McGill

Faculty of Medicine and Health Sciences

Going beyond genetic screening – multi-omics clinical research for the discovery of therapeutic vulnerabilities in hard-to-treat malignancies.

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

Submitted on May 2023 by:

Georgia Mitsa

Division of Experimental Medicine

McGill University, Montreal

Even if I don't finish, we need others to continue.

~Terry Fox

I would like to thank my supervisors, Dr. Gerald Batist and Dr. Christoph H. Borchers, as well as the thesis committee members, Dr. Sonia V. del Rincon and Dr. Marc R. Fabian, and my academic advisor Dr. Andrew Mouland for their continued support and expert advice. Thank you to Dr. René P. Zahedi, my scientific advisor for parts of this program, all peers, friends and collaborators for being part of this journey.

Τέλος, το μεγαλύτερο ευχαριστώ ανήκει στα μέλη της οικογένειάς μου. Σας ευχαριστώ για την υποστήριξή σας υπομένοντας τις ατελείωτες ώρες δουλειάς, την κούραση και τα νεύρα μου. Μαμά και μπαμπά, είστε πρότυπα για μένα- σας ευχαριστώ για την κούρασή σας και τις δυσκολίες που περάσατε για να δημιουργήσετε μια καλύτερη ζωή για όλους μας.

Contents

Abstract

Cancer remains a major challenge for modern medicine. It has become evident that cancer is a complex disease involving many coexisting genomic alterations; the molecular consequences of detected genetic and transcriptomic alterations are yet poorly understood. Risk and patient stratification based on the genetic phenotype alone has proven to have limited efficacy on the individual treatment outcome. Multi-omics approaches combine various high-throughput analytical techniques to characterize the genome, transcriptome, proteome and metabolome, as well as other cellular components of individual tumors simultaneously. Multi-omics has emerged as a promising tool for providing a more comprehensive and holistic view of cancer biology, which may lead to the identification of novel therapeutic targets and biomarkers. The aim of this doctoral thesis is to investigate the potential of multi-omics approaches for the identification of therapeutic vulnerabilities in hard-to-treat malignancies.

This thesis presents a systematic review of the literature on multi-omics approaches in clinical research, with a focus on identifying common challenges and opportunities in the field. We present two original research articles on the application of multi-omics for molecular characterization of malignancies. We established an optimized and automatable sample preparation pipeline for mass spectrometry-based quantitative proteomics from limited volume clinical specimens, such as formalin-fixed, paraffin embedded (FFPE) tissue micro array cores (~0.4 mm³ core volume). This accelerated method enables molecular profiling of individual tumors within 24 h, from sample receival to analysis, and does not require special training or equipment, promoting a translation into the clinic. We used this method for a proteomic 'landscaping' of non-invasive breast ductal carcinoma compared to invasive breast ductal carcinoma (IDC). Ductal carcinoma in situ (DCIS) is the most common type (80%) of non-invasive breast lesions. The lack of validated prognostic markers, limited patient numbers and tissue quality significantly impact diagnosis, risk stratification, as well as patient enrolment and results of clinical studies. Our study validated 22 putative biomarkers from independent genetic studies and reveals more than 380 differentially expressed proteins and metabolic vulnerabilities, that can inform new therapeutic strategies for DCIS and IDC. Due to the readily druggable nature of proteins, this study is of high interest for clinical research and the pharmaceutical industry.

Overall, this thesis demonstrates the potential of multi-omics approaches to uncover new therapeutic vulnerabilities in hard-to-treat malignancies. The findings of this research will contribute to the development of personalized and targeted therapies for cancer patients, which can ultimately improve patient outcomes and reduce the burden of cancer on society.

Résumé

Le cancer reste un défi majeur pour la médecine moderne. Il est devenu évident que le cancer est une maladie complexe impliquant de nombreuses altérations génomiques coexistantes; les conséquences moléculaires des altérations génétiques et transcriptomiques détectées sont cependant encore mal comprises. La stratification des risques et des patients sur la seule base du phénotype génétique s'est avérée peu efficace pour les traitements de patients atteints de cancer. Les approches multi-omiques combinent diverses techniques d'analyse à haut débit pour caractériser simultanément le génome, le transcriptome, le protéome et le métabolome, ainsi que d'autres composants cellulaires des tumeurs. La multi-omique est apparue comme un outil prometteur pour fournir une vision plus complète et plus holistique de la biologie du cancer, ce qui peut conduire à l'identification de nouvelles cibles thérapeutiques et de biomarqueurs. L'objectif de cette thèse de Doctorat est d'étudier le potentiel des approches multi-omiques pour l'identification de cibles thérapeutiques dans les tumeurs malignes difficiles à traiter.

Cette thèse présente une revue systématique de la littérature sur les approches multi-omiques en recherche clinique, en mettant l'accent sur l'identification des défis et des opportunités majeurs dans ce domaine. Nous présentons deux articles de recherche originaux sur l'application de la multi-omique à la caractérisation moléculaire des tumeurs malignes. Nous avons établi un protocole de préparation d'échantillons optimisé et automatisable pour la protéomique quantitative basée sur la spectrométrie de masse à partir d'échantillons cliniques de volume limité, tels que les micro-carottes de tissus fixés en formaldéhyde et enrobés de paraffine (FFPE) (carotte de ~0,4 mm^3). Cette méthode accélérée permet d'établir le profil moléculaire de tumeurs individuelles en 24 heures, de la réception de l'échantillon à l'analyse, et ne nécessite pas de formation ou d'équipement spécial, ce qui favorise son application en clinique. Nous appliquons cette méthode afin de comparer l'expression protéomique du carcinome canalaire mammaire non invasif à celui du carcinome canalaire mammaire invasif (CMI). Le carcinome canalaire in situ (CCIS) est le type le plus courant (80 %) de lésions mammaires non invasives. L'absence de marqueurs pronostiques validés, le nombre limité de patientes et la qualité des tissus ont un impact significatif sur le diagnostic, la stratification des risques, le recrutement des patientes et les résultats des études cliniques. Notre étude valide 22 biomarqueurs putatifs issus d'études génétiques indépendantes et révèle plus de 380 protéines différentiellement exprimées et cibles métaboliques, qui peuvent mener à de nouvelles stratégies thérapeutiques pour les CCIS et les CMI. Comme il est assez aisé de développer des médicaments ciblant spécifiquement une protéine, cette étude est d'un grand intérêt pour la recherche clinique et l'industrie pharmaceutique.

Dans l'ensemble, cette thèse démontre le potentiel des approches multi-omiques pour découvrir de nouvelles cibles thérapeutiques dans les tumeurs malignes difficiles à traiter. Les résultats de cette recherche contribuent au développement de thérapies personnalisées et ciblées pour les patients atteints de cancer, ce qui peut en fin de compte apporter des traitements plus efficaces et mieux tolérés par les patients et alléger le fardeau du cancer pour la société.

Contribution to Original Knowledge and Contribution of Authors

In accordance with McGill Guidelines, the candidate chose to present the results of this thesis in a manuscript-based format. A doctoral thesis submitted to McGill University must include the text of a minimum of two manuscripts published, submitted or to be submitted for publication. Included peer-reviewed manuscripts have not been changed, i.e., they are identical to the published or submitted versions, except for font/size to meet the format of the thesis for consistency and homogeneity. Copyright permissions have been requested where applicable.

The first manuscript (Chapter 2) describes a quantitative proteomics method that was developed and holistically optimized by the candidate to enable clinical proteomics research on formalin-fixed, paraffin embedded (FFPE) specimens. The main objective of this manuscript was to develop a streamlined FFPE-proteomics sample preparation workflow for improved protein extraction from needle-core-sized specimens, i.e., collected during disease diagnosis, as these samples are generally well annotated and collected at different stages of the disease progression. The developed method was optimized to promote clinical translation for a standardized, semi-automated and broad clinical application. The candidate is the first author of this published manuscript (PMID: 35457260); conceptualization, experiments, and data analysis were performed by the candidate. The co-authors provided resources and advice in their respective areas of expertise.

The second manuscript (Chapter 3) describes results from a proteomic characterization of FFPE clinical specimens from pre-invasive ductal carcinoma (DCIS), the most common type (80%) of non-invasive breast lesions. To date, there is no precision oncology treatment available for patients diagnosed with DCIS. Generally, studies on DCIS are challenging due to limited patient numbers and tissue quality. Five key morphological features, inter-observer variability, intra-tumor heterogeneity and a lack of prognostic markers impact clear diagnosis, risk stratification and treatment options. More importantly, these factors also limit patient enrollment and final results of clinical studies on this matter. The candidate employs the developed FFPE-proteomics method presented in Chapter 2, and identifies more than 380 proteomic and metabolic vulnerabilities, that can inform new diagnostic and therapeutic strategies for pre-invasive ductal carcinoma. Due to the readily druggable nature of proteins and metabolites, the findings of this study propose alternative use of FDA-approved drugs, such as antibiotics and nonsteroidal anti-inflammatory drugs (NSAID). It is worth highlighting, that as a result of this study, a highly multiplexed quantitative assay has been developed reflecting signature pathways and proteins (n=90) for cancer progression and metabolic reprogramming that can be applied to a broad range of malignancies. The candidate is the first author of this manuscript that was submitted for publication, and initiated this clinical study; conceptualization, sample retrieval, experiments and data analysis were conducted by the candidate. The co-authors provided resources and advice in their respective areas of expertise.

The candidate created the infrastructure for her larger-scale discovery studies, including the establishment of collaborations and cooperations with national and international research groups and industrial partners. She installed instruments purchased for the purposes of this research study, was responsible for their repair and maintenance, as well as training of new employees and trainees. The candidate developed specialized application methods and trained other students and peers in using these methods for their own studies and fee-for-service projects. These methods are available as Standard Operating Procedures (SOPs) in the Borchers lab. The candidate obtained specialized training for advanced data analysis of -omics data, which was self-initiated and partially self-taught, and obtained certificates for Research Biobanking and GCP for Clinical Trials with Investigational Drugs and Biologics (ICH Focus).

The candidate presented her doctoral research work in Canadian and international conferences and symposia (poster presentation and talks), is the first author/co-author in 5 peer reviewed publications, and 2 submitted manuscripts. Three (3) publications, where two of which the candidate is first author, are in preparation and expected to be submitted in the following months.

Contribution to peer reviewed publications *not included* in this Thesis:

- Sobsey C, Froehlich B, Mitsa G, Ibrahim S, Zahedi RP, d Bruin E, Borchers CH, Batist G. *Proteomic analysis of PIK3CA-mutated tumours identifies protein networks correlated with clinical benefit of capivasertib genetically pre-selected patients.* Manuscript submitted
- Mitsa G, Richard VR, Majedi Y, Lafleur J, Aguilar-Mahecha A, Basik M, Borchers CH. *Evaluation of a "plug and play" nanoflow liquid chromatography system for MS-based proteomic characterization of clinical FFPE specimens.* Expert Review of Proteomics. 2023;20(4-6):87-92. doi: 10.1080/14789450.2023.2219844
- Ibrahim S, Lan C, Chabot C, Mitsa G, Buchanan M, Aguilar-Mahecha A, Elchebly M, et al. *Precise Quantitation of PTEN by Immuno-MRM: A Tool To Resolve the Breast Cancer Biomarker Controversy*. Anal. Chem. 93, 10816-10824 (2021). doi: 10.1021/acs.analchem.1c00975
- Sobsey CA, Ibrahim S, Richard VR, Gaspar V, Mitsa G, Lacasse V, Zahedi RP, Batist G, Borchers CH. *Targeted and Untargeted Proteomics Approaches in Biomarker Development.* Proteomics 20, e1900029 (2020). doi: 10.1002/pmic.201900029
- Blank-Landeshammer B, Richard VR, Mitsa G, Marques M, LeBlanc A, Kollipara L, et al. *Proteogenomics of Colorectal Cancer Liver Metastases: Complementing Precision Oncology with Phenotypic Data*. Cancers (Basel) 11(2019). doi: 10.3390/cancers11121907.

Contribution to publications in preparation *not included* in this Thesis:

- Mitsa G, Nagaria TS; Merza R, Sobsey C, Smolar Bocher J, Spatz A, Zahedi RP, Tourcotte S, Batist G, Borchers CH. Clinical cancer research on metastases from refractory colorectal cancer using a multi-omics approach to improve current precision oncology treatment.
- Mitsa G, Gaither C, Popp R, Zahedi RP, Borchers CH. *Internal TMT-based calibration curve allows for absolute quantitation of plasma proteins by parallel reaction monitoring (PRM).*
- Richard VR, Mitsa G, Esghi A, Chaplygina D, Qaswari D, Lee D, Thevis M, Borchers CH. *Non-invasive micro-sampling and targeted MS-based longitudinal blood proteome profiling to establish intra-individual protein reference ranges of athletes.*

Contribution to scientific journals as subject matter expert peer-reviewer (2019-2021)

- International Journal of Cancer
- Journal of Proteomics

List of Abbreviations

BCA	bicinchoninic acid assay
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CV	coefficient of variation
DCIS	ductal carcinoma in situ, breast ductal carcinoma in situ, ductal breast carcinoma in situ
DDA	data dependent acquisition
FASP	filter-aided sample preparation
FDR	false-discovery rate
FFPE	formalin-fixed, paraffin embedded
H&E	Hematoxylin & Eosin
ICPC	International Cancer Proteogenome Consortium
IDC	invasive ductal carcinoma
IHC	Immunohistochemistry
LFQ	label-free quantitation
mCRC	metastatic colorectal cancer
MS	mass spectrometry
NCI	US National Cancer Institute
NSCLC	non-small cell lung cancer
PAC	protein aggregation capture
PO	precision oncology
PRM	parallel reaction monitoring
PSM	peptide-spectrum matches
PTMs	post-translational modifications
RAC-BCA	reducing agent compatible bicinchoninic acid assay
STRAP	suspension trapping
TMA	tissue micro array
TMT	tandem mass tag

List of Figures and Tables

Chapter 1: General Introduction and Literature Review

1 Paradigm shift of clinical research

Genomic and transcriptomic molecular research has led to the discovery of driver alterations such as *EGFR*, *KRAS*, *BRAF*, *PDL-1* and *ERBB2* (also *HER2*), which facilitated precision oncology treatment with targeted agents for subtypes of colorectal cancer, metastatic breast cancer and non-small cell lung cancer, among other malignancies.[1,2] Although genetic screening provides a clear cellular blueprint of what *might* happen, risk and patient stratification based on the genetic phenotype alone has proven to have limited efficacy on the individual treatment outcome.[3] The vast majority of identified somatic mutations are likely passengers without oncogenic function, affect multiple genes or are not readily druggable. The molecular consequences of detected genetic and transcriptomic alterations are yet poorly understood and difficult to model with deep learning algorithms.

Multi-omic clinical cancer research encompasses methods integrating mass spectrometry (MS)-based measurements of protein abundance and post-translational modifications (PTMs), such as phosphorylation, as well as quantitation of metabolic active compounds and complement genetic and epigenetic data. As proteins and metabolites are the biologically active compounds, multi-omic data on tumors provides information on what *has* happened and therefore gives a 'real-world' snapshot of an individual tumor. This added layer of molecular information can be used to (i) better characterize genetic alterations, to better distinguish between driver and passenger mutations, (ii) it can be used to better understand molecular escape mechanisms of tumors leading to therapy resistance and (iii) can ultimately be used to identify (new) therapeutic vulnerabilities of cancer subtypes.

Early proteogenomic studies, conducted under auspices of the US National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC) and the International Cancer Proteogenome Consortium (ICPC), have already delivered new biological insights, actionable targets, and commented on future approaches for cancer diagnostics and treatment choices.[4-6] In this chapter, we will discuss current approaches in clinical research, with a focus on identifying common challenges and opportunities in the -omics field.

1.1 Genomics

DNA sequencing dates back to the 1970s, when Frederik Sanger and colleagues introduced the chain-termination method (Sanger sequencing), which uses radiolabelled chemical analogues of the deoxyribonucleotides and gel electrophoresis to determine the order of nucleic acids in biological samples.[7] Since then, large endeavors have been made to develop faster, more sensitive and more cost-effective techniques to study genes and their functions, often in relation to disease. Termed as next generation sequencing (NGS), these techniques share key principles behind Sanger sequencing, but use improved approaches for clonal amplification and detection (Figure 1), allowing automated high-throughput DNA-sequencing, while requiring less starting material and performing on the single-cell level.[8] Third generation and forth generation (3G and 4G, respectively) systems allow for real-time monitoring of nucleotide incorporation, and bypass the DNA amplification step. These systems are currently the fastest NGS systems on the market but are still quite expensive and error prone.[9,10] The generation of large data volumes further poses challenges for data management and data science. Consequently, second generation (2G) sequencing platforms are the most extensively used techniques, both for research and clinical use. Considerations for selection of a platform, their advantages and limitations are comprehensively reviewed by Gupta and Verma.[11]

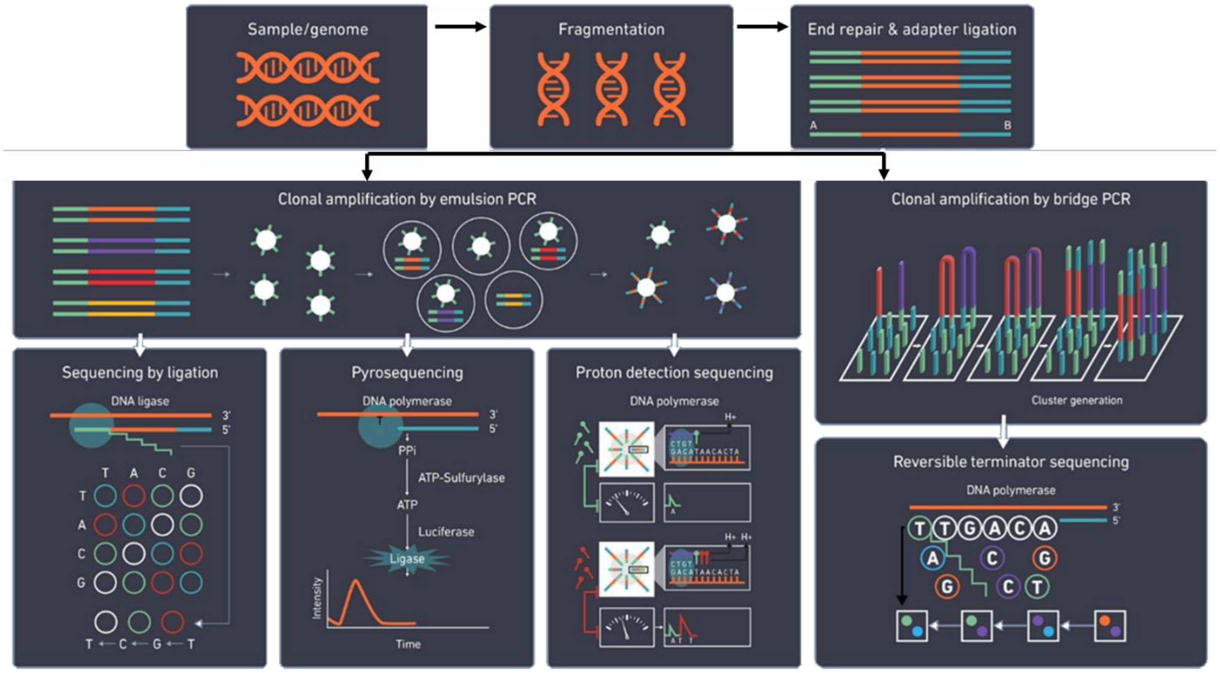

Figure 1: Schematic overview of next generation sequencing (2G) platforms. (originally created by Technology Networks 9)

NGS has been used for clinical research since the early 2000s, where methods like whole genome sequencing (WGS) and whole exome sequencing (WES) are being used for the study of genes and their involvement in diseases. Several cancer genomics databases and projects including the Human Genome Project, The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC) and the Catalog of Somatic Mutations in Cancer (COSMIC), have been initiated to generate a nearly complete coverage of the human genome and its alterations. Genomics studies inform about single nucleotide variants (SNVs), copy number variations (CNVs), somatic and germline mutations (mutational landscapes/signatures), and tumor mutational burden (TMB), which can be used to support diagnostic, preventative, and therapeutic strategies.

However, many of the genomic studies reported in the aforementioned databases are performed on untreated cancers and do not contain clinical annotations, therefore genetic events cannot be linked to specific cancer types, prognoses, or treatment response.[12] The Cancer Moonshot^{SM} Research Initiatives have since been formed to address this shortcoming, and aim to collect as much -omics and non-omics (e.g., clinical) information on disease states as possible.[13]

In the clinical setting, genomics is primarily used to identify somatic/germline mutations and for pharmacogenomics using traditional immunohistochemistry (IHC), quantitative polymerase chain reaction (qPCR), hybridization capture technique, or NGS panels (target-specific NGS).[14] The U.S. Food and Drug Administration (FDA) has approved NGS panels, such as Foundation One Dx, Oncotype Dx/DCIS or MammaPrint,[15] which focus on established cancer-associated genes (genetic hotspots) and frequently represent a cost-efficient solution for companion diagnostics due to the relatively lower demand in terms of data handling, storage, and analysis. Targeted genomic tests are the “gold standard” in clinical practice, and are an integral part of molecular tumor boards, where the identified genetic events are evaluated for guidance of the treatment regimens. Standard-of-care therapies, therefore, attribute to the diversity of sequence variants, inter-/intratumor heterogeneity and cellular plasticity.[16]

In conclusion, genomics has revolutionized cancer research and has provided tools for and insights into molecular biology. Single-gene defects have been discovered, and gene panels are available as companion-diagnostic tools to guide targeted treatments. However, the effect of genome-centred treatments has not been as expected.[17,18] Clinical phenotypes are more complex than the ‘one gene, one protein, one function’ paradigm implies. It has become clear that genomic changes and molecular mechanisms driving disease development and progression, involve other molecular layers, such as the transcriptome and the proteome. Poor correlation between independent genomics, transcriptomics, and proteomics studies shows that the clinical phenotype, including protein expression and pathway activity, is not describable by NGS alone, as these events are regulated on a co-transcriptional and post-translational level.

1.2 Transcriptomics

If we exclude epigenetic and conformational changes, the genome of an individual is mostly stable. Different phenotypes are the result of dynamic adjustments of the transcriptome (and proteome) to environmental stimuli.[19] According to the National Human Genome Research Institute, the transcriptome includes all of the gene readouts present in a cell,[20] and includes protein-coding mRNA, as well as non-coding tRNA, microRNA, ribosomal RNA, and long non-coding RNA.[21] According to the GENCODE database, the human transcriptome is comprised of more than 89,000 protein-coding mRNA isoforms.[22] These isoforms are results of co-transcriptional events, such as RNA-editing and alternative splicing. The global functional impact of these co-transcriptional events remains to be elucidated and requires proteome level information.[21]

Transcriptomics, or RNA-sequencing, has developed alongside genomics and uses similar instrumental platforms.[23] As an important part of NGS, RNA-sequencing enables differential gene expression analysis and provides information on gene fusions or allele-specific expression patterns, as well as information on mutational load and mutational signatures.[12,14,23] Although studies estimate that mRNA expression can serve as a surrogate for protein expression at steady state,[21] comparative studies targeting the transcriptome and proteome show a poor correlation between those two -omics fields.[24,25] This poor correlation is likely attributed to the complex translational machinery as response to pathway activation, and/or post-transcriptional modifications on the protein level, that cannot be assessed by DNA- or RNA-sequencing.[21,24,26]

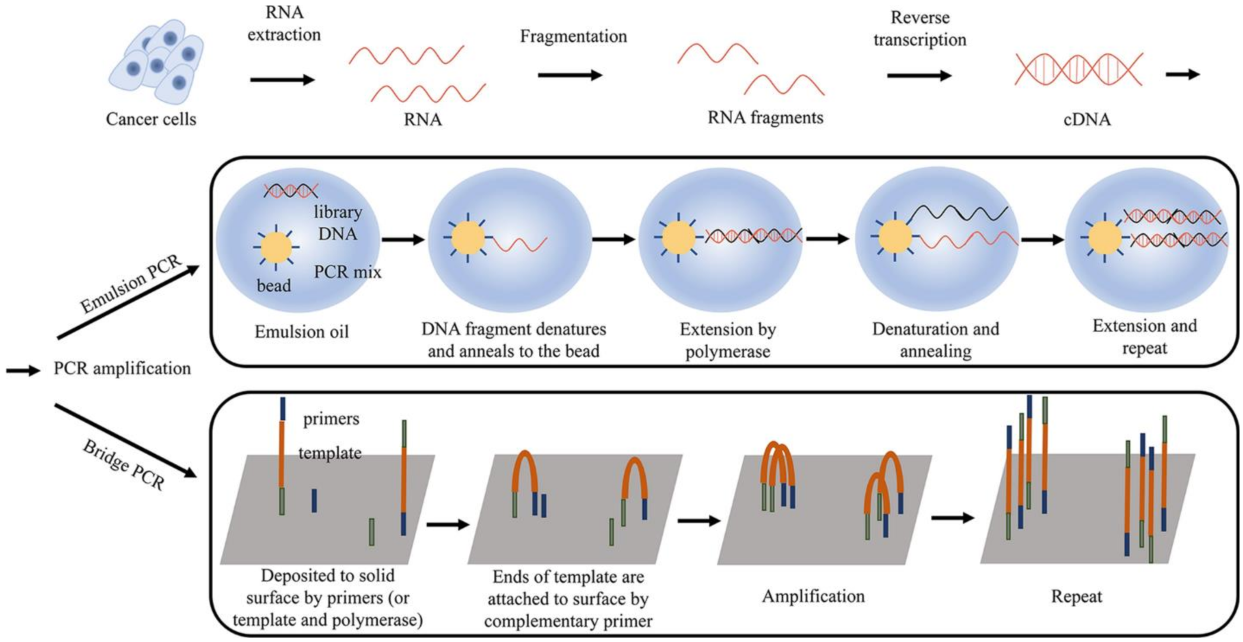

Figure 2: Schematic overview of RNA extraction and amplification steps for RNA-sequencing.[23]

1.3 Proteomics

Cancer has been widely considered a disease of the genome; a result of genetic mutations and chromosomal abnormalities that lead to genomic instability.[16] Proteins, however, are the genomic 'end products' that perform most biochemical functions and are the targets of most FDA-approved drugs.[27] Protein expression is the result of upstream genetic and epigenetic processes; cancer is, therefore, inherently a proteomic disease and quantitative proteomics is as important as the identification of (epi)genetic alterations.[28] Proteomics provides valuable information on the functional level, that neither DNA- nor RNA-sequencing can assess.

Proteomic studies have successfully described the clinical course of patients on the protein level and provide valuable molecular information on potential reasons for treatment resistance. While proteomics supports the treatment recommendations of DNA/RNA-based analyses in up to 57%,[16] mass spectrometry-based proteomics has the ability to refine genomic and transcriptomic guided lines of treatment, by providing unparalleled coverage of additional actionable therapeutic targets that are differentially expressed in various disease states and which are highly specific to the individual clinical case. The integration of quantitative proteomics into tumor profiling programs, and ultimately into clinical interdisciplinary molecular tumor boards, seems to be of special importance in cases with (extensively pretreated) advanced-stage-cancer or 'rare' malignancies, for which therapeutic options have been exhausted or are not yet available.

MS-based proteomics provides information on post-translational modification, such as phosphorylation, glycosylation, ubiquitination, nitrosylation and other proteoforms with unparalleled clinical sensitivity and specificity. Proteogenomic studies have associated DNA mutations -- either obtained from genomic databases or experimentally -- to protein signaling and provided mechanistic insights into genetic driver mutations.[1,2] Moreover, proteomics provides information on pathway activity, thus enabling the identification of new therapeutic vulnerabilities and allowing molecular subtyping of malignancies.[1,4,29] The Clinical Proteomic Tumor Analysis Consortium (CPTAC) and the Human Proteome Organization (HUPO) has defined guidelines from sample collection to data analysis and evaluation, in order to address variations in the proteomics data sets and to ensure high-quality data sets for clinical applications. To promote integrative studies, standard data submission and dissemination pipelines have been established.[30]

The increased functional diversity of proteins -- due to co-transcriptional and post-translational modification -- as well as the vast dynamic range of expression are technically challenging and require specialized sample preparation and data acquisition techniques. The dynamic range of proteins in eukaryotic cell lines and tissues spans 6-7 orders of magnitude, and 12 orders of magnitude in body fluids. There are currently more than 200 known post-translational modifications.[27] Consequently, most proteomic studies focus on technology development for improved sample preparation and quantitation of targets in limited amounts of material and/or where they are of low abundance. As for genomics and transcriptomics, data management and data analysis is another bottleneck, especially with the onset of technological advances.

Quantitative proteomics can be divided into two main applications: (i) untargeted proteomics, mainly used for discovery studies, and (ii) targeted proteomics for verification and validation studies. Untargeted quantitative proteomics applications are designed for in-depth unbiased quantitation of the global proteome and include several techniques (e.g., label-free, stable-isotope labeling, as well as chemical or metabolic labeling) with two acquisition modes: data-dependent acquisition (DDA), data-independent acquisition (DIA). Figure 3 gives an overview of available applications, which have been comprehensively described elsewhere.[28,31] Untargeted proteomics is limited by a low reproducibility and a high rate of missing values for low abundance proteins/peptides. This problem can partially be overcome by data imputation, labeled approaches, and by DIA applications.

Targeted proteomics methods include multiple reaction monitoring (MRM) and parallel reaction monitoring (PRM). Both methods provide high sensitivity, specificity, accuracy and reproducibility with high sample-throughput capabilities, and are the preferred methods for verification and validation studies, e.g., for biomarker development. MRM/PRM assay development is relatively time consuming, as it encompasses the selection of unique proteotypic peptides and the optimization of the LC-MS parameters (see Figure 4). To improve detection capabilities over the aforementioned wide dynamic range, special sample preparation techniques are required. These techniques include, but are not limited to, immunoaffinity enrichment to enrich low abundance proteins or proteoforms, immunoaffinity depletion to deplete high abundance interferences, and fractionation to reduce sample complexity. These methods need to be optimized for each target. Furthermore, targeted proteomics must follow strict guidelines to ensure accuracy and reproducibility. More recently, 4D Proteomics using (trapped) ion mobility spectrometry (TIMS or IMS) together with parallel accumulation-serial fragmentation (PASEF) are being explored for even more selective separation, identification, and quantitation of peptides and proteins on the single cell level.[32]

In conclusion, mass spectrometry-based protein analysis can complement and refine genomics/transcriptomics studies and has great potential for biomarker development.

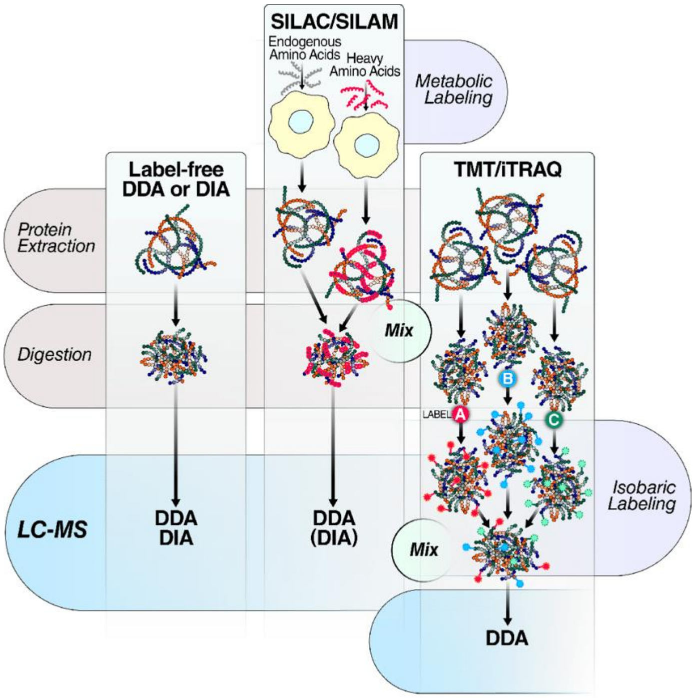

Figure 3: Overview of currently available untargeted proteomics approaches.[31]

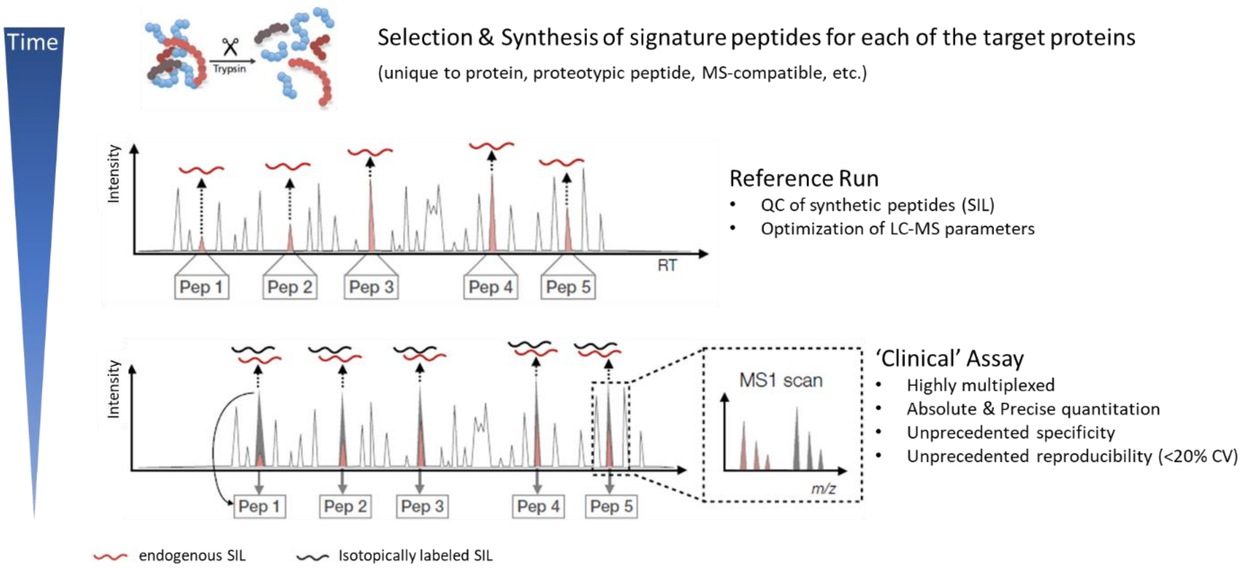
Time
Trypsin
Selection & Synthesis of signature peptides for each of the target proteins
(unique to protein, proteotypic peptide, MS-compatible, etc.)
Intensity
RT
Pep 1
Pep 2
Pep 3
Pep 4
Pep 5
Reference Run
• QC of synthetic peptides (SIL)
• Optimization of LC-MS parameters
Intensity
MS1 scan
m/z
Pep 1
Pep 2
Pep 3
Pep 4
Pep 5
'Clinical' Assay
• Highly multiplexed
• Absolute & Precise quantitation
• Unprecedented specificity
• Unprecedented reproducibility (<20% CV)
endogenous SIL
Isotopically labeled SIL

Figure 4: Overview of steps in the development of targeted proteomics approaches. (modified from 33)

2 Mass spectrometry for biomarker development

Biomarker development starts with the identification of molecular targets (e.g., genes, proteins, metabolites) that show significant changes between biological and pathogenic processes, and/or pharmacological responses using validated analytical techniques. Despite the technological advances in -omics methods for fast paced, large scale biomarker discovery, more than 99% of the published cancer biomarkers/genomic assays fail to enter clinical practice.[7] The majority of the identified targets are eliminated during analytical validation. Traditional validation methods, such as antibody-based immunoassays (e.g., Western Blot), enzyme-linked immunosorbent assays (e.g., ELISA), and/ or immunohistochemistry (IHC) lack strict performance metrices such as sensitivity, selectivity, precisions, accuracy, multiplexity, and/or high-interlaboratory reproducibility. Moreover, these techniques are limited in identifying/quantifying modified sequences, i.e., sequences that have genetic mutations or post-translational modifications. Therefore only a few targets reach the clinical validation and utility phase, where they are evaluated in -- or used as endpoints of -- clinical trials to demonstrate the relevance of the assay and the proposed biomarker for individual disease management.

Mass spectrometry (MS) has been a well-established technique for many years in clinical chemistry, where it has been used for the quantitation of metabolites and hormones. More recently, it has been acknowledged as a powerful and versatile tool for integrative clinical research and biomarker development, as it grants access to the profiling of small molecules, endogenous proteins and peptides in a vast dynamic range, and provides valuable insights into molecular changes in a single time-point or over a course-of-time. Targeted MS, specifically targeted proteomics using multiple reaction monitoring (MRM), has been chosen as Nature's method of the year in 2012, representing a standardized, highly reproducible and versatile platform for accurate quantitation of molecular targets in a variety of specimens (i.e., matrices) and dynamic ranges. Targeted MS can bridge the gap between discovery and successful verification and validation of biomarkers, thus facilitating their clinical translation.

3 Hypothesis and Objectives

Cancer is a complex disease involving many coexisting genomic alterations with rising evidence, that the vast majority of somatic mutations has likely little to no specific oncogenic function. We therefore hypothesize that (i) quantitative proteomics will better elucidate the clinical phenotype of a tumor, as proteins are biologically active products of gene expression, and (ii) complementing clinical genotyping with comprehensive phospho-/proteomic data as part of clinical proteomics will better define therapeutic guidance.

The primary objective of this study is to go beyond the identification of genomic driver aberrations by complementing genomics data on tumors with quantitative measurements of the proteome in order to (i) improve the prediction of therapy response, (ii) obtain a deeper insight into tumor biology and resistance mechanisms, and (iii) identify (novel) targets for precision oncology treatment.

This project is funded by the Terry Fox Research Institute, the Quebec Cancer Consortium and Genomic Applications Partnership Program from Genome Canada/Quebec and is part of the Cancer Moonshot Initiative.

4 Ethics

This research project has been reviewed and approved by the Research Ethics Board of the CIUSSS West Central Montreal (Project 2020-1752). The studies were conducted in accordance with Good Clinical Practice, the guiding principles of the Declaration of Helsinki, and all applicable regulatory requirements.

Preface Chapter 2

Clinical tissue samples are typically archived as formalin fixed and paraffin embedded (FFPE) tissue blocks. Proteomic analysis of FFPE tumor tissue specimens has gained interest in the past 5 years due to technological advances and improved biobanking for clinical trials. The ‘real-world’ implementation of clinical proteomics to these specimens, however, is hampered by tedious sample preparation steps and long instrument acquisition times of more than 2 hours.

To advance the translation of quantitative proteomics into the clinic, the candidate developed a streamlined, automatable procedure that enables quantitative proteomic analysis of limited volume clinical samples, such as core needle biopsies, that can be collected over the time course of disease management. This method can be used for both clinical research, i.e., biomarker development or clinical trials, and as companion diagnostic device in a clinical setting, as it facilities sample processing from extraction and analysis of clinical biomarkers within 24 hours from pathological assessment, and does not require specialized equipment or expertise.

This method is used for integrative translational research conducted at the Segal Cancer Research Centre, by partners of the Marathon of Hope Cancer Centres Network and other research groups.

Author Contributions: The following manuscript was written by the candidate and reviewed and edited by Rene P. Zahedi. Experiments and data analysis were conducted by the candidate under supervision by Rene P. Zahedi, Christoph H. Borchers and Gerald Batist. Qianyu Guo and Samuel E.J. Preston generated tumor specimens. Adriana Aguilar-Mahecha, Naciba Benlimame, Mark Basik, Alan Spatz, Vincent Lacasse, Wilson H. Miller Jr., and Sonia V. del Rincon provided other resources.

Chapter 2: A Non-Hazardous Deparaffinization Protocol Enables Quantitative Proteomics of Core Needle Biopsy-Sized Formalin-Fixed and Paraffin-Embedded (FFPE) Tissue Specimens

Georgia Mitsa[1,2], Qianyu Guo[1,3], Christophe Goncalves[3], Samuel E. J. Preston[3], Adriana Aguilar-Mahecha[3], Naciba Benlimame[4], Mark Basik[1,3], Alan Spatz[3,4], Vincent Lacasse[2,4], Gerald Batist[1,3,5], Wilson H. Miller, Jr.[1,3,6], Sonia V. del Rincon[1,3], René P. Zahedi[2,*,†,‡] and Christoph H. Borchers[2,3,*]

1. Division of Experimental Medicine, McGill University, Montreal, QC H4A 3J1, Canada; georgia.mitsa@outlook.com (G.M.); qianyu.guo@rocketmail.com (Q.G.); mark.basik@mcgill.ca (M.B.); gerald.batist@mcgill.ca (G.B.); wilsonmiller@gmail.com (W.H.M.J.); sonia.delrincon@gmail.com (S.V.d.R.)
2. Segal Cancer Proteomics Centre, Lady Davis Institute, Jewish General Hospital, McGill University, Montreal, QC H3T 1E2, Canada; vincent.lacasse2@mail.mcgill.ca
3. Gerald Bronfman Department of Oncology, Jewish General Hospital, McGill University, Montreal, QC H4A 3T2, Canada; goncalves.christophe@gmail.com (C.G.); samuel.preston@mail.mcgill.ca (S.E.J.P.); nanaaguilar@gmail.com (A.A.-M.); alan.spatz@mcgill.ca (A.S.)
4. Research Pathology Core Facility, Department of Pathology, Segal Cancer Centre, Lady Davis Institute, Jewish General Hospital, McGill University, Montreal, QC H3T 1E2, Canada; nbenlimame@jgh.mcgill.ca
5. Exactis Innovation, 5450 Cote-des-Neiges, Suite 522, Montreal, QC H3T 1Y6, Canada 6 Rossy Cancer Network, McGill University, Montreal, QC H3H 1E8, Canada

* Correspondence: rene.zahedi@umanitoba.ca (R.P.Z.); christoph.borchers@mcgill.ca (C.H.B.); Tel.: +1-204-789-3639 (R.P.Z.); +1-514-340-8222 (ext. 7886) (C.H.B.)

†Current address: Department of Internal Medicine, University of Manitoba, Winnipeg, MB R3E 0Z2, Canada. ‡ Current address: Manitoba Centre for Proteomics and Systems Biology, Winnipeg, MB R3E 3P4, Canada.

Received: 24 March 2022 Accepted: 11 April 2022 Published: 18 April 2022

Published in International Journal of Molecular Science (doi: 10.3390/ijms23084443)

Abstract

Most human tumor tissues that are obtained for pathology and diagnostic purposes are formalin-fixed and paraffin-embedded (FFPE). To perform quantitative proteomics of FFPE samples, paraffin has to be removed and formalin-induced crosslinks have to be reversed prior to proteolytic digestion. A central component of almost all deparaffinization protocols is xylene, a toxic and highly flammable solvent that has been reported to negatively affect protein extraction and quantitative proteome analysis. Here, we present a 'green' xylene-free protocol for accelerated sample preparation of FFPE tissues based on paraffin-removal with hot water. Combined with tissue homogenization using disposable micropestles and a modified protein aggregation capture (PAC) digestion protocol, our workflow enables streamlined and reproducible quantitative proteomic profiling of FFPE tissue. Label-free quantitation of FFPE cores from human ductal breast carcinoma in situ (DCIS) xenografts with a volume of only 0.79 mm^3 showed a high correlation between replicates ($r^2 = 0.992$) with a median %CV of 16.9%. Importantly, this small volume is already compatible with tissue micro array (TMA) cores and core needle biopsies, while our results and its ease-of-use indicate that further downsizing is feasible. Finally, our FFPE workflow does not require costly equipment and can be established in every standard clinical laboratory.

Keywords: clinical proteomics; tumor tissues; FFPE; quantitative proteomics; core needle biopsy; cancer research; molecular pathology; breast ductal carcinoma; in situ cancer

1 Introduction

Most human tumor tissues that are obtained for pathology and diagnostic purposes are formalin-fixed and paraffin-embedded (FFPE).[1] FFPE allows the preservation of tissues in a "life-like" state while preserving spatial features and keeping them accessible for subsequent downstream analyses, such as immunohistochemistry (IHC), and genomic (hotspot) sequencing, without the requirement of expensive equipment for sample storage.[2-5] Vast FFPE tissue archives are available in clinics around the globe. These represent an invaluable resource for precision oncology and clinical research[1,6] because the archives often include pathological, clinical, and outcome data that are linked to the clinical samples, which have often been collected from a patient during different stages of disease.

The MS-based proteomic analysis of FFPE samples has gained increasing attention during the past decade, not only because of improved protocols for the extraction of proteins, but also because of substantial improvements in the overall sensitivity of MS instrumentation and workflows.[7] These technological advances now enable the proteomic profiling of minute sample amounts. Both system-wide 'discovery' data on aberrant protein expression/signaling pathway activity and targeted data providing actual protein concentrations for selected protein targets can provide important phenotypic information from FFPE samples that cannot be extracted from genomic screening or from IHC staining, and which may be in disagreement with genomic information.[8-10] Nevertheless, quantitative proteomics of FFPE samples, i.e., minuscule FFPE cores (down to 1 mm in diameter) or thin FFPE slices (down to 5 µm thick), which are key to utilizing the full potential of FFPE proteomics, are still far from routine. This can be partially attributed to the challenges in obtaining a well-defined and homogenous tissue sample, which may require micro-/macro-dissection to enrich for the area of interest following examination by a pathologist of a hematoxylin and eosin (H&E) stained representative slide. Once a well-defined FFPE sample has been obtained, paraffin has to be removed in order to make the sample amenable to MS-based proteomics, as paraffin interferes with the subsequent steps of proteomic sample preparation and MS analysis.[11] In addition, formalin-induced crosslinks have to be reversed prior to protein extraction, which is then typically followed by a bottom-up proteomic workflow. A central component of almost all deparaffinization protocols is xylene, which is a toxic and highly flammable solvent.[12,13] It has also been reported that xylene may negatively affect protein extraction and quantitative proteomic analysis.[3,14] Most studies agree that heat, the choice of detergents and protein denaturants, as well as the pH of the extraction buffer and physical agitation, are important parameters affecting the efficacy of protein extraction.[3,15]

To make quantitative proteomics of FFPE cores and slides more streamlined and easier to automate for use in the clinic, as well as to expand its use beyond the current focus on retrospective studies, we present here an accelerated and efficient workflow for FFPE proteomics that does not require xylene and which can be set up in any standard clinical laboratory, as it requires neither costly equipment nor special training (Figure 1). The absence of xylene in our protocol is in line with the principles of "green chemistry" [16] as it avoids the use of chemicals that are both hazardous to nature and involve a hazardous synthesis.

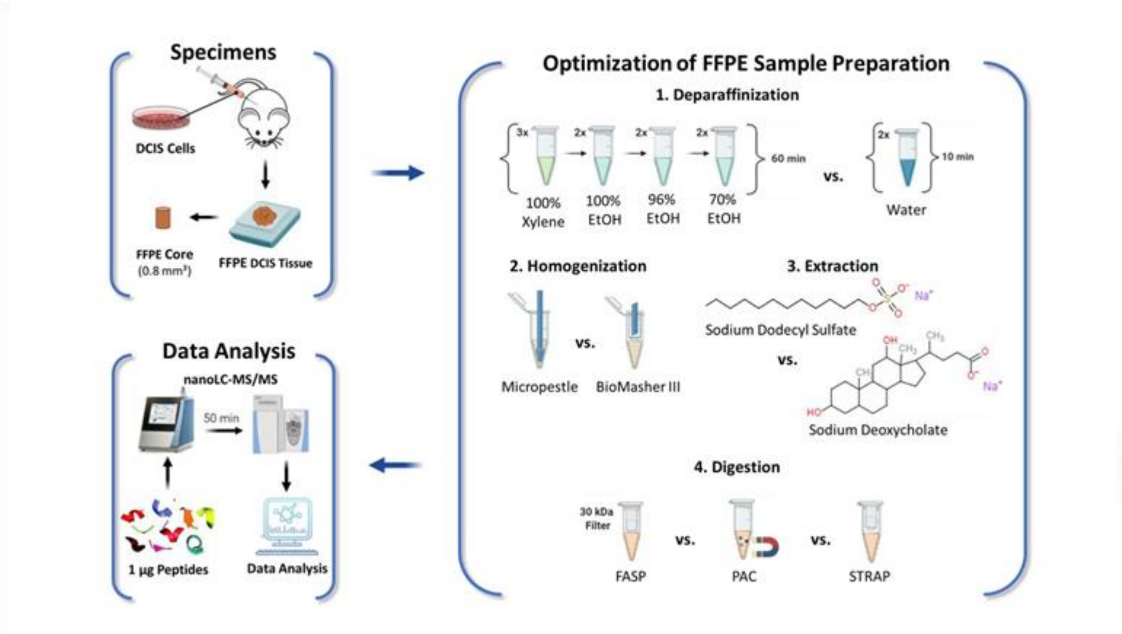

Figure 1. Experimental Design. Xenografts were generated from human DCIS cells and tumors were resected after 1.5 weeks, followed by formalin-fixation and paraffin-embedding, as described in.[17] One-millimeter-diameter FFPE cores were used to optimize individual steps of the FFPE sample preparation: (1) deparaffinization, (2) homogenization, (3) extraction, and (4) digestion. Peptide samples were analyzed by nano-LC-MS/MS label-free quantitation (LFQ) to compare the performance of the evaluated protocols for each step of the sample preparation workflow.

2 Results

2.1 Water-Based Deparaffinization Competes with the Gold-Standard Xylene and Takes Only a Fraction of the Time

The initial step, common to all FFPE sample preparation protocols, is deparaffinization, and the protocol used in most laboratories is essentially the reversal of the paraffinization procedure, comprising many steps that cannot be readily automated and are time-consuming: e.g., sequential washing steps with xylene and decreasing concentrations of ethanol (100%, 96%, and 70%). Our goal was to develop a simpler and safer protocol.

We, therefore, compared the standard xylene-procedure to a deparaffinization method that is based on washes with hot water (*depW*) without toxic solvents. The use of water for deparaffinization had been first suggested in 2016 by Kalantari et al. for DNA extraction,[18] and 2017 by Mansour et al. for Western blot analysis.[14]

After deparaffinization with either xylene (*depX*) or hot water (*depW*), the samples were homogenized using a disposable micropestle (*homMP*) in a sodium deoxycholate (SDC)-based buffer (*exSDC*) and digested using FASP. Details can be found in Section 4.3.1.

Our data show that two cycles of short incubation (~5 min) in hot water is sufficient to efficiently retrieve proteins from FFPE cores, thus avoiding laborious successive washes with organic solvents (8 steps, 60 min). Water-based deparaffinization (*depW*) yielded, on average, 89 ± 17 µg of total protein per mg (dry weight) of FFPE core, compared to 97 ± 23 µg using xylene (*depX*; unpaired *t-test* $p = 0.54$, Figure 2A).

Database searches performed on individual samples led to the identification of 933 ± 26 (*depX*) and 835 ± 80 (*depW*) proteins, 6778 ± 294 (*depX*) and 5925 ± 608 (*depW*) peptides, and 8400 ± 352 (*depX*) and 7244 ± 815 (*depW*) peptide-spectrum matches (PSM), respectively, from 23,008 ± 561 (*depX*) and 22,764 ± 1378 (*depW*), acquired MS/MS spectra. A quantitative comparison of the five *depW* and *depX* replicates using label-free quantitation (LFQ) led to the quantitation of 1502 (*depW*) and 1521 (*depX*) unique proteins across the five replicates, with intra-method CVs of 16.8% (*depW*) and 16.9% (*depX*; Figure 2B). The LFQ-derived intensities of 1495 proteins that were quantified between the two methods show a good agreement between *depW* and *depX* (correlation coefficient R = 0.94, slope = 0.97, Figure 2C). In general, the choice of the deparaffinization method did not significantly impact the recovery of hydrophobic or hydrophilic proteins (see Supplementary Table S1); however, almost half of the quantified high-abundance cytosolic ribosomal proteins showed poorer recoveries with *depW* (Figure 2D; Benjamini–Krieger adjusted *p* value <0.01, median *depW/depX* = 0.25), while membrane and nuclear proteins of interest in cancer biology, such as TOMM5,[19,20] TOMM7,[19,21] RAB18 (RAS related protein),[22,23] and nuclear BCCIP (BRCA2 interacting protein),[22,24] had significantly better recoveries using *depW* (adj. $p < 0.01$, *depW/depX*= >2.08). Notably, other important cancer proteins, such as EGFR (adj. $p = 0.20$, *depW/depX* = 1.36), EIF4E (adj. $p = 0.13$, *depW/depX* = 1.19), or AKT1S1 (adj. $p = 0.31$, *depW/depX* = 1.71) seem to show better recoveries using water-based deparaffinization, but this was not statistically significant.

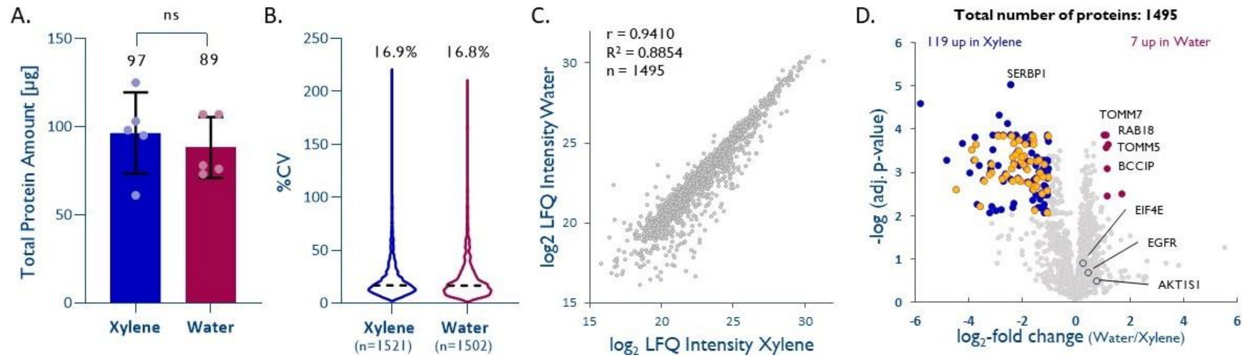

Figure 2. Water-based deparaffinization is a 'green' alternative. (A) Total protein extracted after deparaffinization with either water (*depW*) or xylene (*depX*) (*n* = 5, unpaired *t-test, p* = 0.54). (B) Intra-method %CVs based on quantified proteins, median %CV are given. (C) Pearson correlation plot based on all proteins quantified by both methods. (D) Volcano plot highlighting proteins significantly enriched by either method (Benjamini–Krieger multiple hypothesis testing, FDR 1%). Cytosolic ribosomal proteins significantly enriched with *depX* are shown in orange. EIF4E, EGFR, and AKT1S1 are highlighted in dark grey.

2.2 Efficient Tissue Homogenization Using Micropestles

Next, we evaluated different tools for the homogenization of small tissue samples. We compared the total protein amounts extracted with a disposable, autoclavable micropestle (*homMP*) or a BioMasher III (*homBM*) which is a micropestle with a filter unit that is commonly used in genomics studies.[25-28] The samples were deparaffinized using *depW*, homogenized with either *homMP* or *homBM* in *exSDC* and digested using *FASP*. Details can be found in Section 4.3.1.

Based on the bicinchoninic acid assay (BCA), both methods yielded similar protein amounts (95 ± 19 µg for *homMP* vs. 90 ± 13 µg for *homBM*; Figure 3A). LFQ of the five replicates per method enabling the quantitation of 1405 (*homMP*) and 1364 (*homBM*) unique proteins showed that both methods are equally reproducible (%CVs of 20.9% and 21.0% for *homMP* and *homBM*, respectively; Figure 3B). LFQ-derived normalized abundances of 1410 proteins that were quantified by both methods showed a good correlation (r = 0.94, slope = 0.98, Figure 3C), with a tendency toward higher intensities for *homMP* (Figure 3D). The proteins with significantly differential recovery (Figure 3D), however, do not seem to indicate a role of pI, hydrophobicity, subcellular localization, or molecular weight in the enrichment/depletion with either method (see Supplemental Table S2). Based on a slightly better overall performance, ease of use, and lower cost, we prefer the disposable micropestle to the BioMasher III.

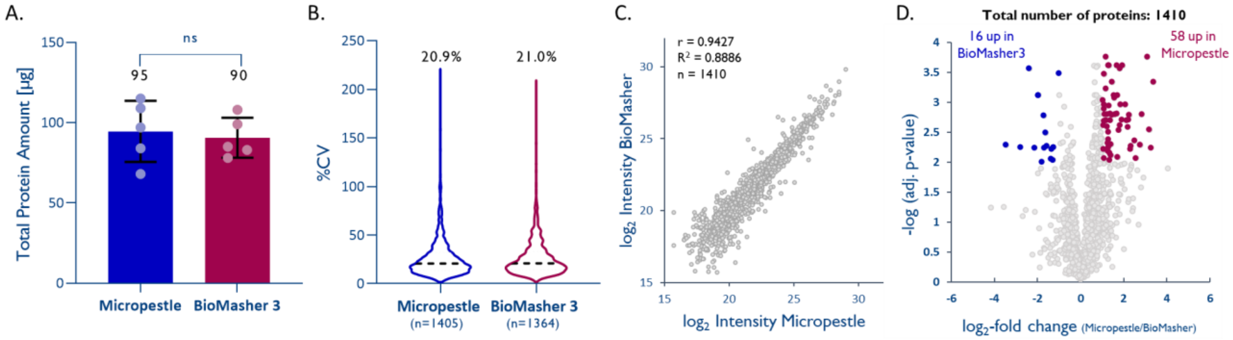

Figure 3. Efficient tissue homogenization using micropestles. (A) Total protein extracted from 1 mm cores with a dry-weight < 1 mg (*n* = 5; unpaired *t-test; p* = 0.70). (B) Intra-method %CVs based on all quantified proteins, median %CV are given. (C) Pearson correlation plot based on all quantified proteins. (D) Volcano plot highlighting proteins that were significantly enriched by one method (multiple hypothesis testing using the FDR-based approach by Benjamini–Krieger, FDR 1%).

2.3 Improved Protein Extraction with Sodium Deoxycholate (SDC)

After mechanical cell disruption using a disposable micropestle (*homMP*), preceded by deparaffinization using hot water (*depW*), the homogenate was incubated in different extraction buffers, and digested using *FASP*.

Formalin fixation of tissue preserves proteins in their native structures, but studies suggest that several modifications may occur during the fixation process, and these seem to progress over time.[29-31] Lysine methylation and methionine oxidation are the most frequent protein modifications observed in FFPE tissue.[30,31] These protein modifications can ultimately lead to protein–protein, DNA–protein, and/or RNA–protein crosslinking (e.g., Schiff base reaction). Combining heat[32,33] with high concentrations of detergents is considered to be the most effective approach for protein extraction from FFPE tissue and the reversal of protein-crosslinking that had been induced during the fixation process. We, therefore, compared a standard SDS buffer (*exSDS*), as used in many studies,[3,15] to an SDC-based buffer (*exSDC*).[34,36] *exSDC* contains TCEP, a potent denaturing agent, which seems to improve the denaturation of FFPE-preserved proteins and facilitates downstream processing. We use the identification of peptides with lysine methylation as an indicator of effective denaturation and decrosslinking in FFPE samples. We observed 4.95 ± 0.62% of these modified peptides with *exSDS* and 3.62 ± 0.26% with *exSDC*. This indicates that both lysis buffers are effective for decrosslinking, but *exSDC* seems to be marginally better.

In our hands, the *exSDC* buffer yielded significantly higher amounts of total protein (unpaired *t-test, p* = 0.0341, Figure 4A) than the SDS buffer, which is thought to be the 'gold-standard'. LFQ data on 1748 (*exSDC*) and 1765 (*exSDS*) quantified unique proteins also showed a higher reproducibility for *exSDC* than for *exSDS* (%CV of 14.7% vs. 21.3%) (Figure 3B). A direct comparison using LFQ enabled the quantitation of 1722 unique proteins quantified by both methods and showed a slight tendency towards higher intensities, and consequently, are more likely to be a better extraction efficiency for *exSDC* (Figure 3C,D). A closer look into the physicochemical properties of proteins exclusively quantified in each method reveals that *exSDC* is significantly better for extraction of larger (unpaired *t-test, p* = 0.0001), more acidic (unpaired *t-test, p* = 0.005), and more hydrophilic proteins (unpaired *t-test, p* = 0.0001) than *exSDS* (see Supplemental Table S3).

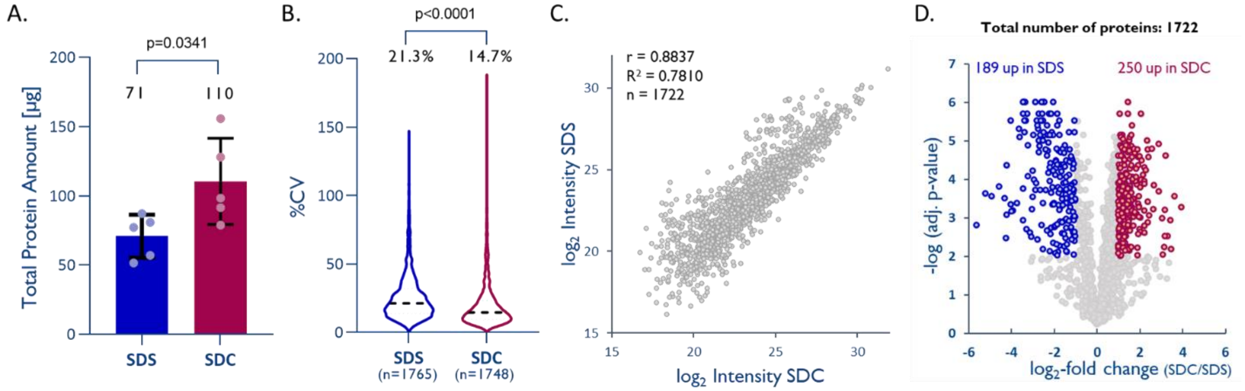

Figure 4. An SDC–TCEP-based buffer improves overall protein recovery from FFPE tissues. (A) Total protein extracted from 1 mm cores with dry-weight < 1 mg (*n* = 5; unpaired *t-test; p* = 0.0341). (B) Intra-method %CVs based on all quantified proteins, median %CVs are given (unpaired *t-test; p* < 0.0001). (C) Pearson correlation plot based on all proteins quantified with both methods. (D) Volcano plot highlighting proteins that were significantly enriched by either method (multiple hypothesis testing using FDR-based approach by Benjamini–Krieger, FDR 1%).

2.4 PAC and STRAP Are Good Alternatives to FASP

Next, we evaluated the efficacy of different digestion techniques for FFPE-proteomics, namely, filter-aided sample preparation (*FASP*),[37,38] which is still one of the most-widely used sample preparation methods, in addition to the more recent protein-aggregation capture (*PAC*)[39] and suspension trapping using micro spin columns (*STRAP*)[40] that enable a simpler sample preparation with improved parallelization and automation capabilities.[31,41-44] For each method, 20 µg of total protein were digested after water-based deparaffinization, micropestle homogenization, and SDC-based extraction (*depW/homMP/exSDC*).

Important differences in the three digestion procedures are the (i) time and temperature of incubation with trypsin, and (ii) the substrate-to-protein ratio. While *FASP* typically involves overnight digestion at 37 °C, *PAC* and *STRAP* digestions are usually performed for 3 h at 37 °C and 47 °C, respectively. *FASP* and *PAC* samples were digested with a substrate-to-trypsin ratio of 20:1, while *STRAP* samples were digested with a substrate-to-trypsin ratio of 10:1, following the manufacturer's instructions. Importantly, although it worked well for conventional samples, in our hands, the standard *PAC* protocol led to a poor performance for FFPE samples, which seemed to result from an overall poor recovery of proteins/peptides. We were able to compensate for this by adjusting the standard protein-to-bead ratio from 1:4 to 1:12, which led to substantially higher signals in the analysis, comparable to the other two methods. Notably, for *STRAP*, the buffer was brought to a final concentration of 5% SDS before loading, as recommended by the manufacturer. When comparing the three digestion protocols, *FASP* yielded a clearly higher proportion of fully tryptic peptides (79%) compared to *PAC* and *STRAP* (both 70%; Figure 5A). LFQ of the individual methods showed that *FASP* also had the lowest intra-method %CV (15.6%, *n* = 1496) followed by *PAC* with 16.9% (*n* = 1482) and *STRAP* with 17.7% (*n* = 1496; Figure 5B). A quantitative comparison of the three methods based on LFQ led to the quantitation of 1436 unique proteins quantified by all methods. Unsupervised hierarchical clustering based on the normalized abundances of these proteins (multiple hypothesis testing using FDR-based approach by Benjamini–Krieger, FDR 1%, Figure 5C) shows good agreement between the three methods. A closer look into the physicochemical properties of proteins that were exclusively recovered with one of the methods shows that *PAC* seems to be better suited for small and more-acidic proteins, although not statistically significant (see Supplementary Table S4). A pair-wise comparison of the three methods by LFQ shows that *FASP* results in a slightly better overall protein recovery than *PAC*, while both methods are superior to *STRAP* (see Figure 5D–F).

Thus, *PAC* is a good alternative to the considerably more laborious and time-consuming *FASP*, is also easily scalable to the amount of protein, and has already been fully automated using liquid-handling systems.[41,45-48]

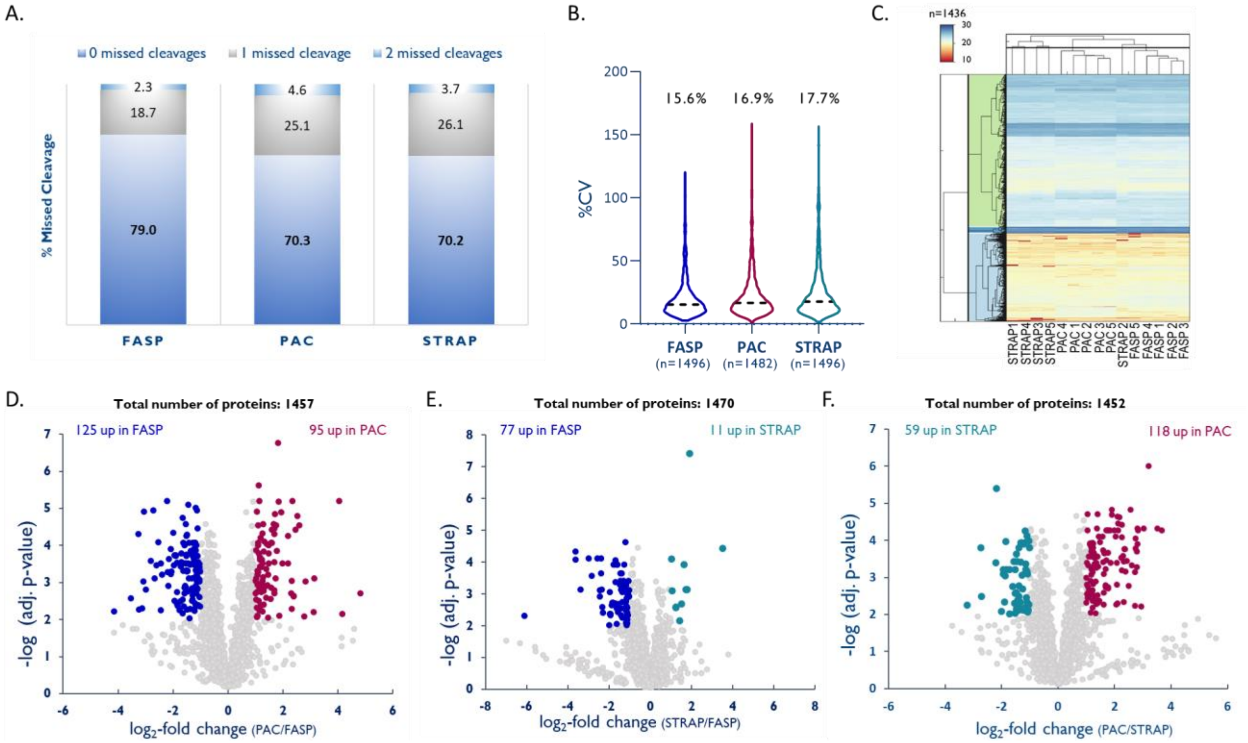

Figure 5. Comparison of PAC and STRAP with FASP. (A) Efficacy of tryptic digestion shown by percentage of missed cleavages. (B) Intra-method %CVs based on all quantified proteins, median %CVs are given. (C) Hierarchical clustering [49] of all quantified proteins, colors reflect log2-normalized abundances. (D–F) Volcano plots highlighting proteins that were significantly enriched by either method (multiple hypothesis testing using FDR-based approach by Benjamini–Krieger, FDR 1%).

3 Discussion

In this paper, we present a streamlined and efficient protocol for quantitative proteomics of FFPE cores based on a novel, ‘green’, and non-hazardous water-based method for deparaffinization (*depW*) prior to quantitative proteomic analysis, and show that harsh, non-MS compatible detergents, such as SDS, are not required for efficient retrieval of proteins from archived FFPE material. The water-based deparaffinization results in paraffin-removal that is six times as fast as the conventional xylene-based protocol and enables robust proteomic profiling from less than 1 mg of FFPE core tissue (dry weight with wax) for clinical research. To our knowledge, this is the first report of a protocol for efficient protein retrieval from 1 mm-diameter core punches (~0.8 mm^3 tissue volume). This volume is comparable to tissue micro array (TMA) cores used to build TMAs for IHC analysis, and also to core needle biopsies, thus enabling proteomic analysis of clinical samples that were not amenable to clinical research due to very limited tissue availability, i.e., ductal carcinoma in situ, where tumor areas of interest are too small for analyses beyond IHC.

Interestingly, several membrane receptors with central roles in cancer biology showed significantly better recoveries ($p < 0.01$, ≥2-fold) using our water-based method compared to xylene-based protocols (e.g., TOMM5/7, RAB18, BCCIP).[50] Other relevant cancer targets, such as translational elongation and initiation factors (e.g., EIF4E), membrane receptors (e.g., EGFR), as well as RNA binding proteins, also showed a tendency towards better recovery, albeit not statistically significant. These results may reflect the differential impact of water- and several rounds of ethanol-washes on the partial removal of either soluble or hydrophobic proteins.

More recent protocols for proteomics of FFPE samples include the use of Adaptive Focused Acoustics (AFA) for efficient deparaffinization and decrosslinking,[40,41,43,51] but the availability of AFA systems that can handle multiple samples in standard laboratories is limited because of the high cost of both the instrumentation and the required consumables. In our hands, sonication of the small sample amounts used in this study results in a high risk of sample loss, even when a single tube is used. A heat-based homogenization/lysis in SDC buffer with TCEP, together with autoclavable micropestles, enabled effective decrosslinking, homogenization, and extraction, using a simple protocol. A BCA kit compatible with reducing agents (RAC-BCA) is recommended to account for interferences in the colorimetric detection and quantitation of total protein.

The translation of FFPE-based proteomics into high-throughput (clinical) applications requires fast, reproducible, scalable, and (semi)automated workflows. We, therefore, performed a direct comparison of more recent techniques for protein extraction and digestion (i.e., *PAC* and *STRAP*), to the ‘gold-standard’ filter-assisted sample preparation (*FASP*).

Our data reveal differences in the efficiency of tryptic digestion and the feasibility for use on low sample amounts (scalability). Although *FASP* performs slightly better in our hands, it is a very laborious procedure including several washing steps that cannot be automated. It has been shown to work best for a protein range between 10 µg to 100 µg.[52,53] *STRAP* is a very attractive sample preparation method because the available cartridge formats cover a wide range of total protein amounts (1–100 µg, 100–300 µg or ≥300 µg).[40] It also allows tryptic digestion in as little as 3 h, with comparable digestion efficiency to overnight digestion used for *FASP*. In our hands, however, *STRAP* was slightly less reproducible than *PAC* and *FASP*, with *PAC* being the best choice for very low sample amounts. *PAC* requires tryptic digestion for only 3 h and has already been successfully automated in several laboratories.[41,45-48] Moreover, a variety of bead-chemistries are available (e.g., amine-reactive, carboxylic, HILIC, etc.) which might be adapted based on specific research needs. Notably, to enable efficient PAC-based sample preparation for FFPE samples, we had to increase the recommended protein-to-bead ratio by a factor of three, likely as a result of FFPE-matrix effects that reduced the protein binding capacity.

4 Materials and Methods

We used FFPE tissue from human ductal breast carcinoma in situ (DCIS) xenografts to optimize the main steps of FFPE sample preparation: (i) deparaffinization, (ii) homogenization, (iii) protein extraction, and (iv) proteolytic digestion (see Figure 1). To avoid a systematic bias derived from tissue heterogeneity, we used 1 mm-diameter cores that had been obtained from a single FFPE block for each step (i–iv) of the protocol to be optimized, and randomly assigned these to the different protocols in order to have a total of 5 replicates per condition. Thus, any statistically significant differences observed should result from methodical differences rather than tissue heterogeneity.

Each method was evaluated based on the total protein yield (RAC-BCA), the number of identified peptides and proteins (qualitative MS), as well as the reproducibility and the enrichment/depletion of proteins (quantitative MS).

4.1. Chemicals and Reagents

All chemicals and reagents were purchased from Sigma Aldrich (St. Louis, Michigan, USA) unless otherwise stated. For sample homogenization, two types of micropestles were acquired, one from Sigma Aldrich (#BAF199230001) and one Optima Inc., Glencoe, IL, USA, (#320302). Filter-aided sample preparation (*FASP*)[37,38] was conducted on Microcon® Centrifugal Filters (30 kDa molecular cut-off, Merck KGaA #MRCF0R030, purchased through Sigma Aldrich. For bead-based sample preparation, ferromagnetic beads with MagReSyn® Amine functional groups (ReSyn Biosciences, Gauteng, South Africa) were used. For sample preparation using suspension trapping, S-Trap[51] micro-cartridges with a binding capacity of <100 µg total protein were purchased from ProtiFi (Farmingdale, NY, USA, #CO2-micro-80).

The total protein concentration was determined using a reducing-agent-compatible Pierce BCA Protein Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA, #23250) following the manufacturer's instructions.

4.2 Source of Specimens

Method development was performed using 1 mm-diameter FFPE cores (~0.8 mm^3 tissue volume) of xenografts from human ductal breast carcinoma in situ (DCIS, Figure 6). A total of 1 × 10^5 human DCIS cells (MCF10DCIS.com; Wake Forest University, Winston-Salem, NC, USA) were injected into the mammary fat pads of an athymic nude mouse. DCIS tumors were resected after 1.5 weeks, were formalin-fixed, paraffin-embedded, and stored under ambient conditions (~1 year).[17]

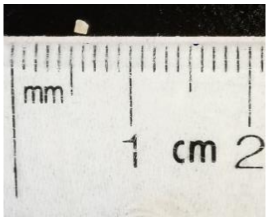

Figure 6. Representative size of FFPE core used in this study. After deparaffinization, the core volume was approximately 0.4 mm^3.

4.3 Sample Preparation of Core Needle Biopsy-Sized Specimens

4.3.1 Optimization of Deparaffinization

Recovering proteins from FFPE tissue requires prior deparaffinization with the vast majority of protocols being based on xylene, followed by a series of washes with decreasing amounts of ethanol for tissue rehydration.[2,31,42] We compared a standard protocol for xylene/ethanol washes (*depX*) used in most clinical laboratories for paraffin removal, to a 'green' and solvent-free deparaffinization approach based on hot deionized water (80 °C; *depW*) using 5 FFPE cores per protocol.

(a) *depX*:[54] The samples were washed with 1 mL of 100% xylene and incubated for 10 min at room temperature (RT), followed by centrifugation at 14,000x g for 2 min and disposal of the supernatant, followed by another 2 repetitions. Then, the samples were washed twice each with 1 mL of 100%, 96%, and 70% ethanol, followed by incubation for 1 min at RT and centrifugation as above.

(b) *depW*:[(modified from 14)] The samples were washed 2x with 500 µL of hot deionized water and incubated for 1 min at RT under vigorous vortex mixing. Each washing step was followed by centrifugation at 20,000x g for 5 min at 4 °C. The supernatant, containing paraffin either floating on the liquid surface or stuck to the wall of the tube (Figure 7), was discarded and the deparaffinized and rehydrated core was transferred to a clean LoBind Eppendorf tube.

Each core was mechanically disrupted using a micropestle (Sigma Aldrich) in 250 µL of 2% sodium deoxycholate (SDC), 50 mM Tris-HCl, 10 mM tris(2-carboxyethyl)phosphine (TCEP), pH 8.5, followed by sequential incubation on an Eppendorf ThermoMixer C (purchased from VWR International, Mississauga, ON, CA) for 20 min at 99 °C (700 rpm) and for 2 h at 80 °C (900 rpm). The samples were cooled down on ice for 5 min, followed by RAC-BCA protein determination using 9 µL aliquots. Free cysteines were alkylated using 30 mM iodoacetamide (IAA) for 30 min at room temperature (RT), protected from light, followed by a quench with 10 mM dithiothreitol (DTT) for 15 min at RT. Tryptic digestion was performed by *FASP* [37,38] with slight modifications.[55] Briefly, lysate corresponding to 20 µg of total protein was diluted to 450 µL with freshly prepared 8 M Urea,100 mM Tris, pH 8.5 [56] and loaded onto a 30 kDa Microcon filter. The sample was centrifuged for 25 min at 13,500x g and the eluate was discarded, followed by three washes using 100 µL of the same buffer and three washes with 50 mM ammonium bicarbonate (AmBic). Finally, 100 µL of digestion buffer comprising 1:20 (w/w) trypsin:protein in 0.2 M guanidine-hydrochloride (GuHCl), 50 mM AmBic, 2 mM CaCl2 were added and the sample was incubated at 37 ◦C for 14 h. The tryptic peptides were recovered by centrifugation for 15 min at 13,500x g, followed by two additional washes using 50 µL of 50 mM AmBic and 50 µL of ultrapure water. The collected peptide sample was dried under vacuum and reconstituted in 0.1% formic acid (FA) for nano-LC-MS/MS.

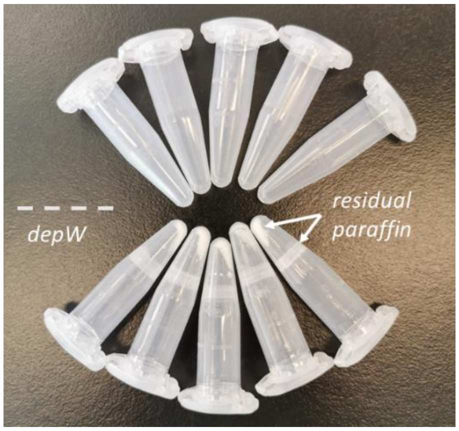

Figure 7. Representative tubes after deparaffinization. The molten paraffin in the *depW* approach forms a layer on the surface of the hot water and residual paraffin 'flakes' precipitate to the bottom of the tube during centrifugation.

The deparaffinized tissue floats in the water and can be easily transferred into a new tube for further sample preparation.

4.3.2 Optimization of Tissue Homogenization

250 µL of extraction buffer (2% SDC, 50 mM Tris-HCl, 10 mM TCEP, pH 8.5) were added to each H_2O-deparaffinized (*depW*) tissue core (*n* = 10) for mechanical cell disruption and homogenization using either a disposable micropestle (*homMP*; Sigma Aldrich, #BAF199230001) or a BioMasher III (*homBM*; Optima Inc., #320302). After homogenization, the samples were digested using FASP as described above, dried under vacuum, and reconstituted in 0.1% FA for nano-LC-MS/MS.

4.3.3 Optimization of Protein Extraction

Two buffers were compared: 4% SDS (*w*/*v*), 150 mM NaCl, 50 mM Tris-HCl, pH 8.5 (*exSDS*)[3,15,39] and 2% SDC, 50 mM Tris-HCl, 10 mM TCEP, pH 8.5 (*exSDC*). Then, 250 µL of either *exSDS* or *exSDC* buffer were added to each deparaffinized core. Samples were homogenized and digested as described for *homMP*. Samples were dried under vacuum and reconstituted in 0.1% FA for nano-LC-MS/MS.

4.3.4 Optimization of Tryptic Digestion

Three strategies for proteolytic digestion were compared: *FASP* [1,2] using a 30 kDa MW cut-off filter,[2,57] protein aggregation capture (*PAC*),[39] and suspension trapping using S-Trap micro-cartridges (*STRAP*).[40, 51] The cores were deparaffinized using H_2O (*depW*) and homogenized using a disposable micropestle (*homMP*) in 2% SDC, 50 mM Tris-HCl, 10 mM TCEP, pH 8.5 (*exSDC*).

(a) *FASP* was performed as described above.

(b) *PAC* was performed using amine microparticles (MagReSyn) based on Batth et al.[39] For 20 µg of protein lysate, 12 µL of microparticle stock solution (20 µg/mL) were equilibrated with 100 µL of 70% ACN, briefly vortexed and placed on a magnetic rack to remove the supernatant. This step was repeated another two times. Next, the protein extracts were added to the beads and the sample was adjusted to a final concentration of 70% ACN, thoroughly vortexed and incubated for 10 min at RT without shaking. The following washing steps were performed on a magnetic rack without disturbing the protein/bead aggregate. The supernatants were discarded, and the beads were washed on the magnetic rack with 1 mL of 95% ACN for 10 s, followed by a wash with 1 mL of 70% ACN without disturbing the protein/bead aggregate. The tubes were removed from the magnetic rack, 100 µL of digestion buffer (1:20 (*w*/*w*) trypsin:protein in 0.2 M GuHCl, 50 mM AmBic, 2 mM $CaCl_2$) were added and the samples were incubated at 37 °C for 3 h. After acidification with trifluoroacetic acid (TFA) to a final concentration of 2%, the tubes were placed on the magnetic rack for 1 min, followed by removal of the supernatant. To remove residual beads, the samples were centrifuged at 20,000x *g* for 10 min. The supernatants were dried under vacuum and reconstituted in 0.1% FA for nano-LC-MS/MS.

(c) *STRAP* digestion was performed according to the manufacturer's instructions.[51] Lysate corresponding to 20 µg of total protein was acidified to a final concentration of 1.2% phosphoric acid. SDS was added to a final concentration of 2% followed by a 7-fold dilution with STRAP binding buffer (90% methanol, 100 mM Tris-HCl, pH 7.1). The sample was loaded onto the STRAP and centrifuged at 4000x *g* for 1 min, followed by three washes with 150 µL binding buffer, with the spin-column being rotated by 180° between centrifugation steps. Then, 200 µL of STRAP digestion buffer, comprised of 1:10 (*w*/*w*) trypsin:protein in 0.2 M GuHCl, 50 mM AmBic, 2 mM $CaCl_2$ were added to the STRAP, which was briefly spun on a benchtop centrifuge to assure saturation of the column material with the digestion buffer. The flow-through was loaded again on top of the column. The sample was incubated at 47 °C for 3 h. Peptides were eluted by sequential elution (1000x g, 1 min) using 40 µL of 50 mM AmBic, 40 µL of 0.1% FA, and 35 µL of 50% ACN, 0.1% FA. The collected peptide sample was dried under vacuum and reconstituted in 0.1% FA for nano-LC-MS/MS.

4.4 Data Analysis

All samples were analyzed by data dependent acquisition (DDA) using an Easy-nLC 1200 (Thermo Fisher Scientific, Waltham, MA, USA) coupled to a Q Exactive Plus (Thermo Fisher Scientific) mass spectrometer that was operated with a Nanospray Flex ion source (Thermo Fisher Scientific). To minimize systematic errors, all samples from one experimental set (e.g., comparison of *FASP/PAC/STRAP*) were injected in a randomized order. Then, 1 µg of digested protein were pre-concentrated on an AcclaimPepMap 100 C18 pre-column (Thermo Fisher Scientific, 3 µm particle size, 75 µm inner diameter × 2 cm length) and separated on an AcclaimPepMap 100 C18 main column (Thermo Fisher Scientific, 2 µm particle size, 75 µm inner diameter x 25 cm length) using a 50 min binary gradient (A: 0.1% FA; B: 84% ACN in 0.1% FA) at a flow rate of 300 nL/min. B was increased from 3-17% until min 30 and from 17-40% until min 20. Full MS scans were acquired from m/z 350-1500 at a resolution of 70,000 with an automatic gain control (AGC) target value of 1×10^6 and a maximum injection time of 50 ms. The 15 most intense precursor ions (charge states +2, +3, +4) were isolated with a window of *m*/*z* 1.2 and fragmented using a normalized collision energy of 28; the dynamic exclusion was set to 40 s. MS/MS spectra were acquired at a resolution of 17,500, using an AGC target value of 2×10^4 and a maximum injection time of 64 ms.

MS raw data were processed using Proteome Discoverer 2.4 (PD, Thermo Fisher Scientific). Database searches were performed using SequestHT and a human Swissprot database (January 2019; 20,414 target entries). Label-free quantitation (LFQ) was performed using the Minora feature detector node, Percolator was used to calculate posterior error probabilities. Database searches were performed using trypsin as enzyme with a maximum of 2 missed cleavages. Carbamidomethylation of cysteine (+57.021 Da) was set as fixed modification and oxidation of methionine (+15.995 Da), and lysine methylation (+14.016 Da, +28.031 Da, and +42.047 Da) as variable modifications.[21] Mass tolerances were set to 5 ppm for precursor- and 0.02 Da for product-ions. The data were filtered to a false discovery rate (FDR) < 1% on the peptide and protein levels. Only proteins that were (i) identified with at least two unique peptides and (ii) quantified in at least 3 out of 5 replicates of at least one of the methods to be compared, were considered for the quantitative comparison. Protein LFQ data obtained from Proteome Discoverer were normalized based on summed protein intensities to correct for differences in sample loading. For proteins passing the abovementioned criteria, missing protein intensity values were imputed using 1.5x the minimum observed intensity for this particular sample. The obtained normalized abundances were used for unpaired *t-tests* (two tailed, 95% confidence) and Pearson correlation analyses. Differential expression analysis was performed on log2-transformed normalized abundance data with multiple hypothesis testing using a false discovery approach by Benjamini–Krieger false discovery rate (FDR 1%). Proteins having *q*-values of < 0.01 and absolute log2 fold-changes >1 were considered as differential between tested workflows. Statistical analysis was performed using GraphPad Prism 9 (San Diego, CA, USA).

MS data files are publicly available through the ProteomeXchange Consortium via the PRIDE partner repository[58] with the dataset identifier PXD031946.

5 Conclusions

Based on our results, we recommend the following protocol for quantitative proteomics of small FFPE core samples: (i) water-based deparaffinization (*depW*) followed by homogenization using a micropestle (*homMP*) in an SDC-based extraction buffer containing reducing agents, e.g., DTT or TCEP (*exSDC*). In particular, for protein amounts of <10 µg, we recommend PAC-based digestion, while amounts of up to 100 µg work well with *FASP*. Importantly, based on our results, the tendency for better recoveries of subsets of proteins with certain properties might further influence the choice of sample preparation—for example, small and more acidic proteins showed better recoveries using PAC (see Supplemental Table S4).

This ’green’ FFPE proteomics workflow requires less starting material than used in comparable studies, i.e., <1 mg dry-weight of a FFPE core with wax, and enables robust proteomic profiling of FFPE tumor tissues (average RSD <20%) for clinical research. The micro-volumes of FFPE tissue used successfully in this study show that our workflow is well-suited for small tissue samples preserved in FFPE blocks such as core needle biopsies, and also small tissue cores such as those used to build TMAs. Diseases where tissue areas of interest are very limited, e.g., in pre-invasive cancer, are currently only amenable to histological assessment for diagnostic purposes. Our protocol enables clinical research and molecular characterization of such diseases, as only 1 mm in diameter cores are required. Furthermore, our workflow does not require any special equipment, is suitable for any standard hospital clinical laboratory, and can be automated, thereby facilitating high-throughput analysis for clinical research.

We have already used this protocol successfully for quantitative proteomics of FFPE cores and sections in breast cancer patient-derived xenografts for precise quantitation of PTEN,[59] AKT1/2, and PIK3CA (publication in preparation), as well as in clinical samples of non-small cell lung cancer (NSCLC) for quantitative assessment of proteins in the PDL1-axis (publication in preparation). Tissue areas of interest as small as ~1 mm^3 and, based on the total amounts of extracted protein, even considerably smaller volumes of FFPE tissues should be compatible with this workflow.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms23084443/s1.

Author Contributions: The manuscript was written by G.M. and R.P.Z. Experiments and data analysis were conducted by G.M. under supervision of R.P.Z., C.H.B. and G.B. Q.G., C.G., and S.E.J.P. generated the tumor specimens. Resources were provided by V.L., A.A.-M., N.B., M.B., A.S., W.H.M.J. and S.V.d.R. The manuscript was written with contributions from all authors. All authors have read and agreed to the published version of the manuscript.

Funding: We are grateful to Genome Canada for financial support through the Genomics Technology Platform (GTP: 264PRO) and the Quebec Cancer Consortium, Ministère de l'Économie et de l'Innovation, Quebec. We are also grateful for financial support from Genome Canada and Genome Quebec through the Genomics Applications Partnership Program (GAPP: PD-L1), and from the Terry Fox Research Institute. CHB is also grateful for support from the Segal McGill Chair in Molecular Oncology at McGill University (Montreal, Quebec, Canada), and for support from the Warren Y. Soper Charitable Trust and the Alvin Segal Family Foundation to the Jewish General Hospital (Montreal, Quebec, Canada). This research was also funded by a grant from the Canadian Institutes for Health Research (PJT-156269 to WHM and SVR). SEJP was supported by a fellowship from the Fonds de recherche du Québec–Santé. VL was supported by a fellowship from the Fonds de recherche du Québec–Santé (grant number: 288230). MB is grateful to FRQS —Reseau de Recherche Cancer for financial support, and the Weekend to End Breast Cancer Foundation for biobank support.

Institutional Review Board Statement: All animal experiments were conducted according to the regulations established by the Canadian Council of Animal Care, under protocols approved by the McGill University Animal Care and Use Committee.

Informed Consent Statement: Not applicable.

Data Availability Statement: MS data files are publicly available through the ProteomeXchange Consortium via the PRIDE partner repository[52] with the dataset identifier PXD031946**.**

Conflicts of Interest: C.H.B. is the Chief Scientific Officer of MRM Proteomics, Inc. R.P.Z. is the Chief Executive Officer of MRM Proteomics Inc. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Gaffney, E.F.; Riegman, P.H.; Grizzle, W.E.; Watson, P.H. Factors that drive the increasing use of FFPE tissue in basic and translational cancer research. Biotech. Histochem. 2018, 93, 373–386.

2. Foll, M.C.; Fahrner, M.; Oria, V.O.; Kuhs, M.; Biniossek, M.L.; Werner, M.; Bronsert, P.; Schilling, O. Reproducible proteomics sample preparation for single FFPE tissue slices using acid-labile surfactant and direct trypsinization. Clin. Proteom. 2018, 15, 11.
3. Maes, E.; Broeckx, V.; Mertens, I.; Sagaert, X.; Prenen, H.; Landuyt, B.; Schoofs, L. Analysis of the formalin-fixed paraffin-embedded tissue proteome: Pitfalls, challenges, and future prospectives. Amino Acids 2013, 45, 205–218.
4. Maes, E.; Valkenborg, D.; Mertens, I.; Broeckx, V.; Baggerman, G.; Sagaert, X.; Landuyt, B.; Prenen, H.; Schoofs, L. Proteomic analysis of formalin-fixed paraffin-embedded colorectal cancer tissue using tandem mass tag protein labeling. Mol. Biosyst. 2013, 9, 2686–2695.
5. Ralton, L.D.; Murray, G.I. The use of formalin fixed wax embedded tissue for proteomic analysis. J. Clin. Pathol. 2011, 64, 297–302.
6. Seyhan, A.A.; Carini, C. Are innovation and new technologies in precision medicine paving a new era in patients centric care? J. Transl. Med. 2019, 17, 114.
7. Sobsey, C.A.; Ibrahim, S.; Richard, V.R.; Gaspar, V.; Mitsa, G.; Lacasse, V.; Zahedi, R.P.; Batist, G.; Borchers, C.H. Targeted and Untargeted Proteomics Approaches in Biomarker Development. Proteomics 2019, 20, e1900029.
8. Alvarez-Chaver, P.; De Chiara, L.; Martinez-Zorzano, V.S. Proteomic Profiling for Colorectal Cancer Biomarker Discovery. Methods Mol. Biol. 2018, 1765, 241–269.
9. Choi, C.H.; Chung, J.Y.; Kang, J.H.; Paik, E.S.; Lee, Y.Y.; Park, W.; Byeon, S.J.; Chung, E.J.; Kim, B.G.; Hewitt, S.M.; et al. Chemoradiotherapy response prediction model by proteomic expressional profiling in patients with locally advanced cervical cancer. Gynecol. Oncol. 2020, 157, 437–443.
10. Toomey, S.; Carr, A.; Mezynski, M.J.; Elamin, Y.; Rafee, S.; Cremona, M.; Morgan, C.; Madden, S.; Abdul-Jalil, K.I.; Gately, K.; et al. Identification and clinical impact of potentially actionable somatic oncogenic mutations in solid tumor samples. J. Transl. Med. 020, 18, 99.
11. Faoláin, E.Ó.; Hunter, M.B.; Byrne, J.M.; Kelehan, P.; Lambkin, H.A.; Byrne, H.J.; Lyng, F.M. Raman spectroscopic evaluation of efficacy of current paraffin wax section dewaxing agents. J. Histochem. Cytochem. 2005, 53, 121–129.
12. Kandyala, R.; Raghavendra, S.P.C.; Rajasekharan, S.T. Xylene: An overview of its health hazards and preventive measures. J. Oral Maxillofac. Pathol. 2010, 14, 1–5.
13. European Chemical Agency. Xylene (Last Updated: 16 April 2020). Available online: https://echa.europa.eu/brief-profile/-/briefprofile/100.014.124 (accessed on 11 August 2021).

14. Mansour, A.G.; Khalil, P.A.; Bejjani, N.; Chatila, R.; Dagher-Hamalian, C.; Faour, W.H. An optimized xylene-free protein extraction method adapted to formalin-fixed paraffin embedded tissue sections for western blot analysis. Histol. Histopathol. 2017, 32, 307–313.

15. Magdeldin, S.; Yamamoto, T. Toward deciphering proteomes of formalin-fixed paraffin-embedded (FFPE) tissues. Proteomics 2012, 12, 1045–1058.

16. Anastas, P.T.; Warner, J.C. Green Chemistry: Theory and Practice; Oxford University Press: Oxford, UK, 1998.

17. Guo, Q.; Li, V.Z.; Nichol, J.N.; Huang, F.; Yang, W.; Preston, S.E.J.; Talat, Z.; Lefrere, H.; Yu, H.; Zhang, G.; et al. MNK1/NODAL Signaling Promotes Invasive Progression of Breast Ductal Carcinoma In Situ. Cancer Res. 2019, 79, 1646–1657.

18. Kalantari, N.; Bayani, M.; Ghaffari, T. Deparaffinization of formalin-fixed paraffin-embedded tissue blocks using hot water instead of xylene. Anal. Biochem. 2016, 507, 71–73.

19. The Human Protein Atlas. Available online: https://www.proteinatlas.org/ (accessed on 11 August 2021).

20. Maertens, A.; Tran, V.P.; Maertens, M.; Kleensang, A.; Luechtefeld, T.H.; Hartung, T.; Paller, C.J. Functionally Enigmatic Genes in Cancer: Using TCGA Data to Map the Limitations of Annotations. Sci. Rep. 2020, 10, 4106

21. Sekine, S.; Wang, C.; Sideris, D.P.; Bunker, E.; Zhang, Z.; Youle, R.J. Reciprocal roles of Tom7 and OMA1 during mitochondrial import and activation of PINK1. Mol. Cell 2019, 73, 1028–1043.e5.

22. Alnouti, Y. Bile Acid Sulfation: A Pathway of Bile Acid Elimination and Detoxification. Toxicol. Sci. 2009, 108, 225–246.

23. Zhong, K.; Chen, K.; Han, L.; Li, B. MicroRNA-30b/c inhibits non-small cell lung cancer cell proliferation by targeting Rab18. BMC Cancer 2014, 14, 703.

24. Meng, X.; Liu, J.; Shen, Z. Genomic structure of the human BCCIP gene and its expression in cancer. Gene 2003, 302, 139–146.

25. Fujikawa, T.; Miyata, S.-I.; Iwanami, T. Convenient detection of the citrus greening (huanglongbing) bacterium ‘Candidatus Liberibacter asiaticus’ by direct PCR from the midrib extract. PLoS ONE 2013, 8, e57011.

26. Alqaydi, M.; Roy, R. Quantitative and qualitative study of STR DNA from ethanol and formalin fixed tissues. Forensic Sci. Int.2016, 262, 18–29.

27. Yamamoto, T.; Nakashima, K.; Maruta, Y.; Kiriyama, T.; Sasaki, M.; Sugiyama, S.; Suzuki, K.; Fujisaki, H.; Sasaki, J.; Kaku-Ushiki, Y. Improved RNA extraction method using the BioMasher and BioMasher power-plus. J. Vet. Med. Sci. 2012, 74, 1561–1567.

28. Yamamoto, T.; Ushiki, Y.; Hara, S.; Hall, W.W.; Tsukagoshi-Nagai, H.; Yokoyama, T.; Tagawa, Y.; Sata, T.; Yamakawa, Y.; Kinoshita, N. An advantageous method utilizing new homogenizing device Bio-Masher and a sensitive ELISA to detect bovine spongiform encephalopathy accurately in brain tissue. J. Virol. Methods 2008, 149, 316–325.
29. Zhang, Y.; Muller, M.; Xu, B.; Yoshida, Y.; Horlacher, O.; Nikitin, F.; Garessus, S.; Magdeldin, S.; Kinoshita, N.; Fujinaka, H.; et al. Unrestricted modification search reveals lysine methylation as major modification induced by tissue formalin fixation and paraffin embedding. Proteomics 2015, 15, 2568–2579.
30. Sprung, R.W.; Brock, J.W.C.; Tanksley, J.P.; Li, M.; Washington, M.K.; Slebos, R.J.C.; Liebler, D.C. Equivalence of Protein Inventories Obtained from Formalin-fixed Paraffin-embedded and Frozen Tissue in Multidimensional Liquid Chromatography-Tandem Mass Spectrometry Shotgun Proteomic Analysis. Mol. Cell. Proteom. 2009, 8, 1988–1998.
31. Coscia, F.; Doll, S.; Bech, J.M.; Schweizer, L.; Mund, A.; Lengyel, E.; Lindebjerg, J.; Madsen, G.I.; Moreira, J.M.; Mann, M. A streamlined mass spectrometry–based proteomics workflow for large-scale FFPE tissue analysis. J. Pathol. 2020, 251, 100–112.
32. Metz, B.; Kersten, G.F.A.; Baart, G.J.E.; de Jong, A.; Meiring, H.; ten Hove, J.; van Steenbergen, M.J.; Hennink, W.E.; Crommelin, D.J.A.; Jiskoot, W. Identification of Formaldehyde-Induced Modifications in Proteins: Reactions with Insulin. Bioconjugate Chem. 2006, 17, 815–822.
33. Shi, S.R.; Liu, C.; Balgley, B.M.; Lee, C.; Taylor, C.R. Protein extraction from formalin-fixed, paraffin-embedded tissue sections: Quality evaluation by mass spectrometry. J. Histochem. Cytochem. 2006, 54, 739–743.
34. Proc, J.L.; Kuzyk, M.A.; Hardie, D.B.; Yang, J.; Smith, D.S.; Jackson, A.M.; Parker, C.E.; Borchers, C.H. A quantitative study of the effects of chaotropic agents, surfactants, and solvents on the digestion efficiency of human plasma proteins by trypsin. J. Proteome Res. 2010, 9, 5422–5437.
35. Lin, Y.; Lin, H.; Liu, Z.; Wang, K.; Yan, Y. Improvement of a sample preparation method assisted by sodium deoxycholate for mass-spectrometry-based shotgun membrane proteomics. J. Sep. Sci. 2014, 37, 3321–3329.
36. Erde, J.; Loo, R.R.; Loo, J.A. Enhanced FASP (eFASP) to increase proteome coverage and sample recovery for quantitative proteomic experiments. J. Proteome Res. 2014, 13, 1885–1895.
37. Wisniewski, J.R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. Nat. Methods 2009, 6, 359–362.
38. Manza, L.L.; Stamer, S.L.; Ham, A.J.; Codreanu, S.G.; Liebler, D.C. Sample preparation and digestion for proteomic analyses using spin filters. Proteomics 2005, 5, 1742–1745.

39. Batth, T.S.; Tollenaere, M.A.X.; Rüther, P.L.; Gonzalez-Franquesa, A.; Prabhakar, B.S.; Bekker-Jensen, S.H.; Deshmukh, A.S.; Olsen, J.V. Protein aggregation capture on microparticles enables multi-purpose proteomics sample preparation. Mol. Cell. Proteom. 2019, 18, 1027–1035.

40. Wilson, J.; Mayyappan, V.; Narducci, D.N.; Neely, B.; Laugharn, J.; Pappin, D. Universal Sample Processing of Multiple Sample Types for Reproducible Proteomic Sample Preparation. HUPO. 2018. Available online: https://cdn.shopify.com/s/files/1/0271/1964/8832/files/HUPO-ProtiFi-Covaris-poster-final.pdf (accessed on 11 August 2021).

41. Müller, T.; Kalxdorf, M.; Longuespée, R.; Kazdal, D.N.; Stenzinger, A.; Krijgsveld, J. Automated sample preparation with SP3 for low-input clinical proteomics. Mol. Syst. Biol. 2020, 16, e9111.

42. Buczak, K.; Ori, A.; Kirkpatrick, J.M.; Holzer, K.; Dauch, D.; Roessler, S.; Endris, V.; Lasitschka, F.; Parca, L.; Schmidt, A.; et al. Spatial tissue proteomics quantifies inter- and intra-tumor heterogeneity in hepatocellular carcinoma. Mol. Cell. Proteom. 2018, 17, 810–825.

43. Schweizer, L.; Coscia, F.; Müller, J.; Doll, S.; Wierer, M.; Mann, M. AFA-sonication Followed by Modified Protein Aggregation Capture (APAC) Enables Direct, Reproducible and Non-toxic Sample Preparation of FFPE Tissue for Mass Spectrometry based Proteomics. Covaris Appl. Note-M020141. Available online: https://d24ci5y4j5ezt1.cloudfront.net/wp/wp-content/uploads/2020/06/M020141.pdf (accessed on 11 August 2021).

44. Hughes, C.S.; Moggridge, S.; Müller, T.; Sorensen, P.H.; Morin, G.B.; Krijgsveld, J. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. Nat. Protoc. 2019, 14, 68–85.

45. Stoychev, S.; Govender, I.; Naicker, P.; Gerber, I.; Jordaan, J.; Pauw, M.; Tabb, D.; Arribas Diez, I.; Norregaard Jensen, O.; Baath, T.; et al. Development of a fully automated high throughput magnetic workflow for phosphoproteome profiling. HUPO. 2019. Available online: https://resynbio.com/wp-content/uploads/2020/01/HUPO_2019_Poster_SS_Phos_coupled_clean-up.pdf (accessed on 11 August 2021).

46. Tape, C.J.; Worboys, J.D.; Sinclair, J.; Gourlay, R.; Vogt, J.; McMahon, K.M.; Trost, M.; Lauffenburger, D.A.; Lamont, D.J.; Jørgensen, C. Reproducible automated phosphopeptide enrichment using magnetic TiO2 and Ti-IMAC. Anal. Chem. 2014, 86, 10296–10302.

47. Martínez-Val, A.; Bekker-Jensen, D.B.; Steigerwald, S.; Stoychev, S.; Gerber, I.; Jordaan, J.; Bache, N.; Olsen, J.V. Fast and reproducible phosphoproteomics using MagReSyn Amine and Ti-IMAC HP magnetic beads and the Evosep One. Tech. Note. Available online: https://www.evosep.com/wp-content/uploads/2020/03/Phospho-app-note-Evosep-Resyn-A5layout-v5-lores.pdf (accessed on 11 August 2021).

48. Leutert, M.; Rodriguez-Mias, R.A.; Fukuda, N.K.; Villén, J. R2-P2 rapid-robotic phosphoproteomics enables multidimensional cell signaling studies. Mol. Syst. Biol. 2019, 15, e9021.
49. Nolte, H.; MacVicar, T.D.; Tellkamp, F.; Krüger, M. Instant Clue: A Software Suite for Interactive Data Visualization and Analysis. Sci. Rep. 2018, 8, 12648.
50. The Human Proteome Atlas. Available online: https://www.proteinatlas.org (accessed on 23 February 2022).
51. Marchione, D.M.; Ilieva, I.; Devins, K.; Sharpe, D.; Pappin, D.J.; Garcia, B.A.; Wilson, J.P.; Wojcik, J.B. HYPERsol: High-Quality Data from Archival FFPE Tissue for Clinical Proteomics. J. Proteome Res. 2020, 19, 973–983.
52. Sielaff, M.; Kuharev, J.; Bohn, T.; Hahlbrock, J.; Bopp, T.; Tenzer, S.; Distler, U. Evaluation of FASP, SP3, and iST Protocols for Proteomic Sample Preparation in the Low Microgram Range. J. Proteome Res. 2017, 16, 4060–4072.
53. Wisniewski, J.R. Filter Aided Sample Preparation—A tutorial. Anal. Chim. Acta 2019, 1090, 23–30.
54. Boellner, S.; Becker, K.F. Reverse Phase Protein Arrays-Quantitative Assessment of Multiple Biomarkers in Biopsies for Clinical Use. Microarrays 2015, 4, 98–114.
55. Shema, G.; Nguyen, M.T.N.; Solari, F.A.; Loroch, S.; Venne, A.S.; Kollipara, L.; Sickmann, A.; Verhelst, S.H.L.; Zahedi, R.P. Simple, scalable, and ultrasensitive tip-based identification of protease substrates. Mol. Cell. Proteom. 2018, 17, 826–834.
56. Kollipara, L.; Zahedi, R.P. Protein carbamylation: In vivo modification or in vitro artefact? Proteomics 2013, 13, 941–944.
57. Tanca, A.; Abbondio, M.; Pisanu, S.; Pagnozzi, D.; Uzzau, S.; Addis, M.F. Critical comparison of sample preparation strategies for shotgun proteomic analysis of formalin-fixed, paraffin-embedded samples: Insights from liver tissue. Clin. Proteom. 2014, 11, 28.
58. Vizcaino, J.A.; Deutsch, E.W.; Wang, R.; Csordas, A.; Reisinger, F.; Rios, D.; Dianes, J.A.; Sun, Z.; Farrah, T.; Bandeira, N.; et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. 2014, 32, 223–226.
59. Ibrahim, S.; Lan, C.; Chabot, C.; Mitsa, G.; Buchanan, M.; Aguilar-Mahecha, A.; Elchebly, M.; Poetz, O.; Spatz, A.; Basik, M.; et al. Precise quantitation of PTEN by immuno-MRM: A tool to resolve the breast cancer biomarker controversy. Anal. Chem. 2021, 93, 10816–10824.

Preface Chapter 3

Proteomics has the potential to address the chiasm between biomarker discovery and successful verification of biomarkers (see Chapter 1). Here we demonstrate the invaluable importance of proteomics as part of integrative translational research by applying our improved FFPE-proteomics procedure on a study to unlock the proteomic secrets of ductal carcinoma in situ - the most prevalent non-invasive breast lesion. Using quantitative proteomics, both with discovery and targeted approaches, we not only confirm genomic biomarkers suggested by independent transcriptomic studies but identify more than 300 additional biomarkers that could inform much needed strategies for disease management of pre-malignant lesions. With this work, we further provide an assay for clinical validation and utility testing, that meets all regulatory requirements and can be readily incorporated into clinical practice guidelines.

Author Contributions: The following manuscript was conceptualized by the candidate and her supervisors, Gerald Batist and Christoph H. Borchers. Experiments, methodology and formal data analysis was performed by the candidate. The manuscript was written by the candidate, and reviewed/edited by Adriana Aguilar-Mahecha, Mark Basik, Gerald Batist and Christoph H. Borchers. Livia Florianova curated and selected clinical cases and specimens included in this study. Josiane Lafleur retrieved the specimens and relevant clinical data from the patient files.

Chapter 3: Clinical proteomics reveals vulnerabilities in non-invasive breast ductal carcinoma and drives personalized treatment strategies.

Georgia Mitsa[1, 2], Livia Florianova[3], Josiane Lafleur[4], Adriana Aguilar-Mahecha[4], Rene P. Zahedi[5, 6], Sonia V del Rincon[1,4], Mark Basik[7], Christoph H Borchers[1, 2, 3, 4, 8]*, Gerald Batist[1, 4, 8, 9]*

1. Division of Experimental Medicine, McGill University, Montreal, QC
2. Segal Cancer Proteomics Centre, Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC
3. Department of Pathology, Segal Cancer Centre, Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, Montreal, QC
4. Segal Cancer Centre, Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC
5. Manitoba Centre for Proteomics and Systems Biology, Winnipeg, MB
6. Department of Internal Medicine, University of Manitoba, Winnipeg, MB
7. Department of Oncology and Surgery, Lady Davis Institute for Medical Research, Jewish General Hospital, McGill University, Montreal, QC
8. Department of Oncology, McGill University, Montreal, QC
9. Exactis Innovation, Montreal, QC

Correspondence

Gerald Batist, MD
Lady Davis Institute for Medical Research
Jewish General Hospital, McGill University
Montréal, Quebec, H3T 1E2, Canada
gerald.batist@mcgill.ca
Tel.: +1 514-340-8222 ext.5418

Christoph H Borchers, Ph.D.
Segal Cancer Proteomics Centre, Lady Davis Institute for Medical Research
Jewish General Hospital, McGill University
Montréal, Quebec, H3T 1E2, Canada
christoph.borchers@mcgill.ca
Tel.: +1 514-340-8222 ext.7886

Manuscript under consideration for publication in Nature Communication: 14. June 2023

Available as pre-print at bioRxiv: 2023.2007.2011.548580. doi: 10.1101/2023.07.11.548580

Abstract

Ductal carcinoma in situ (DCIS) is the most common type (80%) of non-invasive breast lesions. The lack of validated prognostic markers, limited patient numbers and variable tissue quality significantly impact diagnosis, risk stratification, patient enrolment, and results of clinical studies. We performed label-free quantitative proteomics on 50 clinical formalin-fixed, paraffin embedded biopsies, validating 22 putative biomarkers from independent genetic studies. Our comprehensive proteomic phenotyping reveals more than 380 differentially expressed proteins and metabolic vulnerabilities, that can inform new therapeutic strategies for DCIS and IDC. Due to the readily druggable nature of proteins and metabolites, this study is of high interest for clinical research and pharmaceutical industry. To further evaluate our findings, and to promote the clinical translation of our study, we developed a highly multiplexed targeted proteomics assay for 90 proteins associated with cancer metabolism, RNA regulation and signature cancer pathways, such as Pi3K/AKT/mTOR and EGFR/RAS/RAF.

1 Introduction

Ductal carcinoma in situ (DCIS) is a pre-invasive (stage 0) neoplastic lesion that is associated with a ~10-fold elevated risk of developing invasive breast cancer, e.g., invasive ductal carcinoma (IDC).[1] Due to this increased risk, patients diagnosed with DCIS undergo aggressive treatment with breast conserving surgery or total mastectomy with optional adjuvant therapy, i.e., radiation or endocrine therapy.

Studies, however, show that if left untreated, only 20-50% of DCIS patients will progress to IDC. [2-5] This has led to global concerns regarding overtreatment of DCIS patients, the resulting high economic burden for the healthcare system and, most importantly, a high psychological burden for the patients. Tools and expression signatures to predict invasive progression for better informed clinical decision making are required and many international trials are currently enrolling patients with DCIS for non-surgical management by active surveillance, e.g., LORIS, LORD and LARRIKIN. [6] The COMET trial (NCT02926911) in the US is targeting histologically confirmed low-risk DCIS for a comparison of surgery to monitoring and endocrine therapy.

At present, the diagnosis of DCIS is based on calcifications observed during mammography screenings and histological assessment of tissue biopsies, i.e., formalin-fixed and paraffin embedded (FFPE) needle core biopsies. Five morphological key features, high intra-tumor heterogeneity, poor inter-observer agreement, [7-10] and the lack of validated prognostic markers significantly impact clear diagnosis and risk stratification, as well as patient enrolment and final results of clinical studies.

There is currently no precision oncology treatment available for patients diagnosed with DCIS. Post-operative (adjuvant) therapy is guided by immunohistochemistry (IHC) assays for estrogen and progesterone receptor status (ER and PR), HER2 expression status (by fluorescence in situ hybridization, FISH), as well as BRCA1/2 mutation status. Clinical multigene assays, such as Oncotype DX/DCIS, MammaPrint or PreludeDx DCIS, are sometimes used to clinically predict recurrence risks of patients but are not standard and only guide the use of adjuvant therapy.

Generally, DCIS studies are limited by patient number and tissue quality. Recent genomic landscaping studies on individual DCIS lesions identified putative biomarkers associated with progression towards IDC and give insights into the underlying cancer biology. Multi-omic profiling of DCIS, however, is still challenging since DCIS and IDC lesions are mostly studied in FFPE-preserved samples; pure DCIS lesions can be very small in size as they are usually from minimally invasive needle core biopsies, and access to pure IDC lesions is limited, as most surgically removed IDC lesions also present in-situ components and may follow effective neoadjuvant therapy.

The current study makes use of our recently published FFPE-proteomics method that facilitates proteomic profiling on FFPE-preserved tissue cores. [11] In a cohort of carefully curated patients treated with DCIS and IDC at the Segal Cancer Centre of the Jewish General Hospital (JGH) in Montreal (n=51) we investigate changes in the protein expression of 29 pure DCIS lesions, 18 pure IDC lesions, 13 mixed-type lesions (IDC with in-situ components), and 9 cases where pure DCIS and pure IDC is present in different lesions in the same patient, either synchronously or metachronously (see Fig. 1). Data from recently published independent gene expression studies investigating the progression from DCIS to IDC were used to complement the label-free protein expression data. Since FFPE preservation eliminates up to 85% of metabolites, [12-16] we used Quantitative Systems Metabolism (QSM™) technology from Doppelganger Biosystem GmbH, Germany, an AI-driven metabolic analysis using proteomics data, [17] for a comprehensive profiling of the central metabolism/energy metabolism. Guided by these results, we developed a highly multiplexed parallel-reaction monitoring (PRM) assay for precise quantitation of 90 proteins, that are associated with cancer metabolism, RNA regulation and major cancer growth-associated pathways, such as PI3K/AKT/mTOR and EGFR/RAS/RAF.

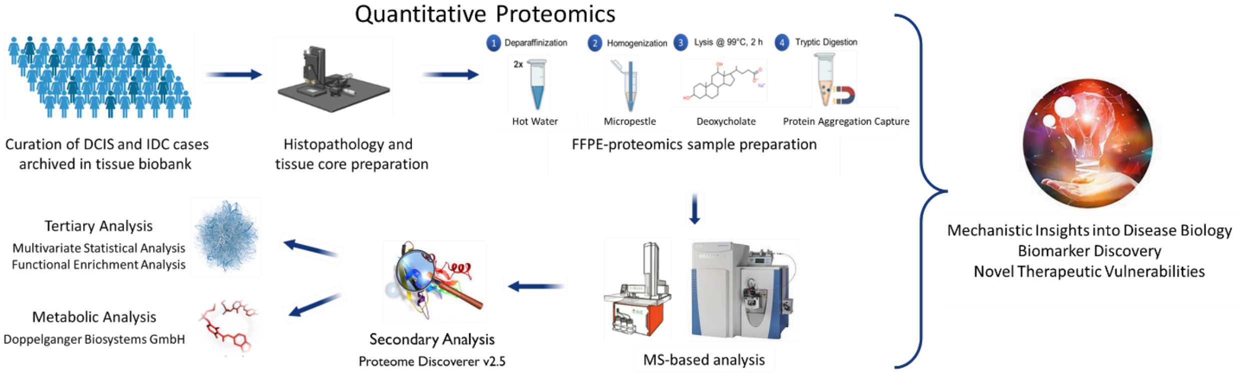
Quantitative Proteomics
Curation of DCIS and IDC cases archived in tissue biobank
Histopathology and tissue core preparation
1 Deparaffinization
2 Homogenization
3 Lysis @ 99°C, 2 h
4 Tryptic Digestion
2x
Hot Water
Micropestle
Deoxycholate
Protein Aggregation Capture
FFPE-proteomics sample preparation
Mechanistic Insights into Disease Biology
Biomarker Discovery
Novel Therapeutic Vulnerabilities
Tertiary Analysis
Multivariate Statistical Analysis
Functional Enrichment Analysis
Metabolic Analysis
Doppelganger Biosystems GmbH
Secondary Analysis
Proteome Discoverer v2.5
MS-based analysis

Figure 5: Experimental Design. Label-free quantitative proteomics was performed in a cohort of carefully curated patients treated with DCIS and IDC (n=50) to investigate changes in the protein expression of 29 pure DCIS lesions, 18 pure IDC lesions, 13 mixed-type lesions (IDC with in-situ components), and 9 cases where pure DCIS and pure IDC is present in different lesions in the same patient, either synchronously or metachronously. The protein extraction of FFPE tissue cores (1 mm diameter, ~0.8 mm³ tissue volume) used an optimized FFPE-proteomics protocol published here. [11] The samples were analyzed on 'plug-and-play' platform built for standardization in clinical proteomics, and the data was processed using state-of-the-art data analysis tools, including machine learning/AI-driven algorithms for improved and higher confidence mechanistic insights.

2 Results

2.1 DCIS and IDC are highly heterogeneous tumor phenotypes but build two distinct clusters in sparse Partial Least Squares Regression for Discrimination Analysis (sPLS-DA).

Several genomic centered studies have reported that both DCIS and IDC tumor phenotypes are highly heterogeneous, [8-10,18,19] hampering clinical diagnosis but also limiting statistical power and robust assay development to complement clinical diagnosis. Using a streamlined FFPE-proteomics workflow [11] with a standard label-free mass spectrometry (MS)-based data analysis, we quantified more than 2800 proteins at a 1% false discovery rate (FDR) on the protein and peptide level. Using less than 1% of the total protein extracted from a single 1-mm FFPE tissue core, we cover 6 orders of magnitude of the DCIS/IDC proteome (Fig. 2a). Notably, the proteome of the two ductal breast cancer disease states seems to be clearly differential from each other, as an sPLS-DA shows two distinct clusters between the study cohorts (Fig. 2b). The sPLS-DA is a statistical method used for extracting and selecting important features from high-dimensional data to discriminate between different groups, while simultaneously considering sparsity to improve interpretability and reduce overfitting. [20] Based on the available clinical data (non-omics data) and small sample size, we are not in the position to infer any underlying patterns or biological relationships leading to this clustering on the protein level. Nevertheless, the top 10 features driving the proteomic variability between DCIS and IDC seem to reflect high transcriptional activity, extracellular matrix remodeling and inflammation processes (Figs. 2c and 2d).

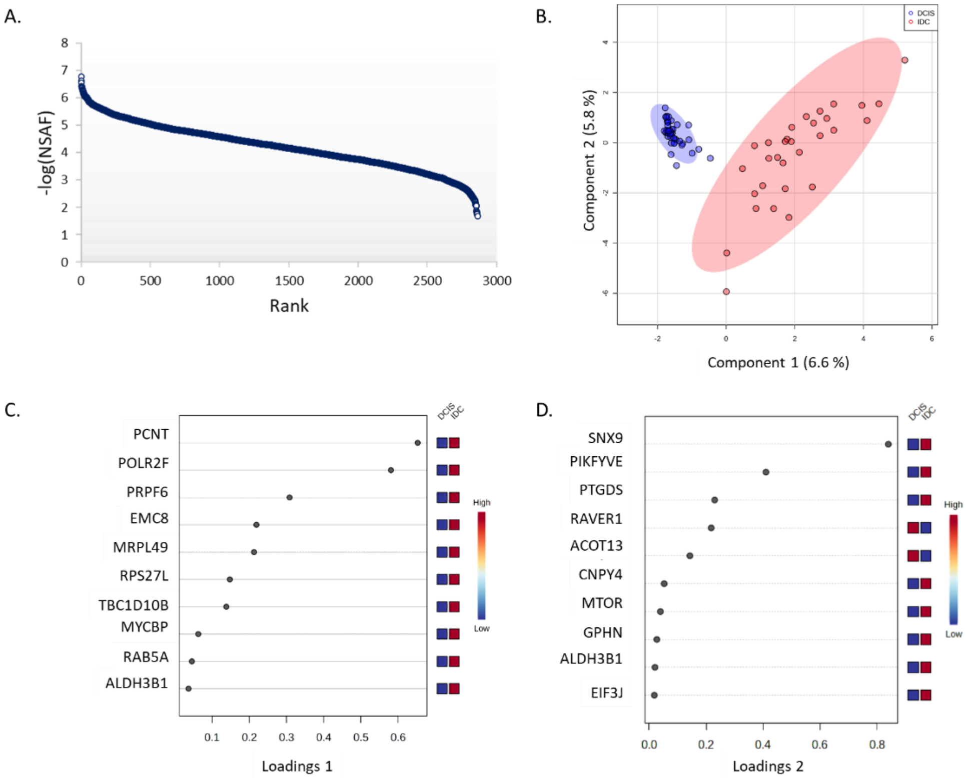

Figure 2: Data Quality and Evaluation of Variability. (A) Dynamic range of ~2,860 proteins quantified in ductal breast cancer, at a 1% false discovery rate. All –log10 values were based on normalized spectral abundance factor (NSAF) values, which were used to normalize the spectral count. High NSAF values represent a high level of expression. 6 orders of magnitude of the DCIS/IDC proteome are covered using ~1% of the total sample and a standard data dependent acquisition (DDA) method without fractionation. (B) Sparse Partial Least Squares Regression for Discrimination Analysis (sPLS-DA) showing good clustering of the two study groups. The oval shape represents 95% confidence intervals. Interquartile and ROUT method identified no outlier samples. (C/D) Loading plots of the sPLS-DA, showing proteins/genes that drive the variability and clustering between DCIS and IDC. The right x-axis shows expression levels of these drivers in the DCIS/IDC samples.

2.2 MS-based proteomics complements and supports independent genomic/transcriptomic studies of DCIS to IDC progression.

Study of progression of DCIS to IDC has mainly used gene expression analysis or IHC/FISH on the protein level. Recent studies have demonstrated significant misalignment between genome and even transcriptome and the ultimate protein levels, and IHC is poorly quantitative. [21-24] Therefore, we sought to confirm these findings using direct measurement of proteins. We compared MS-based label-free proteomics data with 49 differentially expressed genes identified by three recent larger-scale independent genomics/transcriptomics studies [9,25,26] and found 22 overlapping genes (see Table 1). Proteomics data identified gene products of *FOXA1*, *POSTN*, *THBS2*, *CA12*, *FN1* and *ALDH1* as differentially expressed proteins (DEPs, unpaired t-test, $p<0.05$).

The proteomics data shows lower *FOXA1* (Forkhead Box A1) expression in pure DCIS compared to pure IDC ($p<0.0001$), and increased expression in mixed-type DCIS compared to pure DCIS ($p=0.03$) suggesting a protective function of FOXA1. The loss or silencing of *FOXA1* observed in DCIS seems to promote cell migration and invasion. Interestingly, forced expression of *FOXA1* in MCF-7 (IDC cell line) inhibits growth, and controls cell plasticity by repressing the basal-like phenotype. [27,28] Genetic studies associate *FOXA1* with heterochromatin remodeling, particularly affecting hormone receptor transcription, [29] and regulation of the cell cycle with *BRCA1*. [30,31] Evidence of FOXA1 involvement in tumor progression on the (epi-)genetic, transcriptomic, and proteomic level warrants further investigation of FOXA1 as clinical biomarker and its clinical utility for DCIS risk stratification.

POSTN (Periostin), THBS2 (Thrombospondin 2), and FN1 (Fibronectin) mediate cell-cell and cell-matrix interactions. POSTN, a downstream effector of β-catenin, activates PI3K/AKT and ERK pathways. [32] In DCIS, these proteins have lower expression levels compared to IDC ($p<0.03$, $p<0.04$, $p<0.03$, respectively), indicating stromal remodeling in DCIS to IDC progression.

CA12 (Carbonic Anhydrase 12) regulates the tumor microenvironment and metabolic pathways, [33-35] with lower protein levels in pure DCIS compared to pure IDC ($p<0.0001$). Loss of CA12 activity likely creates a more favorable environment for malignant cell growth and progression towards IDC.

High *ALDH1* (Aldehyde Dehydrogenase 1) expression characterizes cancer stem cells associated with tumorigenesis, metastatic behavior, and poor outcomes. [36,37] While an IHC-based profiling of DCIS did not associate ALDH1 with breast cancer events, [9] our MS-based analysis on paired DCIS/IDC lesions, does show a significantly higher concentration of ALDH1 in DCIS compared to IDC lesions (p=0.01), supporting findings from stem cell biology that ALDH1 might be a functional and prognostic biomarker of tumorigenesis in DCIS.

Having access to 'real-world' mixed-type lesions, the most prevalent clinical phenotype of breast ductal carcinoma, we were in the unique position to investigate the proteome of DCIS lesions that are likely active in the transition to IDC, depleted from inter-tumor heterogeneity. Comparing pure DCIS to mixed-type DCIS lesions revealed significantly lower protein levels of KRT5, KRT14, KRT6B, and CEACAM5 in pure DCIS lesions (p<0.05), indicating stromal remodeling as a key feature in the progression from pre-cancer to invasive cancer, with prognostic value for DCIS management. High expression of KRTs is linked to good prognosis in breast cancer, while lower levels are associated with invasive tumor proliferation. [38-40] CEACAM5 (also CEA) expression has context-dependent impact and a protective function in breast cancer, with potential usefulness in disease monitoring. [9,41,42] Similarly, comparing pure IDC to mixed-type IDC lesions showed a loss of KRT expression in mixed-type IDC (p<0.05), suggesting a protective role of KRTs and marker of progressiveness in DCIS.

UniProt ID	Gene Symbol	IDC vs DCIS p value	IDC mixed vs DCIS mixed p value	IDC paired vs DCIS paired p value	DCIS pure vs DCIS mixed p value	IDC pure vs IDC mixed p value
P55317	FOXA1	<0.0001	0.1584	na	0.0277	0.2796
Q9BV36	MLPH	0.9937	0.9549	0.1108	0.1053	0.2391
O43570	CA12	<0.0001	na	0.2684	0.9932	na
P02751	FN1	0.1223	0.0584	0.0246	0.3442	0.5544
P08123	COL1A2	0.2946	0.7485	0.2399	0.5725	0.6809
Q15063	POSTN	0.006	0.6849	0.0293	0.1870	0.3989
P35442	THBS2	<0.0001	0.3087	0.0431	0.5240	0.9132
Q02487	DSC3	0.2055	0.7676	0.4584	0.0608	na
P13647	KRT5	0.6350	0.6893	0.4811	0.0008	0.0019
P02533	KRT14	0.3695	0.195	0.7197	0.0026	0.0006
P04259	KRT6B	0.5053	na	0.5849	0.0092	0.0491
P19012	KRT15	0.0949	0.1307	0.3695	0.2278	0.5593
Q05682	CALD1	0.1707	0.9676	0.8580	0.1800	0.4176
P51884	LUM	0.2691	0.3142	0.8070	0.6572	0.5536
P46777	RPL5	0.7106	0.1105	0.1934	0.1819	0.0578
P05154	SERPINA5	0.1831	na	0.7365	0.4077	0.2000
P06401	PGR	0.6468	0.4329	0.1435	0.3075	0.3864
P04626	HER2	0.4177	0.2724	0.8663	0.3003	0.7259
P00403	COX2	0.2184	0.6032	0.1624	0.4270	0.1042
P00352	ALDH1	0.4049	0.4452	0.0125	0.3963	0.9840
P16070	CD44	0.3565	0.8431	0.1683	0.2674	0.9394
P06731	CEACAM5	0.0592	na	na	0.0253	na

Table 1: Overlapping molecules from independent gene expression studies and this proteomic profiling. Differential expression values of 22 proteins corresponding to genes proposed in the literature as biomarkers for DCIS to IDC progression. In red statistically significant entities with a student t-test p-value <0.05, in blue entities close to the set p-value. na = 'not applicable'; the protein was not quantified in that dataset.

2.3 Loss of basal membrane stability, inflammatory processes, and epithelial-to-mesenchymal transition (EMT) identified as key events driving DCIS progression.

Having confirmed the results of genomic/transcriptomic studies in this setting using direct MS-based protein measurements, we turned to a global proteomics approach to discover further features of the DCIS-IDC scenario.

Differential expression analysis of more than 2800 proteins identified in pure DCIS compared to IDC, revealed ~388 DEPs using an unpaired t-test with post-hoc Benjamini-Hochberg FDR method for multiple hypothesis testing (q<0.01) and at least a 2-fold-change in protein expression between DCIS and IDC (Fig. 3 a, and Supplemental Table 1). To reduce the inter-patient variability, we compared proteomic profiles of DCIS and IDC lesions from the same patients (n=9). Ten differentially expressed proteins (DEPs) were identified: ILK, ITGA4, GPRC5A, FNTA, SCPEP1, EPB41L3, and SORBS1 were upregulated in DCIS, while ACAP1, ATP6V0A1, and KPRP were upregulated in IDC (Fig 3 b, and Supplemental Table 2).

ILK, an integrin linked kinase, regulates integrin signaling and is associated with tumor growth and metastasis. [43,44] ITGA4 mediates cell-cell adhesions and is linked to cancer progression, inflammatory reactions, and ECM stemness. [45-47] GPRC5A acts as an oncogene or tumor suppressor in different cancers. [48-50] Androgen receptor-regulated FNTA enhances KRAS signaling and might be involved in tumorigenesis. [51-56] SCPEP1 is associated with cancer development, growth and metastasis. [57-59] EPB41L3 is a tumor suppressor involved in apoptosis and cell cycle regulation. [60-63] Decreased expression in DCIS was observed for ATP6V0A1, which plays a role in pH homeostasis and tumor cell invasion. [64-66] ACAP1, which is associated with cell proliferation, migration, and immune infiltration in tumors. [67-69] Loss of ACAP1 could indicate impaired immune response in IDC progression. KPRP, involved in keratinocyte differentiation, [70,71] might contribute to invasiveness when its expression is lost in DCIS.

Overall, proteomic profiling of DCIS identified more than 380 putative biomarkers (protein level) to clinically profile DCIS lesions for risk stratification and disease management. The association of the differentially expressed proteins quantified in this study with hallmarks of cancer, such as remodeling of the tumor microenvironment (e.g., ILK, ITGA4, SCPEP1), escape of apoptosis (e.g., ILK, GPRC5A, FNTA, EPB41L3), deregulation of apical junction and energy metabolism (e.g., ATP6V0A1, KPRP, ITGA4), as well as inflammation and immune response processes (e.g., ACAP1, ITGA4) (Fig. 3c), warrants further investigation. Further, most of the identified DEPs are readily druggable and re-purposing of FDA-approved anti-inflammatory drugs and antibiotics pose interesting treatment options for DCIS.

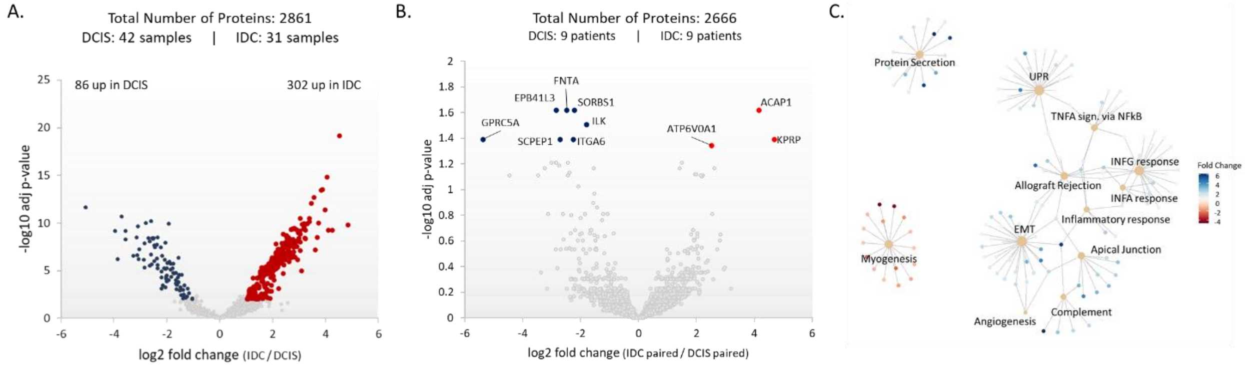

Figure 3: Differential Expression Analysis reflects loss of basal membrane stability, inflammatory processes, and epithelial-mesenchymal transition as key events towards DCIS to IDC progression. (A) Volcano plot of the proteome of pure IDC compared to pure DCIS lesions showing 388 differentially expressed proteins (unpaired t-test with post-hoc Benjamini-Krieger analysis p<0.01, abs log2 fold change >2). (B) Volcano plot of the proteome of paired IDC lesions compared to paired DCIS lesions showing 10 differentially expressed proteins (unpaired t-test with post-hoc Benjamini-Krieger analysis p<0.05, abs log2 fold change >2). (C) Molecular networks representing up-/downregulated pathways in IDC compared to DCIS lesions. UPR: Unfolded Protein Response, EMT: Epithelial-to-Mesenchymal Transition.

2.4 EIF2 and PI3K/Akt/mTOR signaling pathway potentially drive IDC phenotype development through dysregulation of central energy metabolism in cancer.

A deeper look into the molecular relationships of all the DEPs we've identified by functional enrichment analysis and gene set enrichment analysis, confirms the previously reported loss of basal layer integrity and epithelial to mesenchymal transitions (EMT) as key events supporting IDC. Figure 4a shows cancer hallmarks that are predominant for the IDC- and DCIS phenotype, highlighting the dysregulation of cell metabolism as a key event in the DCIS-phenotype. Proteomic profiling using MS-based techniques revealed metabolic vulnerabilities in DCIS that can provide insights into tumorigenic metabolic mechanisms, that were missed by genomic/transcriptomic analysis alone.

Functional Enrichment Analysis using IPA identifies mitochondrial dysfunction, granzyme A signaling, glucocorticoid receptor signaling and sirtuin signaling as significantly enriched (p-value of overlap <0.01) in our proteomics dataset, suggesting a dysregulation of glucose metabolism, through a shift from oxidative phosphorylation (i.e., tricarboxylic acid (TCA) cycle) to aerobic glycolysis (Figs. 4b and 4c). [72]

Aerobic glycolysis is also known as Warburg Effect and is characterized by high glucose uptake and glycolytic conversion of glucose to lactate to meet the high energy demands of proliferating cells. [73] During glycolysis, glucose is converted to pyruvate. Cytosolic pyruvate can either enter the TCA cycle for oxidative phosphorylation (OXPHOS) and ATP-production or be converted to lactate. Under normoxia, the metabolic fate of cytosolic pyruvate, and thus glucose metabolism, is regulated by pyruvate dehydrogenase complex (PDH) and lactate dehydrogenase (LDH), where the PDH reaction is favored. [73,74] PI3K/AKT signaling can modulate the metabolic fate of pyruvate as an upstream regulator of PDH and LDH, creating “pseudo-hypoxic” conditions that favor pyruvate conversion to lactate. The pivotal role of PI3K/AKT as an upstream regulator in metabolic reprogramming is comprehensively reviewed by Hoxhaj et al. [75] and involves the interaction with other proliferating signaling pathways, such as MAPK and mTOR. Our proteomic analysis of DCIS identified several differentially expressed molecules involved in glycolysis, hypoxia-mediated reactions and PI3K/AKT/mTOR signaling (Fig. 4c) which warrant further investigation.

Metabolomic profiling of FFPE specimens is challenging, because ~85% of metabolites are washed out during the preservation procedure. To nevertheless gain insights into metabolic changes occurring towards IDC progression, we conducted an AI-based metabolic profiling using QSM™ technology, which is supported by more than 500 publications. [17] Clear metabolic differences between DCIS/IDC lesions from the same patient (paired DCIS/IDC) were identified, but due to the large variability and small sample size (n=9) metabolic differences between the groups were hard to assess. A multitude of functional markers with direct causal relation to ATP production capacity and utilization of glucose were nevertheless identified (Tbl. 2). These findings confirm the dysregulation of energy metabolism towards IDC progression and suggest that the energy demand of transforming pre-invasive cells (DCIS-phenotype) is mainly achieved by fatty acid metabolism and lactate production.

To further evaluate and promote the translation of our findings into the clinic, we developed a highly multiplexed targeted MS-assay for absolute quantitation of 90 signature peptides, associated with cancer metabolism, central energy metabolism, RNA regulation and members of the PI3K/AKT/mTOR, EIF2 and EGFR/RAS/RAF signaling pathways. A complete list of peptides included in this assay is provided in Supplemental Table 4. The results of the PRM assay are depicted as STRING network (Figure 4d), where the differential expression is represented by the node color, and the absolute fold-change by the node size. These findings correlate well with the previously discussed observations from label-free proteomics and independent genomics/transcriptomics study, showing that DCIS tumors have a tendency towards loss of metabolic functions. Albumin (ALB) is significantly higher expressed in the DCIS phenotype compared to the IDC phenotype (q value = 0.03). Studies associated low albumin levels with changes in the tumor microenvironment to more favorable conditions for disease progression and tumor migration, suggesting that serum albumin levels might have a prognostic value for cancer. [76,77] Other studies discuss albumin as a potent marker for inflammation and the nutritional status of patients, where low albumin levels correlate with inflammatory processes resulting in higher morbidity and poor prognosis. [78,79] Our results support these findings, and highlight remodeling of the tumor microenvironment, environmental stress (i.e., malnutrition, which inhibits EIF2 signaling) [80] and inflammatory processes as key events towards IDC progression.

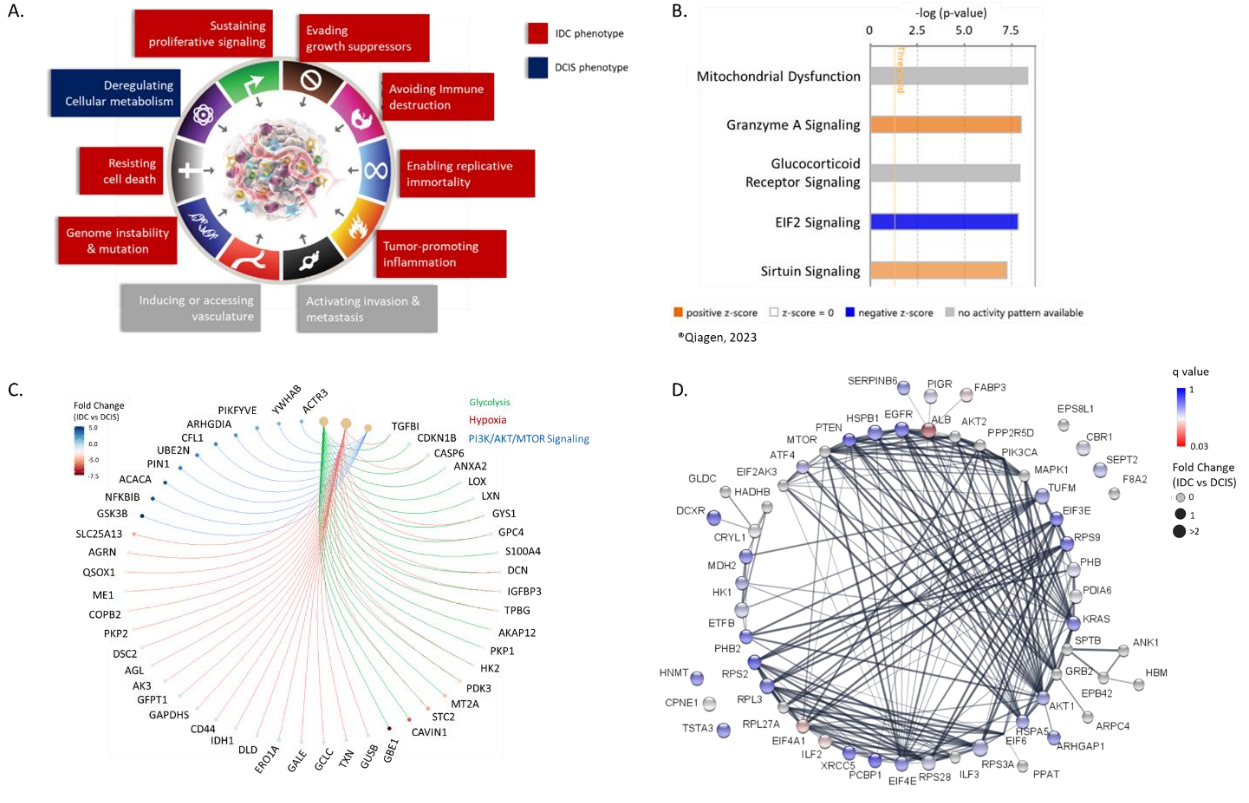

Figure 4: Dysregulation of central energy metabolism is a key event in the DCIS tumor phenotype. (A) Graphical representation of hallmarks of cancer (modified from [81]) characteristic for proteomic tumor profiling of DCIS and IDC tumors. (B) Top 5 canonical pathways from Ingenuity Pathway Analysis on differentially expressed proteins in 42 DCIS

and 31 IDC tumors. (C) Signature proteins potentially driving DCIS progression through glycolysis, hypoxia (or "pseudo-hypoxia") and PI3K/AKT/mTOR pathway, identified by gene set enrichment analysis. (D) STRING Network showing the protein expression profile of signature proteins, associated with cancer metabolism, RNA regulation and major cancer pathways, such as PI3K/AKT/mTOR and EGFR/RAS/RAF. Absolute concentration of the proteins was determined by parallel reaction monitoring. The color of the nodes represents q values from multiple-hypothesis testing using unpaired t-tests with post hoc correction using the Benjamini-Krieger FDR method (1% FDR). The node size represents the fold-change. Grey nodes were not quantified, either because no SIS/NAT was available or because there were more than 60% missing values. Edges represent physical and/or functional interaction partners based on the STRING database.

Table 2: List of putative metabolic biomarkers identified by AI-based metabolic profiling of DCIS and IDC specimens from the same patient.

ATP Production Capacity	CPT2, ACADM, HCDH, NDUBA, NDUBB, NDUV1, NDUV2, NDUS1, NDUS2, QCR1, QCR2, CY1, UCRI, QCR6, QCR7, QCR8, ATPA, ATPB, ATPD, ATP5H, ATP5I, ATPO, ADT3, MCEE, MUTA, THIK, THIM, ECHB, THIL, ODPA, ODPB, ODP2, DLDH, CISY, ACON, IDH3A, ODO2, SUCA, SUCB2, FUMH, MDHM, ACPM, NDUA2
Glucose Utilization	HXK1, ALDOC, PGAM1, ENOG

3 Discussion

Clinical research on DCIS has been limited due to low sample numbers, high inter-tumor heterogeneity and low tissue quality, as most DCIS lesions derive from diagnostic needle-core-biopsies and are FFPE embedded. Although genetic/transcriptomic studies of DCIS progression provide a cellular blueprint of what might happen, genes cannot be readily targeted for therapy and post-translational modification cannot be assessed by genetic screening alone. Quantitative proteomics can complement and confirm genetic changes and provides a deeper look into the 'real-life' tumor phenotype. The readily druggable nature of proteins makes quantitative proteomics studies attractive for clinical research. Additionally, mass spectrometry-based studies allow both (i) discovery studies for comprehensive tumor profiling and (ii) validation studies in a highly multiplexed manner, with unprecedented accuracy, specificity, and sensitivity.

We established a label-free quantitative proteomics pipeline suitable for needle-core biopsy sized FFPE specimens and performed a comprehensive proteomic phenotyping of DCIS and IDC using less than 1% of the total extracted protein material. We cover 6 orders of magnitude of the disease proteome and identify more than 380 differentially expressed proteins that identify classical hallmarks of cancer, reflective for high transcriptional activity, extracellular matrix remodeling and inflammation processes as key events towards IDC progression. We further identify dysregulation of glucose metabolism as a key event in the transition from pre-invasive to invasive carcinoma. Guided by these results, we developed a highly multiplexed parallel-reaction monitoring (PRM) assay for precise quantitation of 90 proteins, that are associated with cancer metabolism, RNA regulation and major cancer pathways, such as PI3K/AKT/mTOR and EGFR/RAS/RAF. We applied this assay to generate an activation profile of these signature proteins for proliferation and metabolic remodeling in cancer in ‘real world’ clinical samples and were able to support observations from label-free proteomics data with absolute concentrations in the amol range, facilitating the translation of our findings into the clinic. Notably, proteomics profiling revealed that FDA-approved drugs, such as antibiotics and NSAID, may be repurposed for DCIS and IDC treatment, as they have been shown to control and target proteins identified as key events towards IDC progression.

It is important to highlight, that this study design is applicable to many diseases with limited sample volumes and low tissue quality, as it requires only a fraction of the total sample amount allowing discovery and validation studies in the same sample cohort. In our opinion, clinical proteomics is a versatile tool for comprehensive tumor phenotyping, able to capture a ‘real-life’ snapshot of tumor phenotypes, representative of post-translational modifications and epigenetic changes. More than 99% of published clinical biomarkers/genomic assays fail to enter clinical practice, [82] but we show here that complementing genomics and transcriptomics studies with proteomics data, and vice versa, will help create a better understanding of underlying disease mechanisms and will better inform the selection of biomarker candidates and patient enrolment for clinical studies, ultimately improving the quality and final results of clinical trials.

4 Methods

All chemicals and reagents were purchased by Sigma Aldrich (St. Louis, MI, USA) unless otherwise specified. Sequencing grade trypsin (Promega, P/N V511A) was used for the generation of tryptic peptides.

4.1 Clinical Specimens

Clinical specimens were obtained from patients who consented for tissue biobanking part of the Jewish General Hospital Breast Biobank (protocol 05-006). The study was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and was approved by the Jewish General Hospital Research Ethics Board.

A total of 50 clinical cases of patients diagnosed and treated with DCIS and/or IDC at the JGH were carefully curated by a pathologist with expertise in breast cancer to select lesions meeting inclusion criteria for mass spectrometry-based (MS-based) analysis, i.e., at least 30% tumor cellularity and less than 10% necrosis. The patients were of Caucasian ethnicity ranging from 22 to 82 years of age at first diagnosis (median age 52 years). The patients were followed for a period of 1 to 18 years (median 8 years). During the period of follow-up, 43 patients have no evidence of disease, and 1 patient has metastatic disease, while 8 patients died from cancer. The cohort comprises 29 cases with pure DCIS lesions, 18 cases with pure IDC lesions, 13 cases with mixed-type lesions (IDC with in-situ components) and 9 cases with synchronous/metachronous DCIS and IDC. Clinical data for the patients is available upon request.

4.2 Sample Preparation

1 mm-diameter tissue cores (~0.8 mm^3 tissue volume) were prepared from FFPE-blocks enriching for DCIS or IDC only tumor cells. Excessive paraffin was trimmed off using a clean scalpel blade. Protein extraction was performed following our developed FFPE-proteomics workflow for core needle biopsies. Briefly, paraffin was removed by incubation with hot water (~80 °C). Each deparaffinized core was mechanically disrupted using a micropestle (Sigma Aldrich, #BAF199230001) in 250 µL of 2% sodium deoxycholate (SDC), 50 mM Tris-HCl, 10 mM tris(2-carboxyethyl)phosphine (TCEP), pH 8.5, followed by sequential incubation on an Eppendorf ThermoMixer C for 20 min at 99 °C (1100 rpm) and for 2 h at 80°C (1100 rpm). Samples were cooled down on ice for 1 minute before a 15-minute centrifugation at 21,000x g (4 °C) to remove cell debris. The supernatant was collected into a Protein LoBinding tube (Eppendorf, Germany) and the total protein concentration was determined using a Pierce Reducing Agent Compatible BCA kit (RAC-BCA, Thermo Scientific, P/N 23252) following the manufacturer's instructions. Free cysteine residues were alkylated with iodoacetamide to a final concentration of 30 mM and incubation for 30 minutes at room temperature, protected from light.

For 2 µg of protein lysate, 2 µL of ferromagnetic beads with MagReSyn® Hydroxyl functional groups (ReSyn Biosciences, Gauteng, South Africa, 20 µg/mL) were equilibrated with 100 µL of 70% ACN, briefly vortexed and placed on a magnetic rack to remove the supernatant. This step was repeated another two times. Next, the protein extracts were added to the beads and the sample was adjusted to a final concentration of 70% ACN, thoroughly vortexed and incubated for 10 min at room temperature without shaking. The following washing steps were performed on a magnetic rack without disturbing the protein/bead aggregate. The supernatants were discarded, and the beads were washed on the magnetic rack with 1 mL of 95% ACN for 10 s, followed by a wash with 1 mL of 70% ACN without disturbing the protein/bead aggregate. The tubes were removed from the magnetic rack, 100 µL of digestion buffer (1:20 (w/w) trypsin:protein in 0.2 M GuHCl, 50 mM AmBic, 2 mM $CaCl_2$) were added and the samples were incubated at 37 °C for 12 h. After acidification with trifluoroacetic acid (TFA) to a final concentration of 2%, the tubes were placed on the magnetic rack for 1 min, followed by removal of the supernatant. To remove residual beads, the samples were centrifuged at 20,000x g for 10 min.

4.3 Preparation of spiking solutions for the response curve and absolute quantitation

In order to promote translation of our findings and to validate LFQ Abundances with a more precise targeted MS approach we developed a multiplexed parallel reaction monitoring (PRM) method to quantify 90 proteins in FFPE specimens, measuring the concentration of a unique signature peptide for each protein. All 90 peptides were measured in a single LC-MS/MS run. Two equimolar synthetic peptide mixtures (100 fmol/µg of each peptide) were prepared in 30% ACN with 0.1% formic acid (FA) in water (w/v); one mixture contained unlabeled peptides (light or NAT peptides), and the second mixture contained stable isotope labeled standard peptides (heavy or SIS peptides). The light peptide mixture was used to develop the highly multiplexed PRM assay with optimized peptide-specific parameters, such as collision energy and charge state, while the heavy peptide mixture was used for normalization, serving as spiking solution and internal standard for clinical samples.

Quantitation was performed using a 7-point response curve consisting of a variable amount of light peptides, ranging from 0.41 to 250 fmol (three orders of magnitude), and a constant amount of SIS peptides (50 fmol). Digested bovine serum albumin (BSA, 0.01 µg) was used as surrogate matrix of the response curve. To determine the limit of detection (LOD), a double blank sample was prepared. The blank sample consisted of 0.01 µg BSA digest spiked with 50 fmol of the SIS-mixture and analyzed before and/or directly after the highest calibrant level of the response curve. For quantitation of endogenous protein in the patient samples, 50 fmol of SIS peptide were spiked into 1 µg total digested tissue protein, as determined by RAC-BCA.

4.4 Data analysis

1 µg digested protein was pre-concentrated on EV2001 C18 Evotips and separated on a heated (40 °C) EV1137 column (15 cm x 150 µm, 1.5 µm particle size) using Evosep's "Extended meth-od" (15 samples-per-day (SPD)). The samples were analyzed by data dependent acquisition (DDA) mode, on a Q Exactive Plus Orbitrap mass spectrometer operated with a Nanospray Flex ion source (both from Thermo Fisher Scientific), connected to an Evosep One HPLC (Evosep Bio-systems, Odense, Denmark). Full MS scans were acquired over the mass range from m/z 350 to m/z 1500 at a resolution of 70,000 with an automatic gain control (AGC) target value of 1×10^6 and a maximum injection time of 50 ms. The 15 most-intense precursor ions (charge states +2, +3, +4) were isolated with a window of 1.2 Da and fragmented using a normalized collision energy of 28; the dynamic exclusion was set to 30 s. MS/MS spectra were acquired at a mass resolution of 17,500, using an AGC target value of 2×10^4 and a maximum injection time of 64 ms.

Chromatographic separation of all PRM runs was performed with the same equipment and buffers as described above. The Q Exactive Plus was operated in PRM mode at a resolution of 35,000. Target precursor ions were isolated with the quadrupole isolation window set to m/z 1.2. An AGC target of $3x10^6$ was used, allowing for a maximum injection time of 110 ms. Data was acquired in time-scheduled mode, allowing a 2-min retention-time window for each target. Full MS scans were acquired in parallel at low resolution (m/z 17,500) with an AGC target value of $1x10^6$ and a maximum injection time of 50 ms, to ensure sample quality.

MS data files are publicly available through the ProteomeXchange Consortium via the PRIDE partner repository[83] with the dataset identifier PXD040782. The synthetic peptides selected for this PRM assay were validated by others, information is available through National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC) Assay Portal (assays.cancer.gov).

4.5 Data Processing and Differential Expression Analysis

MS raw data were processed using Proteome Discoverer 2.5 (Thermo Fisher Scientific). Database searches were performed using SequestHT with Multi-Peptide Search (MPS) and a human Swissprot database (January 2019; 20,414 target entries). Label-free quantitation (LFQ) was performed using the Minora feature-detector node within Proteome Discoverer, and the Percolator software was used to calculate posterior error probabilities. Database searches were performed using trypsin as enzyme with a maximum of 2 missed cleavages. Carbamidomethylation of cysteine (+57.021 Da) was set as a fixed modification, and oxidation of methionine (+15.995 Da) as variable modifications. Mass tolerances were set to 10 ppm for precursor ions and 0.02 Da for product ions. The data were filtered to a false discovery rate (FDR) <1% at the peptide and protein levels. Only proteins that were (i) identified with at least one protein unique peptide and (ii) quantified in ≥60% of replicates of at least one of the study groups, were considered for the quantitative comparison. Protein LFQ data obtained from Proteome Discoverer was normalized based on summed protein intensities to correct for differences in sample loading. Missing protein intensity values were imputed using 1.5x the minimum observed intensity for this particular sample. The obtained normalized abundances were used for unpaired t-tests (two tailed, 95% confidence) and differential expression analysis on log2-transformed data with multiple hypothesis testing using the Benjamini-Krieger false-discovery approach (FDR 1%). Proteins having q-values of <0.01 and absolute log2 fold-changes >1 were considered as differential between test-ed groups. Statistical analysis was performed using GraphPad Prism 9 (San Diego, CA, USA).

Raw PRM data were analyzed using Skyline (v22.2.0.351). [84] Correct peak integration and visual verification of detected peaks was performed manually for each target, and the three to four highest and most stable transitions were selected for quantitation. A linear regression model with $1/x^2$ weighting using the SIS/NAT ratio of each target peptide was used for the calculation of concentrations. Only calibration levels meeting following criteria were accepted for response curve generation and regression analysis; precision average <20% CV per calibration level, accuracy average between 80% and 120% per calibrant level, quantified in at least 3 consecutive calibrant levels. The LOD describes the smallest concentration of the target peptide (analyte) that is likely to be reliably distinguished from instrument noise and at which detection is feasible. To determine the LOD we use replicate injections from a double blank sample, i.e., fixed concentration of the SIS-peptide in surrogate matrix. The average concentration of the double blank plus 3.3x the standard deviation of the blank replicates is used to calculate the lowest detectable concentration for each peptide. The limit of quantitation (LOQ) describes the lowest concentration at which the analyte can not only be reliably detected, but at which above mentioned precision and accuracy criteria are met. Here the LOQ was defined as the lowest calibration level for each peptide. Proteins/Peptides with more than 60% missing values were excluded from downstream analysis.

4.6 Functional Enrichment Analysis

Functional Enrichment Analysis was performed using the 'Core Analysis' function within Ingenuity Pathway Analysis (Qiagen, Inc., content version: 81348237, release Date: 2022-09-15). [85] Ingenuity Knowledge Base was used as reference set, allowing direct and indirect relationships. Only molecules having expression p-values <0.05 and absolute log2 fold-changes of >1 were considered for the core analysis. All other settings were kept with default parameters.

4.7 Gene Set Enrichment Analysis

A pre-ranked Gene Set Enrichment Analysis (GSEA) was performed using GSEA v4.3.2 (Broad Institute, Inc.) software. The gene list was ranked by differential expression using SIGN function within Excel with calculated log2 fold-change and p-value from an unpaired t-test. A hallmark gene set Molecular Signature Database (MSigDB v2022.1) [86] was used as references gene set. The search allowed 1000 permutations, with set sizes between 15 and 500 genes. Pathways were collapsed to remove redundancy and to increase selectivity and specificity. Data was visualized using the clusterProfiler[87] package within R.

4.8 Metabolic Analysis

Protein expression data from paired DCIS/IDC cases was sent to Doppelganger Biosystems Inc (Berlin, Germany) for metabolic profiling using Quantitative Systems Metabolism (QSM™) technology. [17]

Supplemental Materials: Supplemental Tables T1-T4 with detailed quantitative proteomics data and results from differential expression analysis.

Author Contributions: Conceptualization, G.M., G.B., C.H.B; methodology, G.M., R.P.Z.; sample preparation and formal analysis, G.M.; writing-original draft preparation, G.M.; writing-review and editing, G.M.,

A.A.-M., M.B., G.B., C.H.B.; providing of resources, L.F., J.L., A.A-M, M.B., C.H.B. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by Genome Canada through the Genomics Technology Platform (264PRO). The JGH breast biobank is supported by funds from the Quebec Breast Cancer Foundation and Réseau Recherche Cancer. GM thanks the McGill Faculty of Medicine for student funding through the Ruth and Alex Dworkin Fellowship.

Acknowledgement: CHB is grateful to Genome Canada for financial support through the Genomics Technology Platform (264PRO). CHB is also grateful for support from the Segal McGill Chair in Molecular Oncology at McGill University (Montreal, Quebec, Canada) and for support from the Terry Fox Research Institute, the Warren Y. Soper Charitable Trust, and the Alvin Segal Family Foundation to the Jewish General Hospital (Montreal, Quebec, Canada). GM would like to acknowledge the patients and their families for their consent to archive biological specimens for clinical research. GM would like to thank Dr. Naciba Benlimame, Lilian Canetti and Dr. James Saliba for their guidance and constructive discussions.

This work was done under the auspices of a Memorandum of Understanding between McGill and the U.S. National Cancer Institute's International Cancer Proteogenome Consortium (ICPC). ICPC encourages international cooperation among institutions and nations in proteogenomic cancer research in which proteogenomic datasets are made available to the public. This work was also done in collaboration with the U.S. National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC).

Institutional Review Board Statement: The study was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and was approved by the Jewish General Hospital Research Ethics Board.

Informed Consent Statement: Clinical specimens were obtained from patients who consented for tissue biobanking part of the Jewish General Hospital Breast Biobank (protocol 05-006).

Data Availability Statement: MS data files are publicly available through the ProteomeXchange Consortium via the PRIDE partner repository[83] with the dataset identifier PXD040782. Clinical data and H&E staining of clinical specimens used in this study are available upon request.

Conflicts of Interest: CHB is the CSO of MRM Proteomics, Inc. and the CTO of Molecular You. The other authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript.

References

1. Mannu, G.S., et al. Invasive breast cancer and breast cancer mortality after ductal carcinoma in situ in women attending for breast screening in England, 1988-2014: population based observational cohort study. BMJ 369, m1570 (2020).
2. Kumar, A.S., Bhatia, V. & Henderson, I.C. Overdiagnosis and overtreatment of breast cancer: Rates of ductal carcinoma in situ: a US perspective. Breast Cancer Res. 7, 271 (2005).
3. Sanders, M.E., Schuyler, P.A., Dupont, W.D. & Page, D.L. The natural history of low-grade ductal carcinoma in situ of the breast in women treated by biopsy only revealed over 30 years of long-term follow-up. Cancer 103, 2481-2484 (2005).
4. Jones, J.L. Overdiagnosis and overtreatment of breast cancer: Progression of ductal carcinoma in situ: the pathological perspective. Breast Cancer Res. 8, 204 (2006).
5. Collins, L.C., et al. Outcome of patients with ductal carcinoma in situ untreated after diagnostic biopsy: results from the Nurses' Health Study. Cancer 103, 1778-1784 (2005).
6. Morrissey, R.L., Thompson, A.M. & Lozano, G. Is loss of p53 a driver of ductal carcinoma in situ progression? Br. J. Cancer 127, 1744-1754 (2022).
7. Badve, S., et al. Prediction of local recurrence of ductal carcinoma in situ of the breast using five histological classifications: a comparative study with long follow-up. Hum. Pathol. 29, 915-923 (1998).
8. Badve, S. & Gökmen-Polar, Y. Tumor Heterogeneity in Breast Cancer. Adv. Anat. Pathol. 22, 294-302 (2015).
9. Badve, S.S., et al. Multi-protein spatial signatures in ductal carcinoma in situ (DCIS) of breast. Br. J. Cancer 124, 1150-1159 (2021).
10. Gerdes, M.J., et al. Single-cell heterogeneity in ductal carcinoma in situ of breast. Mod. Pathol. 31, 406-417 (2018).
11. Mitsa, G., et al. A Non-Hazardous Deparaffinization Protocol Enables Quantitative Proteomics of Core Needle Biopsy-Sized Formalin-Fixed and Paraffin-Embedded (FFPE) Tissue Specimens. Int. J. Mol. Sci. 23, 4443 (2022).
12. Buszewska-Forajta, M., et al. Paraffin-embedded tissue as a novel matrix in metabolomics study: optimization of metabolite extraction method. Chromatographia 82, 1501-1513 (2019).
13. Cacciatore, S. & Loda, M. Innovation in metabolomics to improve personalized healthcare. Annals of the New York Academy of Sciences 1346, 57-62 (2015).
14. Cacciatore, S., et al. Metabolic Profiling in Formalin-Fixed and Paraffin-Embedded Prostate Cancer TissuesMetabolic Profile in FFPE Tissues. Mol. Cancer Res. 15, 439-447 (2017).

15. Dannhorn, A., et al. Evaluation of Formalin-Fixed and FFPE Tissues for Spatially Resolved Metabolomics and Drug Distribution Studies. Pharmaceuticals 15, 1307 (2022).

16. Neef, S.K., et al. Optimized protocol for metabolomic and lipidomic profiling in formalin-fixed paraffin-embedded kidney tissue by LC-MS. Anal. Chim. Acta 1134, 125-135 (2020).

17. Berndt, N., Kann, O. & Holzhütter, H.G. Physiology-based kinetic modeling of neuronal energy metabolism unravels the molecular basis of NAD(P)H fluorescence transients. J. Cereb. Blood Flow Metab. 35, 1494-1506 (2015).

18. Nachmanson, D., et al. The breast pre-cancer atlas illustrates the molecular and micro-environmental diversity of ductal carcinoma in situ. NPJ Breast Cancer 8, 6 (2022).

19. Nagasawa, S., et al. Genomic profiling reveals heterogeneous populations of ductal carcinoma in situ of the breast. Commun Biol 4, 438 (2021).

20. Sorochan Armstrong, M.D., de la Mata, A.P. & Harynuk, J.J. Review of Variable Selection Methods for Discriminant-Type Problems in Chemometrics. Frontiers in Analytical Science 2(2022).

21. Jain, A.P., et al. Pan-cancer quantitation of epithelial-mesenchymal transition dynamics using parallel reaction monitoring-based targeted proteomics approach. J. Transl. Med. 20, 1-13 (2022).

22. Chakraborty, S., Hosen, M.I., Ahmed, M. & Shekhar, H.U. Onco-multi-OMICS approach: a new frontier in cancer research. BioMed research international 2018(2018).

23. Dunn, J., et al. Integration and comparison of transcriptomic and proteomic data for meningioma. Cancers (Basel) 12, 3270 (2020).

24. Harnik, Y., et al. Spatial discordances between mRNAs and proteins in the intestinal epithelium. Nature metabolism 3, 1680-1693 (2021).

25. Rebbeck, C.A., et al. Gene expression signatures of individual ductal carcinoma in situ lesions identify processes and biomarkers associated with progression towards invasive ductal carcinoma. Nat Commun 13, 3399 (2022).

26. Dettogni, R.S., et al. Potential biomarkers of ductal carcinoma in situ progression. BMC Cancer 20, 119 (2020).

27. Bernardo, G.M., et al. FOXA1 represses the molecular phenotype of basal breast cancer cells. Oncogene 32, 554-563 (2013).

28. Bernardo, G.M. & Keri, R.A. FOXA1: a transcription factor with parallel functions in development and cancer. Biosci. Rep. 32, 113-130 (2012).

29. Seachrist, D.D., Anstine, L.J. & Keri, R.A. FOXA1: A Pioneer of Nuclear Receptor Action in Breast Cancer. Cancers (Basel) 13(2021).

30. Williamson, E.A., et al. BRCA1 and FOXA1 proteins coregulate the expression of the cell cycle-dependent kinase inhibitor p27(Kip1). Oncogene 25, 1391-1399 (2006).
31. Wolf, I., et al. FOXA1: Growth inhibitor and a favorable prognostic factor in human breast cancer. Int. J. Cancer 120, 1013-1022 (2007).
32. Liu, T., Zhou, L., Xiao, Y., Andl, T. & Zhang, Y. BRAF Inhibitors Reprogram Cancer-Associated Fibroblasts to Drive Matrix Remodeling and Therapeutic Escape in Melanoma. Cancer Res. 82, 419-432 (2022).
33. Barnett, D.H., et al. Estrogen receptor regulation of carbonic anhydrase XII through a distal enhancer in breast cancer. Cancer Res. 68, 3505-3515 (2008).
34. Li, Y., et al. High expression of carbonic anhydrase 12 (CA12) is associated with good prognosis in breast cancer. Neoplasma 66, 420-426 (2019).
35. Ning, W.R., et al. Carbonic anhydrase XII mediates the survival and prometastatic functions of macrophages in human hepatocellular carcinoma. J. Clin. Invest. 132(2022).
36. Charafe-Jauffret, E., et al. Aldehyde dehydrogenase 1-positive cancer stem cells mediate metastasis and poor clinical outcome in inflammatory breast cancer. Clin Cancer Res 16, 45-55 (2010).
37. Douville, J., Beaulieu, R. & Balicki, D. ALDH1 as a functional marker of cancer stem and progenitor cells. Stem Cells Dev. 18, 17-25 (2009).
38. Han, W., Hu, C., Fan, Z.J. & Shen, G.L. Transcript levels of keratin 1/5/6/14/15/16/17 as potential prognostic indicators in melanoma patients. Sci. Rep. 11, 1023 (2021).
39. Saha, S.K., et al. KRT19 directly interacts with beta-catenin/RAC1 complex to regulate NUMB-dependent NOTCH signaling pathway and breast cancer properties. Oncogene 36, 332-349 (2017).
40. Saha, S.K., Yin, Y., Chae, H.S. & Cho, S.-G. Opposing Regulation of Cancer Properties via KRT19-Mediated Differential Modulation of Wnt/β-Catenin/Notch Signaling in Breast and Colon Cancers. in Cancers (Basel), Vol. 11 (2019).
41. Bechmann, M.B., Brydholm, A.V., Codony, V.L., Kim, J. & Villadsen, R. Heterogeneity of CEACAM5 in breast cancer. Oncotarget 11, 3886-3899 (2020).
42. Yang, C., et al. Down-regulation of CEACAM1 in breast cancer. Acta Biochim Biophys Sin (Shanghai) 47, 788-794 (2015).
43. Hannigan, G.E., et al. Regulation of cell adhesion and anchorage-dependent growth by a new beta 1-integrin-linked protein kinase. Nature 379, 91-96 (1996).
44. Delcommenne, M., et al. Phosphoinositide-3-OH kinase-dependent regulation of glycogen synthase kinase 3 and protein kinase B/AKT by the integrin-linked kinase. Proc. Natl. Acad. Sci. U. S. A. 95, 11211-11216 (1998).

45. Kinashi, T. Overview of integrin signaling in the immune system. Methods Mol. Biol. 757, 261-278 (2012).

46. Mo, J., et al. The early predictive effect of low expression of the ITGA4 in colorectal cancer. J. Gastrointest. Oncol. 13, 265-278 (2022).

47. Pulkka, O.P., et al. Clinical relevance of integrin alpha 4 in gastrointestinal stromal tumours. J Cell Mol Med 22, 2220-2230 (2018).

48. Zhou, H. & Rigoutsos, I. The emerging roles of GPRC5A in diseases. Oncoscience 1, 765-776 (2014).

49. Qian, X., Jiang, C., Shen, S. & Zou, X. GPRC5A: An emerging prognostic biomarker for predicting malignancy of Pancreatic Cancer based on bioinformatics analysis. J. Cancer 12, 2010-2022 (2021).

50. Yang, L., Zhao, S., Zhu, T. & Zhang, J. GPRC5A Is a Negative Regulator of the Pro-Survival PI3K/Akt Signaling Pathway in Triple-Negative Breast Cancer. Front. Oncol. 10, 624493 (2020).

51. Chen, J., et al. Androgen receptor-regulated circ FNTA activates KRAS signaling to promote bladder cancer invasion. EMBO reports 21, e48467 (2020).

52. Tian, J., et al. circ-FNTA accelerates proliferation and invasion of bladder cancer. Oncol. Lett. 19, 1017-1023 (2020).

53. Cox, A.D. & Der, C.J. Farnesyltransferase inhibitors and cancer treatment: targeting simply Ras? Biochimica et Biophysica Acta (BBA)-Reviews on Cancer 1333, F51-F71 (1997).

54. Head, J. & Johnston, S.R. New targets for therapy in breast cancer: farnesyltransferase inhibitors. Breast Cancer Res. 6, 1-7 (2004).

55. Rowinsky, E.K., Windle, J.J. & Von Hoff, D.D. Ras protein farnesyltransferase: a strategic target for anticancer therapeutic development. J. Clin. Oncol. 17, 3631-3652 (1999).

56. Sebti, S.M. & Hamilton, A.D. Farnesyltransferase and geranylgeranyltransferase I inhibitors and cancer therapy: lessons from mechanism and bench-to-bedside translational studies. Oncogene 19, 6584-6593 (2000).

57. Fish, L., et al. Cancer cells exploit an orphan RNA to drive metastatic progression. Nat. Med. 24, 1743-1751 (2018).

58. Santhekadur, P.K. & Kumar, D.P. RISC assembly and post-transcriptional gene regulation in Hepatocellular Carcinoma. Genes Dis 7, 199-204 (2020).

59. Pan, X., Wang, Y., Lübke, T., Hinek, A. & Pshezhetsky, A.V. Mice, double deficient in lysosomal serine carboxypeptidases Scpep1 and Cathepsin A develop the hyperproliferative vesicular corneal dystrophy and hypertrophic skin thickenings. PLoS One 12, e0172854 (2017).

60. Gu, Y.Y., et al. HDAC10 Inhibits Cervical Cancer Progression through Downregulating the HDAC10-microRNA-223-EPB41L3 Axis. J. Oncol. 2022, 8092751 (2022).

61. Jiang, W. & Newsham, I.F. The tumor suppressor DAL-1/4.1B and protein methylation cooperate in inducing apoptosis in MCF-7 breast cancer cells. Mol. Cancer 5, 4 (2006).
62. Tuerxun, G., et al. Over-expression of EPB41L3 promotes apoptosis of human cervical carcinoma cells through PI3K/AKT signaling. Acta Biochim. Pol. 69, 283-289 (2022).
63. Zeng, R., et al. EPB41L3 is a potential tumor suppressor gene and prognostic indicator in esophageal squamous cell carcinoma. Int. J. Oncol. 52, 1443-1454 (2018).
64. Aoto, K., et al. ATP6V0A1 encoding the a1-subunit of the V0 domain of vacuolar H+-ATPases is essential for brain development in humans and mice. Nature communications 12, 2107 (2021).
65. Cotter, K., Stransky, L., McGuire, C. & Forgac, M. Recent Insights into the Structure, Regulation, and Function of the V-ATPases. Trends Biochem Sci 40, 611-622 (2015).
66. Capecci, J. & Forgac, M. The function of vacuolar ATPase (V-ATPase) a subunit isoforms in invasiveness of MCF10a and MCF10CA1a human breast cancer cells. J. Biol. Chem. 288, 32731-32741 (2013).
67. Wang, N., Zhu, L., Xu, X., Yu, C. & Huang, X. Integrated analysis and validation reveal ACAP1 as a novel prognostic biomarker associated with tumor immunity in lung adenocarcinoma. Comput Struct Biotechnol J 20, 4390-4401 (2022).
68. Yi, Q., Pu, Y., Chao, F., Bian, P. & Lv, L. ACAP1 Deficiency Predicts Inferior Immunotherapy Response in Solid Tumors. in Cancers (Basel), Vol. 14 (2022).
69. Zhang, J., Zhang, Q., Zhang, J. & Wang, Q. Expression of ACAP1 Is Associated with Tumor Immune Infiltration and Clinical Outcome of Ovarian Cancer. DNA Cell Biol 39, 1545-1557 (2020).
70. Lee, W.H., et al. Molecular cloning and expression of human keratinocyte proline-rich protein (hKPRP), an epidermal marker isolated from calcium-induced differentiating keratinocytes. J. Invest. Dermatol. 125, 995-1000 (2005).
71. Liu, Q., et al. Genome-wide association analysis reveals regulation of at-risk loci by DNA methylation in prostate cancer. Asian J Androl 23, 472-478 (2021).
72. Boland, M., Chourasia, A. & Macleod, K. Mitochondrial Dysfunction in Cancer. Front. Oncol. 3(2013).
73. National Cancer Institute. New Clarity on the Warburg Effect. (2022).
74. Wu, Z., et al. OMA1 reprograms metabolism under hypoxia to promote colorectal cancer development. EMBO Rep 22, e50827 (2021).
75. Hoxhaj, G. & Manning, B.D. The PI3K-AKT network at the interface of oncogenic signalling and cancer metabolism. Nat. Rev. Cancer 20, 74-88 (2020).
76. Fujii, T., et al. Implications of Low Serum Albumin as a Prognostic Factor of Long-term Outcomes in Patients With Breast Cancer. In Vivo 34, 2033-2036 (2020).

77. Soeters, P.B., Wolfe, R.R. & Shenkin, A. Hypoalbuminemia: Pathogenesis and Clinical Significance. JPEN J. Parenter. Enteral Nutr. 43, 181-193 (2019).
78. Galata, C., et al. Role of Albumin as a Nutritional and Prognostic Marker in Elective Intestinal Surgery. Can. J. Gastroenterol. Hepatol. 2020, 7028216 (2020).
79. von Meyenfeldt, M. Cancer-associated malnutrition: an introduction. Eur. J. Oncol. Nurs. 9 Suppl 2, S35-38 (2005).
80. Roux, P.P. & Topisirovic, I. Signaling Pathways Involved in the Regulation of mRNA Translation. Mol Cell Biol 38(2018).
81. Hanahan, D. Hallmarks of Cancer: New Dimensions. Cancer Discov. 12, 31-46 (2022).
82. Kern, S.E. Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. Cancer Res. 72, 6097-6101 (2012).
83. Vizcaino, J.A., et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. 32, 223-226 (2014).
84. MacLean, B., et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26, 966-968 (2010).
85. Krämer, A., Green, J., Pollard, J., Jr. & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics 30, 523-530 (2014).
86. Liberzon, A., et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 1, 417-425 (2015).
87. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS: A Journal of Integrative Biology 16, 284-287 (2012).

Chapter 4: Conclusion, General Discussion and Future Direction

This thesis aims to enhance the use of quantitative protein mass spectrometry in clinical research by developing (targeted) methods for the quantification of proteins in patient biopsy specimens. The results complement current clinical practice for stratification of patients with hard-to-treat malignant diseases, such as breast ductal carcinoma. The underlying hypothesis is that MS-derived data can enhance therapies driven by genomic data alone, as these may often be insufficient since genomic data may fail to identify aberrant pathway activities in specific tumors.

This work has explored the paradigm shift in clinical research from genetic screening to multi-omics approaches for the discovery of therapeutic vulnerabilities in hard-to-treat malignancies. The thesis emphasizes the limitations of genetic screening alone and highlights the importance of integrating it with proteomic data to gain a more comprehensive understanding of cancer biology and to identify potential targets for therapy. This work aimed to demonstrate the potential of multi-omics and mass spectrometry-based proteomics for integrative clinical research and biomarker development, and presents two original proteomics research papers where the data not only complements genomic/transcriptomic data, but also reveals new therapeutic vulnerabilities that can contribute to the development of personalized or targeted therapies for cancer patients, whose disease does not respond to standard-of-care treatments or for whom no (personalized) treatment options are currently available.

The thesis begins by discussing the advancements in genomics research, particularly the development of next-generation sequencing (NGS) technologies. While genomics has been instrumental in identifying driver alterations and guiding targeted therapies, it has been acknowledged that the clinical outcome is not solely determined by genetic mutations. The complex molecular consequences of these alterations and their impact on treatment response are still poorly understood. Therefore, the integration of multi-omic data, including proteomics and metabolomics, is crucial for a more accurate and complete characterization of tumors, and for the identification of therapeutic vulnerabilities.

Transcriptomics is then introduced as an essential component of multi-omics research, providing insights into dynamic changes in gene expression in response to environmental stimuli. While mRNA expression can serve as a surrogate for protein expression at steady state, the correlation between transcriptomics and proteomics data is often poor. This discrepancy highlights the need to consider post-transcriptional modifications and protein-level regulation, which cannot be assessed through genomics or transcriptomics alone.

The thesis further emphasizes the significance of proteomics in understanding cancer biology and identifying potential therapeutic targets. Proteins are the functional end products of the genome and are directly involved in cellular processes. Mass spectrometry-based proteomics enables the quantification of proteins, post-translational modifications, and pathway activity, providing valuable insights into the mechanisms of treatment resistance and the molecular subtyping of malignancies. The integration of proteomics into tumor profiling programs and molecular tumor boards can refine treatment recommendations and offer new therapeutic options, particularly for advanced-stage, rare, or pre-malignant cancers with limited treatment options.

Biomarker development is another important aspect discussed in the thesis. Traditional validation methods, such as antibody-based immunoassays, have limitations in terms of sensitivity, selectivity, and the ability to detect modified sequences. Mass spectrometry, particularly targeted proteomics using multiple reaction monitoring (MRM) or parallel reaction monitoring (PRM), is highlighted as a powerful tool for biomarker development. MS-based approaches offer accurate quantitation of molecular targets in a variety of specimens and dynamic ranges, bridging the gap between discovery and successful verification and validation of biomarkers.

This work argues that cancer is inherently a proteomic disease and quantitative proteomics is as important as the identification of genetic alterations and should be integrated into clinical decision making and molecular tumor boards. This work presents a streamlined and automatable procedure for quantitative proteomic analysis of clinical tissue samples, particularly formalin-fixed and paraffin-embedded (FFPE) specimens, which represents a significant advancement in the field of clinical proteomic research. The traditional hurdles of tedious sample preparation steps and long instrument acquisition times have been addressed, paving the way for the translation of quantitative proteomics into clinical practice. The method allows the analysis of limited volume clinical samples, such as core needle biopsies, for instance, collected over the course of disease management. This not only enables clinical research activities, including biomarker development and clinical trials, but also positions the method as a potential companion diagnostic device in a clinical setting. Its ability to process samples from extraction to analysis of clinical biomarkers within 24 hours from pathological assessment is a crucial advantage, facilitating rapid decision-making and personalized patient care.

Furthermore, the streamlined procedure does not require specialized equipment or expertise, making it accessible and feasible for implementation in various clinical laboratories. Its applicability has been demonstrated in integrative translational research conducted at the Segal Cancer Research Centre,[34-37] as well as by partners of the Marathon of Hope Cancer Centres Network and other research groups (publications in preparation), underlining its potential for widespread use. By bridging the gap between biomarker discovery and successful verification, proteomics emerges as a powerful tool within integrative translational research. The improved FFPE-proteomics procedure showcased in this thesis exemplifies the invaluable importance of proteomics in unlocking the proteomic secrets of diseases such as ductal carcinoma in situ, the most prevalent non-invasive breast lesion. Through both discovery and targeted approaches, quantitative proteomics not only confirms the genomic biomarkers identified in independent transcriptomic studies but also uncovers more than 380 additional proteomics and metabolic biomarkers. This wealth of information has the potential to inform much-needed strategies for disease management, particularly in pre-malignant lesions. The next step in this study would be the design of larger scale and more refined validation studies, taking hormone receptor status and other clinical variables into consideration. In addition, functional studies to examine the hypotheses and findings of this fundamental cancer study need to be designed and performed. Importantly, the work presented in this thesis provides an assay that not only demonstrates clinical validation but also meets regulatory requirements and can be readily incorporated into clinical practice guidelines. All -omics and non-omics data have been made available on recognized data-sharing platforms to improve accessibility of the translational research performed here. This further solidifies the clinical utility of the developed method and enhances its potential for real-world impact in improving patient outcomes.

The candidate has co-authored several articles reviewing current challenges in MS-based proteomics for biomarker development and developing targeted MS methods to advance clinical proteomics.

The article published in *Cancers*[34] is of particular interest as it points to a potential therapeutically relevant discordance between genomic and proteomic data in colorectal cancer liver metastases and represents the starting point of this PhD study. Notably, for one of the patients in this study, the proteomic data indicates that targeted cancer therapy might have been beneficial, even though genomic data alone excluded the patient from precision oncology approaches. [34] The validation of this intriguing result in a larger cohort was anticipated within this PhD project and might have important implications for future treatment decisions in colorectal cancer, leading to improved outcomes and providing promising therapeutic options for patients who currently have little hope of receiving state-of-the-art targeted therapies. The following paragraphs will discuss details of this project to highlight current limitations and challenges of multi-omics studies, and to present approaches the candidate has chosen to overcome these bottlenecks.

The candidate established an optimized and semi-automated proteogenomic pipeline which integrates state-of-the-art precise MS and genomic sequencing to quantify protein expression, including mutated proteins and protein phosphorylation in individual tumors. All sample preparation steps were systematically developed and optimized by the candidate to fit the purposes of this study and the limited sample volumes. New equipment was purchased, and methods were developed for label-free and TMT-labeled high-pH reversed-phase fractionation. A commercial TMT-labeling kit was optimized for a cost-effective, robust labeling of peptides for TMT-labeled larger-scale (phosho-)proteomics studies. A liquid handling robot was utilized for the development of an automated phosphopeptide enrichment method.

Collaborations with local clinical biobanks were established and a total of 97 fresh frozen surgical biopsies of metastatic liver lesions were selected by the candidate for the purposes of this study -- based on (i) RAS genotyping, the prevalent approved clinical biomarker for clinical decision making in colorectal cancer, (ii) the disease stage at time of collection, and (iii) the response to standard-of-care treatment. Cryosections of each biopsy, embedded in optimal cutting temperature (OCT) medium were prepared by the candidate and were stained with hematoxylin and eosin (H&E) for pathological examination. Of the recruited specimens, 42 were excluded due to non-compliance with preset thresholds, i.e., at least 30% tumor cellularity, less than 10% necrosis. Thus, 55 specimens were included for comprehensive proteogenomic analysis (discovery cohort). The study population was split into two study groups: (1) KRAS mutation negative (KRAS wt, n=32) and (2) KRAS mutation positive (KRAS mut, n=23). The patients in each study group were matched to age at time of diagnosis, sex, and sidedness of primary tumor, guided by findings from clinical trials conducted on colorectal cancer. From all clinical samples five 8-µm thick sections were prepared and stored along with H&E-stained sections for future analyses, such as validation studies and functional experiments for hypothesis testing.

A simultaneous extraction for DNA, proteins and metabolites from available clinical samples was not successful due to the low sample volumes. A collaboration with a leading company for targeted metabolomics assays was established by the candidate for future projects of the same type. The metabolomics team of the Borchers lab is currently developing untargeted tissue metabolomics pipelines, including data processing, which have proven to be difficult. An additional collaboration for AI-based metabolic profiling, used in the work presented in Chapter 3 of this thesis, was self-initiated by the candidate.

Although the genomic, proteomic, and phosphoproteomic analyses of this fundamental integrative cancer study have been completed, data analysis is challenging and requires specialized tools and statistical approaches. In particular, the genomic data analysis for the purposes of this study has proven to be difficult as somatic mutations identified by WES need to be translated into mutated protein sequences, including known protein isoforms. Next, signature proteotypic peptides need to be identified that are unique to the mutated protein and integrated into a 'genomic-centred' FASTA database. This database can then be used for proteomics data analysis to identify and quantify canonical (wildtype) and mutated proteins in the proteome of liver metastases from colorectal cancer. Recently published tools, such as QUILTS,[38] SeqTailor[39] and ProteoDisco[40] are either not suitable for this purpose or technical difficulties have been encountered. Collaborations with leading data scientists in the field had been established but had to be suspended due to restrictions beyond the candidate's control. The newly formed bioinformatics team at the Borchers lab is currently developing an algorithm for -omics data analysis which we hope will overcome these problems.

Preliminary data analysis conducted by the candidate, however, is intriguing and reveals a total of 659,748 exome altering events with an average alignment rate to the hg38 genome of 99.9%. In a preliminary data-processing approach, a genomic-centred FASTA database was created containing 33,009 mutated proteotypic peptides, unique to 11,610 proteins; 97.8% are missense mutations, and 2.2% are stop-gain mutations. An attempt to include stop-loss mutations to the database was not successful, because the resulting mutated protein sequence could not be determined. In the event of a stop-loss mutation, the mRNA transcript will end in a unique 3'UTR-sequence. This region is usually not translated into a protein sequence so we expect that the expressed protein would be truncated or nearly normal, if transcribed. The candidate has been working with the builders of SeqTailor,[39] a computational tool for the translation of genomics data into proteomic sequences and vice versa, to include this type of mutation in their algorithm. Although it is unclear whether stop-loss mutations are relevant for protein expression due to the lack of proteomics studies on this matter, stop-loss events have been described to be relevant for gene expression and cancer research by others.[41-43] Guided by these recent genomic findings, we therefore decided to include copy number variant (CNV) analysis of the genomics data, to aid in variant interpretation, and to estimate whether stop-loss mutations might be relevant for our study.

Data processing and analysis of this challenging multi-omics study is still ongoing, but first results show clear genomic and proteomic differences between the study groups of heavily pre-treated metastatic disease, for which treatment options have been exhausted. This work will present multi-layered information on mechanisms leading to acquired treatment resistance and will reveal new therapeutic strategies that would have not been accessible by either of the -omics fields alone.

In summary, this PhD thesis argues for a paradigm shift in clinical research towards multi-omics approaches, particularly the integration of proteomics data, in order to uncover therapeutic vulnerabilities in hard-to-treat malignancies, such as pre-malignant breast ductal carcinoma and treatment resistant metastatic disease in colorectal cancer. The limitations of genetic screening alone are widely acknowledged, and this thesis provides a comprehensive overview of the advancements in genomics, transcriptomics, and proteomics research. The importance of multi-omic data integration, biomarker development, and the potential of mass spectrometry-based proteomics are emphasized throughout the thesis. This research has significant implications for precision oncology and the development of personalized treatment strategies for patients with challenging malignancies.

References

1. Rodriguez, H., Zenklusen, J.C., Staudt, L.M., Doroshow, J.H. & Lowy, D.R. The next horizon in precision oncology: Proteogenomics to inform cancer diagnosis and treatment. Cell 184, 1661-1670 (2021).
2. Mani, D.R., et al. Cancer proteogenomics: current impact and future prospects. Nature Reviews Cancer 22, 298-313 (2022).
3. Jia, P. & Zhao, Z. Impacts of somatic mutations on gene expression: an association perspective. Brief. Bioinform. 18, 413-425 (2017).
4. Vasaikar, S., et al. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell 177, 1035-1049.e1019 (2019).
5. Alvarez, M.J., et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat. Genet. 48, 838-847 (2016).
6. Mertins, P., et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 534, 55-62 (2016).
7. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences 74, 5463-5467 (1977).
8. Anaparthy, N., Ho, Y.J., Martelotto, L., Hammell, M. & Hicks, J. Single-Cell Applications of Next-Generation Sequencing. Cold Spring Harb. Perspect. Med. 9(2019).
9. Gkazi, A. An Overview of Next-Generation Sequencing. in Technology Networks Genomics Research (2022).
10. Heather, J.M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1-8 (2016).
11. Gupta, N. & Verma, V.K. Next-Generation Sequencing and Its Application: Empowering in Public Health Beyond Reality. in Microbial Technology for the Welfare of Society (ed. Arora, P.K.) 313-341 (Springer Singapore, Singapore, 2019).
12. Patel, S.K., George, B. & Rai, V. Artificial Intelligence to Decode Cancer Mechanism: Beyond Patient Stratification for Precision Oncology. Front. Pharmacol. 11, 1177 (2020).
13. Barlas, S. The White House Launches a Cancer Moonshot: Despite Funding Questions, the Progress Appears Promising. P t 41, 290-295 (2016).
14. Noh, K.W., Buettner, R. & Klein, S. Shifting Gears in Precision Oncology-Challenges and Opportunities of Integrative Data Analysis. Biomolecules 11(2021).
15. U.S. Food and Drug Administration. Nucleic Acid Based Tests.

16. Wahjudi, L.W., et al. Integrating proteomics into precision oncology. Int. J. Cancer 148, 1438-1451 (2021).
17. Prasad, V. Perspective: The precision-oncology illusion. Nature 537, S63-S63 (2016).
18. Letai, A. Functional precision cancer medicine—moving beyond pure genomics. Nat. Med. 23, 1028-1035 (2017).
19. Kumar, D., et al. Integrating transcriptome and proteome profiling: Strategies and applications. Proteomics 16, 2533-2544 (2016).
20. National Human Genome Research Institute. Transcriptome Fact Sheet.
21. Bludau, I. & Aebersold, R. Proteomic and interactomic insights into the molecular basis of cell functional diversity. Nature Reviews Molecular Cell Biology 21, 327-340 (2020).
22. National Human Genome Research Institute. GENCODE Release (version 43). (02.2023).
23. Hong, M., et al. RNA sequencing: new technologies and applications in cancer research. J. Hematol. Oncol. 13, 166 (2020).
24. Ghazalpour, A., et al. Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet. 7, e1001393 (2011).
25. Paolillo, C., Londin, E. & Fortina, P. Next generation sequencing in cancer: opportunities and challenges for precision cancer medicine. Scandinavian Journal of Clinical and Laboratory Investigation 76, S84-S91 (2016).
26. Roux, P.P. & Topisirovic, I. Signaling Pathways Involved in the Regulation of mRNA Translation. Mol Cell Biol 38(2018).
27. Ebhardt, H.A., Root, A., Sander, C. & Aebersold, R. Applications of targeted proteomics in systems biology and translational medicine. Proteomics 15, 3193-3208 (2015).
28. Ryu, J. & Thomas, S.N. Quantitative Mass Spectrometry-Based Proteomics for Biomarker Development in Ovarian Cancer. Molecules 26(2021).
29. Zhang, B., et al. Clinical potential of mass spectrometry-based proteogenomics. Nature Reviews Clinical Oncology 16, 256-268 (2019).
30. Vizcaino, J.A., et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. 32, 223-226 (2014).
31. Pino, L.K., Rose, J., O'Broin, A., Shah, S. & Schilling, B. Emerging mass spectrometry-based proteomics methodologies for novel biomedical applications. Biochem. Soc. Trans. 48, 1953-1966 (2020).
32. Vasilopoulou, C.G., et al. Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts. Nature communications 11, 331 (2020).

33. van Bentum, M. & Selbach, M. An Introduction to Advanced Targeted Acquisition Methods. Mol. Cell. Proteomics 20(2021).

34. Blank-Landeshammer, B., et al. Proteogenomics of Colorectal Cancer Liver Metastases: Complementing Precision Oncology with Phenotypic Data. Cancers (Basel) 11(2019).

35. Constance Sobsey, B.F., Georgia Mitsa, Sahar Ibrahim, Rene Zahedi, Elza d Bruin, Christoph Borchers, Gerald Batist. Proteomic analysis of PIK3CA-mutated tumours identifies protein networks correlated with clinical benefit of capivasertib genetically pre-selected patients. Clinical Cancer Research (2023).

36. Ibrahim, S., et al. Precise Quantitation of PTEN by Immuno-MRM: A Tool To Resolve the Breast Cancer Biomarker Controversy. Anal. Chem. 93, 10816-10824 (2021).

37. Sahar Ibrahim, G.M., Cathy Lan, Catherine Chabot, Marguerite Bu-chanan, Adriana Aguilar-Mahecha, Mounib Elchebly, Oliver Poetz, Alan Spatz, Mark Basik, Gerald Batist, René P. Zahedi, Christoph H. Borchers. Precise quantitation of PTEN by immuno-MRM: a tool to resolve the breast cancer biomarker controversy. Anal. Chem. (2021).

38. Ruggles, K.V., et al. An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. Mol Cell Proteomics 15, 1060-1071 (2016).

39. Zhang, P., et al. SeqTailor: a user-friendly webserver for the extraction of DNA or protein sequences from next-generation sequencing data. Nucleic Acids Res. 47, W623-W631 (2019).

40. van de Geer, W.S., van Riet, J. & van de Werken, H.J.G. ProteoDisco: a flexible R approach to generate customized protein databases for extended search space of novel and variant proteins in proteogenomic studies. Bioinformatics 38, 1437-1439 (2022).

41. Dhamija, S., et al. A pan-cancer analysis reveals nonstop extension mutations causing SMAD4 tumour suppressor degradation. Nat. Cell Biol. 22, 999-1010 (2020).

42. Waarts, M.R., Stonestrom, A.J., Park, Y.C. & Levine, R.L. Targeting mutations in cancer. J. Clin. Invest. 132(2022).

43. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 39, e118-e118 (2011).