal Cancer Screening: Recommendations for Physicians

and Patients from the U.S. Multi-Society Task Force on Colorectal Cancer. *Am. J. Gastroenterol.* **112**, 1016–1030 (2017).

18. Rank, K. M. & Shaukat, A. Stool Based Testing for Colorectal Cancer: an Overview of Available Evidence. *Curr. Gastroenterol. Rep.* **19**, 39 (2017).
19. Carethers, J. M. Review: Systemic treatment of advanced colorectal cancer: Tailoring therapy to the tumor. *Therap. Adv. Gastroenterol.* **1**, 33–42 (2008).
20. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
21. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
22. de la Chapelle, A. Genetic predisposition to colorectal cancer. *Nat. Rev. Cancer* **4**, 769–780 (2004).
23. Valle, L. *et al.* Update on genetic predisposition to colorectal cancer and polyposis. *Mol. Aspects Med.* (2019). doi:10.1016/J.MAM.2019.03.001
24. Vermeulen, L. *et al.* Wnt activity defines colon cancer stem cells and is regulated by the microenvironment. *Nat. Cell Biol.* **12**, 468–476 (2010).
25. Tauriello, D. V. F. *et al.* TGFβ drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature* (2018). doi:10.1038/nature25492
26. Qu, X. *et al.* Integrated genomic analysis of colorectal cancer progression reveals activation of EGFR through demethylation of the EREG promoter. *Oncogene* **35**, 6403–6415 (2016).
27. Spano, J.-P. *et al.* Impact of EGFR expression on colorectal cancer patient prognosis and survival. *Ann. Oncol.* **16**, 102–108 (2005).
28. Schatoff, E. M., Leach, B. I. & Dow, L. E. Wnt Signaling and Colorectal Cancer.

Curr. Colorectal Cancer Rep. **13**, 101–110 (2017).

29. Fang, J. Y. & Richardson, B. C. The MAPK signalling pathways and colorectal cancer. *Lancet Oncology* (2005). doi:10.1016/S1470-2045(05)70168-6
30. Noser, J. A. *et al.* The RAS/Raf1/MEK/ERK signaling pathway facilitates VSV-mediated oncolysis: implication for the defective interferon response in cancer cells. *Mol. Ther.* **15**, 1531–6 (2007).
31. Shi, Y. & Massagué, J. Mechanisms of TGF-beta signaling from cell membrane to the nucleus. *Cell* **113**, 685–700 (2003).
32. Bachman, K. E. & Park, B. H. Duel nature of TGF-beta signaling: tumor suppressor vs. tumor promoter. *Curr. Opin. Oncol.* **17**, 49–54 (2005).
33. Calon, A. *et al.* Dependency of Colorectal Cancer on a TGF-β-Driven Program in Stromal Cells for Metastasis Initiation. *Cancer Cell* **22**, 571–584 (2012).
34. Redston, M. Carcinogenesis in the GI tract: From morphology to genetics and back again. *Modern Pathology* (2001). doi:10.1038/modpathol.3880292
35. Liu, X., Lazenby, A. J. & Siegal, G. P. Signal transduction cross-talk during colorectal tumorigenesis. *Advances in Anatomic Pathology* (2006). doi:10.1097/01.pap.0000213046.61941.5c
36. Grady, W. M. & Carethers, J. M. Genomic and Epigenetic Instability in Colorectal Cancer Pathogenesis. *Gastroenterology* (2008). doi:10.1053/j.gastro.2008.07.076
37. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).
38. Duffy, M. The Urokinase Plasminogen Activator System: Role in Malignancy.

Curr. Pharm. Des. **10**, 39–49 (2004).

39. Su, S.-C., Lin, C.-W., Yang, W.-E., Fan, W.-L. & Yang, S.-F. The urokinase-type plasminogen activator (uPA) system as a biomarker and therapeutic target in human malignancies. *Expert Opin. Ther. Targets* **20**, 551–566 (2016).
40. Smith, H. W. & Marshall, C. J. *Regulation of cell signalling by uPAR*. *Nature Reviews Molecular Cell Biology* **11**, 23–36 (Nature Publishing Group, 2010).
41. Lester, R. D., Jo, M., Montel, V., Takimoto, S. & Gonias, S. L. uPAR induces epithelial-mesenchymal transition in hypoxic breast cancer cells. *J. Cell Biol.* (2007). doi:10.1083/jcb.200701092
42. Montuori, N. *et al.* Urokinase type plasminogen activator receptor (uPAR) as a new therapeutic target in cancer. *Transl. Med. @ UniSa* (2016).
43. Boonstra, M. C. *et al.* uPAR-targeted multimodal tracer for pre- and intraoperative imaging in cancer surgery. *Oncotarget* (2015). doi:10.18632/oncotarget.3680
44. Cantor, D., Slapetova, I., Kan, A., McQuade, L. R. & Baker, M. S. Overexpression of αvβ6 integrin alters the colorectal cancer cell proteome in favor of elevated proliferation and a switching in cellular adhesion that increases invasion. *J. Proteome Res.* (2013). doi:10.1021/pr301099f

Chapter 4: Antagonism of metastasis in late-stage Colorectal Cancer (CRC)

4.1 Overview

Urokinase plasminogen activator receptor (uPAR) and integrin αvβ6 independently play an important role in metastatic progression, they are widely accepted as a marker for poor prognosis in most of the epithelial carcinoma [1]. We have discovered multiple evidence supporting membrane protein-protein interaction between uPAR and integrin αvβ6 [2–5]. We identified six prospective binding sites on uPAR for the integrin αvβ6 binding, by proximity ligation, peptide array and *in silico* structural modelling [4]. Synthetic peptides from these binding regions were generated and utilized as an interference peptide (iPEPs) against the cancer biologies that are reportedly driven by uPAR and integrin αvβ6.

Treatment of the CRC cell model expressing both integrin αvβ6 and uPAR demonstrated only two of the six iPEPs were able to bind to the cell surface and induce morphological changes [3]. High throughput mass spectrometry (MS) analysis was utilized to quantify the proteome altered by these two proteins in a CRC cell model.

Preliminary results on iPEPs binding activity to the cell surface and their ability to induce morphological changes were a part of PhD thesis submitted previously by our team member [3].

4.2 References

1. Smith, H. W. & Marshall, C. J. *Regulation of cell signalling by uPAR*. *Nature Reviews Molecular Cell Biology* **11**, 23–36 (Nature Publishing Group, 2010).
2. Saldanha, R. G. *et al.* Proteomic identification of lynchpin urokinase plasminogen activator receptor protein interactions associated with epithelial cancer malignancy. *J. Proteome Res.* (2007). doi:10.1021/pr060518n
3. Cantor, D. I. The αvβ6 integrin plays an integral role in colorectal cancer metastasis. *PhD thesis, Macquarie Uni* (2016).
4. Ahn, S. B. *et al.* Characterization of the interaction between heterodimeric αvβ6 integrin and urokinase plasminogen activator receptor (uPAR) using functional proteomics. *J. Proteome Res.* (2014). doi:10.1021/pr500849x
5. Sowmya, G. *et al.* A site for direct integrin αvβ6·uPAR interaction from structural modelling and docking. *J. Struct. Biol.* (2014). doi:10.1016/j.jsb.2014.01.001

4.3 uPAR-based interference peptides (iPEPs) inhibit cancer metastatic phenotypes in CRC cell model expressing the integrin β6

Subash Adhikari[1], David Cantor[2], Seong Beom Ahn[1], Abidali Mohamedali[3], Janne Lehtiö[4,5], Edouard C. Nice[6] and Mark S. Baker[1*]

[1]Department of Biomedical Sciences, Faculty of Medicine and Health Science, [2]Australian Proteome Analysis Facility, [3]Department of Molecular Sciences, Faculty of Science and Engineering, Macquarie University, Sydney, Australia. [4]Department of Oncology-Pathology, Karolinska Institutet, SE-17121 Solna, Sweden, [5]Clinical Proteomics Mass Spectrometry, Science for Life Laboratory, SE-17177 Stockholm, Sweden. [6]Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia.

* Corresponding Author: Mark S. Baker, Level 1, 75 Talavera Road, Macquarie University, 2109, Australia, +61 2 9850 8211, mark.baker@mq.edu.au

Abstract

Urokinase plasminogen activator receptor (uPAR) and integrin αvβ6 independently play essential roles in cancer metastasis. These proteins have both been implicated in the epithelial-mesenchymal transition (EMT), during which cancer cells acquire the ability to escape from their site of origin and spread to distal locations. The acquisition of the EMT phenotype significantly impacts the metastasis, that is responsible for most cancer-related deaths. A range of different biochemical evidence has been used to confirm the interaction between uPAR and the epithelially-restricted integrin αvβ6 (referred to

forthwith as uPAR-αvβ6), including cellular fluorescence co-localisation, *in vitro* functional proteomics analysis and *in silico* structural modelling studies. Synthetic peptides derived from the putative primary sequence found in proposed binding sites to integrin αvβ6 were utilized here as interference peptide (iPEPs) against cancer biologies reportedly driven by uPAR and αvβ6 individually or collectively. With the aid of high-resolution quantitative data-dependant acquisition (DDA) mass spectrometry (MS)-based proteomics, we also demonstrated that iPEPs interfering with uPAR-αvβ6 formation could "repurpose" many cellular biochemical pathways expressed during invasion and metastasis.

Introduction

Colorectal cancer (CRC) is the fourth most common cancer worldwide [1]. Although CRC is curative with colonoscopy in combination with adjuvant therapies (chemo-, radio- or immuno-) in the earliest AJCC stage I/II clinical early stages, a lack of population-compliant, early-stage diagnostic screening moieties results in later-stage III/IV diagnosis where cancer has already metastasized to lymph nodes and/or distal organs [2].

CRC is initially characterized by a series of mutations in genes involved in cellular proliferation and migration, eventually allowing cancer cells to acquire a metastatic phenotype. Mutations in adenomatous polyposis coli (APC) genes, RAS-family genes and TP53 [3] are commonly observed in CRC metastasis. These have been linked biochemically to alterations in the mitogen-activated protein kinases (MAPK), epidermal growth factor receptor (EGFR), p53 and transforming growth factor-beta (TGFβ) signalling pathways [4]. It is also well established that the RAS-family mutational landscape in CRC correlates with downstream signalling partially driven by the urokinase plasminogen activator receptor (uPAR) interactome [5,6]. uPAR confers downstream

signalling via i) proteolytic pathways dependent on plasminogen activation system (PAS) and ii) proteolysis cascade independent pathways mediated through interaction with lateral membrane located protein partners. Correlation of elevated uPAR levels during inflammation, tissue remodelling and in many epithelial carcinomas has been clinically established [7]. Elevated uPAR levels are proposed to be a prognostic indicator of poor CRC survival [8,9].

uPAR's fundamental and best established biochemical role is to act as a multi-domain GPI-anchored cell surface membrane receptor for the zymogen serine protease urokinase plasminogen activator (uPA) [10]. uPAR is composed of three domains (D1, D2 and D3) forming a globular structure, with the D2 and D3 domains interconnected by a short linker region. External uPAR D2 and D3 domain structures on the reverse backside of the protease-binding central cavity are primarily involved in all lateral interactions [11].

uPAR is universally accepted to playing a vital role in the PAS-mediated proteolytic cascade. PAS-dependent plasminogen activation to plasmin induces the EMT, enhances the release of growth factors and activates pro-metalloproteases that collectively contribute to increased cellular proliferation, invasion and migration [12]. uPAR can also signal via proteolysis-independent pathways upon interaction with ancillary proteins like vitronectin (Vn) that prime cells for EMT [13]. Similarly, uPAR regulates cell survival and distal metastasis through interaction with transmembrane ligands such as integrins, G-protein-coupled receptors (GPCRs) and receptor tyrosine kinases (RTKs) [13].

We have recently demonstrated the central role of uPAR in CRC by superimposing a comprehensive proteome data onto a unifying cancer framework proposed by and

recently revised by cancer researchers Douglas Hanahan and Robert Weinberg. This framework is universally known as the Hallmarks of Cancer (HoCs) [14,15].

Proteomics analysis was performed to quantify uPAR-dependent signalling changes between CRC HCT116 wild type cells expressing basal uPAR levels and HCT116^{uPARAS} cells carrying a uPAR antisense construct resulting in decreased (↓~43%) uPAR expression levels. The data shows that reductions in uPAR expression negate many cancer-related signalling pathways, some responsible for metastasis. In particular, this data supports a major contribution of uPAR in mediating multiple HoCs, particularly associated with; (i) invasion/metastasis, (ii) resisting cell death and (iii) sustaining proliferation. Furthermore, the decrease in uPAR was found to suppress many metastasis-related components of the uPAR interactome, including caveolin, EGFR, Vn and integrin β4 [16].

Integrins are a group of 24 covalently linked heterodimers composed of 18 α and eight β subunits [17] Integrin αvβ6 is a spatially and cell-restricted heterodimer consisting of both the αv and β6 subunits. Integrin β6 subunit, levels of which are determined by transcriptional activity of β6 integrin gene (ITGB6) are exclusively expressed in the epithelial cells (e.g., during wound healing, tissue remodelling and cancer) and is responsible for the regulated formation of the αvβ6 integrin heterodimer [18].

Extracellular and transmembrane domains of integrin αvβ6 are involved in activation of TGFβ and initiation of EMT respectively, whereas the αvβ6 integrin's cytoplasmic domain regulates proliferation, migration and downstream production of matrix metalloproteinase (MMPs) [19]. Integrin αvβ6 interacts with disintegrin and the MMP family of proteins and RGD motif-containing extracellular ligands such as Vn, fibronectin and tenascin-C.

Interaction of integrin αvβ6 with these proteins mediates cell-cell and/or cell-ECM adhesion, which confers traction to a cell during cell motility [20]. Expression of the integrin αvβ6 promotes signalling pathways associated with cell proliferation, migration, invasion and metastasis [19]. Similar to uPAR, integrin αvβ6 levels increase during inflammation, tissue remodelling and epithelial carcinomas. Integrin αvβ6 expression in CRC, ovarian and breast cancer is correlated with the poor patient survival [21,19].

We have previously corroborated multiple evidence on the involvement of integrin αvβ6 in CRC. Our past studies have shown the presence of strong expression of integrin αvβ6 throughout late-stage CRC tumors by immunological histopathological analysis, where expression of integrin αvβ6 was restricted to the surface of epithelial cells, at the invasive front on both benign and malignant tumors [22,23]. We further investigated the role of integrin αvβ6 signalling in CRC metastasis by establishing a direct link between integrin αvβ6 with extracellular signal-regulated kinase (ERK2) signaling in the colon cancer cell and animal models [24]. Downregulation of β6 was responsible for inhibition of MAPK activity and tumor growth *in vivo*. ERK2 bound only to integrin β6 subunit in CRC cells expressing integrin αvβ6, similarly deletion of the ERK2 binding site resulted in inhibition of tumor growth [24].

We further identified multiple components of the CRC metastasome with pull-down assays. Integrin αvβ6 was found to be co-immunoprecipitated along with uPAR in human ovarian cell line OVCA429. Reverse co-immunoprecipitation with the integrin β6 subunit antigen identified uPAR as one protein associated with integrin αvβ6. Inhibition of uPAR and integrin αvβ6 using monoclonal antibodies suppressed uPA-dependent ERK1/2 phosphorylation and cell proliferation. This data suggests some role for the interaction between integrin and uPAR mediating uPA-dependent cell proliferation. Likewise, uPAR

was found to signal through the formation of an active multi-protein complex that included integrin αvβ6 [25]. Subsequent to those studies, we further investigated the systemic effects of enhanced αvβ6 expression. This study showed inducing integrin αvβ6 in the cell-like SW480 that lacked endogenous β6, with a β6 vector construct enhanced cellular proliferation, invasion and migration, compared to a mock cell line lacking β6 expression [26]. This provides convincing evidence on a biochemically relevant interaction between uPAR and αvβ6. Our earliest data on co-immunoprecipitation of uPAR and αvβ6 provided proof these proteins act as a complex [25]. Then we substantiated the αvβ6 integrin-binding site for uPAR. Finally, we employed crystal structure analysis of uPAR-αvβ6 uPAR [27,28] even though we could only model αvβ6 on a close protein analogue, namely αvβ3. Hence, we performed an *in silico* structural modelling based on separate uPAR and integrin αvβ6 structure co-ordinates, to identify putative residues involved in uPAR-αvβ6 [29]. The putative binding sites for uPAR-αvβ6 were further experimentally verified based on a peptide array approach and proximity ligation studies, leading to the identification of 6 putative binding sites for αvβ6 to uPAR sequences [30]. Synthetic uPAR peptides generated from these six uPAR αvβ6 binding sites were utilized as an interference peptide (iPEPs) to examine their ability to antagonize uPAR-mediated metastatic biologies. Of these six iPEPs selected, two iPEPs (iPEP2 and iPEP6) bound explicitly to and induced actin spicule morphological changes CRC SW480 cells carrying a stable β6 vector construct and referred to as $SW480^{OE}$ when compared to the cells lacking β6, namely SW480 cells carrying an empty “mock” vector construct referred to as $SW480^{mock}$. In detail, iPEP2 and iPEP6 inhibited cell proliferation by up to 60% and they decreased cell migration and invasion by ~25% when compared to $SW480^{mock}$ cells as controls. Many iPEP2/6-induced

changes appear to be driven by a switch in TGFβ signalling where cells switch from being SMAD-dominant (canonical TGFβ signalling) to MAPK-dominant (non-canonical TGFβ signalling) [31].

In this study, I extended previous knowledge regarding the role of iPEPs-induced signalling changes on CRC metastatic phenotypes, with the aid of high-resolution MS analysis. Combinations of tandem-mass-tag (TMT), immobilized pH gradient isoelectric focusing (IPG-IEF) fractionation based on high-resolution isoelectric focussing (HiRIEF) [32] and high-resolution MS analysis of iPEP-treated CRC cells allowed us to reach unprecedented depth in proteome coverage. Data complement results from previous cellular assays. The combination of TMT-HiRIEF-LCMS has previously been utilized for in-depth high-throughput quantitation of the cellular proteome [32–34].

Materials and Methods

Cell lines

The SW480 CRC cell line lacking endogenous β6 expression was established by Leibovitz *et al.* [35]. Derivatives of this cell line containing either a stable full-length β6 construct (i.e., $SW480^{\beta6OE}$) or an empty vector (i.e., $SW480^{Mock}$) transfection were obtained from our collaborator Prof. Michael Agrez, University of Newcastle. These cells were utilized to assess the ability of iPEPs to inhibit uPAR and integrin αvβ6 processes.

Interference peptides

Two rationally-designed iPEPs (iPEPs 2 and iPEPs 6) based on our previous data regarding the putative binding site of integrin αvβ6 on uPAR were utilized for the antagonism of uPAR-αvβ6 biologies. Biotinylated iPEPs and respective randomly "scrambled" iPEPs isoforms were sourced commercially from AUSPEP, Melbourne, Australia.

Interference peptides (iPEPs) treatment.

All SW480 cell lines ($SW480^{\beta6OE}$, $SW480^{Mock}$) were incubated in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% foetal bovine serum (FBS), containing 500µg/mL selective agent geneticin (G418 sulphate) and grown until 80% confluent. Cells were washed twice with 1xPBS, followed by 12hr starvation in serum-free DMEM media containing 500µg/mL geneticin (SF media). Upon starvation, cell lines were washed twice with 1X PBS and treated in triplicate with either 10µM iPEP2 or iPEP6 supplemented on SF media, 10µM scrambled iPEP2 or iPEP6 supplemented on the SF media or were untreated controls. All treatments were performed for 6 hr at 37ºC in 5% CO_2 generating a total of 30 samples.

Proteomics sample preparation

Upon treatment for 6hr, cells were washed twice with ice-cold 1X PBS and lysed with whole-cell sodium deoxycholate lysis buffer (100mM TEAB, 1% sodium deoxycholate) under a probe sonicator (Branson Sonifier 450; 10 bursts at 40% amplitude, output 2 setting, repeated 3x). Lysates were heated to 95°C for 5 min to inactivate proteases. Sodium deoxycholate was removed by precipitation upon acidification of lysates to a final

concentration of 0.1% formic acid (FA) and centrifuged twice at 14,000g for 10 mins. Cell lysates were reduced (final 5mM DTT for 30mins at 60°C) and alkylated (final 14mM IAA for 30mins at RT on dark) prior to tryptic digestion. The lysates were digested with trypsin overnight at 37°C/800 rpm with a 1:30 ratio of trypsin and total protein. Digested peptides were acidified to a final concentration of 0.1% FA, dried and stored at -80°C until MS analysis.

TMT labelling and HiRIEF IPG-IEF fractionation

MS analysis was performed at the Clinical Proteomics Mass Spectrometry facility, Karolinska Institutet/ Karolinska University Hospital/ Science for Life Laboratory. Thirty different samples obtained from iPEPs study were labelled with 3 sets of TMTs 10plex (Thermo Scientific) as per the manufacturer's recommendations. TMT labelling was performed using 30µg peptides in each channel and pooled TMT labelled peptides (300µg) were separated on a 24cm pH3-10 IPG-IEF strip utilizing HiRIEF as previously described [32]. A plastic cast with 72 linear wells was placed over the pI-IPG separating strip, and peptides were transferred from the pI-IPG strip to the solvent front upon incubating in Milli-Q water for 30mins. Peptides were transferred to 72 well plates using a prototype liquid handling robot (GE Healthcare BioSciences AB). Robotic extraction was repeated twice with i) 35 % ACN and ii) 0.1% FA in ACN respectively. Extracted peptides were dissolved in 3% ACN and 0.1% FA for LC-MS/MS application, upon drying in a Speed-Vac.

LC-MS/MS analysis

Peptide separation was performed on the Ultimate 3000 RSLC nano system coupled to a Q-Exactive (QE) MS (Thermo Fischer Scientific, San Jose, CA, USA). Acclaim PepMap nanotrap column (C18, 3µm, 100Å, 75µm x 20mm, Thermo Scientific) was utilized as a pre-column, and peptide separation was performed on a similar column 75µm x 50cm inner diameter containing 2 µm C18 beads. Peptides were eluted from the column on a gradient ranging between 6% and 37% of mobile phase B (90% acetonitrile (ACN), 5% dimethyl sulfoxide (DMSO), 0.1% FA) for 30 to 90 mins at a flow rate of 0.23 µL/min, depending on the IPG-IEF sample complexity. Mobile phase A contained 5% DMSO in 0.1% FA. QE was operated on a data-dependent mode with a survey scan at a resolution of 70,000 between 400-1600m/z. Maximum injection time was set at 100ms for a target of $1x10^6$ ions. Precursor fragmentation of the top 10 precursors was performed by higher-energy collisional dissociation (HCD). HCD fragments were generated with 30% normalized collision energy and AGC of $1x10^5$ at an injection time of 140ms.

Data processing and statistics

RAW MS data files were searched with SEQUEST-HT node [36] on Proteome Discoverer (v 2.3) (Thermo Scientific) against the SwissProt human dataset containing 20,431 sequences. Precursor mass tolerance of 10 ppm and fragment mass tolerance of 0.2 Da was applied for tryptic peptides search. The search was filtered to define protein identification based on minimum two unique, non-missed cleaved tryptic peptides of that contain a minimum of seven amino acid residues. Percolator [37] based target-decoy strategy was utilized to set protein level FDR at 1%. Quantitative analysis was performed

with the proteome discoverer quantification module with quantification ratios normalized against the total peptide amount.

Differentially-expressed proteins between iPEPs treated and non-treated on both SW480$^{\beta6OE}$ and SW480^{Mock} cells were quantified by unpaired t-test using p-value thresholds of <0.05 and fold change limits of >1.5 up- or down-regulation for each protein. The Cytoscape [38] plugin ClueGO [39] and CluePedia [40] were used to perform pathway and cellular localization enrichment. DAVID [41] was utilized for pathway analysis and superimposition of the differentially expressed proteins in the KEGG [42] pathways.

Results

Previous work from our collaborative team has established the role of integrin β6 subunit in CRC metastatic progression [31]. Expression of integrin β6 subunit alone (in SW480$^{\beta6OE}$ cells) increased cell proliferation, migration and invasion by 65%, 159% and 100% respectively when compared against basal wild-type SW480^{mock} cells. Integrin β6 was found to mediate these effects in a ligand-dependent manner. Treatment of the SW480$^{\beta6OE}$ cells with the ligand zymogens (i.e., latent TGFβ (binds to □v□6), and plasminogen; Plg (binds to cell-surface Plg receptors)) further enhanced proliferation, migration and invasion (i.e., latent TGFβ alone ↑ 24%, Plg alone ↑ 12% and latent TGFβ+Plg together ↑ 21%) when compared to the SW480^{mock} cells lacking integrin β6 subunit expression [31]. Based on our studies on the biological relevance of uPAR and αvβ6 in CRC, we have highlighted that these biological alterations are likely mediated by the uPAR·αvβ6 interactome [26,29–31,43].

Furthermore, we demonstrated that the uPAR·αvβ6 interactome in the presence of active TGFβ or active plasmin increased proliferation, migration and invasion of CRC cells

(SW480$^{\beta6OE}$) [31]. Antagonism of TGFβ or plasmin activity using either the antagonist SB431542 or aprotinin, respectively, significantly reduced proliferation in SW480$^{\beta6OE}$ cells. The antagonist SB431542 inhibited pro-invasive properties of plasmin and/or TGFβ [31]. Interestingly, in that study, treatment with either iPEP2 and/or iPEP6 also inhibited proliferation of CRC cell (i.e., the iPEP2 and iPEP6 significantly inhibited ligand-dependent proliferation in SW480$^{\beta6OE}$ cells compared to the untreated or scrambled peptide treatment controls) [31]. Additionally, iPEPs were able to nullify β6 integrin subunit mediated pro-invasive effects. Collectively, antagonization of the uPAR·αvβ6 association by iPEPs has been demonstrated by our team to inhibit proliferation, migration and invasion of CRC cells in a manner that appears to reflect similar changes exerted by TGFβ antagonists [31].

HiRIEF enables high-throughput quantitation of proteomic changes in uPAR·αvβ6 metastasome

With the evidence of reduced patient survival upon co-expression of the uPAR·αvβ6, I sought to antagonize the association with the aid of antagonist peptides. For this purpose, six (6) rationally-designed iPEPs were synthesised based on our previous work [30] and were utilized to assess antagonism of uPAR·αvβ6. Binding of these peptides on cell surfaces and subsequent changes in cell morphology was visualized by fluorescence microscopy. Fluorescence microscopy demonstrated that only iPEP2 and iPEP6 bound to the cell surfaces while the remaining 4 iPEPs and all scrambled iPEP controls did not bind or affect cell morphology [31].

Similarly, iPEP2 and iPEP6-induced changes in the distribution of actin-cytoskeleton produced subsequent morphological changes. Both iPEP2 and iPEP6 were able to

independently induce a more stellar (“star-like”) morphology only on cells expressing both the proteins uPAR and the integrin β6 subunit, but not in cells lacking integrin β6 subunit expression or integrin β6-expressing cells treated with scrambled iPEPs. We have evidence some evidence concerning the role of TGFβ and proliferative MAPK signalling as potentially driving many of these changes [31] however, the complete systems-level molecular mechanism driving these changes has yet to be established.

To determine how iPEP2 and iPEP6 induce overall proteome changes, a quantitative proteome analysis was performed after exposing CRC cells to iPEP2 and/or iPEP6 or their respective scrambled peptides. IPG-IEF-based HiRIEF proteomic methods enable high-throughput quantitative analysis through the significant reduction of peptide complexity [32–34]. In our study, a combination of isobaric TMT 10-plex peptide labelling and HiRIEF-LC-MSMS allowed us to identify a total of 7,618 proteins across all iPEP-treated and non-treated SW480 cells. Of the total HiRIEF-based protein identifications 3,752 (49%) proteins were membrane related proteins **(supplementary table 4.1)**, which confirms that HiRIEF is capable of efficient peptide fractionation of the incredibly complex plasma and other membrane proteomes. This ability to identify membrane proteins without enrichment allowed us to perform an in detail analysis of the interactome linked to the membrane protein-protein association between uPAR and integrin αvβ6.

Integrin αvβ6 mediates metastatic behaviours in this CRC cell model

In an effort to identify the proteome altered by the integrin αvβ6 expression, we performed pathway enrichment using the Cytoscape [38] plugins ClueGO [39] and CluePedia [40]. Three hundred and seventy-nine significantly altered proteins were identified between $SW480^{\beta6OE}$ and $SW480^{mock}$ cells lines, based on a fold change of ±1.5 times when

significance p<0.05 was considered **(supplementary table 4.1)**. This analysis was found to possess pathways enriched for behaviours including cell migration, regulation of programmed cell death, embryogenesis, regulation of cell communication and other signalling pathways, as shown in **figure 1** by colour coded network node annotations. Each process is represented by a different node colour in the network.

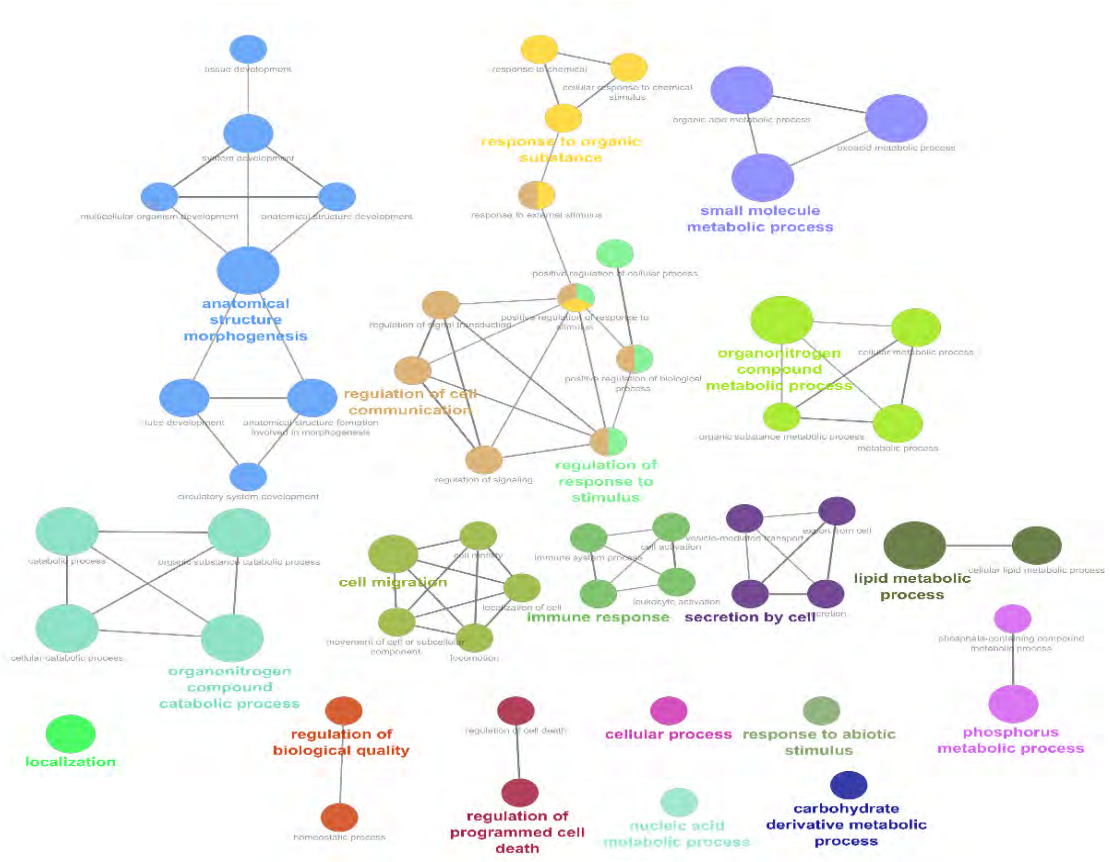


Figure 1: **Functional analysis of the proteome levels associated with integrin β6 expression:** Differentially expressed proteins between SW480$^{\beta6OE}$ and SW480^{mock} cells were enriched with Cytoscape plugin ClueGO. Processes associated with embryogenesis, regulation of cell communication, regulation of programmed cell deaths,

cell migration, immune response, response to stimulus and metabolic processes were found to be altered in SW480$^{\beta6OE}$ cells compared to SW480^{mock} cells, as represented by the colour coded node annotations.

Processes associated with morphogenesis, cell communication and cell migration were identified to be the most significantly altered pathways between SW480$^{\beta6OE}$ and SW480^{mock} cells (**supplementary table 4.2**) as shown in **figure 2**. The enrichment analysis pairs with previous observations concerning the role of integrin αvβ6 in driving multiple metastatic phenotypes such as proliferation, migration and invasion in functional cell assays. This recapitulated previous observations [31] and extended them using more comprehensive proteome analysis.

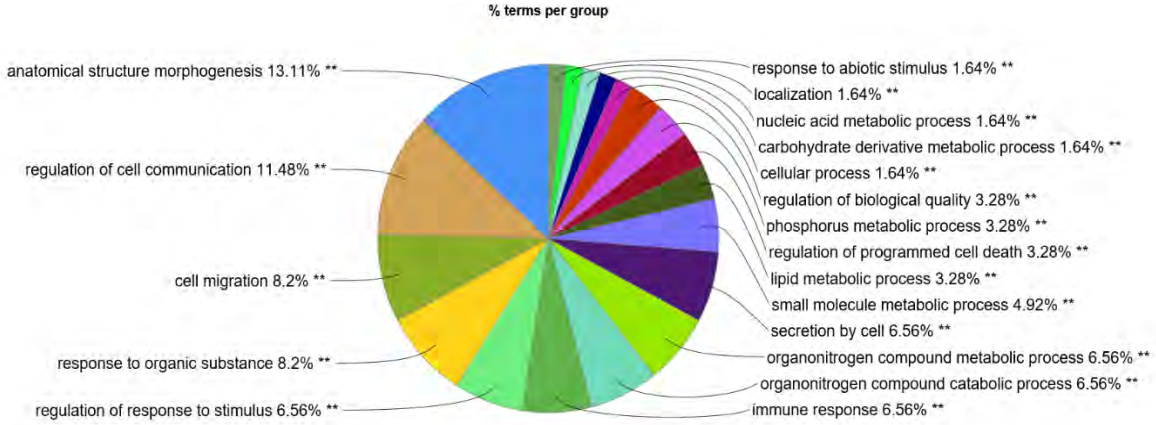


Figure 2: Overview of the pathways altered by the integrin αvβ6 expression: Integrin αvβ6 primarily alters the processes associated with the cell morphological studies, cell communication and cell migration. The observation aligns with previous observations by Cantor et al., [31] using cell-based assays, which confirmed the involvement of integrin β6 with an increase in basal level proliferation, migration and invasion.

Among 379 differentially-expressed proteins, 27 proteins were found to exhibit a minimum 3 fold change in SW480$^{\beta6OE}$ cells, with 23 upregulated and 4 downregulated (**supplementary table 4.1**). The highest fold changes associated with the SW480$^{\beta6OE}$ cell proteome was observed in the integral membrane protein GPR180 (↑ 8.47 fold change, GPR180, UniProt: Q86V85) and Granzyme B (↓ 8.33 fold change, GRAB, UniProt: P10144), when compared against the proteome of SW480^{mock} cells.

GPR180 gene is primarily expressed in the vascular smooth muscle cell biology and is thought to be involved in vascular remodelling [44]. Recently the GPR180 gene has been identified to be critically involved in breast cancer [45]. Similarly, GRAB is serum protease that induces apoptosis via activation of caspase cascades during cell-mediated immune response [46]. GRAB-mediated apoptosis is utilized by the natural killer cells and cytotoxic T lymphocytes to eliminate tumorous, virally-infected and/or allogeneic cells [47] and loss of individual granzymes is associated with increased cancer risk [48].

KEGG [42] pathway enrichment of differentially expressed proteins in/on SW480$^{\beta6OE}$ against SW480^{mock} cells identified retinol metabolism (hsa00830), metabolic pathways (hsa01100), tyrosine metabolism (hsa00350) and fatty acid degradation (hsa00071) as the top four (4) pathways involved in GRAB signalling, compared to SW480^{mock} cells.

Antagonization of the uPAR-αvβ6 mediated by iPEPs 2/6 alters processes associated with metastatic phenotypes in CRC cell model

Proteome differences between individual iPEP2- and iPEP6-treated SW480$^{\beta6OE}$ cells was quantified against the proteome of SW480^{mock} cells. iPEP2-treated proteomes were found to contain 273 differentially-expressed proteins at a minimum 1.5 fold change and p-

values ≤0.05 **(supplementary table 4.1)**. These proteins were found to be associated with epithelial cell migration, axon development, artery morphogenesis, cell-substrate adhesion, wound healing responses and negative regulation of proteolysis, as shown in colour node annotation in **figure 3.**

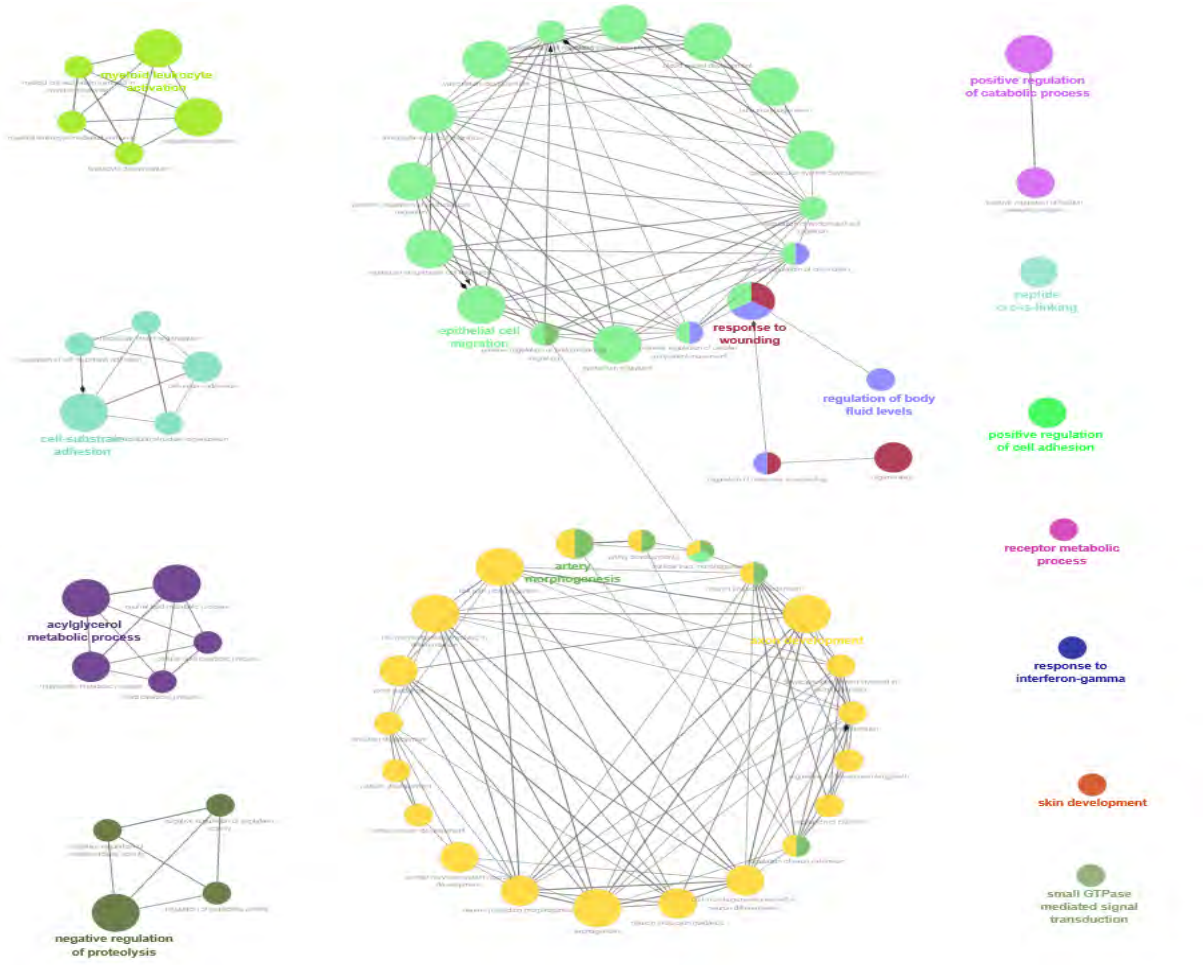


Figure 3: Processes associated with the differential proteome expression in SW480$^{\beta 6OE}$ cells upon iPEP2 treatment. Differentially expressed proteins between iPEP2 treated SW480$^{\beta 6OE}$ and SW480^{mock} cells were enriched to identify the processes altered with the proteome changes. Analysis with 273 differentially expressed proteins (min. 1.5 fold change and $p < 0.05$) enriched the processes associated with "*cell migration, axon development, artery morphogenesis, cell-substrate adhesion, response to wound healing and negative regulation of proteolysis activity*". Each colour in the network represents a separate process, while coloured node annotation represents the primary process.

DAVID-based pathway enrichment analysis[41] of differentially-expressed proteins identified proteoglycans in cancer (KEGG: hsa05205), focal adhesion contacts (KEGG: map04510), ECM-receptor interactions (KEGG: hsa04512), complement and coagulation cascades (KEGG: map04610) and HIF-1 signalling pathway (KEGG: map04066) as the top 5 pathways affected **(supplementary table 4.3)**. It comes as no surprise that most of these pathways are independently and intimately involved in metastatic cancer biologies.

For example, proteoglycans are key tumor microenvironment cell surface and pericellular molecular effectors involved in processes long-associated with tumor progression [49,50]. Similarly, the formation and turnover of focal adhesions play an essential role in cell migration and invasion during cancer where they form specialized contact points between the cell and the extracellular matrix, where actin filaments bundles are anchored to the transmembrane integrin receptors. Focal adhesion signalling culminates in actin cytoskeletal reorganization which facilitates cell motility [51,52]. Proteoglycan and integrin mediate cell and ECM interactions was another enriched pathway that mediates the direct or indirect regulation of the adhesion, migration, invasion and apoptosis processes [19]. Likewise, the complement system contains a number of central serine proteolytic cascades which mediate innate immunity [53]. Recent discoveries in the function of the complement system have identified roles in cell-cell and cell-extracellular communication, cell migration and proliferation in cancer [54–56]. Equally, hypoxia-inducible factor 1 (HIF-1) modulates cell signalling by activating transcription of genes involved in glucose metabolism that allows metastatic cells to fulfil their increased glucose uptake requirements [57].

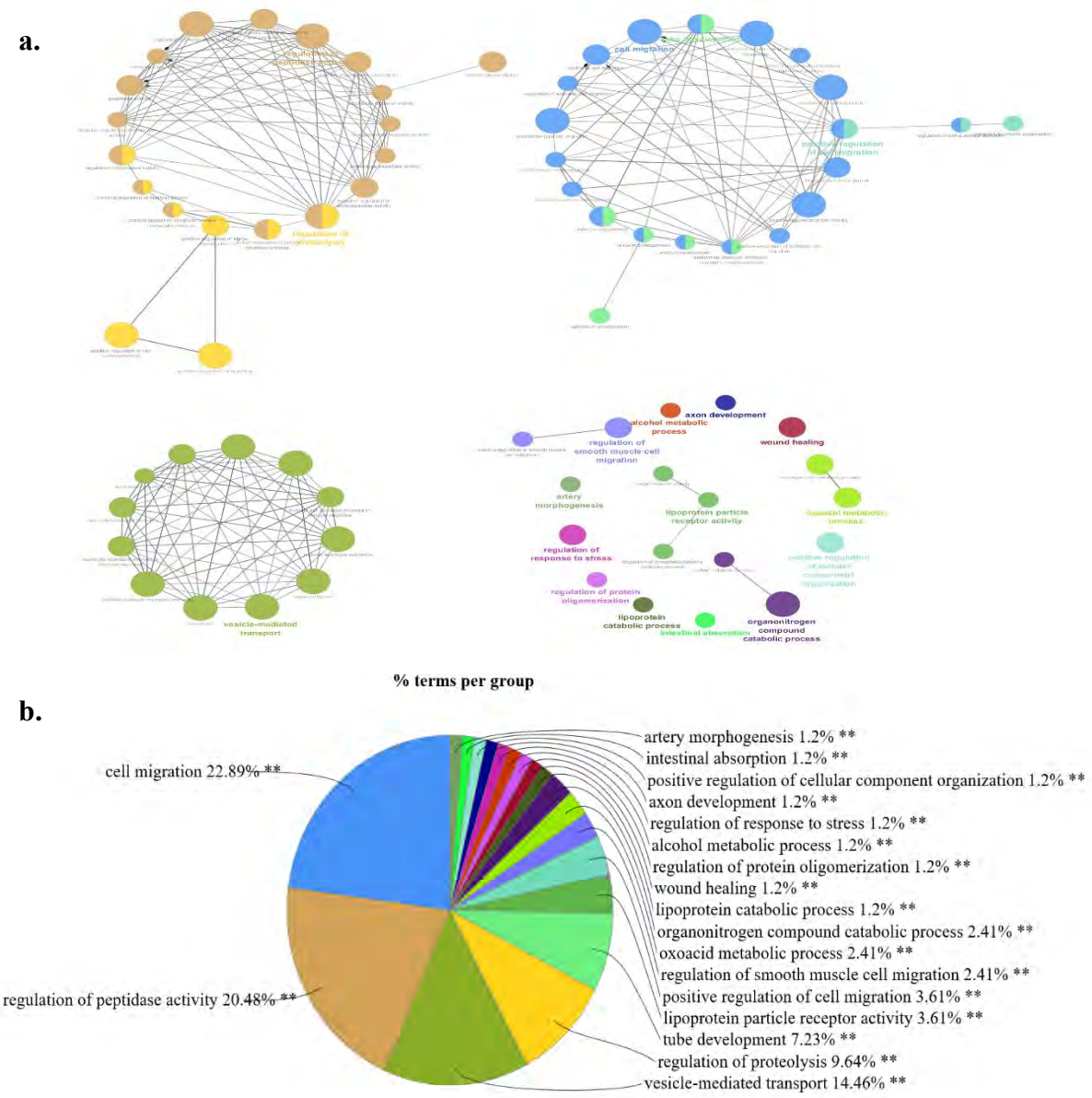


Figure 4: Processes altered by the treatment of iPEP6 on SW480$^{\beta 6OE}$ cells. Differentially expressed proteins upon iPEP6 treatment on SW480$^{\beta 6OE}$ cells against SW480^{mock} cells were enriched to identify the processes associated with the altered proteome. Enrichment of 464 differentially expressed proteins (min. 1.5 fold change and $p < 0.05$) generated an enrichment network specific to the processes associated with "cell migration, regulation of peptidase activity, versicle mediated transport and regulation of

proteolysis", represented as a node annotation in colour (4a) and the pie chart (4b). Enrichment network constitutes subprocess for all applicable processes.

Quantitative proteome analysis identified 464 significant proteins (minimum ± 1.5-fold change and p-value<0.05) out of 7,618 proteins, from iPEP6-treated SWS480$^{\beta 6OE}$ cells against non-treated SW480^{mock} cells **(supplementary table 4.1)**. iPEP6 treatment altered processes associated with cell migration, regulation of proteolysis/peptidase activity, vesicle-mediated transport, as shown in **figure 4**. Twenty-four of these proteins were characterized by a minimum 3 fold change in protein levels against non-treated SW480^{mock} controls. Similar to the iPEP2 study, GPR180 was identified as the top upregulated (↑ 9.6) protein, whereas GRAB (↓ 9.6) was the most downregulated protein upon iPEP6 treatment.

Importantly, DAVID-based pathway analyses identified that iPEP2 and iPEP6 treatment affected a common set of biological processes. In detail, the top five (5) processes altered by iPEP6 treatment were ECM-receptor interaction (KEGG: hsa04512), proteoglycans in cancer (KEGG: hsa05205), complement and coagulation cascades (KEGG: hsa04610), glutathione metabolism (KEGG: hsa00480) and focal adhesion (KEGG: hsa04510) (**supplementary table 4.4**). Glutathione (GSH) metabolism was identified to be a significantly altered process only in the iPEP6-treated CRC cell line proteome. GSH metabolism is primarily involved in maintaining cellular redox homeostasis and nutrient metabolism. However, roles in cellular processes like differentiation, proliferation and migration have been established and disturbances in reduced GSH homeostasis has been linked to cancer progression [58,59].

Discussion

Expression of the integrin β6 subunit repurposes signalling events based on cellular assays leading to enhanced proliferative signalling [31]. Our previous study has established that cells expressing full-length integrin β6 subunit display increased endogenous ERK2 phosphorylation. This is not surprising given it is one of integrin αvβ6 key signaling ligands and equally an end-product of MAPK proliferative signalling [24]. Significant increases in ERK 1/2 phosphorylation were observed upon expression of full-length integrin β6, and phosphorylation levels were further enhanced after treatment with latent TGFβ1 and or Plg [24]. This evidence suggests the presence of non-canonical signalling involved in the translation of the TGFβ and plasmin activity leading to the enhanced phosphorylation of ERK 1/2.

Likewise, integrin β6 subunit expression also enhanced basal SMAD2 and Akt1/2/3 phosphorylation levels, however latent TGFβ and Plg were able to inhibit the phosphorylation, indicating a switch in MAPK-dependent proliferating signalling to SMAD/Akt signalling. This data demonstrates the uPAR·αvβ6 metastasome contributes to trans-activation of the MAPK pathway [31].

These proliferative effects are further exacerbated in the presence of the uPAR-αvβ6 interactome, interactome ligand zymogens like latent TGFβ and/or Plg. This enhanced proliferative signalling is inhibited either by the zymogen inhibitors or interference iPEPs. Previous studies by Cantor et al., in our team have conclusively identified signalling changes are mediated via non-canonical MAPK proliferative signalling pathways [31].

HiRIEF aids in the generation of high-stringency MS data

We have performed a comprehensive proteomic analysis to observe effects of rationally-designed interference peptides on biological effects of the uPAR·αvβ6 interactome. A well-established IPG-IEF based HiRIEF analysis was utilized for TMT label-based quantitative proteomics analysis identifying 7,618 proteins and high MS stringency criterion applied for identification of proteins (i.e., minimum of two no miss-cleaved peptides of at least seven amino acid residues length, as peptides of longer length, tend to result in low false discovery rate (FDR)). Peptides identified with the stringency standards are more specific and unique to a protein, elimination of the missed cleavages and requirement of two peptides further increases the FDR sensitivity [60]. Increase in the stringency levels decreases the number of the protein identified from MS analysis however, the number does not translate into the biological relevance of the data. Number of protein identified from a proteomics data is a dynamic value that is highly dependent on the protein search metrics and applied FDR, as exemplified previously in the identification of olfactory receptors [61].

High-Quality proteomics data reveal new insights into critical interactions.

A significant 3,752 (49%) proteins identified with HiRIEF were found to be membrane proteins. A comprehensive peptide fractionation into 72 subsets adopted by HiRIEF decreased proteome complexity by orders of magnitude, allowing identification of many low abundance and hydrophobic proteins. The HiRIEF methodology has previously been utilized in the identification of a significant percentage of the human proteome using proteogenomics [32,33,62,63].

The ability of HiRIEF to achieve in-depth proteome coverage allowed high-throughput identification of proteome changes induced by integrin β6 subunit expression and subsequent treatment with interference iPEPs. A comparison of the proteome levels between $SW480^{\beta6OE}$ and $SW480^{mock}$ cells proteome enriched pathways associated with focal adhesion, ECM receptor, Wnt signalling and proteoglycan cancer biology (figures 1 and 2). It is interesting to speculate that the observed focal adhesion and regulation of the actin cytoskeleton changes might be linked to the observation of a more 'stellar' $SW480^{\beta6OE}$ morphology. The regulation of focal adhesion and ECM-receptor interaction processes is known to mediate integrin-dependent adhesion, activation of growth factors and inhibition of apoptosis which subsequently contribute to cell proliferation, invasion and migration [64].

Similarly, iPEP2 and iPEP6 treatment altered common biological processes including "proteoglycans in cancer (KEGG: hsa05205), focal adhesion (KEGG: map04510), ECM-receptor interaction (KEGG: hsa04512), complement and coagulation cascades (KEGG: map04610)". HIF-1 signalling pathway was identified as one of the top 5 processes only in iPEP2 treatment whilst glutathione metabolism (KEGG: hsa00480) was only identified on iPEP6 treated proteome. Collectively, these processes are known to be involved in cell proliferation, adhesion, migration, invasion and distal metastasis, which are commonly involved in metastatic carcinoma. These signalling changes align with a previous cell treatment assays performed by Cantor et al. [31], who identified the role of integrin β6 in inducing the metastatic phenotypes which are eventually nullified in the presence of iPEPs.

iPEPs as potential clinical agents against metastatic progression

Whilst the iPEP-induced proteome displays a characteristic signature of the inhibited metastasome that pairs with the previous cell assay [31]. However, iPEPs cell treatment was performed at a final concentration of 10μM in cell culture media, and further studies are required to assess their activity at lower concentrations. Similarly, processes induced by iPEP treatment should be validated with orthogonal methodologies to complement MS-based data. Likewise, the ability of iPEPs to resist proteases upon oral administration or intravenous injection, internalization into the tumor cells and ability to inhibit pro-metastatic effects *in vivo* has to be studied to establish iPEPs as a potential agent targeting late-stage metastasis.

Conclusions

In summary, the study presents the pro-metastatic role of integrin αvβ6 in CRC cell model, which can eventually be nullified by interference iPEPs. This data substantiates iPEPs as an attunable potential therapeutic agent against late-stage CRC metastasis potentially by antagonizing association between membrane proteins like uPAR and integrin αvβ6.

Author contributions

M.S.B conceived the idea. M.S.B, E.N, S.B.A and J.L designed the experiments. S.A performed the proteomics experiments whilst S.B.A and D.C performed the background functional cell assays. S.A performed proteomics analysis with inputs from S.B.A, A.M and M.S.B. All authors contributed in preparing and revising the manuscript. HiRIEF-LC-

MS analysis was performed by the Clinical Proteomics Mass Spectrometry facility, Karolinska Institutet/ Karolinska University Hospital/ Science for Life Laboratory.

Acknowledgement

The authors acknowledge ECR fellowship 15/ECF/1-38 from Cancer Institute NSW (S.B.A), NHMRC for project grant #1010303, Cancer Council NSW RG19-04, RG10-04 & RG08-16 (M.S.B & E.C.N) and the iMQRES funding from Macquarie University, Skipper Jacobs travel award and Sydney vital research scholar award (S.A).

References

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
2. Seyfried, T. N. & Huysentruyt, L. C. On the origin of cancer metastasis. *Crit. Rev. Oncog.* **18**, 43–73 (2013).
3. de la Chapelle, A. Genetic predisposition to colorectal cancer. *Nat. Rev. Cancer* **4**, 769–780 (2004).
4. Tauriello, D. V. F. *et al.* TGFβ drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature* (2018). doi:10.1038/nature25492
5. Eden, G., Archinti, M., Furlan, F., Murphy, R. & Degryse, B. The urokinase receptor interactome. *Curr. Pharm. Des.* **17**, 1874–89 (2011).
6. Schweiger, M. R., Hussong, M., Röhr, C. & Lehrach, H. Genomics and epigenomics of colorectal cancer. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* (2013). doi:10.1002/wsbm.1206
7. Kugaevskaya, E. V., Gureeva, T. A., Timoshenko, O. S. & Solovyeva, N. I. The Role of the Urokinase-Type Plasminogen Activator System In Tumor Progression.

Biochem. (Moscow), Suppl. Ser. B Biomed. Chem. **13**, 97–112 (2019).

8. Falanga, A. & Marchetti, M. Hemostatic biomarkers in cancer progression. *Thromb. Res.* (2018). doi:10.1016/j.thromres.2018.01.017
9. Lippert, S. *et al.* Copenhagen uPAR prostate cancer (CuPCa) database: Protocol and early results. *Biomark. Med.* (2016). doi:10.2217/bmm.15.114
10. Noh, H., Hong, S. & Huang, S. Role of urokinase receptor in tumor progression and development. *Theranostics* **3**, 487–95 (2013).
11. Huai, Q. *et al.* Structure of human urokinase plasminogen activator in complex with its receptor. *Science* **311**, 656–659 (2006).
12. Smith, H. W. & Marshall, C. J. *Regulation of cell signalling by uPAR*. *Nature Reviews Molecular Cell Biology* **11**, 23–36 (Nature Publishing Group, 2010).
13. Di Mauro, C. *et al.* Urokinase-type plasminogen activator receptor (uPAR) expression enhances invasion and metastasis in RAS mutated tumors. *Sci. Rep.* (2017). doi:10.1038/s41598-017-10062-1
14. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
15. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
16. Ahn, S. B. *et al.* Proteomics Reveals Cell-Surface Urokinase Plasminogen Activator Receptor (uPAR) Expression Impacts Most Hallmarks of Cancer. *Proteomics* (2019). doi:10.1002/PMIC.201900026
17. Humphries, J. D., Byron, A. & Humphries, M. J. Integrin ligands at a glance. *J. Cell Sci.* (2006). doi:10.1242/jcs.03098
18. Koivisto, L., Bi, J., Häkkinen, L. & Larjava, H. Integrin αvβ6: Structure, function

and role in health and disease. *International Journal of Biochemistry and Cell Biology* (2018). doi:10.1016/j.biocel.2018.04.013

19. Niu, J. & Li, Z. *The roles of integrin αvβ6 in cancer*. *Cancer Letters* **403**, 128–137 (2017).
20. Hynes, R. O. Integrins: Bidirectional, allosteric signaling machines. *Cell* **110**, 673–687 (2002).
21. Lian, P.-L. *et al.* Integrin αvβ6 and matrix metalloproteinase 9 correlate with survival in gastric cancer Observational Study. *World J. Gastroenterol.* (2016). doi:10.3748/wjg.v22.i14.3852
22. Ahn, S. B. *et al.* Correlations between integrin αvβ6 expression and clinico-pathological features in stage B and stage C rectal cancer. *PLoS One* (2014). doi:10.1371/journal.pone.0097248
23. Ahmed, N., Riley, C., Rice, G. E., Quinn, M. A. & Baker, M. S. αvβ6 integrin-A marker for the malignant potential of epithelial ovarian cancer. *J. Histochem. Cytochem.* (2002). doi:10.1177/002215540205001010
24. Ahmed, N. *et al.* Direct integrin αvβ6-ERK binding: Implications for tumour growth. *Oncogene* (2002). doi:10.1038/sj/onc/1205286
25. Saldanha, R. G. *et al.* Proteomic identification of lynchpin urokinase plasminogen activator receptor protein interactions associated with epithelial cancer malignancy. *J. Proteome Res.* (2007). doi:10.1021/pr060518n
26. Cantor, D., Slapetova, I., Kan, A., McQuade, L. R. & Baker, M. S. Overexpression of αvβ6 integrin alters the colorectal cancer cell proteome in favor of elevated proliferation and a switching in cellular adhesion that increases invasion. *J.*

Proteome Res. (2013). doi:10.1021/pr301099f

27. Liu, M. *et al.* Crystal structure of the unoccupied murine urokinase-type plasminogen activator receptor (uPAR) reveals a tightly packed DII–DIII unit. *FEBS Letters* (2019). doi:10.1002/1873-3468.13397
28. Xu, X. *et al.* Crystal structure of the urokinase receptor in a ligand-free form. *J. Mol. Biol.* (2012). doi:10.1016/j.jmb.2011.12.058
29. Sowmya, G. *et al.* A site for direct integrin αvβ6·uPAR interaction from structural modelling and docking. *J. Struct. Biol.* (2014). doi:10.1016/j.jsb.2014.01.001
30. Ahn, S. B. *et al.* Characterization of the interaction between heterodimeric αvβ6 integrin and urokinase plasminogen activator receptor (uPAR) using functional proteomics. *J. Proteome Res.* (2014). doi:10.1021/pr500849x
31. Cantor, D. I. The αvβ6 integrin plays an integral role in colorectal cancer metastasis. *PhD thesis, Macquarie Uni* (2016).
32. Branca, R. M. M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* **11**, 59–62 (2014).
33. Panizza, E., Branca, R. M. M., Oliviusson, P., Orre, L. M. & Lehtiö, J. Isoelectric point-based fractionation by HiRIEF coupled to LC-MS allows for in-depth quantitative analysis of the phosphoproteome. *Sci. Rep.* (2017). doi:10.1038/s41598-017-04798-z
34. Orre, L. M. *et al.* SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. *Mol. Cell* (2019). doi:10.1016/j.molcel.2018.11.035
35. Leibovitz, A. *et al.* Classification of Human Colorectal Adenocarcinoma Cell Lines.

Cancer Res. (1976).

36. Tabb, D. L. The SEQUEST Family Tree. *Journal of the American Society for Mass Spectrometry* (2015). doi:10.1007/s13361-015-1201-3
37. Spivak, M., Weston, J., Bottou, L., Käll, L. & Noble, W. S. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. *J. Proteome Res.* (2009). doi:10.1021/pr801109k
38. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
39. Bindea, G. *et al.* ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp101
40. Bindea, G., Galon, J. & Mlecnik, B. CluePedia Cytoscape plugin: Pathway insights using integrated experimental and in silico data. *Bioinformatics* (2013). doi:10.1093/bioinformatics/btt019
41. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* (2009). doi:10.1038/nprot.2008.211
42. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).
43. Saldanha, R. G., Xu, N., Molloy, M. P., Veal, D. A. & Baker, M. S. Differential proteome expression associated with urokinase plasminogen activator receptor (uPAR) suppression in malignant epithelial cancer. *J. Proteome Res.* (2008).

doi:10.1021/pr800357h

44. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* (2014). doi:10.1074/mcp.M113.035600

45. Mosly, D., Turnbull, A., Sims, A., Ward, C. & Langdon, S. Predictive markers of endocrine response in breast cancer. *World J. Exp. Med.* (2018). doi:10.5493/wjem.v8.i1.1

46. Prager, I. *et al.* NK cells switch from granzyme B to death receptor–mediated cytotoxicity during serial killing. *J. Exp. Med.* (2019). doi:10.1084/jem.20181454

47. Rousalova, I. & Krepela, E. Granzyme B-induced apoptosis in cancer cells and its regulation (review). *International Journal of Oncology* (2010). doi:10.3892/ijo-00000788

48. Cullen, S. P., Brunet, M. & Martin, S. J. Granzymes in cancer and immunity. *Cell Death and Differentiation* (2010). doi:10.1038/cdd.2009.206

49. Iozzo, R. V. & Sanderson, R. D. Proteoglycans in cancer biology, tumour microenvironment and angiogenesis. *Journal of Cellular and Molecular Medicine* (2011). doi:10.1111/j.1582-4934.2010.01236.x

50. Nikitovic, D. *et al.* Proteoglycans-Biomarkers and targets in cancer therapy. *Frontiers in Endocrinology* (2018). doi:10.3389/fendo.2018.00069

51. Leube, R. E., Moch, M. & Windoffer, R. Intermediate filaments and the regulation of focal adhesion. *Current Opinion in Cell Biology* (2015). doi:10.1016/j.ceb.2014.09.011

52. Annis, M. G. *et al.* Integrin-uPAR signaling leads to FRA-1 phosphorylation and

enhanced breast cancer invasion. *Breast Cancer Res.* **20**, 9 (2018).

53. Ricklin, D., Reis, E. S. & Lambris, J. D. Complement in disease: a defence system turning offensive. *Nature Reviews Nephrology* (2016). doi:10.1038/nrneph.2016.70

54. Afshar-Kharghan, V. The role of the complement system in cancer. *Journal of Clinical Investigation* (2017). doi:10.1172/JCI90962

55. Carmona-Fontaine, C. *et al.* Complement Fragment C3a Controls Mutual Cell Attraction during Collective Cell Migration. *Dev. Cell* (2011). doi:10.1016/j.devcel.2011.10.012

56. Strey, C. W. *et al.* The proinflammatory mediators C3a and C5a are essential for liver regeneration. *J. Exp. Med.* (2003). doi:10.1084/jem.20030374

57. Semenza, G. L. HIF-1: upstream and downstream of cancer metabolism. *Current Opinion in Genetics and Development* (2010). doi:10.1016/j.gde.2009.10.009

58. Bansal, A. & Celeste Simon, M. Glutathione metabolism in cancer progression and treatment resistance. *Journal of Cell Biology* (2018). doi:10.1083/jcb.201804161

59. Traverso, N. *et al.* Role of glutathione in cancer progression and chemoresistance. *Oxidative Medicine and Cellular Longevity* (2013). doi:10.1155/2013/972913

60. Sinitcyn, P., Rudolph, J. D. & Cox, J. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annu. Rev. Biomed. Data Sci.* **1**, 207–234 (2018).

61. Adhikari, S., Sharma, S., Ahn, S. B. & Baker, M. S. In Silico Peptide Repertoire of

Human Olfactory Receptor Proteome on High-Stringency Mass Spectrometry. *J. Proteome Res.* acs.jproteome.8b00494 (2019). doi:10.1021/acs.jproteome.8b00494

62. Zhu, Y. *et al.* Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat. Commun.* (2018). doi:10.1038/s41467-018-03311-y

63. Pernemalm, M. *et al.* In-depth human plasma proteome analysis captures tissue proteins and transfer of protein variants across the placenta. *Elife* (2019). doi:10.7554/eLife.41608

64. McLean, G. W. *et al.* The role of focal-adhesion kinase in cancer - A new therapeutic opportunity. *Nature Reviews Cancer* (2005). doi:10.1038/nrc1647

Chapter 5: Thesis Discussion and Future Directions

This thesis contributes in the generation of new knowledge regarding a number of discrete initiatives under the aegis of the **H**uman **P**roteome **P**roject (HPP), namely the chromosome-centric HPP (C-HPP) and biology/disease HPP project (B/D-HPP). Contributions to C-HPP knowledge was achieved through;

i) an update on the status of the past and current missing proteins (MPs; PE2+PE3+PE4 proteins) defined by neXtProt [1], and

ii) determining one potential explanation for the paucity of MS-based PE1 identifications for multi-transmembrane domain-containing human proteins (i.e., ≥2 transmembrane domains; TMDs), where a higher proportion (880/2,129 or 41%) are currently PE2-4 MPs.

Analysis of the neXtProt missing proteins over the years allowed the identification of multiple protein families that have failed to obtain sufficient MS evidence for their PE1 assignment. Differentiation of the missing protein families based on their rate of identification will aid in designing targeted MS approach for their potential PE1 assignment. The analysis concluded that ORs were the largest group of missing proteins that lack any MS evidence. In silico analysis of multi-TMD containing proteins indicated that membrane hydrophobicity and distribution of proteolytic sites play a role in restricted identification of these protein groups. This signifies the requirement of either adoption of

non-conventional MS approach or relaxation of the stringency requirements for specific missing proteins as per the HPP “exceptions committee” criteria for PE1 assignment. During the in silico analysis, multiple missing proteins failed to generate theoretical peptides qualifying the HPP PE1 guidelines. These missing proteins are the probable candidates qualifying the PE1 assignment, based on the exceptions committee criteria. Overall, contribution to the C-HPP would aid in the identification of the missing proteins groups that lacked significant PE1 assignment, prediction of the theoretical peptidome of any multi-TMD containing proteins and identification of potential missing proteins candidate that could qualify for PE1 assignment based on upcoming exceptions criteria.

Contributions to the B/D-HPP was substantiated through antagonism of a putative cancer metastasis-related membrane protein-protein interaction between urokinase plasminogen activator receptor (uPAR) and integrin αvβ6. There, rationally designed uPAR interference peptides (iPEPs) based on the primary sequence binding sites of uPAR for integrin αvβ6 were utilized to assess their ability to abrogate the metastatic progression and to determine what proteome changes they affected. The analysis was complemented with identification of the uPAR interactome via a text mining approach to better understand the uPAR signalling networks.

uPAR interactome revealed its pleiotropic nature, interacting with different classes of proteins. Role of uPAR interaction with many of these partners has already been identified in inflammation, wound healing and metastatic cancer progression. Reduction in the levels of the uPAR in CRC cell model was found to impair metastatic signalling signatures. Moreover, the definition of the epitope binding site of ligand on the uPAR resides could

be exploited for antagonization studies aiding in the precision treatment. In this effort, antagonization of uPAR interactome was exemplified by a synthetic peptide derived from the binding site of integrin αvβ6 on the uPAR. Disruption of uPAR and integrin αvβ6 interaction by the synthetic peptide resulted in the inhibition of several metastatic phenotypes in colon cancer cell model. Comprehensive proteomics analysis demonstrated that either reduction of the uPAR expression levels or antagonization of the uPAR interactome with the aid of synthetic peptides negates multiple hallmarks of the metastatic progression. This presents uPAR as an attenable target against CRC and epithelial cancer metastasis and uPAR and integrin αvβ6 interaction as a potential clinical target against late-stage metastasis.

The thesis offers future directions for the field on the improved strategies for membrane protein solubilization, utilization of “confetti” approach for membrane digestion, optimization of LCMS strategies to attain more in-depth coverage, and adoption of communal analysis of publicly available MS data for identification of novel MS peptides. The thesis also emphasizes the need for identification and validation of novel markers. Finally, the thesis presents iPEPs as imaging probe and therapeutic agent, applicable to multiple epithelial carcinomas.

5.1 Challenges in identification of multi transmembrane domain-containing proteins by high stringency mass spectrometry

Identification of multi-TMD-containing membrane proteins with high stringency MS remains problematic, primarily due to technical challenges on solubilization, proteolytic digestion, chromatographic separation and MS identification of at least two 9 or more amino acid containing unitypic non-nested cleaved peptides from the N- and C-termini,

inter-TMD loops and TMDs [2,3]. Equally, incompatibility of with a traditional shotgun MS workflow results in the under-representation of peptides originating from the TMD regions, when compared to the peptides derived from the soluble loop and strand domains [4].

Due to the improbable tryptic peptide generation from TMD regions, we hypothesised that the experimental tryptic peptide yield from TMD-containing proteins should be comparable to the theoretical (*in silico*) peptides originating from concatenated soluble domains (i.e., N- and C- terminal strands and loop domain). To test my hypothesis, I performed an *in silico* analysis to construct a tryptic peptide repertoire derived solely from the TMD and soluble domains of all 404 human seven TMD-containing olfactory receptors (OR). ORs are the ideal protein family on which to perform *in silico* analysis as they entirely lack HPP MS-based evidence to support their protein existence and the *in silico* analysis could potentially decipher the reasons behind this paucity in MS-based identification. Analysis of the tryptic peptidome revealed that only 235/404 (58%) ORs were able to generate tryptic peptides that could meet the current HPP guidelines for protein existence 1 (PE1) assignment (i.e., generate a minimum two or more uniquely-mapping, non-nested, experimentally-derived peptides that are of 9 or more amino acid residues in length).

Manuscript 1 presents an *in silico* analysis on seven TMD-containing ORs to assess their ability to generate peptides qualifying the PE1 assignment criteria from their concatenated soluble domains [5]. The inability of specific ORs to the generate peptides qualifying HPP criteria, from their complete sequence (i.e., including TMDs) presents the definitive rationale behind the paucity in the identification of ORs by an MS approach. The biological facts that ORs have low transcription rates are expressed in a limited number

of cells and that each cell only expresses any single (i.e., 1/404) OR at a given time further complicates the search for ORs by MS.

In **Manuscript 2,** I expanded previous OR family analysis to all human multi-TMD containing proteins. Analysis of the TMD number, length, hydrophobicity and presence of the arginine and lysine (R/K) residues was performed to assess their ability to generate suitably long and unitypic tryptic peptides. The *in silico* tryptic peptide repertoire of multi-TMD containing proteins identified a subset of proteins that are also not able to generate peptides as per HPP MS-based PE1 assignment requirements. Future PE1 assignment of these proteins by MS data may require the adoption of non-conventional MS analysis methods, i.e., utilization of alternate protease/s digestion or inclusion of these proteins under the "exceptions" committee now considering these types of proteins.

5.2 Mass spectrometry-based proteomics analysis

The HPP has recognized MS as one of its fundamental resource technology pillars along with antibody/affinity, knowledgebase and pathology pillars. MS is an indispensable component of current high-throughput proteomic analysis which enables determination of an unbiased map of protein abundances, interactions, sequence modifications and splice variants [6, 7]. MS-based proteomic analysis can broadly be classified into i) discovery approaches that aim at identification and characterization of a complete proteome from a protein sample, and ii) targeted approaches that aim at the identification of specific proteins of interest utilizing predefined peptide model [8].

MS analysis can either be performed with i) top-down or ii) bottom-up approaches. A top-down approach involves an analysis of intact proteins for protein identification allowing discernment of the proteins' primary structure, as the data generation does not involve

proteolytic agents [9]. Bottom-up proteomics (also referred to as “shotgun”) involves proteolytic digestion of the proteins to peptides and their analysis by LC-MS/MS. Protein identification from shotgun data is often performed by matching theoretical spectra with experimental spectra, commonly referred to as peptide-spectrum-matches (PSM) [10].

The PSM method provides a qualitative and quantitative measure for protein abundances across samples either using a label-free or label-based methods [11]. Label-based methods rely on the incorporation of isotopologues into peptide or protein which can be distinguished by MS. Label-free techniques adopt a comparative analysis based on the number or intensity of the acquired ion spectra [12], different label-free quantification strategies along with a workflow for targeted analysis in Skyline [12] is described in **Manuscript 3**. Label-free quantification was applied in **Supplementary Manuscript 1** to decipher the systemic proteome changes associated with a decrease in uPAR levels in wild-type and genetically-modified CRC cell line (HCT116)**.** Quantitative proteomics analysis was performed between cells carrying a uPAR antisense construct and wildtype cells expressing high basal levels of uPAR [13]. Comprehensive proteomics study demonstrated for the first time that reduction in uPAR expression levels negates multiple hallmarks of cancers (HoCs) associated with metastasis upon suppression of uPAR [13].

Similarly, label-based high-throughput proteomics was performed in **Manuscript 5,** which adopted a combination of tandem mass tagging (TMT), immobilized pH gradient isoelectric focusing (IPG-IEF) fractionation based on high-resolution isoelectric focusing (HiRIEF) methods with liquid chromatography and LC-MS. This analysis aimed at the identification of proteome differences associated with exposure of cells to rationally designed interference peptides (iPEPs). CRC cells that either express ($SW480^{\beta 6OE}$) or

lack integrin β6 ($SW480^{mock}$) were examined. Since SW480 cells lack endogenous integrin β6 expression, overexpression of the integrin β6 in $SW480^{\beta6OE}$ cells also allowed us to study the uPAR and integrin αvβ6 mediated signalling in CRC metastasis.

5.3 Identification of membrane and low abundant proteins

The mass-to-charge (m/z) ratio of parent and daughter ions acts as a surrogate for protein identification during bottom-up approaches. Intensity and mass of coeluting peptide species are recorded in the MS mode whereas, the fragmentation pattern is identified during the MS/MS mode [14]. Mass and fragmentation pattern provides the peptide identity, whereas intensity on parent ion provides a measure of abundance [14].

MS has a remarkable sensitivity and resolution in the identification of high abundant ion species. However, the identification of low abundant proteins is challenging due to the masking effects imparted on low abundant proteins by the high abundant proteins [15]. The phenomenon is more pronounced in data-dependent approaches (DDA) approach, where only selected number of most intense parent peptide species are selected for MS/MS fragmentation to generate daughter ions. The process results in bias favouring high abundant peptide species [15]. The identification of low abundant peptide species could be enhanced by either i) enrichment strategies to remove specific high abundant proteins or by ii) fractionation strategies which reduce sample complexity by dividing the proteome over multiple subsets.

5.4 Colorectal cancer (CRC)

Globally 1.8 million patients were diagnosed with CRC during 2018; 75 % of these patients are predicted to succumb to the disease by the end of 5 years [16]. CRC is more

often diagnosed during the later stages of the disease when cancer has likely spread to distal organs [17].

CRC, like most other cancers, is characterized by multiple gene mutations that allow cancer cells to acquire a metastatic phenotype. A series of gene mutations responsible for the transformation of the benign polyps to adenocarcinoma and metastatic cancer has been characterized. Multiple signalling and proteome level changes associated with these mutations are thought to define the CRC metastatic phenotypes. Of the many proteins families involved in the CRC progression and metastasis, integrins (e.g., αvβ1, αvβ3, αvβ4 and αvβ6), proteolytic enzymes (uPA, uPAR, PAI-1, MMPs), growth factors (TGFβ, EGFR, VEGF), components of the MAPK signalling (ERK 1/2, RAS, JAN), components of the Notch and Wnt signalling are thought or known to be the primary mediators of CRC metastasome [18, 19].

5.4.1 Urokinase plasminogen activator receptor in CRC

Of the multiple pathways involved in CRC metastasome, plasminogen activation system (PAS) mediated proteolysis is thought to be a significant factor driving the CRC metastasome. PAS is driven by urokinase plasminogen receptor (uPA) and uPAR interactome [20]. Elevated uPAR expression level is evident during tissue remodelling, inflammation and epithelial carcinoma. It is thought that uPAR promotes tumorigenesis either by participating in i) PAS dependent proteolytic cascade or ii) proteolysis independent cascade upon interaction with its lateral partners [13,21,22]. The uPAR interactome consists of multiple lateral partners that are involved in metastatic carcinoma and was discussed in **manuscript 4**. A text-mining approach was adopted to identify

uPAR partners, their subcellular locations and the methods utilized to decipher interactions. Of the total 111 uPAR interactions extracted from the PSICQIC [23] platform, binding site data for 12 uPAR ligands were identified and mapped on the uPAR primary sequence. Binding site analysis showed that uPA was likely and not unexpectedly the most significant uPAR partner followed by the vitronectin.

A further observation was that integrins were a significant component of the uPAR interactome as they provide uPAR with signal specificity. Similar to uPAR, integrin expression is elevated during the inflammation, tissue modelling and multiple epithelial carcinomas [24]. Integrins facilitate linkage of the cytoskeleton and ECM upon the formation of focal contacts, which are involved in cytoskeletal reorganization and cell motility [25].

5.4.2 Functional association between integrin αvβ6 and uPAR in CRC

Modulating the expression levels of either uPAR or the integrin αvβ6 has a profound effect on cell behaviour. For example, elevated levels of the integrin β6 in CRC cell model carrying β6 construct (SW480 cells carrying β6 construct, referred as $SW480^{OE}$) shows enhanced proliferation, invasion and migration, compared to a cell lacking the β6 expression [26]. Similarly, decreasing the expression levels of the uPAR in CRC cell model (HCT116 cells carrying uPAR antisense construct) results in the inhibition of the uPAR interactome components involved in CRC metastasis [13].

The Baker research team at Macquarie University has made a significant contribution to defining the independent and related roles of uPAR and integrin αvβ6 in human cancer. Our previous study has confirmed the strong expression of integrin αvβ6 during late CRC stages based on a tissue immunohistochemistry [27,28]. Co-purification of uPAR from an

ovarian cell line (OVCA429 cell) by reverse immunoprecipitation with the integrin β6 antigen provided evidence for uPAR and integrin αvβ6 functioning together in the CRC metastasome [29]. Our continued interest in investigating a direct interaction between uPAR and integrin αvβ6 lead us to examine the precise epitope binding site of integrin αvβ6 with uPAR. *In silico* structural modelling based on uPAR crystal structure and models of integrin αvβ6 enabled us to discern a putative binding site location between these two proteins [30]. These binding site identifications were experimentally confirmed with the aid of peptide array and proximity ligation assays identifying six putative binding sites between uPAR and αvβ6 [31].

5.4.3 Antagonization of the uPAR and integrin αvβ6 interaction.

Synthetic peptides generated from the putative binding site of integrin αvβ6 on uPAR were utilized as interference peptides (iPEPs) to antagonize the uPAR and integrin αvβ6 interaction. Two iPEPs (iPEP2 and iPEP6) were able to induce morphological changes in $SW480^{OE}$ cells. Similarly, iPEP2 and iPEP6 treatment decreased cell proliferation, migration and invasion in $SW480^{OE}$ cells, when compared to $SW480^{mock}$ cells. These SW480 cells lack endogenous β6 expression, as previously shown by Cantor *et al.* [26].

In the study associated with **manuscript 5**, high-resolution MS analysis in combination with TMT and HiRIEF was performed to identify if these iPEPs induced proteome changes in $SW480^{\beta6OE}$ cells and control $SW480^{mock}$ cells. Comparative quantitative proteomics analysis was performed between the proteome of non-treated $SW480^{mock}$ and $SW480^{\beta6OE}$ cells. iPEPs were found to mediate signalling associated with embryogenesis, cell differentiation, apoptotic signalling and wound healing. These results are congruent with

previous biochemical and signalling analysis performed by our team, which demonstrated that iPEPs induced morphological changes and inhibited proliferation, invasion and migration on these same cell lines.

5.5 Future directions

5.5.1 Characterization of the complete human proteome

The HPP aims to "*generate a map of the protein-based molecular architecture of the human body and become a resource to help elucidate biological and molecular function and advance diagnosis and treatment of diseases*" (https://hupo.org/human-proteome-project). The HPP has achieved considerable progress in the identification and characterization of multiple proteins over the years. However, paucity in the identification of membrane proteins has hindered the HPP goal of characterizing the complete human proteome by MS. We have identified several multi-TMD containing proteins that cannot generate peptides required for PE1 assignment. These proteins are unlikely to be identified by conventional MS approaches and would require specialized and targeted approaches for PE1 assignment. Strategies that could aid in the identification of these missing membrane proteins are summarized below;

i) strategies for optimization of membrane proteins solubilization

ii) utilization of proteases compatible with the membrane preparations, such as CNBr or the so-called "confetti" multi-protease digestion [32]

iii) membrane enrichment prior to LCMS [13]

iv) optimization of the LCMS strategy to suit identification of the MS methods; such as multidimensional protein or peptide fractionation to achieve more in-depth

coverage [33], utilization of heated chromatography [34], adoption of targeted approaches

v) informatics approach: communal reanalysis of the publicly available MS data to identify novel peptides from MPs or through the combination of “stranded peptides” from multiple MS repositories [35]

vi) identification of the proteins that are not theoretically capable of generating peptides as per the HPP guidelines [5] and assigning them exceptions criteria or define the existence based on non-MS evidence

In summary, many missing proteins cannot generate peptides qualifying the HPP MS-based PE1 criteria, any attempt to identify these proteins by traditional MS strategy may not be purposeful. However, assignment of alternate stringency criteria like “exceptions lists” as defined by HPP or their identification by non-MS methods such as genomics or interactomics may result in subsequent PE1 assignment of these “distinctive” proteins. Similarly, MS-based identification of proteins that can theoretically generate the peptides qualifying PE1 status can be enhanced with membrane enrichment, optimization of MS and other protein inference strategies.

5.5.2 Validation of the CRC metastatic markers

Whilst the results from iPEPs quantitative proteome changes correlate with the literature and previous cell-based assays from our team, further confirmation with orthogonal technologies is needed to validate these findings. These assays will aid in the elucidation of the role of uPA, uPAR, integrin αvβ6 independently and in when they act in concert.

Rational avenues that could be utilized in the validation of iPEP-induced biology has been exemplified below

5.5.2.1 Dose-dependent zymogen treatments

We performed the iPEPs treatment of the CRC cell model at a concentration of 10μM iPEPs and scrambled iPEPs supplemented in the media. Although the treatment resulted in the identification of plausible biomarkers, achieving the concentration *in vivo* is not cost-effective. iPEPs bound to the $SW480^{OE}$ when they were treated with 10μM iPEP in the growth media; the concentration was adopted to assess the quantitative proteome differences. Proteome changes associated with the lower concentrations of the iPEPs needs to be studied. Moreover, additional studies are required to recapitulate the biological system *in vivo* with the inclusion of agonists and antagonists of the uPAR interactome in iPEPs treatment.

5.5.2.2 Imaging the course of tumor action

Expression of the uPAR interactome in late-stage CRC can be utilized as an imaging target apart from its role as a putative therapeutic target. Anti-uPAR peptides conjugated with the radiolabelled probe allows imaging the severity of the metastatic progression. Multiple radio probes exist that are targeted against uPAR and have shown promising results in mouse models [36].

5.5.2.3 Establishment of a mouse model

Validation of the iPEPs biology will require translation of the treatment to an animal model. Mouse model carrying the CRC phenotypes could be utilized to assess the suitability of

the iPEPs to function *in vivo*. Examination on the viability of the iPEPs upon intravenous injection should be performed to confirm their stability *in vivo*. Analysis of the immune system activation should be monitored for any immunogenic response on the host system to control co-morbidity.

5.6 Summary

This thesis has demonstrated

i. **The importance of the membrane topology in high-stringency MS identification of the multi-TMD containing proteins.**

Multi-TMD containing proteins are composed of TMDs, extracellular and intracellular domains. Among these domains, TMDs are characterized by the sparse distribution of tryptic sites located within the membrane hydrophobic environment, which restricts trypsin activity and subsequent peptide generation for MS analysis.

Non-TMD domains generate most of the peptide complement during experimental MS analysis. However, they require a minimum of two proteolytic sites to release peptides upon proteolytic digestion, with an exception for N- and C-terminal strands. Lack of two proteolytic sites in the domain loop results in the formation of “stranded peptides” that remain anchored to the membrane and is not available for MS analysis.

We have identified multiple membrane proteins that cannot generate peptides qualifying the HPP MS-based PE1 assignment, these proteins may never be identified at standard HPP MS stringency requirements. PE1 assignment of these proteins requires derivation of protein existence information from non-MS techniques such as genomics and protein-protein interaction.

ii. Role of uPAR in driving multiple hallmarks of cancer

Comprehensive membrane proteome analysis of a CRC cell model with reduced uPAR expression demonstrated that uPAR is involved in multiple processes associated with the hallmark of cancer, particularly those involved in the metastatic progression, resisting apoptosis and enhanced proliferative signalling. Components of these signalling networks can be utilized in defining novel targets against metastatic progression. uPAR interactome remains as a promising target against diagnostic and prognostic markers in multiple epithelial carcinomas.

iii. The capacity of a rationally designed interference peptide in abrogating metastatic phenotypes

Antagonization of a putative membrane protein-protein interaction between uPAR and integrin αvβ6 by iPEPs alter the proteome towards inhibition of metastatic phenotypes. IPEPs altered proteome is involved in processes associated with cell embryogenesis, differentiation, wound healing and inhibition of apoptotic signalling. Similarly, results from our previous studies showed the capability of iPEPs to induce changes in cell morphology and inhibit proliferation, invasion and migration. These results collectively present iPEPs as an attunable therapeutic target against CRC and multiple other epithelial cancers.

In summary, this thesis has demonstrated an informatics approach to predict the ability of any multi-TMD containing proteins to generate peptides as per the current MS-based PE1 assignment criteria. Similarly, the thesis presents the utility of high-throughput proteomics analysis in identification and antagonization of uPAR interactome to define novel late-stage CRC biomarkers.

Findings of this thesis will aid in i) formulating strategies for identification of current missing proteins and ii) providing a framework for future initiatives on detection and treatment of CRC.

5.7 References

1. Gaudet, P. *et al.* The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* **45**, D177–D182 (2017).
2. Helbig, A. O., Heck, A. J. R. & Slijper, M. Exploring the membrane proteome-Challenges and analytical strategies. *Journal of Proteomics* (2010). doi:10.1016/j.jprot.2010.01.005
3. Trötschel, C. & Poetsch, A. Current approaches and challenges in targeted absolute quantification of membrane proteins. *Proteomics* .(2015). doi:10.1002/pmic.201400427
4. Josic, D. Strategies for Complete Proteomic Analysis of Hydrophobic Proteins in Complex Biological Samples–Hyde-and Seek. *J. Data Mining Genomics Proteomics* (2014). doi:10.4172/2153-0602.1000e111
5. Adhikari, S., Sharma, S., Ahn, S. B. & Baker, M. S. In Silico Peptide Repertoire of Human Olfactory Receptor Proteome on High-Stringency Mass Spectrometry. *J. Proteome Res.* acs.jproteome.8b00494 (2019). doi:10.1021/acs.jproteome.8b00494
6. Ong, S.-E., Foster, L. J. & Mann, M. Mass spectrometric-based approaches in quantitative proteomics. *Methods* **29**, 124–130 (2003).

7. Nilsson, T. *et al.* Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* **7**, 681–685 (2010).

8. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* (2006). doi:10.1126/science.1124619

9. Savaryn, J. P., Catherman, A. D., Thomas, P. M., Abecassis, M. M. & Kelleher, N. L. The emergence of top-down proteomics in clinical research. *Genome Med.* **5**, 53 (2013).

10. Gillet, L. C., Leitner, A. & Aebersold, R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu. Rev. Anal. Chem.* (2016). doi:10.1146/annurev-anchem-071015-041535

11. Craig, R., Cortens, J. C., Fenyo, D. & Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* (2006). doi:10.1021/pr0602085

12. Noor, Z., Adhikari, S., Ranganathan, S. & Mohamedali, A. Quantification of Proteins From Proteomic Analysis. in *Encyclopedia of Bioinformatics and Computational Biology* 871–890 (Elsevier, 2019). doi:10.1016/B978-0-12-809633-8.20677-8

13. Ahn, S. B. *et al.* Proteomics Reveals Cell-Surface Urokinase Plasminogen Activator Receptor (uPAR) Expression Impacts Most Hallmarks of Cancer. *Proteomics* (2019). doi:10.1002/PMIC.201900026

14. Matthiesen, R. & Bunkenborg, J. Introduction to Mass Spectrometry-Based Proteomics. in *Methods in molecular biology (Clifton, N.J.)* **1007**, 1–45 (Humana Press, Totowa, NJ, 2013).

15. Rauniyar, N. Parallel reaction monitoring: A targeted experiment performed using high resolution and high mass accuracy mass spectrometry. *International Journal of Molecular Sciences* (2015). doi:10.3390/ijms161226120

16. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).

17. Edwards, B. K. *et al.* Annual Report to the Nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. *Cancer* (2014). doi:10.1002/cncr.28509

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

19. Valle, L. *et al.* Update on genetic predisposition to colorectal cancer and polyposis. *Mol. Aspects Med.* (2019). doi:10.1016/J.MAM.2019.03.001

20. Di Mauro, C. *et al.* Urokinase-type plasminogen activator receptor (uPAR) expression enhances invasion and metastasis in RAS mutated tumors. *Sci. Rep.* (2017). doi:10.1038/s41598-017-10062-1

21. Smith, H. W. & Marshall, C. J. *Regulation of cell signalling by uPAR*. *Nature Reviews Molecular Cell Biology* **11**, 23–36 (Nature Publishing Group, 2010).

22. Eden, G., Archinti, M., Furlan, F., Murphy, R. & Degryse, B. The urokinase receptor interactome. *Curr. Pharm. Des.* **17**, 1874–89 (2011).

23. del-Toro, N. *et al.* A new reference implementation of the PSICQUIC web service. *Nucleic Acids Res.* **41**, W601–W606 (2013).

24. Danen, E. H. J. Integrins: An Overview of Structural and Functional Aspects. (2013).

25. Hamidi, H. & Ivaska, J. Every step of the way: integrins in cancer progression and metastasis. *Nat. Rev. Cancer* **18**, 533–548 (2018).

26. Cantor, D., Slapetova, I., Kan, A., McQuade, L. R. & Baker, M. S. Overexpression of αvβ6 integrin alters the colorectal cancer cell proteome in favor of elevated proliferation and a switching in cellular adhesion that increases invasion. *J. Proteome Res.* (2013). doi:10.1021/pr301099f

27. Ahn, S. B. *et al.* Correlations between integrin αvβ6 expression and clinico-pathological features in stage B and stage C rectal cancer. *PLoS One* (2014). doi:10.1371/journal.pone.0097248

28. Ahmed, N. *et al.* Overexpression of αvβ6 integrin in serous epithelial ovarian cancer regulates extracellular matrix degradation via the plasminogen activation cascade. *Carcinogenesis* (2002). doi:10.1093/carcin/23.2.237

29. Saldanha, R. G. *et al.* Proteomic identification of lynchpin urokinase plasminogen activator receptor protein interactions associated with epithelial cancer malignancy. *J. Proteome Res.* (2007). doi:10.1021/pr060518n

30. Sowmya, G. *et al.* A site for direct integrin αvβ6·uPAR interaction from structural modelling and docking. *J. Struct. Biol.* (2014). doi:10.1016/j.jsb.2014.01.001

31. Ahn, S. B. *et al.* Characterization of the interaction between heterodimeric αvβ6 integrin and urokinase plasminogen activator receptor (uPAR) using functional proteomics. *J. Proteome Res.* (2014). doi:10.1021/pr500849x

32. Vit, O. & Petrak, J. Integral membrane proteins in proteomics. How to break open

the black box? *Journal of Proteomics* (2017). doi:10.1016/j.jprot.2016.08.006

33. Orre, L. M. *et al.* SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. *Mol. Cell* (2019). doi:10.1016/j.molcel.2018.11.035

34. Blackler, A. R., Speers, A. E. & Wu, C. C. Chromatographic benefits of elevated temperature for the proteomic analysis of membrane proteins. *Proteomics* **8**, 3956–64 (2008).

35. Baker, M. S. *et al.* Accelerating the search for the missing proteins in the human proteome. *Nat. Commun.* **8**, 14271 (2017).

36. Persson, M. *et al.* First-in-human uPAR PET: Imaging of Cancer Aggressiveness. *Theranostics* **5**, 1303–16 (2015).