

# Dissertation

submitted to the  
Combined Faculty of Natural Sciences and Mathematics  
of the Ruperto Carola University Heidelberg, Germany  
for the degree of

**Doctor of Natural Sciences**

Presented by  
Torsten Michael Müller  
M.Sc. Biomedical Science and Technology  
born in Idar-Oberstein, Germany

Oral examination on the 2<sup>nd</sup> of November 2020



# Automated sample preparation for streamlined proteomic profiling of clinical specimens

Referees:

Prof. Dr. Britta Brügger

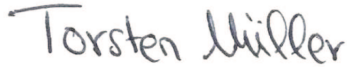
Prof. Dr. Jeroen Krijgsveld





## Declaration

Herewith I declare that I have written and submitted this dissertation myself and, in this process, have not used any other sources than those indicated. I hereby declare that I have not applied to be examined at any other institution, nor have I used this dissertation in this or any other format any other institution as an examination paper, nor submitted it to any other faculty as a dissertation

  
-----

Torsten Michael Müller



## Abstract

The genetic information of all life is encoded within DNA molecules that are translated into functional entities, so-called proteins. They are responsible for operating and controlling a vast array of molecular mechanisms in any biological system and ubiquitous in (patho)physiology as a result. Besides, proteins are the primary target of drugs and can have a central role as biomarkers for diagnostic, prognostic, or predictive purposes. Here, many regulatory mechanisms and spatiotemporal influences prevent an accurate prediction of a proteins' abundance and its associated functionality based on the genome information alone. Nowadays, it has become possible to measure and quantify thousands of proteins simultaneously, however, involving comprehensive sample preparation procedures. Currently, no universally standardized method enables a routine application of proteome profiling in a clinical environment.

In this thesis, an automated workflow for the efficient processing of the most common and quantity-limited specimens is described. In order to demonstrate the usefulness of the end-to-end pipeline, which was termed autoSP3, it was applied to the proteome profiling of histologically defined and WHO recognized growth patterns of pulmonary adenocarcinoma (ADC) that currently have a limited clinical implication. *Secondly*, we investigated the proteome composition of a molecularly well-defined cohort of Ependymoma (EPN) pediatric brain tumors. Despite the availability of substantial NGS data and their ability to differentiate nine distinct subgroups, the majority of tumors remained without a functional insight. Here, the proteome profiling could provide a missing link and emphasize several subgroup-specific protein targets.

In summary, this thesis describes the optimization of SP3 and its automation into a robust and cost-efficient pipeline for quantity-limited sample preparation and biological insight into the proteome composition of ADC growth patterns and EPN tumor subgroups.



## Zusammenfassung

Die genetische Information, welche in der DNS eines jeden Lebewesen's codiert ist, wird übertragen in funktionellen Einheiten, so genannte Proteine. Diese sind verantwortlich fuer den Betrieb und die Kontrolle zahlreicher molekularer Mechanismen in jedem biologischen System. Dadurch sind Proteine allgegenwärtig in der (Patho)-physiologie. Zusätzlich sind Proteine der Hauptangriffspunkt der meisten klinischen Arzneimittel und sie können eine zentrale Rolle als Biomarker fuer diagnostische, prognostische oder prädiktive Zwecke einnehmen. Da die Abundanz eines jeden Proteins und die damit zusammenhängende Funktion von zahlreichen regulatorischen Mechanisms sowie räumlichen und zeitlichen Faktoren abhängt, ist es kaum möglich diese allein anhand der genetischen Information vorherzusagen. Heutzutage ist es möglich tausende von Proteinen gleichzeitig zu messen und zu quantifizieren. Bisher gibt es allerdings keine universelle und standardisierte Methode, welche eine routinierte Anwendung in einem klinischen Umfeld ermöglichen würde.

In dieser Doktorarbeit wird eine automatisierte Methode zur effizienten Prozessierung der am häufigsten verwendeten und mengenlimitierten Proben typen beschrieben. Um die allgemeine Nützlichkeit dieser Methode zu demonstrieren, welche autoSP3 genannt wurde, wurde sie in zwei realistischen Szenarios angewendet. Zunächst wurde sie verwendet um Unterschiede in der Proteinzusammensetzung von Lungenkarzinomen mit verschiedenen Wachstumsmustern zu untersuchen, welche nach WHO Richtlinien histologisch klassifiziert wurden. Darüber hinaus wurde eine Kohorte von Ependymoma (EPN) Gehirntumoren, welche bei Kindern und Jugendlichen vorkommen, auf Ihre Proteinzusammensetzung untersucht. Bisher konnten diese basierend auf NGS Daten in neun individuelle Untergruppen klassifiziert werden, aber für die Mehrheit existiert bisher keine funktionelle Erklärung. Die tumorspezifischen Proteinprofile bieten die Möglichkeit potentielle Ursachen, Mechanismen oder Angriffspunkte für Therapien aufzudecken.

Zusammenfassend beschreibt diese Doktorarbeit die Optimierung und Automatisierung von SP3 zu einem robusten und kosteneffizienten Prozess sowie dessen Anwendung und daraus folgende biologische Erkenntnisse zu Lungenkarzinom-Wachstumsmustern und EPN-Gehirntumoren.



## List of Abbreviations

<b>ABC</b>	Ammonium bicarbonate
<b>ACN</b>	Acetonitrile
<b>ADC</b>	Adenocarcinoma
<b>AFA</b>	Adaptive focused acoustics
<b>ATRTs</b>	Atypical teratoid rhabdoid tumors
<b>autoSP3</b>	Automated single-pot, solid-phase-enhance sample preparation
<b>BBB</b>	Blood-brain barrier
<b>BCA</b>	Bicinchoninic acid assay
<b>BMBF</b>	Federal Ministry for Education and Research
<b>CAA</b>	Chloroacetamide
<b>ChIP-seq</b>	Chromatin immunoprecipitation-sequencing
<b>CIMP</b>	CpG island methylator phenotype
<b>CNS</b>	Central nervous system
<b>CNV</b>	Copy-number variation
<b>CSF</b>	Cerebrospinal fluid
<b>ctDNA</b>	Circulating tumor DNA
<b>CV</b>	Coefficient of variation
<b>DMEM</b>	Dulbecco's modified Eagle's medium
<b>DTT</b>	Dithiothreitol
<b>EASyM</b>	European Association of Systems Medicine e.V
<b>ECM</b>	Extracellular matrix
<b>EDTA</b>	Ethylenediaminetetraacetic acid
<b>EGFR</b>	Epidermal growth factor receptor
<b>EPN</b>	Ependymoma
<b>ER</b>	Endoplasmic reticulum
<b>ESI</b>	Electrospray ionization
<b>EtOH</b>	Ethanol
<b>EV</b>	Extracellular vesicle
<b>FA</b>	Formic acid
<b>FASP</b>	Filter-aided sample preparation
<b>FBS</b>	Fetal Bovine serum
<b>FFPE</b>	Formalin-fixed and paraffin-embedded
<b>FWHM</b>	Full width half maximum
<b>GDPR</b>	General data protection regulations
<b>GIC</b>	Glioma initiating cell

## List of Abbreviations

<b>GO</b>	Gene ontology
<b>GRAVY</b>	Grand average of hydropathy
<b>GSEA</b>	Gene set enrichment analysis
<b>GST</b>	General systems theory
<b>GTR</b>	Gross total resection
<b>H&amp;E staining</b>	Hematoxylin and Eosin staining
<b>H3K27</b>	Histone 3 lysine 27
<b>H3K27me3</b>	Histone 3 lysine 27 trimethylation
<b>HCD</b>	Higher-energy collision-activated dissociation
<b>HILIC</b>	Hydrophilic interaction chromatography
<b>HIPPA</b>	Health insurance and portability and accountability act
<b>iBAQ</b>	Intensity-based absolute Quantification
<b>ICD-11</b>	WHO international classification of diseases
<b>ICD-O</b>	International classification of diseases with focus on oncology
<b>InDels</b>	Small insertions and deletions of nucleotide sequences
<b>LC</b>	Liquid Chromatography
<b>LC-MS</b>	Liquid Chromatography coupled to Mass Spectrometry
<b>LFQ</b>	Label-free quantification
<b>MALDI</b>	Matrix-assisted laser desorption ionization
<b>MOFA</b>	Multi-omics factor analysis
<b>MRI</b>	Magnetic resonance imaging
<b>MS</b>	Mass spectrometry
<b>MS2</b>	Tandem mass spectrometry
<b>MVP</b>	Multivesicular bodies
<b>NGS</b>	Next-generation sequencing
<b>NHGRI</b>	National Human Genome Research Institute
<b>NMR</b>	Nuclear magnetic resonance
<b>NSCLC</b>	Non-small-cell Lung Cancer
<b>NTA</b>	Nanoparticle tracking analysis
<b>PBS</b>	Phosphate-buffered saline
<b>PCA</b>	Principal component analysis
<b>PCR</b>	Polymerase chain reaction
<b>PF-EPN-A</b>	Posterior fossa ependymoma A
<b>PF-EPN-B</b>	Posterior fossa ependymoma B
<b>PFS</b>	Progression-free survival
<b>PF-SE</b>	Posterior fossa sub-ependymoma
<b>PIC</b>	Protease-inhibitor cocktail
<b>PRC2</b>	Polycomb repressive complex 2



<b>PTM</b>	Post-translational modification
<b>QE HF</b>	Q-Exactive High Field Mass Spectrometer
<b>RIPA</b>	Radioimmunoprecipitation assay
<b>RSLC</b>	Rapid Separation Liquid Chromatography
<b>RT-PCR</b>	Real time-Polymerase chain reaction
<b>SCLC</b>	Small-cell Lung Cancer
<b>SDS</b>	Sodium dodecylsulfate
<b>SE</b>	Subependymoma
<b>SNV</b>	Single nucleotide variation
<b>SP3</b>	Single-pot, solid-phase-enhanced sample preparation
<b>SP-EPN</b>	Spine ependymoma
<b>SP-MPE</b>	Spine myxopapillary ependymoma
<b>SP-SE</b>	Spine Sub-ependymoma
<b>ST-EPN-RELA</b>	Supratentorial ependymoma RELA-fusion positive
<b>ST-EPN-YAP1</b>	Supratentorial ependymoma YAP1-fusion positive
<b>STR</b>	Subtotal resection
<b>ST-SE</b>	Supratentorial sub-ependymoma
<b>SV</b>	Large structural variation
<b>T2</b>	Top 2 most abundant precursor ions trigger MS2
<b>T20</b>	Top 20 most abundant precursor ions trigger MS2
<b>TCEP</b>	Tris(2-carboxyethyl)phosphine
<b>TEM</b>	Transmission electron microscope
<b>TFA</b>	Trifluoroacetic acid
<b>TMT</b>	Tandem mass-tag
<b>t-SNE</b>	t-distributed stochastic neighbor embedding
<b>WES</b>	Whole-exome sequencing
<b>WGS</b>	Whole-genome sequencing
<b>WHO</b>	World Health Organization
<b>μPAC</b>	Micro pillar-array column



# Table of Content

<b>DECLARATION</b> .....	<b>I</b>
<b>ABSTRACT</b> .....	<b>III</b>
<b>ZUSAMMENFASSUNG</b> .....	<b>V</b>
<b>LIST OF ABBREVIATIONS</b> .....	<b>VII</b>
<b>TABLE OF CONTENT</b> .....	<b>XI</b>
<b>1. INTRODUCTION</b> .....	<b>1</b>
1.1. SYSTEMS MEDICINE .....	1
1.1.1. <i>Definition and emergence</i> .....	1
1.1.2. <i>Next-generation sequencing (NGS): status quo in the clinic</i> .....	3
1.2. MASS SPECTROMETRY-BASED PROTEOME PROFILING .....	6
1.2.1. <i>Definition and emergence</i> .....	6
1.2.2. <i>Status quo: proteomics in systems medicine</i> .....	8
1.2.3. <i>Clinical translation: bottlenecks and requirements</i> .....	10
1.3. ONCOLOGY .....	13
1.3.1. <i>Cancer: definition and epidemiology</i> .....	13
1.3.2. <i>Carcinogenesis</i> .....	14
1.3.3. <i>Lung cancer</i> .....	16
1.3.4. <i>Pediatric brain tumor: ependymoma</i> .....	18
1.3.4.1. <i>Definition and epidemiology</i> .....	18
1.3.4.2. <i>Molecular classification</i> .....	20
1.4. AIM OF THIS STUDY .....	24
<b>2. MATERIALS</b> .....	<b>25</b>
2.1. CHEMICALS AND OTHER MATERIALS.....	25
<b>3. EXPERIMENTAL METHODS</b> .....	<b>32</b>
3.1. MASS SPECTROMETRY METHODS .....	32
3.1.1. <i>Liquid chromatography column setup</i> .....	32
3.1.2. <i>Liquid chromatography gradients and data-dependent acquisition (DDA)</i> .....	32
3.1.3. <i>Proteomics data processing</i> .....	34

## Table of Content

3.2.	METHODS TAKEN FROM JOINT PUBLICATIONS.....	35
3.2.1.	<i>Methods taken from “Single-pot, solid-phase-enhanced sample preparation for proteomics experiments.”</i> .....	36
3.2.1.1.	Single-pot, solid-phase-enhanced sample preparation (SP3) bead preparation .....	36
3.2.1.2.	SP3 protein clean-up.....	36
3.2.1.3.	SP3 peptide clean-up.....	37
3.2.2.	<i>Methods taken from “Automated sample preparation with SP3 for low-input clinical proteomics.”</i> .....	37
3.2.2.1.	Cell culture of HeLa cells.....	37
3.2.2.2.	HeLa protein standard preparation .....	38
3.2.2.3.	Pulmonary adenocarcinoma (ADC) sample collection.....	38
3.2.2.4.	Automated SP3 protocol (autoSP3) .....	39
3.2.2.5.	Quantitative proteomics analysis of FFPE tissue .....	40
3.2.2.6.	Proteomics data acquisition .....	41
3.2.2.7.	Proteomics data processing.....	42
3.2.2.8.	Intra-day and inter-day precision .....	42
3.2.2.9.	Sensitivity of autoSP3 .....	43
3.2.2.10.	Assessment of cross-contamination.....	43
3.2.3.	<i>Methods taken from „EZHIP/CXorf67 mimics K27M mutated oncohistones and functions as an intrinsic inhibitor of PRC2 function in aggressive posterior fossa ependymoma.”</i> .....	43
3.2.3.1.	Cell Culture of HEK293T cells.....	43
3.2.3.2.	Production of Lentiviral particles and generation of stable cell lines .....	44
3.2.3.3.	Co-Immunoprecipitation for mass spectrometry and western blot .....	44
3.2.3.4.	Nuclear extraction and western blot analysis.....	45
3.2.3.5.	Protein digestion and SP3 peptide clean-up of Co-IP samples .....	45
3.2.3.6.	Mass spectrometry data acquisition.....	46
3.2.3.7.	Mass spectrometry data processing.....	46
3.3.	ADDITIONAL EXPERIMENTAL METHODS .....	47
3.3.1.	<i>Cell culture of stable cell lines</i> .....	47
3.3.2.	<i>Cell culture of patient-derived EPN tumor cell lines</i> .....	48
3.3.3.	<i>Additional methods for lysis of cells and tissue for protein extraction</i> .....	49
3.3.3.1.	RapiGest SF Surfactant protein extraction.....	49
3.3.3.2.	Urea-based protein extraction .....	49
3.3.4.	<i>DNA extraction</i> .....	50
3.3.5.	<i>E. coli spike-in sample preparation</i> .....	51
3.3.6.	<i>Agarose Gels for DNA visualization</i> .....	51
3.3.7.	<i>SDS-Gels for protein visualization</i> .....	51
3.3.8.	<i>Cell-surface labeling and protein enrichment</i> .....	52
3.3.9.	<i>Desalting and clean-up of peptide samples</i> .....	54
3.3.10.	<i>High pH reversed-phase fractionation of proteomic samples</i> .....	54

3.3.11.	<i>Extracellular vesicle sample preparation</i> .....	56
3.3.11.1.	Patient-derived EPN cell culture .....	56
3.3.11.2.	Exosome and microvesicle isolation.....	56
3.3.11.3.	Immunogold electron microscopy .....	57
3.3.11.4.	Nanoparticle tracking analysis (NTA) .....	58
3.3.11.5.	Qubit protein quantification assay.....	58
3.3.11.6.	Protein digestion and SP3 protein clean-up of EV samples.....	58
3.4.	ADDITIONAL DATA ANALYSIS .....	59
3.4.1.	<i>Differential expression</i> .....	59
3.4.2.	<i>Gene set enrichment analysis (GSEA) and gene ontology (GO)</i> .....	59
3.4.3.	<i>t-SNE and umap</i> .....	59
3.4.4.	<i>Gene- and protein expression correlation</i> .....	59
3.4.5.	<i>Copy number variation (CNV) correlation to gene- and protein expression</i> .....	60
3.4.6.	<i>Multi-omics factor analysis (MOFA)</i> .....	60
<b>4.</b>	<b>RESULTS</b> .....	<b>61</b>
	COPYRIGHT DISCLAIMER .....	61
4.1.	PROTEOMIC PROFILING IN SYSTEMS MEDICINE .....	62
4.1.1.	<i>Cell- or tissue lysis and protein extraction</i> .....	62
4.1.2.	<i>Single-pot, solid-phase-enhanced sample preparation (SP3)</i> .....	66
4.1.2.1.	Optimization of protein binding.....	67
4.1.2.2.	Capacity and reproducibility of SP3 .....	69
4.1.3.	<i>LC-MS data acquisition</i> .....	71
4.1.3.1.	Library generation and LC optimization .....	72
4.1.3.2.	Match-between-runs and optimal quantification .....	74
4.2.	AUTOMATED SP3 (AUTO-SP3) .....	78
4.2.1.	<i>Establishment of autoSP3: generic sample preparation</i> .....	79
4.2.2.	<i>Evaluation of autoSP3 precision</i> .....	84
4.2.3.	<i>Assessment of autoSP3 sensitivity</i> .....	88
4.2.4.	<i>autoSP3 and challenging specimens</i> .....	90
4.2.4.1.	End-to-end sample preparation .....	91
4.2.4.2.	Pulmonary adenocarcinoma (ADC) FFPE.....	92
4.3.	EPENDYMOMA (EPN) BRAIN TUMORS.....	96
4.3.1.	<i>Proteome profiles of molecular subgroups</i> .....	97
4.3.2.	<i>Subgroup-specific putative marker proteins</i> .....	99
4.3.2.1.	CXorf67 (EZHIP): an intrinsic inhibitor of PRC2 in PF-EPN-A.....	99
4.3.3.	<i>Protein- and gene expression</i> .....	102
4.3.3.1.	Signature gene translation to proteins.....	106

## Table of Content

4.3.3.2.	Differential expression determines signature proteins .....	109
4.3.3.3.	Genetic structural aberrations (CNVs) to phenotype .....	113
4.3.3.4.	ST-EPN-RELA cell surface-proteome sub-classification .....	116
4.3.4.	<i>Exosome cargo characterization in ST-EPN-RELA</i> .....	120
4.3.5.	<i>Perspective view on multi-omics data integration</i> .....	124
<b>5.</b>	<b>DISCUSSION</b> .....	<b>127</b>
5.1.	LARGE-SCALE PROTEOME PROFILING ENABLED BY AUTOSP3 .....	127
5.2.	THE ADDED VALUE OF PROTEOME PROFILING .....	132
5.2.1.	<i>Molecular characterization of lung adenocarcinoma (ADC) growth patterns</i> .....	132
5.2.2.	<i>Proteome profiles extent ependymoma (EPN) molecular classification</i> .....	133
5.3.	RE-EVALUATION OF THE STATUS QUO: CLINICAL PROTEOMICS .....	136
<b>6.</b>	<b>REFERENCES</b> .....	<b>141</b>
<b>7.</b>	<b>ACKNOWLEDGMENTS</b> .....	<b>155</b>
<b>8.</b>	<b>TABLE OF FIGURES</b> .....	<b>157</b>
<b>9.</b>	<b>SUPPLEMENTARY FIGURES</b> .....	<b>159</b>

# 1. Introduction

## 1.1. Systems Medicine

### 1.1.1. Definition and emergence

The interdisciplinary field of systems medicine emanates from the translation of systems biology to medical research and routine clinical practice<sup>1,2</sup>. It is a systems-oriented approach that aims to combine the multifaceted network of factors (e.g., genes, transcripts, proteins, metabolites, family history, and environmental factors) that define and influence the function and development of the human body to improve disease diagnostics, prognostics, and to develop innovative therapies<sup>3,4</sup>.

Early phases of most scientific fields were coined mainly by the concept of reductionism, in which one attempts to reduce every instance of a system to its individual, constituent parts<sup>5,6</sup>. It hypothesizes that understanding the simple parts suffices to draw upward causation to explain all overarching phenomena or mechanisms that are crucial for understanding the whole system itself<sup>5,6</sup>. This method has been successful in physics and chemistry because physicochemical properties and resulting physical laws down to the atomic level can explain most problems or questions. Then and now, reductionism models were also used to explain many fundamentals in biology to understand living processes, and it remains to be the predominant concept in classical medicine approaches<sup>7-9</sup>. Here, clinicians aim to break down a problem or disease phenotype to its single-cause, which has proven utility in cases where an individual factor, such as a bacterial infection, is responsible for the disease. However, this concept quickly becomes challenged by I) the heterogeneity of most diseases<sup>10</sup>, such as cancer, with a complex patho-phenotype and without a single causative factor, II) the sheer complexity of biological systems, organisms or patients, being comprised of several networks of factors, signaling pathways, multi-layer interaction, and dynamic spatiotemporal features, and III) the variable influence of environmental factors<sup>7,11</sup>. Taken together, this complex and dynamic interaction of multi-layer factors within a system and with its environment leads to a yet non-predictable behavior that cannot be explained by the individual parts of a system<sup>7,12</sup>.

The foresight about the importance of a systems-oriented approach dates back to Aristotle, one of the first inquiring minds in philosophical and scientific history, who wrote: “*The*

## Introduction

*totality is not, as it were, a mere heap, but the whole is something besides the parts.*" (Aristotle's *Metaphysics: Book VIII, 1045a.8–10*)<sup>13</sup>. His primal understanding of the essential complexity of biological systems can be seen as the philosophical foundation of modern systems biology as it emerged throughout the 20<sup>th</sup> century<sup>9</sup>. In 1926, Jan Christian Smuts, a South African statesman and philosopher, introduced the concept of holism antithetic to reductionism, which in a broader sense of Aristotle's words states that a system is more than the sum of its parts<sup>14,15</sup>. The theory was further extended at the end of the 1960s by Ludwig von Bertalanffy, who is thought to have created the term of systems biology by introducing his general systems theory (GST)<sup>9,16</sup>. In its essence, he describes that every system is composed of the sum of its structure and functional purpose, the environmental and temporal influences, and its spacial boundaries<sup>9,16</sup>. This steady change to a systems-oriented mindset was manifested and driven by numerous breakthrough discoveries throughout the 20<sup>th</sup> century, including the discovery that DNA makes up the genetic material of the chromosome in 1944 (Oswald Avery)<sup>17</sup>, the finding of the structure of DNA in 1953 (James Watson, Francis Crick, and Rosalind Franklin)<sup>18–20</sup>, and the Sanger sequencing technology providing the first DNA genome of an entire organism in 1977 (Frederick Sanger)<sup>21</sup>.

Nowadays, systems biology has matured into its own independent, inter-disciplinary field<sup>22–24</sup>. It involves not only modern analytical technologies to generate comprehensive data but also gained momentum through the rapid development of computational hardware and software, providing the performance capacity and mathematical models to store, handle, and analyze the data<sup>24</sup>. This advanced technological toolbox of systems biology approaches enables the study of (patho)physiological processes on a complex molecular level beyond single, linear parameters<sup>22</sup>. Thus, the complexity of biological systems and the sheer amount of qualitative and quantitative data became impossible to handle without a systems biology-based approach<sup>22–24</sup>. In a clinical context, the integration of systems biology marks the beginning of systems medicine as a new discipline<sup>10</sup>. Since then, it is rapidly evolving and growing to an integration of all fields of expertise from bioinformatics- and statistics to basic biology research, mathematics, and classical medicine to generate, collect, and interpret data comprising molecular, behavioral (lifestyle), environmental, as well as family data.



Systems medicine, first introduced by B.Z. Zeng<sup>25</sup> and T. Kamada<sup>26</sup> in 1992, promises to surpass the limits of reductionism and leverages our understanding of (patho)physiological processes to positively impact therapy development and decision making, diagnostics, and prognostics<sup>7,27</sup>. It was Hippocrates of Kos (460-377 B.C.) who said: *“It is more important to know what sort of person has a disease than to know what sort of disease a person has”*<sup>28</sup>. Thus, he recognized early on the importance of the individual and the necessity of a personalized medicine approach, including tailored diagnostics and treatment<sup>29</sup>. To achieve this, the translation of modern technologies towards clinics and tackling remaining bottlenecks are gaining momentum since the beginning of the new century<sup>27,30</sup>. Systems medicine promises to not only measurably improve patients’ health and treatment outcomes but also offers aid in improving the efficacy of drug discovery and development through better disease and patient characterization<sup>27,30,31</sup>. The future of medicine is at a tipping point.

### **1.1.2. Next-generation sequencing (NGS): status quo in the clinic**

After the discovery of the DNA structure in 1953<sup>18–20</sup>, it took more than two decades, until 1977, for the development of first sequencing methods, namely the Maxam & Gilbert sequencing<sup>32</sup> and the Sanger sequencing<sup>21</sup>. The latter of the two became the predominant method for the following years up until today, because of less handling requirements of toxic chemicals and radioisotopes<sup>33</sup>. Since then, Sanger sequencing technologies have rapidly evolved at an unprecedented speed with the immense support of the Human Genome Project, which initiated in 1990<sup>33</sup>. As the largest collaborative biological project, it resulted in the completion of the first human genome sequence in 2004, a significant milestone for the field of systems biology<sup>34,35</sup>. However, the sequencing of an entire genome required extensive amounts of time and resources, which lead to a new funding initiative by the National Human Genome Research Institute (NHGRI) aiming to reduce the human genome sequencing cost to about 1000 US-Dollar within the next ten years<sup>36</sup>. This call has led to an explosion of “next-generation” sequencing (NGS) methods, including faster instruments, chemicals, tools, bioinformatics data analysis, and protocols within the last 15 years that enable scientists to address all sorts of basic genetics or clinically relevant questions<sup>37–44</sup>. NGS has become the mainstream acronym for essentially every very-high-throughput sequencing technique or methods involving sequencing, that allow millions of observations

## Introduction

in a single run at significantly reduced costs as compared to 2004<sup>33</sup>. In more recent years, third-generation instruments and methods (also considered NGS) were commercialized that are even faster and more accurate, that require less DNA or RNA input material, generate lower error rates and fewer artifacts, at lower costs<sup>45-47</sup>. A higher standardization and automated sample handling, as well as improved bioinformatic tools and pipelines, continuously support these developments<sup>33,45</sup>. Taken together, scientists nowadays have a vast toolbox of NGS applications that enable the study of entire genomes (DNA-seq)<sup>34,35</sup>, the transcriptome (RNA-seq)<sup>48</sup>, and targeted fractions of both beyond the determination of a nucleotide sequence<sup>40,47,49,50</sup>.

In 2010, the first large-scale study of human genetic variation was published, providing evolutionary insights at a population scale to understand the impact of genetic differences on our (patho)physiology<sup>51</sup>. This massive sequencing effort extends our knowledge of the functional consequences of mutations, providing a link between genotype and phenotype concerning health and disease. This type of study has the potential to improve the precision of diagnosis, the classification of a disease state or subtype, and to provide accurate prognosis or even identify potentially druggable mutations for individual patients<sup>51</sup>. In a study by Ashley, E.A. et al. in 2010, for example, the genome of a patient with a family history of vascular disease and sudden early death was assessed<sup>52</sup>. The whole-genome sequencing (WGS) analysis and integration with clinical features pointed to an increased genetic risk for myocardial infarction and discovered three rare gene variants that are clinically associated with sudden cardiac death. Thus, relevant and personalized information could be retrieved using WGS analysis<sup>52</sup>.

In practice, genetic disease diagnostics is focused on panels of genes that are sequenced, which are associated with a clinical phenotype<sup>53,54</sup>. This approach is limited to well-characterized monogenic illnesses in which a single mutation can explain the disease<sup>55-58</sup>. For less straight-forward applications, the WGS<sup>52-54</sup> or targeted approach<sup>55-58</sup> is neither necessary nor beneficial and replaced by whole-exome sequencing (WES)<sup>59,60</sup>, in which solely the coding regions of the genome are sequenced. The human exome represents a small fraction of the entire genome but comprises >85% of disease-causing gene variants<sup>61,62</sup>. In 2011, Worthey E.A. et al. used WES to identify a causative mutation in a male child with Crohn disease-like illness where comprehensive clinical evaluation did not

yield in a definitive diagnosis<sup>63</sup>. The sequencing analysis identified a novel missense mutation in the X-linked inhibitor of apoptosis gene, which was previously not correlated to Crohn disease. That followed, the child was diagnosed with an X-linked inhibitor of apoptosis deficiency and treated in concordance with the respective recommendation guidelines<sup>63</sup>. Thus, the exome sequencing led to a valuable, life-saving therapeutic decision, highlighting the potential of NGS approaches in a clinical setup beyond standard diagnostics.

Furthermore, NGS enables the identification of single nucleotide variants<sup>64</sup>, for example, somatic or germline mutations and structural variants<sup>65-68</sup>, such as inversions, translocations, or gene copy number alterations. Other sequencing-based gene expression analysis (RNA-seq) enable the identification and quantification of rare transcripts, alternative splicing variants, or newly synthesized (nascent) transcripts<sup>48,69</sup>. Techniques to profile protein-DNA interactions, using chromatin immunoprecipitation followed by sequencing (ChIP-seq)<sup>70,71</sup>, and epigenetic marks<sup>72,73</sup>, have been developed driven by the continuous evolution of the NGS field<sup>33</sup>. The latter, epigenetics, defines functionally relevant changes to the genome without an alteration in the nucleotide sequence<sup>74</sup>. In particular, the epigenome of DNA methylation, a mark for silencing of transcription, has been extensively studied using methods such as bisulfite sequencing<sup>73,75</sup>. Here, the DNA is treated with bisulfite before routine sequencing, which leads to a conversion of the base cytosine to uracil. The majority of DNA methylation events occur at cytosine and remain unaffected during the bisulfite treatment, leading to methylation status dependent alterations of the nucleotide sequence that can be readout<sup>50,76</sup>.

The impressive developments in NGS technologies and the ever-increasing number of applications and molecular profiling studies highlight the potential impact in improving a patient's health<sup>38,40,42</sup>. The analysis of cell-free, circulating DNA isolated from liquid biopsies, for example, offers an attractive, low-invasive approach for the discovery of disease biomarkers<sup>77,78</sup>. The genetic information contained in circulating tumor DNA (ctDNA) might be relevant for cancer diagnostics, progression or relapse monitoring, and guiding therapy decisions<sup>78</sup>. Altogether, NGS promises improvements in patient stratification, risk assessment for genetic diseases, and the capability to identify multiple mutations in a variety of cases, such as oncology<sup>38,42</sup>. This holds for the broad field of

## Introduction

oncology<sup>79</sup>, but also other diseases, such as Parkinson's disease<sup>80</sup>, Alzheimer's disease<sup>81</sup>, and cardiovascular diseases<sup>82</sup>.

Nevertheless, neither WGS nor WES or RNA-seq is currently established in routine diagnostics, and the implementation is only slowly progressing with a few examples of clinical translation<sup>83,84</sup>. Known obstacles for clinical translation are false positive and false negative results as well as low sensitivity in the detection of early-stage cancer<sup>84</sup>. Other limitations are practical demands, such as fast turn-around times from receiving patient samples to providing analyzed data, the establishment of necessary infrastructure, and the overall costs<sup>84</sup>. Besides, the sheer amount of data, the requirements to ensure high data quality and reproducibility, the associated bioinformatic analysis, the handling and storing of data, the medically relevant interpretation, and the clarification of ethical concerns are significant challenges for a successful implementation of any systems medicine approach that need systematic problem-solving<sup>85,86</sup>. Other obvious bottlenecks of NGS profiling are I) the challenges in distinguishing between a driver and a passenger mutation (explained in chapter 1.3.2.)<sup>87</sup>, or II) the low correlation of gene expression and protein expression and the consequential phenotype<sup>88-90</sup>. This is because proteins are dynamically regulated by numerous post-transcriptional mechanisms and post-translational modifications (PTMs), such as phosphorylation<sup>88-90</sup>. The function of a protein additionally depends on its subcellular localization<sup>91</sup>, the interaction of proteins in complexes<sup>92</sup>, or their half-life's<sup>93</sup>, which explains why transcript abundance does not need to correlate with a protein's abundance and activity<sup>90</sup>. Knowing that proteins are responsible for a vast number of biological functions, this makes them an essential factor for understanding a biological system and its phenotype<sup>94,95</sup>. The function of proteins and their behavior, however, remain invisible for the NGS technology, leading to an incomplete molecular picture without the proteome<sup>96</sup>.

## **1.2. Mass spectrometry-based proteome profiling**

### **1.2.1. Definition and emergence**

The field of proteomics, termed by Marc Wilkins in 1996, describes the large-scale study of proteins, the functional workhorses of any living system, cells or whole organisms<sup>97</sup>. In particular, proteins are the molecular entities or biomolecules encoded in the genome that operate and control a vast array of processes from replication of DNA and cell cycle, to

signaling tasks, immune response, and cell differentiation<sup>94,95</sup>. In the form of the cytoskeleton, proteins provide the mechanical stability of cells, support the information and molecule transport, lead responses to internal and external stimuli, and catalyze biochemical reactions<sup>94,95</sup>.

The existence of proteins as distinct biomolecules is recognized since the mid-18<sup>th</sup> century, described by the Dutch chemist Gerardus Johannes Mulder and named by the Swedish chemist Jöns Jacob Berzelius<sup>98</sup>. Since then, it took years of research to understand the structure, function, and complexity of proteins. In 1951, Linus Pauling et al., for example, described the helices structure with indirect evidence in particular proteins<sup>99</sup>, whereas in 1956, Walter Kauzmann significantly contributed to the understanding of protein folding with his work about structural factors in protein denaturation<sup>100</sup>. Just before, in 1949, it was Frederick Sanger, who sequenced the first protein, namely insulin, and established the link of proteins being amino acid sequences<sup>101,102</sup>. The first complete structure was unraveled for the myoglobin protein molecule by Sir John Cowdery Kendrew in 1958<sup>103</sup>. These and numerous other discoveries throughout the 20<sup>th</sup> century have significantly enhanced our knowledge about proteins, the 3<sup>rd</sup> downstream layer of the genetically encoded information following DNA and RNA, as explained within the central dogma of biological systems. Studying the entire set of proteins, however, remains a challenging task owing to technical difficulties to measure proteins and the highly variable and complex environment of protein expression, its regulation, and their interaction<sup>94,95,104</sup>.

The framework for large-scale protein measurements, as we know it today, was set by crucial developments in the field of mass spectrometry, a technique to determine the mass-to-charge ratio of ions<sup>105</sup>. Besides others, this includes the development of quadrupole ion traps, so-called Paul traps, in the 1950s by Wolfgang Paul and Hans Georg Dehmelt to trap charged particles in electric fields<sup>106</sup>. Therefore, they were later recognized with the Nobel Prize in Physics in 1989 (shared with Norman Foster Ramsey, Jr. for the invention of the separated oscillatory field method to precisely measure time and frequency)<sup>106</sup>. In the same year, Prestage J.D. et al. described the linear ion trap providing higher ion storage capacity and faster scan rates<sup>107</sup>. Another significant milestone was the employment of electrospray as a soft ionization method to produce ions from large molecules, such as proteins or peptides, with minimal fragmentation or degradation during the liquid-to-gas-phase

## Introduction

transition<sup>108</sup>. Electrospray ionization (ESI)<sup>108</sup> was developed in 1984 by John Bennett Fenn. He was awarded the Nobel Prize in Chemistry in 2002 (shared with Koichi Tanaka for the development of MALDI - matrix-assisted laser desorption ionization<sup>109</sup>, another soft ionization technique, and Kurt Wüthrich for the development of NMR - nuclear magnetic resonance spectroscopy to identify 3D structures of biological macromolecules<sup>110</sup>). In 1999, Alexander Makarov presented his proof-of-principle for the first orbital ion trap mass analyzer, the Orbitrap, a derivative of the earlier Kingdon trap (1923) or the modified Knight configuration (1981)<sup>111</sup>. The Orbitrap provided unprecedented, sensitive, and robust mass accuracy and high resolution. Then and nowadays, this marks a kick-off for continuous technological improvements that pave the way for modern analytical mass-spectrometry (MS)-based proteomics. The first Orbitrap mass spectrometer became commercially available in 2005 and has since remained the chief MS technology in proteomics<sup>112</sup>.

In the last 15 years, MS-based proteomic technologies have matured into a powerful tool allowing robust, reliable, and comprehensive proteome profiling in cells and tissues<sup>94,95</sup>. This is the result of parallel developments in mass spectrometric instrumentation that continues to gain speed and sensitivity<sup>113-116</sup>, in liquid chromatographic technology to separate proteins and peptides directly interfaced with MS<sup>117,118</sup>, and in data analysis pipelines for reliable protein identification and quantification<sup>119,120</sup>. Various workflows have been developed for comparative analyses across many samples using, for example, isobaric labels allowing sample multiplexing, or label-free approaches and short liquid chromatography (LC) gradients<sup>118,121,122</sup>. Collectively this has propelled proteomic studies in multiple areas of basic, mechanistic, and systems biology, using in-depth and quantitative proteomic profiles to understand spatial and temporal aspects of proteome organization and dynamics in a wide variety of conditions<sup>123</sup>. In 2014, Mathias Wilhelm et al. and Min-Sik Kim et al., released the first drafts of the human proteome<sup>124,125</sup>.

### **1.2.2. Status quo: proteomics in systems medicine**

The speed, sensitivity, robustness, and general accessibility of present-day proteomic technologies have an increasing appeal for clinical applications, for various reasons: I) underlying mechanisms of many diseases are still unclear, where proteome-level information will increase the mechanistic insight of (patho)physiological processes<sup>94,126,127</sup>; II) proteins are the primary targets of almost all current drugs, and insight in their function

will help to understand how drugs impact on cellular processes. Many drugs act unspecifically on multiple rather than a single protein target, thus making it a crucial factor to deconvolute the mechanism of action to gain confidence and improve drug discovery<sup>128</sup>; III) for many diseases there is a persistent lack of robust protein biomarkers for diagnostic, prognostic, or predictive purposes<sup>1,3</sup>. Liquid biopsies, such as blood or urine, are again particularly promising for protein-based biomarker discovery due to their non-invasiveness as compared to tissue biopsies obtained through surgery<sup>77,78</sup>. The proteome provides a unique insight that can also complement NGS-derived data, and with the current state of development, it is the consequential next step in studying biological systems, animal or cell models, and patient specimens<sup>126,129</sup>.

Despite the demand for in-depth proteome-level information, its value, and promise to bridge the blind spot between DNA/RNA and phenotype<sup>126,127</sup>, most recent applications in a clinical context have been limited to highly specialized workflows or individual proteins<sup>130,131</sup>. Many routine laboratory tests for diagnostics and therapy decision-making are based on proteins<sup>132,133</sup>. Immunohistochemistry, for example, is the main procedure in pathology for disease entity classification through staining of individual proteins, which in turn highlights their role in clinical practice<sup>133,134</sup>. In contrast, MS-based proteomics enables the global identification and quantification of thousands of proteins simultaneously<sup>94</sup>. Here, the proteomic field benefits from the head start of NGS technologies, in which it has already become clear that complex biological disease entities cannot be explained simply by their genetic alterations and transcriptional response alone<sup>1,33,127</sup>. Instead, an integration of multiple “omics” layers, including clinical data and proteomics, holds promise to extend our understanding of (patho)physiological processes and to gain insight into its clinical utility<sup>1,33,127,135</sup>. Latest systems medicine programs, such as the “Obama Precision Medicine” initiative, clearly propose an incentive to establish MS-based profiling of proteins and other biomolecules and to integrate with NGS-based and clinical data<sup>136</sup>.

Several promising case studies utilizing global proteome or phosphoproteome profiling and data integration in a clinical context are available<sup>137</sup>, which guided a therapy decision or helped subclassify a specific disease entity. In 2018, for example, Doll S. et al. have applied proteomic profiling to a chemorefractory patient with a rare urachal carcinoma for whom all previous treatment options failed<sup>130</sup>. Comparing the protein profiles in tumor tissue to

## Introduction

its surrounding, they identified differentially expressed candidates, of which one, namely lysine-specific histone demethylase 1 (LSD1), an epigenetic regulator, was in focus as a potential target in ongoing drug development attempts. Backed up with NGS and clinical data, their finding sufficiently convinced the tumor board to propose a personalized treatment with an LSD1 targeting drug. In another example, in 2018, Archer T.C. et al. applied quantitative proteome and phosphoproteome profiling to a cohort of 45 primary medulloblastoma specimens, a common pediatric brain tumor, to identify potential therapeutic targets<sup>131</sup>. Despite the low mutation rates in pediatric tumors and highly similar RNA expression, they identified extensive heterogeneity in molecular mechanisms, representing the functional state of the cancer cells, within the World Health Organization (WHO)-accepted subgroups for medulloblastoma. The membrane protein CD47, for example, was significantly enriched, suggesting that anti-CD47 therapies might be particularly successful within the respective subgroup. Furthermore, the PTM readout of phosphorylation status in MYC revealed its distinct activity in certain tumors irrespective of the expression level. The activity of MYC upregulates many genes, some of which are involved in cancer formation and cell proliferation. The utility for a clinical implication remains to be shown, but the integrative data promise a new perspective for understanding tumor biology and guiding therapy.

MS-based proteomics in clinical systems medicine is a promising trend, yet remains challenging to implement as a routine application<sup>126–129</sup>. Considering recent pioneering measures and technological maturation, it seems to be a question of time until robust instrumentation, broad training, computational efforts, and standardization will facilitate the day-to-day molecular proteome profiling of individual patients.

### **1.2.3. Clinical translation: bottlenecks and requirements**

The field of proteomics is faced with significant analytical challenges due to the sheer complexity of protein expression, their interaction, and regulation<sup>94</sup>. They can be highly variable in their spatiotemporal expression across different tissue or cell types and an organism's lifetime<sup>138–140</sup>. The expression of proteins varies tremendously during cell differentiation<sup>141</sup>, early development<sup>142</sup>, during cell cycle phases<sup>143,144</sup>, or in a disease, such as cancer<sup>145</sup>. Unlike the proteome, the nucleotide sequence of the genome stays constant and static over time, whereas individual genes or genomic regions can be more actively



transcribed or repressed<sup>146</sup>. Proteins can differ in their abundance by several orders of magnitude, complicating the identification and quantification of low abundant proteins in the presence of high abundant proteins<sup>104</sup>. The broad dynamic range constitutes a sensitivity issue for sample preparation as well as MS instrumentation, lacking a methodology to amplify protein samples similar to the polymerase chain reaction (1983 by Kary Mullis) for nucleotide sequences<sup>104,147</sup>. The transcription of a gene to mRNA does not allow a sufficient prediction of protein expression and the underlying phenotype<sup>90</sup>. Transcripts might be inefficiently translated or quickly degraded by other post-transcriptional regulatory mechanisms. On top of differential synthesis rates of individual transcripts, proteins vary significantly in their half-life or become stabilized through functional interaction in protein complexes. As previously mentioned, some proteins can be post-translationally modified by, for example, phosphorylation, which determines their activity and might be crucial for oncogenic drivers. All of the above contribute to the technical and functional complexity of proteomics, presenting both challenges and promises for systematic proteome profiling.

The successful implementation of MS-based proteomics in a clinical environment has not materialized yet, primarily because of additional requirements that need to be met on top of those in a research environment alluded to above (e.g., proteome coverage, sensitivity)<sup>126-129</sup>. This mostly pertains to I) the ability to analyze many (possibly hundreds) samples uninterruptedly and robustly in order to achieve sufficient statistical power in patient cohorts<sup>117,118</sup>, II) simplify the workflow, thereby removing the need for personnel with cutting-edge expertise and technical skills in proteomics<sup>148,149</sup>, III) achieving an adequate turn-around time from receiving samples to the generation of a complete proteome profile analysis<sup>126,129</sup>, and IV) cost-effectiveness of the workflow<sup>128</sup>. Most of these bottlenecks can be resolved simultaneously by automation, avoiding manual handling and thereby eliminating the risk of error and variability, while at the same time enabling longitudinal standardization irrespective of the number of samples. Although liquid chromatography coupled to mass spectrometry (LC-MS) has nowadays been sufficiently standardized to achieve excellent performance across hundreds of samples<sup>150</sup>, preceding sample preparation is often still highly cumbersome, involving multiple steps to extract, purify, and digest proteins before subsequent LC-MS<sup>94,104</sup>. In an ideal scenario, this

## Introduction

procedure should be streamlined into an automated pipeline that accepts processing conditions for any sample type, thereby facilitating universal applicability. Despite the range of existing sample preparation methods<sup>151–158</sup>, very few satisfy these demands to universally accommodate the different requirements imposed by various clinical tissue types. For instance, blood cells can be lysed under more mild conditions than fresh frozen tissue, while formalin-fixed and paraffin-embedded (FFPE) tissue requires harsh detergent-based methods to extract proteins efficiently<sup>159</sup>. Many currently available sample preparation methods have demonstrated their great utility in many application areas of proteomics<sup>92,139,160–167</sup>. However, they also come with some drawbacks. For instance, stage tips<sup>154</sup>, and its derivative iST<sup>156</sup>, do not tolerate detergents commonly used in proteomics, thereby restricting their generic use. Other approaches involve extensive handling procedures such as filtration<sup>151,156,158</sup>, centrifugation<sup>151,156,158</sup>, precipitation<sup>153</sup>, and electrophoresis<sup>152</sup> that are difficult to standardize or scale-up, or that lead to undesirable sample losses. The latter is especially important because the majority of realistic clinical scenarios are limited to minute amounts of an available specimen, highlighting the demand for universal and sensitive methods<sup>149,168</sup>.

With large-scale, multi-omics molecular profiling comes a considerable time and resource investment in computational data handling, storing, and integration<sup>135,169,170</sup>. Clinicians need to be aware and willing to utilize comprehensive data for therapeutic decision making. This, however, requires bioinformatic analysis due to the immense amount of data, which practically cannot easily be interpreted by a single person<sup>169</sup>. Therefore, clinical implementation of a systems medicine approach, including proteomics, needs to be an inter-disciplinary coordination between scientists, medical doctors, bioinformaticians, and others. Logistical challenges for dedicated instruments, working space, personnel, and others add up to the list of requirements<sup>128,129</sup>. Ethical issues concerning the collection and interpretation of ‘big data’ are another critical aspect that needs to be discussed to find binding agreements<sup>1,171,172</sup>. Insurance companies and health care providers need to be involved in order to ease the translation of clinical decisions and tumor boards to approved cost reimbursements.

We are still far from decoding the full complexity of biological systems, and many diseases, such as cancer, remain poorly understood. Each tumor is unique, and most exhibit a diverse

cellular and molecular heterogeneity, illustrating the need for systematic profiling. Here, MS-based proteomics will provide a crucial, thus far missing, bridge between genome, transcriptome, and phenotype.

### **1.3. Oncology**

In this thesis, two specific cancer entities, namely lung cancer and ependymoma (EPN) pediatric brain cancer, were used to showcase the applicability of our workflow and to illustrate the added value of proteome profiling. Both are introduced in the following chapters among a general framework of cancer, its epidemiology and emergence.

#### **1.3.1. Cancer: definition and epidemiology**

Neoplasia defines the uncontrolled and excessive growth of cells and tissue. The abnormal proliferation of cells typically leads to the formation of a tumor<sup>173</sup>. Neoplasms can be described in four main classes that are defined and recognized by the WHO international classification of diseases (ICD-11), namely: benign, in-situ, malignant, and of unknown behavior<sup>174,175</sup>. Malignant neoplasms are more commonly known as cancer and the focus of oncology (ICD-O), the medical branch dealing with its prevention, diagnosis, and treatment<sup>176</sup>. Cancer cells can invade surrounding tissue or organs and spread to distant parts of the body via the blood and lymph system<sup>177,178</sup>. This process is called metastasis and denotes a significant cause of cancer-related death. Several main types of cancer can be distinguished based on their cells of origin; for example, I) carcinomas emerge from epithelial cells in the skin or within tissue covering internal organs; II) sarcoma describes cancer beginning in the bone, fat, muscle, blood vessels, or connective tissue; III) multiple myeloma and lymphoma define cancer types beginning in cells of the immune system, and IV) leukemia originates from bone marrow and causes abnormal blood cells<sup>173,175,176</sup>.

Besides cardiovascular diseases, cancer is the second most prevalent cause of death worldwide, with approximately 9.6 million cases of deaths and a total burden of 18.1 million new cases in 2018 (according to the WHO)<sup>179,180</sup>. It is estimated that 38.4% of men and women will be diagnosed with cancer during their lifetime, while the prevalence significantly increases with age<sup>181</sup>. Children and young adults under 14 years of age account for roughly 1% of cancer deaths worldwide<sup>181,182</sup>. Men have a 20% higher chance than women to develop cancer<sup>179</sup>. The 5-year relative survival across all cancer types in adults

## Introduction

has increased from 50.3% between 1970 and 1977 to 67% between 2007 and 2013<sup>183–185</sup>. Improved early diagnosis and better treatment options are positively contributing factors to this development.

The most common types of cancer in both males and females are lung cancer and breast cancer, with approximately 2.09 million cases in 2018, respectively. That followed are colorectal cancer and prostate cancer with 1.8 million and 1.28 million annual cases in 2018 (according to the WHO). Lung cancer was the leading cause of cancer-related death, with approximately 1.76 million cases in 2018, followed by colorectal cancer (~862.000 cases), stomach cancer (~783.000 cases), liver cancer (~782.00 cases), and breast cancer (~627.000 cases). The highest mortality rate for either male or female prevails from lung cancer (22%) and breast cancer (15%), respectively<sup>173,179</sup>. In children, brain tumors, lymphomas, and leukemia are the most commonly diagnosed types of cancer. Brain tumors remain the leading cause of cancer-related death in children<sup>186</sup>.

### 1.3.2. Carcinogenesis

Carcinogenesis describes the transition from a normal cell into a cancer cell<sup>177,178</sup>. This process is characterized by cellular, genetic, and epigenetic changes and consequential abnormal cell proliferation and division. Homeostatic cells exhibit a fine regulation of growth and programmed cell death (apoptosis). During carcinogenesis, this order is disrupted in a stepwise process during which a cell acquires distinct traits enabling a continuous, abnormal proliferation. Initially proposed by Douglas Hanahan and Robert A. Weinberg in 2000, these traits are widely accepted as the “hallmarks of cancer”<sup>177,178</sup>. They encompass the following eight essential alterations to a cell's physiology that are necessary to breach the anti-cancer defense mechanisms and shared among all types of cancer: I) self-sufficient in growth signals, II) insensitive to growth-inhibitory signals, III) ability to evade apoptosis, IV) limitless replicative potential, V) sustained angiogenesis, the process of blood vessel formation, VI) metastasizing capabilities, VII) the reprogramming of a cell's energy metabolism, and VIII) the ability to avoid immune destruction. Hanahan and Weinberg suggest that these hallmarks are individually acquired during tumor development and only collectively cause cancer. Further, they define genomic instability, tumor-promoting inflammation, and the tumor microenvironment as “enabling characteristics” that contribute to genetic diversity and the acquisition of all hallmark traits.

This framework illustrates that carcinogenesis is a multi-step transition of distinct cellular mechanisms and pathways from their physiological conditions to deregulation, upsetting the integrity between proliferation and cell death<sup>177,178</sup>.

The accumulation of mutations in the genome is an evolutionary process that likely leads to carcinogenesis<sup>187</sup>. Several types of mutations exist ranging from single nucleotide variations (SNVs), large structural variations (SVs), such as copy-number variations (CNVs), and small insertions and deletions of nucleotide sequences (InDels)<sup>188,189</sup>. Most of them occurring in human cancer are so-called “passenger mutations” because they do not trigger a disease phenotype<sup>190</sup>. The role of these uninvolved mutations remains poorly understood, whereas an increasing body of scientists suspect that they might have a crucial involvement in pathophysiology. In contrast, mutations that cause a selective growth advantage or increased survival for the cell are called “driver mutation”<sup>190</sup>. Genes that carry a driver mutation are grouped into two classes, namely oncogenes and tumor suppressor genes<sup>191</sup>. Both play a crucial role in carcinogenesis.

Oncogenes typically upregulate cell proliferation and survival<sup>191</sup>. They are characterized by a dominant gain-of-function mutation that leads to its constant activation or overexpression<sup>192,193</sup>. In some cases, mutations in oncogenes result in altered proteins with a novel, tumor-promoting property<sup>194–196</sup>. A prominent example is an amplifying point mutation in the gene coding for the AKT protein, a serine/threonine-protein kinase (AKT1, AKT2, and AKT3)<sup>197</sup>. Under physiological conditions, it is involved in an array of different processes from metabolism, proliferation, to angiogenesis. It contributes to the regulation of cell survival via the phosphorylation of MAP3K5, an apoptosis signal-related kinase, which is activated upon oxidative stress. Its decreasing activity triggered by the AKT overexpression prevents apoptosis, one of the acquired hallmark traits. On the other hand, tumor suppressor genes are characterized by a repressive loss-of-function mutation<sup>191</sup>. They are often involved in maintaining the integrity of cell proliferation or protection against genomic instability. During carcinogenesis, they are often disabled by cancer-promoting genetic alterations leading to an inactivation of their regulatory impact. The cellular tumor antigen p53 (TP53), for example, acts as a tumor suppressor in many cancer types as it is involved in growth arrest, apoptosis, or cell cycle regulation<sup>191</sup>. Its inactivation causes insensitivity to anti-growth signals and the ability to evade apoptosis.

## Introduction

Despite the theoretical understanding of carcinogenesis, the coherence between a genotype, its influencing environmental factors, and corresponding cancer or disease states often remain unclear<sup>177,178</sup>. Here, the proteome composition has the potential to provide the missing link to understand the impact of mutations, the mechanisms of hallmark trait acquisition, and the breaching of anti-cancer defenses as a result.

### 1.3.3. Lung cancer

Primary lung cancer arises from respiratory epithelial cells and thereby classifies as carcinoma with uncontrolled growth in the lung<sup>175,198</sup>. Metastasis that spread to the lung from other parts of the body are considered as secondary lung cancer. The most common age of diagnosis is 65 years or older, with an average of around 70 years<sup>181,199,200</sup>. A small number of cases are diagnosed per year at an age younger than 45 years. About 85%, the vast majority of cases of lung cancer are caused by long-term tobacco smoking<sup>181,200</sup>. The number of diagnosed cases per year is continuously declining, together with the increasing trend of non-smokers. Other causes are frequent exposure to dust, asbestos, paint, or general air pollution. On average, lung cancer has a 5-year survival rate of approximately 18.6% (>55% at early detection). The occurrence of lung cancer is categorized based on the size and appearance of the tumor mass and the malignant cells' morphology. The WHO classification of lung cancer comprises two main types, namely small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC), based on the cell type of cancer origin<sup>175,176</sup>. In comparison, SCLC comprises significantly smaller cells and features the ability to metastasize, making it a highly malignant tumor rapidly. SCLC is rarely seen in non-smokers and accounts for roughly 13% of lung cancer cases worldwide<sup>181,200</sup>. On the other hand, NSCLC represents roughly 87% of all cases<sup>201</sup>.

NSCLC can be sub-divided into three major pathologic subtypes: I) squamous cell carcinoma, II) large cell carcinoma, and III) adenocarcinoma (ADCs)<sup>175</sup>. The latter is the most common histological lung cancer subtype accounting for roughly 60% of all NSCLC (~38% of all lung cancer cases). They are known for their heterogeneous clinical, radiologic<sup>162</sup>, molecular<sup>202-204</sup>, and morphological<sup>205</sup> features. Thus far, five distinct histological growth patterns have been recognized by the 2015 WHO Classification of Lung Tumors<sup>206</sup>. These growth patterns, which are reported in any pathology report, have been proposed for tumor grading according to the predominant pattern of a tumor: lepidic (low grade; group

1), acinar and papillary (intermediate grade; group 2), and solid and micropapillary (high grade; group 3)<sup>198,201</sup>. Applying this grading system led to the observation of significant differences regarding prognosis<sup>207</sup> and prediction of benefit from adjuvant chemotherapy<sup>208</sup>, where patients with lepidic ADC or micropapillary ADC were associated with the most favorable or worst prognosis, respectively.

The current standard of care for lung cancer is highly dependent on the stage of diagnosis, the type of mutation, and the potential spread of metastasis<sup>198,201</sup>. In an early stage, for example, a maximal-safe surgical resection can be facilitated by a still confined area of cancerous cells (a localized disease has a 5-year survival rate of approximately 52%). In severe cases that do not allow surgery, cancerous cells might be targeted by radiotherapy if it is tolerable considering the patients' health<sup>201,209</sup>. Platinum-based chemotherapy is likely the next stage of treatment in cases that already developed metastasis. The spreading and growth of cancer cells might be slowed down by drug therapies that target specific changes in the cancer cell microenvironment<sup>210</sup>. Unfortunately, lung cancer frequently does not cause symptoms until cancerous cells have spread to other parts of the body, leading to an overall bad prognosis.

A vast number of studies aimed to bring genetic factors associated with lung cancer to light. These efforts have led to several different discoveries. In 2006, for example, Lu Y. et al. performed a meta-analysis of seven microarray studies and identified a 64-gene signature, which is predictive for lung cancer reoccurrence in stage I NSCLC patients<sup>211</sup>. In other genomic studies, a tumors' responsiveness to chemotherapies could be predicted, or the association between genomic alterations and distinct growth advantages was elucidated<sup>189</sup>. Further, the risk of developing lung cancer correlates to frequently observed polymorphisms on chromosome 5, 6, and 15<sup>212–214</sup>. Increasing mutations rates in epidermal growth factor receptor (EGFR) were linked to NSCLC ADC of patients<sup>215</sup>. Supported by molecular characterization, treatment plans using EGFR tyrosine kinase inhibitors (gefitinib and erlotinib) have demonstrated improved clinical outcomes by slowing down disease progression<sup>215</sup>. In another study in 2009, Boutros PC. et al. identified a six-gene expression signature (STX1A, HIF1A, CCT3, HLA-DPB1, RNF5, and MAFK) with prognostic value for NSCLC patients, which could be validated in four distinct testing datasets<sup>216</sup>. They propose that a clinical implementation using RT-PCR analysis of the six genes can provide a quick

## Introduction

readout about good or poor prognosis. However, they observed that many proposed gene signatures from immense efforts in lung cancer research appear to have a lack of overlap, illustrating the need for an improved molecular characterization. Additional proteome profiling efforts may achieve a complementary or improved insight.

To this end, there were no efforts to perform a molecular classification and characterization of growth patterns of ADCs (NSCLC), the most common lung cancer type, on the proteome level. In most invasive ADCs, more than one of the previously mentioned growth patterns can coincide<sup>204,217</sup>, which further highlights the need to understand functional differences and clinical implications of histological heterogeneity better. In this thesis, proteome profiling of ADC growth patterns was performed for the first time.

### **1.3.4. Pediatric brain tumor: ependymoma**

#### 1.3.4.1. Definition and epidemiology

The majority of pediatric brain tumors are classified as gliomas accounting for roughly 52.9% of all cases<sup>218</sup>. Brain tumors are the second most common type of tumor occurring in children<sup>219</sup>. Other types of pediatric brain tumors are medulloblastomas (15-20%), originating from immature or embryonal cells, choroid plexus tumors (10-20%), germ cell brain tumors (4%) or atypical teratoid rhabdoid tumors (ATRTs; 1-2%)<sup>218</sup>. Gliomas are primary tumors that arise from glial cells<sup>220,221</sup>. Oligodendrocytes, astrocytes, ependymal cells, and microglia comprise the four types of glial cells in the central nervous system (CNS). The majority of gliomas are typically named corresponding to the glial cell type that is histologically most similar but which not necessarily reflect the tumor origin. The main types include astrocytomas, ependymomas, oligodendrogliomas, brainstem gliomas, and mixed gliomas that are comprised of several types of glial cells. They are further categorized by four different tumor grades, from least severe (low-grade: grade I and II) to highly malignant (high-grade gliomas: grade III and IV), and by their anatomical location within the CNS<sup>175,176</sup>. Astrocytomas are the most common glioma tumor in children, accounting for 33.2% of all pediatric cases<sup>218</sup>.

Ependymomas are the third most common type of glioma tumor in the CNS in children<sup>222</sup>. They account for 10.4% of all pediatric glioma cases and arise from ependymal cells or so-called radial glial cells. They constitute a specialized type of epithelium to line the ventricular system of the brain and the spinal cords' central canal, allowing a continuous



flow of cerebrospinal fluid (CSF)<sup>223,224</sup>. Ependymal cells are further involved in CSF production and secretion. Ependymomas most frequently occur at an average age of 5 years and 35 years, in children and adults, respectively<sup>225</sup>. In adults, they are rarely diagnosed and only account for 1.9% of all primary brain tumors<sup>226</sup>. While the majority of incidences in children occur intracranially (90% within the brain), most ependymal tumors in adults arise in the spine<sup>227,228</sup>. The clinical outcome for ependymoma patients also varies between children and adults with a 10-year survival of 64% and >80%<sup>229</sup>. Despite the ability of ependymomas to spread, they are rarely observed to metastasize beyond the CNS<sup>230</sup>.

Following the glioma-related grading system, the most recent WHO classification distinguishes ependymomas into main categories of subependymoma (grade I), myxopapillary ependymoma (grade II), anaplastic ependymoma (grade III), and ependymoma, RELA-fusion positive (grade II and III)<sup>231,232</sup>. The tumor cells can be well separated from the surrounding healthy cells by histological examination and exhibit features of true ependymal rosettes and perivascular pseudorosettes<sup>233</sup>. Most tumors are low-grade (grade II), while anaplastic ependymoma (grade III) are often additionally characterized by an increased cellular density, necrosis, and microvascular proliferation, without compromising the typical tissue pattern<sup>233</sup>. Myxopapillary ependymomas are further characterized by papillary formation with a mucinous core and most commonly arise in the spine. Low-grade subependymomas have a spherical phenotype and consist of uniform cells in a fibrillary stroma with cystic degeneration<sup>226</sup>.

The standard of care for ependymal tumors is a gross total resection (GTR) via surgery followed by optional chemotherapy and focal radiotherapy<sup>234-238</sup>. The latter has been linked to reduced tumor mass, increased overall survival rates, and benefits in the prevention of ependymoma recurrence<sup>235,239</sup>. The complete removal of the tumor mass is typically confirmed using postoperative magnetic resonance imaging (MRI)<sup>240</sup>. In many cases, GTR surgery in the brain or spine of children remains challenging, with a high risk of side effects or damaging healthy tissue<sup>241-243</sup>. This may be due to the tumor location, or a not well-differentiated growth and infiltration into healthy parenchyma. Some highly malignant ependymoma types may spread through the CSF to other parts of the CNS and typically require radiation therapy. Here, the surgeon aims for a maximal safe surgical resection (subtotal resection, STR), which can still significantly reduce the tumor mass to

## Introduction

increase the efficacy of the subsequent radiotherapy<sup>241,242</sup>. The degree of tumor resection via GTR or STR has been shown as the main prognostic factor in children and adults<sup>234–236,238</sup>.

Not every incidence of ependymoma can be treated easily, especially children younger than three years of age have shown severe side-effects upon focal radiotherapy, resulting in neurocognitive deficiencies, surrounding tissue abnormalities, or increased likelihood of secondary cancer development<sup>244–246</sup>. For highly malignant, anaplastic ependymomas, the overall recurrence rate remains high, showing a median progression-free survival (PFS) of only 2.3 years<sup>247</sup>. This further highlights an obvious need for novel treatment strategies and a better understanding of the ependymoma-related pathophysiology. Both are addressed by molecular characterization, including insight into the proteome composition.

### 1.3.4.2. Molecular classification

The molecular characterization of brain tumors<sup>248,249</sup> has an increasing appeal to improve diagnosis and the WHO classification, which conventionally relies mostly on histopathological examination and staging into grade I to IV<sup>175,231,232</sup>. While routinely applied for many diseases, this histological grading presents a common problem for many brain tumors<sup>250–252</sup>. Low inter-observer reproducibility and ambiguous results are frequent in ependymomas, due to their diverse clinical behavior and highly challenging histopathological features<sup>250–252</sup>. Increasing numbers of studies additionally rely on molecular characterization, such as gene expression profiling or DNA methylation profiling (e.g., CNS-PNET)<sup>253,254</sup>. The latter has been established to enable robust and reproducible evaluation of brain tumors beyond the hitherto existing classification. Capper D. et al. recently implemented a DNA methylation-based classifier with which a brain tumor can be assigned to a distinct methylation class by comparison to a reference cohort of 2801 brain tumors<sup>255</sup>. In their validation cohort of 1104 tumors, the initial histopathological diagnosis was changed in 12% of all cases based on the assigned methylation class, thereby improving the diagnostic accuracy. Despite a few examples in which the WHO accredited the incorporation of molecular features to extend the conventional classification, histopathology remains the established standard of diagnostics<sup>231,232</sup>. In 2015, Kristian Pajtler et al. revealed that ependymal tumors are comprised of at least nine molecular subgroups utilizing DNA methylation profiling and additional gene expression data<sup>73</sup>. The

subgroups are equally distributed among the compartments of the CNS (ST: supratentorial, PF: posterior fossa, and SP: Spine) and named accordingly: SP-EPN, SP-MPE, SP-SE, PF-EPN-A, PF-EPN-B, PF-SE, as well as ST-EPN-RELA, ST-EPN-YAP1, and ST-SE. Despite these immense efforts, the subgroup-specific oncogenic driver and functional background remain largely unknown and only few or no recurrent genetic events were observed<sup>72,73,227,256–259</sup>. This highlights the potential for complementary proteome profiling to shed light on currently unknown EPN subgroup-specific biology.

The SP-SE group predominantly encompasses low-grade I subependymomas (SE) with most incidences in adults<sup>226,260</sup>. They exhibit a characteristic deletion of chromosome arm 6q with an otherwise stable genome<sup>73</sup>. Nevertheless, they usually have an excellent prognosis and outcome, even in highly malignant grade III anaplastic EPNs. SP-MPE ependymal tumors are mostly grade I myxopapillary ependymomas, which also primarily occur in adults. Despite a vast number of chromosomal instabilities, including gains and losses of entire arms, they instead display a favorable clinical outcome. Most grade II and III ependymal tumors in the spine are classified as SP-EPN. They typically feature a deletion of chromosome arm 22q, harboring the tumor suppressor gene NF2 that is frequently mutated or lost in spine ependymal tumors<sup>261</sup>. NF2 codes for the Merlin protein and is involved in tumor suppression by restricting proliferation and promoting apoptosis.

The second anatomical region with the occurrence of ependymomas is the posterior fossa. It is part of the intracranial space and contains the brainstem (medulla oblongata, pons, mid- and hindbrain) and cerebellum<sup>262</sup>. Subependymomas within the posterior fossa (PF-SE) again have a distinctive methylation pattern, showcasing similar prevalence and clinical characteristics as SP-SE tumors without the deletion of chromosome arm 6q<sup>73</sup>. PF-EPN-A and PF-EPN-B subgroups are comprised of grade II and III tumors with vast molecular and clinical differences<sup>263–265</sup>. PF-EPN-A tumors predominantly occur in children with high reoccurrence and invasive growth patterns. On the other hand, PF-EPN-B tumors display a more benign phenotype with antithetic characteristics occurring mostly in adults, low rate of recurrence, non-invasive growth, and a resulting reasonable 10-year survival rate of 88%<sup>73,260,263</sup>. Molecularly, PF-EPN-A tumors present a balanced genome besides a prominent gain of chromosome arm 1q<sup>260,266</sup>. This gain has been linked to poor prognosis in some independent studies<sup>266–269</sup>. Another unique characteristic of PF-EPN-A tumors is

## Introduction

the lack of the repressive histone mark H3K27me3<sup>265,270-272</sup>. Correspondingly, these tumors show increased expression of several genes involved in various carcinogenic processes, such as angiogenesis, growth-factor pathways, and receptor tyrosine kinase signaling. In comparison, PF-EPN-B tumors neither lack the H3K27me3 mark nor do they show the associated genes upregulated<sup>265,270,272</sup>. Further, PF-EPN-A tumors display an increased CpG methylation pattern (CpG island methylator phenotype: CIMP-positive) of promoter regions of polycomb repressive complex 2 (PRC2) target genes. CIMP-positive cancer types are often associated with worst disease-free survival after primary treatment and worse overall survival<sup>273</sup>. The PRC2 complex has histone methyltransferase activity and primarily functions to trimethylate the H3K27 for the silencing of genomic regions<sup>274,275</sup>.

The remaining site of ependymal tumor occurrence is the supratentorial region of the brain. The reasonably large area contains the cerebrum, which consists of both hemispheres of the cerebral cortex, the hippocampus, basal ganglia, and the olfactory bulb<sup>262</sup>. The cerebral cortex is the most prominent site of neural integration in the CNS and has a pivotal role in consciousness, awareness, memory, language, and other crucial functions<sup>262</sup>. The ST regions again present a grade I subependymoma (ST-SE) molecular subgroup with an overall good outcome, the highest prevalence in adults, and otherwise similar characteristics to the SP- and PF-SE tumors<sup>73</sup>. A similar good outcome is observed in ST-EPN-YAP1 tumors that are predominantly comprised of grade II and III tumors<sup>73</sup>. They are characterized by focal aberrations on chromosome 11, resulting in the dominating and the less recurrent fusion genes, YAP1-MAMLD1 and YAP1-FAM1188<sup>73,276</sup>. YAP1 is a transcriptional regulator taking part in proliferation and suppression of apoptotic genes. The Hippo signaling pathway is known to inhibit YAP1 to allow cellular control of organ size and tumor suppression<sup>277,278</sup>. ST-EPN-RELA tumors are driven by other distinct gene fusions involving C11orf95 and RELA, an effector of the NF-kappa-B transcription factor complex<sup>73,259</sup>. The complex is involved in a vast number of cellular processes and metabolism<sup>279</sup>. The NF-kappa-B/RELA activation has been associated with carcinogenesis and a negative correlation with patient survival in a series of different tumor entities<sup>279-282</sup>, such as breast cancer<sup>283</sup>, prostate cancer<sup>284</sup>, and leukemia<sup>285</sup>. The fusion is thought to result from a local chromothripsis on chromosome 11, a single, massive mutational rearrangement in a confined genomic region<sup>259,265</sup>. ST-EPN-RELA tumors encompass grade

II and III ependymomas and mostly occur in children. Together with PF-EPN-A, they show the worst overall prognosis with a 10-year survival rate of 50%, while accounting for the majority of ST ependymal tumors (70%)<sup>73,260,286</sup>. Importantly, ST-EPN-RELA constitutes the only molecular subgroup that is already accredited and included in the latest WHO classification of tumors in the CNS<sup>231,232</sup>.

Although, in the past and present, extensive efforts are undertaken to elucidate the biology of EPNs and to find potential therapeutic implications, the majority of subgroups are still poorly understood and without a specific treatment possibility<sup>73,260</sup>. Here, proteomic profiling has the potential to enhance our insight into unknown biological functions on the level of molecular mechanisms that drive pathophysiologic conditions. Ensuing, the proteome composition may facilitate the discovery of new drug targets, subgroup-specific biomarkers, or provide an extension of the current classification system for EPNs. Collectively, the proteome holds promise to complement the yet incomplete molecular picture along with previous molecular characterization efforts.

#### **1.4. Aim of this study**

Mass spectrometry (MS)-based proteomic technologies have evolved to allow global profiling across thousands of proteins. Due to previously mentioned bottlenecks in proteomics workflows, the routine application of proteome profiling has not yet been implemented in a clinical context complementary to other next-generation sequencing techniques. The aim of this study was 2-fold:

I) the technical establishment and implementation of an automated, universal workflow for routine protein sample preparation from a wide range of clinically relevant input material. Specifically, we aimed to include challenging to handle sample types, such as FFPE tissue, or quantity-limited samples. The workflow's performance was demonstrated by assessing the precision, longitudinal robustness, and sensitivity. We further aimed to evaluate and optimize all relevant parameters to allow a deep proteome profiling with optimal quantification and rapid turn-around times.

II) the applicability of our workflow to a clinical question: we used our automated workflow to process a pulmonary ADC (FFPE) cohort, comprising all histologically defined growth patterns that are accredited by the WHO. Currently, these growth patterns have a limited clinical implication. To the best of our knowledge, a proteomic characterization, including the functional assessment of molecular mechanisms between different ADC growth patterns, did not exist until now. Simultaneously, we aimed to illustrate the potential of proteome profiling in another realistic scenario. Specifically, we utilized a cohort of EPN pediatric brain tumors, an entity of primary tumors within the CNS of children and young adults. Recently, the existence of nine distinct molecular subgroups has been shown, whereas, for the majority of subgroups, a functional explanation is still lacking. In this study, we used a subset of an EPN cohort (Pajtler et al., 2015)<sup>73</sup> to investigate the proteome composition across all nine molecular subgroups. Altogether, this provides a rich molecular dataset to explore the utility of proteomic data in combination with other NGS data to enhance our understanding of EPN biology and potential clinical implications.

## 2. Materials

### 2.1. Chemicals and other materials

Here, all used chemicals, reagents, equipment, and consumables are listed in alphabetically ordered groups. The materials used solely by collaborators are not explicitly listed:

Reagent/Resource	Reference or Source	Identifier or Catalog Number
<b>Experimental Models</b>		
<i>HeLa cells (H. sapiens)</i>	ATCC (Wesel, Germany)	ATCC CCL-2
<i>MCF7 cells (H. sapiens)</i>	ATCC (Wesel, Germany)	ATCC
<i>HEK-293 (H. sapiens)</i>	ATCC (Wesel, Germany)	ATCC
<i>ISTMEL-1 (H. sapiens)</i>	Obtained from colleagues for protein quantification	N/A
<i>UACC-62 (H. sapiens)</i>	Obtained from colleagues for protein quantification	N/A
RPMI-7951 ( <i>H. sapiens</i> )	Obtained from colleagues for protein quantification	N/A
<i>A375 (H. sapiens)</i>	Obtained from colleagues for protein quantification	N/A
Patient-derived EPN cell lines ( <i>H. sapiens</i> )	Obtained from collaborator	N/A
Pulmonary adenocarcinoma (ADC) FFPE specimens ( <i>H. sapiens</i> )	Thoraxklinik at Heidelberg University (NCT; project: # 1746; # 2818)	N/A
Ependymoma patient fresh-frozen tissue ( <i>H. sapiens</i> )	N/A	N/A
<b>Chemicals, Enzymes and other reagents</b>		
1,2- Cyclohexanedione	Sigma-Aldrich (Steinheim, Germany)	Ref: C101400; Lot: STBF6948V
100 x glutamine stock solution	Life Technologies (Darmstadt, Germany)	25030081
2-Chloroacetamide (CAA)	Sigma-Aldrich (Steinheim, Germany)	22790; Lot: BCBN8771V
4x Laemmli Buffer	Bio-Rad Laboratories GmbH (Feldkirchen, Germany)	Ref: 1610747; Lot: 64261673
6x DNA loading dye	Thermo Scientific (Schwerte, Germany)	Ref: R0611; Lot: 00652028
Acetic acid glacial	Biosolve Chemicals (Dieuze, France)	000107413185; Lot: 1061651
Acetic acid glacial	Fisher Scientific (Schwerte, Germany)	UN2789; Lot: 1679445
Acetonitrile (ACN)	Biosolve Chemicals (Dieuze, France)	0001204101BS; Lot: 1274241
Agarose	Sigma-Aldrich (Steinheim, Germany)	Ref: A9539; Lot: SLBT5972
Ammonium bicarbonate (ABC)	Fluka Analytical (Munich, Germany)	Ref: 40867; Lot: I1620
Ammonium formate	Sigma-Aldrich (Steinheim, Germany)	Ref: 70221-25G, Lot: BCBV1667
Ammonium formate	Biosolve Chemicals (Dieuze, France)	Ref: 0001904153BS; Lot: 1323041
Ammonium hydroxide solution	Fluka Analytical (Munich, Germany)	Ref: 44273-100ML-F, Lot: BCBQ0888V
Aniline	Sigma-Aldrich (Steinheim, Germany)	Ref: 242284; Lot: STBH5612

## Materials

Benzonase	Merck (Darmstadt, Germany)	71206-3; Lot: 3271105
Bovine Serum Albumin (BSA)	Fisher Scientific (Schwerte, Germany)	BP9702; Lot: 190211-0662
cOmplete, EDTA-free Protease Inhibitor Cocktail	Roche Diagnostics (Mannheim, Germany)	40694200; Lot: 05056489001
Dithiothreitol (DTT)	Biomol GmbH (Hamburg, Germany)	04010.25; Lot: 4001
Dulbecco's Modified Eagle Medium (DMEM) with high glucose and no glutamine	Life Technologies (Darmstadt, Germany)	11960085
E. coli standard	Bio-Rad Laboratories GmbH (Feldkirchen, Germany)	1632110
Ethanol (EtOH)	Merck (Darmstadt, Germany)	34852
Ethanol (EtOH) absolute	VWR International GmbH (Darmstadt, Germany)	20821.310; Lot: 18K144019
Ethylenediaminetetraacetic acid (EDTA)	Sigma-Aldrich (Steinheim, Germany)	Ref: E9884; Lot: BCBZ8264
EZ-Link Alkoxyamine-PEG4-Biotin	Thermo Scientific (Schwerte, Germany)	Ref: 26137
Fetal Bovine Serum (FBS)	Life Technologies (Darmstadt, Germany)	10270106
Formaldehyde solution (37%)	Merck (Darmstadt, Germany)	K46701403519; 1.04003.1000
Formaldehyde solution (w/v) (16%) Methanol-free	Thermo Scientific (Schwerte, Germany)	Ref: 28908; Lot: TL2688131
Formic acid (FA)	Biosolve Chemicals (Dieuze, France)	0006914143BS; Lot: 1297891
Gene Ruler 1kb DNA Ladder	Thermo Scientific (Schwerte, Germany)	Ref: SM1331; Lot: 00663462
GeneRuler 100 bp Plus DNA Ladder	Thermo Scientific (Schwerte, Germany)	Ref: SN0321; Lot: 00303113
GlutaMAX HEPES supplement	Thermo Fisher Scientific (Braunschweig, Germany)	10564011
High Capacity Neutravidin Agarose Resin	Thermo Scientific (Schwerte, Germany)	Ref: 29204; Lot: TE269779
Hydrochloric acid (37%)	Merck (Darmstadt, Germany)	K51884217943; 1.00317.1011
LCMS-grade water	Biosolve Chemicals (Dieuze, France)	00232141B1BS
MagReSyn Amine Beads	ReSyn Biosciences (Edenvale, South Africa)	NA
MagReSyn HILIC Beads	ReSyn Biosciences (Edenvale, South Africa)	NA
Methanol (MeOH)	Biosolve Chemicals (Dieuze, France)	0013684101BS; Lot: 1277161
Paramagnetic beads for SP3 (Sera-Mag Speed Beads A and B)	Fisher Scientific (Schwerte, Germany)	24152105050250 & 44152105050250
Penicillin-Streptomycin (P&S) mix	Life Technologies (Darmstadt, Germany)	15140122
Phosphate buffered saline (PBS)	Life Technologies (Darmstadt, Germany)	70011051
Pierce HeLa standard	Thermo Fisher Scientific (Braunschweig, Germany)	88328
Pierce LTQ Velos ESI Positive Ion Calibration Solution	Thermo Scientific (Schwerte, Germany)	Ref: 88323; Lot: UE283806
Precision Plus Protein Dual Color Standard	Bio-Rad Laboratories GmbH (Feldkirchen, Germany)	Ref: 61-0374; Lot: 004030A
Precision Plus Protein standard	Bio-Rad Laboratories GmbH (Feldkirchen, Germany)	161-0374
Protease inhibitor cocktail (PIC)	Sigma-Aldrich (Steinheim, Germany)	5056489001



ProteaseMax Enhancer	Surfactant, Trypsin	Promega (Madison, WI, USA)	Ref: V2072; Lot: 000340968
Proteinase K		Thermo Scientific (Schwerte, Germany)	Ref: EO0491, Lot: 00521266
RapiGest SF Surfactant		Waters Corporation (Milford, USA)	Ref:186001861; Lot: 190231
Sequencing grade modified trypsin		Promega (Madison, WI, USA)	V5111; Lot: 0000379610
Silver nitrate		Sigma-Aldrich (Steinheim, Germany)	S8157; Lot: MKBZ5510V
Sodium carbonate		Sigma-Aldrich (Steinheim, Germany)	Ref: S7795; Lot: BCBT4969
Sodium Chloride		Sigma-Aldrich (Steinheim, Germany)	Ref: S1679; Lot: SLBN3273V
Sodium Cyanoborohydrate		Sigma-Aldrich (Steinheim, Germany)	Ref: 156159; Lot: SHBH1335V
Sodium meta-Periodate		Thermo Scientific (Schwerte, Germany)	Ref: 20504, Lot: TI273898
Sodium pyruvate		Thermo Fisher Scientific (Braunschweig, Germany)	11360070
Sodium thiosulfate anhydrous		Merck (Darmstadt, Germany)	K48623312707; 1.06512.0250
Sodium-dodecylsulfate (SDS)		Applichem (Darmstadt, Germany)	A0675
Sodium-dodecylsulfate (SDS) 20%		Bio-Rad Laboratories GmbH (Feldkirchen, Germany)	Ref: 1610418; Lot: 64245485
SYBR Safe DNA Gel Stain		Invitrogen (Carlsbad, USA)	Ref: S33102; Lot: 2053914
TAE buffer (50 mM EDTA, 2 M Tris, 1 M glacial acetic acid)		Self-made (chemicals listed separately)	
Tartrazine		Sigma-Aldrich (Steinheim, Germany)	T0388; Lot: MKCB1542V
Tissue Lysis Buffer		Covaris, Inc. (Woburn, USA)	Lot: R001595
Triethylammonium bicarbonate		Sigma-Aldrich (Steinheim, Germany)	T7408; Lot: BCBX6381
Trifluoroacetic acid (TFA)		Biosolve Chemicals (Dieuze, France)	0020234131BS; Lot: 1273961
Tris(2-carboxyethyl)phosphine (TCEP)		Sigma-Aldrich (Steinheim, Germany)	C4706
Triton X-100		Sigma-Aldrich (Steinheim, Germany)	Ref: T8787-250ML, Lot: SLBW7103
Trypsin/Lys-C Mix, Mass Spec grade		Promega (Madison, WI, USA)	V5073; Lot: 0000351191
Trypsin-EDTA (0.25%)		Life Technologies (Darmstadt, Germany)	25200056
TWEEN20		Sigma-Aldrich (Steinheim, Germany)	Ref: P9416-100ML, Lot: SLBS9921
UltraPure Tris		Invitrogen (Carlsbad, USA)	Ref: 15504-020; Lot:8309093
Urea		Bio-Rad Laboratories GmbH (Feldkirchen, Germany)	161-0731

### Software

Limma moderated t-statistics (R package version 3.36.3)	<a href="https://support.bioconductor.org/p/6124/">https://support.bioconductor.org/p/6124/</a>
MaxQuant (version 1.5.1.2)	<a href="https://www.maxquant.org/">https://www.maxquant.org/</a>
MOFA	<a href="https://rdrr.io/bioc/MOFA/man/MOFA.html">https://rdrr.io/bioc/MOFA/man/MOFA.html</a>
Perseus (version 1.6.1.3)	<a href="https://maxquant.net/perseus/">https://maxquant.net/perseus/</a>
R (version 3.5.1)	<a href="https://www.r-project.org/">https://www.r-project.org/</a>
R package fgsea (version 1.6.0)	Sergushichev, A. A. An algorithm for fast pre-ranked gene set enrichment analysis using cumulative statistic calculation. <i>bioRxiv</i> 60012 (2016). doi:10.1101/060012

## Materials

REACTOME pathway database Gene sets using ReactomePA R package (version 1.24.0)	Yu, G. & He, Q. Y. ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. <i>Mol. Biosyst.</i> <b>12</b> , 477–479 (2016).	
t-SNE analyses were performed using R package tsne (version 0.1-3)	van der Maaten, Laurens, Hinton E., G. Visualizing Data using t-SNE. <i>J. Mach. Learn. Res.</i> <b>164</b> , 10 (2008).	
vworks automation control software	<a href="https://www.agilent.com/en/products/software-informatics/automation-solutions/vworks-automation-control-software">https://www.agilent.com/en/products/software-informatics/automation-solutions/vworks-automation-control-software</a>	
Instrumentation/ Equipment		
1 mL tissue dounce homogenizer	Wheaton (DWK Life Science Inc.)	N/A
-80°C Freezer	Eppendorf - New Brunswick (Edison, USA)	U725-G Innova
Agarose-Gel running chamber	Biostep GmbH (Burkhardtsdorf, Germany)	GH140318005
Aspiration System	Integra Bioscience GmbH	Integra Vacusafe
Automated Cell Counter	Bio-Rad Laboratories GmbH (Feldkirchen, Germany)	TC20
Bioruptor Pico	Diagenode SA (Seraing, Belgium)	SN:P-152703
Branson Digital Sonifier	Branson Ultrasonic Corporation (USA)	NA
Bravo liquid handling system	Agilent Technologies (Santa Clara, USA)	<a href="https://www.agilent.com/en/products/automated-liquid-handling/automated-liquid-handling-platforms/bravo-automated-liquid-handling-platform">https://www.agilent.com/en/products/automated-liquid-handling/automated-liquid-handling-platforms/bravo-automated-liquid-handling-platform</a>
Cell Culture Centrifuge	Thermo Scientific (Schwerte, Germany)	Heraeus MegaFuge 40
Cell Culture Laminar Flow Hood	Thermo Scientific (Schwerte, Germany)	MaxiSafe 2020
Centrifuge	Eppendorf (Hamburg, Germany)	Centrifuge 5424
Centrifuge 5424 Rotor	Eppendorf (Hamburg, Germany)	F-45-32-5-PCR
Centrifuge 5424 Rotor	Eppendorf (Hamburg, Germany)	FA-45-24-11 (Eppi's)
Centrifuge 5430R Rotor	Eppendorf (Hamburg, Germany)	FA-45-48-11 (Eppi's)
Centrifuge 5430R Rotor	Eppendorf (Hamburg, Germany)	A-2-MTP (Plates)
CO <sub>2</sub> Incubator	Thermo Scientific (Schwerte, Germany)	HeraCell Vios 160i
Cooling Centrifuge	Eppendorf (Hamburg, Germany)	Centrifuge 5430R
Covaris Cap Strip Seal holder	Covaris, Inc. (Woburn, USA)	500608 (Strip Caps)
Covaris Foil Seal holder	Covaris, Inc. (Woburn, USA)	500608 (Foil)
Custom-made SP3 magnet	EMBL, Heidelberg	N/A
DynaMag-2 magnetic stand	Life Technologies (Darmstadt, Germany)	N/A
Easy NanoLC 1200	Thermo Fisher Scientific (Braunschweig, Germany)	NA
Heraeus MegaFuge 40 Rotor	Thermo Scientific (Schwerte, Germany)	75003180
High pH HPLC System (Infinity 1260)	Agilent Technologies (Santa Clara, USA)	1260/1290 Infinity

HotSleeve 25 cm Smart Column Heater	Analytical Sales & Services, Inc. (Flanders, USA)	Ref: HSI-25L
Ice machine	Ziegma Eismaschinen GmbH (Isernhagen, Germany)	SN:151759
Incubator	Thermo Scientific (Schwerte, Germany)	HeraTherm
LE220R-plus Focused-ultrasonicator	Covaris, Inc. (Woburn, USA)	500578
MAGNUM FLX enhanced universal magnet	ALPAQUA (Beverly, USA)	<a href="https://www.alpaqua.com/Products/Magnet-Plates/Magnum-FLX">https://www.alpaqua.com/Products/Magnet-Plates/Magnum-FLX</a>
Mastercycler	Eppendorf (Hamburg, Germany)	Eppgradient S
Microscale	Sartorius Lab Instruments (Göttingen, German)	MSA125P-000-DA
Minicentrifuge	neoLab (Heidelberg, Germany)	3-1810
MiniChiller (Picoruptor)	Diagenode SA / Huber (Seraing, Belgium)	NA
MonoSleeve Column Heater	Analytical Sales & Services, Inc. (Flanders, USA)	NA
Multi-image Light Cabinet	Alpha Innotech Corporation (San Leandro, USA)	NA
Multi-Rotator	Grant-bio Instruments (Royston, UK)	PTR-35
Nanodrop 1000 Spectrophotometer	Thermo Scientific (Schwerte, Germany)	N/A
NanoQuant Plate Reader	Tecan (Männedorf, Switzerland)	Infinite M200pro
Orbital shaking station	Agilent Technologies (Santa Clara, USA)	Variomag Teleshake
PCR cycler with lid heating (CHB-T2-D ThermoQ)	Hangzhou BIOER Technologies (Binjiang, China)	CHB-T2-D ThermoQ
Polymax 2040 Platform shaker	Heidolph Instruments GmbH & Co. KG (Schwabach, Germany)	Polymax 2040
Power Supply	Bio-Rad Laboratories GmbH (Feldkirchen, Germany)	PowerPac Universal
Pressure Bomb	Nanobaume-Western fluids (Wildomar, USA)	N/A
Primovert Microscope	Carl Zeiss Microscopy GmbH (Oberkochen, Germany=)	N/A
Probe Sonicator horn	Branson Ultrasonic Corporation (USA)	102C, SN: OBU15091229G
Q-Exacte HF Orbitrap mass spectrometer	Thermo Fisher Scientific (Braunschweig, Germany)	NA
Scale	Sartorius Lab Instruments (Göttingen, German)	MSE2202S-000-D0
SDS-Gel running chamber	Bio-Rad Laboratories GmbH (Feldkirchen, Germany)	Mini-Protean Tetra System
Solid-state cooling systems, Thermo Cube	Covaris, Inc. (Woburn, USA)	SN005576
SPD111V Rotor	Thermo Scientific (Schwerte, Germany)	RH40-11 (Eppi's)
SPD111V Rotor	Thermo Scientific (Schwerte, Germany)	(Plates)
SpeedVac Concentrator	Thermo Scientific (Schwerte, Germany)	Savant SPD111V
Stemi 305 Microscope	Carl Zeiss Microscopy GmbH (Oberkochen, Germany=)	SN: 3943000950

## Materials

ThermoMixer C	Eppendorf (Hamburg, Germany)	5382000015
Ultrapure Water System	Thermo Scientific (Schwerte, Germany)	SN:41801405
Ultrasonic Cleaner	VWR International GmbH (Darmstadt, Germany)	USC-T
Universal Vacuum System	Thermo Scientific (Schwerte, Germany)	UVS400A
Vacuum manifold	Waters Corporation (Milford, USA)	S/N 2327
Vortex	Scientific Industries (Bohemia, USA)	Vortex-Genie 2
Water bath	Thermo Scientific (Schwerte, Germany)	SWB25
WCS 2.0 Water Pump	Covaris, Inc. (Woburn, USA)	SN005516
<b>Kits</b>		
Pierce BCA Protein assay	Thermo Fisher (Karlsruhe, Germany)	Ref: 23225; Lot: SL258365
Pierce Quantitative Colorimetric Peptide assay	Thermo Fisher (Karlsruhe, Germany)	Ref: 23275; Lot: TK273137B
Pierce Albumin Depletion Kit	Thermo Fisher (Karlsruhe, Germany)	Ref: 85160; Lot: TH269851
<b>Consumption material</b>		
10 µL tips	Gilson (Limburg, Germany)	Ref: F171100
1000 mL tips	Neptune Scientific (San Diego, USA)	Ref: BT1250
200 µL tips	Gilson (Limburg, Germany)	Ref: F171300
250 microliter tips	Agilent Technologies (Santa Clara, USA)	19477-002
8-row reservoir 32 mL/row	Agilent Technologies (Santa Clara, USA)	201260-100
96 AFA-TUBE TPX Plate	Covaris, Inc. (Woburn, USA)	520272
96-well SuperPlates, skirted	Thermo Scientific (Schwerte, Germany)	AB-2800
Acclaim PepMap C18, 5 µm, 100 Å, 100 µm x 2 cm)	Thermo Fisher Scientific (Braunschweig, Germany)	164564-CMD
Acclaim PepMap RSLC C18, 2 µm, 100 Å, 75 µm x 50 cm	Thermo Fisher Scientific (Braunschweig, Germany)	11342103
AFA-tube TPX Strip Caps	Covaris, Inc. (Woburn, USA)	500639
BioPureSPE Midi 96-well plate Proto 300 C18	The Nest Group, Inc. (Southborough, USA)	Part #: HNS S18V-M; Lot: BN1176-2E697
Drain Caps	Porvair Sciences Ltd. (Wrexham, UK)	ML42115C M5
Drain Caps	Fisherbrand (Schwerte, Germany)	219005; Lot: 030961
Gemini 3 µm C18 110 Å, LC Column 100 x 1 mm	Phenomenex (Aschaffenburg, Germany)	SN: H15-233964
HyperSep C18 unendcapped 96-well plate	Thermo Scientific (Schwerte, Germany)	60300-425, Lot: 517021-BE
MicroLute Combinatorial 96-deep well plate	Porvair Sciences Ltd. (Wrexham, UK)	Lot: 031043
Micro-pillar array columns (µPAC) 50 cm	Pharmafluidics (Ghent, Belgium)	N/A
Microplate 96-well (e.g., BCA)	Greiner Bio-one GmbH (Frickenhausen, Germany)	Ref: 655101
Microplate 96-well conical bottom (High pH)	Thermo Scientific (Schwerte, Germany)	Ref: 249946; Lot: 1253565
Millex-GS 0.22 µm filter	Merck (Darmstadt, Germany)	SLGS033SB

## Materials

Mini-Protean TGX Gels (10-well comb)	Bio-Rad Laboratories (Feldkirchen, Germany)	GmbH	Ref: 456-1084; Lot: L006936A
Mini-Protean TGX Gels (15-well comb)	Bio-Rad Laboratories (Feldkirchen, Germany)	GmbH	Ref: 456-1086; Lot: L006940A
nanoEase MZ Peptide BEH C18 130 Å, 1.7 µm, 75 µm x 250 mm	Waters Corporation (Milford, USA)		Ref: 186008795
Oasis Prime HLB µElution Plate	Waters Corporation (Milford, USA)		Part #: 186008052; Lot: 010737089A
oneTUBE-10 AFA Strip	Covaris, Inc. (Woburn, USA)		520225
PCR Foil Seal	4titude Ltd. (Berlin, Germany)		4ti-0550
PCR-8 stripes	Ratiolab GmbH (Dreieich, Germany)		Ref: 8610040; Lot: 8610040- 463668
PicoTip Emitter	New Objective, Inc. (Woburn, USA)		FS360-20-10-D-20
Reprosil-Pur Basic C18 for analytical columns	Dr. Maisch GmbH (Ammerbuch, Germany)		NA
Sealing Mats	Thermo Scientific (Schwerte, Germany)		AB-0675
Spin-X 0.45 µm filter	Corning Incorporated (Salt Lake City, USA)		Lot: 17418000
X-Pierce film	Sigma-Aldrich (Steinheim, Germany)		Z721646-50EA

### 3. Experimental Methods

#### 3.1. Mass spectrometry methods

The following mass spectrometry methods have been used throughout the study. However, several parameters and instrumental settings were evaluated and modified during this project to achieve optimal performance. Here, most method details are described while varying parameters between experiments are specified within individual paragraphs, if applicable:

##### 3.1.1. Liquid chromatography column setup

Peptides were separated using an Easy NanoLC 1200 fitted with a two-column setup up comprised of a trap column and an analytical column. While the trapping column (Acclaim PepMap C18, 5  $\mu\text{m}$ , 100  $\text{\AA}$ , 100  $\mu\text{m}$  x 2 cm) (Thermo Fisher Scientific) remained mostly constant over time, several analytical columns (**Figure 4E**) were used within this work: initially Acclaim PepMap RSLC C18, 2  $\mu\text{m}$ , 100  $\text{\AA}$ , 75  $\mu\text{m}$  x 50 cm (Thermo Fisher Scientific) was used. Subsequently, we used self-packed analytical columns with Reprosil-Pur Basic C18, 1.9  $\mu\text{m}$ , 100  $\text{\AA}$ , 75  $\mu\text{m}$  x 40 cm material, which was packed into fused silica with an uncoated Pico-Tip Emitter with a 10  $\mu\text{m}$  tip (New Objective) using a pressure bomb (Nanobaum). Here, the spray voltage was set to 2.5 kV to compensate for electrification at the T-piece connection between the trap column, the waste line, and the analytical column. A 50 cm micro pillar-array column ( $\mu\text{PAC}$ , Pharmafluidics) was used in a single-column setup at flow rates between 300 nL/min and 750 nL/min. Finally, we achieved the best performance using a nanoEase MZ Peptide BEH C18 Column, 130  $\text{\AA}$ , 1.7  $\mu\text{m}$ , 75  $\mu\text{m}$  x 250 mm (Waters Corporation). The outlet of the analytical column was directly coupled to an Orbitrap Fusion (Thermo Fisher Scientific) or an Orbitrap Q-Exactive HF (Thermo Fisher Scientific) mass spectrometer via a Pico-Tip Emitter 360  $\mu\text{m}$  OD x 20  $\mu\text{m}$  ID; 10  $\mu\text{m}$  tip (New Objective) and a spray voltage of 2 kV.

##### 3.1.2. Liquid chromatography gradients and data-dependent acquisition (DDA)

The samples were loaded with a constant flow of solvent A at a maximum pressure of 800 bar, onto the trapping column. The maximum pressure was set to 600 bar for the nanoEase MZ Peptide BEH C18 columns. (Waters Corporation). The  $\mu\text{PAC}$  column was limited to a maximal pressure of 200 bar. Solvent A was ddH<sub>2</sub>O, 0.1% (v/v) formic acid (FA) and solvent

B was 80% acetonitrile (ACN) in ddH<sub>2</sub>O, 0.1% (v/v) FA. Peptides were eluted via the analytical column at a constant flow of 300 nL/minute, at 55°C (between 300 nL/min and 750 nL/min for the  $\mu$ PAC). The ion transfer capillary temperature was set to 275°C. Throughout this study, several different gradient lengths were used, in which all settings remained as described for the one hour and 10 minutes method unless otherwise stated in the corresponding paragraphs:

**1 hour 10 minutes:** During elution, the percentage of solvent B was increased linearly from 3 to 8% in 4 minutes, then from 8% to 10% in 2 minutes, then from 10% to 32% in 17 minutes, and then from 32% to 50% in a further 3 minutes. Finally, the gradient was finished with 8 minutes at 100% solvent B, followed by 11 minutes at 96% solvent A. Full scan MS spectra with a mass range of  $m/z$  350 to 1500 were acquired in the Orbitrap with a resolution of 60.000 full width half maximum (FWHM). The ion filling time was set to a maximum of 32 ms with an automatic gain control target of  $3 \times 10^6$  ions. The top 2 or 20 most abundant ions per full scan were selected for a tandem MS ( $MS^2$ ) acquisition. For  $MS^2$  scans, the resolution was set to 15.000 FWHM with automatic gain control of  $1 \times 10^5$  ions and a maximum fill time of 50 ms. The isolation window was set to  $m/z$  2.0, with a fixed first mass of  $m/z$  110, and stepped collision energy (n)ce of 26. The intensity threshold was set to  $2 \times 10^4$  and isotopes, unassigned charges, charge 1, charge 5 to 8, and >8 were excluded. The dynamic exclusion list was set with a maximum retention period of 15 seconds.

**1 hour 25 minutes:** During elution, the percentage of solvent B was increased linearly from 4 to 5% in 1 minute, then from 5% to 27% in 30 minutes, and then from 27% to 44% in a further 5 minutes. Finally, the gradient was finished with 10.1 minutes at 95% solvent B, followed by 13.5 minutes at 96% solvent A. Full scan MS spectra with a mass range of  $m/z$  300 to 1500 were acquired. The ion filling time was set to a maximum of 50 ms with an automatic gain control target of  $3 \times 10^6$  ions. The top 10 most abundant ions per full scan were selected for an  $MS^2$  acquisition. For  $MS^2$  scans, the resolution was set to 15.000 FWHM with automatic gain control of  $5 \times 10^4$  ions and a maximum fill time of 50 ms. The isolation window was set to  $m/z$  1.6, with a fixed first mass of  $m/z$  120, and stepped collision energy (n)ce of 28. The intensity threshold was set to  $1 \times 10^5$  and isotopes, unassigned

## Experimental Methods

charges, and charges of 1 and >8 were excluded. The dynamic exclusion list was set with a maximum retention period of 60 seconds.

**2-hours:** During the elution, the percentage of solvent B was increased linearly from 3 to 8% in 4 minutes, then from 8% to 10% in 2 minutes, then from 10% to 32% in a further 68 minutes, and then to 50% B in 12 minutes. Finally, the gradient was finished with 8 minutes at 100% solvent B, followed by 11 minutes 97% solvent A. The dynamic exclusion was set to 25 seconds.

**3-hours:** During the elution, the percentage of solvent B was increased linearly from 3 to 8% in 4 minutes, then from 8% to 10% in 2 minutes, then from 10% to 32% in a further 118 minutes, and then to 50% B in 22 minutes. Finally, the gradient was finished with 8 minutes at 100% solvent B, followed by 11 minutes 97% solvent A. The dynamic exclusion was set to 35 seconds.

**4-hours:** During the elution, the percentage of solvent B was increased linearly from 3 to 8% in 4 minutes, then from 8% to 10% in 2 minutes, then from 10% to 32% in a further 175 minutes, and then to 50% B in 25 minutes. Finally, the gradient was finished with 8 minutes at 100% solvent B, followed by 11 minutes 97% solvent A. The dynamic exclusion was set to 80 seconds.

### 3.1.3. Proteomics data processing

Raw files were processed using MaxQuant (version 1.5.1.2)<sup>287,288</sup>. The search was performed against the human Uniprot database (20170801\_Uniprot\_homo-sapiens\_canonical\_reviewed; 20,214 entries) using the Andromeda search engine<sup>289</sup> with the following search criteria: enzyme was set to trypsin/P with up to 2 missed cleavages. Carbamidomethylation (C) and oxidation (M) / acetylation (protein N-term) were selected as a fixed and variable modifications, respectively. First and second search peptide tolerances were set to 20 and 4.5 ppm, respectively. Protein quantification was performed using the label-free quantification (LFQ) algorithm of MaxQuant. LFQ intensities were calculated separately for different parameter groups using a minimum ratio count of 1, and the minimum and the average number of neighbors of 3 and 6, respectively. MS<sup>2</sup> spectra were not required for the LFQ comparison. On top, intensity-based absolute quantification (iBAQ) intensities were calculated with a log fit enabled. Identification transfer between



runs via the matching between runs algorithm was allowed with a match time window of 0.3 minutes. Peptide and protein hits were filtered at a false discovery rate of 1%, with a minimum peptide length of 7 amino acids. The reversed sequences of the target database were used as a decoy database. All remaining settings were set as default in MaxQuant. LFQ values were extracted from the protein groups table and log<sub>2</sub>-transformed for further analysis. No additional normalization steps were performed, as the resulting LFQ intensities are normalized by the MaxLFQ procedure<sup>287</sup>. Proteins that were only identified by a modification site, the contaminants, as well as the reversed sequences, were removed from the data set. All consecutive steps were performed in Microsoft Excel, Perseus (version 1.6.1.3)<sup>290</sup>, and the software environment R (version 3.5.1).

### 3.2. Methods taken from joint publications

The following methods have been taken partially or in their entirety from joint publications, as listed below. Every section that was not written entirely by me is indicated with quotation marks:

Hughes, C. S., Moggridge, S., **Mueller, Torsten**, Sorensen, P. H., Morin, G. B., Krijgsveld, J. (2019). „**Single-pot, solid-phase-enhanced sample preparation for proteomics experiments.**” *Nature Protocols* 14: 68-85.

**Mueller, Torsten**, Kalxdorf, M., Longuespée, R., Kazdal, D., Stenzinger, A., Krijgsveld, J. (2020). “**Automated sample preparation with SP3 for low-input clinical proteomics**”. *Molecular Systems Biology* 16(1): e9111.

Hübner, J. M., **Mueller, Torsten**, Papageorgiou, D. N., Mauermann, M., Krijgsveld, J., Russell, R. B., Ellison, D. W., Pfister, S. M., Pajtler, K. W., Kool, M. (2019). „**EZH1/CXorf67 mimics K27M mutated oncohistones and functions as an intrinsic inhibitor of PRC2 function in aggressive posterior fossa ependymoma.**” *Neuro Oncology* 21(7): 878-889.

### **3.2.1. Methods taken from “Single-pot, solid-phase-enhanced sample preparation for proteomics experiments.”**

The evaluation and optimization of the original single-pot, solid-phase-enhance sample preparation (SP3) method (Hughes et al. 2014)<sup>149</sup> have led to an improved protocol version as comprehensively described for different protein input and working volume scenarios in Hughes et al., 2019<sup>291</sup>. In this study, the majority of applications were carried out with 10 µg or less in a working volume smaller than 50 µL.

3.2.1.1. Single-pot, solid-phase-enhanced sample preparation (SP3) bead preparation  
Magnetic beads were prepared by combining 20 µL of both, Sera-Mag Speed Beads A and B (Fisher Scientific, Germany), and washing them one time with 160 µL ddH<sub>2</sub>O and two times with 200 µL ddH<sub>2</sub>O, and re-suspending them in 20 µL ddH<sub>2</sub>O for a final working concentration of 100 µg/µL. The washing steps were carried out using an in-house designed and built magnetic rack for two PCR 8-stripes or in the case of larger volumes in a DynaMag 2 magnet rack (Life technologies). For higher numbers of samples, the preparation of magnetic beads was carried out multiple times to provide at least 2 µL per sample. The pre-washed magnetic beads were combined to a single-tube and vortexed before proceeding with the protein clean-up protocol.

#### 3.2.1.2. SP3 protein clean-up

In brief, 10 µg or less of extracted protein were added to PCR tubes in a total volume of 10 µL lysis buffer (1% sodium dodecylsulfate (SDS), 100 mM ammonium bicarbonate (ABC), pH 8.5). 2 µL of pre-washed magnetic beads, as well as 12 µL 100% ACN, were added to each sample to reach a final concentration of 50% ACN. Protein binding to the beads was allowed for 18 minutes, followed by 2 minutes incubation on a magnetic rack to immobilize beads. The supernatant was removed, and beads were washed two times, with 200 µL of 80% ethanol (EtOH) and one time with 180 µL of 100% ACN. Beads were resuspended in 15 µL of 100 mM ABC and sonicated for 5 minutes in a water bath. Finally, sequencing-grade trypsin was added in an enzyme:protein ratio of 1:20 (e.g., 5 µL of 0.1 µg/µL trypsin in ddH<sub>2</sub>O), and beads were pushed from the tube walls into the solution to ensure efficient digestion. Upon overnight or 4 hours incubation at 37°C and 1000 rpm in a table-top thermomixer, samples were acidified by adding 5 µL of 5% trifluoroacetic acid (TFA) and brief vortexing. Beads were immobilized on a magnetic rack, and peptides were recovered by

transferring the supernatant to new PCR tubes. If necessary, samples were diluted by adding 0.1% FA to reach a suitable peptide concentration of approximately 1 µg/10 µL. At lower peptide concentrations, the entire sample volume was injected. MS injection-ready samples were stored at -20 C.

#### 3.2.1.3. SP3 peptide clean-up

In brief, 10 µL of pre-washed beads and 100% ACN were added to each sample to a final concentration of 95% ACN. Peptides were allowed to bind to the beads for 18 minutes in a thermocycler at 750 rpm, followed by 2 minutes incubation on a magnetic rack (Life technologies, DynaMag 2) to immobilize the beads. The supernatant was removed, and beads were washed 2x with 800 µL of 100% ACN. Beads were air-dried for 2 minutes at 37°C, resuspended in 17 µL of 0.1% FA, and sonicated in a VWR Ultrasonic Cleaner USC-T water bath for 5 minutes. Finally, samples were vortexed, quick-centrifuged, and placed into a magnetic rack to allow a clean transfer of the peptide-containing supernatant to a new reaction tube. MS injection-ready samples were stored at -20°C.

### **3.2.2. Methods taken from “Automated sample preparation with SP3 for low-input clinical proteomics.”**

#### 3.2.2.1. Cell culture of HeLa cells

HeLa cells were cultured in regular DMEM medium (Gibco, Life Technologies) supplemented with 10% fetal bovine serum (Gibco, Life Technologies), 1% of a 100 x penicillin and streptomycin mix (Gibco, Life Technologies), and 1% of 100 x glutamine stock solution (Gibco, Life Technologies). Upon establishment of a stable culture, cells were harvested using trypsin and counted using Bio-Rad TC20 automated cell counter. Cell pellets were stored at -80°C until further use.

For showing the use of the Bravo application starting from limited, small numbers of cells, HeLa cells were harvested, counted, resuspended in lysis buffer (1% SDS, 100 mM ABC pH 8.5), and directly transferred to a 96-well plate. The total volume for different numbers of cells was adjusted using lysis buffer (1% SDS, 100 mM ABC pH 8.5). The entire 96-well plate was sonicated in a water bath for 10 minutes, followed by Benzonase (~40 Units) enzymatic cleavage of DNA and RNA for 15 minutes at 37°C. Subsequently, the buffer was adapted to a final concentration of 1% SDS, 100 mM ABC, 10 mM tris(2-carboxyethyl)phosphine

## Experimental Methods

(TCEP), and 40 mM chloroacetamide (CAA) including protease inhibitor cocktail (PIC) before incubation for 5 minutes at 95°C. The plate was allowed to cool to 23°C before it was transferred to the Bravo deck for the SP3 processing, as described in the “automated SP3 protocol” section.

### 3.2.2.2. HeLa protein standard preparation

Cell pellets of ~11.9 million cells were resuspended in 1 mL of lysis buffer (1% SDS, 100 mM ABC pH 8.5, and 50 µL 25x PIC) and probe sonicated for 5 times 20 seconds at a frequency of 10% using a Branson Sonifier. Cell lysates were kept on ice in-between cycles to avoid overheating. DNA or RNA contaminants were cleaved using 250 Units of Benzonase for 15 minutes at 37°C and 750 rpm. Subsequently, the buffer was adapted to a final concentration of 1% SDS, 100 mM ABC, 10 mM TCEP, and 40 mM CAA, including PIC, before incubation for 5 minutes at 95°C in a CHB-T2-D ThermoQ heating device (Hangzhou BIOER Technologies). Reduced and alkylated proteins were quantified using a bicinchoninic acid assay (BCA) assay and stored at -20°C until further use in manual and automated SP3 processing.

### 3.2.2.3. Pulmonary adenocarcinoma (ADC) sample collection

All pulmonary ADC specimens used for this study were obtained from the Thoraxklinik at Heidelberg University and diagnosed according to the criteria of the 2015 WHO Classification of lung tumors at the Institute of Pathology at Heidelberg University<sup>206</sup>. Tissue procession to formalin-fixed and paraffin-embedded (FFPE) tissue sections was carried out by the tissue bank of the National Center for Tumor Diseases (NCT; project: # 1746; # 2818) in accordance with its ethical regulations approved by the local ethics committee.

A multiregional sample set consisting of 2-4 samples of eight tumors was constructed as described previously<sup>292</sup>. In short, a formalin-fixed central section of each tumor was segmented into multiple 5 x 5 mm regions according to a Cartesian grid. Ink marks ensured the retention of the original orientation of each segment during sample processing. Tumor regions considered for analysis were selected in accordance with the tumor size (larger tumor corresponds to more regions), different histological growth patterns as well as sufficient tumor cell content ( $\geq 10\%$ ). An experienced pathologist determined the histological growth pattern with the predominant portion in each segment. For each tumor,

two to four different growth patterns were excised. Samples were analyzed in replicates using one 5  $\mu\text{m}$  section after deparaffinization as input, respectively. For deparaffinization, the sections were incubated for 20 minutes at 80°C, followed by three times 8 minutes incubation in Xylol and EtOH, consecutively. Finally, the sections were incubated in ddH<sub>2</sub>O for 30 minutes before the tissue was scratched off and collected in a well. Replicates were excised as consecutive cuts of the same region having the highest possible similarity.

#### 3.2.2.4. Automated SP3 protocol (autoSP3)

As a reference, the SP3 protocol was carried out manually, as described in the corresponding paragraph 4.2.1.2 (Hughes 2019)<sup>291</sup>. In the automated version of the SP3 protocol, the Bravo system is programmed to process 96 samples simultaneously, carrying out all handling steps including reduction and alkylation of proteins, aliquoting of magnetic beads, protein clean-up by SP3, protein digestion, and peptide recovery. The core SP3 protocol is available in combination with reduction and alkylation either as a single-step using a TCEP/ CAA mix for 5 minutes at 95°C (**Figure 6D**) or as a two-step protocol using, for example, dithiothreitol (DTT)/ CAA consecutively with 30 minutes incubation for each reaction at 60°C and 23°C, respectively (**Figure 6D**). A shortened version is available that consists of the core SP3 protocol while omitting on-deck reduction and alkylation (**Figure 6D**), saving time due to slow heating of the heating block (altogether taking one hour for heating and cooling), instead performing this off-deck (taking 5 minutes and 30 seconds to reach working temperature and 5 minutes for incubation) in a PCR thermocycler (CHB-T2-D ThermoQ, Hangzhou BIOER Technologies) prior to initiation of the automated protocol. In addition, the PCR thermocycler provides lid heating, which prevents any unwanted evaporation or variation in the sample volume. This latter protocol (Protocol C, **Figure 6D**) was used in the work presented here. Each protocol is designed for a starting sample volume of 10  $\mu\text{L}$ , which can easily be varied in the protocol files to add respective amounts of organic solvent to reach higher than 50% and to remove the resulting volume after protein binding. Next, either protocol A, B, or C (**Figure 6D**) aliquot 5  $\mu\text{L}$  of a suspension of washed magnetic beads to protein samples previously collected in a 96-well plate. Different to the manual protocol (bead working concentration 100  $\mu\text{g}/\mu\text{L}$ ), the suspension of washed beads is prepared to have a working concentration of 50  $\mu\text{g}/\mu\text{L}$  to allow more robust pipetting. Next, the respective volume of 100% ACN (20  $\mu\text{L}$  in A; 25  $\mu\text{L}$  in B, 15  $\mu\text{L}$  in C) is

## Experimental Methods

added to each sample followed by 18 minutes incubation off the magnetic rack with cycles of agitation at 1500 rpm and 100 rpm for 30 seconds and 90 seconds, respectively. Upon binding of the proteins to the beads, the sample plate is incubated on the magnetic rack for further 5 minutes to allow magnetic trapping of beads inside each well. Here, the beads will form a ring at the wall of each well, slightly above the bottom. The removal of any supernatant in the protocol is performed using well-specific tips in two consecutive steps to ensure complete liquid removal. Next, beads are washed two times with 200  $\mu\text{L}$  of 80% EtOH and one time with 171.5  $\mu\text{L}$  of 100% ACN. Due to the limited 200  $\mu\text{L}$  pipetting volume of the Bravo and the limited reagent space, the respective washing volumes of 80% EtOH and 100% ACN were added in 4 and 7 consecutive steps of 50  $\mu\text{L}$  and 24.5  $\mu\text{L}$ , respectively, with in-between shaking at 500 rpm or 250 rpm for 30 seconds. Upon removal of residual washing solvents, the beads are resuspended in 35  $\mu\text{L}$  of 100 mM ABC and 5  $\mu\text{L}$  of 0.05  $\mu\text{g}/\mu\text{L}$  pre-prepared trypsin in 50 mM acetic acid to avoid autolysis. Of note, in the dilution series experiments, the trypsin amount was reduced to avoid abundant peptide features resulting from its autolysis. In a final shaking step at 1500 rpm for 60 seconds, the trypsin solution is mixed with the sample, and the plate is transferred to the heating deck position for incubation at 37°C. Subsequently, the plate was manually sealed and transferred to a PCR cycler to avoid lid condensation during a 4-hour incubation at 37°C. Next, after completion of either protocol A, B, or C and exchange of used pipette tips, a short protocol is provided for peptide acidification and recovery of LC-MS injection-ready samples to a new 96-well plate (**Figure 6D**). Alternatively, as used in this study, peptide acidification and recovery can be performed manually. Therefore, each sample was acidified by adding 5  $\mu\text{L}$  of 5% TFA solution, sonicated in a water bath for 5 minutes to swirl the settled beads, and incubated on a magnetic rack for further 2 minutes. Finally, the peptide-containing supernatant was recovered into a new 96-well plate without transferring the beads. If necessary, samples were either diluted or directly frozen at -20°C until MS acquisition. Optionally peptide quantification assays (colorimetric assay kit, Thermo Scientific) were carried out using the Bravo liquid handling system.

### 3.2.2.5. Quantitative proteomics analysis of FFPE tissue

For proteomic analysis, 5  $\mu\text{m}$  FFPE tissue sections were collected in stripes of 8 PCR tubes, centrifuged at 15.000  $\times g$  for 10 minutes to ensure that FFPE slices are at the bottom of the

tube, and stored at 4°C until further processing. Next, each tissue section was carefully reconstituted in 20 µL lysis buffer (4% SDS, 100 mM ABC, pH 8.5), sonicated at 4°C for 25 cycles of 30 seconds on and 30 seconds off in a Pico Bioruptor, and heated for one hour at 95°C. Samples were spun down and subjected to a second round of sonication and heating. The Pico Bioruptor (Diagenode SA) was equipped with a house-made tube holder, which allows the simultaneous processing of 28 samples. Subsequently, PCR tubes were centrifuged at 15.000 x g for 3 minutes, and the buffer was adjusted to a final concentration of 1% SDS, 100 mM ABC, 10 mM TCEP, and 40 mM CAA, including PIC. Samples were heated for 5 minutes at 95°C to denature proteins and to reduce and alkylate cysteine residues. Cooled to RT and again centrifuged at 15.000 x g for 3 minutes, 10 µL of each sample was further processed by our automated SP3 sample clean-up procedure, as described above. Here, protein digestion was allowed for 16 hours overnight before stopping the reaction by acidification to 0.5% with TFA. The peptide-containing supernatant was recovered to a new 96-well plate without transferring the beads. MS injection-ready samples were stored at -20°C, and about 25% of each sample was later used for data acquisition.

#### 3.2.2.6. Proteomics data acquisition

For HeLa standard measurements, samples were diluted with solvent A (0.1% FA in ddH<sub>2</sub>O) to enable the injection of 1 µg in 10 µL volume. Peptides were separated using the Easy NanoLC 1200 fitted with a trapping (Acclaim PepMap C18, 5 µm, 100 Å, 100 µm x 2 cm) and an analytical column (Acclaim PepMap RSLC C18, 2 µm, 100 Å, 75 µm x 50 cm). The outlet of the analytical column was coupled directly to a Q-Exactive HF Orbitrap mass spectrometer (Thermo Fisher Scientific). Data were acquired using the one hour 25 minutes method as described in chapter 3.1.2.

For FFPE lung ADC measurements, about 25% of each sample was used for direct injection. Peptides were separated using the Easy NanoLC 1200 fitted with a trapping (Acclaim PepMap C18, 5 µm, 100 Å, 100 µm x 2 cm) and a self-packed analytical column (Reprosil-Pur Basic C18, 1.9 µm, 100 Å, 75 µm x 40 cm). The outlet of the analytical column was coupled directly to a Q-Exactive HF Orbitrap (Thermo Fisher Scientific) mass spectrometer. Data were acquired using the 2-hours methods, as described in chapter 3.1.2.

## Experimental Methods

### 3.2.2.7. Proteomics data processing

The data processing was carried out as comprehensively described in paragraph 4.1.3. Additional analyses were performed as follows: the differential expression analysis of the ADC samples was performed using Limma moderated t-statistics (R package version 3.36.3)<sup>293</sup>. Here, the technical replicates and the patient-dependent batch effect were taken into account within the applied model. Proteins with a Benjamini-Hochberg-adjusted p-value lower than 0.05 and an absolute log<sub>2</sub>-fold change higher than 1 were considered as significantly changing. The resulting lists of significantly regulated proteins were subjected to a gene ontology (GO)-term enrichment analyses using the STRING: functional protein association network database<sup>294</sup>. The gene set enrichment analyses (GSEA) were performed using R package fgsea<sup>295</sup> (version 1.6.0) with a p-value ranking of proteins, gene sets defined by the REACTOME pathway database (R package ReactomePA version 1.24.0)<sup>296</sup>, the minimum size of gene sets set to 15, the maximum size of gene sets set to 500, and the number of permutations set to 10.000. The t-distributed stochastic neighbor embedding (t-SNE) analyses were performed using R package tsne<sup>297</sup> (version 0.1-3) with a perplexity set to 2 and the number of iterations set to 5000.

### 3.2.2.8. Intra-day and inter-day precision

To test the precision of SP3 sample handling, we followed guidelines of the European Pharmacopoeia and the European Medicines Agency for the number of replicates necessary to validate our method<sup>298,299</sup>. Specifically, we validated automated SP3 by an intra-day and inter-day component by processing a total of six 96-well plates with 10 µg protein of a HeLa batch lysate in each well in the morning and the afternoon of three different days, over roughly one month, resulting in a total of 575 individual samples. Five randomly picked samples per plate (10 samples per day) were selected for direct LC-MS analysis on the day of sample generation and a second technical-repeat injection of all 30 samples in a single batch acquisition. The number of samples per plate to be analyzed was chosen as a fair compromise to determine the precision of our sample processing with a reasonable amount of data acquisition time. The selected samples allowed the evaluation of the inter-day precision and intra-day precision while taking different processing times, plates, and buffers into account (robustness). The second technical injection in one batch allowed us to evaluate the influence of longitudinal MS performance. Lastly, for the



comparison of manual SP3 sixteen times, 10 µg protein of a HeLa batch lysate were processed manually at the bench.

#### 3.2.2.9. Sensitivity of autoSP3

To evaluate the lower limit of processing capabilities of the Bravo SP3 setup, we generated starting material dilution series as follows: A) a dilution series of our standard HeLa protein stock, ranging from 10 µg to ~5 ng in 1:2 dilution steps (10 µg, 5 µg, 2.5 µg, 1.25 µg, ~625 ng, ~312 ng, ~156 ng, ~78 ng, ~39 ng, ~19 ng, ~10 ng, and ~5 ng). The dilution series was generated and processed in four replicates on the same 96-well plate (12 concentrations and n=4). B) a dilution series starting from small numbers of counted cells that were directly transferred to a 96-well plate, ranging from 10.000 down to 10 cells. The dilution series was generated and processed in two plates à four replicate series (7 concentrations and n=8). Here, the European Pharmacopoeia recommends a minimum of three concentrations à three replicates<sup>298</sup>. In addition, two empty control injections were performed upfront of the data acquisition of each dilution series. The dilution series were measured in blank-interspaced blocks from lowest to highest concentrated samples to avoid potential carry over between injections.

#### 3.2.2.10. Assessment of cross-contamination

To assess potential cross-contamination between samples, we processed 24 wells of 10 µg standard HeLa protein stock interspaced with 24 empty controls. Seven peptide-containing samples and eleven empty controls were randomly selected for direct LC-MS analysis. The number of samples to be analyzed was chosen as a fair compromise to determine potential carry over between wells during our sample processing with a reasonable amount of data acquisition time.

### **3.2.3. Methods taken from „EZHIP/CXorf67 mimics K27M mutated oncohistones and functions as an intrinsic inhibitor of PRC2 function in aggressive posterior fossa ependymoma.”**

#### 3.2.3.1. Cell Culture of HEK293T cells

“HEK293T cells were cultured in regular DMEM medium (Gibco, Life Technologies) supplemented with 10% fetal calf serum. The medium was exchanged every second day, and cells were split at least once per week.”

## Experimental Methods

### 3.2.3.2. Production of Lentiviral particles and generation of stable cell lines

“Lentiviral constructs were generated by replacing the Ngn2 gene of the FUW-TetO-Ngn2-T2A-puromycin construct published by Zhang et al. with DNA sequences encoding for the CXorf67 full-length protein or CXorf67 truncates carrying a C-terminal FLAG-HA-tag. Lentiviruses were produced by co-transfecting lentiviral constructs with psPAX2 and pMD2.G into low-passage HEK293T cells using FugeneHD (Promega). The medium was replaced 24 hours after transfection. On the next day, lentivirus-containing supernatant was harvested and passed through a 0.45 µm filter before being directly added to the target cells. To allow induction of gene expression by administration of Doxycycline, cells were additionally co-transduced with a rtTA carrying lentivirus. 24 hours after infection, protein expression was induced by addition of 1 µg/mL of Doxycycline followed by selection with Puromycin. HEK293 cells were selected with 1 µg/mL of Puromycin. For continuous protein expression, Doxycycline was replenished every two to three days.”

### 3.2.3.3. Co-Immunoprecipitation for mass spectrometry and western blot

“For co-immunoprecipitation (Co-IP) followed by mass spectrometry (MS) analysis, cells were resuspended in lysis buffer (20 mM Tris-HCl pH 8, 200 mM NaCl, 1 mM EGTA, 1 mM ethylenediaminetetraacetic acid (EDTA), 1% Triton X-100), vortexed and incubated on ice for 30 minutes. Subsequently, cellular debris was pelleted by centrifugation and the supernatant was transferred to a new tube. The supernatant was then pre-cleared for one hour at 4°C using mouse IgG agarose beads. Afterwards, beads were pelleted by centrifugation and the supernatant was again transferred to a separate tube. The supernatant was then incubated with FLAG-M2 affinity gel overnight at 4°C. Mouse IgG agarose beads used for pre-clearing were washed twice with lysis buffer and then twice with PBS. Next, the beads were resuspended in 30 µL of elution buffer (50 mM NH<sub>4</sub>HCO<sub>3</sub>, 15 mM DTT, 0.1% SDS) and boiled at 95°C for 5 minutes. Eluted proteins were saved for MS analysis to determine the protein background. The next day, proteins were eluted from the FLAG-M2 affinity gel using the same procedure as for the mouse IgG agarose beads followed by MS analysis. For Co-IP followed by western blot analysis, proteins were only eluted from the FLAG-M2 affinity gel using 40 µL of western blot elution buffer (10 µL of 4x NuPAGE™ LDS Sample Buffer, 4 µL of 10x NuPAGE™ Sample Reducing Agent and 26 µL of lysis buffer).”

#### 3.2.3.4. Nuclear extraction and western blot analysis

“For the separation of nuclear and cytoplasmic fractions, cells were harvested, washed once with ice-cold PBS, and incubated in 2ml of swelling buffer (10 mM Tris-HCl pH 6.8, 5 mM KCl, 1 mM MgCl<sub>2</sub>) on ice for 20 minutes. Next, cell membranes were ruptured using a douncer. Nuclei were spun down at 1000 x g for 10 minutes at 4°C, and the supernatant was saved as the cytoplasmic fraction. The nuclei were washed once with swelling buffer and pelleted again via centrifugation. Then, nuclei were resuspended in Laemmli buffer (62.5 mM Tris-HCl pH 6.8, 10% Glycerol, 3% SDS, 150 mM DTT, 250 Units Benzonase) and cooked at 95°C for 10 minutes. Finally, insoluble debris was removed by centrifugation, and the supernatant was saved as the nuclear fraction.

Whole-cell lysates, nuclear, and cytoplasmic extracts or eluted fractions from Co-immunoprecipitation experiments were separated on a 4-12% Bis-Tris gradient gel (Invitrogen) followed by transfer onto a 0.2 µm PVDF membrane. The membrane was then blocked for 30 minutes, with 5% milk in Tris-buffered saline-Tween 0.05% (TBS-T). Primary antibody incubation was performed overnight at 4°C. The next day, the membrane was washed three times with TBS-T, followed by incubation with a secondary HRP-conjugated antibody for one hour at 23°C. Finally, the membrane was washed three times with TBS-T and covered in ECL Western Blotting Detection Reagent (GE Healthcare Life Sciences) followed by detection of chemiluminescence using the Intas Chemostar ECL Imager device (Intas Science Imaging). Primary antibodies used for western blot analysis were targeted against H3K27me3 (ab6002, abcam, 1:1000), histone H3 (ab1791, abcam, 1:5000), FLAG-tag (F1804, Sigma-Aldrich, 1:1000), EZH2 (D2C9, Cell Signaling Technology, 1:1000), SUZ12 (D39F6, Cell Signaling Technology, 1:1000), EED (09-774, Merck, 1:1000), β-tubulin (#2146, Cell Signaling Technology, 1:1000), or Lamin B1 (ab16048, abcam, 1:1000). Secondary antibodies used were goat anti-mouse-HRP (ab6789, abcam, 1:5000) and goat anti-rabbit-HRP (ab6721, abcam, 1:3000).”

#### 3.2.3.5. Protein digestion and SP3 peptide clean-up of Co-IP samples

Samples obtained from the Co-Immunoprecipitation (stored at -20°C) were reduced with DTT (10 mM final concentration) at 45°C for 30 minutes. Subsequently, proteins were alkylated using 40 mM final concentration of CAA at 23°C for 30 minutes. Reduced and alkylated proteins were digested overnight at 37°C using 0.65 µg sequencing-grade

## Experimental Methods

modified trypsin in 100 mM ABC. Next, samples were further processed by the SP3 peptide clean-up procedure (Hughes *et al.*, 2014; Hughes *et al.*, 2019)<sup>149,291</sup> as briefly described in paragraph 3.2.1.3. MS injection-ready samples were stored at -20°C.

### 3.2.3.6. Mass spectrometry data acquisition

Peptides were separated using an Easy NanoLC 1200 fitted with a trapping (Acclaim PepMap C18, 5 µm, 100 Å, 100 µm x 2 cm) and a self-packed analytical column (Reprosil-Pur Basic C18, 1.9 µm, 100 Å, 75 µm x 40 cm). The C18 material was packed into fused silica with an uncoated Pico-Tip Emitter with a 10 µm tip (New Objective) using a Nanobaum pressure bomb. The outlet of the analytical column was coupled directly to an Orbitrap Fusion (Thermo Fisher Scientific) mass spectrometer. Solvent A was ddH<sub>2</sub>O, 0.1% (v/v) FA and solvent B was 80% ACN in ddH<sub>2</sub>O, 0.1% (v/v) FA. The samples were loaded with a constant flow of solvent A at a maximum pressure of 800 bar, onto the trapping column. Peptides were eluted via the analytical column at a constant flow of 0.3 µL/minute, at 55°C, using the 2-hours gradient described in chapter 3.1.2. Peptides were introduced into the mass spectrometer at a positive spray voltage of 2.5 kV. The ion transfer tube temperature was set at 275°C. Full scan MS spectra with a mass range of m/z 375 to 1500 were acquired in the Orbitrap with a resolution of 120,000 FWHM. The filling time was set to a maximum of 50 ms with an automatic gain control target of 1 x 10<sup>6</sup> ions. Intensities were filtered at a threshold of 5 x 10<sup>3</sup>. The dynamic exclusion list was set with a maximum retention period of 40 seconds and a mass tolerance of 10 ppm, high and low, respectively. Isotopes, unassigned charges, and charges of 1, 5 to 8, and >8 were excluded. MS<sup>2</sup> scan properties were set to use the quadrupole isolation mode with a window of m/z 1.6. Higher-energy collision-activated dissociation (HCD) was selected as an activation type at a percentage collision energy of 33%. MS<sup>2</sup> scans were performed in the ion trap at a rapid scan rate with a first mass at m/z 120 and an automatic gain control target of 1 x 10<sup>4</sup> ions. The maximum injection time was set to 50 ms with ion injection for all available parallelizable time. MS<sup>2</sup> spectra were acquired in a centroid data type.

### 3.2.3.7. Mass spectrometry data processing

Raw files were processed using MaxQuant (version 1.5.1.2)<sup>287,288</sup>. The search was performed against the human Uniprot database (201708\_Uniprot\_homo-sapiens\_canonical\_reviewed; 20214 entries) using the Andromeda search engine<sup>289</sup> with

the following search criteria: enzyme specificity was set to trypsin/P with up to 2 missed cleavages. Carbamidomethylation (C) was selected as a fixed modification; oxidation (M) and acetylation (protein N-term) were set as variable modifications. The first and second search peptide tolerances were set to 20 and 4.5 ppm, respectively. The protein quantification was performed using the label-free quantification algorithm of MaxQuant. LFQ intensities were calculated using a minimum ratio count of 1, and a minimum and an average number of neighbors of 3 and 6, respectively. MS/MS were required for the LFQ comparison, and the stabilization of large LFQ ratios was enabled. iBAQ intensities were calculated with a log fit. Peptide and protein hits were filtered at a false discovery rate of 1%, with a minimum peptide length of 7 amino acids. The reversed sequences of the target database were used as a decoy database. The remaining parameters of MaxQuant were left at the default settings. LFQ values were extracted from the protein Groups table and  $\log_2$ -transformed for further analysis. iBAQ values were extracted from the protein Groups table and  $\log_{10}$ -transformed for further analysis. The MaxQuant protein groups' output table was filtered for contaminants, reverse hits, and hits only identified by site. No additional normalization steps were performed, as the resulting LFQ intensities are normalized by the MaxLFQ procedure<sup>287</sup>. LFQ intensities were  $\log_2$ -transformed, and ratios were calculated for each construct over its respective IgG control counterpart. Proteins without a positive ratio in the full-length experiment and those without a positive ratio in all three construct experiments were filtered from the protein list. Subsequently, LFQ ratios were uploaded in Perseus (v. 1.5.3.0)<sup>290</sup>, and hierarchical clustering was performed using Euclidean distances with an average linkage for both row and column tree clustering, respectively.

### **3.3. Additional experimental methods**

#### **3.3.1. Cell culture of stable cell lines**

*The culturing of A375, RPMI-7951, UACC-62, and ISTMEL-1 cell lines was carried out by Dr. Gertjan Kramer.*

HeLa, HEK-293, and MCF7 were cultured in regular DMEM medium (Gibco, Life Technologies) supplemented with 10% fetal bovine serum (Gibco, Life Technologies), 1% of a 100 x penicillin & streptomycin mix (Gibco, Life Technologies), and 1% of 100 x glutamine stock solution (Gibco, Life Technologies). A375, RPMI-7951, UACC-62, and ISTMEL-1 cells

## Experimental Methods

were cultured in DMEM medium (Gibco, Life Technologies) supplemented with 10% fetal calf serum (Gibco, Life Technologies), 1 mM sodium pyruvate (Thermo Fisher Scientific), 25 mM HEPES (Thermo Fisher Scientific), 2 mM GlutaMAX (Gibco, Life Technologies), and 1% of a 100 x penicillin & streptomycin mix (Gibco, Life Technologies). Upon establishment of a stable culture, cells were harvested using trypsin and counted using Bio-Rad TC20 automated cell counter. Cell pellets were stored at -80°C until further use.

### **3.3.2. Cell culture of patient-derived EPN tumor cell lines**

*The culturing of patient-derived EPN tumor cell lines was carried out by either Dr. Jens Huebner (Global EPN and CXorf67 related experiments), Dr. Kendra Maaß, or Mieke Roosen (Extracellular vesicle related experiments) from the collaborating groups of Prof. Marcel Kool and Prof Kristian Pajtler at the DKFZ.*

In total, four different patient-derived EPN tumor cell lines were available corresponding to two out of nine EPN subgroups: namely BT214 and EPD210 from PF-EPN-A, as well as BT165 and EP1NS from ST-EPN-RELA. The PF-EPN-A cell lines were cultured in 1:100 Laminin-coated (Sigma-Aldrich) flasks in NeuroCult NS-A basal medium (Stem Cell Technologies) supplemented with 2 mM L-Glutamine (Gibco, Life Technologies), 75 µL/mL Bovine serum albumin (BSA), 10% NeuroCult NS-A proliferation supplement (Stem Cell Technologies), and 1x antibiotic/antimycotic reagent (Life Technologies). Growth factors were added to 50 mL aliquots of medium: 20 ng/mL recombinant human EGF and FGF (both from Peprotech). The ST-EPN-RELA cell lines were cultured in Geltrex-coated (Life Technologies) flasks in Neurobasal medium A (Life Technologies) supplemented with 1 µg/mL Heparin (Sigma-Aldrich), 2 mM L-Glutamine (Gibco, Life Technologies), and 1% penicillin & streptomycin mix (Gibco, Life Technologies). Growth factors were added to 50 mL aliquots of medium: 1 mL of B-27 supplement minus vitamin A (Life Technologies), 20 ng/mL recombinant human EGF, and FGF (both from Peprotech).

For splitting, the ST-EPN-RELA BT165 cell line and both PF-EPN-A cell lines were detached using Accutase (Sigma-Aldrich) for 5 minutes at 23°C. Only the ST-EPN-RELA EP1NS cell line was detached using Accumax (Thermo Fisher Scientific) for 5 minutes at 37°C. All cell lines were cultured at 37°C at 5% CO<sub>2</sub>. All cell lines were regularly tested for mycoplasma. Cell pellets were stored at -80°C until further use.

### 3.3.3. Additional methods for lysis of cells and tissue for protein extraction

#### 3.3.3.1. RapiGest SF Surfactant protein extraction

Each vial of 1 mg RapiGest SF Surfactant was dissolved in 1 mL of 50 mM triethylammonium bicarbonate (TEAB) and mixed thoroughly to achieve a 0.1% solution. For large sample batches, multiple vials of RapiGest SF Surfactant were dissolved and combined. PIC (Roche Diagnostics) was added to 1x final concentration before adding 100  $\mu$ L to each tissue sample. Samples were kept on-ice and probe sonicated for 2 times 15 seconds at 10% frequency using a probe sonicator (Branson). The sample viscosity was used as quality control for sufficient DNA shearing. Subsequently, lysates were centrifuged at 15,000 x g for 30 minutes at 4°C (Eppendorf 5430R centrifuge) to pellet residual cell- or tissue debris. The protein content was determined using a BCA protein assay (Pierce) according to the manufacturer's instructions. Based on the BCA results across all samples, the smallest possible volume for a certain amount of protein, e.g., 10 to 20  $\mu$ g, was selected for further processing. The same amount of protein per sample was transferred to PCR 8-stripes and balanced to the same volume with 50 mM TEAB. Next, samples were incubated for 5 minutes at 95°C, cooled to 23°C, and reduced with a final concentration of 5 mM DTT for 30 minutes at 60°C. Upon incubation, samples were quickly vortexed and spun down prior to alkylation with a final concentration of 15 mM CAA for 30 minutes at 23°C. Proteins were digested overnight using a 1:50 ratio of trypsin (in 50 mM TEAB) to protein at 37°C and 500 rpm. On the next morning, the digestion reaction was stopped by acidification to 0.5% TFA, followed by 30 minutes incubation at 37°C. Subsequently, samples were centrifuged at 15,000 x g for 30 minutes to pellet the precipitated RapiGest SF Surfactant. The peptide-containing supernatant was transferred to new tubes for storage at -20°C until data acquisition.

#### 3.3.3.2. Urea-based protein extraction

For Urea-based protein extraction, a 10 M Urea stock solution was prepared as follows. In total, 24 g of Urea was mixed with 4 mL of 1 M ABC and topped up to 40 mL volume with ddH<sub>2</sub>O. The mixture was fully dissolved by additional vortexing and heating with warm tap water. One complete PIC tablet was added, and the solution was subsequently filtered through a 0.22  $\mu$ m syringe filter (Millex-GS). The filtered stock solution was aliquoted in 2 mL tubes and stored at -80°C.

## Experimental Methods

Cell or tissue samples were resuspended in 100  $\mu$ L 10 M Urea and mixed at 800 rpm for 10 minutes at 23°C. Subsequently, samples were either sonicated in a water bath or using a probe sonicator (Branson) for 2 times 15 seconds at 10% frequency. Here, samples were kept on-ice to avoid overheating and carbamylation of cysteines as a result. Samples were centrifuged at 18.214 x g (Eppendorf 5430R centrifuge) for one hour at 10°C, and the resulting supernatant was transferred to a new 2 mL tube. Precaution was taken to not transfer any of the residual sticky DNA at the tube bottom. The protein content was determined using a BCA protein assay (Pierce) according to the manufacturer's instructions. On the basis of the BCA results across all samples, the smallest possible volume for a certain amount of protein, e.g., 10 to 20  $\mu$ g, was selected for further processing. Samples were further reduced with a final concentration of 5 mM DTT for 30 minutes at 40°C. Upon incubation, samples were quickly vortexed and spun down prior to alkylation with a final concentration of 15 mM CAA for 30 minutes at 23°C. Before protein digestion, samples were diluted to a final Urea concentration below 1.6 M for enzyme compatibility. Proteins were digested overnight using a 1:50 ratio of trypsin (in 50 mM TEAB) to protein at 37°C and 500 rpm. On the next morning, the digestion reaction was stopped by acidification to 0.5% TFA, followed by 30 minutes incubation at 37°C. The resulting peptide samples were further cleaned up and desalted using the Oasis protocol described in chapter 3.3.9.

### **3.3.4. DNA extraction**

DNA was extracted from mouse kidney tissue (~ 2 mg wet weight). The tissue pieces were cut in a glass Petri dish on dry-ice using a commercial razor blade. The cut tissue parts were transferred within 600  $\mu$ L TNES buffer (10 mM Tris-HCL pH 7.5, 400 mM NaCl, 100 mM EDTA, and 0.6% SDS) and 35  $\mu$ L Proteinase K (Thermo Scientific) to a 1 mL Dounce homogenizer. Samples were digested overnight in a 2 mL tube at 50°C after 20 pestle strokes for complete tissue solubilization. Next, 166.7  $\mu$ L 6 M NaCl was added, and samples were vigorously vortexed for 20 seconds and centrifuged at 20.238 x g for 10 minutes at 23°C. The resulting supernatant was transferred and mixed with 800  $\mu$ L ice-cold 100% EtOH. The tube was inverted several times to ensure sufficient gentle mixing and starting of DNA precipitation. The samples were further centrifuged at 15147 x g for 20 minutes at 4°C. The supernatant was discarded, and the DNA pellet was washed with 500  $\mu$ L 100% EtOH and several times inversion of the sample tube. The washing step was performed for a second



time with 70% EtOH before the remaining EtOH was entirely removed by spinning down and discarding the supernatant and air-drying the DNA pellet for 5 minutes. The pellet was resuspended in ddH<sub>2</sub>O, and the amount of extracted DNA [ng/μL] was measured using a Nanodrop 1000 spectrophotometer device to read the absorbance at 260/280 nm and 260/230 nm. (Thermo Fisher Scientific). Besides, samples were checked by loading different amounts on agarose gels, as described previously.

### **3.3.5. E. coli spike-in sample preparation**

*E. coli* lyophilized sample (Bio-Rad) was resuspended in ddH<sub>2</sub>O to achieve a stock concentration of 2 μg/μL. 100 μL (200 μg) were incubated at 95°C for 5 minutes, followed by reduction and alkylation using DTT (10 mM final concentration) at 37°C for one hour and CAA (40 mM final concentration) at 23°C for 45 minutes at 500 rpm. Reduced and alkylated proteins were digested overnight at 37°C in a table-top thermomixer at 700 rpm using sequencing-grade modified trypsin (Promega) in ddH<sub>2</sub>O. Upon overnight protein digestion, each sample was acidified to a final concentration of 1% TFA (Biosolve Chimie). Subsequently, stocks of spike-in samples were prepared with a constant amount of HeLa peptides and increasing spike-ins of 0%, 3%, 4.5%, 6%, 7.5%, and 9% *E. coli* peptides (n= 3). MS injection-ready samples were stored at -20°C.

### **3.3.6. Agarose Gels for DNA visualization**

Agarose gels were prepared by dissolving 1.2 g of agarose (Sigma) in 100 mL TAE buffer (50 mM EDTA, 2 M Tris, and 1 M glacial acetic acid) in an Erlenmeyer flask. The solution is heated for 2 minutes in a commercial microwave and cooled to room temperature. Immediately upon reaching near 23°C SYBR safe DNA stain mix (Invitrogen) was added to the gel solution. Eight-well combs were inserted, and gels were poured in a gel running device (Biostep GmbH) to polymerize.

Samples were mixed in a 1:6 ratio with loading dye (Thermo Scientific). Combs were removed, and samples were loaded. Gels typically ran for about 30 minutes at 140 V. DNA marker (Thermo Scientific) was used in each gel.

### **3.3.7. SDS-Gels for protein visualization**

Samples were incubated for 10 minutes at 95°C with a 1x final concentration of Laemmli buffer stock solution (Bio-Rad Laboratories, Inc.) supplemented with 50 mM DTT.

## Experimental Methods

Subsequently, samples were cooled to room temperature before loading into SDS-gels (either a 10-comb or a 16-comb commercial solutions). Precision plus protein standard (Bio-Rad Laboratories, Inc.) was used in each gel. Gels were either run at 120 or 160 V for a time range of 45 to one hour and 45 minutes until the running front almost reached the gel bottom. Gels were run in Mini-Protean Tetra system chambers (Bio-Rad Laboratories, Inc.) using a Power PAC universal power supply (Bio-Rad Laboratories, Inc.). Gels were removed from the running chamber and washed once with ddH<sub>2</sub>O and fixed for one hour at room temperature in 50% EtOH, 10% acetic acid, and 40% ddH<sub>2</sub>O. Three consecutive ddh<sub>2</sub>O washes removed the fixation solution before leaving the gel in ddh<sub>2</sub>O overnight for rehydration.

That followed, SDS-gels were washed twice with ddh<sub>2</sub>O and sensitized using 0.02% sodium thiosulfate for 1 minute at room temperature and three ddh<sub>2</sub>O washes. The silver staining was performed for 20 minutes at 4°C using a 0.1% silver nitrate solution with 0.02% formaldehyde added right before use. After an additional three consecutive ddh<sub>2</sub>O washes, gels were developed using 3% sodium carbonate with freshly added 0.05% formaldehyde. The development was terminated quickly at sufficient signal intensity by a single ddh<sub>2</sub>O wash, followed by 5 minutes incubation in 5% acetic acid. Gels were stored for a short term in 1% acetic acid. Gels were digitalized using a scanner.

### **3.3.8. Cell-surface labeling and protein enrichment**

The cell-surface labeling and protein enrichment was performed using an adapted version of Kalxdorf et al., 2017<sup>300</sup>. In brief, frozen ST-EPN-RELA ependymoma tissue samples with an average wet weight of 9.16 mg were transferred in 500 µL 1x PBS to a Dounce homogenizer for six gentle pestle strokes. The resulting homogenate and residual solid structures were transferred to a 1.5 mL tube and centrifuged at 1000 x g for 1 minute at 23°C to remove the supernatant. The remaining pellet was washed once with 1 mL of 1x PBS before performing a second round of centrifugation at 1000 x g for 1 minute at 23°C. The supernatant was again discarded, and the cell-debris pellet was resuspended in 1 mL of ice-cold 1x PBS with 1 mM sodium metaperiodate to oxidize carbohydrates for ten minutes with occasional gentle mixing on-ice and in the dark. Each sample was washed once with ice-cold 1x PBS and centrifugation at 1000 x g for 1 minute at 23°C to remove the supernatant. Next, biotinylation was performed with 400 µL of 1x PBS, 1 mM EZ-Link

Alkoxyamine-PEG4-Biotin, and 10 mM Aniline for ten minutes and occasional gentle mixing on-ice and in the dark. Subsequently, the supernatant was removed, and samples were washed once more with ice-cold 1x PBS. Upon centrifugation at 1000 x g for 1 minute at 23°C and discarding of the supernatant, the labeled cell-surface protein samples were frozen at -80°C until further processing.

In the meanwhile, the high capacity neutravidin-agarose resin was blocked to achieve trypsin-resistance by an in-house developed protocol. In brief, 210 µL of neutravidin-agarose bead slurry was transferred to a 2 mL tube. Subsequently, the slurry was washed three consecutive times with 1 mL of 1x PBS, 0.1% Tween, and centrifugation at ~100 x g for 1 minute to remove the supernatant. Unless otherwise stated, all following washing steps were performed in the same way (three times, 1 mL of 1x PBS, 0.1% Tween, and centrifugation at ~100 x g for 1 minute). The beads were resuspended in 800 µL 1x PBS, 0.1% Tween plus additional 200 µL 1 M NaOH (final pH > 12). The solution was further transferred to a new 2 mL tube containing 8.6 mg of 1,2-cyclohexadione (CDH), inverted several times to dissolve the CDH completely, and incubated with constant stirring in a PTR-35 multi-rotator (Grant-bio Instruments) for 5 hours at 25°C. Upon incubation, the bead slurry was centrifuged at ~100 x g for 1 minute, and the supernatant was discarded. The beads were further washed three times and resuspended in 500 µL of 200 mM sodium cyanoborohydride in 1x PBS and 500 µL 4% formaldehyde in 1x PBS, followed by two hours of incubation at 23°C with occasional vortexing. The reaction was stopped by adding 500 µL of 1 M Tris-HCl, pH 7.6, and three consecutive washes. Finally, the washed beads were resuspended in 210 µL 1x PBS, 0.1% Tween. Protease-resistant neutravidin-agarose beads were stored at 4°C until further processing.

Previously labeled cell-surface proteins were thawed, resuspended in 100 µL 4% SDS, 100 mM ABC pH 8.5, and heated for 5 minutes at 95°C. Samples were further probe sonicated (Branson) for 10 seconds at 10% frequency, kept on-ice, and topped with 900 µL 1x PBS. Next, 30 µL of prepared protease-resistant neutravidin-agarose bead slurry was added to a Microlute combinatorial 96 deep-well filter plate (Porvair Sciences Ltd.), sucked through using a vacuum manifold (Waters Corporation), and washed with 1 mL of 1x PBS. The diluted samples were added to the conditioned 96 deep-well plate with the bottom closed using a sealing mat (Porvair Science Ltd.). The top of the plate was closed using a sealing

## Experimental Methods

mat (Thermo Scientific) for incubation with overhead rotation in a PTR-35 multi-rotator (Grant-bio Instruments) for two hours at 23°C. That followed, the liquid was sucked through using a vacuum manifold, and the plate was centrifuged at 314 x g for 1 minute (Heraeus Megafuge 40, Thermo Scientific) to remove the residual liquid. The neutravidin-agarose beads were washed successively with 300 µL steps and sucking through with the vacuum manifold as follows: three times with 400 mM NaCl, 0.4% SDS, 20 mM ABC; eight times with 400 mM NaCl, 20 mM ABC; and eight times with 2 M Urea, 50 mM ABC. Proteins were consecutively reduced and alkylated for 30 minutes at 23°C using 30 µL of 45 mM DTT, 100 mM ABC and 30 µL 100 mM CAA, 100 mM ABC, respectively. The residual liquid was removed by centrifugation at 314 x g for 2 minutes before and after five additional washes with 300 µL 2 M Urea, 50 mM ABC. Proteins were digested on-bead in 60 µL of 1.5 M Urea, 60 mM ABC, and 0.42 µg of trypsin. For overnight digestion at 23°C and 500 rpm on top of a ThermoMixer C (Eppendorf), the top and bottom of the 96 deep-well plate were closed. Peptides were eluted into a new 96-well plate by centrifugation at 314 x g for 2 minutes, followed by a second elution with 50 µL 100 mM ABC. The peptide samples were dried in a vacuum centrifuge at 45°C, resuspended in 100 µL 0.1% FA, and cleaned using a HyperSep C18 plate as described in chapter 3.3.9. The resulting samples were stored at -20°C until data acquisition.

### **3.3.9. Desalting and clean-up of peptide samples**

For the desalting and clean-up of peptide samples, either Oasis PRiME HLB µElution plates (Waters Corporation) or HyperSep C18 (Thermo Scientific) were used as indicated in the individual paragraphs. For both, the packed material was activated through consecutive washes of 100 µL as follows: 100% ACN, then 80% ACN, 0.1% FA, and then 0.1% FA. In each step, the liquid was sucked through using a vacuum manifold. Subsequently, samples were loaded in 0.1% FA and sucked through twice. The bound peptide samples were washed three times with 200 µL 0.1% FA and finally eluted with two times 50 µL 80% ACN, 0.1% FA. Eluted peptides were dried in a vacuum centrifuge at 45°C, resuspended in 50 to 100 µL 0.1% FA depending on the sample amount, and stored at -20°C until data acquisition.

### **3.3.10. High pH reversed-phase fractionation of proteomic samples**

The high pH fractionation of peptide samples was performed using an Agilent Infinity 1260 HPLC system (Agilent) equipped with a Phenomenex Gemini 3 µM, 110 Å, C18, 100 x 1 mm

column (Phenomenex). Solvent A was changed from 20 mM ammonium hydroxide to 20 mM ammonium formate during the course of this project. Unless otherwise stated, all data within this work were generated using ammonium formate. Solvent B was 100% ACN due to its low absorbance at 206 nm, among other organic solvents. Peptides were eluted from the Phenomenex column at a constant flow of 0.1 mL/minute. During the elution, the percentage of solvent B was constant at 0% for 2 minutes, then increased linearly from 0% to 65% in 58 minutes, and then from 65% to 85% in 2 minutes. Finally, the gradient was finished with 5 minutes at 85% solvent B, followed by 8 minutes at 0% solvent B. Fractions were collected during the first 60 minutes of the gradient for every 1.5 minutes, resulting in a total of 40 fractions per sample. Sample pick-up and fraction collection were performed at a constant 4°C. Each sample's peptide map was monitored through the absorbance at 206 nm. The resulting peptide fractions were dried in a vacuum centrifuge at 50°C, further concatenated to either 8, 16, 24, or 32 individual fractions, and resuspended in 0.1% FA. The actual number of concatenated fractions is indicated in the individual paragraphs. Peptide fractions were stored at -20°C until data acquisition.

### 3.3.11. Extracellular vesicle sample preparation

*Every section that was not written entirely by me is indicated with quotation marks. Besides the proteomic sample preparation, the majority of the EV-related work was performed by Dr. Kendra Maaß or Mieke Roosen.*

#### 3.3.11.1. Patient-derived EPN cell culture

“Patient-derived EPN tumor cell lines were used as an experimental model. The ST-EPN-RELA cell lines, BT165 and EP1NS, were cultured in Geltrex-coated (Life Technologies) flasks in Neurobasal medium A (NBA, Life Technologies) supplemented with 1 µg/mL Heparin (Sigma), 2 mM L-Glutamine (Gibco, Life Technologies), 1% penicillin & streptomycin (Gibco, Life Technologies), and growth factors added to aliquots. The growth factors were added to 50 mL aliquots of the above-mentioned supplemented medium in the following concentrations: 1 mL B-27 supplement minus vitamin A (50x, Life Technologies) and 20 ng/mL recombinant human EGF (Peprotech) and 20 ng/mL recombinant human FGF (Peprotech). The PF-EPN-A cell lines, BT214 and EPD210, were cultured in 1:100 Laminin- (Sigma, in PBS) coated flasks in NeuroCult NS-A Basal Medium (Human, Stem Cell Technologies) supplemented with 2 mM L-Glutamine, 75 µg/mL Bovine Serum Albumin (BSA), 10% NeuroCult NS-A proliferation supplement (Human, Stem Cell Technologies), and growth factors added to aliquots. The growth factors were added in the following concentrations to 50 mL aliquots: 20 ng/mL recombinant human EGF and FGF. Human fetal astrocytes were cultured in 1:150 matrix-gel (Corning, in DMEM) coated flasks in DMEM high glucose medium supplemented with 10% fetal calf serum depleted of exosomes, 1% glutamax (Gibco, Life Technologies), 1% sodium pyruvate (Gibco, Life Technologies), and 1% N2 supplement (Gibco, Life Technologies). For splitting, BT165 and the PF-EPN-A cell lines were detached by 5 minutes incubation with Accutase (Sigma) at 23°C. The cell line EP1NS was detached by a 5 minutes incubation with Accumax (Thermo Fisher Scientific) at 37°C. The astrocytes were detached with a 5 minutes incubation with trypsin (Sigma Aldrich) at 37°C. All cell lines were cultured at 37°C with 5% CO<sub>2</sub>. The cells were regularly tested for mycoplasma.”

#### 3.3.11.2. Exosome and microvesicle isolation

“Exosomes and microvesicles were isolated from cell culture supernatant of four cell lines: namely BT214 and EPD210 from PF-EPN-A, as well as BT165 and EP1NS from ST-EPN-RELA.

80 mL of supernatant was collected from confluent flasks. The supernatant was centrifuged at 2000 x g for 20 minutes at 4°C to remove cell debris and apoptotic cells (2k pellet). Afterwards, the supernatant was transferred to ultra-centrifugation tubes (Beckman Coulter) coated with 70% EtOH for ultra-centrifugation in a Beckman Optima L-70 ultra-centrifuge with a SW28 rotor. The supernatant was centrifuged at 10.000 x g for 20 minutes at 4°C (10k pellet). The supernatant was further transferred to new EtOH-coated ultra-centrifugation tubes and centrifuged at 100.000 x g for two hours at 4°C (100k pellet). Afterwards, the supernatant was discarded. The microvesicles (10k pellet) and the exosomes (100k pellet) were vortexed for 1 minute and frozen at -20°C or -80°C depending on the further processing steps. When the vesicles were stained with BODIPY TR ceramide dye (Invitrogen, D7540), 2 µL of dye was added to 100 µL PBS and incubated for 20 minutes at 37°C. The BODIPY TR ceramide dye has absorption and emission maxima of ~589 nm and 617 nm, respectively.

After the initial centrifugation steps, the purity of the pellets was enhanced by loading the samples on an IZON 35 nm qEV single size exclusion column. The first fraction was collected after 1 mL, and the subsequent fractions of 0.2 mL were collected according to the manufacturer protocol. The protein amounts in the different fractions were measured with the Qubit protein assay (Molecular Probes, Life Technologies, Q33211). The vesicles were permeabilized with 0.2% SDS and vortexed for 30 seconds. The remaining steps were carried out according to the manufacturer protocol. The fractions with the peak protein amount were used for further analysis.”

#### 3.3.11.3. Immunogold electron microscopy

“After a glow-discharge in a Baltec SCD005 Sputter Coater, 300 Mesh Formvar-carbon coated Copper grids (Plano) were floated on 10 µL drops of isolated vesicles solution, for 20 minutes at 23°C. After 3x washes with 15 µL PBS drops, the vesicles were blocked with Aurion blocking solution for Au-conjugates (PB) 1:10 in PBS for 20 minutes at 23°C, incubated with primary antibody (Ab) (Mouse- $\alpha$ -CD63, Santa Cruz, diluted 1:50 in PB) solution for 30 minutes at 23°C, washed 6x with PB, incubated with linker-IgG (Rabbit- $\alpha$ -mouse, Dako Denmark, 1:150 in PB) for 40 minutes at 23°C, washed 6x with PB, incubated with Protein-A-Gold 5 nm (UMC Utrecht, PAG 5 nm/S, 1:50 in PB) for 50 minutes, and washed again 6x with PB. After 2x washes with PBS, the samples were fixed in 1%

## Experimental Methods

Glutaraldehyde in PBS for 7 minutes at 23°C, washed again with PBS, and 4x with ddH<sub>2</sub>O. Finally, a negative stain of 1% aqueous uranyl acetate was applied for 2 minutes, and after a ddH<sub>2</sub>O wash, the grids were air-dried. In negative controls, the primary antibody was omitted. All samples were investigated with a Zeiss EM900 transmission electron microscopy (TEM), at an 85.000x magnification. Images taken were further processed with ImageJ (version 1.52a)<sup>301</sup>.”

### 3.3.11.4. Nanoparticle tracking analysis (NTA)

“Two µL of exosomes or MVs were diluted in sterile-filtered PBS and visualized using the LM10 NTA device (Malvern Instruments). Each sample was measured 5 times for 60 seconds (Screen Gain 1.0, camera level 11) to obtain particle concentration and size distribution.”

### 3.3.11.5. Qubit protein quantification assay

“The approximate protein content of exosomes and MVs in the isolation solutions was determined using a Qubit Protein Assay (Molecular Probes, Life Technologies). For this, the vesicles were lysed with 3x Laemmli buffer, and its protein content was later dissolved in 0.2% SDS and vortexed for 30 seconds. For the rest of the assay, the manufacturer’s guidelines were followed.”

### 3.3.11.6. Protein digestion and SP3 protein clean-up of EV samples

The extracellular vesicle fractions described in chapter 3.3.11.2 were further processed using SP3. Therefore, equal protein amounts per sample (previously determined by Qubit assay) were transferred to PCR stripes, and the volume reduced in a vacuum centrifuge at 55°C. Subsequently, samples are resuspended in a small volume (between ~10 µL and 20 µL) of 4% SDS and 100 mM ABC. Samples were sonicated in a Pico Bioruptor (Diagenode SA) using 25 cycles of 30 seconds on and off at 4°C. Subsequently, the buffer was adapted to a final concentration of 1% SDS, 100 mM ABC, 10 mM TCEP, and 40 mM CAA, including PIC, before incubation for 5 minutes at 95°C in a CHB-T2-D ThermoQ heating device. That followed, samples were rested to reach 23°C for subsequent SP3 processing. The SP3 protocol was carried out as previously described in chapter 3.2.1.2. Samples are stored at -20°C and data acquired using the 2-hours method.



### **3.4. Additional data analysis**

#### **3.4.1. Differential expression**

The differential expression analysis of all samples (unless otherwise indicated) were performed using Limma moderated t-statistics (R package version 3.36.3)<sup>293</sup>. If applicable, the technical replicates and the patient-dependent batch effect were taken into account within the applied model. Proteins with a Benjamini-Hochberg-adjusted p-value lower than 0.05 and an absolute log<sub>2</sub>-fold change higher than 1 were considered as significantly changing unless otherwise indicated in the corresponding chapter. The resulting lists of significantly regulated proteins were used for further analysis.

#### **3.4.2. Gene set enrichment analysis (GSEA) and gene ontology (GO)**

Lists of proteins or significantly regulated proteins were subjected to GO-term enrichment analyses using the STRING: functional protein association network database<sup>294</sup>. The GSEA were performed using R package fgsea<sup>295</sup> (version 1.6.0) with a p-value ranking of proteins, gene sets defined by the REACTOME pathway database (R package ReactomePA version 1.24.0)<sup>296</sup> or the Broad Institute gene sets<sup>302</sup>, the minimum size of gene sets set to 15, the maximum size of gene sets set to 500, and the number of permutations set to 10.000.

#### **3.4.3. t-SNE and umap**

The t-SNE analyses were performed using R package tsne<sup>297</sup> (version 0.1-3) with a perplexity set to 2 and the number of iterations set to 5000. The uniform manifold approximation and projection (umap) analyses were performed using the R package umap<sup>303</sup> (version 0.2.3) with neighbors (equivalent to perplexity) were set to 3 and the number of iterations set to 5000.

#### **3.4.4. Gene- and protein expression correlation**

The gene expression data were used from Pajtler et al., 2015<sup>73</sup>. To see gene/protein-specific correlation differences, we calculated the median gene and protein intensities. Next, we plot the mean intensity correlations between both transcriptome and proteome data and determine a mean linear regression model. Subsequently, the linear model is used to determine the deviation for every gene/protein from this general model. Based on the residuals, we could define which genes/proteins are significantly deviating from the model. The residuals are calculated by predicting gene expression from protein expression using

## Experimental Methods

the linear model and subtracting the gene expression from predicted intensities for every gene and every sample. The residuals were normally distributed. Next, proteins were determined that show a significant deviation between the proteome and transcriptome by comparing residuals per gene/protein that deviate from zero. The significance was determined using Limma moderated t-statistics (R package version 3.36.3)<sup>293</sup>. The threshold was set to a Benjamini-Hochberg-adjusted p-value lower than 0.05, and an absolute log<sub>2</sub>-fold change higher than 1 were considered.

### **3.4.5. Copy number variation (CNV) correlation to gene- and protein expression**

The CNV and gene expression data were used from Pajtler et al., 2015<sup>73</sup>. Data were batch corrected using the Limma moderated t-statistics (R package version 3.36.3)<sup>293</sup>. The gender for each sample was predicted from the X- and Y-Chromosome intensities. This was utilized to correct for gender-specific expression changes across all samples. To visualize whether CNVs per EPN subgroup result in changes of gene or protein expression, we calculated the intensities as CNV-varied (CNV > +/- 0.2) and CNV-stable (-0.2 < CNV < 0.2) samples per tumor subtype. Next, we calculate for both gene and protein expression intensities, the ratios between each sample, and the median intensity of the CNV-stable samples. This was ordered by the chromosomal position per gene and their observed intensity relative to the CNV-stable samples. CNV segments were plotted with a line for every sample and chromosomal region, highlighted in colors corresponding to the CNV status: blue= neutral, red= deletion (CNV < -0.15), or green= amplification (CNV > 0.15). Individual dots indicate the mean gene expression at a corresponding genomic locus for each sample (color-coded for its CNV status) relative to the gene expression in the CNV neutral samples.

### **3.4.6. Multi-omics factor analysis (MOFA)**

Proteome data were median normalized. Transcriptome and DNA-Methylation data were batch- and gender-corrected. The top 20% variable proteins (1745 proteins), top 20% variable genes (3890 genes), and the top 1% variable CpG sites (4260 CpGs) were selected for MOFA<sup>135</sup>. The multi-omics factor analysis (MOFA)+ framework was used with default settings and the number of factors set to 15.

## 4. Results

### Copyright Disclaimer

*The work presented in this thesis was carried out by me, Torsten Müller, and under the supervision of Prof. Jeroen Krijgsveld. Throughout this thesis, the pronoun “we” is used to refer to myself, my supervisor, if applicable the helping hand of co-workers, and in some cases, collaborators. Collaborations and their contribution are specifically highlighted in the main text as well as in the method sections. As a general rule, all mass spectrometry-based “proteomics” experiments within this work were performed by me. Dr. Mathias Kalxdorf supported the bioinformatic analysis. Results and experiences outlined in the first chapter, “4.1.2”, involving the optimization of SP3, have contributed to an updated version of the protocol in Hughes et al., Single-pot, solid-phase-enhanced sample preparation for proteomics experiments, Nature Protocols, 2019<sup>291</sup>. The second chapter of this thesis, “4.2”, summarizes the automation of SP3 that has been published in Müller et al., Automated sample preparation with SP3 for low-input clinical proteomics, MSB, 2020<sup>304</sup>. Also, a preprint of this publication is accessible on bioRxiv under the same title<sup>305</sup>. The collaborative project about CXorf67, outlined within chapter, “4.3.2.1”, has been published in Hübner et al., EZHIP/CXorf67 mimics K27M mutated oncohistones and functions as an intrinsic inhibitor of PRC2 function in aggressive posterior fossa ependymoma, Neuro Oncology, 2019<sup>257</sup>. The second collaborative project about extracellular vesicle cargo in ependymoma, outlined within the last chapter, “4.3.4”, has not been published yet. In this thesis, we solely focus on the comparison to our global EPN proteome data. Our collaborators plan to publish this work with a significant contribution of our proteomics data. The remaining results outlined throughout the last chapter, “4.3”, have not been published and are originally presented in this thesis. We note that plans exist to publish these results, irrespective of the embargo period on this thesis, and with significant input from our collaborators with regard to the provision of previously acquired sequencing data as well as clinical context.*

#### 4.1. Proteomic profiling in systems medicine

In the first chapter of this thesis, we systematically evaluated and optimized all relevant steps that are crucial for proteomic sample preparation workflows in light of clinical integration. During this step-wise development of a scalable workflow, we specifically focused on the extraction of proteins from a wide range of sample material and low quantities, the subsequent protein processing to generate peptides, their chromatographic separation, and LC-MS data acquisition. Due to distinct advantages of the in-house developed single-pot, solid-phase-enhanced sample preparation (SP3) method (Hughes *et al.*, 2014; Hughes *et al.*, 2019)<sup>149,291</sup>, we specifically tailored the optimization of preceding steps for SP3 compatibility and its subsequent automation (Chapter 4.2).

Parts of the following chapters, including Figures and Tables, were taken in part or their entirety from the joint publications listed below.

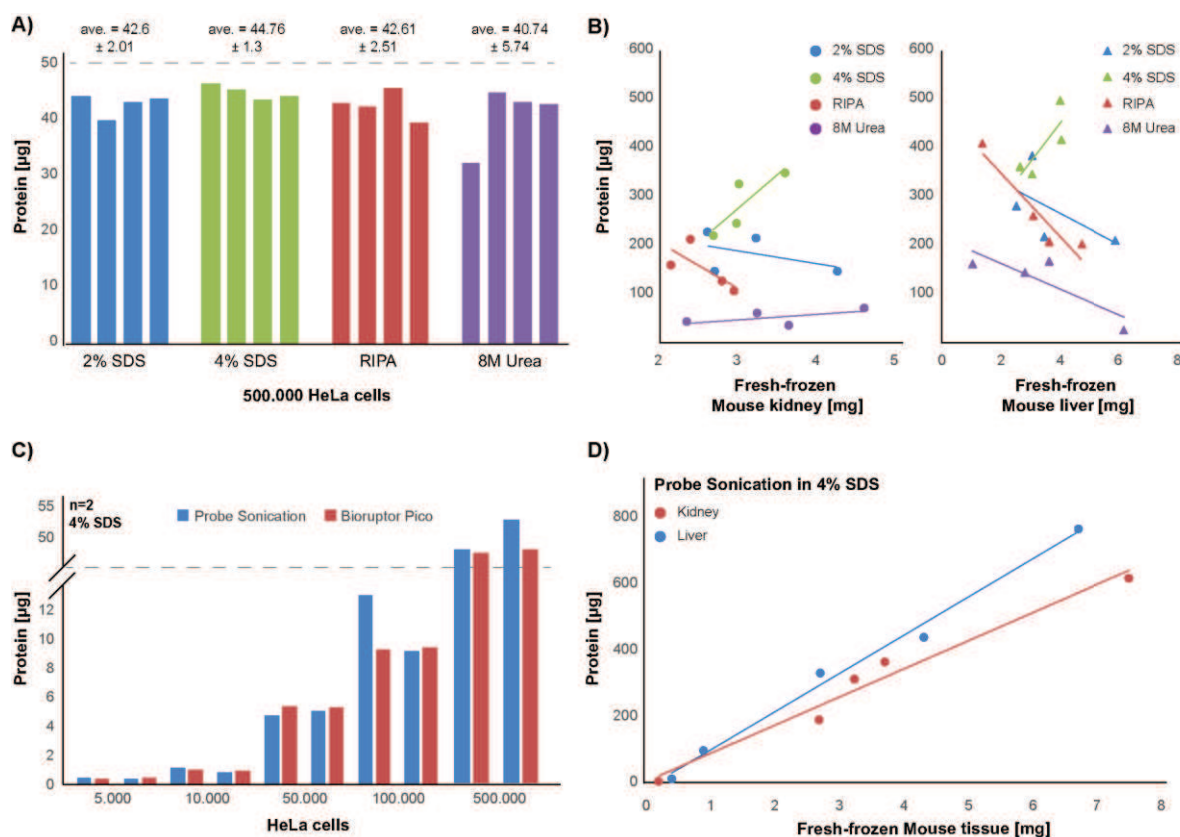
Hughes, C. S., Moggridge, S., **Mueller, Torsten**, Sorensen, P. H., Morin, G. B., Krijgsveld, J. (2019). „**Single-pot, solid-phase-enhanced sample preparation for proteomics experiments.**” *Nature Protocols* 14: 68-85.

**Mueller, Torsten**, Kalxdorf, M., Longuespée, R., Kazdal, D., Stenzinger, A., Krijgsveld, J. (2020). “**Automated sample preparation with SP3 for low-input clinical proteomics.**” *Molecular Systems Biology* 16(1): e9111.

##### 4.1.1. Cell- or tissue lysis and protein extraction

The efficient extraction of proteins is the first critical step in any proteomics methodology. Dependent on the type and quantity of a specimen, different approaches are commonly used in a research environment<sup>151,155,156,158,291,306–308</sup>. They are typically comprised of different lysis buffer combinations and mechanical disruption strategies to facilitate the efficient breakup of cell- or tissue structures and to release proteins. Here, different proteins, such as transmembrane proteins, can significantly differ in their physicochemical properties, requiring different solubilization strategies to avoid any selectivity. In this process, the majority of lysis buffers are comprised of chemicals, such as chaotropes, salts, or detergents, to ensure the disruption of the phospholipid bilayer and that proteins remain in-solution for subsequent proteolytic digestion. For the integration of a method into

clinical practice, a vast number of requirements need to be fulfilled beyond high performance, reproducibility, and cost-efficiency.



**Figure 1: Evaluation of cell- and tissue lysis, followed by protein extraction.** A) Cell lysis facilitated by four different lysis buffers (2% SDS, 4% SDS, RIPA, and 8 M Urea) and quantification of extracted protein mass. B) Fresh-frozen tissue (mouse kidney and liver) lysis facilitated by four different lysis buffers, as in panel A, and quantification of extracted protein mass. C) Lysis of different cell quantities in 4% SDS with additional mechanical disruption using a probe sonicator (blue) or Bioruptor Pico (red), and quantification of extracted protein mass. D) Lysis of different tissue quantities (mouse liver (blue) and kidney (red)) in 4% SDS with additional mechanical disruption using a probe sonicator, and quantification of extracted protein mass.

In an initial attempt, we aimed to avoid any mechanical disruption of cell- or tissue samples to achieve broad applicability without the need for specialized equipment and well-trained personnel, aiming for a lossless integration with SP3 in a single tube. Four commonly used lysis buffers (2% SDS, 4% SDS, radioimmunoprecipitation assay (RIPA) buffer, and 8 M Urea) were used for sample solubilization and protein extraction from 500,000 HeLa cells (**Figure 1A**) and mouse kidney and liver tissue (**Figure 1B**). Consistent protein yields could be extracted from cells irrespective of the lysis buffer and within the expected range of ~0.1 ng protein per cell. This was verified in four randomly selected cell lines, in three varying quantities (5000, 50,000, and 500,000 cells), and using 2% and 4% SDS (**Supplementary Figure 1A**). The amount of protein that could be extracted from fresh-frozen tissue did not

## Results

scale linearly with the input material and conform with yield expectations (~ 10% of tissue weight corresponds to proteins) (**Figure 1B**)<sup>309</sup>. The highest and lowest quantities of protein were extracted using 4% SDS and 8 M Urea, respectively.

Despite the overall reasonable protein yields, the mechanical-free sample lysis appeared to be incompatible with the following SP3 protocol (**Supplementary Figure 1B**). The processing of different protein quantities resulted in >50% sample losses for most conditions. The highest consistency in relative and absolute peptide recoveries were achieved using 4% SDS. During SP3, proteins and nucleic acids compete for the binding capacity of the beads and omitting a proper mechanical DNA and RNA shearing could account for the weak recovery of peptides (**Supplementary Figure 1C**; further discussed in chapter 4.1.2). Indeed, we could show that enzymatic cleavage of nucleic acids can be achieved by using Benzonase (**Supplementary Figure 1C**) and that SP3 recoveries can be improved as a result (further discussed in chapter 4.1.2). This required the adaption of detergent concentrations in the lysis buffer, now including RapiGest SF surfactant, to allow the enzymatic activity of Benzonase. The lower detergent concentrations reduced the protein yield by more than 50% in most conditions and thereby did not qualify for minute amounts of sample (**Supplementary Figure 1D**).

We further assessed mechanical disruption to achieve efficient lysis and protein extraction with sufficient DNA and RNA shearing. Therefore, a classical probe sonicator (Branson) and a Bioruptor Pico (Diagenode SA) were utilized to process varying numbers of HeLa cells (5000, 10.000, 50.000, 100.000, and 500.00 cells) (**Figure 1C**) and different amounts (sub-mg to >7 mg) of fresh-frozen mouse kidney and liver tissue (**Figure 1D**). The quantities of extracted proteins were reproducible and in line with our expectations, while additionally exhibiting a linear correlation between protein yield and tissue input in comparison to mechanical-free lysis. The one-by-one processing using the probe sonicator remains insufficient, taking the anticipated goal of a scalable workflow into account.

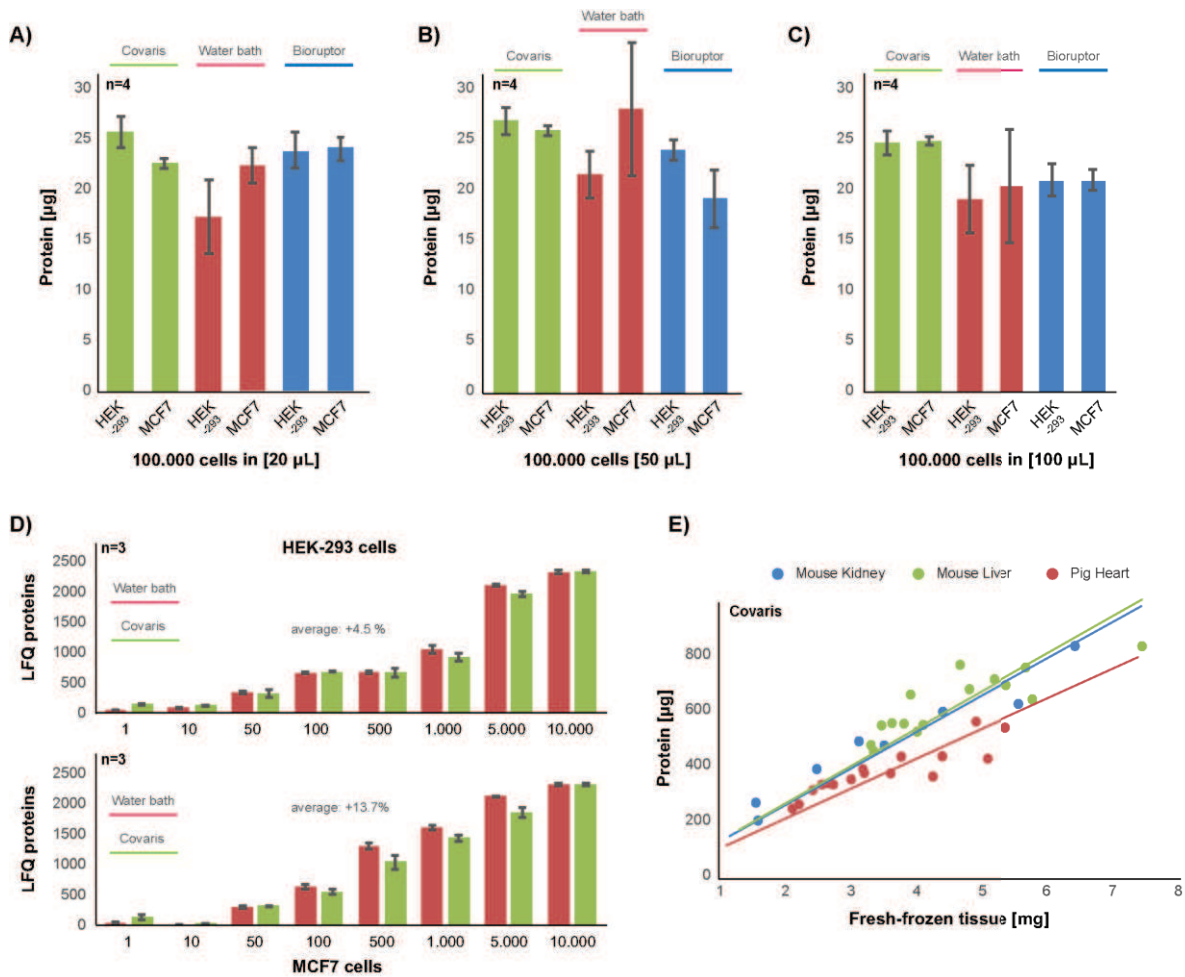
Next, we compared the processing efficiency of all methods accessible to us, which can process several samples at a time, namely a LE220R-plus focused-ultrasonicator (Covaris Ltd, UK), a standard water bath (Thermo Scientific), and the Bioruptor Pico (Diagenode SA). The latter is limited to the simultaneous processing of 32 samples, utilizing a custom-made PCR-tube adapter, while both others can run in 96-well formats. The LE220R-plus focused-

ultrasonicator utilizes adaptive focused acoustics (AFA) to dis-integrate tissue, extract proteins, and enhance the efficiency of DNA shearing by the delivery of highly-controlled and reproducible energy. Hereinafter they are referred to as Covaris, water bath, and Bioruptor. The Covaris achieved the highest protein extraction efficiency and reproducibility from 100.000 HEK-293 cells or MCF7 cells using 4% SDS, irrespective of the processing volume (20  $\mu$ L **Figure 2A**, 50  $\mu$ L **Figure 2B**, and 100  $\mu$ L **Figure 2C**). Overall minimized volumes are desirable to reduce the plastic surface and the associated loss of proteins. The Bioruptor performed equally well at low volumes with comparable standard deviation, for example,  $\sigma$ = 3.11% (1.03%) and 3.21% (1.7%) for processing of HEK-293 (MCF7) cells using the Covaris or Bioruptor. The water bath processing resulted in the lowest protein yield and largest variability ( $\sigma$ = 7.2% [HEK-293] and 3.54% [MCF7]). This is because the water bath sonication suffers from an incomplete shearing of DNA, as opposed to the Covaris (data not shown). Only for minute amounts of cells (1 to 10.000) in small processing volumes (<20  $\mu$ L) this is not evident and results in an average of 4.5% and 13.7% more quantified proteins for HEK-293 and MCF7 cells, respectively, using the water bath and SP3 (**Figure 2D**). All other sample types (fresh-frozen or FFPE tissue, and higher cell numbers) require processing with the Covaris to achieve sufficient lysis. The amount of protein that could be extracted from fresh-frozen tissue (pig heart (n=16), mouse liver (n=16), and mouse kidney (n=8)) scaled linearly with the mass of wet tissue input material, liberating  $\sim$ 100  $\mu$ g protein per mg heart tissue, and  $\sim$ 130  $\mu$ g per mg liver and kidney tissue, as expected from the literature (**Figure 2E**)<sup>309</sup>.

In summary, both mechanical-free and mechanical cell disruption methods achieve effective cell lysis, irrespective of the buffer composition. This does not hold for the extraction of proteins from fresh-frozen tissue where 4% SDS outcompetes all other buffers in terms of protein yield and its linearity. However, the mechanical-free sample lysis is not sufficient to shear nucleic acids and thus remains incompatible with downstream SP3 processing or minute amounts of sample. In contrast, probe sonication or Bioruptor processing, result in reproducible, high and linear protein yields, but remain limited in the sample throughput. The most efficient and high-throughput processing of all sample types and quantities could be achieved using the Covaris, which seamlessly integrates with

## Results

downstream SP3. Altogether, SDS detergent shows the best extraction efficiencies and remains the most frequently used component of common lysis buffers.



**Figure 2: Evaluation of high-throughput cell- and tissue lysis methods followed by protein extraction.** A-C) Lysis of 100,000 HEK-293 and MCF7 cells in 20 µL (A), 50 µL (B), or 100 µL (C) of 4% SDS with additional mechanical disruption using a Covaris LE220R-plus (green), a sonication water bath (red), or a Bioruptor Pico (blue), and quantification of extracted protein mass. D) Lysis of different HEK-293 or MCF7 cell quantities (10,000 to 1 cell) in 4% SDS with additional mechanical disruption using a Covaris LE220R-plus (green) or a sonication water bath (red), followed by SP3 processing and LC-MS. E) Lysis of different tissue quantities (mouse liver (green) and kidney (blue), and pig heart (red)) in 4% SDS with additional mechanical disruption using a Covaris LE220R-plus, and quantification of extracted protein mass. Panel E modified from Mueller et al., *Mol. Syst. Biol.*, 2020.

### 4.1.2. Single-pot, solid-phase-enhanced sample preparation (SP3)

The next step after the extraction of proteins from a specimen, most commonly facilitated by SDS, comprises the proteolytic digestion to peptides<sup>95,104</sup>. In practice, this is limited by the incompatibility of SDS with protease activity and protein digestion as a result. It additionally has an ion suppression feature<sup>310,311</sup>, further highlighting the necessity of a compatible workflow to remove SDS. The SP3 method is a fast and straightforward clean-up procedure for unbiased retrieval and purification of proteins and peptides to remove all



kinds of contaminants, including SDS (Hughes *et al.*, 2014; Hughes *et al.*, 2019)<sup>149,291</sup>. Its broad range of features renders it an attractive solution to tackle common sample preparation bottlenecks and ease the emergence of an automated, routine pipeline for clinical proteomics. For the subsequent automation of the SP3 protocol, we *firstly* went through a series of evaluation and optimization steps to achieve maximal performance.

The method utilizes paramagnetic beads in the presence of an organic solvent (>50% ACN or EtOH) to promote protein binding to the beads, allowing extensive washing to eliminate contaminants. Beyond SDS, this can include other detergents such as Triton X-100 and NP-40, which are commonly used in proteomics experiments or chaotropes and salts. Subsequently, proteins can be digested on the beads without hindrance, and the resulting peptides are thereby released into the aqueous digestion buffer, which is directly compatible with LC-MS analysis (**Figure 3A**). Another distinctive feature of SP3 is its efficiency in protein capture and release, facilitating low- and high-input applications while consistently maintaining in-depth proteome coverage. The combined characteristics of tolerance to detergents, speed and ease of operation, and scalability qualify SP3 as a universal methodology that enables a wide variety of applications. In practice, this includes cases that involve challenging sample types, as diverse as FFPE tissue<sup>168,312</sup> and historical bones<sup>313</sup>. SP3 performs particularly well for low-input applications<sup>307</sup>, for example, allowing the analysis of single human oocytes<sup>314</sup>, and micro-dissected tissue<sup>315,316</sup>.

#### 4.1.2.1. Optimization of protein binding

The principle of SP3 is explained by a mechanism similar to hydrophilic interaction chromatography (HILIC) and aggregation. Increasing the organic proportion of the protein-containing mobile phase induces the formation of a water-rich (aqueous) layer around the hydrophilic surface of the stationary phase, namely carboxylate-modified paramagnetic beads. This phase separation causes the concentration of polar (hydrophilic) side chains of amino acids to the aqueous surrounding of the beads. While in the original protocol, the capture of proteins was performed under acidic conditions, resulting in the protonation of R-COO<sup>-</sup> to neutral R-COOH, we observed that the binding capacity is increased and more reproducibility in a neutral pH environment. In the latter scenario, polar interactions occur between positively charged amine groups of proteins and, for example, arginine and lysine side chains, and the negatively charged carboxylate ions (R-COO<sup>-</sup>) on the bead surface

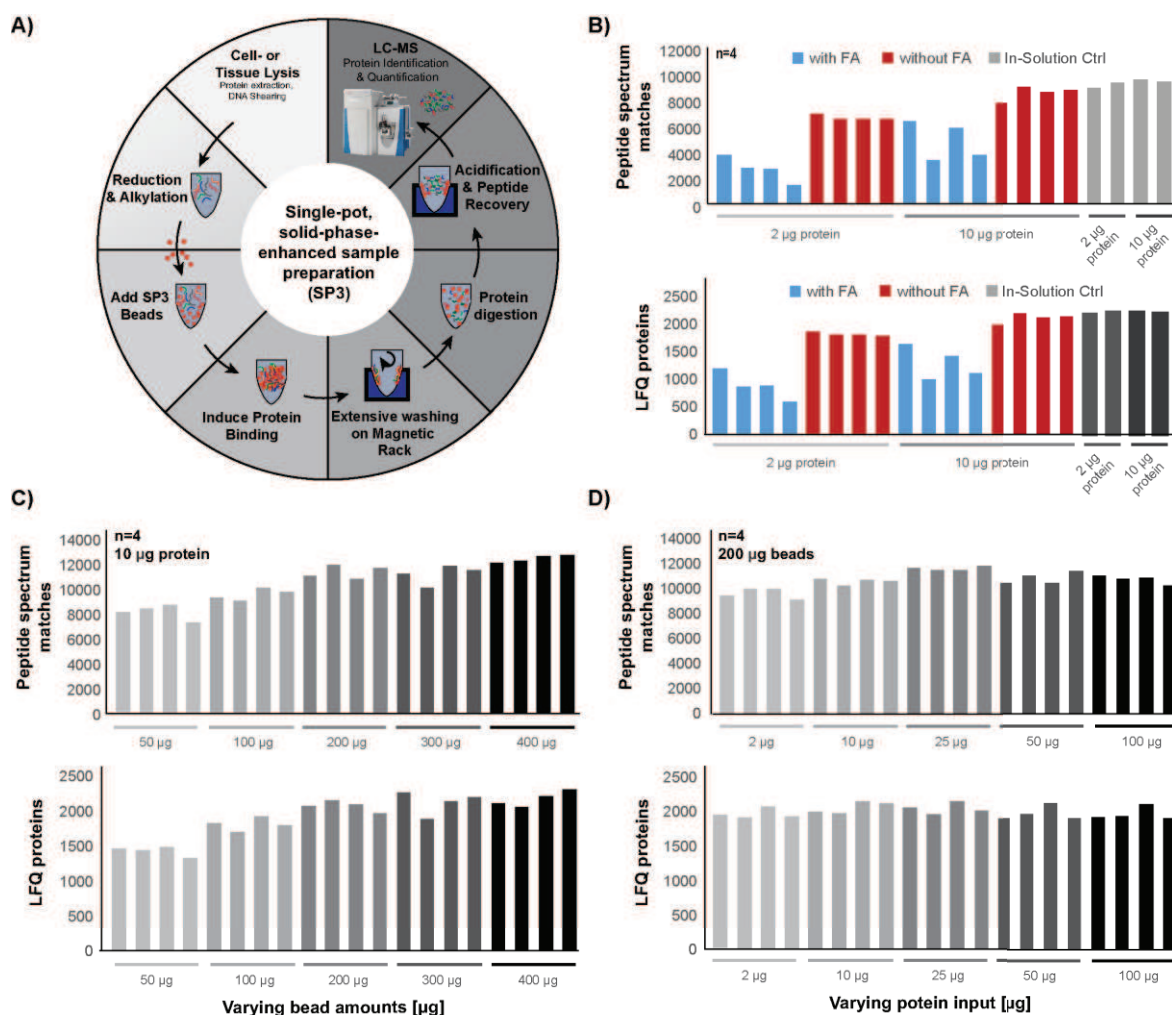
## Results

(**Figure 3B** and **Supplementary Figure 2A**). Together with prolonged incubation times of up to 18 minutes, this adaption of pH increased the efficiency of SP3 as compared to in-solution digestion of 2  $\mu\text{g}$  and 10  $\mu\text{g}$  HeLa protein, resulting in reproducible numbers of peptide spectrum matches and quantified proteins (**Figure 3B**).

On the other hand, DNA and RNA are also attracted to the aqueous solvation layer due to their overall hydrophilic character. We reasoned that the mere proximity of nucleic acids to the bead surface has the potential to influence the protein binding capacity, irrespective of the negatively charged DNA backbone. To show this, we utilized isolated DNA from mouse kidney tissue (**Supplementary Figure 2B**) in different amounts (0.5  $\mu\text{g}$ , 1  $\mu\text{g}$ , 2  $\mu\text{g}$ , and 5  $\mu\text{g}$  DNA) and fragment sizes (fully digested, partially digested, and undigested DNA), generated by both probe sonication and Bioruptor Pico treatment (**Supplementary Figure 2C**). In fact, we demonstrated an increased loss of unbound proteins in the SP3 supernatant with increasing DNA background (**Supplementary Figure 2D**) in a size-dependent manner, where small fragment sizes significantly minimize the interference. Upon change to the aqueous digestion buffer, nucleic acids are released from the beads due to their negative charge repulsion, while the polar protein-bead interaction is only reversed by proteolytic digestion during our protocol. The highest binding efficiency of proteins and its reproducibility can be achieved in the absence of large nucleic acid fragments.

Concerning the planned automation of the SP3 workflow, we further tested two additional types of beads (ReSyn Biosciences, RSA), namely MagReSyn HILIC and MagReSyn Amine. These beads have several potential advantages: I) they are comprised of a hyper-porous polymer matrix, providing an exceptionally high surface area and binding capacity, II) a higher magnetite content to support fast and efficient immobilization on the magnetic rack, and III) the increased sensitivity and correspondingly reduced material consumption to cope with the financial burden in a clinical environment. While the surface chemistry of the HILIC beads is proprietary, the Amine beads are characterized by an amine group ( $\text{NH}_2$ ). They can capture biomolecules, such as proteins or peptides, through polar interaction similar to the carboxylate-modified beads used in the classical SP3 method. Despite the potential advantages, a comparison of all three bead types (Carboxylate-, Amine-, and HILIC beads) and amounts (50  $\mu\text{g}$ , 100  $\mu\text{g}$ , 200  $\mu\text{g}$ , and 250  $\mu\text{g}$ ) for the capture and release of 10  $\mu\text{g}$  protein yielded no significant differences on the level of quantified proteins

(Supplementary Figure 2E). However, at the peptide level, we found at least 10% more identifications using the classical SP3 beads (Supplementary Figure 2F). This could point to incomplete digestion of proteins on the hyper-porous polymer matrix of the MagReSyn beads or a biased recovery and release of peptides into the digestion buffer. We continued with the carboxylate-beads that are used in the classical SP3 method.



**Figure 3: Evaluation and optimization of single-pot, solid-phase-enhanced sample preparation (SP3).** A) Schematic illustration of the SP3 protocol, including sample lysis, reduction & alkylation, protein clean-up, proteolytic digestion, and acidification & peptide recovery. B) Comparison of acidic and neutral pH conditions for the SP3 protein binding step. C) Assessing the protein binding capacity of varying amounts of paramagnetic SP3 beads (50 to 400 µg) by monitoring peptide spectrum matches (PSMs) and the number of quantified proteins. D) Assessing the protein binding scalability for varying amounts of protein inputs (2 to 100 µg) by monitoring peptide spectrum matches (PSMs) and the number of quantified proteins. Panel A modified from Mueller et al., *Mol. Syst. Biol.*, 2020.

#### 4.1.2.2. Capacity and reproducibility of SP3

Upon the optimization of effective protein binding and recovery conditions, we further assessed the capacity of beads (50 µg, 100 µg, 200 µg, 300 µg, and 400 µg) at a fixed protein concentration of 10 µg (Figure 3C). In all conditions, we observed almost no loss of unbound

## Results

proteins (**Supplementary Figure 2G**). The lowest variation in numbers of quantified proteins was observed at a bead to protein ratio of 20:1 (200  $\mu\text{g}$  beads). Further increasing the ratio did not yield more protein quantifications and only added an average of  $\sim 4\%$  peptide spectrum matches. In contrast, lowering to a 5:1 ratio of beads to proteins caused a loss of roughly 25% of quantified proteins. Unless otherwise indicated, we have chosen to use 200  $\mu\text{g}$  beads for most of the following experiments.

In parallel, we evaluated the scalability of the method using varying amounts of protein (2  $\mu\text{g}$ , 10  $\mu\text{g}$ , 25  $\mu\text{g}$ , 50  $\mu\text{g}$ , and 100  $\mu\text{g}$ ) at a fixed concentration of beads (**Figure 3D**). Again, in all conditions, we observed almost no loss of unbound proteins (**Supplementary Figure 2H**). Upon SP3 processing, an equivalent of 500 ng peptides were measured per sample, resulting in consistent numbers of quantified proteins ( $\mu = 1998$  and  $\sigma = 90.2$ ). This indicates a high reproducibility (CV= 4.5%), a nearly complete sample recovery across all LC-MS runs (average peptide intensities CV= 4.84%), and high sensitivity of the method for low protein quantities (2  $\mu\text{g}$ ) with an average peptide intensity of 93.7% compared to the highest observed average (25  $\mu\text{g}$ ) (see also **Figure 2D** for sub- $\mu\text{g}$  protein input). For both, low and high protein input, the number of peptide spectrum matches (PSMs) was slightly decreased (CV= 6.67%), illustrating the importance of optimized ratios between beads, proteins, and the working volume (**Figure 3D**).

We further compared our optimized SP3 method to a standard in-solution digest (without detergents) and the commonly used filter-aided sample preparation (FASP) method<sup>155,158,307</sup>. The aim was not to repeat all proof-of-concept experiments for SP3 but to understand whether our optimization could cause undesired effects. This could manifest itself as a bias towards specific peptides or proteins. The overlap of identified HeLa peptides between all three methods was high ( $\sim 49\%$  shared) (**Supplementary Figure 3A**). Between 5% (FASP) and 8% (In-solution) of peptides were uniquely identified using one method. This is well within the expected range of overlap due to the stochastic nature of LC-MS, in which also technical replicates show similar values. We did not find any significant difference in the distribution of molecular weights (**Supplementary Figure 3B**), compared to the whole proteome (Uniprot), and the average hydrophobicity (GRAVY score)<sup>317</sup> (**Supplementary Figure 3C**). A GO annotation for cellular compartments illustrated a highly similar

distribution of identified proteins between the HeLa in-solution digest and SP3 (**Supplementary Figure 3D**).

All in all, SP3 has room for minor optimization in different conditions in order to achieve optimal performance according to the starting amount of protein and the working volume. However, it is a highly sensitive method that allows the reproducible processing of a variety of sample types and minute amounts without a selective enrichment of specific proteins or peptides. It qualifies as a universal building block for an end-to-end proteomics workflow, resulting in peptides samples compatible with downstream applications, such as tandem mass tag (TMT) labeling and high-pH fractionation, or direct LC-MS.

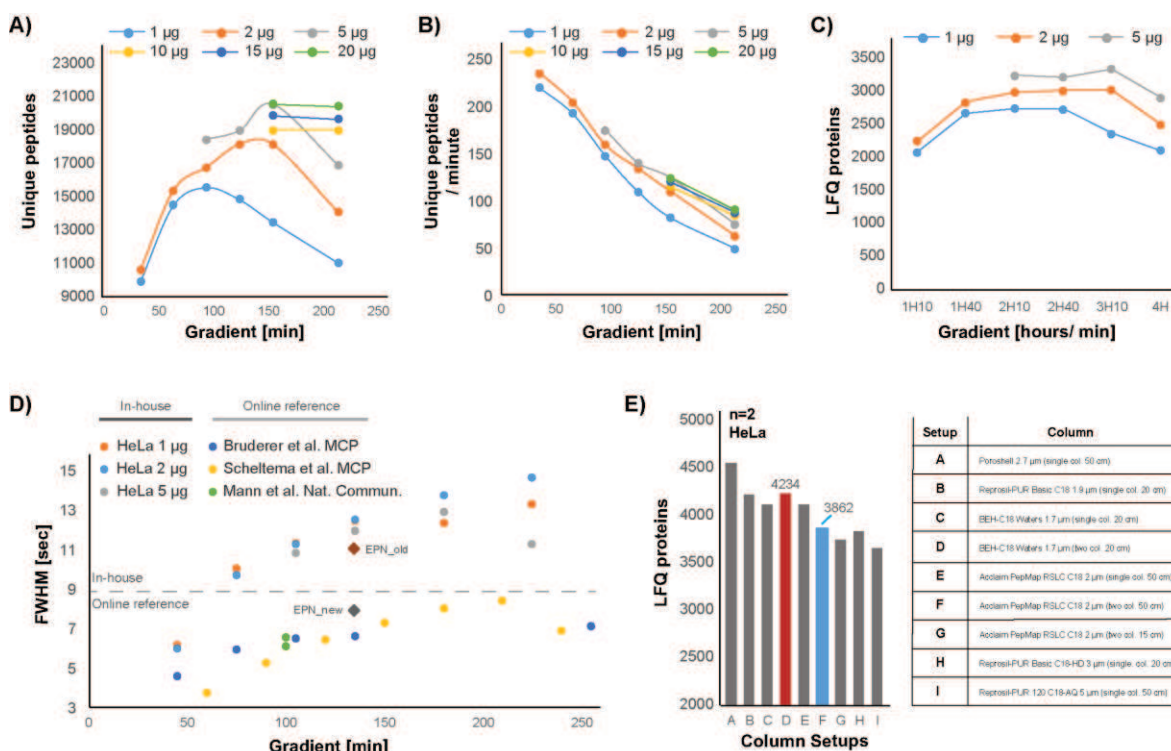
#### 4.1.3. LC-MS data acquisition

During each LC-MS measurement for global proteome profiling, the objective is to analyze the entire set of peptides that are present within a sample. In practice, this is limited by the complexity of samples and the sensitivity and scan rate of the mass spectrometer<sup>94</sup>. One way to improve the depth and coverage of peptide identifications is the increase of measurement time to disperse the analyte over time and provide the instrument with more scanning time. This can be achieved by either longer gradients for every LC-MS run<sup>318</sup> (**Figure 4A** and **Figure 4B**) or, for example, by additional offline high-pH fractionation using a reversed-phase C18 column to further separate peptides into multiple fractions<sup>319</sup>. Each fraction or concatenated fractions of the same sample are measured in consecutive LC-MS runs and compiled using a computer. The analysis time is significantly increased per sample leading to higher peptide coverage. On the other hand, both approaches require a higher sample input (**Figure 4A** and **Figure 4B**), which in practice is often a limiting factor in a clinical environment. Especially here, the balance between data depth and measurement time is a pivotal aspect to generate useful data in acceptable turn-around times.

In proteomic profiling, we are not only interested in consistent and maximized numbers of identifications of peptides or proteins across many samples, but further aim for accurate quantification<sup>94</sup>. Here, a compromise is necessary with the number of samples, their analysis depth, and accurate quantification on the one hand, and the data acquisition time and overall turn-around time on the other hand. To achieve this, we evaluated a hybrid approach to uncouple identification and quantification by 1) *firstly* generating a library of protein and peptide identifications through extensive high-pH fractionation of a

## Results

representative pool of all available samples, and II) *secondly*, focusing on quantification in short LC-MS runs of individual samples. Here, we modified our standard MS methods to collect more data points for every peptide feature and achieve better quantification as a result. The associated loss of peptide identifications in individual short runs was subsequently recovered by the integration of the sample-specific peptide library. This is achieved by matching of identified peptides to unidentified features based on retention time and mass-to-charge ratios<sup>287,288</sup>. This hybrid approach allows fast data acquisition per sample while conserving a good proteome coverage at optimal quantification.



**Figure 4: Optimization of liquid chromatography (LC)-setup for increased peak capacity.** A) Comparison of peptide injection amount (1 µg, 2 µg, 5 µg, 10 µg, 15 µg, and 20 µg) and LC-MS gradient lengths (45, 60, 90, 120, 160, and 220 minutes) for the highest number of unique identified peptides. B) Illustration of panel A to show the identified unique peptides per minute for the different peptide injection amounts and LC-MS gradient lengths. C) Assessment of LC-MS time consumption and proteome depth using three peptide injection amounts (1 µg, 2 µg, 5 µg) and different gradient lengths (45, 60, 90, 120, 160, and 220 minutes). D) Assessment of full width half maximum (FWHM) average peak width for the different peptide injection amounts and LC-MS gradient lengths in panel C. Additional comparison to published datasets from similar LC-MS setups. E) Evaluation of different commercial and self-packed analytical columns for improving the peak width.

### 4.1.3.1. Library generation and LC optimization

Initially, we evaluated and optimized the performance of our Infinity 1260 HPLC system (Agilent) for high-pH fractionation. Sample losses during fractionation were negligible when comparing the number of quantified proteins from concatenated samples to omitted

fractionation with an average ( $n=4$ ) of 1910 and 2085, respectively. (**Supplementary Figure 4A**). This is particularly beneficial in light of quantity-limited material. The high-pH ammonium hydroxide buffer was exchanged for high-pH ammonium formate because it eliminates the need for another clean-up step of the concatenated fractions. Additionally, it was previously demonstrated that low concentrations of ammonium formate could enhance the ionization efficiency during ESI-MS<sup>320</sup>, depicted by increased numbers of peptide spectrum matches (average: 5151 versus 4364) and quantified proteins (average: 2393 versus 2220) per fraction (**Supplementary Figure 4B**). The chromatographic performance remained unaffected by the buffer exchange, as highlighted by the number of peptide sequences solely identified in one or two fractions (**Supplementary Figure 4C**). In an ideal scenario, a sample- or cohort-specific peptide library is generated once, whereas individual patient samples can be matched continuously. We determined the feasible scope of a library by generating multiple examples comparing different numbers of concatenated fractions (8, 16, 24, and 32) and different gradient length (1-hour and 2-hours) (**Supplementary Figure 4D**). The library sizes correlate positively with the instrument time for LC-MS. While it is indisputable that the largest library depth is desired, the optimum balance depends on the overall number of samples and corresponding relative time investment of library generation compared to the acquisition of each sample. In the remaining part of this thesis, the number of fractions and utilized gradient length is indicated for individual experiments.

Following the generation of a deep-proteome library, we first evaluated different gradient lengths and peptide loadings (1  $\mu\text{g}$ , 2  $\mu\text{g}$ , 5  $\mu\text{g}$ , 10  $\mu\text{g}$ , 15  $\mu\text{g}$ , and 20  $\mu\text{g}$ ) to determine the best balance of time consumption and proteome depth in individual runs (**Figure 4A**, **Figure 4B**, and **Figure 4C**). The data illustrate that a high sample input ( $>10 \mu\text{g}$ ) is necessary to benefit from longer gradients (**Figure 4A**). The absolute numbers of unique peptides and their relative identification per minute have guided us to step away from the originally planned 4-hours gradients and further focus on either 1-hour or 2-hours per sample (**Figure 4A** and **Figure 4B**). Consequently, more patient samples can be measured in a shorter time to generate higher statistical power with less sample consumption. As another by-product of the evaluation, we noticed that our average peak width at FWHM was significantly higher ( $\sim 4$  to 6 seconds) compared to published data (**Figure 4D**), irrespective of the gradient

## Results

length. We extensively tested and evaluated different packing materials for self-packaged columns, as well as commercially available columns from different vendors (**Figure 4E**). The transition to a 25 cm BEH-C18 1.7  $\mu\text{m}$  analytical column (Waters Corporation) improved the peak width and peak capacity per minute as a result. In a 1.5-hours method, we could increase the number of quantified proteins and identified peptides by 9.6% and 13.9%, respectively. The LC column setups that were used are described in detail within the corresponding method sections.

At a later stage of this study, we additionally evaluated a recently released technology, namely  $\mu\text{PAC}$  (PharmaFluidics)<sup>321</sup>. It is a novel type of column comprised of highly structured micro-pillars covered with C18 and produced by lithographic etching. The nearly perfect order of the stationary separation bed leads to a uniform flow distribution and low analyte dispersion, resulting in high sensitivity. The  $\mu\text{PAC}$  runs at very low back pressure (<50 bar compared to >650 bar for packed columns) and potentially offers a long lifetime for itself and the LC, which would be superior for clinical applications. Going through a series of trials to evaluate and optimize the LC-MS setup, we were not able to achieve results comparable to our initial setup (data not shown), which is even outperformed after the latest testing of different columns. However, the technology is still in its infancy, and by its continuous development might become a great tool when robustness and lifetime is a key demand.

### 4.1.3.2. Match-between-runs and optimal quantification

Next, we measured replicates of a HeLa digest ( $n=3$ ) with a 2-hours gradient utilizing a standard 'Top-20' or 'Top-2' method. Hereinafter they are referred to as T20 and T2. The latter was designed to generate  $\text{MS}^2$  scans for the two most abundant precursor ions per  $\text{MS}^1$  scan, resulting in a substantially reduced cycle time and more data points ( $\text{MS}^1$  scans) per peptide feature to increase the quantification accuracy. As a consequence, less time remains available for the acquisition of  $\text{MS}^2$  spectra that are used for peptide identification. The T2 method recorded an average of 1.6  $\text{MS}^2$  scans per  $\text{MS}^1$  scan (7.4 in T20). This sacrifice of  $\text{MS}^2$  scans for peptide identification was conserved by using the sample-specific proteome library paired with the MaxQuant matching-between-runs algorithm. In both methods, similar numbers of proteins (**Supplementary Figure 4E**) and peptides (**Supplementary Figure 4F**) were identified with the library approach. Nearly half of the



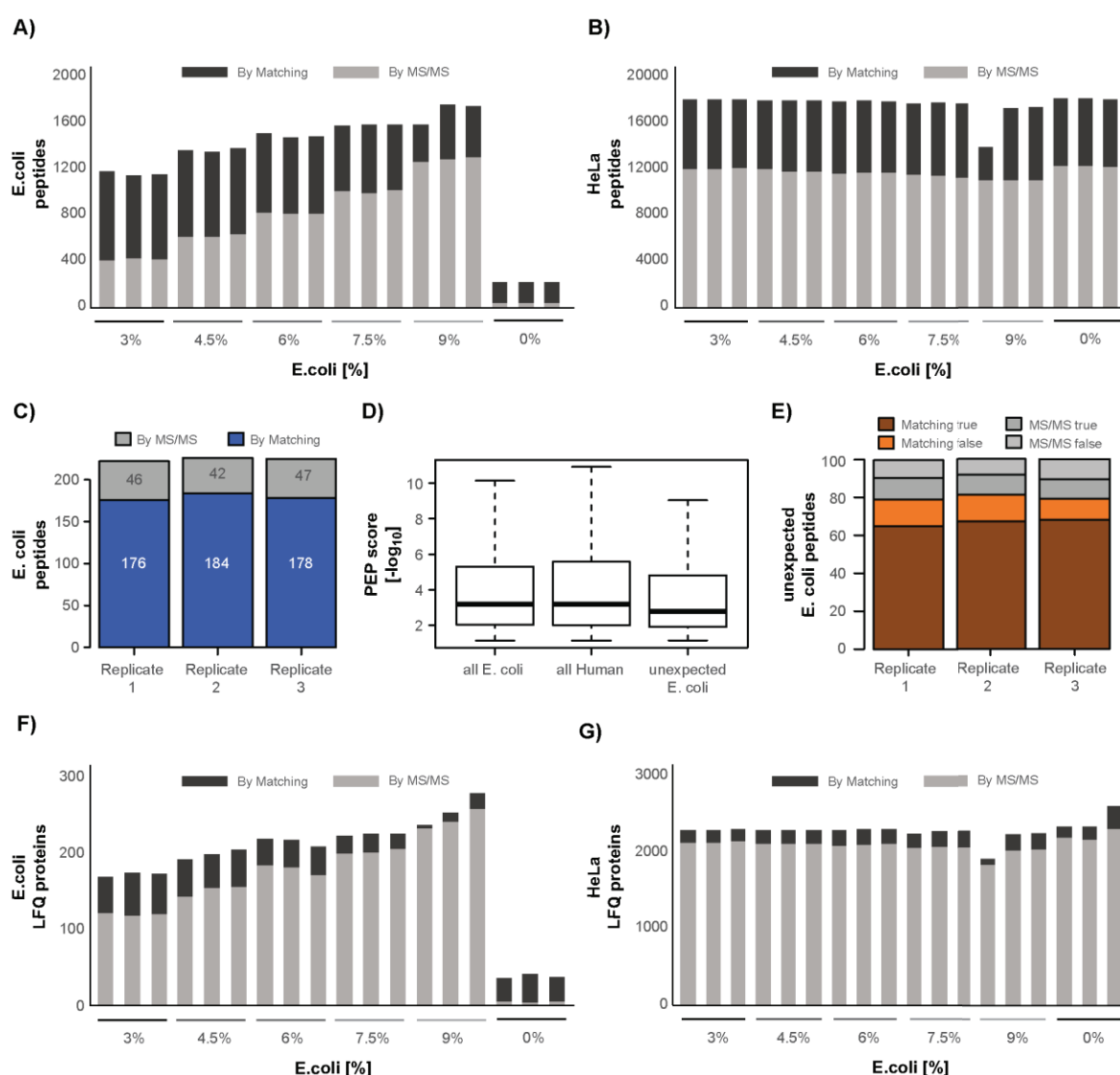
peptide identifications (49.5% in T2) originate from matching to the library (19.07% in T20). The T2-library approach identified an average of 5555 (22%) additional peptides compared to the T20 approach without a library. A comparison of average CV [%] values for identified peptides (T20= 24.6% and T2= 23.7%) and quantified proteins (T20= 13.3% and T2= 12.7%), however, did not reveal a significant difference in quantification accuracy between both methods (**Supplementary Figure 4G**). Outliers were removed for visualization, resulting in marginally reduced average CV [%] values for quantified proteins (T20= 10.7% and T2= 9.6%). Low abundance peptides and proteins are the main effectors that contribute to the minor, negligible difference (**Supplementary Figure 4G**). The increased number of data points per peptide feature in T2 over T20 cannot improve the description of the area under the curve (AUC).

During the early phases of this project, we had access to fresh-frozen tumor samples to establish our sample preparation workflows. This set of 17x pilot samples was subsequently used to compare the T20 and T2 method in a realistic scenario. The data were acquired using a 1-hour and a 2-hours gradient with and without library support. For the 1-hour method, we utilized a peptide library composed of either 16 or extended 32 fractions. On average, we identified and quantified 22% (T20) and 42.8% (T2) more proteins using the 2-hours method. The proportion of peptides identified by matching decreased from roughly 50% (1-hour T20 & T2) to 30% (2-hours T20) and 36% (2-hours T2). Stepping from 1-hour T2 to 2-hours T2 (T20), we observed a decrease in the relative number of missing values by 4.76% (peptide-level) (0.6%) and 8.04% (protein-level) (4.5%). The library-matching approach reduced the percentage of missing values by 10 to 15% on peptide- and protein level. The effect was more pronounced using the library based on 32 fractions. More instrument time and lower sample complexity improve the library depth and only require a one-time higher expenditure of time. In line with our previous observation, the accuracy of quantification, assessed by calculating the CV [%] on peptide- and protein-level, did not reveal any advantage using the T2 method. We identified 4268 additional peptides using 2-hours for data acquisition and 5956 peptides by employing the library-matching. Therefore, the majority of clinical samples within this work have been acquired using the classical T20 method in combination with a sample-specific library, unless otherwise indicated.

## Results

Despite the implementation of the matching-between-runs algorithm in the MaxQuant software and its established usage in the proteomic community, the degree of false transfers and its associated credibility has only recently been investigated by Lim MY. et al., 2019<sup>322</sup>. Similar to their two-proteome model, we had preliminary results from acquiring a constant amount of HeLa peptides with increasing spike-ins of 0%, 3%, 4.5%, 6%, 7.5%, and 9% *E. coli* peptides (n= 3). Here, the transfer of identifications between samples resulted in an average of additional 33.5% human peptides compared to without the matching algorithm. For the *E. coli* spike-ins, the identification transfer rate increased with decreasing *E. coli* peptide concentration from 24% (3% *E. coli*) to 63.9% (9% *E. coli*) (**Figure 5A**). This leveraged the total number of *E. coli* peptides to an average of 1160 (n= 3), which could be identified from as little as 3% spike-in compared to the constant HeLa background (**Figure 5B**). In the pure HeLa measurement (0% *E. coli*), we falsely identified an average of 179 *E. coli* peptides via the matching-between-runs (**Figure 5C**). This corresponds to an average of 0.97% false peptide identification transfers. The majority of these peptide features were identified with a low intensity or a low posterior error probability (PEP) score (**Figure 5D**). Another 42 *E. coli* peptides were identified by MS<sup>2</sup> spectra (**Figure 5C**). Roughly 25% of the unexpected *E. coli* peptides in the pure HeLa measurement, either identified via MS<sup>2</sup> or per matching, could also match to a human protein sequence (**Figure 5E**). This indicates that peptides were likely assigned to the wrong database. In the recent study by Lim MY. et al. (2019)<sup>322</sup>, the authors reported similar percentages of false peptide transfers. However, also in accordance with our observations, the vast majority of these matches did not pass thresholds set within the LFQ calculation in the MaxQuant software. As a result, the number of falsely annotated and quantified *E. coli* proteins (0% *E. coli*: **Figure 5F**) only represent 1.6% of all quantified proteins (0% *E. coli*: **Figure 5F** and **Figure 5G**). We further reduced the allowed time window for matching from default 0.7 minutes to 0.3 minutes (data not shown), which reduced the false transfers by one-third. This gave us sufficient confidence to employ the matching-between-runs functionality and benefit from higher peptide numbers that contribute to protein quantification and less missing values across samples. Recapitulating, we have evaluated and optimized several parameters and data acquisition strategies to achieve the highest proteome coverage with optimal quantification in the least amount of time. We could improve the performance of high-pH fractionation and

evaluated the relation between peptide library depth and time expenditure for its generation. Emerging from a series of experiments and evaluations, we finally settled for the classical data-dependent acquisition strategy with 2-hours per sample and a Top20 spacing. The sample-specific peptide library (32 fractions) approach paired with matching-between-runs (0.3 minutes match time window) was employed, despite that, we could not improve the quantification accuracy by focusing on MS<sup>1</sup> scans. The increased number of peptide identifications, low numbers of missing values across multiple samples, and the low false transfer rate were persuasive. This setup was used for all clinical sample cohorts unless otherwise indicated.



**Figure 5: Two-proteome model for the evaluation of matching-between-runs in MaxQuant.** A-B) The numbers of identified *E. coli* (A) or HeLa (B) peptides by matching-between-runs or per MS<sup>2</sup> in our two-proteome (HeLa, *E. coli*) spike-in series, comprising different amounts of *E. coli* (3%, 4.5%, 6%, 7.5%, 9%, and 0%) with a constant HeLa background. C) The numbers of identified *E. coli* peptides by matching-between-runs or per MS<sup>2</sup> in the pure HeLa samples (0% *E. coli*). D) Global comparison of the posterior error probability

## Results

(PEP) score for all *E. coli*, all human, and all unexpected *E. coli* (IDs in the pure HeLa sample) peptide identifications. E) Distinguishing between true and false identifications via matching-between-runs or per MS<sup>2</sup> based on peptide features that are likely false annotated as they also match to a human protein sequence. F-G) The numbers of identified and quantified *E. coli* (F) or HeLa (G) proteins by matching-between-runs or per MS<sup>2</sup> in our two-proteome (HeLa, *E. coli*) spike-in series, comprising different amounts of *E. coli* (3%, 4.5%, 6%, 7.5%, 9%, and 0%) with a constant HeLa background.

### 4.2. Automated SP3 (autoSP3)

In the first phase of this project, we established and optimized a manual pipeline for proteomic sample preparation and LC-MS data acquisition. Following the 96-well format lysis of all types of specimens for protein extraction and DNA shearing, the core of our workflow comprises the SP3 method to facilitate the handling of detergents and to utilize its unique sensitivity for low-input applications. In the second chapter of this thesis, we focused on exploiting the amenability of protein clean-up and digestion using SP3 to establish a fully automated pipeline that seamlessly integrates with preceding sample lysis and protein extraction using the Covaris ultrasonicator. This potential of SP3 originates from the paramagnetic nature of the employed beads, rendering the possibility to perform the entire procedure on a robotic liquid handling platform. The resulting advantages of hands-free processing can solve several remaining bottlenecks that are important for clinical integration of proteome profiling: I) robustness and reproducibility; II) throughput and turn-around times; III) low costs and simplicity, and IV) a one-for-all method for universal sample preparation.

In other systems biology disciplines, such as genomics, automated sample preparation was introduced almost a decade ago<sup>323</sup> and is now widely used through commercial kits from different vendors. In the field of proteomics, it remains far less common and limited to specific purposes, for example, sub-proteome enrichment (e.g., AssayMap to purify phosphorylated peptides<sup>324</sup>), protein digestion and peptide clean-up<sup>161</sup>, or detergent-free applications, such as plasma proteomics (iST, on an automated system to process plasma and cell lysates<sup>325</sup>).

Parts of the following chapter, including Figures and Tables, were taken in part or their entirety from the joint publication listed below.

**Mueller, Torsten**, Kalxdorf, M., Longuespeé, R., Kazdal, D., Stenzinger, A., Krijgsveld, J. (2020). “Automated sample preparation with SP3 for low-input clinical proteomics”. *Molecular Systems Biology* 16(1): e9111.

#### 4.2.1. Establishment of autoSP3: generic sample preparation

Many different liquid handling systems are available on the market with their pros and cons. Here, we selected a Bravo liquid handling system (Agilent Technologies) for establishing our method because it offers a small bench-top footprint, and it is widely available to many laboratories. The automation of SP3 (autoSP3) required the establishment, optimization, and subsequent validation of a vast number of tasks and associated parameters. This included the positioning of required consumables, the reservoir capacity for reagent and waste volumes, as well as the Bravo accessories, such as a 96-well magnet, an orbital shaker, and a heating block. We had to ensure the accessibility of each consumable, reagent, or respective deck position to allow uninterrupted running of the entire procedure. The full capacity and functionality of the Bravo were utilized to automate the processing of 96 samples simultaneously. As a result, autoSP3 smoothly connects with the preceding extraction of proteins, facilitated by the Covaris (96-well format) or other methods that provide sufficient DNA shearing (**Figure 6A**). The resulting peptides are directly compatible with LC-MS or other downstream applications.

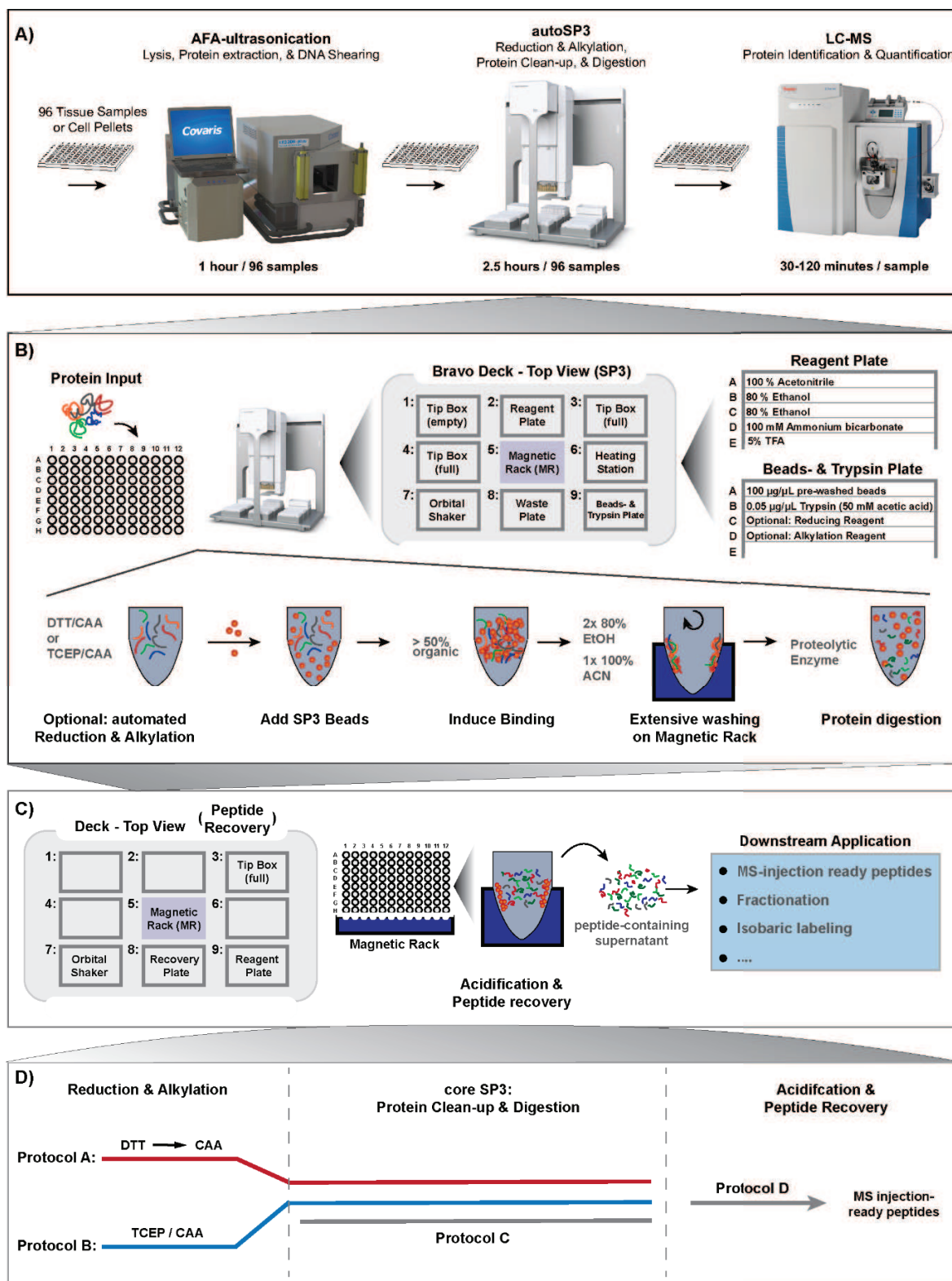
For the development of autoSP3, a HeLa cell lysate was used as an input to evaluate and execute each task from reduction and alkylation of proteins, their clean-up, and proteolytic digestion, to the final recovery of peptides (**Figure 6B** and **Figure 6C**). Initially, we implemented the reduction and alkylation of proteins using DTT and CAA ("Protocol A", **Figure 6D**). However, to minimize the number of protocol tasks and simultaneously decrease the number of reagents, we adapted the procedure for a combined reaction with TCEP and CAA for 5 minutes at 95°C ("Protocol B", **Figure 6D**). The core SP3 protocol ("Protocol C", **Figure 6D**) was programmed with the ability to be executed independently to allow the processing of samples that were reduced and alkylated otherwise. This was also of interest because the Bravo heating accessory is rather inefficient in heating and cooling, taking more than one hour to reach 95°C. Altogether, we provide three protocol options to either integrate reduction and alkylation with SP3 processing in a continuous procedure or to perform this off-deck in any preferred way to enhance speed and flexibility before transferring samples to the Bravo deck. While the latter is favorable in an academic research environment covering all eventualities, the complete workflow is attractive for a clinical setting.

## Results

During the autoSP3 protocol, both the paramagnetic bead stock as well as the enzyme solution (for example, trypsin) and optionally reducing and alkylating reagents, are deposited in a second 96-well plate to ease pipetting of small volumes and to avoid uneconomical dead volumes of expensive reagents (**Figure 6B**). The autolysis of trypsin is prevented during the execution of the protocol by its storage in 50 mM acetic acid, and dilution with an adequate volume of 100 mM ABC at the time of mixing with the protein samples to achieve a digestion-compatible pH range. The addition of reagents or solvents to the samples is performed by successively dispensing row-by-row across the entire 96-well plate. All liquid dispensing heights were adjusted such that the pipette tips never contact the sample surface. In more detail, this alludes to variable dispensing heights along with the entire protocol, corresponding to the sample volume in every step. This eliminates the risk of cross-contamination. The removal of any liquid is carried out with well-specific pipette tips throughout the protocol. In pipetting tasks, additional air plugs are used to prevent spilling of hanging droplets from the tips. The aspirating and dispensing velocities for each step are defined specifically for different liquid classes. Combining row-by-row adding and well-specific removal of solvents and reagents, we were able to establish the SP3 protocol for 96-samples using only two pipette tip boxes, contributing to the overall low costs.

After manual or automated reduction and alkylation of proteins, the autoSP3 protocol either begins or continues with the aliquoting of the paramagnetic bead suspension to each sample (**Figure 6B**). This is achieved by spotting 5  $\mu\text{L}$  beads (50  $\mu\text{g}/\mu\text{L}$  in ddH<sub>2</sub>O) as a droplet to the wall of each well and gently moving them into the sample solution by agitation in the orbital shaking accessory. The bead concentration was optimized as compared to the manual SP3 method (100  $\mu\text{g}/\mu\text{L}$  in ddH<sub>2</sub>O) to improve the pipetting precision. In the next step, protein binding to the beads is induced by the addition of ACN to a final concentration of 50% organic (see also chapter 4.1.2.) (**Figure 6B**). While we could achieve ~8% and ~6% more identified peptides and quantified proteins using EtOH to promote polar interactions with the beads, the pipetting properties of ACN exhibit a better reproducibility (data not shown). The homogenous distribution of beads for the efficient formation of protein-bead aggregates is achieved by continuous alternating between fast and slow agitation rather

than pipette mixing. The latter resulted in a severe sample loss due to the tendency of beads sticking to the pipette tips under the >50% organic condition.



**Figure 6: A schematic overview of the automated single-pot, solid-phase-enhanced sample preparation (autoSP3) workflows.** A) Illustration of the end-to-end workflow from fresh-frozen tissue or cells to injection-ready peptides and LC-MS. B) The overview shows the different steps of the autoSP3 protocol from protein

## Results

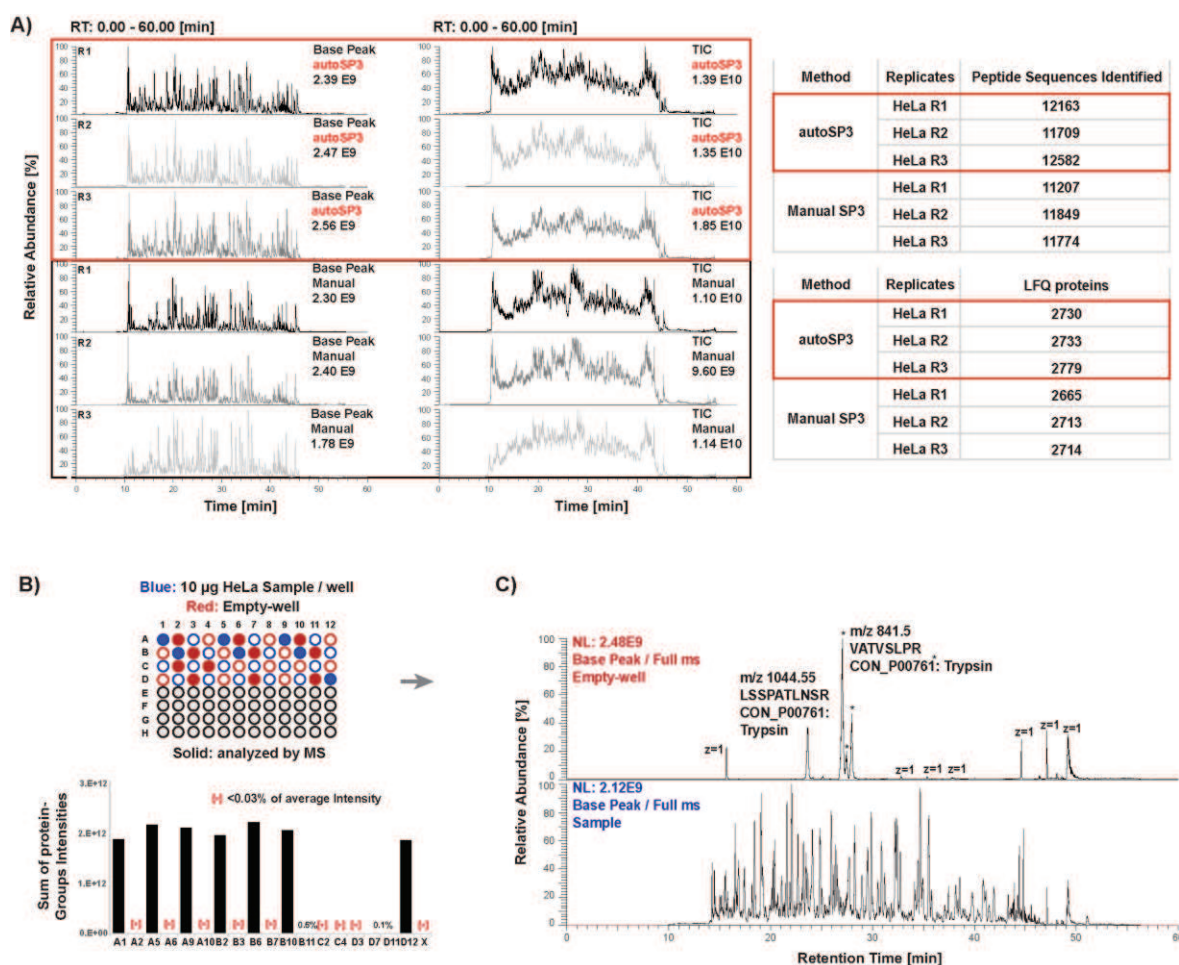
input to enzymatic digestion. The setup of the Bravo deck is shown for the core clean-up protocol. C) The overview shows the steps and Bravo deck setup of the autoSP3 peptide acidification and recovery protocol. The protocol ends with MS injection-ready peptide samples. D) A schematic overview of all available autoSP3 protocol versions. The autoSP3 procedure is provided with three options for reduction and alkylation and with post-digestion peptide recovery. Protocol A: one-step reduction and alkylation using a TCEP/CAA mixture for 5 minutes at 95°C, followed by autoSP3. Protocol B: two-step reduction and alkylation using DTT and CAA consecutively with 30 minutes incubation at 60°C and 23°C, respectively, followed by autoSP3. Protocol C: the core autoSP3 protocol is omitting reduction and alkylation such that the user can flexibly pre-treat manually prepared samples. Protocol D: post-digestion acidification and recovery, delivering MS injection-ready peptides to a new sample plate. Modified from Mueller et al., *Mol. Syst. Biol.*, 2020.

The protein-bead aggregates are rinsed with two times 80% EtOH and one time 100% ACN<sup>291</sup> (**Figure 6B**). The potency of each cleansing task is increased by orbital shaking. For the disposal of the washing solvents, sample plates are stalled on the magnetic rack to allow the beads to settle in a ring shape above the well-bottom. The removal of the wash solvent was split into two consecutive aspiration tasks to minimize the residual liquid volume effectively. Here, EtOH was observed to drain from the sidewall in each well due to its viscosity. Neglecting the second aspiration step reduced the clearance of contaminants and potentially interfered with the subsequent digestion. A low residual concentration of less than 5% ACN (~2  $\mu$ L) might enhance the protein digestion by assisting in protein unfolding and maximizing the accessibility of amino acid sequence cleavage sites as a result. However, this was not experimentally tested, and we tried to avoid exceedingly high residual ACN concentrations not to affect the protease activity. After the last washing step, proteins trapped on the paramagnetic beads are covered in digestion buffer, plates are manually sealed, and transferred to a PCR thermocycler incubation. We chose a thermocycler with lid heating to avoid evaporation during the process. Following enzymatic digestion, resultant peptides were automatically acidified and recovered to a new sample plate. The acidification and peptide recovery tasks were programmed as an independent protocol (**Figure 6C**) because supplying a new set of tips was required.

Initially, we benchmarked the performance of autoSP3 compared to the manual procedure by processing replicates of the same sample. Between both methods, we observed similar ion intensities (base peak and total ion chromatograms) and CV [%] values for numbers of identified peptides (1%) and quantified proteins (1%) (**Figure 7A**). The lack of cross-contamination was demonstrated by running a 96-well plate of HeLa lysate alternating with empty controls. In comparison to sample-containing injections, most control samples had a residual MS intensity of less than 0.03% (**Figure 7B**). The residual intensities were



primarily attributed to autolytic peptides of trypsin (added to each sample, including empty controls), and to (non-peptidic) contaminants with a +1-charge state (**Figure 7C**). This was sharply contrasting with rich chromatograms from protein-containing samples (**Figure 7C**).



**Figure 7: autoSP3 reproducibility and proof of absent cross-contamination.** A) A comparison of three versus three individually processed samples using the manual SP3 protocol or autoSP3. The upper panel (red) shows the base peak and total ion chromatogram (TIC) of three autoSP3 HeLa samples, while the lower panel (black) shows base peak and TIC of three manual SP3 HeLa samples. The number of proteins and peptides identified by either workflow is indicated per replicate. B) Schematic representation of the experimental design to demonstrate the absence of cross-contamination between wells. Half a plate (48 wells) was processed with 10 µg protein of a HeLa batch lysate in every second well (highlighted in blue) interspaced with empty wells as a control (highlighted in red). Randomly selected wells (highlighted in solid) were selected for direct LC-MS. Bar plots of the summed intensities of protein groups across selected samples. A total of seven sample-containing injections were performed, and a total of twelve empty controls. Asterisks indicate intensities < 0.03%. C) Exemplary base peak MS<sup>1</sup> spectrum for an empty control injection (top) and a sample-containing injection (bottom). Modified from Mueller et al., *Mol. Syst. Biol.*, 2020.

We established, optimized, and benchmarked the SP3 protocol on a Bravo liquid handling system, taking care of all sample handling steps<sup>304</sup>. AutoSP3 directly interfaces with 96-well format lysis, protein extraction, and DNA shearing facilitated by the Covaris. Alternatively, the protocol can start from 96 cell- or tissue lysates from any source that provides a

## Results

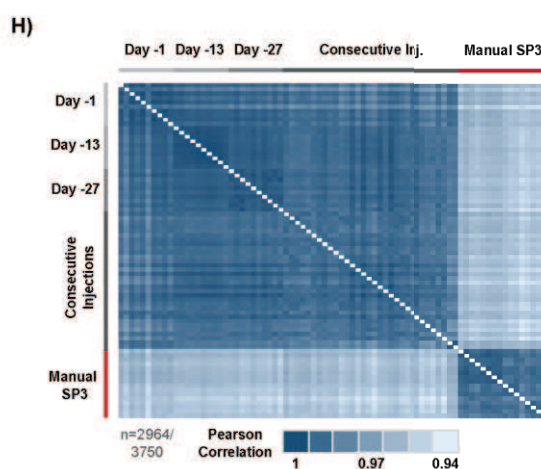
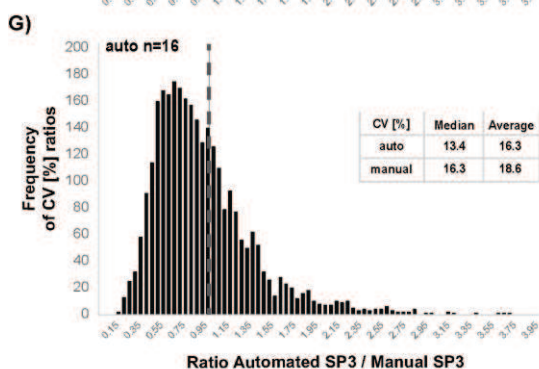
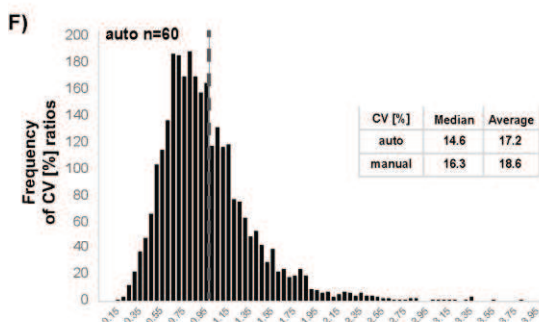
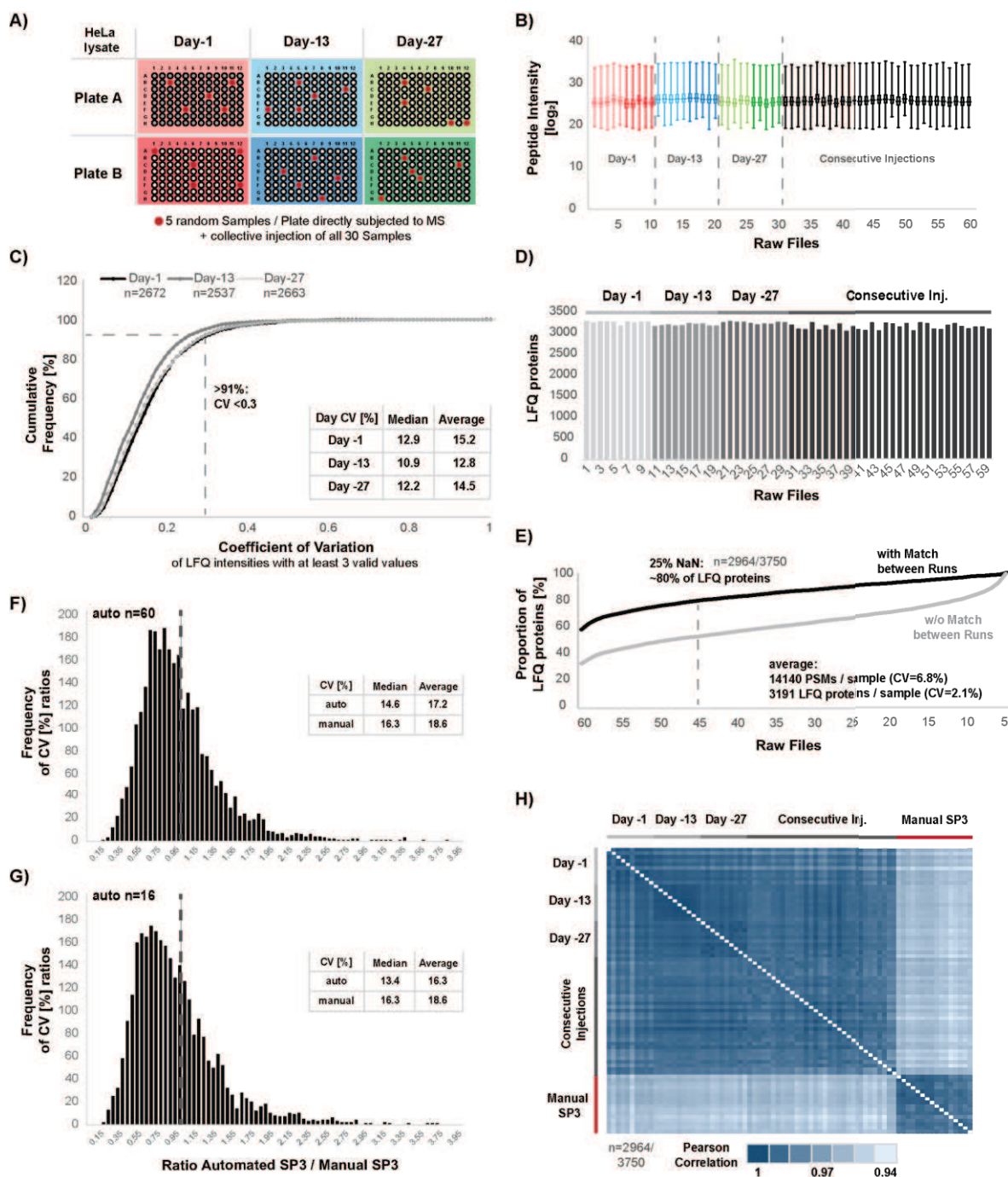
sufficient sample quality. A complete run of the Bravo SP3 protocol takes one hour and 23 minutes for 96 samples (“Protocol C”, **Figure 6D**) with an additional 45 or 65 min for optional reduction and alkylation with DTT/ CAA (“Protocol A”, **Figure 6D**) or TCEP/ CAA (“Protocol B”, **Figure 6D**), respectively. The peptide-containing supernatant can be recovered without further clean-up for direct LC-MS data acquisition or any compatible downstream protocol, such as high-pH fractionation (e.g., for library generation) or TMT labeling. The recovery of samples can be performed using a separate acidification and peptide recovery protocol, which takes about 7.5 minutes to complete (“Protocol D”, **Figure 6D**).

### 4.2.2. Evaluation of autoSP3 precision

Reproducibility and precision are the main focus of any automated procedure. The latter is defined, for example, by the European medicines agency (EMA) as the variability observed within a laboratory<sup>299</sup>. For autoSP3, we assessed both by determining the intra-day precision and the longitudinal inter-day precision throughout one month<sup>326</sup>. Therefore, we utilized reduced and alkylated proteins extracted from HeLa cells with sufficient DNA and RNA shearing. A total of six 96-well plates, corresponding to 576 individual samples, were processed in the morning and in the afternoon of three different days (3 times intra-day) (**Figure 8A**). Choosing “Day-1” and following “Day-13” and “Day-27” for the sample plate processing covered a period of roughly one month and allowed to infer the inter-day precision by correlating data obtained across all six plates. In more detail, the LC-MS analysis was performed for five randomly selected samples per 96-well plate immediately after autoSP3 sample processing on the same day. In the end, the same set of selected samples was re-measured as a coherent batch, resulting in a total of 60 LC-MS runs. This allowed us to determine the potential influence of autoSP3 processing or longitudinal MS performance fluctuations. Taken together, the acquired data allowed us to assess the variability within a single plate, within a single day (two 96-well plates), across three days, as well as including and excluding the variance imposed by the MS performance. As the first indication of reproducibility, the observed intensities of all identified peptides were consistent within and across days (Pearson correlation  $R=0.9$ ) (**Figure 8B**).

To evaluate the data more deeply and assess the intra-day precision at the level of proteins, we initially filtered the list of observations obtained from a single day (10 samples) for at

least three quantified values (3 out of 10). This filtering resulted in 2672, 2537, and 2663 quantified proteins for “Day-1”, “Day-13”, and “Day-27”, respectively (**Figure 8C**). More than 91% of quantified proteins exhibited a CV [%] within each day of less than 30%. The median CV’s [%] per day ranged from 10.9% to 12.9%. This highlights an overall consistent quantification of proteins across sequentially processed replicates originating from different plates within a day (intra-day).



**Figure 8: Longitudinal assessment of autoSP3 performance and reproducibility.** A) A schematic representation of the experimental design. 96 times 10  $\mu$ g protein of a HeLa batch lysate were processed in

## Results

the morning (Plate A) and the afternoon (Plate B) at three different days (Day-1, Day-13, and Day-27) over a month. From each plate, five randomly selected samples were subjected to direct LC-MS analysis (red dots). In addition, all 30 samples (ten per day) were measured in a single combined batch to judge the influence of MS variability B) Box-whisker plots of  $\log_2$ -transformed peptide intensities across all 60 raw files. The color-coding highlights the plate in which each sample was processed. C) Cumulative frequency curve [%] of the observed coefficient of variation (CV) of proteins that were identified and quantified with a minimum of three valid values within each day. Here, the ten raw files of each day are evaluated individually. The resulting median and average CV [%] for each day are shown. D) Bar plot summarizing the number of quantified LFQ protein groups across 60 HeLa samples. Samples originating from different days and the consecutive injections of the same samples are highlighted in grey scales. E) A line chart is showing the proportion of quantified protein groups across all 60 autoSP3 HeLa samples. The data are shown with and without the use of match-between-runs. F) Histogram showing CV's [%] of quantified proteins across all 60 automatically prepared HeLa samples, proportional to CV's [%] of sixteen manually prepared samples. The median and average CV [%] is shown for both automatically and manually prepared samples. A dotted line highlights the ratio of 1. G) Histogram showing CV's [%] of quantified proteins from sixteen randomly selected out of 60 samples, proportional to sixteen manually prepared samples. The median and average CV [%] is shown for both automatically and manually prepared samples. A dotted line highlights the ratio of 1. H) Pearson correlation heatmap of all 60 raw files and an additional sixteen manually prepared HeLa SP3 samples. The displayed data are filtered for 75% data completeness (Table 1). Please note the narrow scaling (1-0.94). Modified from Mueller et al., *Mol. Syst. Biol.*, 2020.

We further determined the inter-day precision of the autoSP3 performance by considering the entire dataset (all 60 LC-MS runs). On top, we manually prepared 16 replicates of the same sample using SP3. The average number of identified peptides and quantified proteins was 14.140 and 3191, respectively (**Figure 8D**). By applying the MaxQuant matching-between-runs feature, we could increase the proportion of consistently (non-missing values) quantified proteins from 33.62% to 58.37%. This increased the number of proteins that are considered for our assessment to  $n=3750$ , with a median and average CV [%] of 18.1% and 20.5%. Hereinafter, we additionally calculated the CV's [%] of quantified proteins with either a minimum of three valid values (3 out of 60;  $n=3688$  proteins) or  $\frac{3}{4}$  valid values (45 out of 60;  $n=2964$  proteins). The latter requirement is equivalent to the minimum data completeness of 75% and covers 78.04% of the entire list of observed and quantified proteins (**Figure 8E**). CV's [%] were calculated for within each day, across days, without the LC-MS performance variability as one batch, and overall 60 measurements (**Table 1**). The comparison of CV's [%] of samples that were analyzed on the day of sample preparation (median 13.3%, **Table 1**) to the acquisition of the same samples as a single batch (median 14.3%, **Table 1**) did not reveal a negative influence driven by the longitudinal LC-MS performance. We obtained excellent CV [%] values irrespective of the time of sample preparation or data acquisition over extended time periods (here four weeks). Across all 60 (+16) LC-MS runs we could see a marginal improvement of median and average CV's [%] between autoSP3 (14.6% and 17.2%;  $n=60$  runs) and manual SP3 (16.3% and 18.6%;  $n=16$

runs). Yet, these numbers compare favorably to what is generally observed for label-free quantitation (20-30%). The frequency distribution of CV [%] ratios between autoSP3 and manual SP3 position towards consistent lower variation using the automated workflow on a per-protein basis (**Figure 8F** and **Figure 8G**).

LFQ Intra- & Inter-day Variability	# of replicates	75% Data Completeness (n=2964)		Minimum of 3 valid values (n=3688)	
		Median CV [%]	Average CV [%]	Median CV [%]	Average CV [%]
within Day -1	10	11.9	14.6	13.3	16.5
within Day -13	10	9.8	12.2	11.3	14
within Day -27	10	10.8	13.5	12.3	15.5
across Days	30	13.3	15.7	16	18.6
w/o MS variability	30	14.3	17.3	17.2	20.1
overall automated	60	14.7	17.4	18.1	20.6
manual SP3	16	16.3	18.6	17.3	20

**Table 1: Summary of the observed coefficient of variation (CV's)** Corresponding to Figure 8, the table summarizes median and average coefficient of variation (CV) [%] values for individual days, across days, with and without the MS-imposed variability, and manual SP3. CV [%] values were calculated with either 75% data completeness requirement (~80% of all available quantified proteins) or with a minimum of three valid values across 60 samples. Modified from Mueller et al., *Mol. Syst. Biol.*, 2020.

Both autoSP3 and manual SP3 yielded robust protein quantification with a Pearson coefficient of higher than 0.97 across all 60 LC-MS runs and an additional 16 manually processed samples (**Figure 8H**). We could not observe a considerable difference between quantified proteins from “Day-1”, “Day-13”, and “Day-27”, respectively. This is a good indication for excellent inter-day precision. The correlation between autoSP3 and manual SP3 data was marginally reduced at >0.94, which is likely reflecting minor differences in both procedures, for example, processing volumes.

High abundant proteins are more reproducibly quantified across the entire autoSP3 dataset. This is highlighted by varying CV [%] value distributions based on protein abundance (**Supplementary Figure 5A**). Intensity bins were defined from highest to lowest as follows: ‘A’: 1-500; ‘B’: 501-1251; ‘C’: 1252-2001; and ‘D’:2002-2964. The CV [%] values were calculated within each bin range. More than >97.5% of proteins in groups ‘A’ and ‘B’ exhibited a CV [%] of less than 30% (median <10%), contrasting to 39.1% (CV<30%) of proteins in group ‘D’. The lowest abundant proteins in group ‘D’ are comprised of the ~1000 proteins recovered by the matching-between-runs feature in MaxQuant. Disabling

## Results

matching-between-runs reduces the number of quantified proteins from  $n=2964$  to  $n=2019$ , while increasing the percentage of proteins with a CV [%] value below 30% to 76.2% in group 'D' (**Supplementary Figure 5B**). A selection of previously described housekeeping proteins<sup>327</sup> and two randomly selected low abundant proteins showed CV [%] values below 5% and 25% for the highest abundant and low abundance proteins for both, within and across all days (**Supplementary Figure 5C**).

Taken together, we could demonstrate the robustness and high performance of SP3 for automated and manual handling. The autoSP3 procedure slightly improved the median CV [%] values, while adding a large sample processing throughput and minimizing hands-on time as a result. The longitudinal performance was highly reproducible for four weeks and irrespective of the LC-MS.

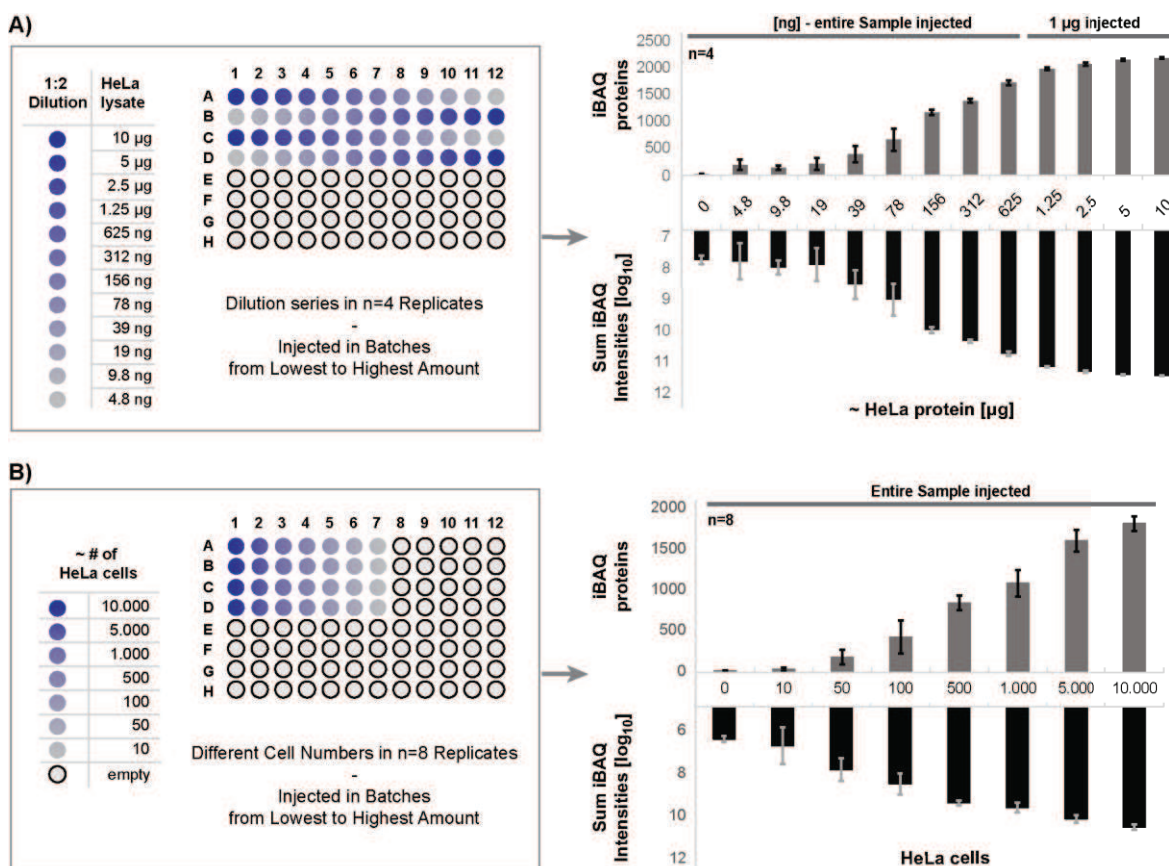
### 4.2.3. Assessment of autoSP3 sensitivity

A permanent challenge in the field of proteomics is the handling and analysis of low-input material<sup>149,307,313,314,328,329</sup>. This originates either from inefficient sample preparation and associated losses of material or from suboptimal liquid chromatography interfaces and confined mass spectrometer performances. In this chapter, we harnessed the unique sensitivity of autoSP3 to handle sub-microgram amounts of protein input. As a key asset of SP3, this was previously demonstrated for manual handling in a number of scenarios<sup>149,314</sup>.

Here, we demonstrated the ability of autoSP3 to handle minute amounts of a sample by processing a 2-fold serial dilution of our HeLa protein stock (10  $\mu\text{g}$  to  $\sim 5$  ng;  $n=4$  per concentration) (**Figure 9A**). The potential of carry-over between samples was eliminated by injecting from the lowest to the highest amount of proteins in blocks with blanks in-between. For the four highest input amounts (10  $\mu\text{g}$  to 1.25  $\mu\text{g}$ ), an estimated equivalent of 1  $\mu\text{g}$  was measured to avoid the overloading of the analytical column. As a result, the number of absolute quantified proteins and their summed intensities reached a plateau, indicative for overall good recovery of peptides from the autoSP3 beads. The remaining samples ( $< 1$   $\mu\text{g}$ ) were sufficient for a single-shot injection. Across the whole range of protein input amounts, we observed narrow error distributions. This illustrated a high degree of reproducibility across all sample amounts. The injection of sub-microgram amounts of the sample was sufficient to quantify, for example, 403 and 681 proteins from



~39 ng and ~80 ng material, respectively. The data illustrate the efficiency of autoSP3 (SP3) to capture and rinse proteins, and to release peptides ready for LC-MS.



**Figure 9: Evaluation of autoSP3 sensitivity.** A) Schematic representation of the experimental design with a 1:2 dilution series of a HeLa batch lysate starting from 10 µg down to 5 ng. The distribution of samples across the 96-well plate is shown. The dilution series was prepared in four replicates, and samples were injected from the lowest to the highest concentration. For the four highest concentrated samples, 1 µg material was injected, whereas, for sub-microgram samples, the entire sample was used. The average number of quantified proteins per sample, as well as the corresponding sum iBAQ intensities, are shown with error bars from the 4 replicates. B) Schematic representation of the experimental design of processing small numbers of HeLa cells. Series of decreasing cell numbers were prepared from 10,000 to 10 cells in 8 replicates. The average number of quantified proteins per sample, as well as the corresponding sum iBAQ intensities, are shown with error bars from the eight replicates. Modified from Mueller et al., *Mol. Syst. Biol.*, 2020.

Taking this one step further, we started a similar experiment with counted numbers of cells to replace the HeLa batch lysate. This represents a more realistic scenario of limited input material, such as applications with patient-derived or FACS-sorted cells. In more detail, we counted a single-cell suspension of HeLa cells and directly transferred equivalents of 10,000 to 10 cells into a 96-well plate, corresponding to a range of 1 µg to 1 ng protein (assuming 0.1 ng protein per cell). To increase the reliability of the experiment, we processed a total of eight replicates per cell number in two independent 96-well plates (**Figure 9B**; see also **Figure 2D**). The entire sample processing, including cell lysis, DNA shearing, and autoSP3,

## Results

were performed without changing the sample plate. For each sample, the entire volume was subjected to LC-MS. Both numbers of absolute quantified proteins, as well as their intensities, scaled with the protein input. For the 1 µg-sample (10.000 cells), we quantified almost 2000 proteins, which is in range with our expectations (compared to **Figure 9A** and **Figure 2D**) and considering the utilized LC-MS setup and gradient length. A great end-to-end reproducibility was further highlighted by the overall narrow error distributions across all replicates and two independent sample plates. This was also seen in a similar experiment using the Covaris for cell lysis and DNA shearing (**Figure 2D**). As little as 100 cells of starting material was sufficient to quantify 459 proteins on average.

In summary, our autoSP3 workflow is capable of reproducibly processing minute amounts of starting material and providing sufficient sample quality. This allows the quantification of several hundreds of proteins from as few as 100 to 1000 cells or below 100 ng protein. Beyond the scope of this work, this opens the path for exciting new applications for which no reliable sample processing was available. This can be part of particular interest in a clinical context, in which sample availability is scarce, but demands in data depth and quality are high. Providing the ability to process these samples in an automated fashion eliminates the inflated issue of reproducibility when handling low amounts of material.

### **4.2.4. autoSP3 and challenging specimens**

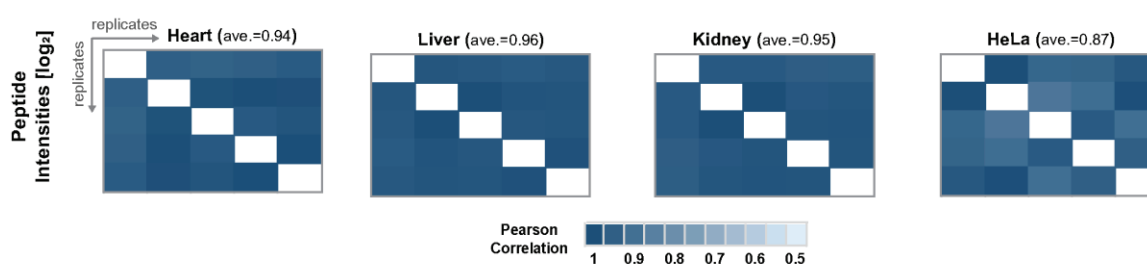
An important requirement to establish a broad involvement of proteome profiling is the intrinsic ability to convert difficult-to-handle samples to high-quality data. FFPE samples comprise the most obvious source of challenging input material<sup>159,168,330–332</sup>. They are the specimen of choice for histopathological diagnosis and routinely collected for cancer patients or other diseases, making it an invaluable resource in translational research. Formalin induces cross-linking between proteins to conserve and stabilize the integrity of the tissue and enables long-term storage. Both linked proteins (linked peptides) and paraffin interfere with global proteome profiling and require a suitable sample preparation strategy. The de-crosslinking of peptides and proteins is commonly achieved by treatment with SDS, which can be efficiently removed before the tryptic digestion by autoSP3 (SP3). Its robustness, sensitivity, and flexibility are additional assets that uniquely qualify autoSP3 (SP3) for the processing of FFPE specimens. Here, we combined the ability to process low amounts of material with a clinical real-world FFPE tissue cohort of pulmonary ADC.



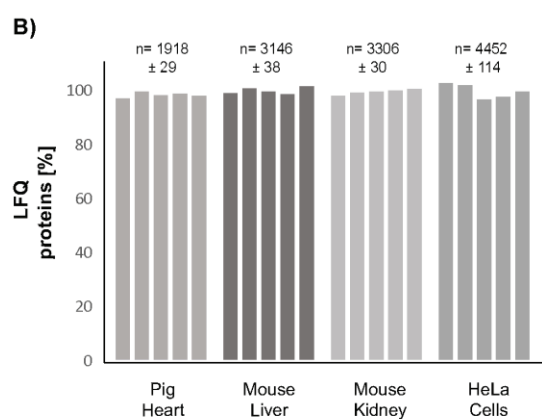
## 4.2.4.1. End-to-end sample preparation

Thus far, the complete workflow from a raw specimen to injection-ready peptides for LC-MS is fully established for cells and fresh-frozen tissue using the Covaris interfaced with autoSP3. To demonstrate the efficient end-to-end processing, we lysed 100.000 HeLa cells (n=15) and varying amounts (1.5 to 7.5 mg wet weight) of different fresh-frozen tissue types, as previously mentioned (**Figure 2E**). Upon extraction of proteins in the Covaris 96-well AFA-tube TPX plates, they were transferred directly to the Bravo liquid handling robot for autoSP3 using the established protocols. From each processed sample type (HeLa cells, pig heart tissue, mouse kidney, and mouse liver tissue), we randomly selected five replicates and continued to acquire LC-MS data, resulting in highly consistent numbers of identified peptides with an average Pearson correlation of higher than 0.94 (**Figure 10A**).

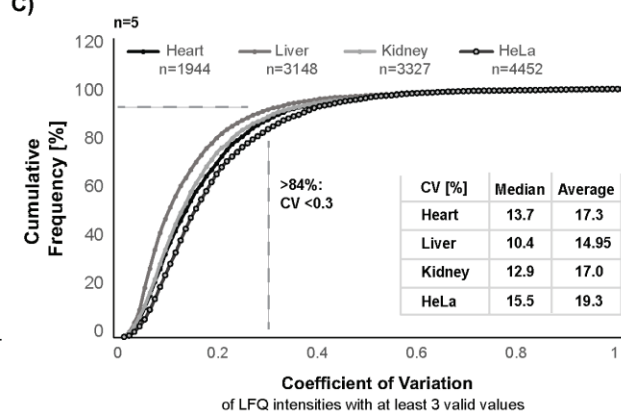
A)



B)



C)



**Figure 10: End-to-end proteome profiling using ultrasonication interfaced with autoSP3.** A) Pearson correlation heatmap of peptide intensities across five replicates of each sample type (heart, liver, kidney, HeLa cells) with the corresponding average. B) The relative number of identified and quantified proteins across the five replicates of each sample type. The average number of identified proteins and the standard deviation across five replicates is shown on top. C) Cumulative frequency curve [%] of the observed coefficient of variation (CV) [%] of proteins that were identified and quantified in at least three out of five replicates in each sample type. The resulting median and average CV [%] for each sample type are shown. Modified from Mueller et al., *Mol. Syst. Biol.*, 2020.

HeLa cells exhibited a marginally lower average at 0.87, which likely reflects variation in manual cell counting and aliquoting of small cell numbers. Similarly, the number of

## Results

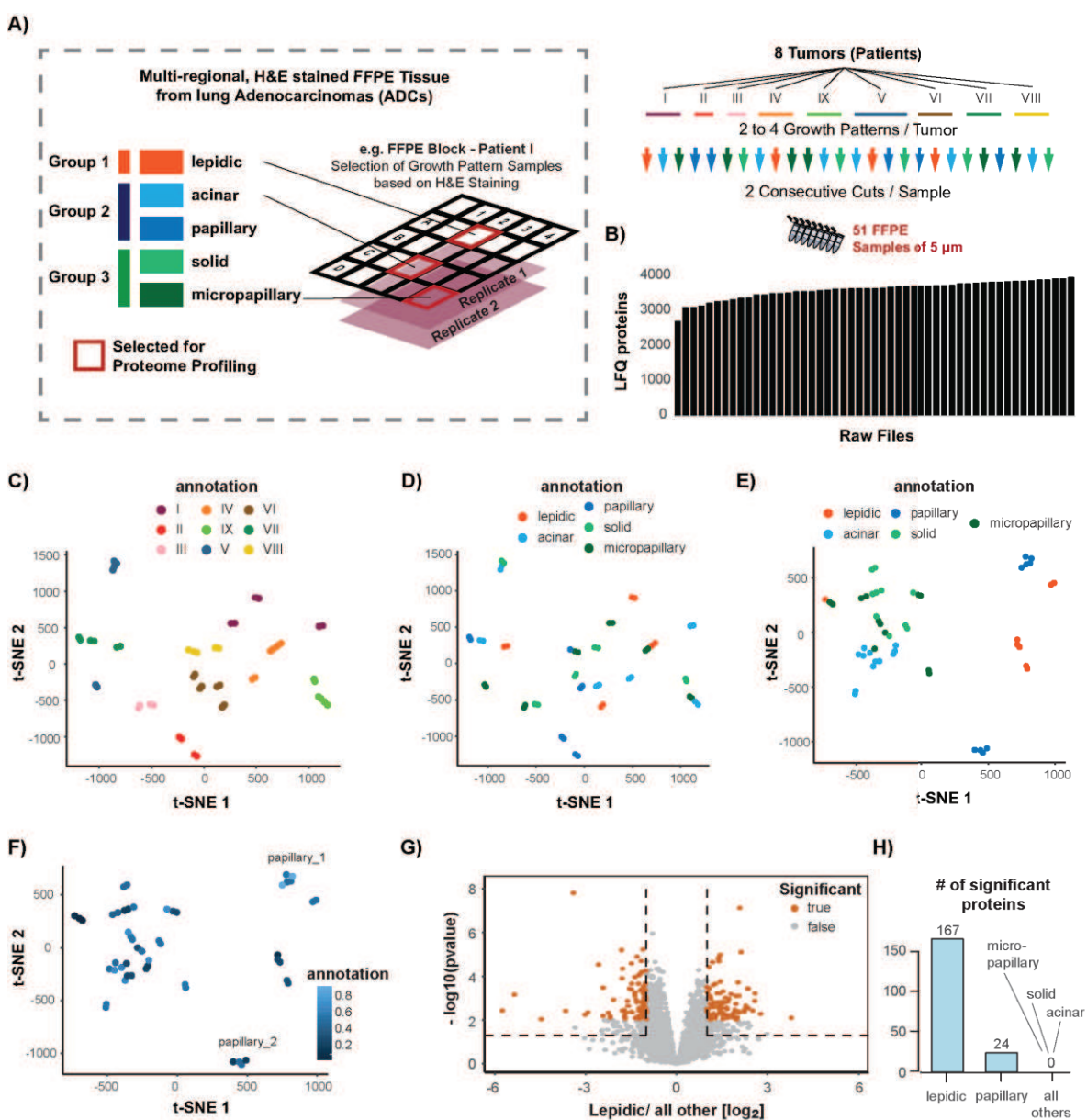
quantified proteins per sample type was highly reproducible (**Figure 10B**). The CV [%] distribution within each sample type revealed that more than 84% of all proteins could be quantified with a CV [%] lower than 30% (**Figure 10C**). Median CV's [%] between 10.4% (liver) and 15.5% (HeLa cells) demonstrated the processing precision spanning the entire procedure from tissue lysis to data acquisition using LC-MS. The end-to-end workflow takes about 3.5 hours for 96 samples, including one hour for ultrasonication for tissue lysis, protein extraction, and DNA shearing. For FFPE tissue, we are currently still working on the full integration of the AFA-based de-paraffinization (Covaris) and the subsequent interfacing with the autoSP3 setup. Therefore, the ADC FFPE cohort was collected in PCR 8-strips and lysed in two batches using the Bioruptor Pico with a customized tube holder. Upon protein extraction with sufficient DNA shearing, we manually transferred all samples in a random order to a 96-well plate to perform autoSP3.

### 4.2.4.2. Pulmonary adenocarcinoma (ADC) FFPE

Five different growth patterns of ADC are recognized by the WHO but yet remain without a comprehensive clinical implication. Gene expression differences have only been identified between lepidic ADCs and all other histologic patterns<sup>217</sup>. The remaining growth patterns (acinar, papillary, solid, and micropapillary) have not been characterized, despite their known higher invasiveness. Here, we aimed to perform proteomic profiling to potentially identify functional differences, a causality for the different growth patterns, or biomarkers and therapeutic targets.

We collected consecutive 5 mm x 5 mm x 5  $\mu$ m sections of tumors blocks that originated from eight different lung cancer patients. Every section was histologically classified (Hematoxylin and Eosin (H&E) staining) to locate and distinguish the multi-regional composition of growth patterns. Subsequently, two to four growth patterns could be selected per tumor, resulting in a total of 51 samples that were processed using our pipeline (**Figure 11A**). All samples were randomized during sample preparation and LC-MS acquisition. On average, we quantified 3576 proteins across the entire cohort (**Figure 11B**) using  $\frac{1}{4}$  of the available sample material. Consecutive sections (biological replicates) exhibited a nearly perfect similarity (**Figure 11C** and **Figure 11D**) despite their 2-fold randomization during the process. This particularly highlights the reproducibility of our workflow. The grouping was mainly driven by the patient of origin (**Figure 11C**) rather than

the respective growth patterns (**Figure 11D**). Taking this into account as a batch-effect, we applied a linear regression model and achieved a rudimental separation of the three superordinate groups as a result (**Figure 11E**). The superordinate groups are defined as: I) lepidic (low grade; group 1), acinar and papillary (intermediate grade; group 2), and solid and micropapillary (high grade; group 3). Both lepidic and papillary growth patterns could now be separated from all other samples. At the same time, consecutive sections with the highest likelihood to be similar were still grouped. The dissimilarity within papillary samples, split into two distinct subclusters and separated from group 2 (acinar), was somewhat unexpected. While the tumor cell content (TCC) might explain this observation, it was rather randomly distributed over all samples (**Figure 11F**).



## Results

**Figure 11: Proteome profiling of tumor growth patterns of pulmonary Adenocarcinoma (ADC) FFPE tissue.** A) Schematic illustration of the sample collection. Samples were collected from eight different patient tumors. For each tumor, sections were processed with hematoxylin & eosin (H&E) staining to locate different growth patterns of lepidic (low-grade; group 1), acinar and papillary (intermediate grade; group 3), and solid and micropapillary (high-grade; group 3). Two to four growth patterns per tumor were selected and sectioned in two consecutive 5  $\mu$ m iterations to provide replicates with the highest possible similarity, resulting in a total of 51 samples (one iteration was missing). B) Bar plot summarizing the number of quantified LFQ protein groups per sample. C) t-distributed stochastic neighbor embedding (t-SNE) analysis of the uncorrected proteome data. The samples are color-coded according to their patient origin. D) Same as in C, now color-coded according to their tumor growth pattern. E) t-distributed stochastic neighbor embedding (t-SNE) analysis of the proteome data corrected via a linear regression model. The different growth patterns are color-coded as in panel A. F) Same as in E, now color-coded for the tumor cell content (TCC) [%] of each sample. G) Volcano plot showing differential expression analysis using Limma moderated t-statistics for the comparison of lepidic samples against all other samples. Proteins passing significance thresholds of  $-\log_{10}$  p-value  $< 0.05$  (Benjamini-Hochberg adjusted) and an absolute  $\log_2$  fold change of  $>1$  are highlighted in orange. H) Summary of significantly expressed proteins in the comparison of each growth pattern against all others. Modified from Mueller et al., *Mol. Syst. Biol.*, 2020.

A comparison between both subclusters of papillary samples (**Supplementary Figure 6A**) revealed 73 differentially expressed proteins (**Supplementary Figure 6B and Supplementary Figure 6C**). Collagen- and extracellular matrix-related gene sets were found to be enriched within papillary\_2 (**Supplementary Figure 6D**), using a gene-set enrichment analysis, which might hint to differing tumor microenvironments. In papillary\_1, we found an overrepresentation of mRNA nonsense-mediated decay and translation. The elimination of dysfunctional mRNAs might show a selective impairment in one of both subclusters. More analyses, beyond the scope of this project, are needed to unravel these phenomena in more detail.

Next, we used a Limma moderated t-statistics differential expression analysis to identify growth pattern-specific proteins (**Figure 11G and Figure 11H**). This was done by comparing the expression profiles of each group against all other groups. The highest number of differentially abundant proteins (167) was found in lepidic tissue (**Figure 11G**). We further subjected these proteins to a gene ontology (GO)-term enrichment analysis (**Supplementary Figure 6E**). Again, collagens were among the high abundant proteins, possibly reflecting a different composition of the extracellular matrix. Lung cancer growth, invasion, and metastasis have previously been linked to collagens<sup>333,334</sup>. Mitochondrial ribosomal proteins (MRPs), previously reported as a predictor of survival and progression with potential prognostic value in NSCLC<sup>335</sup>, were found among the significant proteins. Gene sets were enriched for *metabolism of polyamines* and *glucose metabolism* in all groups compared to lepidic samples (**Supplementary Figure 6F**). The increased capability of polyamine synthesis is linked to accelerated tumor spreading and invasiveness<sup>336</sup>.

Another key characteristic of the majority of NSCLC tumors is the absorption of glucose and metabolism towards anaerobic pathways, which is strongly associated with higher aggressiveness<sup>337</sup>. All of the mentioned observations are in line with the already known high aggressiveness and unfavorable prognosis of intermediate and high grade (group 2 and group 3) growth patterns as compared to low-grade lepidic samples<sup>206</sup>.

We further identified 24 proteins that were significantly different between papillary and all remaining samples. Among them, a subunit of the glycosylphosphatidylinositol transamidase complex, namely PIGT, was overexpressed. This deregulation has been associated with NSCLC in comparison to small cell lung carcinoma and healthy lung tissue. The potential implication for disease diagnostics, prognostics, and therapeutic intervention has been proposed<sup>338</sup>. *Golgi-associated vesicle budding, intra-Golgi, and Golgi-to-ER trafficking*, as well as *retrograde transport at the trans-Golgi network*, were identified among the top 10 most significantly enriched gene sets (**Supplementary Figure 6G**). This might implicate the involvement of the secretory pathway. Taken together, our data suggest that papillary-specific pathology extensively interacts with its environment. A coherence between NSCLC and secreted proteins has been postulated previously<sup>339</sup>. A differentiation between individual ADC growth patterns on the molecular level did not exist up until now. Here, we could identify regulated proteins for lepidic and papillary growth patterns. For the remaining three growth patterns, we could not identify differentially abundant proteins.

The primary purpose of the generated dataset was to demonstrate the applicability of our autoSP3 workflow and showcase the processing of a realistic, clinically-relevant FFPE cohort with quantity-limited material. Despite the randomization of samples during the sample preparation and during the LC-MS analysis, we found a tight grouping of the consecutive FFPE sections. This illustrates a high precision of the autoSP3 protocol. Further, the end-to-end processing virtually eliminates most manual sample handling steps, where an active user is only required for the plate transfer from the Covaris to the Bravo to the PCR cycler for proteolytic digestion, and back to the Bravo. In the end, injection-ready peptides can be recovered into a new sample plate, which is directly compatible with the LC autosampler. The overall hands-on time is reduced to less than 5 minutes. In addition to the technical aspects, our other interest was the molecular characterization of the histologic ADC growth

## Results

patterns. Due to the wide range of different tumor cell contents per sample, we refrain from drawing any conclusion besides confirming previous knowledge. However, we anticipate further experiments with *firstly*, microdissection of FFPE slides to achieve maximal TCC, and *secondly* with a fully established end-to-end processing for FFPE, including the de-paraffinization (Covaris) in combination with autoSP3. Both will add to the biological data quality and enable an improved interpretation of functional differences between ADC growth patterns and their potential clinical implications.

### **4.3. Ependymoma (EPN) brain tumors**

In the last chapter of this thesis, we aimed to demonstrate the added value of large-scale protein expression profiling, complementary to other NGS-layers, for translational research. For this purpose, we utilized a cohort of EPN pediatric brain tumors, which have been extensively characterized by our collaborators on various levels, including genetic, epigenetic, transcriptional, demographical, and clinical data<sup>73</sup>. Nine distinct molecular subgroups were classified based on the DNA methylome, expanding histopathological grading that suffers from a limited clinical utility, poor interobserver reproducibility, and lack of predictive potential for a patients' outcome. Despite the superiority of molecular classification to histological grading, the majority of subgroups still lack a functional explanation, and genetic drivers remain unknown. In many cases, the correlation between cancer entities or states and their corresponding proteome composition is unclear. This also applies to ependymoma and its nine molecular subgroups. Here, the proteome profiling has the prospect to elevate our current understanding of EPN biology and identify actionable, subgroup-specific pathways, and targets.

For proteomic profiling, we utilized fresh-frozen tissue in the range of 3.5 to 6.6 mg wet weight and the following numbers of samples per subgroup: SP-EPN (n= 5), SP-MPE (n= 7), SP-SE (n= 3), PF-EPN-A (n= 24), PF-EPN-B (n= 12), PF-SE (n= 7), as well as ST-EPN-RELA (n= 20), ST-EPN-YAP1 (n= 4), ST-SE (n= 5), and healthy (n= 5). The sample preparation was carried out simultaneously to the evaluation and optimization of the SP3 protocol (Chapter 4.1) and before its automation (Chapter 4.2). The acquisition of the final dataset was performed using the parameters described in the method section. In total, we identified and quantified 8248 proteins in the EPN cohort.

Parts of the following chapter, including Figures and Tables, were taken in part or their entirety from the joint publication listed below.

Hübner, J. M., **Mueller, Torsten**, Papageorgiou, D. N., Mauermann, M., Krijgsveld, J., Russell, R. B., Ellison, D. W., Pfister, S. M., Pajtler, K. W., Kool, M. (2019). „**EZH1P/CXorf67 mimics K27M mutated oncohistones and functions as an intrinsic inhibitor of PRC2 function in aggressive posterior fossa ependymoma.**” *Neuro Oncology* 21(7): 878-889.

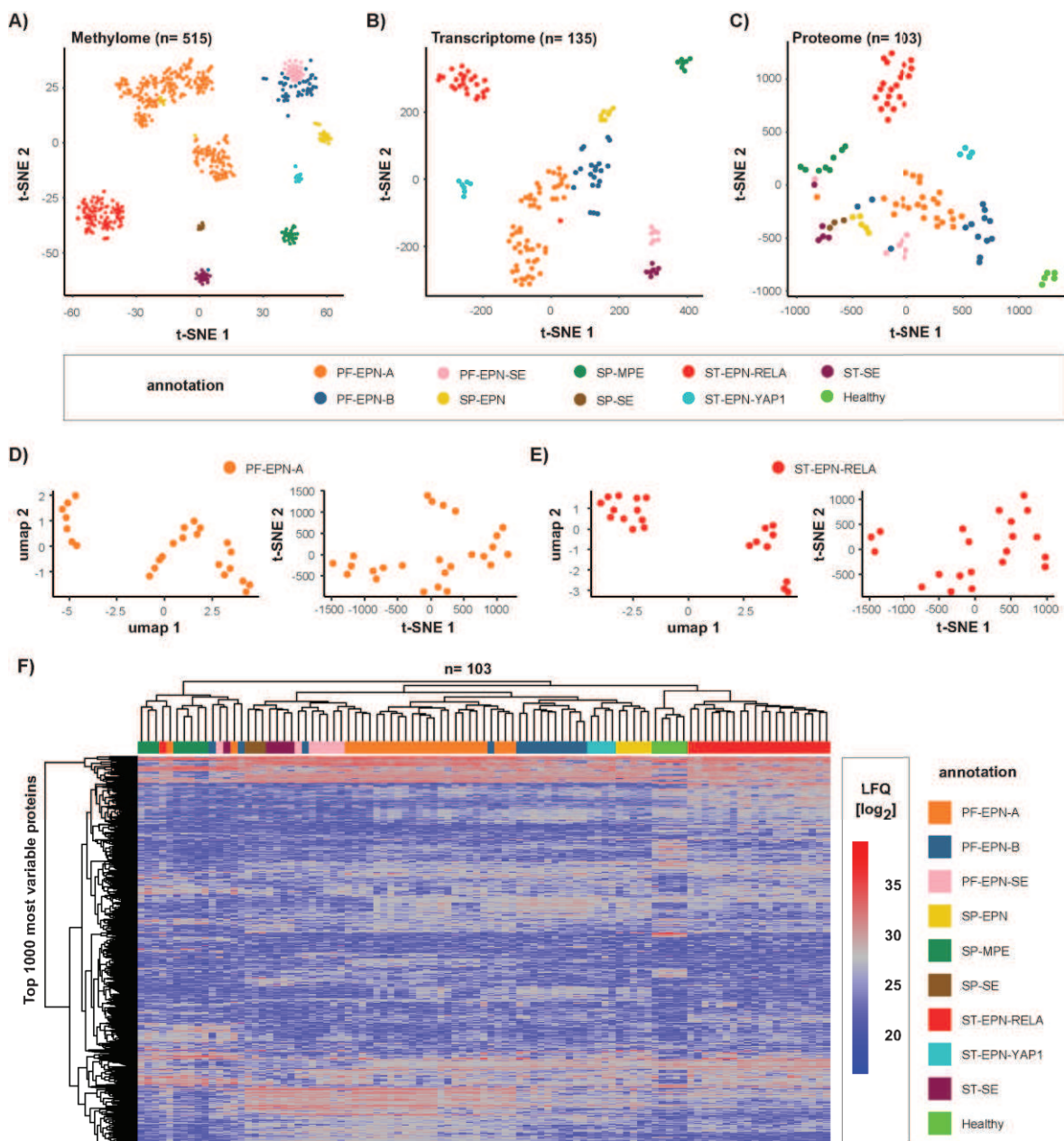
#### 4.3.1. Proteome profiles of molecular subgroups

Protein expression is dynamically regulated in a spatiotemporal manner and exhibits a high complexity as a consequence<sup>90</sup>. Hence, we firstly investigated whether the global proteome information suffices to discriminate between the molecular subgroups as defined based on DNA methylation patterns (n= 515; **Figure 12A**). Affymetrix gene expression profiles (n= 135; **Figure 12B**), as well as proteome profiles (n= 103; **Figure 12C**), result in a fine recapitulation of the classification across all tumor types and CNS compartments. This is further indicated by the observed silhouette scores per omics layer (methylome s= 0.57, transcriptome s= 0.59, and proteome s= 0.43). The clearest separation is observed from the transcriptome data and the methylome data. For the latter, the top 5% of CpG probes with the highest standard deviation were used (~21.000 CpG probes).

Both methylome and transcriptome data exhibit an additional sub-subgroup within PF-EPN-A tumors. Although less obvious, this observation holds for the proteome level when focusing on the individual subgroup (**Figure 12D**). The proteome additionally revealed a subgrouping within ST-EPN-RELA tumors (**Figure 12E**). To demonstrate this, we utilized an unsupervised hierarchical clustering of the top 1000 most variable proteins (**Figure 12F**). The vast majority of samples were clustered according to their molecular subgroup. The separation of subgroups based on their anatomical location (PF, ST, and SP) was less pronounced as compared to the methylome data<sup>73</sup>. For example, SP-EPN samples were clustered close to ST-EPN-YAP1 rather than the remaining SP subgroups. Using the proteome data, we identified 2 and 3 sub-subgroups for ST-EPN-RELA and PF-EPN-A, respectively. A more detailed analysis of underlying differences is provided in chapter 4.3.3.4.

## Results

Altogether, the generated proteome profiles of the EPN cohort seem to recapitulate the previous molecular classification. The differences between anatomical regions are less prominent as compared to the methylome data. Still, it is interesting that all omics-layer lead to a similar subgrouping. This might be expected following the basic principle of silenced or active chromatin regulating the expression of genes and proteins as a result. However, the driving features per omics-layer (underlying genes of driving CpGs, driving transcripts, and driving proteins) are different. Hence, we continue with an in-depth analysis of gene- and protein expression to further investigate subgroup-specific functional implications.





**Figure 12: Molecular classification of ependymoma (EPN) tumors and sub-subclassification.** A-C) t-distributed stochastic neighbor embedding (t-SNE) analysis of EPN methylation patterns (n= 515, top 5% of CpG probes with the highest standard deviation) (A), the transcriptome gene expression (n= 135), and the proteome profiles (n= 103). D-E) umap and t-SNE analysis of PF-EPN-A (D) and ST-EPN-RELA (E) tumors, revealing sub-subgroups in the proteome composition. F) Hierarchical clustering of the top 1000 most variable proteins. Methylome and transcriptome data were provided by our collaborators from Pajtler et al., 2015.

#### 4.3.2. Subgroup-specific putative marker proteins

The comprehensive dataset provides a unique opportunity to identify molecular features, such as genes or proteins, that can be utilized as indicative biomarkers for a (patho)-physiological state or an EPN subgroup. The extraction of biomarkers or functional signatures can be increasingly challenging as a result of more complex datasets. Here, proteins that are exclusively expressed within a single EPN subgroup represent easily accessible candidates. We could identify several uniquely expressed proteins for each subgroup, some of which are additionally found within the healthy reference tissue (**Figure 13A**). The vast majority of the latter group of proteins do not exhibit a significantly changing expression compared to the healthy control group. Other proteins, such as SLC38A1 (in PF-SE), an amino acid transporter, are solely found within a respective subgroup and not in the healthy tissue. SLC38A1 has been associated with proliferation, migration, and tumor progression in other cancer entities but not EPN.

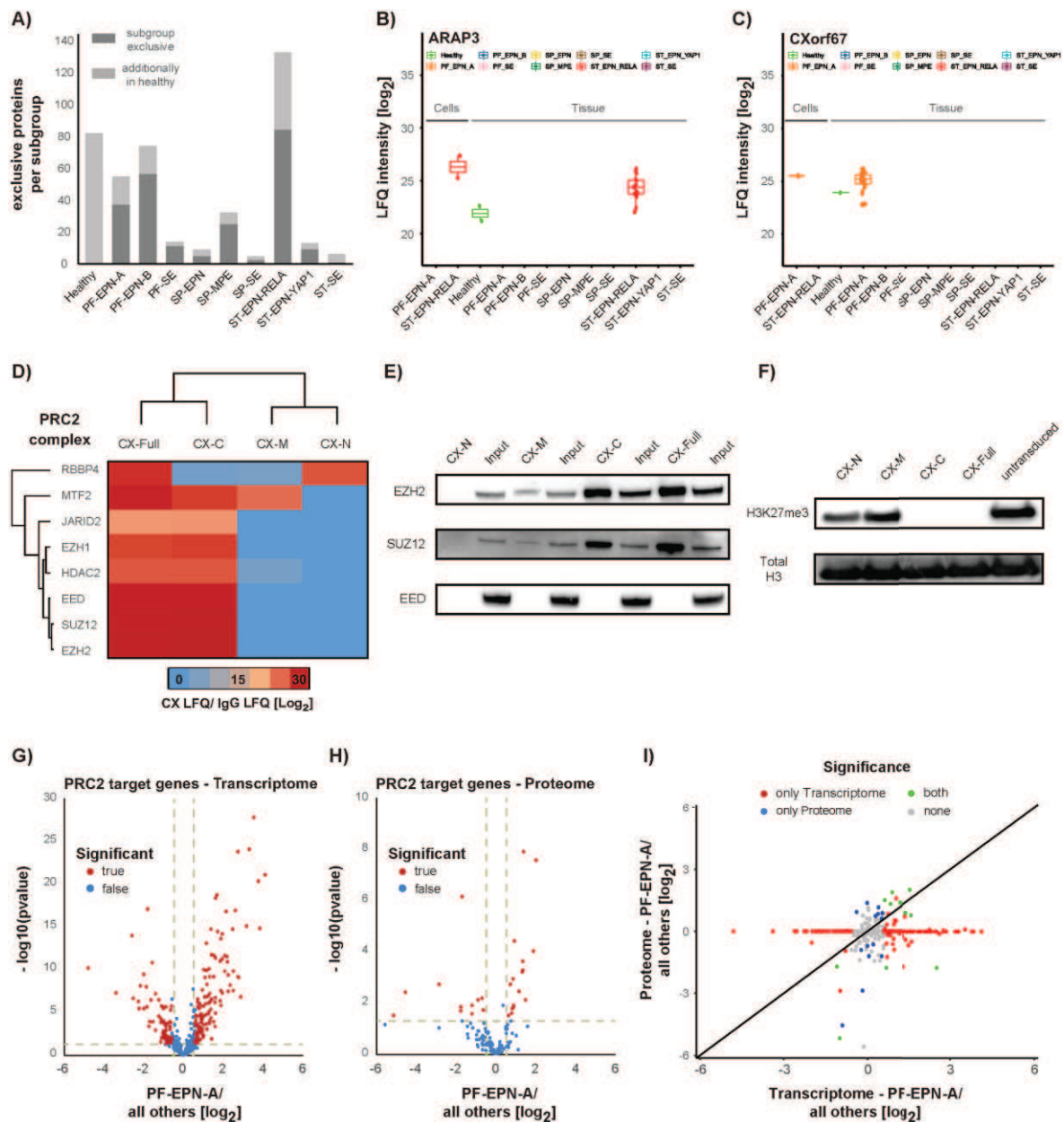
The highest number of unique proteins was identified within ST-EPN-RELA. Among them, we consistently observed ARAP3 (in 15/ 20 samples; **Figure 13B**), a GTPase-activating protein for ARF and RHO family members. It is known as a genuine effector of the phosphoinositide 3-kinase (PI3K) signaling pathway and involved in downstream regulation of angiogenesis<sup>340</sup>. Interestingly, a previous supervised gene expression analysis has suggested ARAP3 as a signature gene for ST-EPN-RELA tumors and a significant enrichment of angiogenesis in a pathway analysis in Pajtler et al., 2015<sup>73</sup>.

##### 4.3.2.1. CXorf67 (EZHIP): an intrinsic inhibitor of PRC2 in PF-EPN-A

The overexpression of chromosome X open reading frame 67 (CXorf67) was previously discovered as a hallmark of PF-EPN-A tumors by our collaborators<sup>265</sup>. Matching to their observation, CXorf67 was solely identified and quantified in PF-EPN-A tumors (in 22/ 24 samples, **Figure 13C**) and overexpressed compared to the healthy control in our proteome data. Statistically, we could not determine a fold change because it was only observed in a

## Results

single healthy sample. Its expression has recently been proposed as a mechanism for the downregulation of the repressive histone H3 lysine 27 trimethylation (H3K27me3) mark, another characteristic feature of these tumors. The negative regulation stems from the interaction of CXorf67 with the enhancer of zeste homolog 2 (EZH2) methyltransferase, a constituent of the polycomb repressive complex 2 (PRC2). The PRC2 histone methyltransferase primarily functions to trimethylate H3K27 to promote the silencing of genomic regions<sup>341</sup>. Its inhibition drives the H3K27 hypomethylation and de-repression of target genes as a result.



**Figure 13: Ependymoma (EPN) subgroup-specific protein expression and CXorf67, an intrinsic inhibitor of PRC2 in PF-EPN-A.** A) The numbers of uniquely expressed proteins for each subgroup (dark grey) and with additional expression in the healthy reference samples (light grey). B-C) Boxplot illustration of ARAP3 (B) and

CXorf67 (C) protein expression across all EPN subgroups. D) PRC2 core component interaction heatmap with the CXorf67 deletion mutants (CX-C, CX-M, and CX-N) and the full length (CX-Full) protein, generated by co-immunoprecipitation. E) Western blot analysis for PRC2 components (EZH2, SUZ12, and EED) against all three deletion mutants and the full-length CXorf67 protein. F) Western blot analysis showing the absence of the H3K27-trimethylation (me3) mark in CX-Full and CX-C. G-H) Differential expression analysis using Limma moderated t-statistics for the comparison of PRC2 target gene expression (G) and protein expression (H) in PF-EPN-A tumors against all others. Genes (G) or proteins (H) passing significance thresholds of  $-\log_{10}$  p-value  $< 0.05$  (Benjamini-Hochberg adjusted) and an absolute  $\log_2$  fold change of  $>1$  are highlighted. I) Correlation analysis of significantly changing PRC2 targets on gene and protein-level to identify specific or common effects. The co-IP and western blot experiments were performed by Dr. Jens Huebner (Huebner et al., 2019). The transcriptome data were provided by our collaborators from Pajtler et al., 2015. Panel D-F were modified from Huebner et al., *Neurooncology*, 2019.

Here, we continued to unravel the precise mechanism of action of CXorf67-mediated inhibition of PRC2. This was done in collaboration with Dr. Jens Hübner, Dr. Marcel Kool, and Dr. Kristian Pajtler<sup>257</sup>. In brief, HEK293T cell lines were transfected with a doxycycline-inducible expression system for the full-length CXorf67 protein and three different CXorf67 deletion mutant constructs as follows: I) amino acids 1 to 150 (N-terminal region), II) amino acids 151 to 300 (Middle region), and III) amino acids 301 to 503 (C-terminal region) (**Supplementary Figure 7A**). Hereinafter they are referred to as CX-Full, CX-N, CX-M, and CX-C. Their selective expression and additional localization to the nucleus was confirmed using western blot analysis (**Supplementary Figure 7B** and **Supplementary Figure 7C**). Subsequently, we performed a co-immunoprecipitation (co-IP) using an anti-FLAG antibody followed by LC-MS analysis to identify putative interaction partners for each construct compared to the CX-Full. The vast majority of identified PRC2 components exclusively interacted with CX-Full and CX-C, indicating a functional domain in the C-terminal region of CXorf67 (**Figure 13D**). The MS results were cross-validated using western blot analysis (**Figure 13E**), which showed an additional marginal interaction of EZH2 and SUZ12 with CX-M. The pull-down of EED could not be validated using western blot analysis.

Next, the MS results were further validated by staining all transduced cell lines for the presence of Flag-tagged proteins (CXorf67, CX-N, CX-M, and CX-C) and H3K27me3. This indeed revealed a hypomethylation of H3K27 in cell lines expressing CX-Full or CX-C, indicating that the C-terminal region of CXorf67 is sufficient for the inhibitory effect on PRC2 (**Supplementary Figure 7D**). The results were cross-validated using western blot analysis showing the absence of the H3K27me3 mark in CX-Full and CX-C (**Figure 13F**). Highly similar transcriptional changes, including enrichment of PRC2 target genes, were observed in CX-Full and CX-C cell lines as a result of the hypomethylation<sup>257</sup>.

## Results

Finally, we utilized the available global gene expression profiles and our proteome profiling data to investigate the effect of CXorf67-mediated inhibition of PRC2 and the de-repression of its target genes compared between molecular subgroups. The list of relevant targets was extracted from MsigDB<sup>342</sup>. Neither on the level of gene expression nor in the proteome profiles, we could observe a statistically significant difference between PF-EPN-A tumors and all others (**Supplementary Figure 8E** and **Supplementary Figure 8F**). For the gene expression data, we did not have access to any healthy reference data or the SP-SE subgroup. Using a Limma moderated t-statistics differential expression analysis, we could find a subset of PRC2 target genes that are regulated in PF-EPN-A tumors compared to all others on the transcriptome (196/585) (**Figure 13G**) and proteome (27/125 identified) (**Figure 13H**) level. The vast majority of transcriptional changes were not identified on the proteome-level (**Figure 13I**). Very few protein targets show a regulation in both omics-layers (n= 12) or exclusively in the protein data (n= 15) (**Figure 13I**). The fact that we see a significant regulation of PRC2 targets in the CX-Full and CX-C might be the artificial overexpression of CXorf67, which is generally low abundant in the PF-EPN-A proteome data. Also, we do not have any global proteome data for the CX-Full and CX-C transduced cell lines.

Recapitulating, we could identify a number of proteins that exhibit subgroup-specific expression. Among them, ARAP3 in ST-EPN-RELA and CXorf67 in PF-EPN-A are interesting examples. The exclusive expression of CXorf67 was already known in part, but the precise mechanism was unraveled in a collaborative project. Here, we identified the putative interaction partners for the N-terminal region, the middle region, and the C-terminal region of CXorf67 (**Supplementary Figure 7A**). Most importantly, the PRC2 core components exclusively interacted with either the full-length protein (CX-Full) or the C-terminal region (CX-C). Using our tumor proteome profiles, we could further investigate the influence of PRC2 inhibition on its targets compared to other EPN subgroups. Other exclusively expressed candidates are under further investigation.

### 4.3.3. Protein- and gene expression

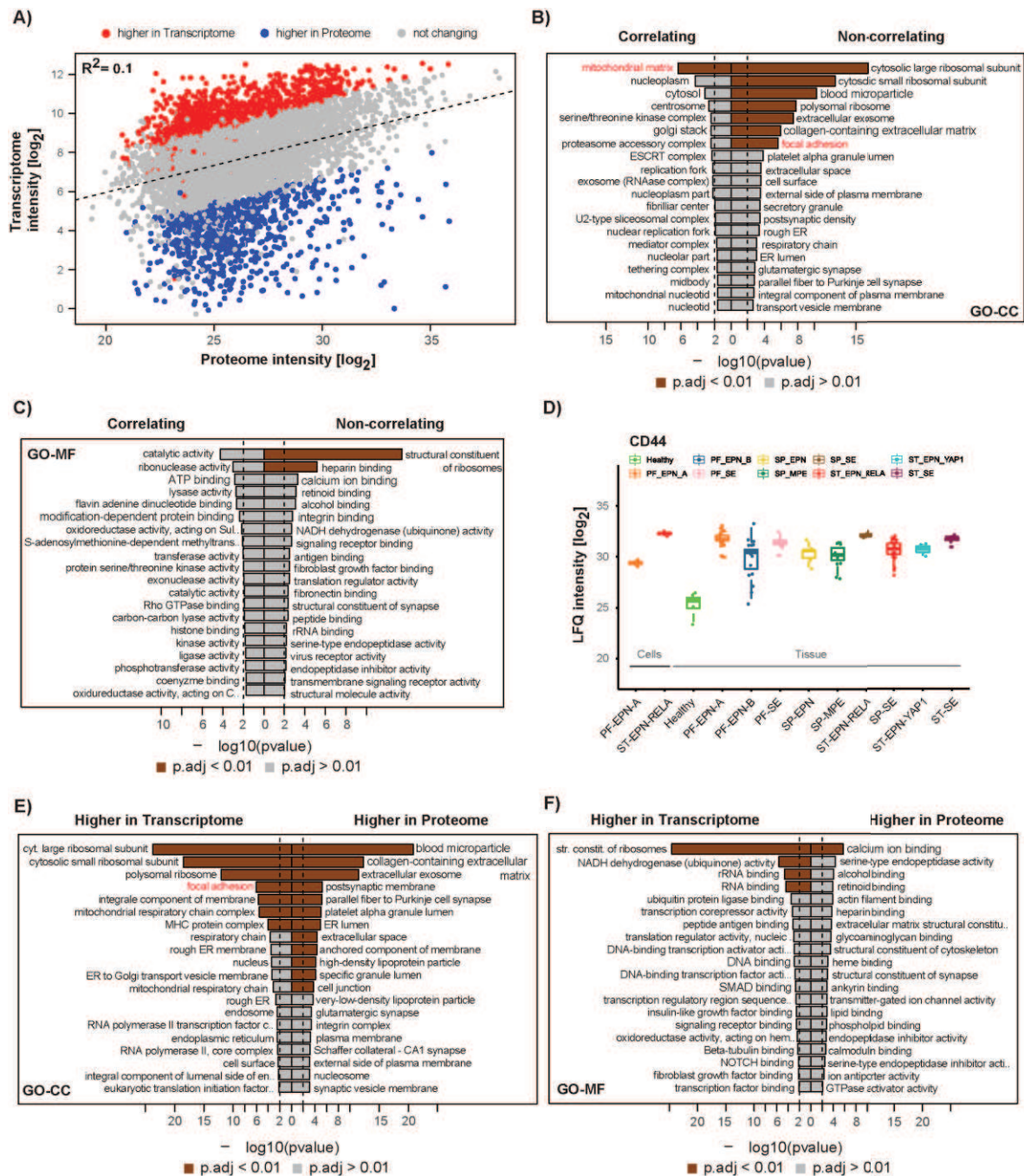
The identification of proteins that are exclusively expressed within a single subgroup presents the ideal scenario for the discovery of biomarkers or functional differences. In practice, exclusively expressed proteins only present the surface of the several hundreds of

identified and quantified observations. In the following chapter, we systematically examined our proteome data with and without the available gene expression data. Initially, we performed a correlation analysis between the previously generated gene expression and our newly generated proteome profiles. The coefficient of determination ( $R^2= 0.1$ ) illustrates a weak overall correlation between intensity data of gene- and protein expression (**Figure 14A**). This observation is expected due to various regulatory mechanisms during protein biosynthesis that lead to a non-linearity from genes to proteins<sup>90</sup>. On the other hand, the additional value of proteome profiles immediately becomes evident by I) proteins that are either non-correlating but significant or II) correlating with their gene expression.

In order to identify both groups of proteins, we utilized the mean intensity correlation between gene and protein expression and calculated a corresponding linear regression model. This allowed us to determine the deviation of each protein from this general model (residuals of true observed and expected intensities) and to perform a gene/protein-wise comparison whether the normally-distributed values are significantly deviating from zero. A detailed description of the process is provided in the respective method section. Using this analysis, we could identify a large number of proteins that deviate in their intensities in one of the two ways: I) higher gene expression as expected from the proteome data (red), or II) higher protein expression as expected from gene expression data (blue) (**Figure 14A**). The remaining set of genes/proteins are statistically not significantly ( $p\text{-value} > 0.05$ ) varying between both expression layers and were thus classified as correlating. Interestingly, the correlating group of proteins showed enrichment of mitochondrial matrix proteins in a cellular component GO analysis (**Figure 14B**). This might reflect a tight regulation of the energy household as an essential mechanism for a cell's functional integrity. Other vital mechanisms are among the non-significant gene sets, for example, centrosome, Golgi stack, proteasome accessory complex, or the replication fork. The non-correlating proteins, on the other hand, exhibited a significant enrichment for structural constituents of ribosomes (GO-molecular function) (**Figure 14C**) and correspondingly the cytosolic large and small ribosomal subunits (GO-CC) (**Figure 14B**). In addition, we found proteins enriched that are localized in extracellular exosomes, secretory granule, cell

## Results

surface, and the collagen-containing extracellular matrix. Some of these observations will be picked up again throughout the thesis.



**Figure 14: Global gene- and protein expression correlation.** A) Global correlation of transcriptome and proteome intensities showcasing an overall weak correlation. B-C) Gene ontology analysis of cellular components (B) or molecular function (C) for correlating and non-correlating genes/proteins. Significance (brown) is defined with a  $\log_{10}$  p-value  $< 0.01$  (Benjamini-Hochberg adjusted). D) Boxplot illustration of CD44 protein expression across all ependymoma (EPN) subgroups. E-F) Gene ontology analysis of cellular components (E) or molecular function (F) for genes/proteins that are either higher in the transcriptome or proteome. Significance (brown) is defined with a  $\log_{10}$  p-value  $< 0.01$  (Benjamini-Hochberg adjusted). The transcriptome data were provided by our collaborators from Pajtler et al., 2015.

Among the deviating genes/proteins, we also found a significant enrichment of focal adhesion proteins. The cell-surface receptor CD44, a non-kinase transmembrane glycoprotein that mediates focal adhesion<sup>343–346</sup>, has previously been linked to PF-EPN tumors as an independent predictor of survival<sup>347</sup>. While the gene expression is reduced for PF-EPN-B and SP-EPN tumors, the protein-level intensities of CD44 are highly upregulated for all tumors compared to the healthy reference (logFC  $\mu$ = 6.2 across all subgroups) (**Figure 14D**). PF-EPN-B tumors show a more variable CD44 protein expression, while the vast majority of samples still exhibit a significant upregulation. The physiological role of CD44 comprises the maintenance of organ and tissue structure, but it also plays various roles in tumor initiation, invasion, and metastasis<sup>343,348</sup>. Its functional role in tumorigenesis is very much in the focus of ongoing research efforts. For example, its expression has been linked to a high expression of the signal transducer and activator of transcription (STAT3) protein and increased cell proliferation as a result. This is in line with the protein expression data for STAT3 in all tumors but SP-MPE (data not shown). The inhibition of CD44-STAT3 complex formation has been reported as a target in breast cancer<sup>349</sup>. In several other cancer entities, CD44 was reported as a potential molecular target for therapy against its tumorigenesis promoting role. Several preclinical and clinical trials are on their way, targeting CD44 expression<sup>348,350,351</sup>.

Next, we evaluated whether we can observe an enrichment of gene sets for genes/proteins, which are either higher or lower expressed in the proteome than expected from the transcriptome. The majority of previously found GO annotations were identified to have a higher expression in the gene expression data and correspondingly did not translate equally to the protein-level (**Figure 14E** and **Figure 14F**). The majority of proteins with an unexpected high expression were localized to the extracellular matrix or extracellular exosomes. Extracellular matrix (ECM)-related proteins are often enriched in cancer entities and can be associated with differences in the tumor microenvironment. The potential role of ECM-related proteins is further discussed throughout the thesis and particularly in chapter 4.3.3.4. Extracellular exosomes are discussed in more detail in chapter 4.3.4.

Beyond the global correlation of gene- and protein expression, we wanted to come back to the evaluation of differences between the molecular subgroups. This was achieved by performing the differential expression and GO-enrichment analysis for each subgroup

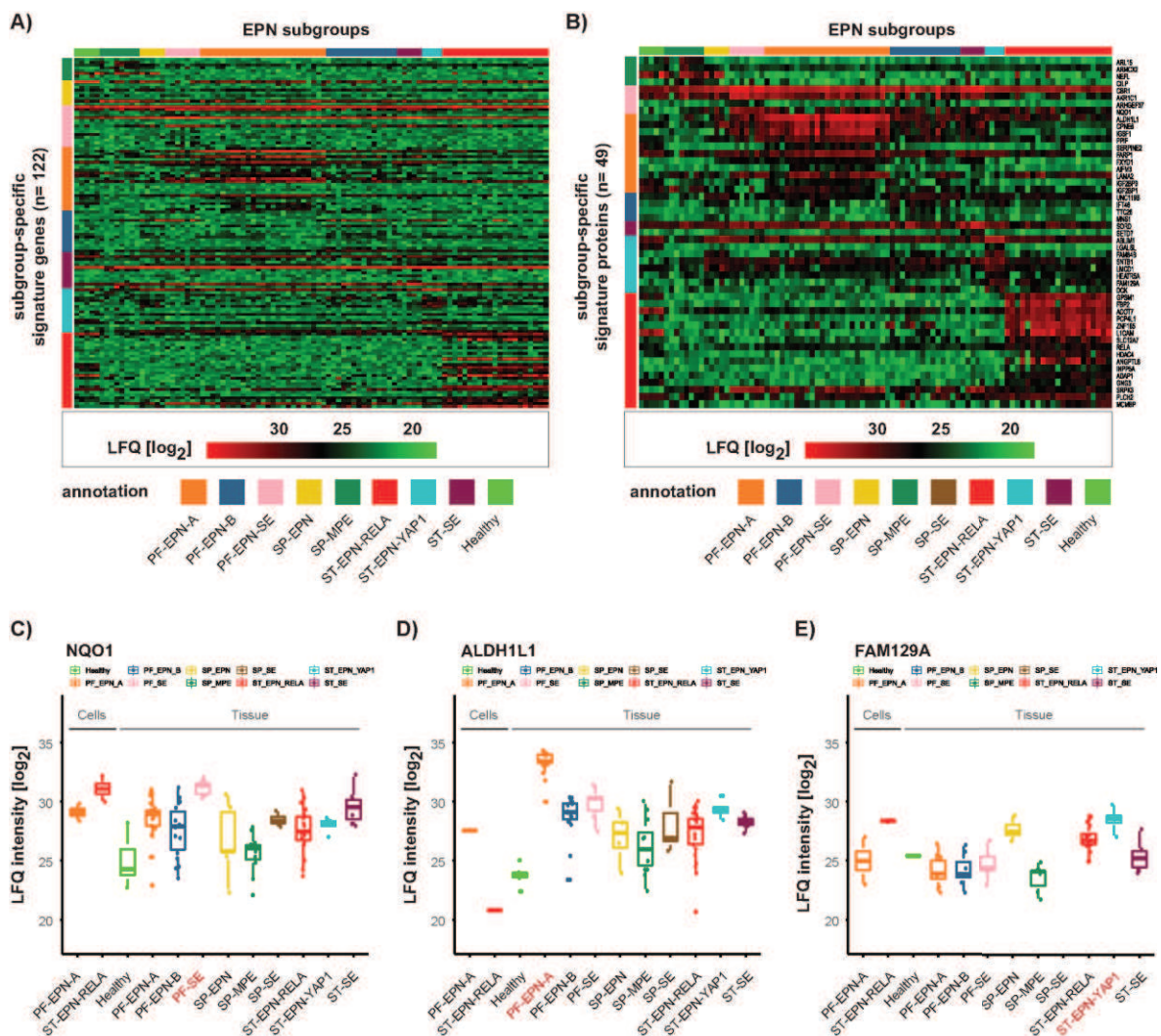


## Results

against all others. Very few subgroup-specific changes were identified (data not shown). This alludes to, for example, the increased intensities of ECM organization proteins in PF-EPN-B or cilium, sodium transport, and axonemal dynein complex in ST-EPN-RELA. ECM proteins were higher on the transcript level for PF-EPN-A and ST-EPN-RELA, which might be associated with their higher aggressiveness and worst prognosis. Since the overall outcome of this analysis was limited, with only a few significant observations, we decided to proceed with a more targeted approach, as outlined below.

### 4.3.3.1. Signature gene translation to proteins

Next, we directed our attention to a shortlist of characteristic genes per subgroup. These were previously determined by our collaborators based on their exclusive gene expression or significant overexpression<sup>73</sup> compared to all other subgroups. Here, we reviewed whether these signature genes translate to the proteome-level, emphasizing their role within a respective subgroup.





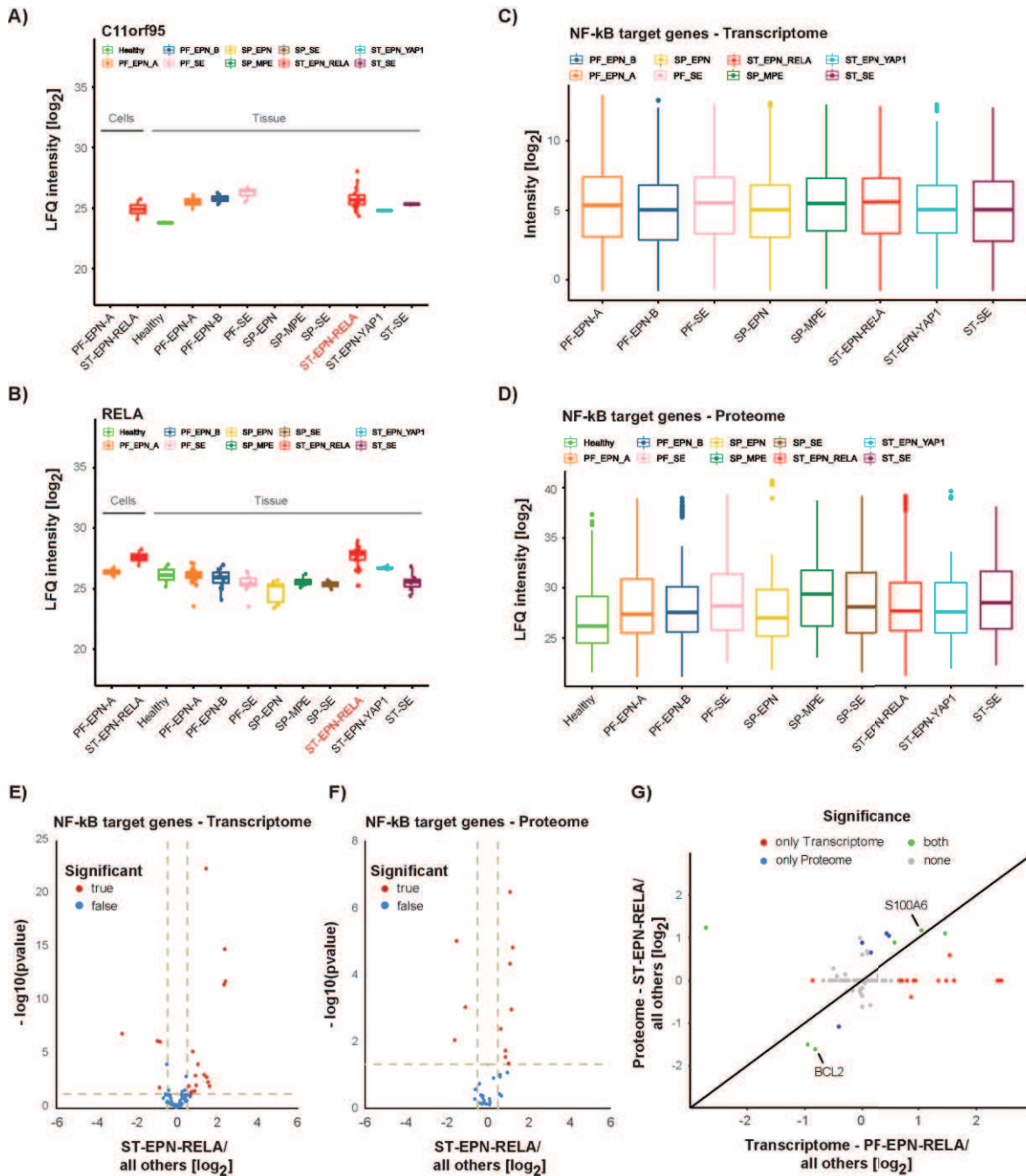
**Figure 15: Translation of signature genes to signature proteins.** A) Heatmap visualization of protein expression of subgroup-specific signature genes (Pajtler et al., 2015). B) Heatmap visualization of proteins that additionally show differential expression (DE) on the proteome-level, as determined by a Limma moderated t-statistics DE analysis. The significance threshold was set at a threshold of Benjamini-Hochberg adjusted p-value < 0.01 and an absolute abundance change of 2-fold. C-E) Boxplot illustration of NQO1 (C), ALDH1L1 (D), and FAM129A (E) protein expression across all EPN subgroups.

Therefore, we extracted the intensities for all identified and quantified proteins corresponding to the subgroup-specific signature genes. In total, 122 of 241 signature genes could be recovered on the proteome-level. The separation of subgroups based on protein intensities of these signature genes is notably less pronounced and suffers from a large proportion of missing values (**Figure 15A**). Here, imputation was used to allow visualization of the data. A large number of proteins do not differ in abundance compared to either the healthy reference or other subgroups. Using a Limma moderated t-statistics differential expression analysis, we could pinpoint those signature genes that translated to statistically significant signature proteins for the individual subgroups (**Figure 15B**). Here, the differential expression analyses were performed between each subgroup and all others without including the healthy reference for which we did not have gene expression data. This reduced the list to 49 signature proteins across eight out of nine subgroups, with no signature proteins for SP-EPN tumors (**Figure 15B**).

The majority of these signature proteins also varied in their expression compared to the healthy reference and showed implications in other cancer entities. For example, quinone oxidoreductase 1 (NQO1) was overexpressed in PF-SE (logFC= 3.11) and linked to reprogramming of glycolysis, proliferation, and metastasis (**Figure 15C**). The downstream effectors and NQO1 itself were suggested as promising therapeutic targets to prevent tumor progression<sup>352</sup>. Another interesting observation was ALDH1L1, a cytosolic dehydrogenase, that was highly overexpressed in PF-EPN-A tumors (logFC= 5.34 compared to all other tumors and logFC= 9.55 compared to healthy tissue) (**Figure 15D**). It is involved in folic acid metabolism and ATP production<sup>353</sup>. A knockdown of ALDH1L1 has previously been shown to reduce the production of ATP by 60% in NSCLC. The PF-EPN-A cell line also exhibits the overexpression of ALDH1L1 and could be utilized for a follow-up experiment using an ALDH inhibitor (e.g., gossypol) to reduce ATP production. In ST-EPN-YAP tumors, the expression of ABLIM1 (logFC= 3.25), FAM84B (logFC= 3.88), SNTB1 (logFC= 5.76), LMCD1 (logFC= 2.36), and FAM129A (logFC= 3.11) were clearly distinguishing from all other

## Results

tumors and healthy. Among them, FAM129A is involved in inhibition of apoptosis and promotion of migration and proliferation in human cancers (**Figure 15E**)<sup>354</sup>.



**Figure 16: Characteristic fusion protein involving C11orf95 and RELA drive oncogenic activation of NF-kB signaling.** A-B) Boxplot illustration of C11orf95 (A) and RELA (B) protein expression across all EPN subgroups. C-D) Global expression of NF-kB target genes per subgroup on the transcriptome-level (C) and proteome-level (D). E-F) Differential expression analysis using Limma moderated t-statistics for the comparison of NF-kB target gene expression (E) and protein expression (F) in ST-EPN-RELA tumors against all others. Genes (E) or proteins (F) passing significance thresholds of  $-\log_{10} \text{ p-value} < 0.05$  (Benjamini-Hochberg adjusted) and an absolute 2-fold change are highlighted. I) Correlation analysis of significantly changing NF-kB targets on gene and protein-level to identify specific or common effects. The transcriptome data were provided by our collaborators from Pajtler et al., 2015.

ST-EPN-RELA tumors are characterized by a recurrent fusion protein involving the uncharacterized C11orf95 and the signature protein RELA (logFC= 1.87) (**Figure 16A** and **Figure 16B**)<sup>259</sup>. The fusion has been linked to driving oncogenic activation of NF- $\kappa$ B-signaling in these tumors. RELA is the principle effector of this signaling pathway and was found among the list of signature proteins in our dataset. Similar to the analysis of PRC2 target genes in PF-EPN-A tumors, we investigated the expression levels for a list of NF- $\kappa$ B-signaling-related target genes<sup>355</sup>. Neither gene nor proteome profiles revealed statistically significant differences between ST-EPN-RELA and all other tumors (**Figure 16C** and **Figure 16D**). For the gene expression data, we did not have access to healthy reference data or the SP-SE subgroup. Interestingly, it seemed that NF- $\kappa$ B target genes are marginally upregulated on the proteome level as compared to healthy (**Figure 16D**). Using a Limma moderated t-statistics differential expression analysis, we could find a subset of NF- $\kappa$ B target genes that are regulated in ST-EPN-RELA tumors compared to all others on the transcriptome (20/83) (**Figure 16E**) and proteome (11/31 identified) (**Figure 16F**) level. A subset of NF- $\kappa$ B targets showed a significant regulation exclusively in the protein data (n= 5) or in both omics-layers (n= 6) or (**Figure 16G**). For example, a known regulator of apoptosis (BCL2) was identified among them. Another protein with strong cancer-related implications is S100A6, which is involved in the regulation of proliferation, invasion, migration, and angiogenesis<sup>356</sup>. The NF- $\kappa$ B complex subunit - NFKB1 was exclusively regulated on the protein level.

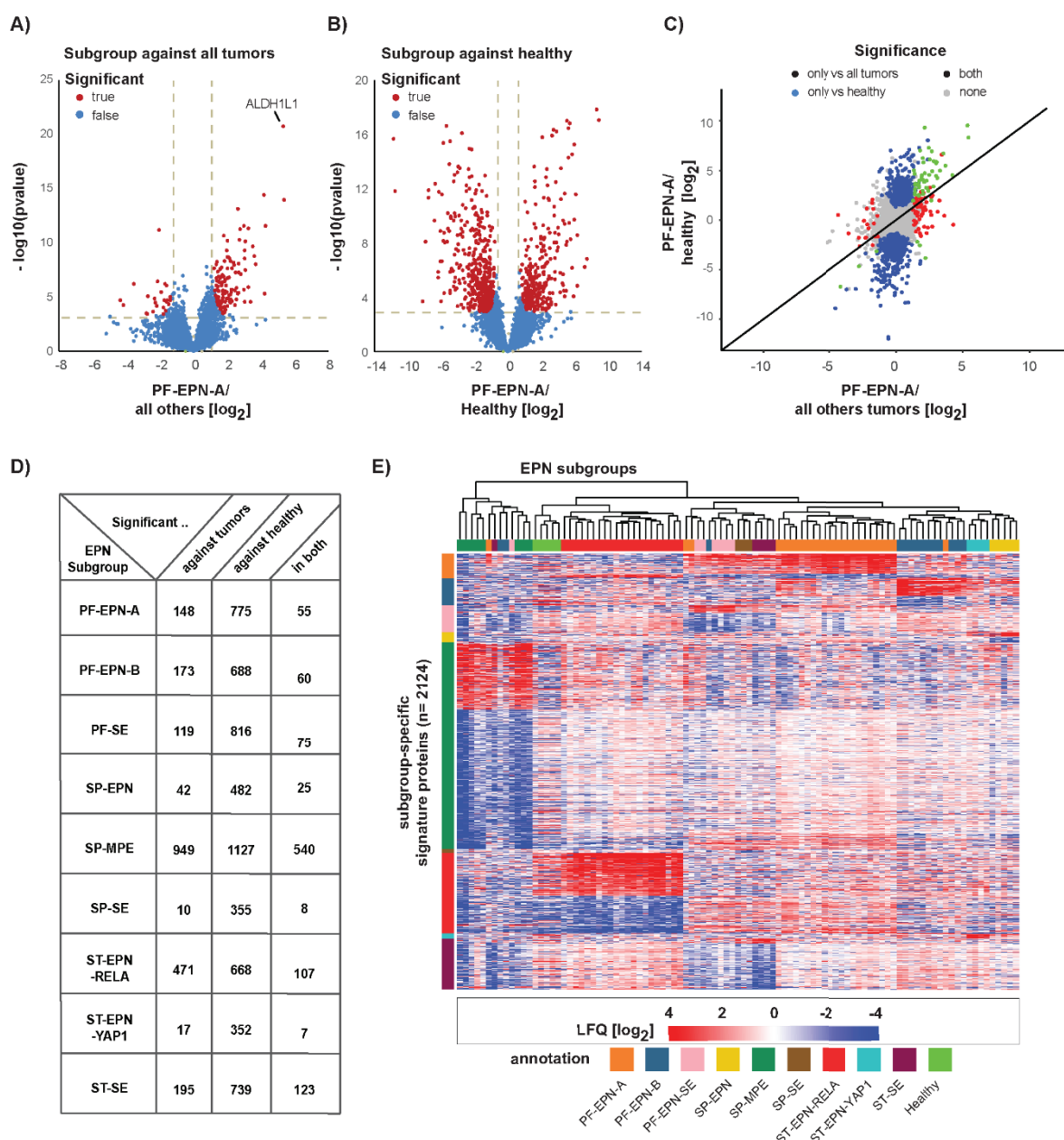
From the preceding analysis, we could identify several proteins (n =49) that indeed follow their subgroup-specific signature gene expression profiles. They are capable of delineating the different tumor subgroups and deserve a closer investigation in the following chapter. A large proportion of signature genes was either not identified on the proteome level (~50%) or did not translate to a signature protein (~40%). In addition, we could show the impact of abnormal NF- $\kappa$ B activation between different tumor subgroups on the proteome level. We continued with a protein-centric analysis to include all of the above and to additionally identify previously unknown protein signatures.

#### 4.3.3.2. Differential expression determines signature proteins

Hitherto assessment of our data was primarily coined by attempting to learn from the gene expression profiles. In the following section, we solely focused on a proteome-centric

## Results

approach, whereas the subgroup annotation is based on the DNA methylome. We performed Limma moderated t-statistics differential expression analysis to compare the proteome profiles of each subgroup, *firstly* against all other subgroups (exemplary **Figure 17A**), and *secondly* against the healthy reference samples (exemplary **Figure 17B**). The significance threshold was defined at a 2-fold change in absolute abundance and a BH-adjusted p-value lower than 0.01. Besides, we required each protein to be quantified at least twice within the compared groups. A correlation analysis revealed whether proteins were only significantly regulated against other tumors, healthy, or both (exemplary **Figure 17C**).



**Figure 17: Subgroup-specific differential expression (DE) analysis reveals signature proteins.** A-B) Exemplary, DE analysis using Limma moderated t-statistics for the comparison of protein expression in each ependymoma (EPN) subgroup against all other subgroups (A) and against the healthy reference (B). Significant proteins, at a threshold of Benjamini-Hochberg adjusted p-value < 0.01, and an absolute abundance change of 2-fold, are highlighted (red). C) Exemplary, correlation analysis of significantly changing proteins against the healthy reference and/or against all other tumor subgroups. D) Summary of DE analysis results for each EPN subgroup. E) Hierarchical clustering analysis using signature protein intensities relative to the mean expression in all other tumors.

Following from this, we identified several novel signature proteins for each tumor subgroup (**Figure 17D**) additionally to those that correlate with signature gene expression (**Figure 15B**). Using signature protein intensities relative to the mean expression in all other tumors for a hierarchical clustering analysis (**Figure 17E**), we could nearly perfectly recapitulate the different EPN subgroups. It follows that we have identified a large number of signature proteins based on our proteome profiling that previously remained undetected. They provide an expansive view of the underlying EPN biology, pathway activation, and potential subgroup-specific therapeutic targets. In order to get a better insight into subgroup-specific differences, we performed GO annotation and GSEA for each comparison. For example, extracellular matrix structural constituents were enriched in ST-EPN-RELA tumors. This is outlined in more detail in chapter 4.3.3.4. Further, the localization of proteins to the plasma membrane or as integral components of the plasma membrane was identified in PF-EPN-A. In both tumors, this might be a common theme towards their higher aggressiveness compared to all others (see also chapter 4.3.3.4). The most stable genomes are observed in SE tumors across all anatomical regions. Interestingly, mRNA splicing via the spliceosome was observed as the only significant term in both PF-SE and ST-SE tumors, compared to all others.

Next, we investigated the top 10 differentially regulated markers per tumor subgroup (**Table 2**), as determined by a differential expression analysis (exemplary **Figure 17A**). The significance threshold was set to an absolute fold change higher than 2-fold and an adjusted p-value < 0.01. Importantly, the top 10 markers were still sufficient to separate the majority of subgroups (**Supplementary Figure 8A** and **Supplementary Figure 8B**). We outline a few interesting hits with potential therapeutic implications. Quinolinic acid phosphoribosyltransferase (QPRT) shows a high expression in human malignant gliomas as well as in PF-EPN-A and PF-SE tumors in our global proteome data (**Supplementary Figure 8C**)<sup>357</sup>. It is involved in utilizing quinolinic acid, which is produced by microglia cells, for NAD(+) synthesis, and its high levels have been linked to increased resistance to oxidative

## Results

stress upon radio-chemotherapy and overall malignancy. QPRT, alkylating agents, or direct NAD(+) synthesis inhibitors have been proposed as therapeutic approaches for gliomas<sup>357</sup>.

Subgroups	PF_EPN_A	PF_EPN_B	PF_SE	SP_EPN	SP_MPE	SP_SE	ST_EPN_RELA	ST_EPN_YAP	ST_SE
Top 10 Signature Proteins	ALDH1L1	TUBB3	FABP7	HSPA6	CTNNA2	ACAN	FBP2	GNAI1	DHX15
	IGSF1	GPR50	HMG1	TNC	PRKCA	CTNNA3	ZNF185	SYNPO	NUCB1
	CPNE6	AGR3	SYVN1	NCAN	SLC3A2	RBP1	IFI30	KIF13A	NUCB2
	PYGM	GAP43	TP53BP1	GPD1	POSTN	TCEA1	GPSM1	FAM84B	RNPS1
	MYLK	CNRIP1	CORO2B	DOCK6	CORO2B	TPRKB	PCP4L1	PDLIM7	SNRNP200
	SNTA1	CYB5R1	SEC16A	NT5E	PTPN11	DPM1	PLCB3	MEIS2	POFUT1
	C1orf198	ALDH3A1	ANAPC1	PPAT	CTNNA1	SART1	PYCR1	PTDSS2	TRIM28
	PPIF	PTBP3	CMAS	COL6A6	XPNPEP1	NUP85	VAV2	ABLIM1	AGRN
	QPRT	MAPT	SERF2	CHUK	AQP4	COPS3	MICALL2	GAS7	PITRM1
	LIN7A	LMOD1	SARNP	ITGA11	CTNND1	SPAG9	CSPG4	UGP2	HNRNPA0

**Table 2: Summary of top 10 significantly differential abundant proteins per Ependymoma molecular subgroup** The top 10 signature proteins per EPN subgroup were determined using a differential expression analysis facilitated by Limma moderated t-statistics. The significance threshold was set to Benjamini-Hochberg adjusted p-value < 0.01 and an absolute abundance change of 2-fold. The subset of proteins corresponds to **Figure 17** and **Supplementary Figure 8**.

Furthermore, we identify an almost exclusive expression of the GPR50 receptor, a member of G protein-coupled receptors (GPCRs), in PF-EPN-A and PF-EPN-B tumors (**Supplementary Figure 8D**)<sup>358</sup>. Of note is that the expression levels are significantly different between both with an average of ~21 [ $\log_2$ ] and ~29 [ $\log_2$ ] in PF-EPN-A and PF-EPN-B, respectively. While GPR50 has previously been associated with ERK signaling, it might also serve as a potential biomarker to differentiate between both EPN subgroups<sup>359</sup>. Another protein, CYB5R1, a NADH-cytochrome b5 reductase that is involved in desaturation and elongation of fatty acids, is upregulated in all tumors except SP-MPE (**Supplementary Figure 8E**). Its highest expression is observed in PF-EPN-B tumors. In a previous study, it was identified as a potential therapeutic target against the development of glioblastoma by systematic genome-wide expression analysis. Using real-time PCR, the authors confirmed CYB5R1 as targetable by the demethylation drug 5'-aza-desoxycytidin<sup>360</sup>. A potential implication in EPN tumors could be investigated further. Furthermore, we found an exclusive upregulation of the exto-5'-nucleotidase (NT5E) in SP-EPN tumors. It has been shown that its blockage can facilitate the suppression of self-renewal, tumor growth, and progression in gliomas. This can be achieved via microRNA-30a treatment, which may present an interesting therapeutic strategy.

Altogether, we could identify novel signature proteins per tumor subgroup that were previously unknown by using our proteome-centric differential expression analysis. Among the top 10 markers per subgroup, we find a number of potentially actionable targets,

partially with implications related to gliomas or other cancer entities. Here, we highlighted a few representative examples to illustrate the added value of proteome profiling on top of or complementary to preceding -omics strategies.

#### 4.3.3.3. Genetic structural aberrations (CNVs) to phenotype

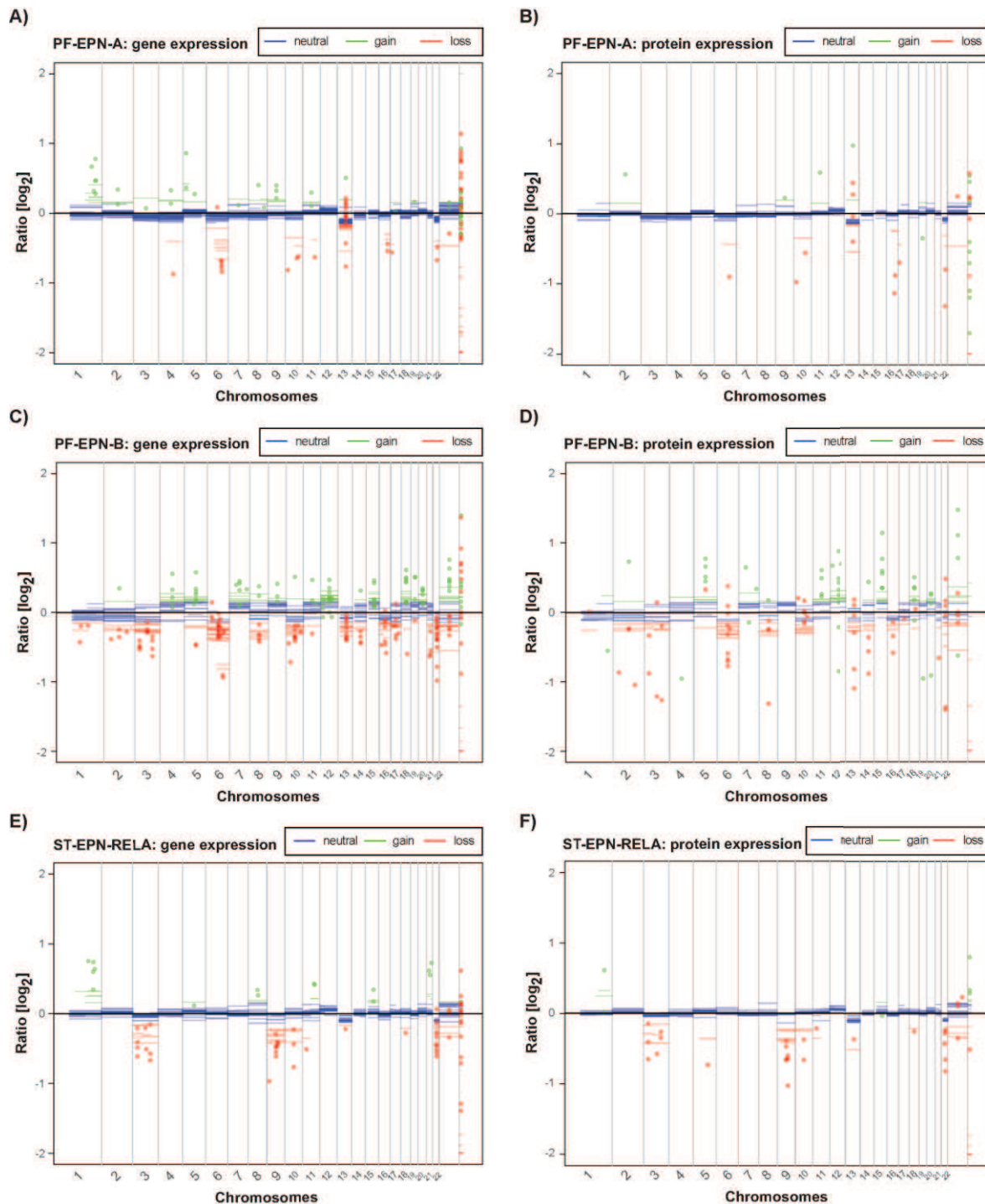
The paucity of recurrent mutations is a common characteristic of many childhood malignancies<sup>361</sup>, such as medulloblastoma<sup>362</sup>, retinoblastoma<sup>363</sup>, glioblastoma<sup>364</sup>, ATRTs<sup>365</sup>, neuroblastoma<sup>366</sup>, and also ependymoma<sup>367</sup>. A limited number of recurrently mutated genes are known for EPNs<sup>259,260,277,367</sup>. This includes a common deletion of CDKN2A in ST-EPN-RELA or a mutation of the NF2 gene in spinal tumors, for example<sup>368</sup>. The latter shows implications in restricting proliferation and promoting angiogenesis, while CDKN2A might play a role in the ST-EPN-RELA characteristic chromothripsis of chromosome 11. In our proteome data, we rarely quantified CDKN2A and found no differential expression of NF2. The majority of EPNs instead suffer from recurrent structural aberrations, including copy number variations (CNVs)<sup>367</sup>. Gains and losses of entire chromosomal arms are frequently observed, but their role and impact remain largely unknown. Here, we utilized our collaborators DNA methylation array-based CNV data to investigate the impact of recurrent structural aberrations on the proteome level. A detailed description of the following analysis is provided in the respective method section. Briefly, we compared the gene and protein expression for every available sample in relation to the observed CNV per sample and chromosomal region. For each sample and all its chromosomes, a line represents its CNV status as neutral (blue), deletion (red), or amplification (green) (exemplary **Figure 18A to 18F**). The mean relative expression values for genes or proteins are highlighted as dots at their genomic locus compared to the expression in CNV neutral samples. This illustration allowed us to obtain a global view of CNV impact on both expression levels for each tumor subgroup.

The gain of chromosome arm (chr) 1q is the most frequently observed CNV in PF-EPN-A. It has shown implications as an important prognostic factor within this tumor subgroup as it correlates with differences in overall survival. The same observation did not hold for PF-EPN-B and ST-EPN-RELA tumors, albeit they have the chr 1q gain in 18% and 24% of all cases, respectively<sup>73</sup>. The effect was restricted to an increased expression for a few genes, whereas the proteome seemed mostly unaffected (**Figure 18A to 18F**). The most



## Results

substantial genomic instability is observed in PF-EPN-B (**Figure 18C** and **Figure 18D**) and ST-EPN-RELA tumors (**Figure 18E** and **Figure 18F**). Especially PF-EPN-B are characterized by a number of aneuploidy events, including monosomy of chr 6 (61%), chr 10 (38.7%), and chr 17 (33.5%), as well as trisomy of chr 5 (31%), chr 8 (23.5%), and chr 18 (51.9%)<sup>73,260</sup>.



**Figure 18: Gene- and protein expression following recurrent structural aberrations.** A-F Exemplary, an illustration of ependymoma (EPN) copy number variation (CNV) and the corresponding gene- and protein-expression for PF-EPN-A (A-B), PF-EPN-B (C-D), and ST-EPN-RELA (E-F). The CNV status per chromosome and subgroup is indicated as neutral (blue), deletion (red), or amplification (green). The mean relative expression



values to the expression in CNV neutral samples are highlighted for genes or proteins as dots at their corresponding genomic locus. The CNV and transcriptome data were provided by our collaborators from Pajtler et al., 2015. Dr. Mathias Kalxdorf generated the CNV plots.

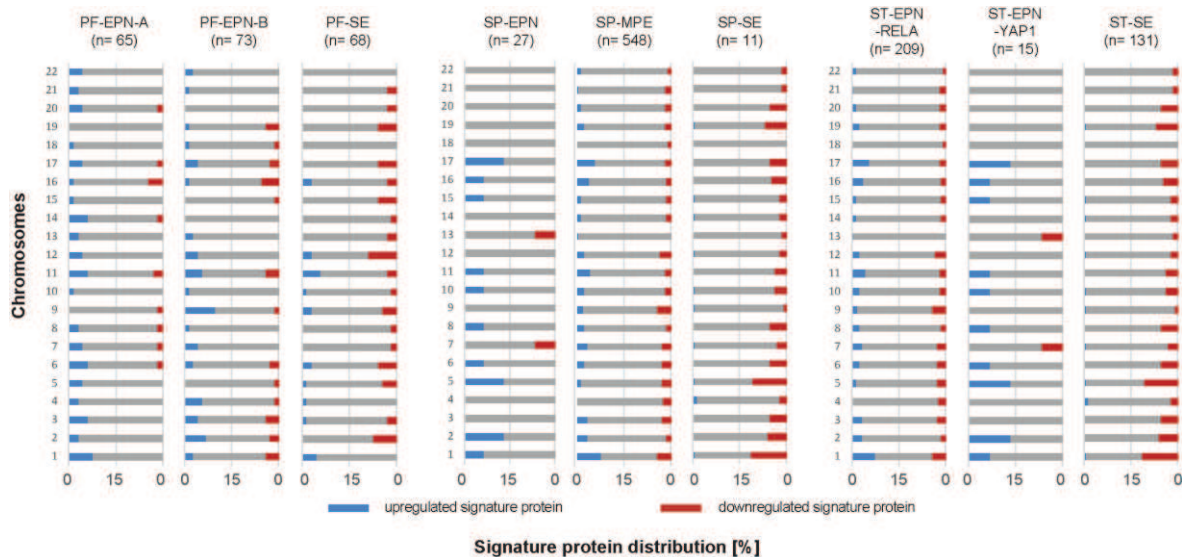
Besides the amplification of chr 8, the majority of these changes seem to translate to the gene expression level. The proteome is less affected with the vast majority of proteins not showing a significantly different abundance. However, few trends can be observed on the proteome level, including a decrease in chr 6 and chr 10, or amplification in chr 5. Further common loss of chr 6 and chr 13q in PF-EPN-B can be more or less observed on both expression levels. The latter has been proposed as a novel marker for PF-EPN-B tumors. Based on our global proteome profiles, we identified two signature proteins, MIPEP and NBEA, located on chromosome 13.

Across all three anatomical regions, the SE tumors exhibit the most stable genomes (data not shown). Neither CNVs nor the expression of genes or proteins seemed to show apparent global trends. For SP-SE tumors, we did not have access to gene expression data. Similarly, ST-EPN-YAP1 tumors are characterized by no obvious changes and they exhibit a mostly stable genome. The most obvious effect on the proteome level of ST-EPN-RELA tumors was the frequent deletion of chr 9 and chr 3 (**Figure 18E** and **Figure 18F**). About 90% of spinal tumors are characterized by loss of chr 22q, which carries the NF2 gene that is frequently mutated in these. This is obvious on the gene expression level for SP-EPNs but does not translate significantly to the protein level, including the NF2 expression. In SP-MPE tumors, chr 9 and chr 18 have a marginal trend following the CNV status. However, as for all other tumors, the vast majority of proteins did not seem to reflect the CNV state in the corresponding subgroup.

Despite that we see a few examples of genes and proteins that follow the global direction of the recurrent structural aberrations, the majority do not. This alludes to the fact that the recurrent structural aberrations do not drive the tumor phenotype alone. Therefore, we continued by matching our subgroup-specific signature proteins, which by definition includes all significantly over- or underexpressed proteins, to their corresponding chr locus to evaluate whether a proportion of them might be explained by CNV patterns (**Figure 19**). The majority of signature proteins seemed rather randomly distributed across the genome. The SE tumors in all compartments were characterized mostly by negative protein fold changes compared to all other tumors. Overall, these findings support the hypothesis that

## Results

no recurrent structural event (for example, SNPs) is the oncogenic driver but rather epigenetic mechanisms because chromosome gains and losses do not translate significantly to the proteome.



**Figure 19: Distribution of signature proteins per subgroup on chromosomes.** Illustration of signature proteins, as determined by a differential expression (DE) analysis, at their genomic locus per subgroup. Proteins with a higher (blue) or lower (red) fold change compared to all other subgroups are highlighted. The blue or red bars are relative to the total number of signature proteins per respective subgroup, as indicated in brackets below subgroup titles.

The main characteristic of EPNs are recurrent structural aberrations. However, genome sequencing efforts have largely failed to identify significantly recurrent mutated genes and oncogenic drivers as a result. Here, we could show that the majority of these structural gains or losses have a limited impact on the global proteome phenotype. This might be explained by buffering mechanisms that could manifest as alterations in the protein synthesis rate or turnover to compensate for the structural gains or losses.

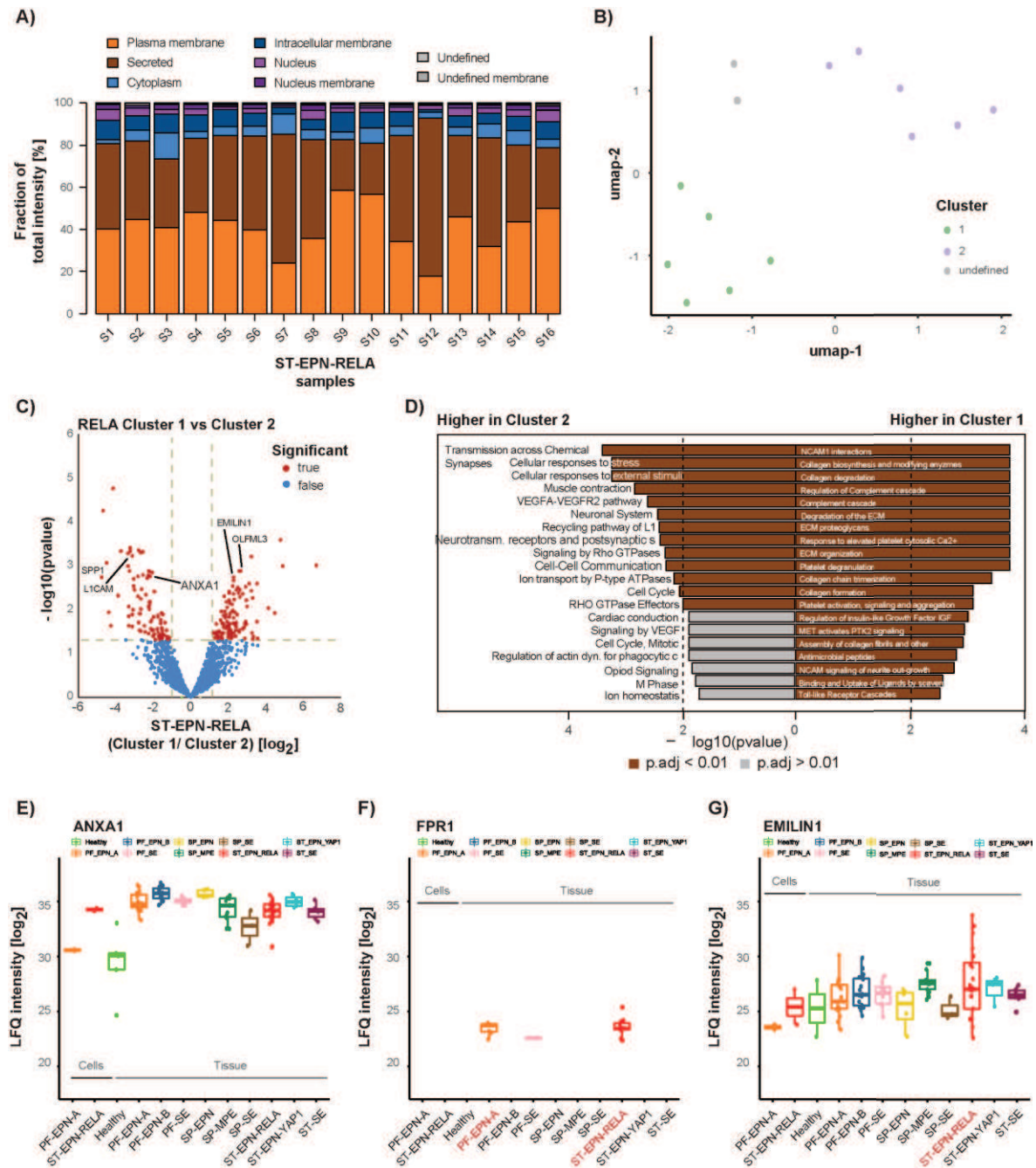
### 4.3.3.4. ST-EPN-RELA cell surface-proteome sub-classification

Plasma membrane and secreted proteins have a crucial role in a vast number of physiological processes<sup>300</sup>. This includes growth factor receptors and G-protein-coupled receptors (GPCRs) for signal transfer from external stimuli or adhesion proteins defining cell shape and motility. Other proteins are involved in the transport of nutrients, salts, or other molecules. They are involved in cell-cell communication and segregation of different tissues, for example. On the other hand, malfunctions in these protein classes are often involved in the acquisition of cancer hallmark traits during tumorigenesis. This can manifest as surface proteome changes that allow a cell to invade surrounding tissue, evade the

immune system response, or gain independence of survival signals. However, transmembrane proteins, especially GPCRs (>25%) and ion channels (>10%), represent the group of proteins that is most frequently targeted by drugs. Simultaneously, about 60% of all FDA-approved drugs target transmembrane proteins demonstrating their importance as potential therapeutic targets<sup>369,370</sup>. Here, we followed up on our previous observation of sub-subgroups in ST-EPN-RELA (2) and PF-EPN-A (3) tumors on the proteome level (see also chapter 4.3.1). Based on preliminary results, we hypothesized that ECM-related proteins might account for the sub-subgrouping, potentially resulting in different levels of aggressiveness and overall outcome. Therefore, we continued by performing a surface proteome enrichment for ST-EPN-RELA tumor samples (n= 16). Briefly, this was achieved by oxidizing carbohydrates of the resuspended cell-debris pellet, comprising plasma membrane fragments, and subsequent biotinylation of the cell-surface proteins. The labeled cell-surface proteins were captured using protease-resistant neutravidin-agarose beads. Reduction, alkylation, and digestion of proteins were performed on-beads, followed by peptide elution in an aqueous buffer and C18-based desalting to achieve LC-MS injection-ready samples.

Upon enrichment, we identified an average of 800 to 1000 plasma membrane proteins per sample. An average of about 80% of the total intensities per sample were attributed to plasma membrane or secreted proteins (**Figure 20A**). Focusing solely on surface proteins, we could identify two different clusters corresponding to our previous observation (**Figure 20B**). Two samples were manually assigned as a third cluster and disregarded in the following analysis, as they were not clearly associated with either cluster. Next, we *firstly* performed a differential expression analysis (**Figure 20C**) and *secondly*, a GSEA of the significantly regulated proteins (**Figure 20D**). In cluster one we found an increased expression of collagen degradation & synthesis, ECM degradation, ECM organization, complement cascade regulation proteins, MET activating PTK2 signaling, and promoting cell motility. The second cluster exhibited an enrichment of proteins involved in transmission across chemical synapses, cellular response to stress, the VEGFA pathway, Rho GTPase signaling, and cell-cell communication. This shows a broad involvement of ECM-related proteins and processes, as well as signaling processes in the sub-subgrouping of ST-EPN-RELA.

## Results



**Figure 20: ST-EPN-RELA cell surface-proteome sub-classification.** A) Illustration of surface-proteome enrichment results. Differentially colored bars represent the fraction of the total intensities originating from the different cellular regions. B) umap analysis showing at least two sub-subgroups of ST-EPN-RELA tumors based on the surface-proteome profiles. C) Differential expression (DE) analysis using Limma moderated t-statistics for the comparison of the surface proteome in ST-EPN-RELA cluster 1 against ST-EPN-RELA cluster 2. Proteins are significant at a threshold of  $-\log_{10}$  p-value  $< 0.05$  (Benjamini-Hochberg adjusted), and an absolute  $\log_2$  fold change of  $>1$  are highlighted. D) Gene set enrichment analysis (GSEA) of differentially regulated proteins comparing cluster 1 and cluster 2. E-G) Boxplot illustration of ANXA1 (E), FPR1 (F), and EMILIN1 (G) protein expression across all EPN subgroups.

Next, we investigated the up- and downregulated proteins in the comparison of both sub-clusters. Since the majority of these proteins were not identified in the global tumor proteome, we could not always compare to the protein's overall expression. We highlight

a few interesting candidates per sub-cluster. For example, SPP1 (Osteopontin) was significantly enriched in cluster 2, and its mRNA expression levels in glioma have previously been reported to rank among the top 10 of all cancer cell lines<sup>371</sup>. Its increased expression in lower-grade gliomas was associated with poor survival. This is in line with the aggressive phenotype of ST-EPN-RELA and PF-EPN-A, which both show a higher expression of SPP1 compared to any other subgroup in our global proteome data. Other proteins enriched in cluster 2 similarly link to poor survival (EZR, ezrin)<sup>372</sup> or cancer cell invasion and aggressiveness (ANXA1, a Ca(2+)-binding protein)<sup>373</sup> (**Figure 20E**). The latter correlates to hypoxia conditions and is highly expressed in various types of malignant tumors, including all EPN subgroups besides PF-SE and healthy tissue. It likely acts via secretion and autocrine signaling to promote aggressiveness and survival. The knockdown of ANXA1 has been shown to mitigate this aggressive phenotype in NSCLC cells<sup>373</sup>. While ANXA1 cannot be directly addressed as a potential therapeutic target, the inhibition of its receptor, FPR1, has been demonstrated to decrease tumor growth and metastasis formation in breast cancer<sup>374</sup>. This is particularly interesting as we see FPR1 exclusively expressed in ST-EPN-RELA and PF-EPN-A tumors in our global proteome data (**Figure 20F**). The inhibition of FPR1 can be efficiently achieved by the immunosuppressive drug Cyclosporin A (CsA), presenting an interesting follow-up experiment.

On the other hand, we also found several potentially interesting proteins enriched in cluster 1. This is, for example, the choline-specific glycerophosphodiesterase (ENPP6) that is often highly expressed in developing oligodendrocytes<sup>375</sup>, or the elastic microfibril interface located protein (EMILIN1)<sup>376</sup> with a crucial role in the tissue microenvironment. It has been shown that lower levels of EMILIN1 facilitate tumor cell trafficking and metastasis, thus implicating its protective role in tumor growth and spread (**Figure 20G**). Here, we observed that it is expressed through most tumors, but exhibits a significantly high variability in PF-EPN-A, PF-EPN-B, and ST-EPN-RELA. Interestingly, the highest expression occurred in some of the ST-EPN-RELA tumors. Lastly, we also found an enrichment of OLFML3 (not detected in global proteome), a secreted scaffold protein with an essential role in early development. It has previously been proposed as a novel therapeutic target for glioblastoma<sup>377</sup>, and its depletion can reduce the intratumoral microglia density with overall survival benefits<sup>378</sup>.

## Results

In summary, utilizing a surface proteome enrichment in ST-EPN-RELA tumors confirmed a sub-subgroup driving impact of ECM-related proteins. We found a number of potentially interesting proteins that vary within ST-EPN-RELA tumors and show implications with aggressive phenotypes. Most interesting among them is ANXA1 and its receptor FPR1, which can be targeted with selective inhibitors to diminish the poor prognosis for ST-EPN-RELA and PF-EPN-A tumors.

### **4.3.4. Exosome cargo characterization in ST-EPN-RELA**

Vesicular trafficking has previously been associated with overall EPN biology<sup>379</sup>. Their cargo is typically a mixture of proteins, lipids, metabolites, and nucleic acids. Importantly, extracellular vesicles (EVs) can cross the blood-brain-barrier (BBB) and are reported in all biological fluids, such as blood or cerebrospinal fluid (CSF)<sup>167,380</sup>. Thus, they present an easily accessible and non-invasive source for biomarker identification to advance diagnostics and prognostics or provide insight into the tumor-specific biology. This could be especially relevant for ST-EPN-RELA and PF-EPN-A tumors that exhibit the worst prognosis and overall survival while lacking a promising therapeutic target. Here, we continued with the optimization of EV isolation from ST-EPN-RELA and PF-EPN-A cell culture supernatant and subsequent characterization of the EPN-related EV protein cargo. This was done in collaboration with Dr. Kendra Maaß, Mieke Roosen, Dr. Marcel Kool, and Dr. Kristian Pajtler.

Briefly, extracellular vesicles are comprised of two main classes, namely exosomes and ectosomes<sup>380</sup>. The latter are generated directly by pinching off a part of the plasma membrane and constitute microvesicles, for example. They have an average diameter of 500 nm. On the other hand, exosomes derive from the endosomal pathway and have an average diameter of 100 nm. Through the formation of early- and late endosomes, they eventually generate so-called, multivesicular bodies (MVs). They are able to either fuse with the lysosome or autophagosome for degradation or with the plasma membrane to release the contained exosomes into the extracellular space. Almost all cells in a biological system release EVs as part of their physiological behavior. The removal of unnecessary constituents from a cell to maintain and regulate homeostasis was thought to be their primary function. However, context-dependent and mechanism-driven accumulation of EV cargo, especially in exosomes, adds to the growing evidence that they are involved in

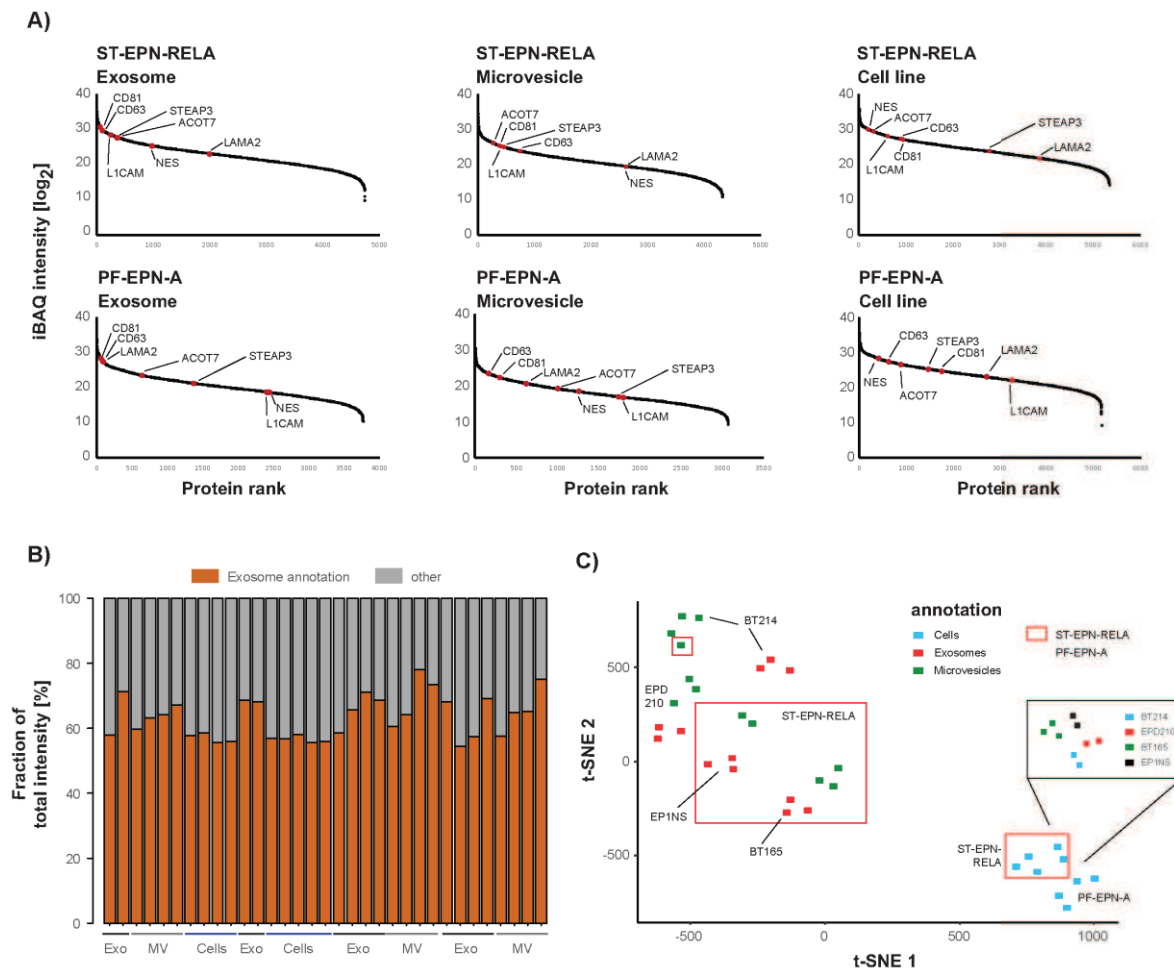
intercellular communication<sup>381,382</sup>. Numerous studies have linked EVs and exosomes in particular to CNS-related diseases, repression of the immune response, cancer progression (proliferation, angiogenesis, and tumor spreading), and other implications<sup>383–387</sup>. Further, it has been shown that the cargo of exosomes can alter the biological response of a target cell in both the close surrounding (paracrine) or distantly located areas (endocrine)<sup>381</sup>.

Here, we utilized all available EPN patient-derived cell lines, representing ST-EPN-RELA and PF-EPN-A tumors. Hereinafter they are referred to as BT165 & EP1NS for ST-EPN-RELA and BT214 & EPD210 for PF-EPN-A. We performed a global proteome profiling of each cell line together with a characterization of the proteomic cargo of isolated microvesicles and exosomes. The data obtained from different cell lines per tumor were combined for the majority of analyses. The isolation was performed from cells supernatant by successive ultracentrifugation and retrieval of fractions for microvesicles (10.000 x g pellet) and exosomes (100.000 x g pellet). The initial optimizations steps, including serum depletion, different coatings for the ultracentrifugation tubes, alternative precipitation methods, and additional purification using a size exclusion column, are not outlined in detail. The final procedure is explained in the corresponding method section. The isolation quality control was performed by our collaborators using immunogold electron microscopy, Qubit protein quantification, and nanoparticle tracking analysis (NTA). The latter showed a peak concentration of particle diameter in range with exosomes, whereas the microvesicle fraction likely presents a rather mixed population of vesicles (data not shown). Furthermore, EVs were specifically gold-labeled using  $\alpha$ -CD63 antibodies to allow their visualization in a transmission electron microscope (TEM) using CD63 as a selective exosome marker (data not shown).

CD63 and CD81 are considered hallmarks of exosomes and significantly enriched compared to microvesicles and global cell line profiles in our proteome data (**Figure 21A**)<sup>380</sup>. Another common marker for exosomes, namely CD24, could not be identified despite its known high expression in EPN tumors. This was due to its overall short sequence with only a single tryptic peptide and our two peptide per protein filter threshold during data analysis. The proportion of identified protein intensities that were annotated according to the GO: cellular component term for extracellular exosomes (ID: GO:0070062) did not show enrichment for the exosome fractions (**Figure 21B**). This was even the case compared to

## Results

the full cell lysates. The loose specificity of exosome annotation databases is a known problem in the field and highlights the necessity of accurate vesicle isolation and quality control workflows. Besides the enrichment of CD63 and CD81, several endoplasmic reticulum (ER)-resident (luminal) proteins were found in the dataset, resulting from the trans-Golgi network and ER contribution in the biogenesis of exosomes and their cargo<sup>382</sup>. On top of our optimized sample preparation, the above results indicate a reliable enrichment of the exosome fraction according to current guidelines of the international society of extracellular vesicles<sup>388</sup>. Using our proteome data, we found an excellent separation of both vesicle types, all cell lines, and the respective tumor subtype (**Figure 21C**).



**Figure 21: Ependymoma (EPN) extracellular vesicle (EV) cargo characterization in ST-EPN-RELA and PF-EPN-A cell lines.** A) Intensity-based absolute quantification (iBAQ) rank distribution for ST-EPN-RELA (EP1NS and BT165) and PF-EPN-A (EPD210 and BT214) cell lines, and isolated exosomes and microvesicle fractions of both. Exosome markers CD63 and CD81, ST-EPN-RELA markers NES and L1CAM, as well as some interesting candidates (STEAP3, ACOT7, and LAMA2), are highlighted in the panel. B) Illustration of the proportion of identified protein intensities (brown) that correspond to the gene ontology (GO)-cellular component (CC) of extracellular exosomes. C) t-distributed stochastic neighbor embedding (t-SNE) analysis of EPN isolated



extracellular vesicles (exosomes and microvesicle) and cell lines (ST-EPN-RELA and PF-EPN-A). Dr. Kendra Maaß and Mieke Roosen performed the isolation and quality control of extracellular vesicles.

Next, we systematically compared vesicle and cell line data to our extended annotation (based on gene expression and extended by proteomics) of subgroup-specific signature proteins. We chose this targeted strategy to circumvent a lacking control cell line, such as astrocytes, which will be added in future experiments to allow a more comprehensive analysis of differentially regulated proteins and pathway enrichment analysis (e.g., GO and GSEA). Until then, we utilized our global proteome data as a guideline. In total, we could identify and quantify 53 out of 69 (PF-EPN-A) and 166 out of 224 (ST-EPN-RELA) signature proteins. Among them, 6 (1) (PF-EPN-A) and 24 (12) (ST-EPN-RELA) were also significantly enriched in exosomes (microvesicles) compared to their cell line of origin (**Supplementary Figure 9A to 9G**). Only by comparing vesicle fractions across tumor subgroups we identified ACOT7, NES, and L1CAM to be enriched in ST-EPN-RELA exosomes (**Figure 21A** and **Supplementary Figure 10A to 10C**). Especially L1CAM and NES represent the tumor biology in the vesicle cargo as they were previously identified as ST-EPN-RELA biomarkers (**Supplementary Figure 10B to 10C**). The potential role of ACOT7, a cytosolic acyl coenzyme A thioester hydrolase, is less evident in an EPN-related context, but it has been suggested as crucial for the physiological brain function<sup>389</sup>.

Another observation was the metalloendopeptidase (STEAP3) upregulation in ST-EPN-RELA exosomes compared to both, its cell line of origin and PF-EPN-A exosomes (**Figure 21A** and **Supplementary Figure 11A**). It is known as a potential effector of the p53 pathway interfacing apoptosis and cell cycle progression<sup>390</sup>. Further, it is indirectly involved in facilitating the secretion of proteins. Interestingly, high expression levels of STEAP3 were previously shown in malignant gliomas, inversely correlating with prognosis and overall poor survival rates<sup>391,392</sup>. In glioma cells, a knockdown of STEAP3 (RNA knockdown) could attenuate aggressive phenotypes, such as cell proliferation and invasion<sup>391</sup>. This was in line with the expression levels of STEAP3 in ST-EPN-RELA exosomes, but neither in the corresponding cell lines nor in the tumor proteome profiles, showing a minus 3.5-fold change compared to all other subgroups (**Figure 21A** and **Supplementary Figure 11A**). This observation might qualify for future follow-up experiments.

Similarly, we identified significant enrichment of LAMA2 and FARP1 (**Figure 21A** and **Supplementary Figure 11B** and **11C**) in PF-EPN-A exosomes by comparing the respective

## Results

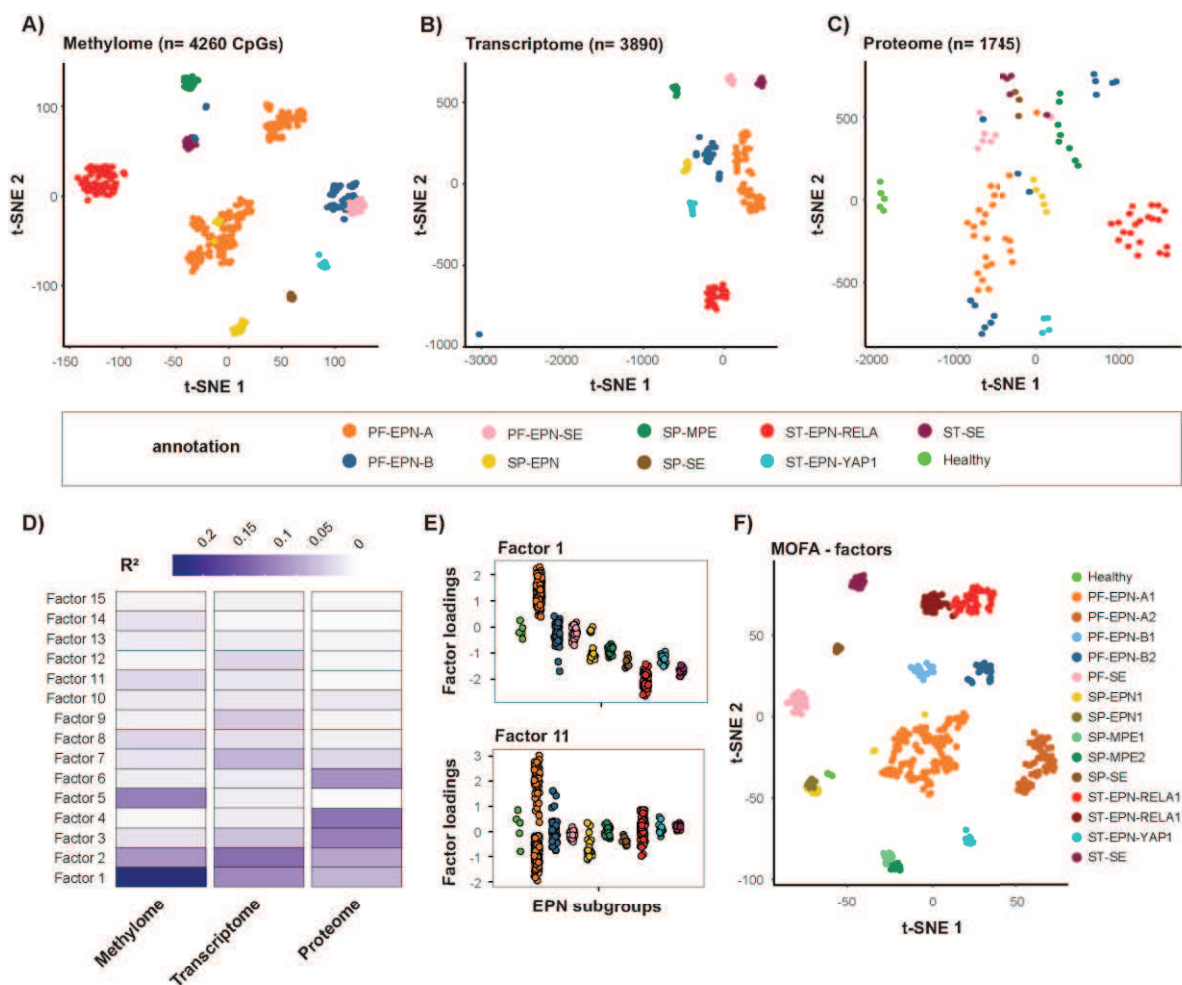
vesicle fractions across tumors or against its cell line of origin (**Supplementary Figure 9A to 9G**)<sup>393</sup>. FARP1 shows localization to the plasma membrane and implications in cell proliferation through MAPK signaling<sup>394</sup>, which was enriched in the tumor GSEA of PF-EPN-A compared to all other subgroups. Glioma initiating cell- (GIC) and oligodendrocyte progenitor cell differentiation were associated with laminin subunit alpha-2 (LAMA2), which has a role in mediating the attachment, migration, and organization of cells into tissues during development<sup>395</sup>.

In summary, together with our collaborators, we established and optimized an efficient exosome isolation protocol. Using our global proteome data as a guideline, we could pinpoint several interesting proteins that exhibit selective enrichment in tumor-specific exosomes. Some of which reflect the known tumor biology, for example, L1CAM. Others, such as STEAP3 or ACOT7, might be of interest for further follow-up studies. In functional assays, our collaborators already generated preliminary results (not part of this thesis) that indicate increased proliferation and migration in ST-EPN-RELA cell lines, endothelial cells, and microglial cells upon exposure to ST-EPN-RELA vesicles. Finally, we are anticipating to extend the current experimental setup to include astrocytes as a control cell line, supernatant from tumor tissue, and exosomes from patient-derived serum samples. This will enable a more comprehensive analysis independent of previous signature proteins.

### **4.3.5. Perspective view on multi-omics data integration**

In the previous chapters, we utilized each -omics layer separately or performed two-dimensional correlation analyses in order to pinpoint common or different features. This approach becomes increasingly challenging with the number and depth of complementary data layers. Identifying complex relations within and across multiple levels and several thousands of data points will require computational methods for the unsupervised integration and examination of these heterogeneous datasets. In this rapidly growing field, only a few approaches are currently available and still require a specialist for its realization in practice. One of these methods, namely MOFA, was developed by a collaborating group for the detection of technical and biological variation in comprehensive multi-omics data<sup>135</sup>. Here, we provide a preliminary perspective of this type of data analysis using the top 1% of most variable CpG probes (4260 CpGs) in the methylome data and the 20% most

variable features in the transcriptome (3890 transcripts) and proteome data (1745 proteins) (Figure 22A to 22C).



**Figure 22: Multi-omics factor analysis (MOFA) achieves higher resolved subgrouping.** A-C) t-distributed stochastic neighbor embedding (t-SNE) analysis of endependymoma (EPN) methylation patterns (n= 515, top 1% of CpG probes with the highest standard deviation) (A), the transcriptome gene expression (n= 135, top 20% of most variable features), and the proteome profiles (n= 103, top 20% of most variable features). D) Multi-omics factor analysis (MOFA) reveals low dimensional factors and their relevance per omics-layer. E) Exemplary, an illustration of genes/proteins that load onto factor 1 and 11 per EPN subgroup, highlighting the separation of PF-EPN-A from all others and in additional two sub-subgroups. F) t-SNE analysis of combined MOFA results, as factors 1 to 15, reveals a more detailed separation. The methyome and transcriptome data were provided by our collaborators from Pajtler et al., 2015. Dr. Mathias Kalxdorf performed the MOFA analysis.

In brief, similar to a principal component analysis (PCA), MOFA performs a dimensionality reduction to infer interpretable factors that describe sources of variation across all data layers. These factors may capture discrete or shared variation within and across the different multi-omics datasets (Figure 22D). The MOFA model explains ~60% of the variation in the methyome data and ~50% in both transcriptome and proteome data. Next, we can highlight whether a specific subgroup of tumors is described by one or multiple of

## Results

these factors, which can be further annotated using GSEA. For example, genes and proteins loading on factor 1 are contributing to the separation of PF-EPN-A and ST-EPN-RELA tumors from all others (**Figure 22E**). Furthermore, the sub-subgrouping of PF-EPN-A tumors is driven by genes and proteins that contribute to factor 11 (**Figure 22E**). The unsupervised identification and integration of all fifteen MOFA-determined factors results in a more detailed separation of the EPN tumors (**Figure 22F**). Neither of the individual -omics layers could identify all substructures independently. The methylome data were insufficient to separate PF-EPN-B and PF-SE, for example. Further subgrouping of SP-MPE, SP-EPN-RELA, PF-EPN-A, and SP-EPN tumors could only be detected in the integrated MOFA approach (**Figure 22F**). GSEA and GO annotation analysis for individual factors or between tumor subgroups (data not shown) recapitulate, for example, that PF-EPN-A and ST-EPN-RELA sub-subgroups are largely driven by ECM degradation, ECM organization, and ECM structural constituents. The heterogeneity of PF-EPN-B has been described previously<sup>264</sup>. For the majority of other EPN subgroups we had insufficient numbers of samples on either the transcriptome or proteome level.

In general, the MOFA analyses perfectly recapitulates and extends the molecular classification of EPNs in an unsupervised fashion. This type of analysis will be especially useful in disease entities with unknown substructures or for the identification of sub-subgroups that were previously not evident from the individual layers.

## 5. Discussion

The large-scale study of protein expression has not yet been implemented into clinical routine<sup>129,137</sup>. Reasons for this are comprised of logistical, ethical, and technical challenges, as outlined in chapter 1.2.3. In this thesis, we approached the technical aspect to ease the introduction of an automated and thus reproducible sample preparation pipeline for a variety of different and quantity-limited, clinical specimen. This was achieved by optimizing and transferring the single-pot, solid-phase-enhanced sample preparation (SP3) protocol onto a liquid handling platform for generic, reproducible, and parallelized proteomic sample preparation, while propagating all its benefits into the workflow. The pipeline resolves several bottlenecks that previously hindered the implementation of proteomics to complement other NGS profiling methods, which are likewise not yet fully integrated into routine clinical application<sup>1,27</sup>. The resulting end-to-end automated workflow (autoSP3) enables systematic proteome profiling for such routine applications, constituting an important step towards its implementation in a clinical or research environment.

### 5.1. Large-scale proteome profiling enabled by autoSP3

Proteomic sample preparation still largely depends on a number of consecutive manual handling and pipetting steps. This includes the lysis of a specimen, the reduction and alkylation of extracted proteins, the subsequent removal of contaminating buffers or salts, and the proteolytic digestion of proteins to peptides. Initially, we evaluated and optimized a series of steps and parameters of the manual SP3 protocol to achieve its maximal performance while aiming for high scalability for subsequent automation of the procedure. The sample lysis and protein extraction comprise the first essential step that is required for almost all types of samples (e.g., fresh-frozen tissue or FFPE tissue). Here, we tried to achieve a one-for-all solution to handle any type of sample, which has not been achieved by any other method. In addition, we aimed for a scalable solution that seamlessly integrates with a 96-well format to avoid a limiting factor early on within our workflow. Here, we failed to omit a mechanical sample disruption to aid the extraction of proteins for two reasons: I) the amount of extracted protein from tissue material was neither linear nor reproducible (**Figure 1B**), and II) the lack of a proper DNA and RNA shearing turned out to be incompatible with the subsequent SP3 procedure. This manifested as protein binding

## Discussion

interferences and reduced peptide recoveries as a result (**Supplementary Figure 1B**). At the same time, enzymatic cleavage of nucleic acids led to reduced protein extraction efficiencies, because of the accompanying reduction of detergent concentrations to allow enzyme activity. This would result in a reduced sensitivity of the workflow, which is a significant drawback, especially for quantity-limited samples, and could dramatically increase the overall costs of the workflow.

The highest protein yield and nucleic acid sheering efficiency was achieved using 4% SDS, as the main buffer constituent, in combination with AFA-based ultrasonication (**Figure 1C** and **1D**). Thus far, we demonstrated proof-of-concept for the multiplexed (**Figure 2D** and **2E**) and highly efficient lysis of cells, fresh-frozen tissue, and a manually de-paraffinized cohort of 51 lung ADC FFPE tissue samples in the presence of 1 or 4% SDS. Despite the good performance, we see the additional potential to improve the current processing settings further. For example, we aim to optimize sonication frequencies and amplitudes in combination with cycle times, cycle length, and lower sample volumes in order to minimize the time needed per sample and maximize the sensitivity for low-input applications. While it currently takes about one hour to process 96 cell- or tissue samples, we can likely reduce this to about 20 minutes. This will enhance the overall throughput and improve the accompanying turn-around times per sample, which will be important when performing this in a clinical environment. Further, we collected preliminary data to show a full integration of AFA-based processing of FFPE tissue without requiring the manual de-paraffinization. Interestingly, this was recently demonstrated in combination with SP3 in an application note by Lisa Schweizer et al., 2020, in collaboration with Covaris<sup>330</sup>. This is an essential step because the WHO classification of tumors and histology primarily rely on FFPE tissue, which is therefore routinely collected in biobanks over decades already<sup>159,168,396</sup>. Thus, this presents an immense resource of samples for retrospective proteome profiling, requiring a suitable method to enhance their accessibility. In addition, we aim to process body fluids (e.g., CSF and blood plasma/serum) to showcase that autoSP3 presents the first method for the preparation of all sample types and low-quantities. Undoubtedly, this marks a significant step towards standardized and routine proteomics applications.

In this thesis, the seamless integration of all steps into an end-to-end automated workflow, comprising multiplexed AFA-based ultrasonication and automated SP3 (autoSP3), was achieved. Peptides resulting from cells, fresh-frozen tissue, or FFPE material could be subjected to LC-MS without any further clean-up. Importantly, the SP3 method does not exhibit a bias towards hydrophobic peptides as the binding happens on the protein level, and digested peptides are released into the aqueous buffer irrelevant of their hydrophobicity (**Figure 3C**). Recently, Tanveer Batth et al., 2019, has proposed protein aggregation as an alternative binding mechanism of SP3<sup>397</sup>. The authors claim that insoluble proteins preferentially precipitate on microparticles (e.g., SP3 beads) irrespective of their surface chemistry. While this is already extensively discussed within the original SP3 publications and its patent application, we could observe apparent differences in the numbers of identified peptides when using different bead types and surface chemistries (**Supplementary Figure 2F**). At the same time, the authors claim that unmodified beads show similar performance, but do not show their data. Further, we find it highly unlikely that polar interactions can be easily reversed by changing to an aqueous buffer composition. We aim to show this in additional experiments. However, most importantly, SP3 (autoSP3) is capable of removing the most frequently used buffer components during sample lysis and protein extraction, adding to the high flexibility and efficiency of protein extraction from a variety of sample types (e.g., SDS facilitates FFPE processing<sup>159,168</sup>). This generally includes SDS or other non-ionic detergents, such as Triton X-100 or NP-40, and the anionic detergent sodium deoxycholate, which aid in extraction and solubilization of proteins including transmembrane proteins. Further, Urea-based buffers can be processed, and they recently gained popularity by leading to an increase in protein yields in specific applications<sup>398–400</sup>. Other acid-labile surfactants, such as RapiGest SF<sup>401</sup> or ProteaseMax SF<sup>402</sup>, are typically less potent for protein extraction and require, for example, precipitation and centrifugation or spin filter columns (e.g., FASP) for sample clean-up after its hydrolysis. The automation of such processes is less straightforward, requiring more sophisticated and expensive equipment on top of a liquid handler. Altogether, the procedure could alleviate many shortcomings that are associated with classical sample preparation protocols and manual handling, by benefiting from all the valuable features of SP3 and the nature of automation. Altogether, uniquely positioning autoSP3 as a building block for routine (clinical) proteomics.

## Discussion

Next, we demonstrated the excellent protein quantification reproducibility achieved with the autoSP3 procedure with a series of 60 HeLa samples, resulting in a median CV [%] of 16.3% over a period of one month. In addition, we achieved a median CV [%] values below 15% when including the entire sample preparation process from sample lysis and protein extraction to LC-MS injection-ready peptides (**Figure 15C**). The resulting advantage is that samples can be processed and measured over extended periods, for example, during longitudinal sample collection or time series, without introducing sample preparation variability. This is particularly important in a realistic clinical environment with ongoing patient enrolment and sample collection in irregular intervals. Indeed, the ADC cohort could showcase an almost perfect grouping of replicate tissue slices, based on their proteome composition and despite randomization, demonstrating consistency and robustness of the procedure. Furthermore, we demonstrated the high sensitivity of autoSP3, a key attribute for clinical workflows, by processing minute amounts of sample and quantifying consistent numbers of proteins, such as roughly 500 proteins from as little as 100 HeLa cells (**Figure 9B**). Especially here, a robust and reproducible sample processing is important to avoid any unnecessary technical variability that has the potential to mask the biological differences of interest. This asset will open up great opportunities for new applications in the routine analysis of rare cell types or overall quantity-limited sample material. For example, the sensitivity of autoSP3 might enable the analysis of small biopsies that were previously inaccessible for proteomics applications. On the other hand, the size of a biopsy could potentially be reduced to improve the tumor cell content or resolution and specificity of the analysis as a result. Here, we also see the potential to further improve the sensitivity of our workflow by reducing overall processing volumes. So far, this was hindered by the available 96-well magnet, which requires a certain digestion volume in order to cover the protein-binding SP3 beads. This could be combined with an upgrade of our current Bravo system to a Bravo 96ST pipetting head that allows reproducible transfer of 0.3 to 70 volumes  $\mu\text{L}$  (as compared to  $\sim 5 \mu\text{L}$  minimal volume).

The final autoSP3 workflow takes about 3.5 hours for the processing of 96 samples simultaneously. This includes all steps from cell- or tissue lysis ( $\sim$ one hour for ultrasonication prior to anticipated optimization) to proteolytic digestion (2.5 hours for autoSP3), and peptide recovery ( $\sim$ 7 minutes). The continuous and parallel operation of the



Covaris LE220R-plus ultrasonicator and the Bravo platform for autoSP3 permits the processing of up to three plates, corresponding to ~300 samples, by a single operator and within a working day. The hands-on time is kept minimal at the same time. High-throughput sample processing contributes to rapid turn-around times that are required for clinical decision making. For example, the NSCLC international guidelines for genetic analysis recommend a turn-around of less than ten days<sup>403</sup>. In other disease entities or clinical scenarios more or even less turn-around may be required or tolerable. The capacity of our autoSP3 setup could already comfortably accommodate very large-scale proteomic studies, and resulting peptide samples may feed into several mass spectrometers, which are currently the remaining limiting factor. On top, we see the additional potential to improve our current setup. The protein reduction and alkylation, for example, is performed at 95°C, which requires one hour (out of 2.5 hours) of autoSP3 for heating and cooling. Here, an increased throughput can be easily achieved by either optimizing the reduction and alkylation conditions at lower temperatures or using a more efficient temperature device. The use of a plate hotel or larger deck-space has the additional potential to increase the workflow capacity to further minimize the hands-on time between individual runs, for example. AutoSP3 could be further improved by preventing evaporation (remains unsolved so far) during the proteolytic digestion to integrate this step on-deck and avoid manual interference. Lastly, the LE220R-plus ultrasonicator is fully compatible with a robotic arm that could facilitate the sample plate transfer between platforms to minimize the need for operator intervention. The implementation of the steps mentioned above has the potential to transform the current protocol in a complete hands-free pipeline that could continuously process many hundreds of samples per day in a robust and reproducible manner.

AutoSP3 has been implemented on a Bravo liquid handling system, which is widely available to many genomics or biochemistry laboratories. The established workflows (Protocol A, B, C, and D described in chapter 4.2.1.) are provided in an online repository for the facile adoption of the method. In addition, we have recently generated methods that allow different starting volumes for samples (up to 25  $\mu$ L) or additional clean-up of peptides. The latter is achieved by binding peptides to the beads at a higher than 95% organic buffer composition. This part of the initial SP3 protocol for peptide purification<sup>149</sup> has recently been established by others on an Eppendorf liquid handling system<sup>404</sup>. While this is not

## Discussion

needed for global protein expression profiling, it can be useful for clean-up of PTM-modified peptides, for example. To further improve the ease of implementation and usage of autoSP3, we are currently working on a user-interface integrated into the Bravo Vworks software. This is done in collaboration with Dr. Mauro Cremonini (Agilent Technologies). Another potential extension of autoSP3 for top-down proteomics, comprises its use for intact protein purification and subsequent MS analysis, as recently shown<sup>329</sup>.

Because of the benefits over previous methods, the SP3 protocol has broad appeal in the field of proteomics. This includes efforts in automation, as recently shown on various platforms, such as a KingFisher liquid handling system with subsequent phosphopeptides enrichment<sup>405</sup>. Furthermore, a recent pre-print study shows the application of autoSP3 on a Hamilton Robotics Microlab STARlet liquid handling system for the fast and low-cost detection of SARS-CoV-2 peptides from clinical samples<sup>406</sup>. The study showcases a short turn-around time, high sample throughput, and cost-efficiency. In this case, the automation additionally reduced the risk of infection during sample preparation and marks an informative example for a clinical application of autoSP3.

### **5.2. The added value of proteome profiling**

The proteome composition is a fundamental part of any biological system and crucial for understanding (patho)physiological conditions or functions. Nevertheless, it remains mostly unused in a clinical routine, as outlined throughout this thesis. Therefore, after technically establishing the autoSP3 workflow<sup>304</sup>, we applied it to process two different clinical cohorts, namely FFPE lung ADC slices and fresh-frozen ependymoma (EPN) brain tumor tissue, in order to demonstrate the added value of global proteome profiling. Further, we highlight the novelty of a number of observations that other NGS methods are inherently blind to or that cannot be predicted from gene expression alone. The profiling of the proteome composition of a disease cohort has the potential to unravel unknown functional consequences or clinically relevant targets or biomarkers.

#### **5.2.1. Molecular characterization of lung adenocarcinoma (ADC) growth patterns**

In this thesis, we showcase the application of autoSP3 to a cohort of 51 ADC FFPE samples for the molecular characterization of tumor growth patterns. In particular, we could demonstrate the ability of FFPE processing with quantity-limited material. In addition, as

expected from pathology, we observed that lepidic samples show a decreased expression of proteins associated with cellular invasion. In comparison to previous microarray gene expression profiling<sup>217</sup>, we identified 167 proteins (compared to 13 genes) with statistically significant differential abundance in lepidic samples compared to all other growth patterns. This shows that the differential proteome composition cannot be predicted from mere gene expression changes alone. A follow-up for a potential implication of any of the 167 differentially regulated proteins (lepidic vs. all others), for their use as a therapeutic target or as a biomarker, might be of interest for further follow-up studies.

However, the high variability in tumor cellularity of the provided ADC samples restricts a more detailed analysis. In the future, we aim to first extend our current workflow to allow the automated de-paraffinization and protein extraction of FFPE samples in combination with sample processing in overall smaller volumes. This will aid the handling of even less starting material. Here, we aim to reach the level of microdissection to collect and process highly concise tumor areas with maximal tumor cell content. This will enable us to perform a more comprehensive analysis of the proteome composition of the different growth patterns and at a higher spatial resolution across the obtained specimen.

### **5.2.2. Proteome profiles extent ependymoma (EPN) molecular classification**

Furthermore, we utilized a cohort of EPN brain tumors with extensive molecular characterization available on various levels<sup>73</sup>. While this built the basis for classification into nine distinct molecular subgroups, the majority of them still lack insight into their functional differences, and the oncogenic driving mechanisms remain unknown (see also chapter 1 and chapter 4.3). Here, we aimed to illustrate the potential of proteome profiling on top of or complementary to other molecular layers.

Starting from small fresh-frozen tissue (<6 mg wet weight), we could quantitatively profile 8248 proteins from 103 tumors that were unevenly distributed over all nine molecular subgroups and the healthy reference tissue. Interestingly, we could achieve a similar separation of the known molecular subgroups based on the tumors' proteome composition (**Figure 12A-C** and **Figure 12F**). However, this annotation is still based on the prior knowledge of the anatomical region and DNA-methylome profiles. Relying on the proteome composition alone would be insufficient to achieve the same grouping without previous knowledge of subgroup-specific protein expression patterns or individual biomarkers. The

## Discussion

methylome and transcriptome result in a clearer separation (average silhouette score= 0.58) compared to the proteome (silhouette score= 0.43). On the other hand, the proteome composition revealed a sub-subgrouping for ST-EPN-RELA and PF-EPN-A, showing a first hint of its added value. In addition, we performed an unsupervised MOFA as a perspective view for scenarios in which molecular subgroups are not yet defined. Methylome, transcriptome, and proteome data together result in a perfect recapitulation of the expected subgroups and further achieve a more detailed sub-subgrouping on top (**Figure 22F**). This specifically highlights the advantage of complementary -omics profiling rather than focusing on individual layers. Multi-omics analysis has been a challenging task due to technical and especially bioinformatical reasons<sup>135,169</sup>. In recent years, however, various groups have invested massive efforts in multi-omics data analysis<sup>169</sup>. While this is still limited to skilled data scientists, the trend is going towards user-friendly solutions that will find rapid adoption in the field.

In this thesis, EPN subgroups were known already and the primary purpose was the supervised identification of functional differences, biomarkers, or even potential therapeutic targets. The uneven distribution of samples per subgroup presented a distinct challenge for the subsequent bioinformatic analysis, including differential expression comparisons. However, this characteristic is a realistic scenario of a disease cohort, in which the sample collection itself can be a limiting factor for rare subtypes. Additionally, in a routine clinical application for molecular profiling analysis, the underlying subtype would be unknown at the time of sample collection. Thus, we inevitably have to deal with inhomogeneous numbers of samples per subgroup. The highest number of samples was available for PF-EPN-A (n= 24) and the lowest for SP-SE (n= 3). Here, we handled this uneven representation of distinct subgroups by utilizing the Limma R/Bioconductor software package for statistical analysis. Benefitting from the large number of samples, Limma performs an analysis of protein quantification value distribution across the entire dataset as an integrated whole rather than focusing on individual comparisons between sample- or group pairs. This allows a more accurate comparison for even small numbers of samples within a subgroup (e.g., SP-SE) against the remaining dataset as a whole (all others). At this stage, the subgroup annotation remains a prerequisite before performing any comparison.

Next, we focused on subgroup-specific proteins that could serve as a biomarker when no prior knowledge on sub-grouping is available. On top, these proteins have the potential to elucidate oncogenic driving mechanisms within a specific tumor subgroup. Here, CXorf67 marks a perfect example, showing exclusive expression in PF-EPN-A tumors<sup>265,272</sup>. While its expression already marks a hallmark for PF-EPN-A tumors, the precise mechanism of CXorf67-mediated inhibition of the PRC2 function was previously unknown. The functional domain of CXorf67 was pinpointed to the C-terminal region, being responsible for binding to the majority of PRC2 components and the inhibition of its methyltransferase activity. We could additionally show the influence of PRC2 inhibition and the associated hypomethylation and de-repression of its targets at the transcriptome and proteome level compared to all other EPN subgroups. Surprisingly, we only found 12 and 15 PRC2 targets that were significantly regulated either on transcriptome and proteome level or solely on the proteome level (**Figure 13I**). Among them, several proteins, such as NCAM1<sup>407</sup> or GPM6B<sup>408</sup>, have previously been associated with brain tumors. Further functional implications of the de-regulation of PRC2 targets remain to be elucidated.

Interestingly, we found that both ST-EPN-RELA and PF-EPN-A show differential expression of ECM proteins between transcriptome and proteome level and all other subgroups while also being associated with higher disease aggressiveness and exhibiting the worst overall prognosis<sup>73</sup>. Furthermore, the proteome composition of both revealed an additional sub-subgrouping that is driven by ECM-related proteins (supported by DE analysis and MOFA). To follow-up on this observation, we performed a cell-surface proteome enrichment of ST-EPN-RELA tumors. Here, we could find several interesting proteins, with FPR1 linked to ANXA1 being the most promising potential target<sup>373,374</sup>. The inhibition of FPR1 has previously resulted in a decrease of tumor growth and metastasis formation<sup>374</sup>. This is particularly interesting as we observed an almost exclusive expression of FPR1 in both subgroups, which urgently call for new treatment plans due to their worst prognosis and poor outcome. Using cell culture experiments, a relevant implication remains to be elucidated, such as IC50 experiments upon treatment with an FPR1 inhibitor (e.g., Cyclosporin A<sup>374</sup> or Cyclosporin H<sup>409</sup>).

Unfortunately, patient-derived cell lines for follow-up experiments are only available for ST-EPN-RELA (2x) and PF-EPN-A (2x). They are rather difficult to handle (e.g., long cell

doubling times) and require specialized, expensive cell culture media. Thus, following up on several observations made throughout this thesis takes rather long and is limited to targets identified in ST-EPN-RELA and PF-EPN-A. Therefore, other subgroup-specific findings cannot be validated yet. However, a more in-depth analysis of the subgroup-specific proteins is still ongoing to derive functional insight into their specific biology. Many of these proteins have previously not been annotated as an EPN subgroup signature on the basis of other -omics layers. Importantly, we could showcase a few examples of signature proteins, such as L1CAM and NES (**Figure 21A**), that are enriched within isolated extracellular vesicles, reflecting the tumor-specific biology<sup>367</sup>. This marks the possibility of subgroup-specific biomarker profiling in a low-invasive manner through blood or CSF sampling. In a clinical environment, this profiling approach could find facile adoption.

Altogether, profiling of the EPN cohorts' proteome composition helped to unravel many previously unknown signature proteins, which were not evident based on the other NGS approaches. Further, as expected, many previously annotated signature genes did not translate to signature proteins and its resulting phenotype. This might influence their importance for the biological interpretation and their functional consequences between different subgroups. A more thorough analysis of the entire dataset is needed and currently ongoing. This already highlights the importance of advanced bioinformatic tools and workflows to support these complex analyses in a systematic approach. Currently, this a limiting factor for a clinical routine as the sheer amount of data and its complexity make it increasingly challenging to find meaningful interpretation. This becomes especially important when turn-around times need to be achieved.

### **5.3. Re-evaluation of the status quo: clinical proteomics**

The quantitative profiling of thousands of proteins across hundreds of samples remains challenging. Yet, massive efforts in the proteomic field towards standardization, simplification, and automation, such as the autoSP3 workflow, are rapidly moving towards its feasibility<sup>148,151,161,304,330,410</sup>. In combination with new-generation mass spectrometers, their sequencing speed, sensitivity, and robustness, and cost-effectiveness of workflows, it is already possible to perform such large-scale profiling experiments in acceptable turn-around times<sup>113,116</sup>. Complementary to other -omics levels, this ability of molecular (proteome) profiling and characterization will build the path to patient-oriented systems

medicine (precision medicine), as described in chapter 1. The anticipated aim of molecular profiling could be a disease (sub)classification, the identification or screening of (predictive) biomarkers, or the functional insight into the (patho)physiology, such as disease progression or relapse<sup>1,2,27</sup>. However, many limitations yet remain to be solved for successful clinical integration. This relates to ethical, legal, logistical, but most importantly, bioinformatic bottlenecks.

Nowadays, the generation of comprehensive data, such as in-depth proteome profiles or even multi-omics data, for the molecular characterization of an individual is becoming feasible from minute amounts of available sample material and at affordable costs<sup>117</sup>. While this can obviously provide a deeper understanding of an individual's molecular make-up and disease phenotype, it is not trivial to extract and interpret the biological and/or clinically relevant results. In the literature, several great examples have emerged that illustrate the utility of proteomics and/or multi-omics data to generate new medical knowledge or identify clinically actionable targets<sup>131,137,139,144,406</sup>. However, these types of analyses typically require a significant expenditure of time spent by a specialist to understand and interpret the data. This becomes even more challenging with multiple -omics layers and additional clinical or health record information about an individual. Here, an increasing number of tools and integrative solutions, such as MOFA, RGCCA, MCIA, iCLUSTER, and others, are becoming available to support the interpretation of complex multi-level data<sup>169</sup>. In an ideal scenario, easy-to-use or automated software tools are key to rapidly extract useful information, allowing fast and straightforward interpretation for a patient's benefit. In foresight, sophisticated software solutions utilizing machine learning algorithms might be useful to identify traits or trends in complex datasets that are not easily observed by manual investigation. Other accompanying bottlenecks are the logistics of data handling and storage. It is not surprising that research groups and especially companies (e.g., Biognosys, Roche Diagnostics, OmicEra Diagnostic) recognize this gap between immense amounts of data and their meaningful interpretation for a clinical utility. Several initiatives (SMART-CARE, CLINSPECT-M, DIASyM, and MSTARs) that were recently funded by the federal ministry of education and research (BMBF, Germany), are envisaging to tackle such remaining bottlenecks.

## Discussion

Another crucial factor for a routine application of personalized molecular profiling is the ethical point of view. Of course, this has been and is extensively discussed in the field and will need the development of clear regulatory systems<sup>1,26</sup>. For example, what and how are the results presented to an individual. What if risk genes or protein expression for specific diseases are detected pre-symptomatically. Individual genes or proteins might reveal an increased predisposition of developing a specific disease. How does a physician deal with incidental observations? In some commercially available genetic testing, for example, offered by Dante Labs (L'Aquila, Italy)<sup>411</sup> genetic counseling with a specialist is already recommended (e.g., by DNAfeed Inc. (San Diego, USA))<sup>412</sup>, but not mandatory in order to provide proper education about potential findings and their implications. For a routine integration of molecular profiling, this individualized counseling will require financing and time of specialists, such as physicians and potentially even psychologists. Consensus agreements need to be established in order to protect the patient's rights and molecular data while simultaneously maintaining the benefits of molecular profiling. Here, the health insurance and portability and accountability act (HIPAA)<sup>413</sup> or the general data protection regulations (GDPR)<sup>414</sup> in the United States or European Union, respectively, define such standard measures for the protection of physical, network, and process security (e.g., data protection= HIPAA or GDPR compliant). On the other hand, this compliance often introduces additional bottlenecks, such as the high maintenance costs and paperwork, and the limited ability for physicians or researchers to perform retrospective (e.g., biobanked tissue) or prospective evaluation of patient samples and resulting data<sup>415</sup>. Informed consent with strict regulations but potentially also the possibility for individualized considerations are key for routine implementation. This can only be achieved by joint agreements between all participating parties ranging from insurance companies and health care providers to scientists, bioinformaticians, and medical doctors, to the individual patient.

Altogether, the advantages of complementary -omics profiling, including proteomics, clearly position it as the key to personalized medicine. Unambiguously, the sheer complexity of such data requires a systems medicine approach for the extraction and interpretation of meaningful results and decision-making guidance in a clinical environment. Beyond a better understanding of biological systems and (patho)physiology



itself, an improved patient stratification and classification, the identification of biomarkers, or new therapeutic targets are among the anticipated aims. In addition, molecular characterization might lead to better therapy decisions, such as avoiding or de-escalating a specific therapy approach for individuals. In 2013 and 2015, the interest in systems medicine has triggered the foundation of the “e:Med” consortium<sup>416</sup> as well as the “European Association of Systems Medicine e.V.” (EASyM)<sup>417</sup>. Both essentially aim to make personalized and systems medicine available to everyone by tackling the major questions and bottlenecks: I) getting together all responsible and relevant fields of expertise, including clinicians, researchers, medical and patient organization, funders, ethic and privacy authorities, and patients. II) Establishing a hands-on training and education framework. III) Developing guidelines for data handling, from storage to analysis and interpretation. IV) Promoting and supporting the implementation of systems medicine “big data” in routine applications. V) Evolving sophisticated computer-based solutions (e.g., machine learning and artificial intelligence) for the analysis of complex (multi-omics) data. Since its inception, the e:Med research and funding concept has resulted in 1410 systems medicine oriented publications (as of 03.06.2020)<sup>416</sup>. With this, we think that the technical framework for large-scale systems medicine data generation can already be actionable. The logistical, ethical, and mainly bioinformatic solutions for a routine implementation are lagging behind. However, these limitations will be tackled and solved in the coming years, while standardization and performance of multi-omics and especially proteomics pipelines will continue to become more sensitive, more reproducible, and easy-to-use.

The future of medicine with personalized -omics profiling and decision-making is at a tipping point from bench to bedside.



## 6. References

1. Apweiler, R. *et al.* Whither systems medicine? *Exp. Mol. Med.* **50**, e453-6 (2018).
2. Kirschner, M. in *Systems Medicine* (eds. Schmitz, U. & Wolkenhauer, O.) 3–15 (Springer New York, 2016). doi:10.1007/978-1-4939-3283-2\_1
3. Tian, Q., Price, N. D. & Hood, L. Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. *J. Intern. Med.* **271**, 111–121 (2012).
4. Schleidgen, S. *et al.* Applying systems biology to biomedical research and health care: A précising definition of systems medicine. *BMC Health Serv. Res.* **17**, 1–16 (2017).
5. Bechtel, W. & Richardson, R. C. *Discovering complexity: Decomposition and localization as strategies in scientific research. Discovering complexity: Decomposition and localization as strategies in scientific research.* (Princeton University Press, 1993).
6. Brigandt, Ingo; Love, A. *The Stanford Encyclopedia of Philosophy - Reductionism in Biology.* (Metaphysics Research Lab, Stanford University, 2017).
7. Ahn, A. C., Tewari, M., Poon, C.-S. & Phillips, R. S. The limits of reductionism in medicine: could systems biology offer an alternative? *PLoS Med.* **3**, e208–e208 (2006).
8. Beresford, M. J. Medical reductionism: lessons from the great philosophers. *QJM An Int. J. Med.* **103**, 721–724 (2010).
9. Tretter, F. 'Systems medicine' in the view of von Bertalanffy's 'organismic biology' and systems theory. *Syst. Res. Behav. Sci.* **36**, 346–362 (2019).
10. Wolkenhauer, O., Auffray, C., Jaster, R., Steinhoff, G. & Dammann, O. The road from systems biology to systems medicine. *Pediatr. Res.* **73**, 502–507 (2013).
11. Mazzocchi, F. Complexity and the reductionism–holism debate in systems biology. *WIREs Syst. Biol. Med.* **4**, 413–427 (2012).
12. Anderson, P. W. More Is Different. *Sci. New Ser.* **177**, 393–396 (1972).
13. Cohen, S. M. *The Stanford Encyclopedia of Philosophy - Aristotle's Metaphysics.* (Metaphysics Research Lab, Stanford University, 2016).
14. Smuts, J. C. . *Holism and Evolution.* (Macmillan And Company Limited, 1926).
15. KERR, J. G. Holism and Evolution. *Nature* **119**, 307–309 (1927).
16. Bertalanffy, L. von. *General system theory; foundations, development, applications.* (New York : G. Braziller, [1969] [©1968], 1969).
17. Avery, O. T., MacLeod, C. M. & McCarty, M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III . *J. Exp. Med.* **79**, 137–158 (1944).
18. WATSON, J. D. & CRICK, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
19. Franklin, R. E. & Gosling, R. G. The structure of sodium thymonucleate fibres. I. The influence of water content. *Acta Crystallogr.* **6**, 673–677 (1953).
20. FRANKLIN, R. E. & GOSLING, R. G. Molecular Configuration in Sodium Thymonucleate. *Nature* **171**, 740–741 (1953).
21. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).
22. Ideker, T., Galitski, T. & Hood, L. A NEW APPROACH TO DECODING LIFE: Systems Biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
23. Ideker, T. & Hood, L. A Blueprint for Systems Biology. *Clin. Chem.* **65**, 342–344 (2019).
24. Chuang, H.-Y., Hofree, M. & Ideker, T. A decade of systems biology. *Annu. Rev. Cell Dev. Biol.* **26**, 721–744 (2010).
25. Zeng, B. Z. On the holographic model of human body. in *1st National Conference of Comparative Studies Traditional Chinese Medicine and West Medicine (Medicine and Philosophy, Guangzhou)* (1992).
26. Kamada, T. System biomedicine: a new paradigm in biomedical engineering. *Frontiers of medical and biological engineering : the international journal of the Japan Society of Medical Electronics and Biological Engineering* **4**, 1–2 (1992).
27. Wang, R.-S., Maron, B. A. & Loscalzo, J. Systems medicine: evolution of systems biology from bench to bedside. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **7**, 141–161 (2015).
28. Hippocrates. Hippocrates Quotes. BrainyQuote.com, BrainyMedia Inc. *BrainyQuote.com* (2020). at <[https://www.brainyquote.com/citation/quotes/hippocrates\\_132701](https://www.brainyquote.com/citation/quotes/hippocrates_132701)>
29. Konstantinidou, M. K., Karaglani, M., Panagopoulou, M., Fiska, A. & Chatzaki, E. Are the Origins of Precision Medicine Found in the Corpus Hippocraticum? *Mol. Diagn. Ther.* **21**, 601–606 (2017).
30. Ayers, D. & Day, P. J. Systems Medicine: The Application of Systems Biology Approaches for Modern Medical Research and Drug Development. *Mol. Biol. Int.* **2015**, 698169 (2015).
31. Butcher, E. C., Berg, E. L. & Kunkel, E. J. Systems biology in drug discovery. *Nat. Biotechnol.* **22**, 1253–1259 (2004).
32. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 560–564 (1977).

## References

33. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30**, 418–426 (2014).
34. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
35. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
36. Schloss, J. A. How to get genomes at one ten-thousandth the cost. *Nat. Biotechnol.* **26**, 1113–1115 (2008).
37. Febrer, M., McLay, K., Caccamo, M., Twomey, K. B. & Ryan, R. P. Advances in bacterial transcriptome and transposon insertion-site profiling using second-generation sequencing. *Trends Biotechnol.* **29**, 586–594 (2011).
38. Morozova, O. & Marra, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255–264 (2008).
39. Imelfort, M. & Edwards, D. De novo sequencing of plant genomes using second-generation technologies. *Brief. Bioinform.* **10**, 609–618 (2009).
40. Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res.* **15**, 1767–1776 (2005).
41. Tripathy, S. & Jiang, R. H. Y. Massively parallel sequencing technology in pathogenic microbes. *Methods Mol. Biol.* **835**, 271–294 (2012).
42. Su, Z. *et al.* Next-generation sequencing and its applications in molecular diagnostics. *Expert Rev. Mol. Diagn.* **11**, 333–343 (2011).
43. Chan, E. Y. Next-generation sequencing methods: impact of sequencing accuracy on SNP discovery. *Methods Mol. Biol.* **578**, 95–111 (2009).
44. Lee, H. & Tang, H. Next-generation sequencing technologies and fragment assembly algorithms. *Methods Mol. Biol.* **855**, 155–174 (2012).
45. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **19**, R227–40 (2010).
46. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
47. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
48. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
49. Maróti, Z., Boldogkői, Z., Tombácz, D., Snyder, M. & Kalmár, T. Evaluation of whole exome sequencing as an alternative to BeadChip and whole genome sequencing in human population genetic analysis. *BMC Genomics* **19**, 778 (2018).
50. Li, Y. & Tollefsbol, T. O. DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol. Biol.* **791**, 11–21 (2011).
51. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
52. Ashley, E. A. *et al.* Clinical assessment incorporating a personal genome. *Lancet (London, England)* **375**, 1525–1535 (2010).
53. Strauss, K. A. *et al.* Genome-wide SNP arrays as a diagnostic tool: clinical description, genetic mapping, and molecular characterization of Salla disease in an Old Order Mennonite population. *Am. J. Med. Genet. A* **138A**, 262–267 (2005).
54. Baker, L., Muir, P. & Sample, S. J. Genome-wide association studies and genetic testing: understanding the science, success, and future of a rapidly developing field. *J. Am. Vet. Med. Assoc.* **255**, 1126–1136 (2019).
55. Yoshida, S. *et al.* Prenatal diagnosis of Gaucher disease using next-generation sequencing. *Pediatr. Int.* **58**, 946–949 (2016).
56. Poujois, A. & Woimant, F. Challenges in the diagnosis of Wilson disease. *Ann. Transl. Med.* **7**, S67–S67 (2019).
57. Petersen, B.-S. *et al.* Targeted Gene Panel Sequencing for Early-onset Inflammatory Bowel Disease and Chronic Diarrhea. *Inflamm. Bowel Dis.* **23**, 2109–2120 (2017).
58. Au, C. H., Wa, A., Ho, D. N., Chan, T. L. & Ma, E. S. K. Clinical evaluation of panel testing by next-generation sequencing (NGS) for gene mutations in myeloid neoplasms. *Diagn. Pathol.* **11**, 11 (2016).
59. Sun, Y. *et al.* Next-generation diagnostics: gene panel, exome, or whole genome? *Hum. Mutat.* **36**, 648–655 (2015).
60. LaDuca, H. *et al.* Exome sequencing covers >98% of mutations identified on targeted next generation sequencing panels. *PLoS One* **12**, e0170843 (2017).
61. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19096–19101 (2009).
62. Rabbani, B., Tekin, M. & Mahdieh, N. The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet.* **59**, 5–15 (2014).
63. Worthey, E. A. *et al.* Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.* **13**, 255–262 (2011).
64. Stasik, S. *et al.* An optimized targeted Next-Generation Sequencing approach for sensitive detection of single nucleotide variants. *Biomol. Detect. Quantif.* **15**, 6–12 (2018).
65. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20**, 117 (2019).
66. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).

67. Tattini, L., D'Aurizio, R. & Magi, A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front. Bioeng. Biotechnol.* **3**, 92 (2015).
68. Neerman, N. *et al.* A clinically validated whole genome pipeline for structural variant detection and analysis. *BMC Genomics* **20**, 545 (2019).
69. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015**, 951–969 (2015).
70. Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42**, 806–810 (2010).
71. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.* **36**, 5221–5231 (2008).
72. Mack, S. C. *et al.* Therapeutic targeting of ependymoma as informed by oncogenic enhancer profiling. *Nature* (2017). doi:10.1038/nature25169
73. Pajtler, K. W. *et al.* Molecular Classification of Ependymal Tumors across All CNS Compartments, Histopathological Grades, and Age Groups. *Cancer Cell* **27**, 728–743 (2015).
74. Harvey, Z. H., Chen, Y. & Jarosz, D. F. Protein-Based Inheritance: Epigenetics beyond the Chromosome. *Mol. Cell* **69**, 195–202 (2018).
75. Johann, P. D. *et al.* Atypical Teratoid/Rhabdoid Tumors Are Comprised of Three Epigenetic Subgroups with Distinct Enhancer Landscapes. *Cancer Cell* **29**, 379–393 (2016).
76. Li, Q., Hermanson, P. J. & Springer, N. M. Detection of DNA Methylation by Whole-Genome Bisulfite Sequencing. *Methods Mol. Biol.* **1676**, 185–196 (2018).
77. Cheng, X., Zhang, L., Chen, Y. & Qing, C. Circulating cell-free DNA and circulating tumor cells, the 'liquid biopsies' in ovarian cancer. *J. Ovarian Res.* **10**, 75 (2017).
78. Volckmar, A.-L. *et al.* A field guide for cancer diagnostics using cell-free DNA: From principles to practice and clinical applications. *Genes. Chromosomes Cancer* **57**, 123–139 (2018).
79. Kamps, R. *et al.* Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification. *Int. J. Mol. Sci.* **18**, (2017).
80. Reynolds, R. H. *et al.* Moving beyond neurons: the role of cell type-specific gene regulation in Parkinson's disease heritability. *npj Park. Dis.* **5**, 6 (2019).
81. Chang, W.-S., Wang, Y.-H., Zhu, X.-T. & Wu, C.-J. Genome-Wide Profiling of miRNA and mRNA Expression in Alzheimer's Disease. *Med. Sci. Monit.* **23**, 2721–2731 (2017).
82. Ounzain, S. *et al.* Genome-wide profiling of the cardiac transcriptome after myocardial infarction identifies novel heart-specific long non-coding RNAs. *Eur. Heart J.* **36**, 353–368 (2014).
83. Ratner, M. Next-generation sequencing tests to become routine. *Nat. Biotechnol.* **36**, 484 (2018).
84. Liu, Z., Zhu, L., Roberts, R. & Tong, W. Toward Clinical Implementation of Next-Generation Sequencing-Based Genetic Testing in Rare Diseases: Where Are We? *Trends Genet.* **35**, 852–867 (2019).
85. Milner, L. C. *et al.* Genomics in the clinic: ethical and policy challenges in clinical next-generation sequencing programs at early adopter USA institutions. *Per. Med.* **12**, 269–282 (2015).
86. Xuan, J., Yu, Y., Qing, T., Guo, L. & Shi, L. Next-generation sequencing in the clinic: promises and challenges. *Cancer Lett.* **340**, 284–295 (2013).
87. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
88. Edfors, F. *et al.* Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.* **12**, 883 (2016).
89. Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? *Nature* **547**, E19–E20 (2017).
90. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).
91. Mulvey, C. M. *et al.* Using hyperLOPIT to perform high-resolution mapping of the spatial proteome. *Nat. Protoc.* **12**, 1110–1135 (2017).
92. Heusel, M. *et al.* Complex-centric proteome profiling by SEC-SWATH-MS. *Mol. Syst. Biol.* **15**, e8438 (2019).
93. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
94. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
95. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
96. Cox, J. & Mann, M. Is proteomics the new genomics? *Cell* **130**, 395–398 (2007).
97. Wilkins, M. R. *et al.* Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It. *Biotechnol. Genet. Eng. Rev.* **13**, 19–50 (1996).
98. HARTLEY, H. Origin of the Word 'Protein'. *Nature* **168**, 244 (1951).
99. PAULING, L., COREY, R. B. & BRANSON, H. R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* **37**, 205–211 (1951).
100. KAUZMANN, W. Structural factors in protein denaturation. *J. Cell. Physiol. Suppl.* **47**, 113–131 (1956).
101. Sanger, F. Species Differences in Insulins. *Nature* **164**, 529 (1949).
102. SANGER, F. The terminal peptides of insulin. *Biochem. J.* **45**, 563–574 (1949).
103. Kendrew, J. C. *et al.* A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature*

## References

- 181**, 662–666 (1958).
104. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711 (2004).
  105. Maher, S., Jjunju, F. P. M. & Taylor, S. Colloquium: 100 years of mass spectrometry: Perspectives and future trends. *Rev. Mod. Phys.* **87**, 113–135 (2015).
  106. Paul, W. Electromagnetic traps for charged and neutral particles. *Rev. Mod. Phys.* **62**, 531–540 (1990).
  107. Prestage, J. D., Janik, G. R., Dick, G. J. & Maleki, L. Linear ion trap for second-order Doppler shift reduction in frequency standard applications. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **37**, 535–542 (1990).
  108. Yamashita, M. & Fenn, J. B. Electrospray Ion Source. Another Variation on the Free-Jet Theme. **434**, 4451–4459 (1984).
  109. Fenn, J. B. Electrospray wings for molecular elephants (Nobel lecture). *Angew. Chem. Int. Ed. Engl.* **42**, 3871–3894 (2003).
  110. Wuethrich, K. The development of nuclear magnetic resonance spectroscopy as a technique for protein structure determination. *Acc. Chem. Res.* **22**, 36–44 (1989).
  111. Makarov, A. Electrostatic Axially Harmonic Orbital Trapping-A High Performance Technique of Mass Analysis. *Anal. Chem.* **72**, 1156–1162 (2000).
  112. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* **40**, 430–443 (2005).
  113. Meier, F. *et al.* Online Parallel Accumulation-Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol. Cell. proteomics* **17**, 2534–2545 (2018).
  114. Vasilopoulou, C. G. *et al.* Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts. *Nat. Commun.* **11**, 331 (2020).
  115. Yu, Q. *et al.* Benchmarking the Orbitrap Tribrid Eclipse for Next Generation Multiplexed Proteomics. *Anal. Chem.* **92**, 6478–6485 (2020).
  116. Hebert, A. S. *et al.* Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Anal. Chem.* **90**, 9529–9537 (2018).
  117. Bian, Y. *et al.* Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC-MS/MS. *Nat. Commun.* **11**, 157 (2020).
  118. Bache, N. *et al.* A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics. *Mol. Cell. proteomics* **17**, 2284–2296 (2018).
  119. Tsai, T.-H. *et al.* Selection of Features with Consistent Profiles Improves Relative Protein Quantification in Mass Spectrometry Experiments. *Mol. Cell. proteomics* **19**, 944 LP-959 (2020).
  120. Huang, T. *et al.* Combining Precursor and Fragment Information for Improved Detection of Differential Abundance in Data Independent Acquisition. *Mol. Cell. proteomics* **19**, 421 LP-430 (2020).
  121. Kelstrup, C. D. *et al.* Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* **17**, 727–738 (2018).
  122. Casey, T. M. *et al.* Analysis of Reproducibility of Proteome Coverage and Quantitation Using Isobaric Mass Tags (iTRAQ and TMT). *J. Proteome Res.* **16**, 384–392 (2017).
  123. Schubert, O. T., Röst, H. L., Collins, B. C., Rosenberger, G. & Aebersold, R. Quantitative proteomics: Challenges and opportunities in basic and applied research. *Nat. Protoc.* **12**, 1289–1294 (2017).
  124. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
  125. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
  126. Frantzi, M., Latosinska, A., Kontostathi, G. & Mischak, H. Clinical Proteomics: Closing the Gap from Discovery to Implementation. *Proteomics* **18**, e1700463 (2018).
  127. Timp, W. & Timp, G. Beyond mass spectrometry, the next step in proteomics. *Sci. Adv.* **6**, eaax8978 (2020).
  128. Kellie, J. F. *et al.* A new era for proteomics. *Bioanalysis* **11**, 1731–1735 (2019).
  129. Vlahou, A. Implementation of Clinical Proteomics: A Step Closer to Personalized Medicine? *PROTEOMICS – Clin. Appl.* **13**, 1800088 (2019).
  130. Doll, S. *et al.* Rapid proteomic analysis for solid tumors reveals LSD1 as a drug target in an end-stage cancer patient. *Mol. Oncol.* **12**, 1296–1307 (2018).
  131. Archer, T. C. *et al.* Proteomics, Post-translational Modifications, and Integrative Analyses Reveal Molecular Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell* **34**, 396–410.e8 (2018).
  132. Dotsey, E. Y. *et al.* A High Throughput Protein Microarray Approach to Classify HIV Monoclonal Antibodies and Variant Antigens. *PLoS One* **10**, e0125581 (2015).
  133. Hirota, S. Differential diagnosis of gastrointestinal stromal tumor by histopathology and immunohistochemistry. *Transl. Gastroenterol. Hepatol.* **3**, 27 (2018).
  134. Bonk, S. *et al.* Prognostic and diagnostic role of PSA immunohistochemistry: A tissue microarray study on 21,000 normal and cancerous tissues. *Oncotarget* **10**, 5439–5453 (2019).
  135. Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
  136. Terry, S. F. Obama's Precision Medicine Initiative. *Genetic testing and molecular biomarkers* **19**, 113–114 (2015).
  137. Doll, S., Gnad, F. & Mann, M. The Case for Proteomics and Phospho-Proteomics in Personalized Cancer Medicine. *Proteomics. Clin. Appl.* **13**, e1800113 (2019).
  138. Lundberg, E. & Borner, G. H. H. Spatial proteomics: a powerful discovery tool for cell biology. *Nat. Rev. Mol. Cell Biol.* **20**, 285–302 (2019).

139. Davis, S., Scott, C., Ansorge, O. & Fischer, R. Development of a Sensitive, Scalable Method for Spatial, Cell-Type-Resolved Proteomics of the Human Brain. *J. Proteome Res.* **18**, 1787–1795 (2019).
140. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Sci.* **27**, 233–244 (2018).
141. Bunina, D. *et al.* Genomic Rewiring of SOX2 Chromatin Interaction Network during Differentiation of ESCs to Postmitotic Neurons. *Cell Syst.* (2020). doi:10.1016/j.cels.2020.05.003
142. Gao, Y. *et al.* Protein Expression Landscape of Mouse Embryos during Pre-implantation Development. *Cell Rep.* **21**, 3957–3969 (2017).
143. Ginno, P. A., Burger, L., Seebacher, J., Iesmantavicius, V. & Schübeler, D. Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape. *Nat. Commun.* **9**, 4048 (2018).
144. Herr, P. *et al.* Cell Cycle Profiling Reveals Protein Oscillation, Phosphorylation, and Localization Dynamics. *Mol. Cell. proteomics* **19**, 608 LP-623 (2020).
145. Sabatier, P., Saei, A. A., Wang, S. & Zubarev, R. A. Dynamic Proteomics Reveals High Plasticity of Cellular Proteome: Growth-Related and Drug-Induced Changes in Cancer Cells are Comparable. *Proteomics* **18**, e1800118 (2018).
146. von Stechow, L. & Olsen, J. V. Proteomics insights into DNA damage response and translating this knowledge to clinical strategies. *Proteomics* **17**, 1600018 (2017).
147. Mullis, K. *et al.* Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 263–273 (1986).
148. Fu, Q. *et al.* Highly Reproducible Automated Proteomics Sample Preparation Workflow for Quantitative Mass Spectrometry. *J. Proteome Res.* **17**, 420–428 (2018).
149. Hughes, C. S. *et al.* Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol. Syst. Biol.* **10**, 757 (2014).
150. Bache, N. *et al.* A Novel LC System Embeds Analytes in Pre-formed Gradients for Rapid, Ultra-robust Proteomics. *Mol. Cell. Proteomics* **17**, 2284–2296 (2018).
151. Milkessa, H. *et al.* S-Trap, an Ultrafast Sample-Preparation Approach for Shotgun Proteomics. *J. Proteome Res.* **17**, 2917–2924 (2018).
152. Huynh, ML, Russell, P, Walsh, B. Tryptic digestion of in-gel proteins for mass spectrometry analysis. *Methods Mol Biol* **519**, 507–513 (2009).
153. Xiaolin, W., Erhui, X., Wei, W., Monica, S. & Mauro, C. Universal sample preparation method integrating trichloroacetic acid/acetone precipitation with phenol extraction for crop proteomic analysis. *Nat. Protoc.* **9**, 362–374 (2014).
154. Rappsilber, J, Ishihama, Y, Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75**, 663–70 (2003).
155. Ludwig, K. R., Schroll, M. M. & Hummon, A. B. Comparison of In-Solution, FASP, and S-Trap Based Digestion Methods for Bottom-Up Proteomic Studies. *J. Proteome Res.* **17**, 2480–2490 (2018).
156. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
157. Guo, T. *et al.* Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nat. Med.* (2015). doi:10.1038/nm.3807
158. Jacek R Wiśniewski, Alexandre Zougman, N. N. & M. M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).
159. Wiśniewski, J. R., Duś, K. & Mann, M. Proteomic workflow for analysis of archival formalin-fixed and paraffin-embedded clinical samples to a depth of 10000 proteins. *Proteomics - Clin. Appl.* **7**, 225–233 (2013).
160. Paulo, J. A., Navarrete-Perea, J. & Gygi, S. P. Multiplexed proteome profiling of carbon source perturbations in two yeast species with SL-SP3-TMT. *J. Proteomics* **210**, 103531 (2019).
161. Kuras, M. *et al.* Assessing Automated Sample Preparation Technologies for High-Throughput Proteomics of Frozen Well Characterized Tissues from Swedish Biobanks. *J. Proteome Res.* **18**, 548–556 (2019).
162. Ma, Y. *et al.* Intra-tumoural heterogeneity characterization through texture and colour analysis for differentiation of non-small cell lung carcinoma subtypes. *Phys. Med. Biol.* **63**, (2018).
163. Zhu, Y. *et al.* Proteomic analysis of single mammalian cells enabled by microfluidic nanodroplet sample preparation and ultrasensitive nanoLC-MS. *Angew. Chemie Int. Ed.* 1–6 (2018). doi:10.1002/anie.201802843
164. Doll, S. *et al.* Region and cell-type resolved quantitative proteomic map of the human heart. *Nat. Commun.* **8**, 1469 (2017).
165. Bekker-Jensen, D. B. *et al.* An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst.* **4**, 587–599.e4 (2017).
166. Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755–765 (2016).
167. Sun, Y. *et al.* Comparative Proteomic Analysis of Exosomes and Microvesicles in Human Saliva for Lung Cancer. *J. Proteome Res.* **17**, 1101–1107 (2018).
168. Hughes, C. S. *et al.* Quantitative Profiling of Single Formalin Fixed Tumour Sections: proteomics for translational research. *Sci. Rep.* **6**, 34949 (2016).
169. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and

## References

- Its Application. *Bioinform. Biol. Insights* **14**, 1177932219899051–1177932219899051 (2020).
170. Pinu, F. R. *et al.* Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. *Metabolites* **9**, 76 (2019).
171. Bertier, G., Carrot-Zhang, J., Ragoussis, V. & Joly, Y. Integrating precision cancer medicine into healthcare-policy, practice, and research challenges. *Genome Med.* **8**, 1–12 (2016).
172. Hood, L., Balling, R. & Auffray, C. Revolutionizing medicine in the 21st century through systems approaches. *Biotechnol. J.* **7**, 992–1001 (2012).
173. WHO. World Health Organization (WHO). Cancer Fact Sheet. (2020). at <<http://www.who.int/en/news-room/fact-sheets/detail/cancer>>
174. Organization, W. H. ICD-10 : international statistical classification of diseases and related health problems : tenth revision.
175. Lancet, T. ICD-11. *Lancet (London, England)* **393**, 2275 (2019).
176. Organization, W. H. *International classification of diseases for oncology (ICD-O) – 3rd edition, 1st revision.*
177. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
178. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
179. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
180. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA. Cancer J. Clin.* **68**, 7–30 (2018).
181. Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, C. K. (eds). SEER Cancer Statistics Review, 1975–2017, National Cancer Institute. Bethesda, MD. (2020). at <[https://seer.cancer.gov/csr/1975\\_2017](https://seer.cancer.gov/csr/1975_2017)>
182. de Magalhães, J. P. How ageing processes influence cancer. *Nature reviews. Cancer* **13**, 357–365 (2013).
183. Jemal, A. *et al.* Annual Report to the Nation on the Status of Cancer, 1975–2014, Featuring Survival. *JNCI J. Natl. Cancer Inst.* **109**, (2017).
184. Rutter, C. M. *et al.* Secular Trends in Colon and Rectal Cancer Relative Survival. *JNCI J. Natl. Cancer Inst.* **105**, 1806–1813 (2013).
185. Roser, Max; Ritchie, H. Cancer. *OurWorldInData.org* (2015). at <<https://ourworldindata.org/cancer>>
186. Smith, M. A., Altekruse, S. F., Adamson, P. C., Reaman, G. H. & Seibel, N. L. Declining childhood and adolescent cancer mortality. *Cancer* **120**, 2497–2506 (2014).
187. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science (80-. ).* **349**, 1483 LP-1489 (2015).
188. Yi, K. & Ju, Y. S. Patterns and mechanisms of structural variations in human cancer. *Exp. Mol. Med.* **50**, 98 (2018).
189. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
190. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
191. Lee, E. Y. H. P. & Muller, W. J. Oncogenes and tumor suppressor genes. *Cold Spring Harb. Perspect. Biol.* **2**, a003236 (2010).
192. Anderson, M. W., Reynolds, S. H., You, M. & Maronpot, R. M. Role of proto-oncogene activation in carcinogenesis. *Environ. Health Perspect.* **98**, 13–24 (1992).
193. Alitalo, K. & Schwab, M. Oncogene amplification in tumor cells. *Adv. Cancer Res.* **47**, 235–281 (1986).
194. Nazarenko, I. *et al.* PDGF and PDGF receptors in glioma. *Ups. J. Med. Sci.* **117**, 99–112 (2012).
195. di Magliano, M. P. & Logsdon, C. D. Roles for KRAS in pancreatic tumor development and progression. *Gastroenterology* **144**, 1220–1229 (2013).
196. Pei, Y. *et al.* An animal model of MYC-driven medulloblastoma. *Cancer Cell* **21**, 155–167 (2012).
197. Testa, J. R. & Tsichlis, P. N. AKT signaling in normal and malignant cells. *Oncogene* **24**, 7391–7393 (2005).
198. Travis, W. D. *et al.* The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **10**, 1243–1260 (2015).
199. Venuta, F. *et al.* Lung cancer in elderly patients. *J. Thorac. Dis.* **8**, S908–S914 (2016).
200. SEER Cancer Stat Facts: Lung and bronchus cancer. National Cancer Institute. Bethesda, MD. *SEER Cancer Stat Facts. National Cancer Institute* (2020). at <<https://seer.cancer.gov/statfacts/html/lungb.html>>
201. Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E. & Adjei, A. A. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin. Proc.* **83**, 584–594 (2008).
202. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613–628 (2017).
203. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
204. Kazdal, D. *et al.* Subclonal evolution of pulmonary adenocarcinomas delineated by spatially distributed somatic mitochondrial mutations. *Lung Cancer* **126**, 80–88 (2018).
205. Annamaria, C. *et al.* Lung cancer histologic and immunohistochemical heterogeneity in the era of molecular therapies: analysis of 172 consecutive surgically resected, entirely sampled pulmonary carcinomas. *Am. J. Surg. Pathol.* **38**, 502–509 (2014).
206. Travis, W. D. *et al.* The 2015 World Health Organization Classification of Lung Tumors. *J. Thorac. Oncol.* **10**, 1243–1260 (2015).



207. Warth, A. *et al.* The novel histologic International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society classification system of lung adenocarcinoma is a stage-independent predictor of survival. *J. Clin. Oncol.* **30**, 1438–1446 (2012).
208. Tsao, M. S. *et al.* Subtype classification of lung adenocarcinoma predicts benefit from adjuvant chemotherapy in patients undergoing complete resection. *J. Clin. Oncol.* **33**, 3439–3446 (2015).
209. Lung Cancer - Non-Small Cell. Cancer.Net. *Cancer.Net* (2020). at <<https://www.cancer.net/cancer-types/lung-cancer-non-small-cell>>
210. in (2002).
211. Lu, Y. *et al.* A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. *PLoS Med.* **3**, e467–e467 (2006).
212. Yoo, S. S. *et al.* Effects of polymorphisms identified in genome-wide association studies of never-smoking females on the prognosis of non-small cell lung cancer. *Cancer Genet.* **212–213**, 8–12 (2017).
213. Truong, T. *et al.* Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. *J. Natl. Cancer Inst.* **102**, 959–971 (2010).
214. Landi, M. T. *et al.* A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am. J. Hum. Genet.* **85**, 679–691 (2009).
215. Sharma, S. V, Bell, D. W., Settleman, J. & Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nat. Rev. Cancer* **7**, 169–181 (2007).
216. Boutros, P. C. *et al.* Prognostic gene signatures for non-small-cell lung cancer. *Proc. Natl. Acad. Sci.* **106**, 2824 LP–2828 (2009).
217. Molina-Romero, C. *et al.* Differential gene expression profiles according to the Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society histopathological classification in lung adenocarcinoma subtypes. *Hum. Pathol.* **66**, 188–199 (2017).
218. Ostrom, Q. T. *et al.* Alex’s Lemonade Stand Foundation Infant and Childhood Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2007–2011. *Neuro. Oncol.* **16 Suppl 1**, x1–x36 (2015).
219. Kaatsch, P. Epidemiology of childhood cancer. *Cancer Treat. Rev.* **36**, 277–285 (2010).
220. Walker, C., Baborie, A., Crooks, D., Wilkins, S. & Jenkinson, M. D. Biology, genetics and imaging of glial cell tumours. *Br. J. Radiol.* **84 Spec No**, S90–S106 (2011).
221. Zong, H., Verhaak, R. G. W. & Canoll, P. The cellular origin for malignant glioma and prospects for clinical advancements. *Expert Rev. Mol. Diagn.* **12**, 383–394 (2012).
222. Purdy, E. *et al.* Ependymoma in children under the age of 3 years: a report from the Canadian Pediatric Brain Tumour Consortium. *J. Neurooncol.* **117**, 359–364 (2014).
223. Shenoy, SS.; Lui, F. Neuroanatomy, Ventricular System. *Treasure Island (FL): StatPearls Publishing* at <<https://www.ncbi.nlm.nih.gov/books/NBK532932/>>
224. Rogers, K. Ependymal cell. Encyclopædia Britannica. *Encyclopædia Britannica, inc.* at <<https://www.britannica.com/science/ependymal-cell>>
225. Kufe, Donald W.; Pollock, Raphael E.; Weichselbaum, Ralph R.; Bast, Robert C. Jr.; Gansler, Ted S.; Holland, James F.; Frei, E. I. *Holland-Frei Cancer Medicine, 8th edition.* (McGraw-Hill Education Ltd, 2010).
226. Wu, J., Armstrong, T. S. & Gilbert, M. R. Biology and management of ependymomas. *Neuro. Oncol.* **18**, 902–913 (2016).
227. Witt, H. *et al.* DNA methylation-based classification of ependymomas in adulthood: implications for diagnosis and treatment. *Neuro. Oncol.* **20**, 1616–1624 (2018).
228. Kilday, J.-P. *et al.* Pediatric ependymoma: biological perspectives. *Mol. Cancer Res.* **7**, 765–786 (2009).
229. Ostrom, Q. T. *et al.* CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2007–2011. *Neuro. Oncol.* **16 Suppl 4**, iv1–63 (2014).
230. Kleinman, G. M., Young, R. H. & Scully, R. E. Ependymoma of the ovary: report of three cases. *Hum. Pathol.* **15**, 632–638 (1984).
231. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).
232. Komori, T. The 2016 WHO Classification of Tumours of the Central Nervous System: The Major Points of Revision. *Neurol. Med. Chir. (Tokyo).* **57**, 301–311 (2017).
233. Pfister, S., Hartmann, C. & Korshunov, A. Histology and molecular pathology of pediatric brain tumors. *J. Child Neurol.* **24**, 1375–1386 (2009).
234. Foreman, N. K., Love, S., Gill, S. S. & Coakham, H. B. Second-look surgery for incompletely resected fourth ventricle ependymomas: technical case report. *Neurosurgery* **40**, 856–60; discussion 860 (1997).
235. Rudà, R., Gilbert, M. & Soffietti, R. Ependymomas of the adult: molecular biology and treatment. *Curr. Opin. Neurol.* **21**, 754–761 (2008).
236. Cage, T. A. *et al.* A systematic review of treatment outcomes in pediatric patients with intracranial ependymomas. *J. Neurosurg. Pediatr.* **11**, 673–681 (2013).
237. Massimino, M. *et al.* Second-look surgery for ependymoma: the Italian experience. *J. Neurosurg. Pediatr.* **8**, 246–250 (2011).
238. Tihan, T. *et al.* The prognostic value of histological grading of posterior fossa ependymomas in children: a Children’s Oncology Group study and a review of prognostic factors. *Mod. Pathol.* **21**, 165–177 (2008).

## References

239. Merchant, T. E. *et al.* Conformal radiotherapy after surgery for paediatric ependymoma: a prospective study. *Lancet. Oncol.* **10**, 258–266 (2009).
240. Koc, K., Anik, I., Cabuk, B. & Ceylan, S. Fluorescein sodium-guided surgery in glioblastoma multiforme: a prospective evaluation. *Br. J. Neurosurg.* **22**, 99–103 (2008).
241. Rutka, J. T. & Kuo, J. S. Pediatric surgical neuro-oncology: current best care practices and strategies. *J. Neurooncol.* **69**, 139–150 (2004).
242. Duffner, P. K., Cohen, M. E. & Freeman, A. I. Pediatric brain tumors: an overview. *CA. Cancer J. Clin.* **35**, 287–301 (1985).
243. Robertson, P. L. *et al.* Incidence and severity of postoperative cerebellar mutism syndrome in children with medulloblastoma: a prospective study by the Children’s Oncology Group. *J. Neurosurg.* **105**, 444–451 (2006).
244. Suc, E. *et al.* Brain tumours under the age of three. The price of survival. A retrospective study of 20 long-term survivors. *Acta Neurochir. (Wien).* **106**, 93–98 (1990).
245. Knab, B. & Connell, P. P. Radiotherapy for pediatric brain tumors: when and how. *Expert Rev. Anticancer Ther.* **7**, S69–77 (2007).
246. Davis, P. C., Hoffman, J. C. J., Pearl, G. S. & Braun, I. F. CT evaluation of effects of cranial radiation therapy in children. *AJR. Am. J. Roentgenol.* **147**, 587–592 (1986).
247. Vera-Bolanos, E. *et al.* Clinical course and progression-free survival of adult intracranial and spinal ependymoma patients. *Neuro. Oncol.* **17**, 440–447 (2015).
248. Duong, C. *et al.* Genomic and Molecular Characterization of Brain Tumors in Asian and Non-Asian Patients of Los Angeles: A Single Institution Analysis. *Brain tumor Res. Treat.* **5**, 64–69 (2017).
249. Boudreau, C. R. & Liau, L. M. Molecular characterization of brain tumors. *Clin. Neurosurg.* **51**, 81–90 (2004).
250. Gilles, F. H. *et al.* Pathologist interobserver variability of histologic features in childhood brain tumors: results from the CCG-945 study. *Pediatr. Dev. Pathol. Off. J. Soc. Pediatr. Pathol. Paediatr. Pathol. Soc.* **11**, 108–117 (2008).
251. Consortium, T. C. B. T. Intraobserver reproducibility in assigning brain tumors to classes in the world health organization diagnostic scheme. *J. Neurooncol.* **7**, 211–224 (1989).
252. van den Bent, M. J. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician’s perspective. *Acta Neuropathol.* **120**, 297–304 (2010).
253. Sturm, D. *et al.* New Brain Tumor Entities Emerge from Molecular Classification of CNS-PNETs. *Cell* **164**, 1060–1072 (2016).
254. Kumar, R., Liu, A. P. Y., Orr, B. A., Northcott, P. A. & Robinson, G. W. Advances in the classification of pediatric brain tumors through DNA methylation profiling: From research tool to frontline diagnostic. *Cancer* **124**, 4168–4180 (2018).
255. Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
256. Gojo, J. *et al.* Telomerase activation in posterior fossa group A ependymomas is associated with dismal prognosis and chromosome 1q gain. *Neuro. Oncol.* **19**, 1183–1194 (2017).
257. Hübner, J.-M. *et al.* EZHIP/CXorf67 mimics K27M mutated oncohistones and functions as an intrinsic inhibitor of PRC2 function in aggressive posterior fossa ependymoma. *Neuro. Oncol.* **21**, 878–889 (2019).
258. Johnson, R. a *et al.* Cross-species genomics matches driver mutations and cell compartments to model ependymoma. *Nature* **466**, 632–636 (2010).
259. Parker, M. *et al.* C11orf95–RELA fusions drive oncogenic NF-κB signalling in ependymoma. *Nature* **506**, 451–455 (2014).
260. Hübner, J.-M., Kool, M., Pfister, S. M. & Pajtler, K. W. Epidemiology, molecular classification and WHO grading of ependymoma. *J. Neurosurg. Sci.* **62**, 46–50 (2018).
261. Ebert, C. *et al.* Molecular genetic analysis of ependymal tumors. NF2 mutations and chromosome 22q loss occur preferentially in intramedullary spinal ependymomas. *Am. J. Pathol.* **155**, 627–632 (1999).
262. Ludwig PE, Reddy V, V. M. Neuroanatomy, Central Nervous System (CNS). *Treasure Island (FL): StatPearls Publishing* at <<https://www.ncbi.nlm.nih.gov/books/NBK442010/>>
263. Witt, H. *et al.* Delineation of two clinically and molecularly distinct subgroups of posterior fossa ependymoma. *Cancer Cell* **20**, 143–157 (2011).
264. Cavalli, F. M. G. *et al.* Heterogeneity within the PF-EPN-B ependymoma subgroup. *Acta Neuropathol.* **136**, 227–237 (2018).
265. Pajtler, K. W. *et al.* Molecular heterogeneity and CXorf67 alterations in posterior fossa group A (PFA) ependymomas. *Acta Neuropathol.* **136**, 211–226 (2018).
266. Mendrzyk, F. *et al.* Identification of gains on 1q and epidermal growth factor receptor overexpression as independent prognostic markers in intracranial ependymoma. *Clin. cancer Res. an Off. J. Am. Assoc. Cancer Res.* **12**, 2070–2079 (2006).
267. Gojo, J. *et al.* Telomerase activation in posterior fossa group A ependymomas is associated with dismal prognosis and chromosome 1q gain. *Neuro. Oncol.* **19**, 1183–1194 (2017).
268. Massimino, M. *et al.* Intracranial ependymoma: factors affecting outcome. *Future Oncol.* **5**, 207–216 (2009).
269. Korshunov, A. *et al.* Molecular staging of intracranial ependymoma in children and adults. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **28**, 3182–3190 (2010).

270. Panwalkar, P. *et al.* Immunohistochemical analysis of H3K27me3 demonstrates global reduction in group-A childhood posterior fossa ependymoma and is a powerful predictor of outcome. *Acta Neuropathol.* **134**, 705–714 (2017).
271. Bender, S. *et al.* Reduced H3K27me3 and DNA hypomethylation are major drivers of gene expression in K27M mutant pediatric high-grade gliomas. *Cancer Cell* **24**, 660–672 (2013).
272. Hübner, J.-M. *et al.* EZHIP/CXorf67 mimics K27M mutated oncohistones and functions as an intrinsic inhibitor of PRC2 function in aggressive posterior fossa ependymoma. *Neuro. Oncol.* **21**, 878–889 (2019).
273. Mack, S. C. *et al.* Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature* **506**, 445–450 (2014).
274. Laugesen, A., Højfeldt, J. W. & Helin, K. Role of the Polycomb Repressive Complex 2 (PRC2) in Transcriptional Regulation and Cancer. *Cold Spring Harb. Perspect. Med.* **6**, a026575 (2016).
275. Lavarone, E., Barbieri, C. M. & Pasini, D. Dissecting the role of H3K27 acetylation and methylation in PRC2 mediated control of cellular identity. *Nat. Commun.* **10**, 1679 (2019).
276. Andreiuolo, F. *et al.* Childhood supratentorial ependymomas with YAP1-MAMLD1 fusion: an entity with characteristic clinical, radiological, cytogenetic and histopathological features. *Brain Pathol.* **29**, 205–216 (2019).
277. Eder, N. *et al.* YAP1/TAZ drives ependymoma-like tumour formation in mice. *Nat. Commun.* **11**, 2380 (2020).
278. Gumbiner, B. M. & Kim, N.-G. The Hippo-YAP signaling pathway and contact inhibition of growth. *J. Cell Sci.* **127**, 709–717 (2014).
279. Oeckinghaus, A. & Ghosh, S. The NF-kappaB family of transcription factors and its regulation. *Cold Spring Harb. Perspect. Biol.* **1**, a000034–a000034 (2009).
280. Xia, Y., Shen, S. & Verma, I. M. NF-κB, an active player in human cancers. *Cancer Immunol. Res.* **2**, 823–830 (2014).
281. Chen, W., Li, Z., Bai, L. & Lin, Y. NF-kappaB in lung cancer, a carcinogenesis mediator and a prevention and therapy target. *Front. Biosci. (Landmark Ed.)* **16**, 1172–1185 (2011).
282. Kim, C. & Pasparakis, M. Epidermal p65/NF-κB signalling is essential for skin carcinogenesis. *EMBO Mol. Med.* **6**, 970–983 (2014).
283. Wang, W., Nag, S. A. & Zhang, R. Targeting the NFκB signaling pathways for breast cancer prevention and therapy. *Curr. Med. Chem.* **22**, 264–289 (2015).
284. Domingo-Domenech, J. *et al.* Activation of nuclear factor-kappaB in human prostate carcinogenesis and association to biochemical relapse. *Br. J. Cancer* **93**, 1285–1294 (2005).
285. Zhou, J., Ching, Y. Q. & Chng, W.-J. Aberrant nuclear factor-kappa B activity in acute myeloid leukemia: from molecular pathogenesis to therapeutic target. *Oncotarget* **6**, 5490–5500 (2015).
286. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System : a summary. *Acta Neuropathol.* **131**, 803–820 (2016).
287. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. proteomics* **13**, 2513–2526 (2014).
288. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
289. Cox, J. *et al.* Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
290. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
291. Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019).
292. Kazdal, D. *et al.* Prevalence of somatic mitochondrial mutations and spatial distribution of mitochondria in non-small cell lung cancer. *Br. J. Cancer* **117**, 220–226 (2017).
293. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
294. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–52 (2015).
295. Sergushichev, A. A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv* 60012 (2016). doi:10.1101/060012
296. Yu, G. & He, Q. Y. ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).
297. van der Maaten, Laurens, Hinton E., G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **164**, 10 (2008).
298. Medicines, E. D. for the Q. of. Technical Guide for the elaboration of monographs. *Eur. Pharmacopoeia* (2011).
299. EMEA. Guidelines for the validation of analytical methods used in residue depletion studies. *Vet. Med. Insp. Eur. Med. Agency* (2009). doi:10.1016/S0140-6736(10)60785-4
300. Kalxdorf, M., Gade, S., Eberl, H. C. & Bantscheff, M. Monitoring Cell-surface N-Glycoproteome Dynamics by Quantitative Proteomics Reveals Mechanistic Insights into Macrophage Differentiation. *Mol. Cell. proteomics* **16**, 770 LP-785 (2017).
301. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
302. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately

## References

- downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
303. McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv Math. Comput. Sci.* **abs/1802.0**, (2018).
304. Müller, T. *et al.* Automated sample preparation with SP3 for low-input clinical proteomics. *Mol. Syst. Biol.* **16**, 1–19 (2020).
305. Müller, T. *et al.* Automated sample preparation with SP3 for low-input clinical proteomics. *bioRxiv* 703413 (2019). doi:10.1101/703413
306. Erde, J., Loo, R. R. O. & Loo, J. a. Enhanced FASP (eFASP) to increase proteome coverage and sample recovery for quantitative proteomic experiments. *J. Proteome Res.* **13**, 1885–95 (2014).
307. Sielaff, M. *et al.* Evaluation of FASP, SP3 and iST Protocols for Proteomic Sample Preparation in the Low Microgram Range. *J. Proteome Res.* [acs.jproteome.7b00433](https://doi.org/10.1021/acs.jproteome.7b00433) (2017). doi:10.1021/acs.jproteome.7b00433
308. Doellinger, J., Schneider, A., Hoeller, M. & Lasch, P. Sample Preparation by Easy Extraction and Digestion (SPEED) - A Universal, Rapid, and Detergent-free Protocol for Proteomics based on Acid Extraction. *bioRxiv* 393249 (2018). doi:10.1101/393249
309. Hulbert, A. J. & Else, P. L. Evolution of mammalian endothermic metabolism: Mitochondrial activity and cell composition. *Am. J. Physiol. - Regul. Integr. Comp. Physiol.* **256**, 63–69 (1989).
310. Botelho, D. *et al.* Top-down and bottom-up proteomics of SDS-containing solutions following mass-based separation. *J. Proteome Res.* **9**, 2863–2870 (2010).
311. Rundlett, K. L. & Armstrong, D. W. Mechanism of signal suppression by anionic surfactants in capillary electrophoresis-electrospray ionization mass spectrometry. *Anal. Chem.* **68**, 3493–3497 (1996).
312. Erich, K. *et al.* Spatial Distribution of Endogenous Tissue Protease Activity in Gastric Carcinoma Mapped by MALDI Mass Spectrometry. *Mol. Cell. Proteomics* 151–161 (2018). doi:10.1074/mcp.RA118.000980
313. Cleland, T. P. Human Bone Paleoproteomics Utilizing the Single-Pot, Solid-Phase-Enhanced Sample Preparation Method to Maximize Detected Proteins and Reduce Humics. *J. Proteome Res.* **17**, 3976–3983 (2018).
314. Virant-Klun, I., Leicht, S., Hughes, C. & Krijgsveld, J. Identification of Maturation-Specific Proteins by Single-Cell Proteomics of Human Oocytes. *Mol. Cell. Proteomics* **15**, 2616–27 (2016).
315. Pellegrini, Davide, Grosso, Ambra del, Angella, Lucia, Giordano, Nadia, D. & Ilaria, Tonazzini, Marialaura, Caleo, Matteo, Cecchini, Marco, McDonnell, L. A. Quantitative Microproteomics Based Characterization of the Central and Peripheral Nervous System of a Mouse Model of Krabbe Disease. *Mol. Cell. Proteomics* 1–57 (2019).
316. DiIillo, M. *et al.* Mass Spectrometry Imaging, Laser Capture Microdissection, and LC-MS/MS of the Same Tissue Section. *J. Proteome Res.* **16**, 2993–3001 (2017).
317. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
318. Wang, H. *et al.* Systematic Optimization of Long Gradient Chromatography Mass Spectrometry for Deep Analysis of Brain Proteome. *J. Proteome Res.* **14**, 829–838 (2015).
319. Yang, F., Shen, Y., Camp 2nd, D. G. & Smith, R. D. High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. *Expert Rev. Proteomics* **9**, 129–134 (2012).
320. Johnson, D., Boyes, B. & Orlando, R. The use of ammonium formate as a mobile-phase modifier for LC-MS/MS analysis of tryptic digests. *J. Biomol. Tech.* **24**, 187–197 (2013).
321. Stadlmann, J. *et al.* Improved Sensitivity in Low-Input Proteomics Using Micropillar Array-Based Chromatography. *Anal. Chem.* **91**, 14203–14207 (2019).
322. Lim, M. Y., Paulo, J. A. & Gygi, S. P. Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *J. Proteome Res.* **18**, 4020–4026 (2019).
323. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12**, R1 (2011).
324. Murillo, J. R. *et al.* Automated phosphopeptide enrichment from minute quantities of frozen malignant melanoma tissue. *PLoS One* **13**, 1–15 (2018).
325. Geyer, P. E. *et al.* Proteomics reveals the effects of sustained weight loss on the human plasma proteome. *Mol. Syst. Biol.* **12**, 901 (2016).
326. Russell, P. Grant, Andrew, N. H. From Lost in Translation to Paradise Found: Enabling Protein Biomarker Method Transfer Using Mass Spectrometry. *Clin Chem.* **60**, 941–944 (2014).
327. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
328. Post, H. *et al.* Robust, Sensitive, and Automated Phosphopeptide Enrichment Optimized for Low Sample Amounts Applied to Primary Hippocampal Neurons. *J. Proteome Res.* **16**, 728–737 (2017).
329. Dagley, L. F., Infusini, G., Larsen, R. H., Sandow, J. J. & Webb, A. I. Universal Solid-Phase Protein Preparation (USP3) for bottom-up and top-down proteomics. *J. Proteome Res.* **18**, [acs.jproteome.9b00217](https://doi.org/10.1021/acs.jproteome.9b00217) (2019).
330. Schweizer, L. *et al.* 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 (M020141). <https://covaris.com/> (2020). at <<https://covaris.com/wp-content/uploads/M020141.pdf>>
331. Spring, C., Spring, C. & Ny, H. HYPER-sol: flash-frozen results from archival FFPE tissue for clinical proteomics. (2019).
332. Buczak, K. *et al.* Spatial tissue proteomics quantifies inter- and intra-tumor heterogeneity in hepatocellular carcinoma. *Mol. Cell. Proteomics* [mcp.RA117.000189](https://doi.org/10.1074/mcp.RA117.000189) (2018). doi:10.1074/mcp.RA117.000189

333. Fang, M., Yuan, J., Peng, C. & Li, Y. Collagen as a double-edged sword in tumor progression. *Tumor Biol.* **35**, 2871–2882 (2014).
334. Hirai, K, Shimada, H, Ogawa, T, Taji, S. The spread of human lung cancer cells on collagens and its inhibition by type III collagen. *Clin Exp Metastasis* **9**, 517–27 (1991).
335. Sotgia, F. & Lisanti, M. P. Mitochondrial markers predict survival and progression in non-small cell lung cancer (NSCLC) patients: Use as companion diagnostics. *Oncotarget* **8**, 68095–68107 (2017).
336. Soda, K. The mechanisms by which polyamines accelerate tumor spread. *J. Exp. Clin. Cancer Res.* **30**, 95 (2011).
337. Giatromanolaki, Alexandra, Siviridis, Efthimios, Arelaki, Stella, Koukourakis, M. I. Expression of enzymes related to glucose metabolism in non-small cell lung cancer and prognosis. *Exp. Lung Res.* **43**, 167–174 (2017).
338. Nagpal, Jatin K, Dasgupta, Santanu, Jadallah, Sana, Chae, Young K, Ratovitski, Edward A, Toubaji, Antoun, Netto, George J, Eagle, Toby Nissan, A. & Sidransky, D. Profiling the expression pattern of GPI transamidase complex subunits in human cancer. *Mod. Pathol.* **21**, 979–991 (2008).
339. L.J., H. *et al.* Proteomic analysis of secreted proteins of non-small cell lung cancer. *Chinese J. Cancer / Ai Zheng* **25**, 1361–1367 (2006).
340. Gambardella, L. *et al.* PI3K signaling through the dual GTPase-activating protein ARAP3 is essential for developmental angiogenesis. *Sci. Signal.* **3**, ra76 (2010).
341. Laugesen, A., Højfeldt, J. W. & Helin, K. Molecular Mechanisms Directing PRC2 Recruitment and H3K27 Methylation. *Mol. Cell* **74**, 8–18 (2019).
342. Ben-Porath, I. *et al.* An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat. Genet.* **40**, 499–507 (2008).
343. Yoshida, T., Matsuda, Y., Naito, Z. & Ishiwata, T. CD44 in human glioma correlates with histopathological grade and cell migration. *Pathol. Int.* **62**, 463–470 (2012).
344. Crosby, H. A., Lalor, P. F., Ross, E., Newsome, P. N. & Adams, D. H. Adhesion of human haematopoietic (CD34+) stem cells to human liver compartments is integrin and CD44 dependent and modulated by CXCR3 and CXCR4. *J. Hepatol.* **51**, 734–749 (2009).
345. Vikesaa, J. *et al.* RNA-binding IMPs promote cell adhesion and invadopodia formation. *EMBO J.* **25**, 1456–1468 (2006).
346. Goodison, S., Urquidi, V. & Tarin, D. CD44 cell adhesion molecules. *Mol. Pathol.* **52**, 189–196 (1999).
347. Shu, C., Wang, Q., Yan, X. & Wang, J. Prognostic and microRNA profile analysis for CD44 positive expression pediatric posterior fossa ependymoma. *Clin. Transl. Oncol. Off. Publ. Fed. Spanish Oncol. Soc. Natl. Cancer Inst. Mex.* **20**, 1439–1447 (2018).
348. Chen, C., Zhao, S., Karnad, A. & Freeman, J. W. The biology and role of CD44 in cancer progression: therapeutic implications. *J. Hematol. Oncol.* **11**, 64 (2018).
349. Chung, S. S. & Vadgama, J. V. Curcumin and epigallocatechin gallate inhibit the cancer stem cell phenotype via down-regulation of STAT3-NFκB signaling. *Anticancer Res.* **35**, 39–46 (2015).
350. Sahin, I. H. & Klostergaard, J. CD44 as a drug delivery target in human cancers: where are we now? *Expert Opin. Ther. Targets* **19**, 1587–1591 (2015).
351. Menke-van der Houven van Oordt, C. W. *et al.* First-in-human phase I clinical trial of RG7356, an anti-CD44 humanized antibody, in patients with advanced, CD44-expressing solid tumors. *Oncotarget* **7**, 80046–80058 (2016).
352. Yang, Y. *et al.* The NQO1/PKLR axis promotes lymph node metastasis and breast cancer progression by modulating glycolytic reprogramming. *Cancer Lett.* **453**, 170–183 (2019).
353. Kang, J. H. *et al.* Aldehyde dehydrogenase inhibition combined with phenformin treatment reversed NSCLC through ATP depletion. *Oncotarget* **7**, 49397–49410 (2016).
354. Zhang, N. *et al.* FAM129A promotes invasion and proliferation by activating FAK signaling pathway in non-small cell lung cancer. *International journal of clinical and experimental pathology* **12**, 893–900 (2019).
355. Pahl, H. L. Activators and target genes of Rel/NF-κappaB transcription factors. *Oncogene* **18**, 6853–6866 (1999).
356. Li, P., Lv, X., Zhang, Z. & Xie, S. S100A6/miR193a regulates the proliferation, invasion, migration and angiogenesis of lung cancer cells through the P53 acetylation. *Am. J. Transl. Res.* **11**, 4634–4649 (2019).
357. Sahm, F. *et al.* The endogenous tryptophan metabolite and NAD+ precursor quinolinic acid confers resistance of gliomas to oxidative stress. *Cancer Res.* **73**, 3225–3234 (2013).
358. Khan, M. Z., He, L. & Zhuang, X. The emerging role of GPR50 receptor in brain. *Biomed. Pharmacother.* **78**, 121–128 (2016).
359. Uhlitz, F. *et al.* An immediate–late gene expression module decodes ERK signal duration. *Mol. Syst. Biol.* **13**, 928 (2017).
360. Ning, T., Cui, H., Sun, F. & Zou, J. Systemic analysis of genome-wide expression profiles identified potential therapeutic targets of demethylation drugs for glioblastoma. *Gene* **627**, 387–392 (2017).
361. Rahal, Z., Abdulhai, F., Kadara, H. & Saab, R. Genomics of adult and pediatric solid tumors. *Am. J. Cancer Res.* **8**, 1356–1386 (2018).
362. Roussel, M. F. & Stripay, J. L. Epigenetic Drivers in Pediatric Medulloblastoma. *Cerebellum* **17**, 28–36 (2018).
363. Schwermer, M. *et al.* Comprehensive characterization of RB1 mutant and MYCN amplified retinoblastoma cell lines. *Exp. Cell Res.* **375**, 92–99 (2019).
364. Muscat, A. M. *et al.* The evolutionary pattern of mutations in glioblastoma reveals therapy-mediated selection.

## References

- Oncotarget* **9**, 7844–7858 (2017).
365. Johann, P. D. *et al.* Atypical Teratoid/Rhabdoid Tumors Are Comprised of Three Epigenetic Subgroups with Distinct Enhancer Landscapes. *Cancer Cell* **29**, 379–393 (2016).
366. Pugh, T. J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat. Genet.* **45**, 279–284 (2013).
367. Pajtler, K. W. *et al.* Molecular Classification of Ependymal Tumors across All CNS Compartments, Histopathological Grades, and Age Groups. *Cancer Cell* **27**, 728–743 (2015).
368. Jünger, S. T. *et al.* CDKN2A deletion in supratentorial ependymoma with RELA alteration indicates a dismal prognosis: a retrospective analysis of the HIT ependymoma trial cohort. *Acta Neuropathol.* (2020). doi:10.1007/s00401-020-02169-z
369. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–72 (2006).
370. Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16**, 19–34 (2017).
371. Chen, J. *et al.* Identification of Secreted Phosphoprotein 1 (SPP1) as a Prognostic Factor in Lower-Grade Gliomas. *World Neurosurg.* **130**, e775–e785 (2019).
372. Zhang, R., Zhang, S., Xing, R. & Zhang, Q. High expression of EZR (ezrin) gene is correlated with the poor overall survival of breast cancer patients. *Thorac. cancer* **10**, 1953–1961 (2019).
373. Bizzarro, V. *et al.* Hypoxia regulates ANXA1 expression to support prostate cancer cell invasion and aggressiveness. *Cell Adh. Migr.* **11**, 247–260 (2017).
374. Vecchi, L. *et al.* Inhibition of the AnxA1/FPR1 autocrine axis reduces MDA-MB-231 breast cancer cell growth and aggressiveness in vitro and in vivo. *Biochim. Biophys. Acta. Mol. Cell Res.* **1865**, 1368–1382 (2018).
375. Morita, J. *et al.* Structure and biological function of ENPP6, a choline-specific glycerophosphodiester-phosphodiesterase. *Sci. Rep.* **6**, 20995 (2016).
376. Danussi, C. *et al.* An EMILIN1-negative microenvironment promotes tumor cell proliferation and lymph node invasion. *Cancer Prev. Res. (Phila.)* **5**, 1131–1143 (2012).
377. Chen, P. *et al.* Circadian Regulator CLOCK Recruits Immune-Suppressive Microglia into the GBM Tumor Microenvironment. *Cancer Discov.* **10**, 371–381 (2020).
378. Jin, Y. & Li, J.-L. Olfactomedin-like 3: possible functions in embryonic development and tumorigenesis. *Chin. Med. J. (Engl.)* **132**, 1733–1738 (2019).
379. Mohankumar, K. M. *et al.* An in vivo screen identifies ependymoma oncogenes and tumor-suppressor genes. *Nat. Genet.* **47**, 878–887 (2015).
380. Kalluri, R. & LeBleu, V. S. The biology, function, and biomedical applications of exosomes. *Science* **367**, (2020).
381. Zhang, Y., Liu, Y., Liu, H. & Tang, W. H. Exosomes: biogenesis, biologic function and clinical potential. *Cell Biosci.* **9**, 19 (2019).
382. Kalluri, R. The biology and function of exosomes in cancer. *J. Clin. Invest.* **126**, 1208–1215 (2016).
383. Pegtel, D. M., Peferoen, L. & Amor, S. Extracellular vesicles as modulators of cell-to-cell communication in the healthy and diseased brain. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **369**, 20130516 (2014).
384. Liu, W. *et al.* Role of Exosomes in Central Nervous System Diseases. *Front. Mol. Neurosci.* **12**, 240 (2019).
385. Kanninen, K. M., Bister, N., Koistinaho, J. & Malm, T. Exosomes as new diagnostic tools in CNS diseases. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1862**, 403–410 (2016).
386. Hill, A. F. Extracellular Vesicles and Neurodegenerative Diseases. *J. Neurosci.* **39**, 9269 LP-9273 (2019).
387. Saeedi, S., Israel, S., Nagy, C. & Turecki, G. The emerging role of exosomes in mental disorders. *Transl. Psychiatry* **9**, 122 (2019).
388. Théry, C. *et al.* Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines. *J. Extracell. vesicles* **7**, 1535750 (2018).
389. Yamada, J. *et al.* Purification, molecular cloning, and genomic organization of human brain long-chain acyl-CoA hydrolase. *J. Biochem.* **126**, 1013–1019 (1999).
390. Lespagnol, A. *et al.* Exosome secretion, including the DNA damage-induced p53-dependent secretory pathway, is severely compromised in TSAP6/Steap3-null mice. *Cell Death Differ.* **15**, 1723–1733 (2008).
391. Han, M. *et al.* Six-Transmembrane Epithelial Antigen of Prostate 3 Predicts Poor Prognosis and Promotes Glioblastoma Growth and Invasion. *Neoplasia* **20**, 543–554 (2018).
392. Zhang, M., Lv, X., Jiang, Y., Li, G. & Qiao, Q. Identification of aberrantly methylated differentially expressed genes in glioblastoma multiforme and their association with patient survival. *Exp. Ther. Med.* **18**, 2140–2152 (2019).
393. Jhunjhunwala, S. *et al.* Diverse modes of genomic alteration in hepatocellular carcinoma. *Genome Biol.* **15**, 436 (2014).
394. Chen, Z.-H. & Wang, L.-H. FARP1 Facilitates Cell Proliferation Through Modulating MAPK Signaling Pathway in Cutaneous Melanoma. *Am. J. Dermatopathol.* **41**, 908–913 (2019).
395. Niibori-Nambu, A. *et al.* Glioma initiating cells form a differentiation niche via the induction of extracellular matrices and integrin  $\alpha V$ . *PLoS One* **8**, e59558–e59558 (2013).
396. Wiśniewski, J. R., Ostasiewicz, P. & Mann, M. High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers. *J. Proteome Res.* **10**, 3040–3049 (2011).
397. Batth, T. S. *et al.* Protein Aggregation Capture on Microparticles Enables Multipurpose Proteomics Sample

- Preparation. *Mol. Cell. proteomics* **18**, 1027–1035 (2019).
398. Ronci, M. *et al.* Protein unlocking procedures of formalin-fixed paraffin-embedded tissues: application to MALDI-TOF imaging MS investigations. *Proteomics* **8**, 3702–3714 (2008).
399. Guo, T. *et al.* Proteome analysis of microdissected formalin-fixed and paraffin-embedded tissue specimens. *J. Histochem. Cytochem. Off. J. Histochem. Soc.* **55**, 763–772 (2007).
400. Bell, L. N. *et al.* Utility of formalin-fixed, paraffin-embedded liver biopsy specimens for global proteomic analysis in nonalcoholic steatohepatitis. *Proteomics. Clin. Appl.* **5**, 397–404 (2011).
401. Föll, M. C. *et al.* Reproducible proteomics sample preparation for single FFPE tissue slices using acid-labile surfactant and direct trypsinization. *Clin. Proteomics* **15**, 11 (2018).
402. Waas, M. *et al.* Combine and conquer: surfactants, solvents, and chaotropes for robust mass spectrometry based analyses of membrane proteins. *Anal. Chem.* **86**, 1551–1559 (2014).
403. Lindeman, N. I. *et al.* Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors. *J. Thorac. Oncol.* **8**, 823–859 (2013).
404. Waas, M., Pereckas, M., Jones Lipinski, R. A., Ashwood, C. & Gundry, R. L. SP2: Rapid and Automatable Contaminant Removal from Peptide Samples for Proteomic Analyses. *J. Proteome Res.* **18**, 1644–1656 (2019).
405. Leutert, M., Rodriguez-Mias, R. A., Fukuda, N. K. & Villén, J. Automated high-throughput proteome and phosphoproteome analysis using paramagnetic bead technology. *bioRxiv* 647784 (2019). doi:10.1101/647784
406. Cardozo, K. H. M. *et al.* Fast and low-cost detection of SARS-CoV-2 peptides by tandem mass spectrometry in clinical samples. *Res. Sq.* (2020).
407. Jayaram, S. *et al.* Identification of a Novel Splice Variant of Neural Cell Adhesion Molecule in Glioblastoma Through Proteogenomics Analysis. *OMICS* **22**, 437–448 (2018).
408. Castells, X. *et al.* Development of a predictor for human brain tumors based on gene expression values obtained from two types of microarray technologies. *OMICS* **14**, 157–164 (2010).
409. Snapkov, I. *et al.* The role of formyl peptide receptor 1 (FPR1) in neuroblastoma tumorigenesis. *BMC Cancer* **16**, 490 (2016).
410. De Graaf, E. L., Pellegrini, D. & McDonnell, L. A. Set of Novel Automated Quantitative Microproteomics Protocols for Small Sample Amounts and Its Application to Kidney Tissue Substructures. *J. Proteome Res.* **15**, 4722–4730 (2016).
411. Dante labs. *Dante Labs c/o* at <<https://www.dantelabs.com/>>
412. DNASystem. *DNASystem Inc.* (2020). at <<https://www.dnasystem.com/>>
413. Edemekong, P., Annamaraju, P. & Haydel, M. Health Insurance Portability and Accountability Act (HIPAA) [Updated 2020 Mar 29]. *Treasure Isl. StatPearls Publ.* (2020).
414. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *OJEU* **L119**, 1–88 (2016).
415. Kayaalp, M. Patient Privacy in the Era of Big Data. *Balkan Med. J.* **35**, 8–17 (2018).
416. e:Med - A systems medicine network. (2013). at <<https://www.sys-med.de/de/>>
417. European Association of Systems Medicine e.V., c/o Joint Research Center for Computational Biomedicine. (2015). at <<https://easym.eu/legal-notice/>>





## 7. Acknowledgments

In the sense of Aristotle's famous quote, this thesis is more than the sum of its words and chapters. It is the result of contributions from my present and former colleagues, more or less successful collaborations, as well as my dear friends and family. Here, I would like to take the opportunity to thank you all for your support.

*"The totality is not, as it were, a mere heap, but the whole is something besides the parts."*

*Aristotle: Book VIII, 1045a.8–10*

*Firstly*, I would like to thank Jeroen Krijgsveld for his supervision and interest in my PhD thesis project. From the beginning, Jeroen was available and eager to support and discuss any progress, no matter how little it was. More than once, he has offered me the opportunity to grow as a scientist through a number of collaborations and the chance to share our work outside the lab. Thank you for your patience and trust.

*Secondly*, I would like to thank my thesis advisory and defense committee members Prof. Ursula Klingmueller, Prof. Britta Bruegger, Prof. Carsten Hopf, and Prof. Stefan Wiemann for their interest in my work and valuable feedback in the past years. Thank you.

Furthermore, I would like to thank all my present and former colleagues of the AG Krijgsveld (B230). Each of you has contributed in one way or another to a unique work environment that allowed me to enjoy most of the time. In particular, I want to thank Gertjan Kramer, Gianluca Sigismondo, and Dimitris Papageorgiou. Undoubtful, all of you have a significant share in my development as a mass spec scientist. Thank you guys for your patience and willingness to share your experiences and to offer a helping hand when the liquid junction or mosquito interferences were approaching our frustration tolerance.

A big hug and thank you goes to Mathias Kalxdorf, without whom this work would not have been possible. You definitely deserve the title as Speedy Mathias: the fastest data analyzer in Heidelberg. Jokes apart, I am grateful for all your help and contribution.

I would like to thank a number of collaborators, including Jens Huebner, Kendra Maaß, Mieke Roosen, Marcel Kool, Kristian Pajtler, Rémi Longuespée, Daniel Kazdal, and Albrecht Stenzinger.

## Acknowledgments

Also, I would like to thank my former supervisors, Yansheng Liu and Dominic Winter, who continue to support my scientific career with critical feedback and valuable discussions. I am glad that we are friends beyond the world of proteomics.

My special gratitude goes to my family. You all have supported me throughout my studies and scientific careers in one way or another. There is no doubt that I would have never reached this point without you. I cannot describe how thankful I am to have you in my life, and being able to come home after a long week of work has been and is something to look forward to, every time. Nothing is better than a long evening in the backyard with your friends, family, and great neighbors.

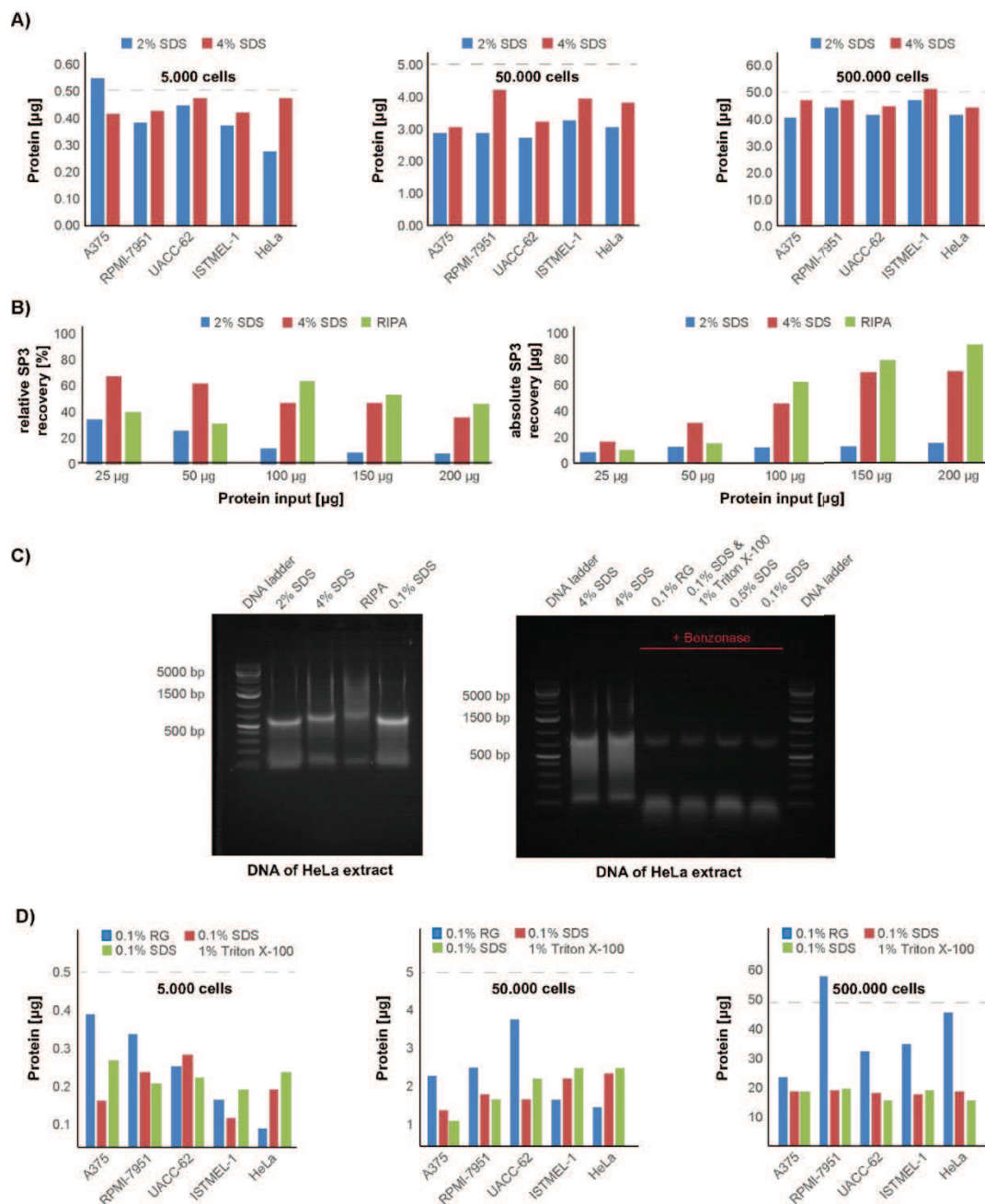
*Finally*, I am deeply thankful to Lorena for being there with me. Even though you failed here and there to give me the freedom to focus on my PhD, you never failed to distract me from work when it was really necessary. For this, I am very thankful, even if that is not always very obvious. Thank you.

## 8. Table of Figures

Figure 1: Evaluation of cell- and tissue lysis, followed by protein extraction. ....	63
Figure 2: Evaluation of high-throughput cell- and tissue lysis methods followed by protein extraction. ....	66
Figure 3: Evaluation and optimization of single-pot solid-phase-enhanced sample preparation (SP3). ....	69
Figure 4: Optimization of liquid chromatography (LC)-setup for increased peak capacity. ....	72
Figure 5: Two-proteome model for the evaluation of matching-between-runs in MaxQuant. ....	77
Figure 6: A schematic overview of the automated single-pot solid-phase-enhanced sample preparation (autoSP3) workflows. ....	81
Figure 7: autoSP3 reproducibility and proof of absent cross-contamination. ....	83
Figure 8: Longitudinal assessment of autoSP3 performance and reproducibility. ....	85
Figure 9: Evaluation of autoSP3 sensitivity. ....	89
Figure 10: End-to-end proteome profiling using ultrasonication interfaced with autoSP3. ....	91
Figure 11: Proteome profiling of tumor growth patterns of pulmonary Adenocarcinoma (ADC) FFPE tissue. ....	94
Figure 12: Molecular classification of ependymoma (EPN) tumors and sub-subclassification. ....	99
Figure 13: Ependymoma (EPN) subgroup-specific protein expression and CXorf67, an intrinsic inhibitor of PRC2 in PF-EPN-A. ....	100
Figure 14: Global gene- and protein expression correlation. ....	104
Figure 15: Translation of signature genes to signature proteins. ....	107
Figure 16: Characteristic fusion protein involving C11orf95 and RELA drive oncogenic activation of NF- $\kappa$ B signaling. ....	108
Figure 17: Subgroup-specific differential expression (DE) analysis reveals signature proteins. ....	111
Figure 18: Gene- and protein expression following recurrent structural aberrations. ....	114
Figure 19: Distribution of signature proteins per subgroup on chromosomes. ....	116
Figure 20: ST-EPN-RELA cell surface-proteome sub-classification. ....	118
Figure 21: Ependymoma (EPN) extracellular vesicle (EV) cargo characterization in ST-EPN-RELA and PF-EPN-A cell lines. ....	122
Figure 22: Multi-omics factor analysis (MOFA) achieves higher resolved subgrouping. ....	125
Table 1: Summary of the observed coefficient of variation (CVs). ....	87
Table 2: Summary of top 10 significantly differential abundant proteins per Ependymoma molecular subgroup. ....	112
Supplementary Figure 1 Optimization of cell lysis and protein extraction conditions tailored for single-pot solid-phase-enhancer sample preparation (SP3). ....	159
Supplementary Figure 2: Optimization of protein binding conditions in single-pot solid-phase-enhancer sample preparation (SP3). ....	160
Supplementary Figure 3: Comparison of single-pot solid-phase-enhancer sample preparation (SP3) to in-solution digest and filter-aided sample preparation (FASP). ....	161
Supplementary Figure 4: Peptide library generation and matching-between-runs. ....	162
Supplementary Figure 5: Correlation of protein abundance and reproducible quantification. ....	163
Supplementary Figure 6: Differential expression analysis and gene-set enrichment of papillary sub-subgroups and lepidic or papillary against all others. ....	164
Supplementary Figure 7: CXorf67-mediated inhibition of PRC2 complex. ....	165
Supplementary Figure 8: Determination of top 10 signature proteins per ependymoma (EPN) subgroup. ....	166
Supplementary Figure 9: Determining differentially abundant proteins between ST-EPN-RELA and PF-EPN-A cell lines and their extracellular vesicle isolates. ....	167
Supplementary Figure 10: Candidate proteins and their expression in extracellular vesicles and the global tumor proteome. ....	168
Supplementary Figure 11: Candidate proteins and their expression in extracellular vesicles and the global tumor proteome. ....	169

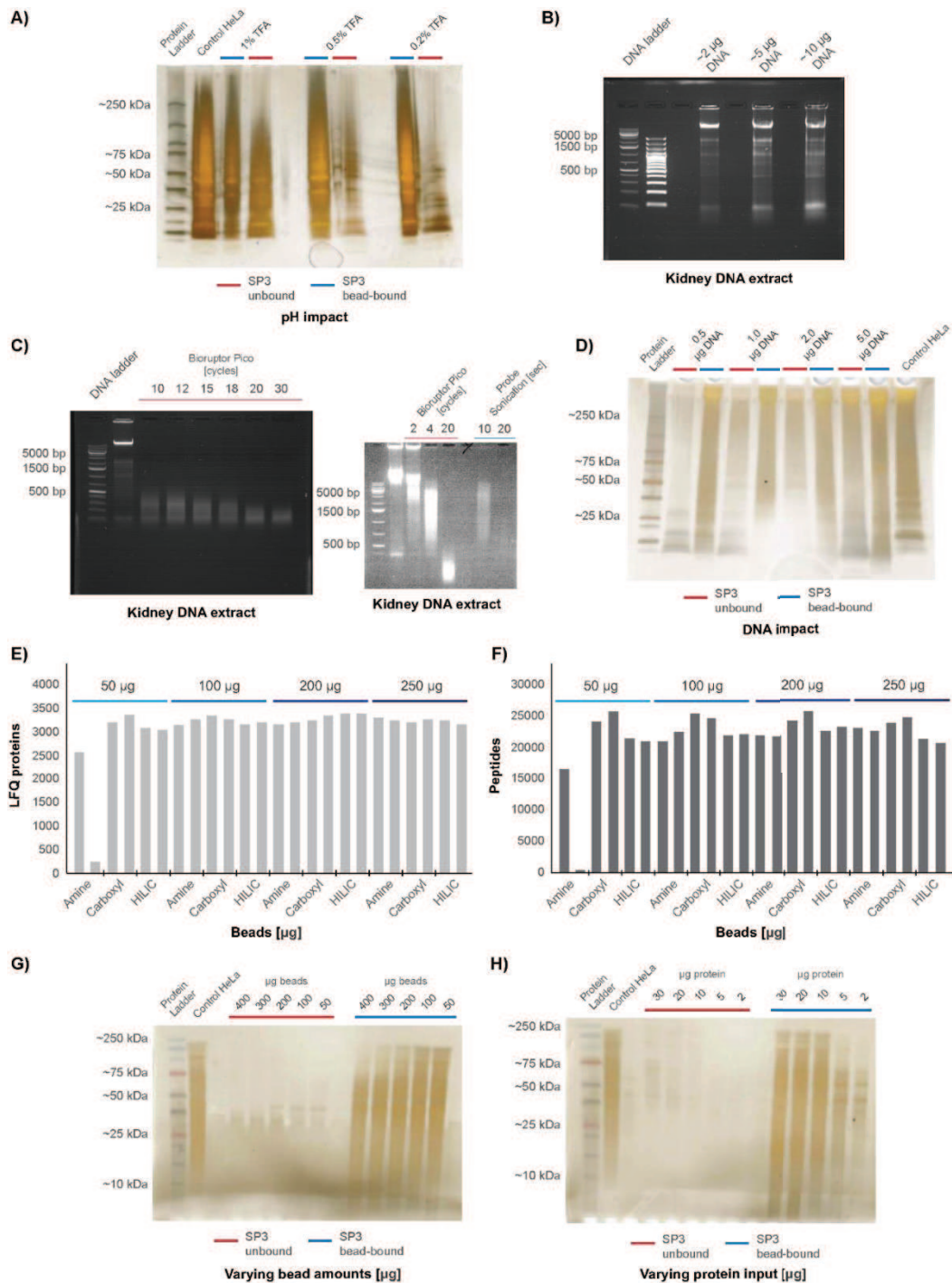


## 9. Supplementary Figures

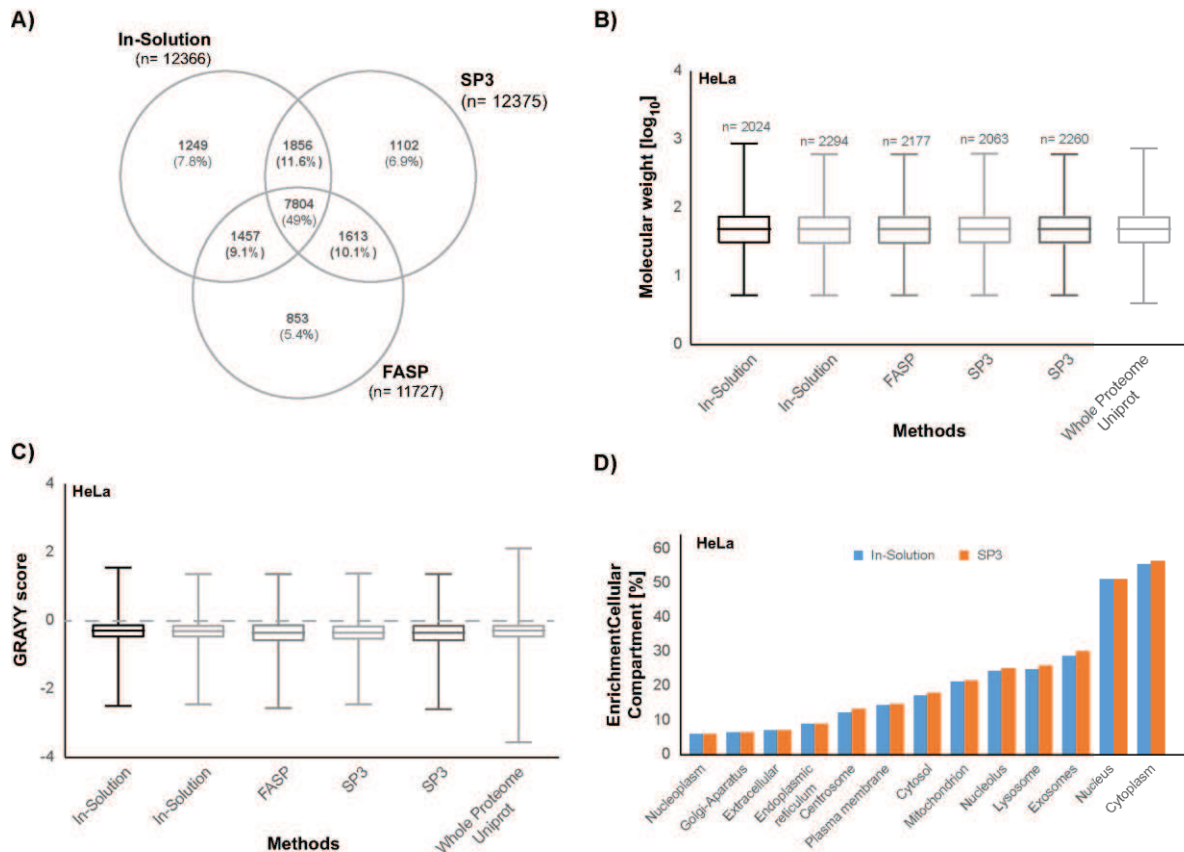


**Supplementary Figure 1: Optimization of cell lysis and protein extraction conditions tailored for single-pot solid-phase-enhancer sample preparation (SP3).** A) Lysis of 5000, 50.000, and 500.000 cells facilitated by 2% SDS (blue) and 4% SDS (red), and quantification of extracted protein mass. Five randomly selected cell lines were used (A375, RPMI-7951, UACC-62, ISTMEL-1, and HeLa). B) Mechanical disruption-free lysis using 2% SDS (blue), 4% SDS (red), or RIPA (green) in combination with SP3 to process varying amounts of protein input (25  $\mu\text{g}$ , 50  $\mu\text{g}$ , 100  $\mu\text{g}$ , 150  $\mu\text{g}$ , and 200  $\mu\text{g}$ ). The relative [%] recovery and the absolute [ $\mu\text{g}$ ] recovery after SP3 processing are shown for each condition. C) Agarose gels to highlight nucleic acid content of a HeLa lysate after mechanical disruption-free processing in different buffer compositions and with and without Benzonase treatment. For the Benzonase treatment, buffer compositions were adapted for enzyme compatibility. D) A) Lysis of 5000, 50.000, and 500.000 cells facilitated by 0.1% RapiGest (RG) (blue), 0.1% SDS (red), 0.1% SDS with 1% Triton X-100 (red), and quantification of extracted protein mass. Five randomly selected cell lines were used (A375, RPMI-7951, UACC-62, ISTMEL-1, and HeLa). A375, RPMI-7951, UACC-62, and ISTMEL-1 cells were cultured, counted, and pelleted by Dr. Gertjan Kramer.

## Supplementary Figures

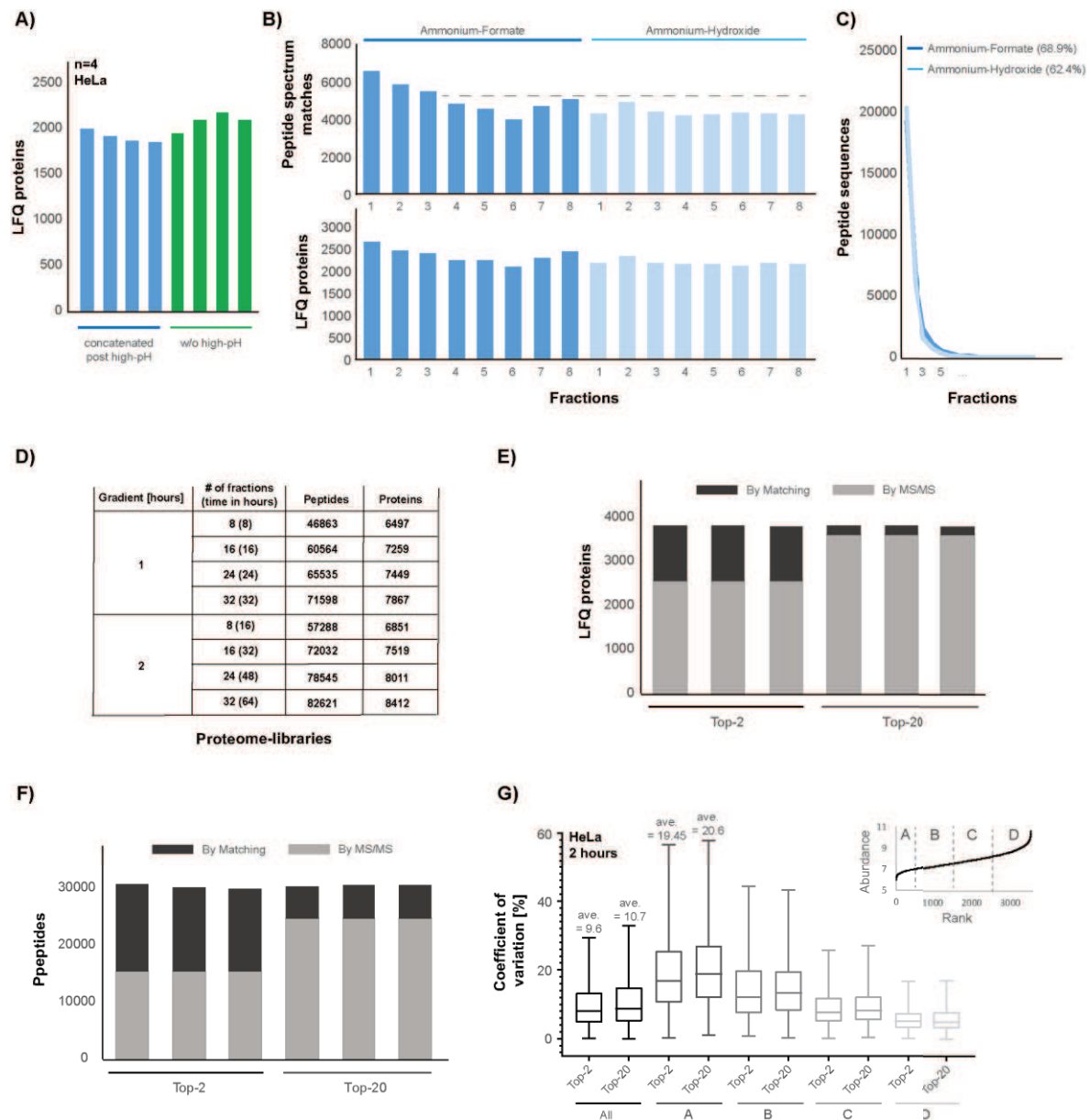


**Supplementary Figure 2: Optimization of protein binding conditions in single-pot solid-phase-enhancer sample preparation (SP3).** A) SDS-page of HeLa input and the SP3 unbound (red) and bead-bound (blue) fraction of proteins using three different acidification concentrations with TFA (1%, 0.5%, and 0.2%). B) Agarose gels to confirm DNA extraction from mouse kidney tissue. C) Agarose gels to illustrate the DNA (mouse kidney) sonication efficiency for the Bioruptor Pico and a probe sonication using different numbers of cycles or sonication times. D) SDS-page of HeLa input and the SP3 unbound (red) and bead-bound (blue) fraction of proteins using four different concentrations of spike-in DNA (mouse kidney) (0.5 µg, 1 µg, 2 µg, and 5 µg). E-F) Number of identified and quantified proteins (E) and peptides (F) with different protein input amounts (50 µg (light blue) to 250 µg (dark blue)) and different paramagnetic bead types (ReSyn Amine, classic carboxyl SP3 beads, ReSyn HILIC). G) SDS-page of constant HeLa input and the SP3 unbound (red) and bead-bound (blue) fraction of proteins using five different concentrations of SP3 beads (400 µg, 300 µg, 200 µg, 100 µg, 50 µg). H) SDS-page of varying HeLa input (30 µg, 20 µg, 10 µg, 5 µg, 2 µg) and the SP3 unbound (red) and bead-bound (blue) fraction of proteins using a constant amount of 200 µg SP3 beads.



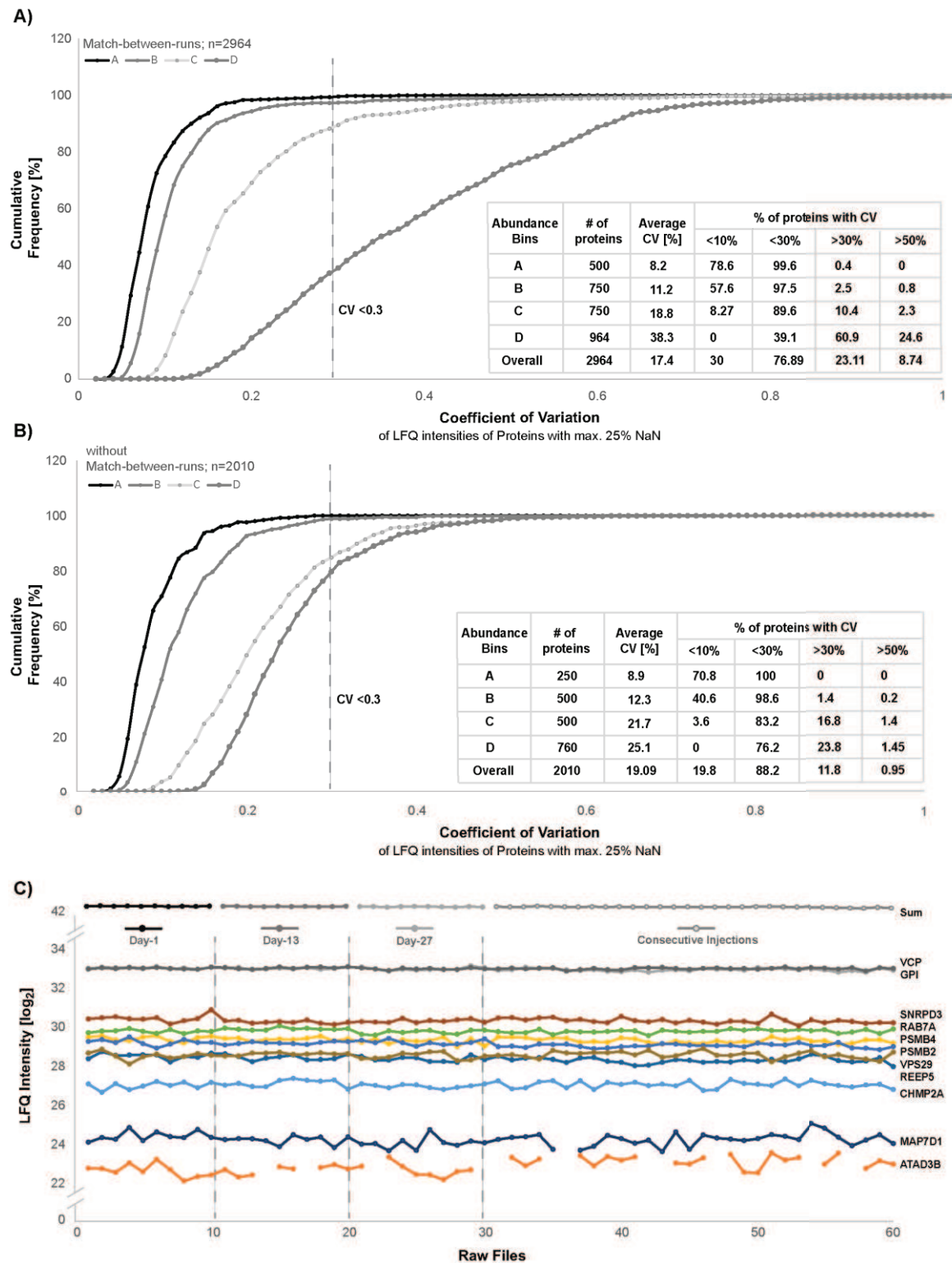
**Supplementary Figure 3: Comparison of single-pot solid-phase-enhancer sample preparation (SP3) to in-solution digest and filter-aided sample preparation (FASP).** A) Venn-diagram of identified HeLa peptides in all three methods. B) Boxplot of molecular weight distribution obtained from all three methods compared to the whole Uniprot human proteome. C) Boxplot of hydrophobicity GRAVY score distribution obtained from all three methods compared to the whole Uniprot human proteome. D) Gene ontology enrichment of cellular compartments for the in-solution digest (blue) and SP3 processed samples (orange), showing an almost identical distribution of protein origins.

## Supplementary Figures



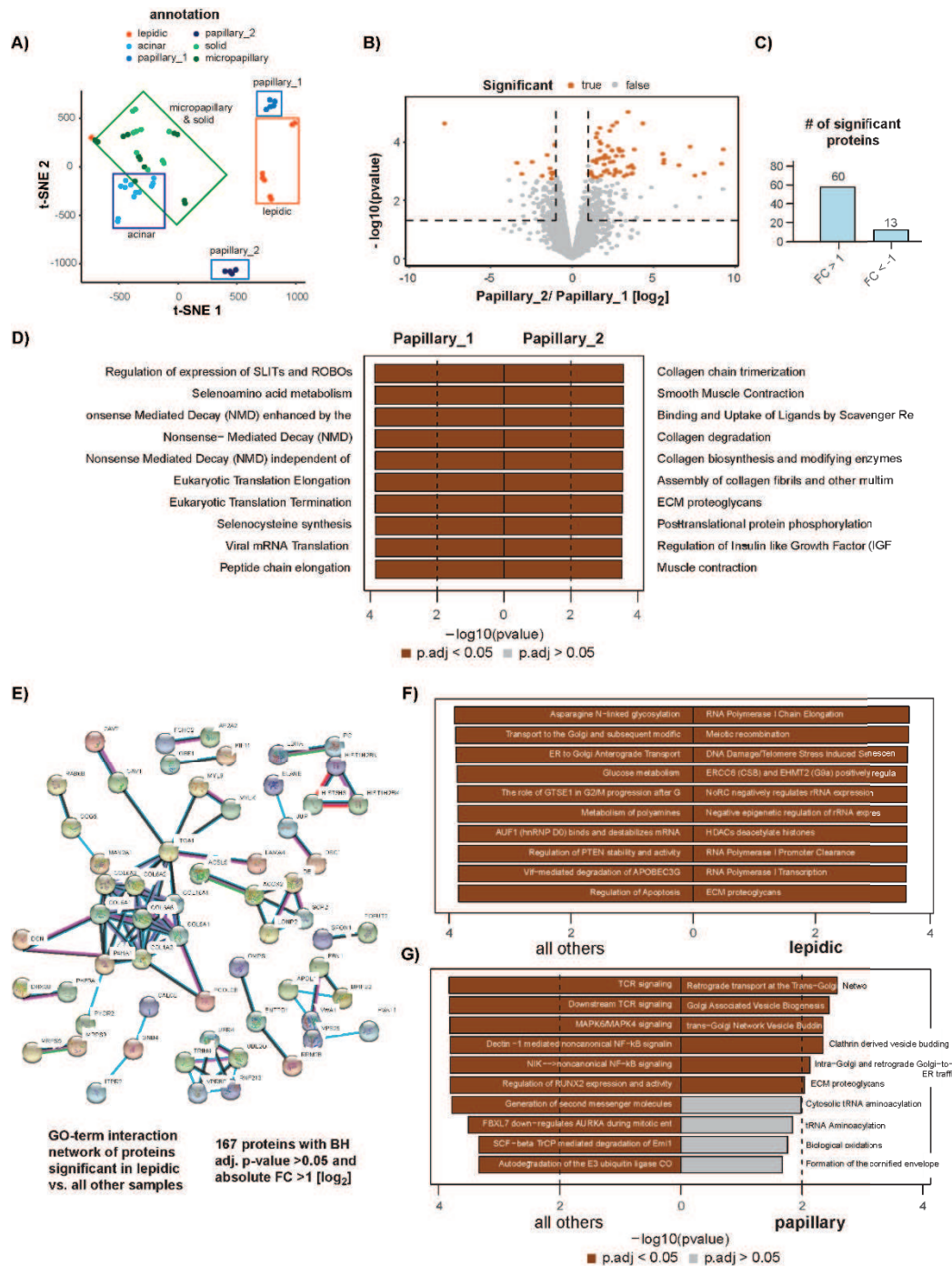
**Supplementary Figure 4: Peptide library generation and matching-between-runs.** A) Comparison of HeLa runs with and without prior high-pH fractionation and concatenation. B) Comparison of high-pH performance with eight fractions per run and with ammonium formate (dark blue) or ammonium hydroxide (light blue) buffer. The number of peptide spectrum matches (upper panel) and identified and quantified proteins (lower panel) are illustrated. C) Separation performance of high-pH runs for ammonium formate (dark blue) or ammonium hydroxide (light blue), shown by the number of fractions in which a peptide sequence can be found. D) Peptide and protein numbers achieved within different libraries using either 1-hour or 2-hours gradients and with different numbers of high-pH fractions (8, 16, 24, 32). E) Comparison of identified and quantified proteins with either matching-between runs (black) or by MS<sup>2</sup> (grey) using a Top-2 or Top-20 method. F) Comparison of numbers of identified peptides with either matching-between runs (black) or by MS<sup>2</sup> (grey) using a Top-2 or Top-20 method. G) Comparison of the coefficient of variation (CV) [%] distribution using a Top-2 or Top-20 method and binned according to the average protein abundance (A (lowest), B, C, and D (highest)).



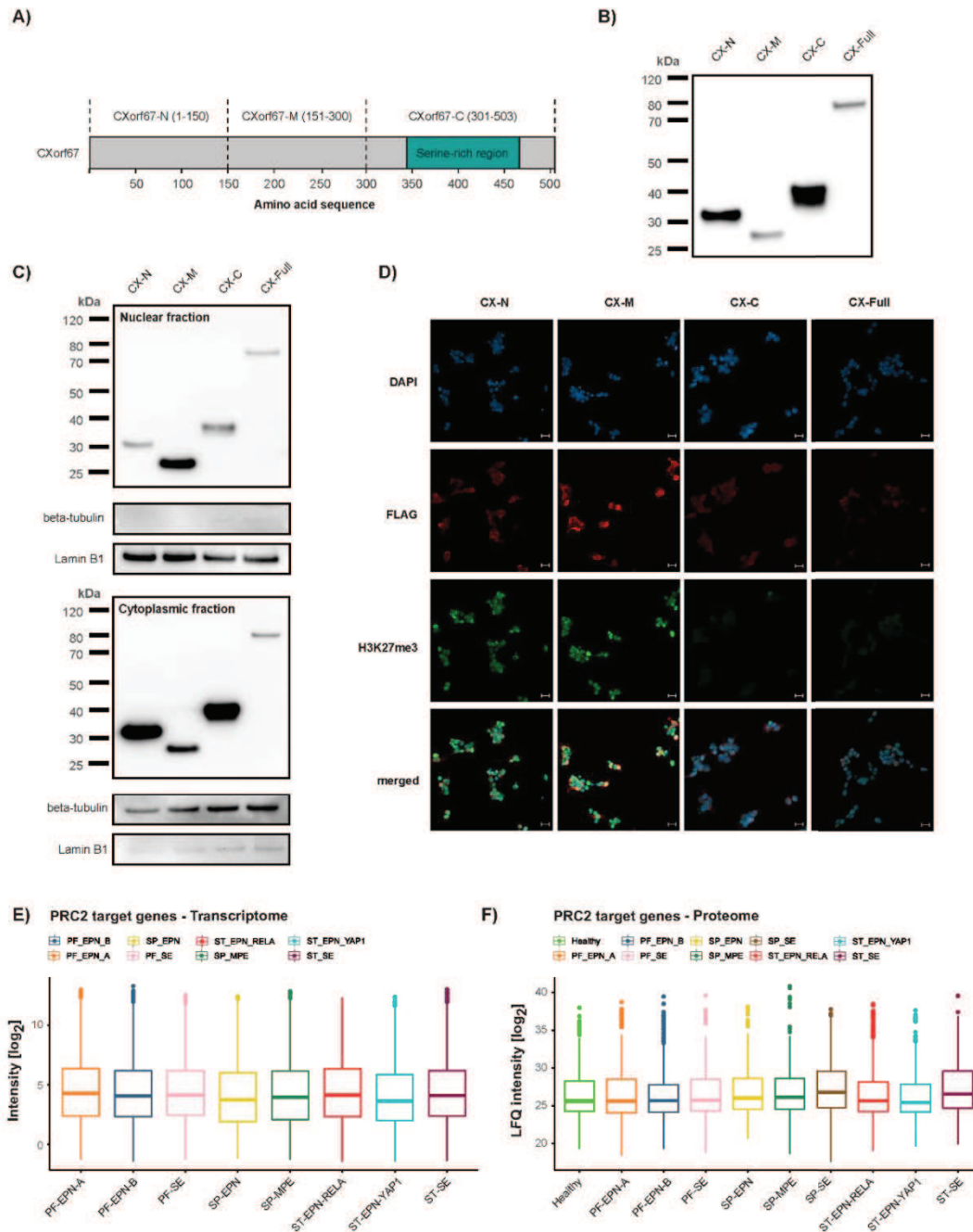


**Supplementary Figure 5: Correlation of protein abundance and reproducible quantification.** A) Four protein abundance bins (A, B, C, and D) were defined and cumulative frequency distributions [%] of the calculated CVs of quantified proteins (including match-between-runs) within each bin are plotted. The corresponding average CV values per group are shown. The table summarizes the percentage of quantified proteins observed with a CV higher or lower than 10%, 30%, and 50% for each abundance bin. B) Same as in A, the data are plotted without the use of match-between-runs. C)  $\log_2$  LfQ intensities of selected individual proteins and the sum of all proteins within a sample are plotted across all 60 measurements. C) Illustration of variation of manually selected housekeeping proteins across the entire protein abundance range and across all 60 raw files. Modified from Mueller et al., *Mol. Syst. Biol.*, 2020.

## Supplementary Figures

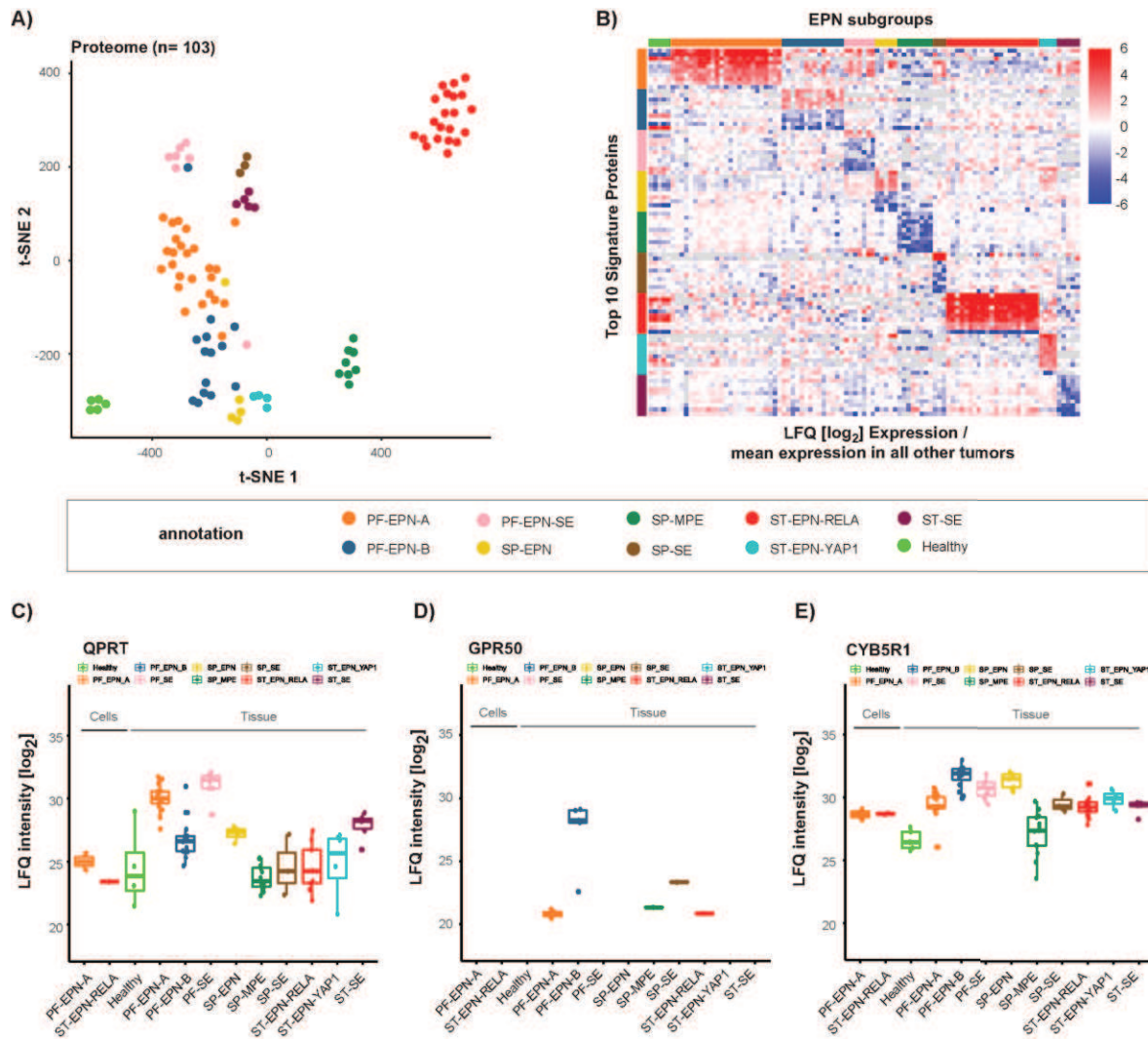


**Supplementary Figure 6: Differential expression analysis and gene-set enrichment of papillary sub-subgroups and lepidic or papillary against all others.** A) t-distributed stochastic neighbor embedding (t-SNE) analysis of the proteome data corrected via a linear regression model. B) Differential expression analysis between subclusters papillary\_1 and papillary\_2 (see A) using Limma moderated t-statistics. Proteins passing significance thresholds of  $-\log_{10} \text{p-value} < 0.05$  (Benjamini-Hochberg adjusted) and an absolute  $\log_2$  fold change  $> 1$  are highlighted in orange. C) The number of differentially expressed proteins in the papillary subcluster comparison. D) Gene set enrichment analysis of p-value ranked proteins for papillary\_1 versus papillary\_2. Gene sets with an adjusted  $-\log_{10} \text{p-value} < 0.05$  were considered significant and are highlighted in dark color. E) STRING network analysis of the 167 significant proteins ( $-\log_{10} \text{p-value} < 0.05$  and an absolute  $\log_2$  fold change  $> 1$ ) in lepidic versus all other samples. F) Gene set enrichment analysis of p-value ranked proteins for lepidic versus all other samples. G) Gene set enrichment analysis of p-value ranked proteins for papillary versus all other samples. In both GSEA analyses, gene sets with an adjusted  $-\log_{10} \text{p-value} < 0.05$  were considered significant and are highlighted in dark color. Modified from Mueller et al., *Mol. Syst. Biol.*, 2020.



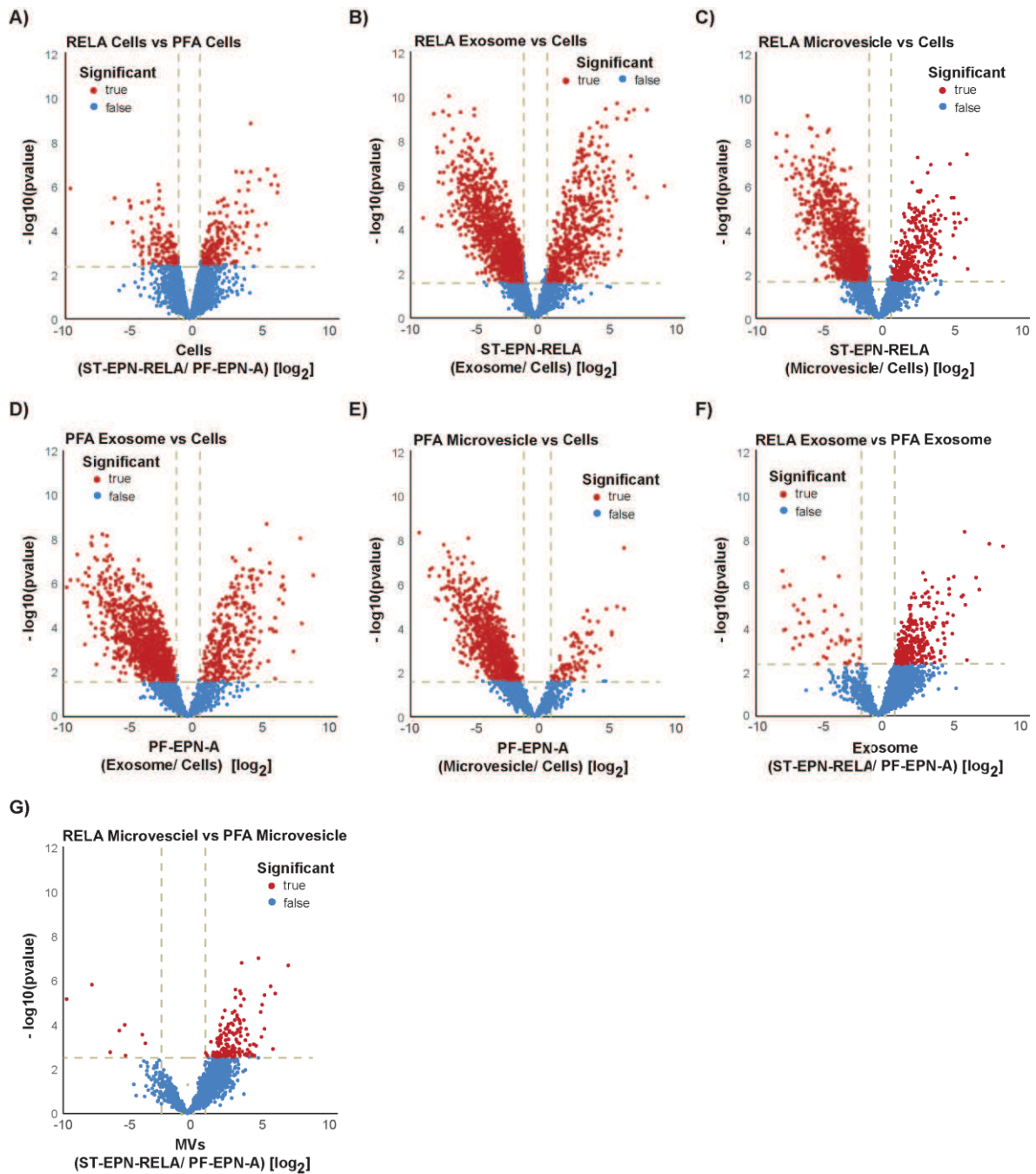
**Supplementary Figure 7: CXorf67-mediated inhibition of PRC2 complex.** A) Schematic illustration of full-length CXorf67 and the three deletion mutants: amino acids 1 to 150 (CX-N), II) amino acids 151 to 300 (CX-M), and III) amino acids 301 to 503 (CX-C). B) Western blot confirming the selective expression of deletion mutants or full-length CXorf67. C) Western blot confirming the additional localization to the nucleus. D) Staining of transduced cell lines for the presence of Flag-tagged proteins (CXorf67, CX-N, CX-M, and CX-C) as well as H3K27me3 mark. E-F) Global expression of PRC2 target genes per subgroup on the transcriptome-level (E) and proteome-level (F). The experiments illustrated in panel A to D were performed by Dr. Jens Huebner. The transcriptome data were provided by our collaborators from Pajtler et al., 2015. Panel A-D were modified from Huebner et al., *Neurooncology*, 2019.

## Supplementary Figures



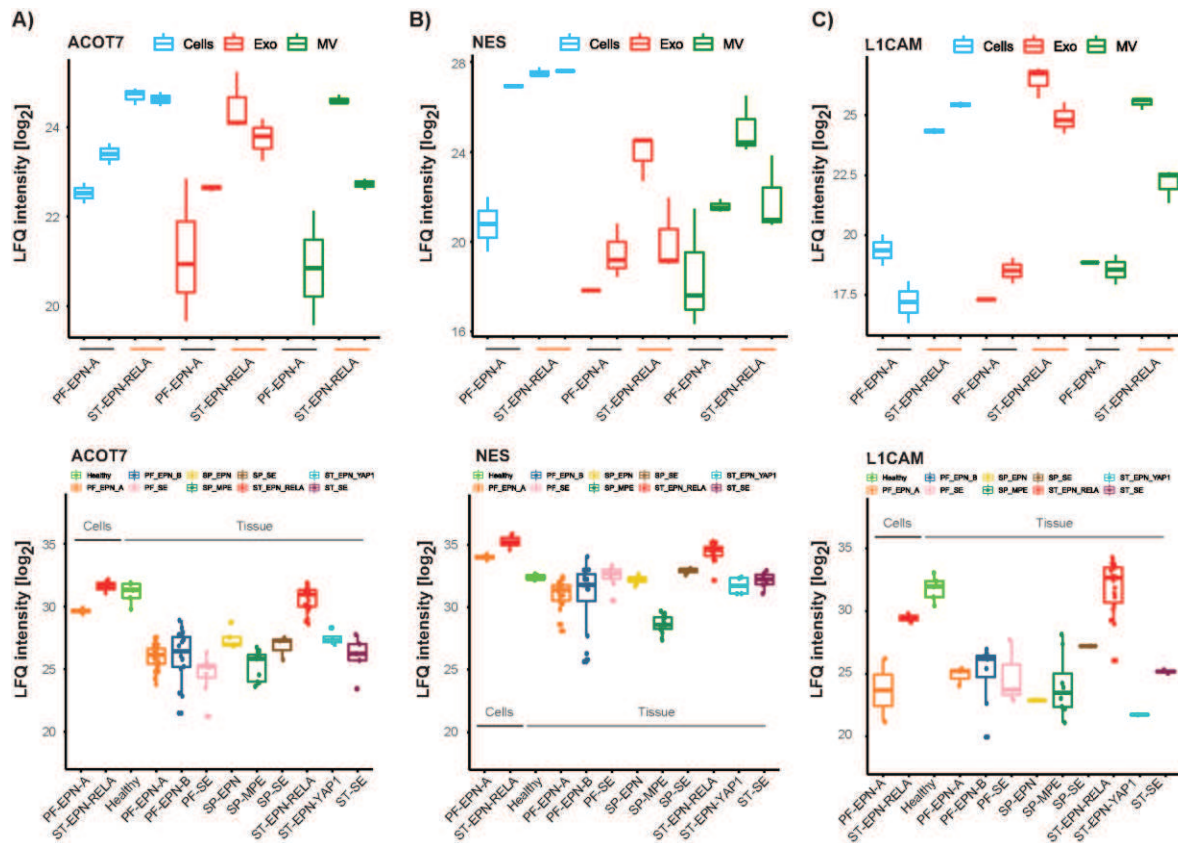
**Supplementary Figure 8: Determination of top 10 signature proteins per ependymoma (EPN) subgroup.** A) t-distributed stochastic neighbor embedding (t-SNE) analysis of the top 10 signature proteins determined by a differential (DE) expression analysis. B) Heatmap illustration of the top 10 signature proteins per EPN subgroup. LFQ expression values [ $\log_2$ ] are shown as a ratio to the mean expression in all other tumors. C-E) Boxplot illustration of QPRT (C), GPR50 (D), and CYB5R1 (E) protein expression across all EPN subgroups in the global tumor proteome.



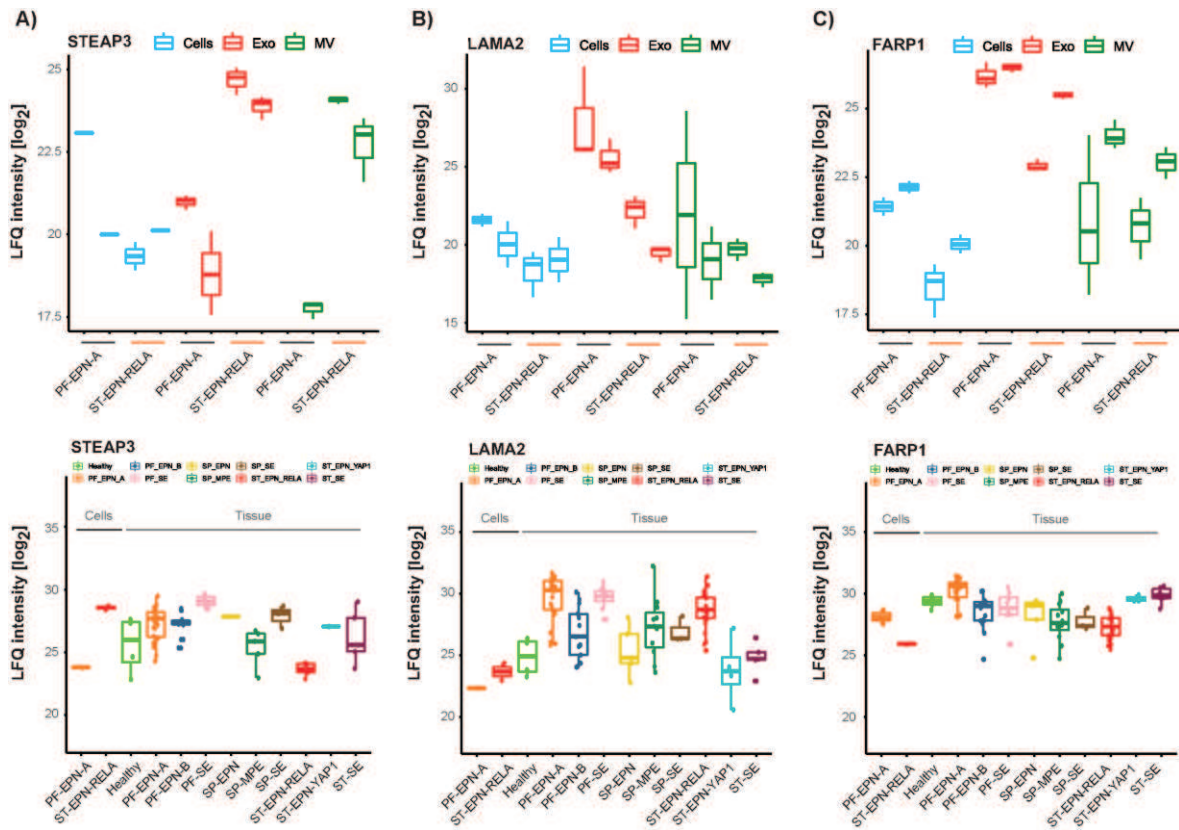


**Supplementary Figure 9: Determining differentially abundant proteins between ST-EPN-RELA and PF-EPN-A cell lines and their extracellular vesicle isolates.** A-G) Differential expression (DE) analysis using Limma moderated t-statistics for the comparison of ST-EPN-RELA and PF-EPN-A cell lines, their extracellular vesicles (exosomes and microvesicles), and between the subgroup-specific vesicle fractions. Proteins are significant at a threshold of  $-\log_{10} p\text{-value} < 0.05$  (Benjamini-Hochberg adjusted), and an absolute  $\log_2$  fold change of  $>1$  are highlighted.

## Supplementary Figures



**Supplementary Figure 10: Candidate proteins and their expression in extracellular vesicles and the global tumor proteome.** A-C) Boxplot illustration of ACOT7 (A), NES (B), and L1CAM (C) protein expression across their cell lines (blue) and extracellular vesicles (exosomes (red) and microvesicle (green)), and across all EPN subgroups in the global tumor proteome.



**Supplementary Figure 11: Candidate proteins and their expression in extracellular vesicles and the global tumor proteome.** A-C) Boxplot illustration of STEAP3 (A), LAMA2 (B), and FARP1 (C) protein expression across their cell lines (blue) and extracellular vesicles (exosomes (red) and microvesicle (green)), and across all EPN subgroups in the global tumor proteome.