From Department of Oncology and Pathology
Karolinska Institutet, Stockholm, Sweden

AT THE INTERSECTION OF PROTEOMICS AND PATHOLOGY: APPLICATION OF MASS SPECTROMETRY-BASED PROTEIN QUANTIFICATION TO HISTOPATHOLOGY AND ANTIBODY VALIDATION

Fabio Socciarelli

Karolinska Institutet

Stockholm 2021

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2021

ISBN 978-91-8016-401-6 Cover illustration: Multiplex immunohistochemistry of usual ductal hyperplasia (UDH) and surrounding invasive carcinoma of the breast (blue Hoechst dye, green MET, red EGFR)

At the intersection of proteomics and pathology: application of mass spectrometry-based protein quantification to histopathology and antibody validation
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Fabio Socciarelli, M.D.

The thesis will be defended in public at Samuelssonsalen,Tomtebodavägen 6, Karolinska Institutet Campus Solna, 10th December 2021 at 9:00 am

Principal Supervisor:
Dr. Henrik J. Johansson, PhD
Karolinska Institutet
Department of Oncology and Pathology
Laboratory of Cancer Proteomics Mass Spectrometry

Co-supervisor(s):
Prof. Janne Lehtiö, PhD
Karolinska Institutet
Department of Oncology and Pathology
Laboratory of Cancer Proteomics Mass Spectrometry

Prof. Johan Hartman, MD PhD
Karolinska Institutet
Department of Oncology and Pathology
Laboratory of Precision Pathology and Tumor Heterogeneity

Opponent:
Dr. William Howat, PhD
Abcam
Department of Validation and Technical Quality

Examination Board:
Dr. Anna Klemm, PhD
Uppsala University
Department of Information Technology
Laboratory of Quantitative Microscopy

Prof. Per Karlsson, MD PhD
University of Gothenburg
Department of Oncology
Division of Oncology

Prof. Bernd Thiede, PhD
University of Oslo
Department of Biosciences
Section for Biochemistry and Molecular Biology

To my wife Angela and my little daughter Olivia, thanks for every moment together and for supporting me through this journey

POPULAR SCIENCE SUMMARY OF THE THESIS

My name is Fabio Socciarelli and I work in the research lab of Prof. Janne Lehtiö, at Karolinska Institutet, SciLifeLab, in Stockholm. The main topic of work is to investigate proteins and their changes in cancer. Proteins can be defined as the "little bricks" of life, since these molecules participate in every aspect of life: proteins constitute tissues, are involved in energy production for the cell, they participate in cell proliferation, etc. Due to their enormous importance for life, we expect them to be altered if there is a disease in a human body, for example, a cancer. Cancer can be simply defined as a disease of human tissues that have ceased to follow the rules they have been assigned. In fact, instead of being finely regulated and strictly localized, cancer tissue proliferates and moves from its normal structure, invading the surrounding tissues and localizing even away from the original position. In this thesis we have adopted a particular technique to study the proteins called "mass spectrometry based protemics". Such technique makes it possible to identify proteins by breaking them in little pieces that are more simple to analyze and identify. Thanks to this approach we have been able to study the protein composition of many tumors, even if in this thesis I will focus only on breast and lung cancer.

In the first project my group and I studied the proteins of breast cancer, the most common cancer in women. Using the mass spectrometry based proteomics method we have identified around 10'000 proteins common in 45 tumors. We have found that breast cancer is not homogeneous, but there are 6 distinct types of breast cancer that show different groups of proteins with differing functions. We then searched for proteins that could be the co-target of anti-cancer drugs and found that some tumors were presenting 2 of them together, called MET and EGFR. This couple of proteins is clinically important, since therapies based on 2 or more drugs are expected to be more effective than therapies based just on a single drug. We have found that this couple is present in 2 different subtypes of breast cancer: basal-like and normal-like. In the so-called "normal-like" subtype this couple was expressed in the so-called "in situ tumor", a portion of tumor that has not invaded the breast tissue and is still confined to the normal breast structures. This is a great advantage, since a cancer therapy based on that couple could be used to kill cancer cells when they are still localized and less dangerous.

In the second project we investigated the proteins present in lung cancer, the most deadly human tumor. We isolated around 10'000 proteins across 141 tumors and we have found that there are 6 different types of lung cancer. Such protein-based classification has given results somewhat similar to a microscopy-based classification and has pointed out important differences between specific classes based on the proteins of the immune system. This finding is interesting: the immune system is the apparatus that helps us not only defeat infections, but also to get tumors under control before they spread. Looking more into such classes of tumors we have discovered that one of the two, the so-called "subtype 2" contains high levels of a protein called PD-L1, a well-known target of a drug that is commonly used for cancer immunotherapy. Such high content of PD-L1 is promising, since it is indicative

that a high percentage of subtype 2 patients will respond to the immunotherapy. Another important subtype investigated is “subtype 3”, since it contains some particular structures called “tertiary lymphoid tissue”. This class of tumors seems to have a better survival rate than other types of lung cancer, indifferent to the kind of therapy is given to the patient. Finally, subtypes 2 and 3 appeared to differ at the microscopy level, confirming a strong difference in growth.

The third project concerns a tool for evaluating the quality of reagents used in diagnostic and research called antibodies. Antibodies are molecules with the shape of a “Y” that are produced by every vertebrate and they have the property of binding to a well-defined area (“epitope”) of a specific target protein. Due to this property, antibodies are used to detect proteins for diagnostics in different kinds of human body fluids and tissues. Unfortunately many antibodies, instead of recognizing solely the epitope of the target protein, bind also other proteins that are not the target protein, working in an unselective way and making it difficult to quantify the expected protein. In our lab we decided to create a method to evaluate antibodies that is based on mass spectrometry based proteomics. The method relies on the measuring the target protein by mass spectrometry, and comparing its signal to the antibody based signal in an in-vitro model of diverse cell lines. As a result of this model we assessed 45 antibodies that showed differing degrees of specificity for the target protein, from highly specific to unspecific. We also observed that antibodies used for diagnostic use generally are of better quality than the ones used for research only, making evident how concerning the use of such reagents without evaluating their performance is.

In summary, in this thesis I have explored the potential of studying proteins with mass spectrometry and relating them to immunohistochemistry in models system and breast and lung cancer tissue structure. Based on the findings within this thesis, I believe that mass spectrometry can contribute to medicine in substantial way by detecting cancer proteins that can be drug targets, help to distinguish distinct types of tumors with different behaviours as well as be a reference to assess how diagnostic reagents work.

ABSTRACT

The paradigm of Precision Medicine relies on the molecular stratification of patients to select "the right patient for the right drug". The genomic approach has tried to answer this question, but large genomic studies have raised more questions than given answers. It has become fundamental to investigate the biological consequences of genetic mutations and how they influence the proteome. The recent development of LC-MS/MS has helped to address this task, giving the possibility of classifying tumors based on their molecular phenotype and the impact of the genome on the proteome. Among this background, the relationship between tumor histology and the proteomic (or molecular classification) is a recently recognized field that has seen its development in the last 6 years. Thanks to the availability of in-situ multiplexing techniques and deep learning tools for histology, it is now easier to connect a molecular aspect to a histological pattern.

One of the main causes for scarce reproducibility in biomedicine is the problem of insufficient antibody validation, especially for FFPE IHC. Many techniques are available for assessing specificity and selectivity of antibodies, showing different advantages and disadvantages, but only orthogonal approaches can give an antibody-free measurement of a specific protein.

The main aims of this thesis have been 1) to investigate the relationship between histology and molecular classification in breast and lung tumors; 2) Develop a proteomic orthogonal validation technique to be used with FFPE IF, based on a dataset of 18 cell lines.

The **paper I** investigated a cohort of 45 breast carcinomas through proteomics, CNA, SNP, mRNA microarray and metabolomics[1]. The comparison of a proteomic-based classification with the mRNA-based PAM50 subtyping showed good agreement between the two approaches. A consensus clustering performed on a subset of highly variant proteins divided the cohort into 6 different clusters, dividing basal-like tumors in 2 groups and fusing some luminal-B with Her2-enriched tumors. A correlation matrix performed to identify the co-expression of drug target showed how MET and EGFR were present together in basal-like and normal-like tumors. An immunohistochemical study confirmed the co-expression of the two proteins in both basal- and normal-like, but with important differences. While the basal-like expressed the two molecules in the invasive component, the normal-like showed the co-expression confined to the DCIS component. A super-resolution microscopy study of both subtypes showed differences in subcellular localization and colocalization between the two.

The **paper II** described the phenotype of NSCLC in terms of molecular classification, proteogenomics, immunology of cancer and differences between different subtypes[2]. A cohort of 141 NSCLC was sampled for LC-MS proteomics, DNA panel sequencing, DNA methylation and RNA-seq. The proteomic classification of the tumors divided the lung cohort in 6 subtypes, largely correlated with the histological aspect (subtypes 1-4 were mostly LUAD, the subtype 5 was mostly composed of LNELCC and the subtype 6 almost exclusively of SqCC). A network analysis of the MS data showed that subtypes 2 and 3 were

mostly enriched in immune-related proteins and there were important differences between them (expression of CD3, CD8 and PD-L1 in the subtype 2, enriched in CD20 the subtype 3 and suggestive for TLSs). A histopathological examination confirmed the presence of TLSs in subtype 3 and IHC analysis showed high levels of PD-L1 in the subtype 2. Moreover, such immunological differences were matched with differences in histological growth pattern (solid for subtype 2, mixed for subtype 3).

The **paper III** illustrated a new method for antibody validation on FFPE IF, based on MS-based orthogonal validation[3]. An 18 cell lines dataset was created that could cover an ample part of the human proteome (diversity cell line set), and the dataset was submitted for LC-MS labelled DDA and for FFPE cell block microarray, with 3 biological replicates for each cell line. As an output of the analysis, we evaluated the reproducibility of both MS and IF data, the correlation between proteomic and AB-based signals. Additionally, we correlated the IF intensities of the 18 cell lines with the entire proteome and then ranked the distribution of correlation coefficients in relation to the target protein, as an estimate measure of specificity and selectivity of the antibody. The analysis of 45 different antibodies showed that the ranking position and the corrispective correlation coefficient were highly related to each other; majority of clones used for IVD showed higher levels of MS-AB correlation if compared to RUO, reflecting a better selection of antibodies used for diagnostic purposes.

The results described above show how a better integration of the two approaches could be useful for improving the diagnostic stratification of patients with cancer, first. The availability of molecular classification, together with a morphological evaluation, will expand the pathologist's toolbox for diagnostics. Secondary, MS-based proteomics can be used for providing a high-throughput way to validate antibodies for IHC diagnostic use and help to develop new biomarkers, in increased demand for PM.

LIST OF SCIENTIFIC PAPERS

I. Johansson HJ, **Socciarelli F**, Vacanti NM, Haugen MH, Zhu Y, Siavelis I, Fernandez-Woodbridge A, Aure MR, Sennblad B, Vesterlund M, Branca RM, Orre LM, Huss M, Fredlund E, Beraki E, Garred Ø, Boekel J, Sauer T, Zhao W, Nord S, Höglander EK, Jans DC, Brismar H, Haukaas TH, Bathen TF, Schlichting E, Naume B; Consortia Oslo Breast Cancer Research Consortium (OSBREAC), Luders T, Borgen E, Kristensen VN, Russnes HG, Lingjærde OC, Mills GB, Sahlberg KK, Børresen-Dale AL, Lehtiö J.
Breast cancer quantitative proteome and proteogenomic landscape.
Nature Communications, 2019 Apr 8;10(1):1600

II. Janne Lehtiö, Taner Arslan, Ioannis Siavelis, Yanbo Pan, **Fabio Socciarelli**, Olena Berkovska, Husen M. Umer, Georgios Mermelekas, Mohammad Pirmoradian, Mats Jönsson, Hans Brunnström, Odd Terje Brustugun, Krishna Pinganksha Purohit, Richard Cunningham, Hassan Foroughi As, Sofi Isaksson, Elsa Arbajian, Mattias Aine, Anna Karlsson, Marija Kotevska, Carsten Gram Hansen, Vilde Drageset Haakensen, Åslaug Helland, David Tamborero, Henrik J. Johansson, Rui M. Branca, Maria Planck, Johan Staaf, and Lukas M. Orre.
Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune evasion mechanisms
Nature Cancer, 2021 (accepted, in press)

III. **Fabio Socciarelli**, Georgios Mermelekas, Rui Mamede Branca, Janne Lehtiö & Henrik J. Johansson.
A method for validation of AntiBodies for immunohistochemistry using quantitative Mass Spectrometry based proteomics: ABMS
Manuscript

SCIENTIFIC PAPER NOT INCLUDED IN THE THESIS

Watanabe, M., Singhal, G., Fisher, F. M., Beck, T. C., Morgan, D. A., **Socciarelli, F.,** Mather, M. L., Risi, R., Bourke, J., Rahmouni, K., McGuinness, O. P., Flier, J. S., & Maratos-Flier, E. **Liver-derived FGF21 is essential for full adaptation to ketogenic diet but does not regulate glucose homeostasis**
Endocrine, 2020 Jan; 67(1), 95–108

CONTENTS

LIST OF ABBREVIATIONS

AB	Antibody
APM	Antigen processing and presentation machinery
AR	Androgen receptor
ASCO	American Society of Clinical Oncology
CNA	Copy number alteration
CNN	Convoluted Neural Network
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CSF	Cerebrospinal fluid
DCIS	Ductal carcinoma in situ
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DNA	Deoxyribonucleic acid
EBM	Evidence-based medicine
ESR1	Estrogen receptor alpha
FDA	Food and Drug Administration
FFPE	Formalin fixed paraffin embedded (tissue or cells)
HIER	Heat-induced epitope retrieval
HPF	High power field (40X object)
HR+	Hormone receptor positive
ICI	Immune checkpoint inhibitor
IDC	Invasive ductal carcinoma
IF	Immunofluorescence
IHC	Immunohistochemistry
ILC	Invasive lobular carcinoma
LC	Liquid chromatography
LCC	Large cell carcinoma
LC-MS	Liquid chromatography – mass spectrometry
LC-MS/MS	Liquid chromatography – tandem mass spectrometry
LNELCC	Lung neuroendocrine large cell carcinoma
LUAD	Lung adenocarcinoma

MRM	Multiple reaction monitoring
MS	Mass Spectrometry
MS1	First mass spectrometer
MS2	Second mass spectrometer
MS/MS	Tandem mass spectrometry
m/z	Mass-to-charge ratio
NGS	Next generation sequencing
NMF	Non-negative Matrix Factorization
NSCLC	Non-small cell lung cancer
OH (CCO)	Ontario Health (Cancer Care Ontario)
PAM50	Prediction Analysis of Microarray 50 (Breast cancer mRNA-based classification)
PGR	Progesterone receptor
PM	Precision medicine
PRM	Parallel reaction monitoring
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RPPA	Reverse phase protein array
RS	Recurrence Score
SCLC	Small cell lung cancer
SNP	Single nucleotide polymorphism
SNR	Signal-to-Noise Ratio
SqCC	Squamous cell carcinoma
SRM	Single reaction monitoring
TCGA	The Cancer Genome Atlas (consortium)
TKI	Tyrosin kinase inhibitor
TMA	Tumor microarray
TMB	Tumor mutation burden
TNM	Tumor Node Metastasis (AJCC cancer staging system)
TLS	Tertiary lymphoid structure
WES	Whole exome sequencing

1 INTRODUCTION

1.1 THE ENCOUNTER OF TWO WORLDS: PROTEOGENOMICS AND PATHOLOGY

The introduction of proteomics in the medical field (and more specifically in pathology) should be considered as an integral component of the paradigm shift in the contemporary medicine: the change from an Evidence-based Medicine (EBM) to a Precision Medicine (PM) platform. The difference is very simple and equally remarkable: from testing (and then treating with) a drug on a heterogeneous population, without knowing who is going to respond to it, to the pre-therapy stratification of patients that will receive the drug, in the attempt of giving "the right medicine to the right patient" (to paraphrase a famous nursing principle[4] that is now considered a precision medicine mantra[5], citing the words of Edward Abrahams), minimizing the inclusion of non-responding patients and trying to divert them to a different therapy[6].

The extraordinary biological knowledge accumulated since the completion of the Human Genome Project, together with major and ongoing technological improvements, have made the development of the contemporary tools for precision medicine possible: first DNA and RNA microarray, then Next Generation Sequencing (NGS) technology for both DNA and RNA and for DNA methylation. Availability of such techniques have paved the way for projects like TCGA, in the attempt to translate into diagnostic and therapeutic strategies the genomic alterations found in cancer. Unfortunately, the TCGA and similar consortia have not been decisive in this aim as expected: in a highly cited paper[7], the Clinical Proteomics Tumor Analysis Consortium (CPTAC) pointed out how such studies generated more questions than answers, given the objective difficulty to connect a majority of the mutations found in TCGA studies into a phenotype, a biological or biochemical function.

Mass spectrometry-based proteomics is among the most important techniques of choice for narrowing the gap between genotype and phenotype, given the possibility of studying the "final product" of genomic and transcriptomic alterations, including mutated proteins that can act as drug targets or neoantigens. This proteomic approach, called proteogenomics, has the possibility of being used as a diagnostic and predictive tool for clinical use, and thus presents itself as an important component of the PM approach[8]. In the next paragraphs I will describe several cancer proteomics examples and their integration with genomic and transcriptomic analysis in a multi-layer fashion, while below in the paragraph 1.2.2 I will discuss LC-MS proteomics.

1.1.1 Breast cancer: from the TNM to multi-omics classification

Breast cancer is the most common neoplasia in women and both sexes, according to GLOBOCAN 2020 statistics[9]. "Breast cancer" is a definition that collects different histological and clinical entities under the same name: this heterogeneity has been well-known from the clinical and histological side for many decades and has been the objective of multiple studies attempting to stratify patients by prognosis or response to therapy[10,11]. Age,

dimension of the tumor[12], positivity for metastasis to lymph- nodes[13], histological grading of the neoplasm[14], positivity for Estrogen Receptor alpha or Progesterone receptor (HR+)[15], amplification of HER2/neu receptor[16] and expression of KI67 protein[17] on IHC have been and are still used to prognosticate patients and guide therapy decisions.

A new level of knowledge came during the 2000s with mRNA microarray studies, in which cohorts of breast cancer patients were studied through unsupervised clustering and identified differences between clinical subgroups[18-20]. Such studies, even if performed in various laboratories, on different patients and with different methodologies (RNA microarray in the beginning, RNA-sequencing after) showed relatively similar results, identifying a so-called basal-like cluster (mostly negative for HR and HER2 [so-called triple negative], high Ki67), two or more "luminal" clusters (generally with expression of HR+, variable KI67 and HER2), one normal-like subgroup and a HER2-enriched cluster (often high HER2 and Ki67)[21].

The molecular classification of breast cancer brought two important consequences: a better understanding of tumor biology and an important contribution to patient stratification[21]. In fact, some multigene tests, derived from these microarray and RNA-sequencing studies, can divide patients by prognosis better than traditional prognostic factors in node-negative HR+ patients, a subgroup already well-known for being difficult to predict and manage[22].

Among the several FDA-cleared tests, Oncotype Dx has been approved for use in HR+, HER2 negative, node-negative breast cancer patients and is the only tool suggested by ASCO[23] to predict response to chemotherapy in this subgroup[24-26]. Based on the assignment of a Recurrence Score (RS), it can assign these patients to three risk groups (low, intermediate and high risk), as demonstrated in the TAILORx trial and previous studies[26-28]. Moreover, the RS can discern patients that will respond to chemotherapy (high risk) and patients that will have little benefit (low and intermediate risk in women beyond 50 years)[29]. The test seems to prognosticate the survival in patients with limited lymph-node involvement, but not with extensive nodal metastasis[27,30].

As elucidated in the precedent paragraphs, the TCGA and CPTAC consortia have performed important studies in the attempt to investigate the mutational landscape and establish connections between the genome expression and its impact on the proteome, before adopting an affinity proteomics approach and then employing the LC-MS based analysis.

The first TCGA paper described the presence of genomic alterations and their distributions across different PAM50 subtypes, the impact of copy number alterations (CNA) on the transcriptome in cis and trans, the clustering based on DNA and its relationship with mRNA data, the classification based on reverse phase protein array (RPPA) in 2 different clusters and the multi-modal classification in 4 subtypes[31]. This TCGA study confirmed the 4-cluster based molecular classification of breast cancer, as had been already proposed in the mRNA-based subtyping studies cited above.

Changing perspective and consortium of study, in 2016 Philipp Mertins and collaborators[32] published the first CPTAC study of breast cancer proteogenomics, analyzing with LC-MS the proteome and the phosphoproteome of a subset of a TCGA breast cancer cohort. Among the voluminous amount of data produced, the researchers performed an unsupervised clustering of the proteomic data and defined 3 different clusters ("basal-enriched", "luminal-enriched" and "stromal-enriched"). In contrast, they were able to recapitulate the PAM50 classification with proteomic quantification. The same consortium published a new article after 4 years with a different cohort of 122 patients including whole exome sequencing (WES), CNA, RNA-seq, proteomics, phosphoproteomics and acetylproteomics[33]. Among the numerous contributions this paper provides, the multi-omics classification using Non-negative Matrix Factorization (NMF) showed good agreement with the PAM50 classification and, comparing to PAM50 itself, highlighted a subgroup of Luminal-A tumors with more aggressive behaviour. Among other important results, the investigation of proteomics of tumor immunity and the proteogenomics of ERBB2+ tumors, clarifying some aspects related to the PAM50 HER2-enriched tumors.

Besides the TCGA and CPTAC projects, many other groups have tried to explore the mutational landscape of breast cancer and the consequences on its phenotype[34,35]: among them, our laboratory has contributed to this topic in a significant way, as described in the Results section of the present thesis.

1.1.2 Lung cancer targeted therapy and LC-MS proteomics

Non-Small Cell Lung Cancer (NSCLC) is among the most common neoplasms and it is considered the second deadliest tumor according to the GLOBOCAN 2020 statistical report[9]. Remarkable progresses have been achieved over the last 20 years in improving the diagnosis and therapy of this disease[36], but the 5-year survival rate is still among the lowest in the oncological field (25% for all stages[37]). Smoking is by far the principal causative agent and it influences the histological and genomic characteristics of lung tumors[38].

NSCLC is a heterogeneous group of diseases that show differences in terms of histology, molecular classification and clinical behavior. The three main histological subtypes of NSCLC are adenocarcinoma (LUAD), squamous cell carcinoma (SqCC) and large cell carcinomas (LCC)[39], with neuroendocrine large cell carcinoma (LNELCC) to be considered as an intermediate form between a NSCLC and a SCLC.

The actual use of tissue-based biomarkers is mainly limited to the metastatic setting (stage IV) in the daily clinical experience[40], with the exception of EGFR and PD-L1 testing for the drugs osimertinib and atezolizumab, that recently received the authorization for use in earlier stages[41,42]. Below I have summarized the biomarker testing recommendations for SqCC and LUAD, as expressed by the ASCO NSCLC molecular testing guidelines for TKI[43] and ASCO and OH (CCO) joint Guidelines of therapy for stage IV NSCLC[44,45]:

• EGFR mutation assessment on NSCLC, important for tyrosine kinase inhibitors (TKIs) therapy. Several alterations can induce the constitutive activation of its catalytic domain,

including deletions of exon 19, a missense mutation on exon 21 (L858R) and a substitution on exon 20 T790M (typical as the patients had therapy already with some TKIs)[46]. Around 27% of adenocarcinomas (and less than 9% of squamous tumors) harbor some EGFR mutation[47,48].

• KRAS mutations on 32% of adenocarcinomas (3% in squamous carcinomas), mutually exclusive with EGFR mutations; when they coexist, KRAS mutation confers resistance to EGFR TKIs[49].

• ROS1 and ALK mutations in lung adenocarcinoma account together for 10% of cases. These mutations are sensitive to second and third generation TKIs (like crizotinib)[50,51].

• Other mutated genes in adenocarcinoma like BRAF, MET (exon 14 skipping mutation), RET and NTKR gene fusions[44,52-56]. HER2 is still under investigation in order to understand its role in NSCLC.

• PD-L1 IHC testing in the context of immune checkpoint inhibitors. Even if several drugs have been proved to be effective in NSCLC, the first-line drugs approved for NSCLC are pembrolizumab (anti-PD-1)[57] and atezolizumab[58](anti-PD-L1), indicated for NSCLC without driver mutation (LUAD and SqCC pembrolizumab, LUAD only atezolizumab) with any PD-L1 IHC result[45].

The TCGA consortium studied the somatic mutations of NSCLC, performing different studies on SqCC and LUAD. The SqCC paper, published in 2012 described the mutational landscape of a cohort of 178 SqCC samples, highlighting some recurrent mutations in TP53 and other proteins, alterations of pathways like squamous differentiation, phosphatydilinositol-3-OH kinase, NFE2L2 and KEAPI, CDKN2A RB1, and proposing a transcriptomic subtyping into 4 classes[59]. A second article published in 2014 performed a similar study involving 230 LUADs describing important gene mutations (some well-known like EGFR and KRAS, other like STK11, NF1, KEAP1), describing aberrant RNA transcripts, individuating candidate driver genes and pathways like RTK/RAS/RAF pathway and identified 3 subtypes based on unsupervised clustering using mRNA[47]. A third paper studied the differences between LUAD and SqCC in terms of somatic mutations, driver alterations and neoantigen expression, highlighting new differential mutations for both subtypes (PPP3CA, DOT1L, and FTSJD1 for LUAD, RASA1 for SqCC) and finding neoepitopes in around 50% of both histotypes[60].

The CPTAC performed 2 different studies for SqCC and LUAD, in which they collected

samples for DNA, RNA-seq, microRNA, DNA methylation, proteomics, phospho- and acetylproteomics (ubiquitinomics for SqCC, too). In the LUAD study they performed NMF analysis based on 4 different data layers and identified 4 different subtypes of LUAD in a cohort of 110 samples, overlapping with the mRNA-based classification and showing different somatic mutations in different subtypes[61]. Moreover, they investigated the impact of CNA and mutations on the proteome and phosphoproteome, identified already known and

putative therapeutic targets (like PTPN11 protein) and investigated the immunology of LUAD. For the SqCC article[62] the authors performed unsupervised clustering on CNA, RNA, protein, phosphoproteome and acetylproteome, clustering the tumor in 5 subtypes. Some of thesubtypes showed correlation with different histological aspects (see the paragraph below dedicated to this topic) and all had specific differences in pathways and mutations. The investigation of the impact of CNA on mRNA and proteins in trans pointed out the effects of chromosome 3q, especially the amplification of TP63 and SOX2, on the genome. Among the important observations, the identification of Np63-low tumors as potential targets for surviving inhibition and the study of drug targets in the SOX2 pathway. Finally, a consensus clustering based on immune signatures showed 3 different immune clusters (hot, warm and cold), with differences in immune pathways and signaling.

1.1.3 High multiplexing approaches for molecular and proteomic spatial/in-situ analysis

The relationship between the histopathological spatial organization of a tumor or tissue and its molecular correlates is a field that has started to be explored only in the last 10 years. Depending on the chosen target molecule several technologies are available, giving researchers the possibility to recognized mRNAs, drugs, small metabolites, proteins, lipids, etc. I will briefly give a overview on the most interesting techniques that researchers have at their disposal:

- In the field of transcriptomics an increasingly popular approach is spatial transcriptomics[63,64], developed by Joakim Lundeberg and collaborators and applied to many biological models, including plants[65]. Such approach consists in capturing in-situ RNA, performing a reverse transcription and sequencing the cDNA with an Illumina dye sequencing. Another frequently used approach is the in-situ RNA sequencing developed by Mats Nilsson and collaborators, a parallel targeted RNA analysis that can be performed on preserved cells and tissues[66].
- Non-proteomic MALDI mass spectrometry imaging (MSI) is a technique used for in situ detection of small molecules, neurotransmitters, drug metabolites and lipids, making important contributions in the field of pharmacology and neurobiology[67].
- Proteomic MALDI mass spectrometry imaging, as stated by the name is performing a label-free proteomic analysis, extracting proteins directly from the tissue. Despite the interesting technological development of the last 5 years, it is still characterized by poor spatial resolution (around 100 μm pixel dimension) and a shallow depth of analysis[68].
- Antibody-based multiplex protein detection, with both fluorescent and MS-based detection system. IHC and mIHC are described in the Method section of this thesis, in this paragraph I am going to discuss high-multiplexing approaches. In the first case the most employed version is the CODEX commercialized by Akoya Biosciences, a platform for multiplex immunofluorescence that makes use of DNA-barcoded primary antibodies that are sequencially detected, the signal acquired and finally

bleached through sequential cycles[69]. Among the MS-based detection, important to mention the MIBI instrument of Michael Angelo[70] and the Hyperion Imaging Mass Cytometry (commercialized by Fluidigm). Both are based on the detection of metal-conjugated primary antibodies, conveying great sensitivity, negligible bleed-through and background and expanding the dynamic range of detection if compared to fluorescence[71].

1.1.4 Connecting the molecular and histopathological phenotype

The connection between the proteome of a tumor and its histopathological picture is an unexplored area of research, leaving many biologically and interesting questions unanswered, such as how different histotypes, morphological patterns and degree of differentiation are reflected in the proteome, what kind of neoantigens can be found, what kind of pathway is differentially activated, etc. The connection of morphology with genomic, transcriptomic, or proteomic characterization has been investigated in different ways, for example employing artificial intelligence for classifying tumors according to different mutations, microsatellite instability or molecular subtypes, as described by several studies[72-76]. There is no doubt that the growth of this field in the last 5 years has been also facilitated by the developments of deep learning in histopathology, considered the main "precision medicine" tool of the future pathologist. The capacity of detecting tumor cells, classifying them in different subtypes, estimating the grading of a tumor, evaluating tumor biomarkers and predicting prognosis has greatly improved, becoming close to the requirements for clinical implementation[77,78].

Some interesting questions concerning the connection between the molecular and pathology phenotype have been addressed in both breast and lung cancer, by CPTAC/TCGA studies and projects carried out in the Lehtiö lab. While the results coming from the Breast Cancer Landscape and the Lung cancer study will be discussed in the Results section of this thesis, I am going to highlight some examples from other studies.

A good example of correlation histopathology – genomic alterations is the TCGA study[79] that has focused on breast invasive lobular carcinoma (ILC), a histotype of breast cancer characterized by discohesive neoplastic cells that have a low grade appearance, Indian file pattern of growth and have a negative expression of E-cadherin[80]. This study shows how 63% of ILC harbor a somatic mutation (often truncating) of the CDH1 gene (E-cadherin), vs 2% of the rest of the breast tumors. However, if considering DNA, RNA and protein data together, almost every ILC had an alteration of CDH1, co-occurring very often with a chromosome 16q loss. Other genes were found differentially expressed between ILC and IDC Luminal-A (since great majority of ILC are Luminal-A PAM50 subtype), were FOXA1, TBX3, RUNX1, GATA3 and PIK3CA.

In lung cancer, STK11-mutated LUAD are connected to the morphological pattern of the neoplastic tissue, as described in the CPTAC study[62]. The authors trained a Convoluted Neural Network (CNN) for distinguishing tumors with STK11 mutation vs STK11 wild type, using images from The Cancer Imaging Archive[81] of 110 patients. This model shows how

STK11-mutated tumors were enriched in tumor epithelium and deprived of immune cells, while the STK11 WT showed mostly an abundant inflammatory component. Another interesting example from the CPTAC lung SqCC study, where multi-omics unsupervised clustering which divided the 108 samples into 5 different molecular subtypes. Of these 5 subtypes, 3 showed connection with tumor morphology, specifically a basaloid connection for "basal-inclusive" tumors, a myxoid connection and fibroblast infiltration for the "epithelial to mesenchymal transition-enriched" subtype and a more typical squamous connection for the "classical" cluster[62].

1.2 MASS SPECTROMETRY VS AFFINITY-BASED PROTEOMICS: HOW TO QUANTIFY A PROTEIN

The following paragraph focuses on the two main approaches to quantify multiple proteins across a relatively large number of samples (MS-based and affinity) and looking into their relationship, especially when the main aim is to validate a MS-discovered protein with an antibody-based technique. This process is finally addressed from a biomarker development perspective, as underlined in the last paragraph.

1.2.1 Mass spectrometry-based approaches

Proteomics is the global analysis of protein content of an organism, tissue or cell line[82]. While in the last 20 years several approaches have been developed to resolve the proteome of an organism even for less abundant proteins, the most common tool is mass spectrometry-based analysis. Factors like speed of analysis, good peptide separation techniques, the availability of in-silico peptide spectra datasets, good reproducibility, reasonable costs per analysis and no need of specific affinity reagents are the main reasons for choosing mass spectrometry (MS) as the common approach for proteomics[83].

While there are many ways of analyzing the protein content of a sample, the most common approach to a proteomic analysis is to extract the proteins from a sample and then digest them with a proteolytic enzyme, most commonly trypsin (so-called bottom-up approach)[84]. Afterwards, the peptides can be separated, referred to as prefractionation, before analysis with liquid chromatography – tandem mass spectrometry (LC-MS/MS). MS spectra are matched to a theoretical database to identify peptides, which are then compiled into proteins. Although analysis of intact proteins by LC-MS is possible in some cases, the large size of some proteins as well as their idiosyncratic properties (e.g., solubility) tend to make their direct analysis extremely difficult, if not borderline impossible. Peptides have chemical properties (size, solubility, ionizability) much more amenable to separation by LC and identification by MS.

Since the mass spectrometer is an instrument that measures the mass-to-charge ratio (m/z) of ionized molecules, the ionization of peptides is a fundamental step. Although different ionizing techniques are available for biological molecules, Electrospray Ionization is one of

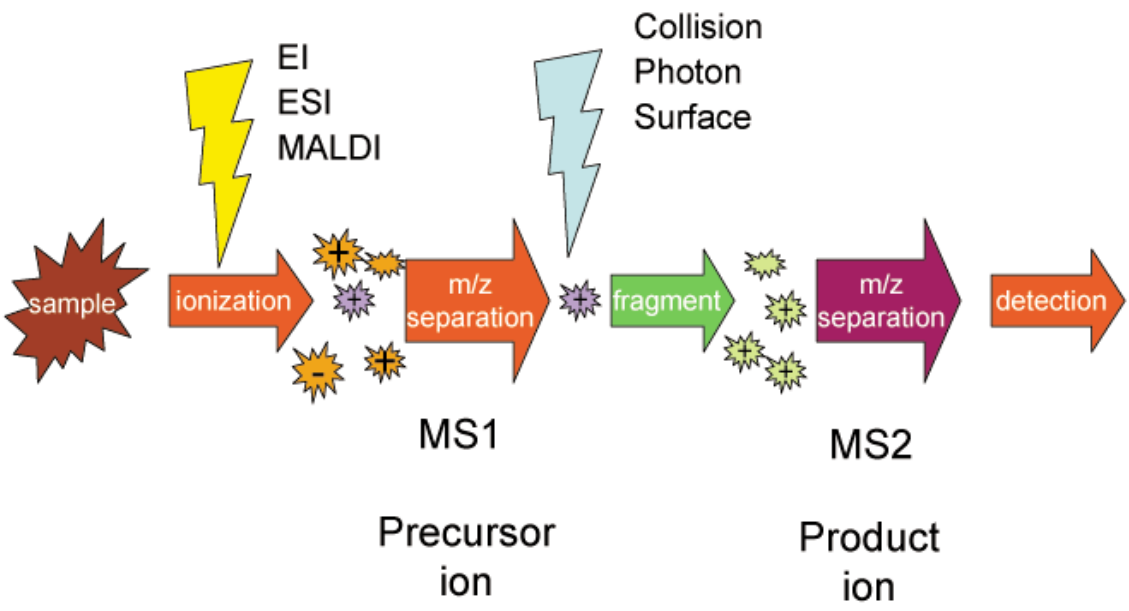

Figure 1: The general functioning of a tandem mass spectrometer is summarized in this scheme: the tryptic peptides are ionized through a "soft" approach (Electrospray ionization in our case), then the mass analyzer performs the separation based on m/z values. Once MS1 precursor ions are detected and selected, the peptides are fragmented through collision against a gas and product ions are produced. In the MS2 part, the fragmented ions are finally analyzed and detected[85]*. https://en.wikipedia.org/wiki/Tandem_mass_spectrometry#/media/File:MS_MS.png. Published 13th May 2006, author K. Murray. Accessed 10th November 2021.*

the most common "soft ionization" techniques that are used for peptides, since it avoids in-source fragmentation of these molecules. Once peptide samples are prepared, a liquid chromatography system connected to the mass spectrometer allows for elution of peptides online during MS analysis (Figure 1). For tandem mass spectrometry acquisition (so-called MS/MS or MS2), the ionized peptides are individually subjected to fragmentation, and the mass difference between fragments allows the reading out of the peptide sequence; peptides take the following path:

• The ion source (in this case the electrospray ionization[86]) applies a large potential difference (a voltage in the range of 1.5-4 kV) to the ion solution, creating an aerosol, from within which, via a process of ion evaporation, peptides pass into dry gas phase state and then enter the mass analyzer section;

• The first mass analyzer, that separates the different ions based on their m/z and can act as a mass filter, allowing for choosing and isolating ions of interest.

• The final mass analyzer and its detector, that will record the m/z and the intensity of the signal across the time of acquisition (called MS1 spectra for these unfragmented peptides and MS2 for spectra after fragmentation, see below)

• The fragmentation component, that will allow to dissociate peptides in smaller fragments, important for peptide sequencing.

• After fragmentation, the resulting ions that are produced are sent again to the final mass analyzer and detector to measure the so-called MS2 spectra.

Depending on how MS1 and MS2 spectra are selected and acquired, there are three different LC-MS/MS modalities, described in the following sections.

1.2.1.1 Data-dependent acquisition (DDA)

DDA is the current method of choice for discovery projects, in which there is no need for previous knowledge about protein presence and abundance in the sample. In this acquisition mode a large m/z interval is scanned (MS1 spectra) and the most intense peptides per scan are isolated in real-time and subsequently isolated and fragmented for [87]. The resulting fragments from every peptide (MS2 spectra) are be used for peptide identification and quantification. The generated MS2 spectra are matched against a database of theoretically in-silico fragmented peptides for peptide identification. In the final step, the identified peptides are used to infer proteins and their abundances[88].

DDA can be performed label-free or, in combination with sample labelling, in different ways (metabolic or chemical labelling)[89,90]. Even if label-free quantification is a cheap and more straightforward option for the analysis of clinical samples[91], peptide labelling has several advantages over the label-free approach: several samples tagged with specific masses and pooled and analyzed together, reducing the instrument time, it increases the technical reproducibility of data by eliminating the batch effect on sample quantification[92]. In comparison, label free quantification is more accurate in terms of protein abundance magnitude, but labelling approaches, as isobaric tags, have higher quantitative precision[93]. The stochastic nature of MS data acquisition yields missing values with both methods that increase with number of samples but can partially be negated by labelling strategies.

1.2.1.2 Targeted acquisition

With targeted MS (Single Reaction Monitoring, Multiple Reaction Monitoring or Parallel Reaction Monitoring) only the acquisition of a limited, number of defined peptides is possible, and previous and comprehensive MS knowledge about the proteins/peptides of interest is required.

Briefly, from a previously acquired DDA dataset (or public database available), peptides are chosen for identification, selecting MS1 features and MS2 fragment ions (in this case called "transitions")[94]. Once the transitions are chosen, the MS acquisition is performed in the selected m/z windows of the peptides, allowing for a very sensitive and selective detection of the targeted peptides across the samples. As consequence, this technique is often employed as a tool for validating DDA discovery data, affinity-based data or for monitoring protein levels[95].

1.2.1.3 Data-independent acquisition (DIA)

DIA is a technique between targeted and DDA modes, in which the MS1 range is systematically and cyclically scanned using predefined sized windows, fragmenting all the peptides isolated simultaneously, independently from the intensity[96]. Subsequently, the isolation and fragmentation of a mixture of peptides yields a complex MS2 spectra that represent all peptides in that window. To identify the peptides from the mixed MS2 spectrum, the transitions are matched to a peptide library generated from DDA data and the protein identities and quantities are computationally inferred[97]. This label-free technique has many advantages over DDA such as improved sensitivity, better reproducibility of data and the magnitude of variation in proteins across several samples[98].

1.2.2 Antibody-based quantification of proteins

There are many reasons that can explain why antibodies are a good choice for protein detection: they are very sensitive, relatively easy to use, they can be employed on different kinds of samples and, most important, they are supposed to be specific to the protein of interest[99]. On the other side, antibodies are well known for harbouring problems of cross-reactivity, a phenomenon in which the antibody recognizes other epitopes with inferior, the same or even higher affinity than the immunogenic epitope[100].

To have an accurate discussion about this topic, it is good to define three important concepts about antibodies:

• Antibody affinity, a thermodynamic concept derived from the application of Mass Action Law to the antibody-antigen reaction, describes the strength of binding between the antibody and the chosen epitope.

• Antibody specificity, a phenomenon for which an antibody is capable of distinguishing among many similar epitopes, even if different because of a very small chemical modification (as elegantly described in 1933 by K. Landsteiner[101]).

• Cross-reactivity (or multi-specificity or promiscuity), a well-known property of antibodies that can recognize epitopes that differ from the immunogenic peptide.

Antibody affinity is a property that varies during the development of B cell response: while the IgM produced by naïve B cells possess a generally low affinity (and high cross-reactivity), the maturation of B cells induces a series of random mutagenesis in the paratope in order to increase the variability of that region and select higher affinity antibodies[102]. At the same time, B cells switch from IgM to IgG production (and in some cases to IgE and IgA). Although affinity maturation increases both antibody affinity and specificity towards the immunogenic epitope, it does not eliminate the cross-reactivity of phenomena, as described below.

As demonstrated, antibodies can show specificity, being capable of differentiating between two epitopes that differ just for a small chemical group. Given these early results, a lock and

key theory was elaborated, supposing rigid and mutually complementary surfaces between the epitope and the paratope[103]. Nonetheless, successive structural studies about antigen-antibody binding have revealed a much more complex situation, in which specific residues in the paratope contribute to the binding, together with electrostatic and shape complementarity, water molecules and other cofactor molecules involved (like heme group in recognizing bacterial antigens[104]).

Even if mechanisms of cross-reactivity are complex and have not been fully elucidated, some important elements have been identified thanks to structural and molecular studies:

• Cross-reactive antibodies bind different epitopes with high specificity, involving each time a different set of hydrogen bonds[105] and amino acid residues[100].

• Disordered domains of proteins show conformational plasticity and these parts are often involved in cross-reaction through molecular mimicry[106].

Conformational rigidity of the paratope is associated with reduced cross-reactivity, while flexible paratopes are capable of adaptation to more epitopes[107]. Even if cross-reactivity is a reaction that involves off-target proteins, nonetheless it will follow the same rules that govern epitope-paratope interactions: the abundance of the off-target epitope will influence the amount of complexes formed with the antibody. Due to the physiological variability of the proteome across different cells and tissue, a contextual variation in cross-reactive binding has to be expected. This phenomenon should be taken in consideration when choosing the right method to validate antibodies.

1.2.3 Validating MS results with IHC

The validation of MS data through antibodies is an important and confirmatory step that is often carried out after a careful selection of proteins. The criteria and the methods used for protein validation are not standardized, since the purpose for validation is quite dependent on the aims of the project. The choice of proteins to be validated should start with a statistically based selection, in which various methods can be applied to point out protein variance. Among the most used I cite Student's t test, ANOVA[108], Limma[109], linear mixed models[110] and DEqMS[111]. The last cited is a specific tool for MS proteomic data developed in my laboratory that takes the number of PSMs for quantification into consideration. Once a list of proteins is chosen for validation with IHC, many issues need to be considered, as described in the section below: specificity and selectivity of primary antibodies, sensitivity of the detection, abundance of the protein, its range of expression and the expected subcellular localization. David Handler and colleagues have written a very interesting paper about validation of MS data, pointing out many flaws that occur during this process (high vs low abundant proteins, total amount of attempted validations)[112].

1.2.4 From a protein to a biomarker

Since the beginning of modern clinical oncology, medical doctors have been searching for symptoms, signs, chemical tests, radiological assessments and histological characteristics that could help to discriminate patients responding to a therapy from non-responders or to distinguish between unfavourable and good prognosis[113]. The heterogeneity of response to treatment is consequential to differences among tumors and among patients in terms of genomes[114], mRNA expression, protein levels, pharmacogenetics[115], immune-system status[116] and previous exposure to infectious agents and toxic agents[117]. From this point of view, the contemporary efforts to build histological and molecular classifications of neoplasms are of vital importance to appreciate the complexity of cancer and allow us to stratify patients with different prognosis in a better way than before.

The term biomarker has been defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic response to a therapeutic intervention" by the Biomarkers Definitions Working Group in 2001[118]. The development of a biomarker from the discovery study to clinical approval is a very long journey that follows specific paths and is necessarily connected to tumour biology and clinical behaviour[119,120]. A successful biomarker discovery project will start with a good experimental design, will go through a good sample collection with sample and data analysis that can identify potential biomarkers as proteins or mRNAs, genetical mutations, metabolites, etc that are differentially expressed between 2 or more subgroups[121].

Once the biomarker discovery phase is concluded, the validation steps occur. Briefly, the biomarker validation is organized in three phases called analytical validation, clinical validation and clinical utility[122]:

• Analytical validation: during this phase, accuracy, precision, analytical sensitivity and specificity, linearity, robustness and reproducibility of the measurement of the biomarker is evaluated[119].

• Clinical validation: in this step, the biomarker is tested for its ability to stratify patients based on different outcomes[120].

• Clinical utility: the biomarker is employed as a clinical test and evaluated for improving the outcome of some patients given an indication of therapy[122].

If every part of this validation path is fulfilled, the biomarker will enter the approval phase through regulatory agencies (e.g. Food and Drug Administration) and will finally enter the market, alone or as a companion for a targeted drug[123].

1.3 ANTIBODY VALIDATION: A PARTIALLY SOLVED PROBLEM

1.3.1 Still a need in 2021?

As stated by numerous articles, the lack of antibody validation is one of the main components of "reproducibility crisis" in biomedicine[124]. Several reasons could explain why scientists employ non validated reagents in their experiments, among them an unjustified trust in the company producing the antibody, lack of time, patience and funding, journals not asking for validation experiments, etc.

What does it mean to validate an antibody? It means to assess its affinity and specificity for the target protein, the presence of cross-reactors and the reproducibility of reaction across different batches of the reagent [125]. Once all these properties of the antibody have been assessed and considered the context in which it has been validated (which tissues and/or cell lines, which technique), the researcher can decide how the reagent should be used. Assessing the specificity of one antibody can be a demanding task, depending on the protein and the intended use. Even more difficult can be the assessment of cross-reactivity against the entire proteome or the proteins expressed in certain normal tissues, tumors or cell lines. As explained in the former paragraph, cross-reactivity is a frequent phenomenon with both monoclonal and polyclonal antibodies and the validation approach needs to take this into account. To produce meaningful validation, we need to adopt specific techniques that make us able to evaluate them in a comprehensive way.

In the past, several approaches were used to validate antibodies for IHC: among the most employed, western blotting (WB) has been the favorite of many scientists and employed in the antibody industry environment. The adoption of WB has been widespread even if it was more supported by empirical evidence of efficacy than extensive studies of validation[126,127]. In the effort of standardizing the approaches to validation and improve the quality of results, in 2016 the International Working Group on Antibody Validation[126] published an article proposing five different "pillars" (concepts) regarding the validation of antibodies:

• Genetic pillar: the antibody is validated on a cell line or tissue in which the target protein has been downregulated using siRNA[128] or knocked out with CRISPR/Cas9

• Orthogonal validation, in which an antibody-based signal is correlated to an antibody-free determination of the same protein detected, using linear correlation as a tool for assessing it[129] (see specific section about orthogonal validation below)

• Independent antibody technique: two or more antibodies against the same protein but recognizing different epitopes are evaluated for producing the same pattern

• Immunoprecipitation-Mass spectrometry: the antibody of interest is incubated with a sample containing the protein of interest, then proteins are digested from the antibody and analysed with the MS, compared to a generical isotype immunoglobulin[130]

• Tagged protein or fluorescent protein expression: using one of these techniques, a protein is expressed in a model that does not contain the protein of interest. The pattern produced by the knocked in protein is then compared to the antibody positivity[131].

While this publication has constituted an important step in establishing which the main approaches we should use for validating any antibody are, the pillars cannot be considered definitive criteria universally adopted. Indeed, extensive work should be done to determine which pillar should be used in which phase of antibody validation, if some pillars are more reliable than other, how the pillar(s) should be implemented (which model, which samples) and if other criteria can be implemented (like subcellular localization through proteomic data)[132,133].

Among the 5 pillars of the IWGAV, the proteomic-based orthogonal validation is among the least implemented criteria. One of the main reasons why proteomic data should be used instead of mRNA data is because of the relatively poor correlation between mRNA and proteins[134], and if compared to proteomic validation, the correlation indexes are lower[135]. The proteomic-based orthogonal validation is discussed in the paragraph 1.3.3

1.3.2 The limits of the actual approaches on human FFPE tissues

As pointed out by many authors, including the International Working Group on Antibody Validation, the validation of antibodies on human FFPE presents several problems that still need to be solved completely, given the impossibility of using genetic techniques and tagged approaches. Each of the rest of the available methods show some drawbacks: the independent antibody validation is based on the comparison between 2 or more primary antibodies directed against different epitopes of the same protein and comparing the staining patterns generated by them. The assumption implicit in this approach is that specific and selective antibodies should show a very similar staining pattern; while this assumption is often widely accepted, it also implies knowledge about the epitope for which the antibody was produced, so that 2 ABs produced against different epitopes can be used. Finally, the lack of knowledge of the epitopes used to produce ABs severely limits independent antibody validation.

1.3.3 Orthogonal validation of antibodies

The orthogonal approach was formalized by Matthias Uhlen and collaborators in 2016[126] as a comparison in which an antibody-free protein quantification is correlated against the antibody-based quantitation for the same protein across several samples. Such method had been already employed on single cases before for validating targeted MS proteomics with ELISA measurements on biological fluids (plasma, urine, CSF, seminal fluid) and FFPE tissue extracts[136-141]. A more direct application of this technique to IHC has been done by the group of David Rimm, testing an anti-EGFR antibody on a panel of cell lines and adopting a fluorescent detection[142]. A more thorough investigation of antibodies and orthogonal validation on WB has been done by the Human Protein Atlas by Fredrik Edfors and collaborators on a set of 53 antibodies, using labelled DDA and targeted quantitation, and

comparing it to mRNA quantification[135]. Moreover, they validated a further set of 711 antibodies using a 2 cell lines approach with labelled DDA quantification.

The implementation of proteomic orthogonal validation has many advantages over several of the other techniques:

- the high number of samples that can be tested in each experiment
- the knowledge of protein levels and magnitude of variation between samples
- the quantification of thousands of proteins for each dataset
- the possibility of relating the AB signal to the rest of the detected proteome for estimation of specificity and sensitivity of the reagent.

1.3.4 A fit-to-purpose approach to antibody validation

Reconnecting to the precedent paragraph about biomarker development and validation, one integral part of such procedure is to perform a thorough analytical validation of the antibody, considering the final use already in these steps. This means defining the final target tissues or tumors, the expected range of expression of the protein in such tissues/tumors and how the antibody can detect it, how heterogeneous the expression is, how reproducible the staining across different antibody's batches and different staining platforms is. In other words, already the analytical phase needs to be shaped around the potential use, considering many factors that could influence its performance. From this point of view, MS-based proteomics can contribute beyond the orthogonal validation in a significant way, given the exciting potential of absolute quantification in defining the upper and lower limits of detection[143].

2 RESEARCH AIMS

The overall aim of this thesis is to investigate how histopathology and mass spectrometry-based proteomics are connected to each other under several aspects. I have focused on three main aspects:

1. Exploration of relationship between proteomic classification and histological characterization of tumors, especially lung and breast cancer (**papers I** and **II**)
2. Development of a mass spectrometry-based orthogonal approach for antibody validation in IHC (**paper III**)
3. Histological and IHC validation of results obtained in a discovery modality with MS proteomics (**paper I, II, III**)

3 MATERIALS AND METHODS

3.1 ETHICAL CONSIDERATIONS

The declaration of Helsinki is a document that lists the ethical principles to be followed regarding human experimentation[144]. Written in 1964, it is considered the most important declaration defending the rights of humans during clinical trials and research. As clearly exposed by Johansson and Lynöe[145], the three main principles to be followed are:

- Requirement of informed consent: the information given to the patient should be comprehensible and adequate (comprising aim of the study, methods, funding, conflict of interests, potential benefits and harms, right of withdrawal consent, right to abstain from the study without consequences). Moreover, the patient should be able to give a voluntary consent and to decide by themself.

- Balance of risks and benefits: the potential benefits deriving from the study should be counterbalanced by the potential harms and a complete assessment done.

- Competence of research: design and methods adopted in the protocol should be adequate for the aim of the study, following laws and regulations of the state in which the experimentation is performed. The research should be accessible to the ethical review committees that approved the study.

The **paper I** and **II** made use of human clinical samples taken from different patient cohorts, as described in the respective sections. Following the Helsinki declaration and as already described in the relative paper, informed consent was obtained by each patient for the collection of cancer samples and clinical information. Moreover, each study here included was submitted to the ethical board for approval and a balance of risks and benefits was done before starting the studies.

The **paper III** only made use of continuous human cell lines commercially available, for which no ethical approval is requested.

3.2 PATIENT COHORTS

The **paper I** is a molecular landscape discovery study in which 45 breast carcinomas, belonging to the OSLO2 cohort, were submitted for proteomics, CNA, microarray, RPPA and metabolomics. The OSLO2 cohort was established in Norway in 2006 and collected samples from 297 patients until 2019[146]. For this study a subgroup of 45 tumors were extracted, choosing 9 tumors for each PAM50 subtype. Immunohistochemical results were validated on a subset of 530 patients of the OSLO1 cohort, a breast cancer cohort established in Norway between 1995 and 1998, including 920 breast carcinomas[147].

The **paper II** generated molecular data from 3 different patients' cohort: for the discovery part, the histopathological examination and the IHC, samples from 141 patients were taken from the Skåne University Hospital in Lund[148], while the validation cohorts used were the

Southern Swedish Lung Cancer Study (now LUCAS cohort, 84 inoperable lung cancer samples) and a Norwegian cohort from Oslo University Hospital of 209 operated lung cancer patients[149].

3.3 MASS SPECTROMETRY BASED-QUANTITATIVE PROTEOMICS

In the paragraph 1.2.1 of this thesis I have described mass spectrometry-based proteomics and how the signal can be acquired using different modalities. In this section I will describe how the data were acquired for the three different papers.

In **paper I** and **paper II** samples were processed for labelled DDA acquisition, using HiRIEF LC-MS/MS. Briefly, samples were lysed using an SDS-based buffer and the proteins extracted. After measuring the protein concentration, proteins were cleaned and digested with trypsin, using the FASP protocol[150] for the **paper I** and following the SP3 protocol developed by the Krijgsveld lab[151] for the **paper II**. After protein digestion peptides were labelled with 10-plex TMT isobaric tags and then separated off-line using immobilized on-gel pH gradient isoelectric focusing (HiRIEF), as described by Branca et collaborators[152]. In **paper II** samples were also prepared for DIA acquisition, a label-free analysis performed after peptide cleaning done with SP3 protocol. Labelled samples were acquired with LC-MS/MS on DDA mode, while label-free data for the paper II were analysed with a DIA modality.

3.4 IMMUNOHISTOCHEMISTRY AND IMMUNOFLUORESCENCE

Immunohistochemistry (IHC) is an in-situ form of affinity-based protein detection. Because of the possibility of being applied on formalin-fixed paraffin-embedded tissues and the possibility of preserving the morphology of tissue, it has gained the favor of histopathologists in both diagnostic and research settings as the main "ancillary" technique for protein studies[153,154]. The main steps of sample preparation and IHC are listed and commented below[155](see also Figure 2):

• Collection of the sample and fixation: the fixation step, necessary to interrupt lytic phenomena of the tissue (or cells), can be carried on with several fixatives, most commonly with buffered formaldehyde. The fixation phase is one of the most important passages in this procedure, since inadequate fixation lengths (too short or too long) will affect the protein detection in an irreversible way[156].

• Dehydration, embedding and cut: at the end of the fixation period, the samples are washed with water, dehydrated with ethanol and xylene, embedded in special paraffin for histology and finally cut in thin sections (around 4 μm).

• Antigen retrieval: after paraffin removal with xylene and rehydration, to recover the antigenicity of the proteins and counteract the action of formaldehyde, sections are submitted

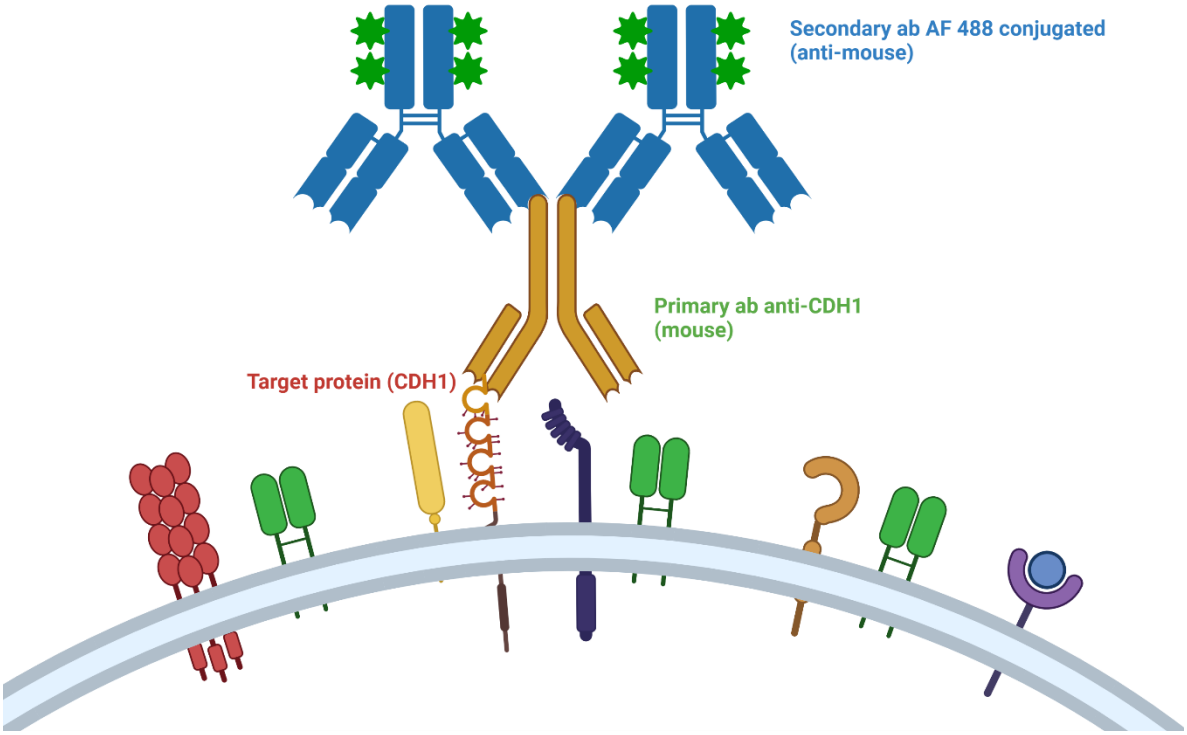

Figure 2: Scheme of fluorescence-detected immunofluorescence against CDH1 (E-cadherin). The primary antibody, raised in mouse, is binding in a specific and selective way to the target protein (in this example CDH1). The secondary antibodies, targeting mouse Abs, are conjugated with a fluorophore (Alexa Fluor 488) and are binding the primary antibody. Created with BioRender.com.

to "Heat-induced epitope retrieval" (HIER)[157]. Briefly, sections are heated in buffers with optimized pH for the antigen, for standardized time and temperature, with the aim of re-exposing the epitopes necessary for the binding with the selected antibodies. Alternative to HIER, it is possible to employ special protease digestions for reaching the same scope.

• Primary antibody incubation: after HIER, sections are incubated with the antibody of choice at a specific dilution (in order to achieve the best signal-to-noise ratio) and optimized time, pH, blocking agent, etc.

• Secondary antibody and detection system: After several washing steps, in order to remove the excess of primary antibody, sections are incubated with the chosen secondary antibody and detection system. Many detection methods can be employed in IHC, employing chromogens or fluorophores, secondary antibodies directly conjugated with enzymes/fluorescent molecules or amplified introducing polymers, etc. The choice of this component is dependent on the abundance (and availability) of the epitope, on the affinity of the primary antibody and on practical aspects of interpretation (the chromogenic detection is much more common in diagnostic pathology)[154].

In the three papers IHC is an important method applied to the different samples. In **paper I** mIHC is used to detect MET and EGFR and characterize their co-expression on 37 breast carcinomas. Seen from another perspective, mIHC is used to validate a co-expression discovered with MS in the basal-like and normal-like tumors. In **paper II** chromogenic IHC is used to describe the type of immune infiltrate and detect tertiary lymphoid structures, together with histology (see below in the immune section). In **paper III** quantitative mIHC is used to measure the AB signal from different cell lines and correlate it to the MS signal.

3.5 HISTOLOGICAL CLASSIFICATION OF BREAST AND LUNG CANCER

3.5.1 The 2019 WHO classification of breast tumors

The WHO classification of breast tumors[158] includes a diverse mixture of neoplasms that can arise in the breast, including metastatic lesions. By far, the most common malignant tumor is the invasive carcinoma no special type (NST), an adenocarcinoma of variable degree of differentiation and that shows no special morphological characteristics. The second most common malignant neoplasm is lobular carcinoma, a special type of breast cancer characterized by scarcely cohesive cells, Indian file growth pattern and lack of expression of E-cadherin. This classification has been used to classify the 45 breast carcinoma samples used in the **paper I.**

3.5.2 The WHO classification of lung and chest tumors

The neoplasms that arise in the lung are histologically categorized according to the WHO lung and chest tumors classification[159]. Among the non-metastatic tumors that can arise in the lung, carcinomas are the most common and are historically divided in small cell lung carcinomas (SCLC) and non-small-cell lung carcinomas (NSCLC), based on the morphological aspect and the speed of progression. Among the NSCLC group, the most common tumors are adenocarcinoma (AC), squamous cell carcinoma (SqCC), large cell carcinoma (LCC) and large cell neuroendocrine carcinoma (LCNEC, this last one shows characteristics in common with SCLC). In the AC group, several growth patterns can be recognized, linked with a different biological behavior and a different 5-years survival[160], including a “carcinoma in situ” form (lepidic pattern). In **paper II** lung carcinomas have been classified according to the 2015 version of the WHO classification and ACs have been subclassified, quantifying the pattern of growth in increases of 5 %, as suggested by Travis and collaborators[39].

3.6 HISTOLOGICAL EVALUATION OF IMMUNITY IN LUNG CANCER

The presence, histological organization and phenotype of immune cells in lung cancer can be assessed in a multitude of ways, depending on what aspect needs to be highlighted. In **paper II** the tumor immune component has been evaluated in the following ways:

1. Chromogenic IHC for CD3, CD8 and CD20 on TMA, following the counting method of Al-Shibli and collaborators[161]. Briefly, the percentage of membrane-positive cells

was calculated (counting at least 100 cells), counting tumor and stroma compartments as different. The cut-off for positivity was set at 1%.

2. Chromogenic IHC for PD-L1 (using the clone 28-8) on TMA sections was interpreted following the interpretation guidelines supplied by the manufacturer[162]. For the TPS (Tumor Proportion Scoring) minimum amount of 100 neoplastic cells was considered for each tumor, considering a weak and partial membrane signal as a minimum positivity. No stromal positivity was considered in the scoring.
3. The presence and quantification of tertiary lymphoid structures was assessed on selected HE whole slides according to Lee HJ et al[163], estimating the amount of peritumoral tissue involved by TLS (little <10%, moderate between 10 and 50%, abundant >50%).

3.7 SUPER-RESOLUTION MICROSCOPY

Super resolution microscopy is a collection of microscopy techniques that can circumvent the physical diffraction limit of light and make it possible to investigate biological processes at sub-microscopic resolution. Among the different approaches, the Stimulated Emission Depletion (STED) microscopy is a point spread function engineering technique that is based on the coupling of a central exciting laser and a doughnut-shaped deactivating laser that inactivates the fluorescent dye in the peripheral portion[164]. Such fluorescent excitation present only in the central part of the laser beam permits a lateral resolution between 10 and 80 nm, well below the diffraction limits of the light[165].

In **paper I** super resolution microscopy has been applied to breast cancer specimens to investigate the co-expression of EGFR and MET and their subcellular localization between basal-like and normal-like tumors.

3.8 IMAGE PROCESSING AND ANALYSIS

Image processing and analysis is a system of computer-based techniques to extract information from images. Depending on the context and the kind of information the user wants to extract, different tools and technique will be applied in a sequential fashion.

In the **paper III** I have elaborated a workflow to extract mean pixel intensity values across the 18 cell lines for many primary antibodies recognizing different protein targets. The main softwares used for this task have been CellProfiler (v 4.0.7), Ilastik (v 1.3.3) and R (v 4.0.3).

1. Image acquisition. The type of microscope and object, the exposure settings and the file format are some of the most important aspects that influence the final results. A wrong choice made at this stage will influence the results in a negative way, making the results unreliable. For this step I have used a Metafer widefield fluorescent slide scanner equipped with a Imager Z.2 and a 20X object. The acquisition has been done with fixed exposure time and autofocus, in 4 fluorescent channels (DAPI for hoechst

33342, FITC for mouse antibodies detection, Texas Red for rabbit antibodies detection and Far Red for cell mask detection). The output images were saved in TIFF format grayscale 8-bit.

2. Pre-processing. Before starting the real image processing, image file needs to be prepared for the segmentation ad measuring steps. Among the many techniques that can be applied at this stage, there are: filtering, color extraction, illumination correction, inversion, enhancement (to increase SNR), stitching, image registration and deconvolution. In my pipeline, after performing an acquisition quality control for saturation, images were corrected for uneven illumination using a "background" illumination function with a B-spline smoothing.
3. Segmentation. In this phase an image is used for identifying "primary objects" that will be used as a starting point to recognize cells, potentially with the help of a mask that will expand in a "secondary object". Usually in a first phase a threshold is calculated (manually or through an algorithm) to assign pixels to background or foreground and transform the image into a binary one. In a second phase another approach is used to separate the different cells from each other, using another algorithm to recognize local minima in the pixel value. If the image is challenging to segment, a machine learning-based approach to classify pixels and objects can be used, like in this case. Briefly, the software Ilastik was trainer to recognize pixels belonging to the nuclei in the DAPI channel, and separate them from the background. In a second step, every nucleus was recognized as an object and exported for subsequent use. Finally, if a cytoplasmic mask is available (or a cell area is calculated, starting from the nucleus), a cytoplasmic area is detected using algorithms for secondary objects detection. In this paper a cytoplasmic mask was available and used for cytoplasmic detection employing the algorithm.
4. Measurement. Once the cells are identified, the cellular area of the image is measured, using different ways of summarizing the pixel intensity values, such as mean, median, standard deviation, etc. In the paper, a background pixel value was calculated for each channel as a lower quartile intensity, in common for all the images belonging to the same slide. Such values were subtracted from the images before measuring the mean pixel value in the antibody channels from the cellulated areas. Additionally, the SNR was calculated for the mouse and rabbit antibodies channel, to check for signal presence.

4 RESULTS AND DISCUSSION

4.1 PAPER I: BREAST CANCER QUANTITATIVE PROTEOME AND PROTEOGENOMIC LANDSCAPE.

This molecular landscape article investigated a cohort of 45 breast carcinomas with multiple levels of omics data: quantitative MS based proteomics, CNA, microarray mRNA, RPPA and metabolomics. The proteomic analysis identified a total of 12645 proteins (gene centric) and 248949 unique peptides, of which 9995 proteins were quantified across all tumors. An unsupervised clustering analysis grouped the tumors in a very similar way to PAM50 classifier, while the same analysis on the proteins quantified showed grouping according to their biology. Using consensus clustering on a subset of highly variant proteins (n = 1334) 6 core tumor clusters were defined (CoTC), showing a partial agreement with PAM50. In fact, basal-like tumors were split into CoTC1 and CoTC2, HER2-enriched were fused with luminal B (CoTC6) and another subset of luminal B formed a separated cluster (CoTC4). A correlation network based on the same subset of proteins showed how the 2 subgroups of basal-like tumors differ by immune response, basal proteins expression and proliferation-related proteins. Similarly, the CoTC3 (luminal-A) and CoTC4 (pure lumina-B) differ by a small immunity network, related to interferon-alpha response, extracellular matrix and proliferation markers.

To find potential drug combinations, the co-expression of drug targets in "druggable" proteome analysis was analysed using the 290 FDA drug targets identified in the cohort (Figure 3A). ESR1, PGR, AR and BCL2 were found to be highly correlating by both RPPA and MS (Figure 3B). Similarly, EGFR and MET were found to be co-expressed in basal- and normal-like carcinomas, as shown by MS and mRNA data on the OSLO2 cohort (and confirmed by the mRNA data from the TCGA dataset[166], Figure 3C, 3D).

Since the normal-like group is debated and novel treatment modalities are needed in basal-like BC, co-expression of EGFR and MET were investigated using a mIHC approach on 37 of 45 tumors available in the OSLO2 cohort. The IHC confirmed the co-expression to be restricted to the basal-like and normal-like subtypes, showing how both proteins were expressed in the invasive tumors in the basal-like, while the normal-like positive tumors were high grade DCIS without an invasive component (Figure 3E). Such results were confirmed on the OSLO1 cohort (Figure 3F), too, showing how just a small number of normal-like

invasive tumors were positive

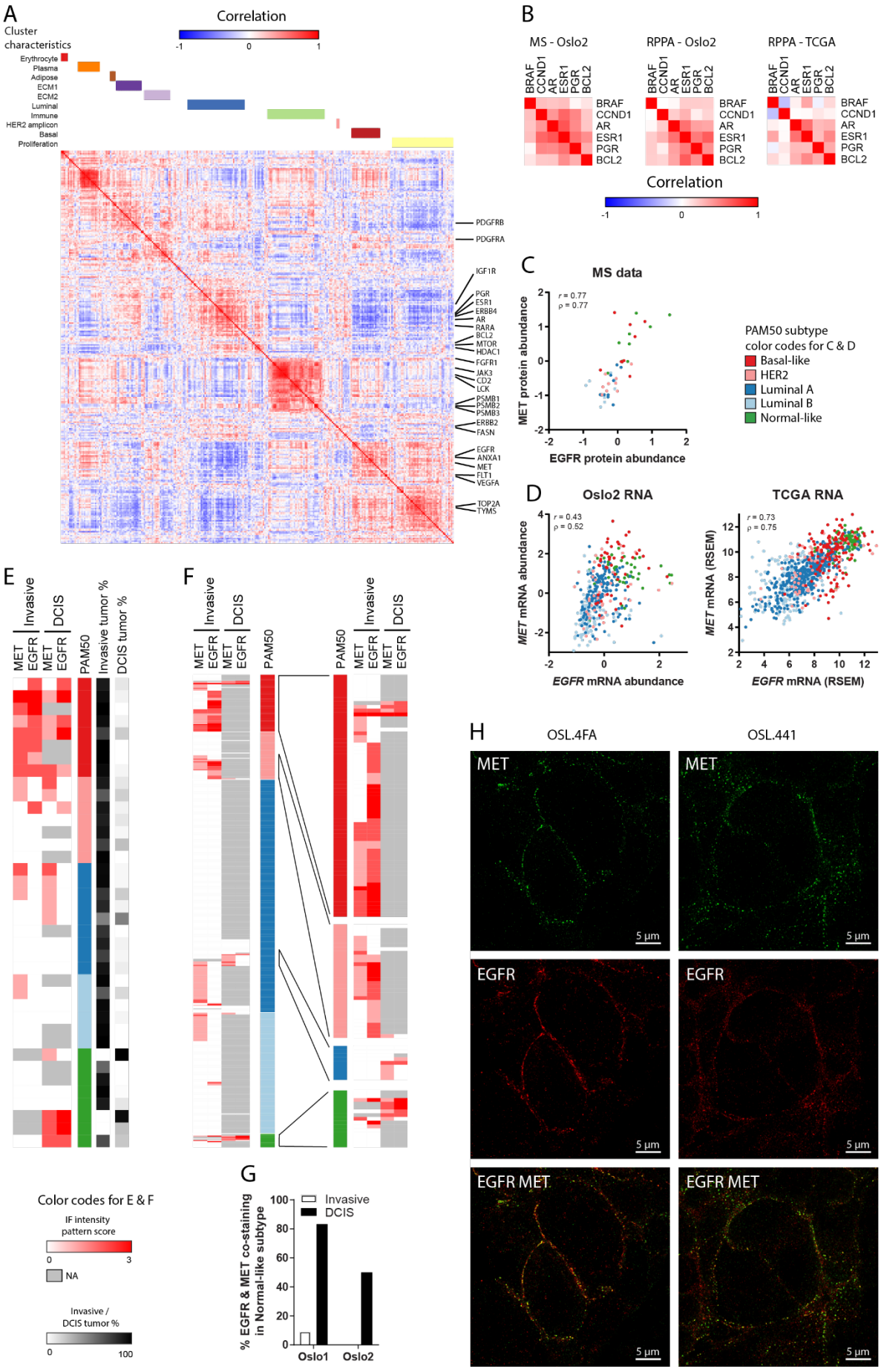

Figure 3: A) Correlation matrix of 290 FDA drug targets detected in the proteomic data; B) correlation matrix data about luminal targets, discovered on the MS data of the Oslo2 cohort, were validated on RPPA data from Oslo2 and the TCGA; C) MET and EGFR co-expression scatter plot, across the 45 tumors of the Oslo 2 cohort. The expression is higher in basal-like and normal-like tumors; E) The co-expression is confirmed on mRNA data from the Oslo2 and TCGA cohorts; e) IHC scoring for the EGFR-MET co-expression in the Oslo2 cohort (whole slides); F) IHC scoring for EGFR and MET co-expression on the validation cohort Oslo1 (TMA); G) co-staining of EGFR and MET in the normal-like subtype, DCIS versus invasive component; H) Super-resolution microscopy of EGFR-MET colocalization in two cases of normal-like DCIS. Reproduced with permission from Springer Nature Publishing Group.

for the co-expression (Figure 3G). Finally, a small number of samples were studied with STED super-resolution microscopy to search for subcellular localization and colocalization of EGFR and MET in basal-like and normal-like samples (Figure 3H). The microscopy showed a different subcellular localization and degree of colocalization, with MET and EGFR more colocalized and expressed on the plasmatic membrane in normal-like samples than basal-like, that showed a more intracellular localization and minor overlapping of the 2 signals.

4.2 PAPER II: PROTEOGENOMICS OF NON-SMALL CELL LUNG CANCER REVEALS MOLECULAR SUBTYPES ASSOCIATED WITH SPECIFIC THERAPEUTIC TARGETS AND IMMUNE EVASION MECHANISMS

The early stage 141 tumors used for the discovery stage were histologically divided in LUAD, SqCC, LCC and LNELCC; two SCLC were added as reference to this dataset. After proteomic analysis with labelled DDA acquisition, 13975 total proteins were identified, with 9793 gene proteins quantified across all samples. Additionally, panel sequencing of cancer associated genes, genome-wide methylation and RNA-seq were performed on these samples. A consensus clustering based on proteomic data showed 6 proteomic subtypes, consistent with a successive NMF clustering based on the same data. Based on this molecular classification, clusters 1-4 were represented mostly by LUAD, while subtype 5 was composed by LNELCC and SCLC. Finally, subtype 6 was exclusively represented by SqCC.A network analysis was performed, based on the protein level differences between subtypes and detected using the DEqMS tool[111]. The results indicated an immune infiltration of subtypes 2 and 3, and a stromal component in subtype 3; such results were confirmed using the ESTIMATE signature[167] for inferring the stromal and immune composition. Such data agreed also with the cell composition evaluation results, since the subtypes 2 and 3 were considered the less “pure” among panel sequencing data evaluation.

The application of immune signatures previously published[168] on the proteomic data showed a differential expression of immune-related proteins between subtype 2 (high in T cells and interferon signaling) and subtype 3 (B cell expression)(Figure 4A). IHC for CD3 and CD8A on a subset of samples showed a good correlation with proteomic quantification. The investigation of the antigen processing and presentation machinery (APM) and the tumor mutation burden (TMB) showed how the subtype 2 was correlated with both APM and TMB high, while the subtype 3 showed high APM but low TMB. The evaluation of PD-L1 IHC on a sub-cohort of patients showed a high expression of the marker in the subtype 2, suggesting a potential group of responders to ICI.

The expression of B cells markers in the subtype 3 suggested the presence of tertiary lymphoid structures (TLS), which presence is associated with a better prognosis and predictive of immunotherapy response[169]. An analysis based on a mRNA signature showed

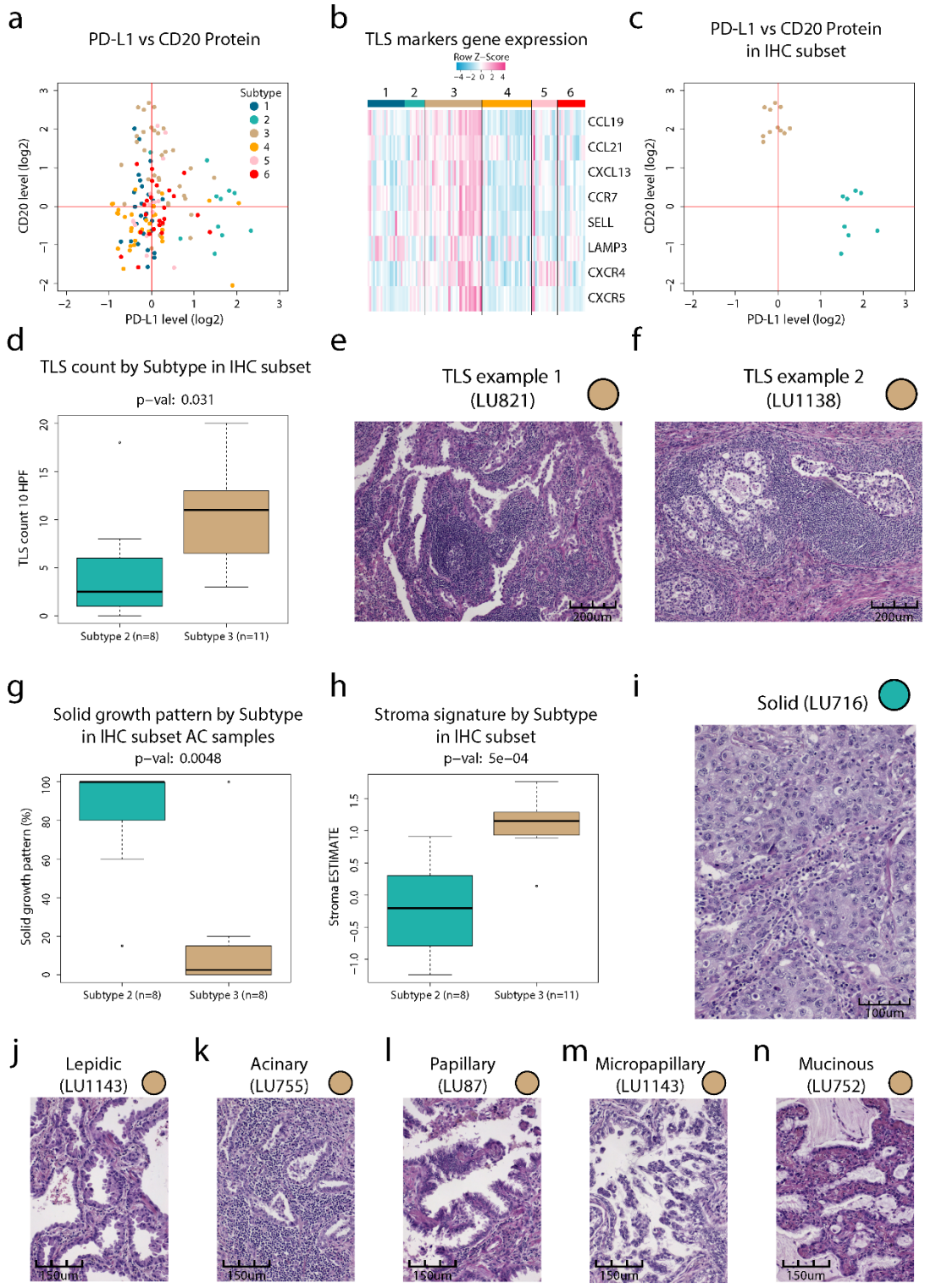

Figure 4: relationship between TLSs, immune markers and growth pattern across subtypes 2 and 3. a) scatterplot of PD-L1 vs CD20 (MS data) across the 6 proteomic subtypes; b) heatmap of mRNA data for TLSs signature, divided by proteomic subtype; c) scatterplot PD-L1 vs CD20 of selected examples of subtype 2 vs subtype 3 tumors; d) Count of TLSs per 10 HPFs,

subtype 2 vs 3; e,f) two examples of TLSs from subtype 3; g) boxplot of solid growth pattern percentage in tumors belonging to subtype 2 vs 3; h) boxplot of stromal signature ESTIMATE, subtype 2 vs 3; i-n) examples of tumor growth patterns from different cases: solid, lepidic, acinar, papillary micropapillary, mucinous. Reproduced with permission from Springer Nature Publishing Group.

high levels of TLSs in a subgroup of subtype 3 (Figure 4B); such finding was confirmed comparing the histology of tumors belonging to the subtype 2 (with high PD-L1) or 3 (high B-cell markers) (Figure 4C, 4D). This histopathological evaluation pointed out important differences in LUAD growth pattern, since subtypes 2 samples showed a largely predominant solid growth, while the subtype 3 showed a mixture of lepidic, acinar, papillary, mucinous, micropapillary and solid patterns (Figure 4E-4N).

4.3 PAPER III: A METHOD FOR VALIDATION OF ANTIBODIES FOR IMMUNOHISTOCHEMISTRY USING QUANTITATIVE MASS SPECTROMETRY BASED PROTEOMICS: ABMS

Eighteen cell lines were chosen to cover a large part of the human proteome, selecting 17 of them based on diverse protein coding mRNA levels in CCLE RNA-sequencing data (containing 1048 cell lines)[138]. Selection was to cell lines present in the DSMZ cell bank (https://www.dsmz.de) and including the in-house cell line A431. Of the 335 cell lines selected, a further selection was based on a hierarchical clustering based on a subgroup of genes that were expressed in at least 4 cell lines and not present in all of them. Finally, based on the mRNA coverage, 18 cell lines were selected manually and used for proteomic analysis and the construction of a cytoblock microarray (CMA) (Figure 5A-5B).

After cultivating and expanding the cell lines, samples were collected and processed for quantitative proteomics using a labelled DDA approach. Briefly, after protein extraction with an SDS-based buffer, protein concentration was measured, and 200 mg digested with trypsin. After labeling 3 replicates of each cell line with a TMT 10-plex isobaric tag, samples were pooled together (using one tag as internal reference between sets) and fractioning the peptides off-line through isoelectric focusing before nanoLC-MS/MS[123].

Cells cultured at the same occasion as for the proteome analysis were used to build a cytoblock microarray. Each cell line was added in triplicate and distributed randomly to build the cytoblock. Such sections were used for developing the IF, that were acquired with an automatized microscope and the mean pixel intensity extracted for each replicate of the cell lines.

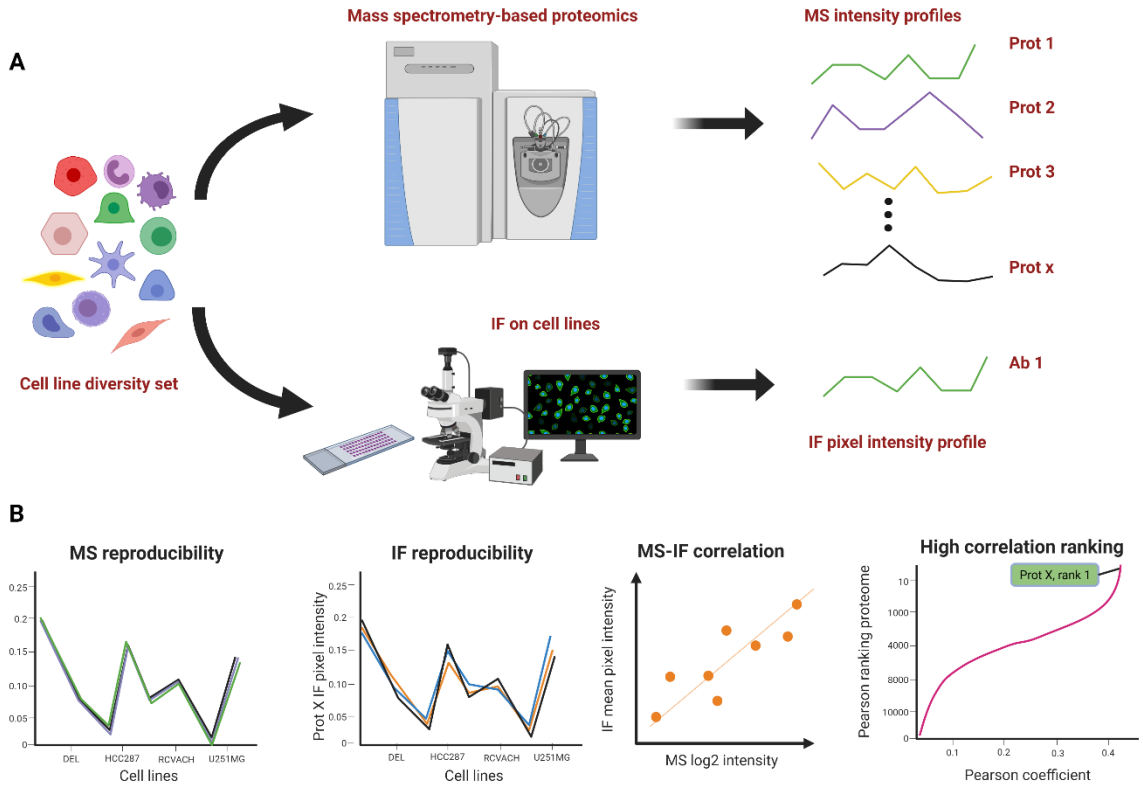

Figure 5: Experimental design of the study. A) The 18 cell lines dataset (diversity set) was submitted for both LC-MS/MS and quantitative IF. From each antibody was obtained an intensity profile that was correlated to the target protein and to the rest of the proteome to study the distribution of the correlations. B) The output of the analysis is shown from the left, with MS and IF reproducibility plots, the MS-IF correlation scatterplot and the coefficient distribution ranking for the target protein. Created with BioRender.com.

After obtaining the signal intensity profiles for both MS and IF data, we evaluated the reproducibility of the intensity signal through the 3 biological replicates (visualized as line plots). Such plots highlighted a variable reproducibility, depending on the specific antibody tested. The next step was to assess the correlation between the MS and the IF across the 18 cell lines, using Pearson and Spearman coefficients of correlation and showing the data on a scatterplot. Finally, an additional way of analyzing data was to correlate the IF profile with the detected proteome and then visualize the target protein ranking in relation to the whole distribution of correlations.

MS-AB and AB-AB correlation were tested over 3 different time points and with different dilutions, which provided confidence that the IF method performed reproducibly. To further investigate the correspondence with other validation approaches we decided to study 3 MKI67 antibodies and 4 PD-L1 (CD274) antibodies, using the independent antibody criterium. We observed, for both antibody groups, a very good concordance between MS-IF correlations and the high AB-AB correlation, measured with Pearson coefficient.

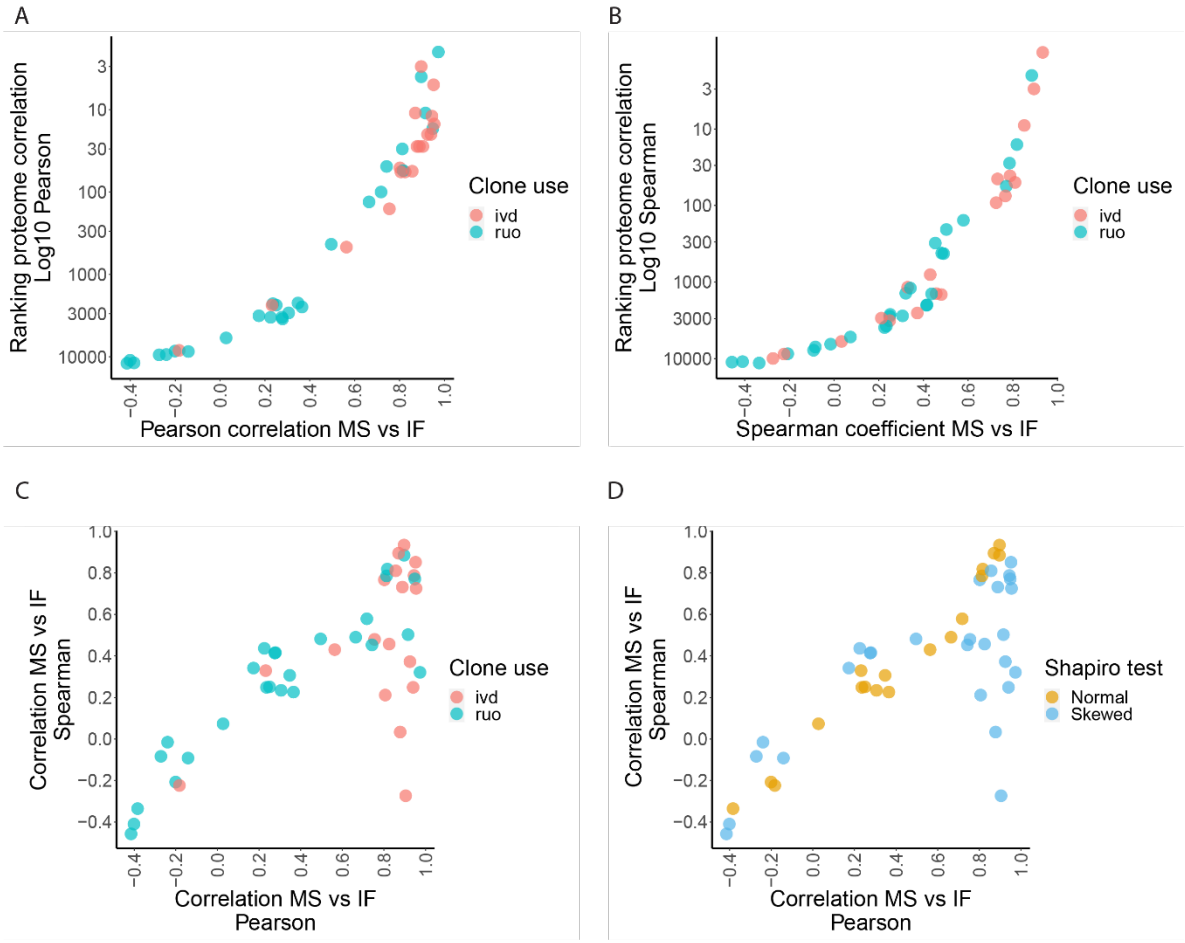

Figure 6: Aggregated data for 45 antibodies. A,B) The MS-IF correlation coefficients were plotted against the target protein ranking in the correlation distribution, for both Pearson (A) and Spearman (B) coefficients. There is an enrichment in IVD clones in the higher correlation values, evident with the Pearson coefficient. Note the logarithmic visualization of the y axis. C,D) Scatter plot of Pearson vs Spearman coefficients that shows the discordance between the two coefficients. It is evident how many high Pearson and low Spearman points are IVD clones (C) and how their data distribution is skewed (Shapiro-Wilk test positive), that justify the difference in coefficient values.

A total amount of 45 different antibodies were tested in this study, with variable MS-IF coefficient values (Figure 6A-6B). The coefficient values were concordant with the ranking of the target protein in relation to the coefficients' distribution. The plot of Pearson against Spearman coefficients of all the 45 antibodies showed discordances between the two coefficients. Such effect was attributed to the skewed distribution of the intensities, as demonstrated by the positivity of these points for the Shapiro-Wilk normality test (alpha = 0.05)(Figure 6D). Finally, a visualization of the Pearson's and Spearman's coefficient plot showed a group of high scoring antibodies, enriched in clones used for IVD (Figure 6C).

5 CONCLUSIONS AND FUTURE PERSPECTIVES

5.1.1 Concluding remarks about the thesis

The work presented in this thesis has the aim to show how pathology and proteomics (and molecular classifications) are strictly interconnected in many ways. In other words, they are simply separate ways to look at the same phenomena. The main conclusions that we can draw from the thesis are the following:

- The **paper I** has shown how investigating co-expression of drug targets has highlighted the co-expression of MET and EGFR in both basal-like and normal-like tumors. In this last group the co-expression was confined to a subset of DCIS high grade, while most of invasive tumors were negative for it. A super-resolution microscopy investigation showed a different degree of colocalization and subcellular distribution between basal-like and normal-like tumors.
- In **paper II** the proteomic-based clustering divided the discovery lung cancer cohort in 6 subtypes, strictly following the histology of tumors (4 subtypes were mostly LUAD, 1 LNELCC and the last SqCC). MS data and molecular signatures highlighted the abundance of immune infiltration in subtypes 2 and 3 and at the same time pointed out differences in immune phenotype (CD8 and Interferon signatures for subtype 2, B cell signature in the subtype 3). PD-L1 IHC showed high expression of the marker in the subtype 2, suggesting this subgroup could benefit from ICI therapy. An histopathological evaluation of selected cases from the subtype 2 and 3 confirmed the abundance of TLSs in the subtype 3 and showed important differences in growth pattern of the tumor (solid pattern in subtype 2, mixed pattern of growth in subtype 3).
- The **paper III** is a technical paper that investigates MS-based orthogonal validation of ABs using IF. I created a dataset of 18 cell lines that was submitted for both labelled DDA and histology, obtaining a MS proteomic quantification and a CMA for IF quantification. To evaluate for antibody performance, both reproducibility of MS and IF data was taken into consideration and the correlation between the two quantifications were evaluate with Pearson and Spearman coefficients. Moreover, we took into consideration the correlation of the IF profile with the rest of the proteome and the rank of the target protein among them. The analysis of 45 antibodies showed a good correlation between Pearson coefficient and top ranking value and highlighted how the highest scoring antibodies were enriched in diagnostic clones. The correlation between independent antibody validation and orthogonal validation was investigated on 4 anti-CD274 and 3 anti-MKI67 antibodies, showing a close performance of both validation criteria.

5.1.2 Future perspectives in pathology

According to Edward Abrahams, the main promise of precision medicine is "to give the right drug to the right patient at the right time"[5]. This principle is reflected in the general trend to increase and improve patient stratification in oncology to assure that patents with a specific molecular alteration will receive the proper drug they can respond to. This has already been translated in an increased involvement of the pathologist in the patient care, a task that requires the specialist to provide not only a histopathological diagnosis but also the search for a set of mutations or expression of proteins that will give the indication to a specific drug therapy. The expansion of knowledge coming from mutational landscape, proteomic and phenotypical studies, coupled with increased availability of -omics technologies such NGS for DNA and RNA sequencing, mIHC for in situ detection of multiple proteins, mass spectrometry-based proteomics for a better characterization of proteins and for drug screening on organoid models will increase the pathologist's toolbox and reshape his role in patient's care.

5.1.3 Future perspectives on the use of affinity reagents

The ongoing problem of antibody validation is raising several questions that are waiting to be answered, first what should be the right way to validate antibodies? The second question that is being asked is who should validate the antibodies for the final use, a problem easy to address. It is the opinion of the author that the 2 main stakeholders involved in this are antibody producers and the final user, and both should participate in the process. Antibody producers (and sellers) need to guarantee that the application-specific performance is as expected, validating the antibody for the specified application with a good degree of certainty. If we consider IVD antibodies the manufacturer should be also responsible for the tissue-specific application and validate the antibody on the target tissue/tumor. At the same time, the end user should be accountable for assessing the performance of the antibody as expected and responsible for validating eventual new applications for research. A third option would be the institution of third-party entities responsible for antibody validation[170]. Such entities should be responsible for performing a standardized set of experiments, evaluate them according to well-established guidelines and publish the results in a public repository. Even if some attempts of third-party validation have been done by Euromabnet[171], there is no standard approach that has been proposed to regularize the antibody validation process.

6 ACKNOWLEDGEMENTS

A long and hard journey is now reaching its conclusion and it is time to thank all the people that in many ways have been walking along me in this 6 year-long trip. The PhD has brought me important and indelible experiences that I will never forget, during which I had to deal with 1000 different difficulties, and I had to learn really a lot, much more than I would have expected (and I have struggled much more I was expecting!).

I want to start thanking my main supervisor **Henrik Johansson** for being patient with me, teaching me mass spectrometry, how to analyze data from scratches, the attention to the detail (often God/Devil is in the details!) and changing/challenging many times my points of view. It has been thanks to his guidance if I have managed to survive across many experiments, write this thesis and becoming finally a scientist (I hope)!

A big thank you to my co-supervisor, "the boss of the lab" **Janne Lehtiö**, for his unique insights about science and a point of view completely different from mine (and financing my PhD program). Personal growth goes through resolving challenges (and surviving PhD programs), thanks for all of this and for the great suggestions I got from you!

Many thanks to **Mattias Vesterlund**, thanks for being my chairperson and for being a good scientist that never gives up! Great thanks to **Helena**, thanks for fixing all the 100 problems I got in the lab about administration! **Jenny** and **AnnSofi**, thanks for very nice discussions at lunch and about science!

Thanks to **Lukas Orre** for involving me in the lung cancer project and supporting me in the clinical proteomics, always great scientific discussions! Many thanks to **Maria Pernemalm** too, always great to have you around!

Dear **Santtu**, thank you for all the help, the good brainstorming and the suggestions I got during the last year, it has been really great working with you, seeing things from a different perspective is what makes people grow! Hope we put in practice asap some good ideas we thought about!

A great thanks to **Rui** for teaching me Mass Spectrometry and having great chats about science, history and many other fun topics!

A big thanks to **Rozbeh** for being a great example of scientist, for inspiring me and for sharing with me several spitfballs at the department parties (hope to repeat it soon!)

Thank you **Xiaofang** for all the help in the wetlab and for sharing with me tips about Chinese food! An enormous thanks to you **Yan** for everything you have done for me and for the fun moments, so happy to have you again in Scilifelab, I hope to see you more often here or at Mulle Mecks Lekpark! And thanks to you **Lena**, so many great moments in the lab and outside with you, hope to see you soon! Dear **Yanbo**, thanks for sharing with me night shift,

very good scientific discussions and being a great person! Thanks to **Lingjie** for being fun to be around and for helping me with the first MS sample preps!

A great thanks to **Mann** and **Husen** for helping me with coding and bioinformatic questions, hope to see you in the lab more often! A big thanks to **Luay**, I enjoyed our BBQs, hope to organize something soon in the next months! Dear **Nidhi**, it is always a pleasure to have you around in the lab and outside! **David**, thanks for all the discussions we had and for being a great person to be around!

A special thanks to **Georgios**, my friend I will never forget our fun together in so many occasions, not only the wetlab (where I learnt a lot from you) but also outside between drinks, barbeques and fun, hope to repeat asap when I am back, including kajaking (this time without sinking)! Dear **Aida**, thanks for being a fantastic friend and super supportive (and cooking very well), hope Ariana and Olivia can meet soon again! Thanks to you **Eduardo** my friend, from your S. Bestiale! I will never forget our whistling in the wetlab, the screaming during the move in the van and sharing picture of traditional dishes! Dear **Ioannis** and **Haris**, thanks for the fun you shared with me, for the cinemas in the Supernova and for remaining often until late in the lab! Hope to have more drinks with you guys soon (especially you Ioannis, that never want to go out)!

Thanks to you **Taner**, it has been great to have you around until you left for the "dark side" (the industry), hope to see you often, kajaking again with you (without sinking)! A big thanks to **Olena**, **Isabelle**, **Ali**, **Nooria**, **Sebastian, Ida** and **Yi**, it is always great to have you around the lab or going out with you guys, hope to repeat it soon! Thanks to you **Jorrit** for being my long lasting neighboor, sharing with me your culinary knowledge, answering my coding questions and showing me the funniest sides of the internet!

An enormous thanks to **Jan Mulder** and **Nick Mitsios** for hosting me in your fantastic lab, sharing your knowledge, patience, and fun with me, giving me the opportunity to work on my project, having great scientific discussions and fun moments with me! I will always be grateful for all of this, thank you for everything!

Dear **Laura** and **Christian**, it has been amazing to meet you guys, a lot of fun and very good barbeques, hope to see you soon (more often than we actually do!)

A great thanks to **Päivi** for being supportive with me, giving me so many ideas about mIHC and organizing my visit to the Kallioniemi's lab in Helsinki. Many thanks to **Emilie**, **Elisabeth** and **Francesco**, you guys helped me with reagents, ideas and good conversations!

Dear **Amena**, it has been a pleasure to work with you and fun to be around in general, I have enjoyed our collaboration! Thanks to you **Madeleine**, **Linnea**, **Hakim** and **Rajitha**, always great to have you around!

Many thanks to you **Ioannis**, **Dimitris** and **Theo**, I hope to start again to collaborate with you, I love doing research and sharing ideas about projects with you!

Ho un grande debito di riconoscenza verso di voi **Elisa**, **Claudia** e **Davide**! Grazie per esserci stati sempre, avermi insegnato sempre qualcosa, avermi anche (e soprattutto) sopportato ed essere stati grandi amici, non lo dimenticherò mai! Grazie tantissimo a voi **Carlo** e **Luca**, come posso dimenticare le serate al Fasching a sentire la Süperstar Orkestar o al Melt (o da nonna)?? Spero di rivedervi prestissimo (e più spesso)! Cara **Guya** e **Giovanni**, peccato ci abbia separato il COVID, speriamo di riuscire a vederci il più presto possibile, mi mancate!

And, finally, a big thank you to my wife **Angela** and my little daughter **Olivia** for being very supportive and believing in me, being around a person doing his PhD is hard, sometimes annoying and takes so much patience!

7 REFERENCES

1 Johansson, H. J. *et al.* Breast cancer quantitative proteome and proteogenomic landscape. *Nature communications* **10**, 1600, doi:10.1038/s41467-019-09018-y (2019).

2 Lehtiö, J. *et al.* Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune evasion mechanisms. *Nature Cancer* **Accepted in press** (2021).

3 Socciarelli, F., Mermelekas, G., Branca, R. M., Lehtiö, J. & Johansson, H. J. A method for validation of AntiBodies for immunohistochemistry using quantitative Mass Spectrometry based proteomics: ABMS. *Manuscript* (2021).

4 Hanson, A. & Haddad, L. M. in *StatPearls* (© 2021, StatPearls Publishing LLC., 2021).

5 Abrahams, E. Right drug-right patient-right time: personalized medicine coalition. *Clin Transl Sci* **1**, 11-12, doi:10.1111/j.1752-8062.2008.00003.x (2008).

6 Chow, N., Gallo, L. & Busse, J. W. Evidence-based medicine and precision medicine: Complementary approaches to clinical decision-making. *Precision Clinical Medicine* **1**, 60-64, doi:10.1093/pcmedi/pby009 (2018).

7 Ellis, M. J. *et al.* Connecting Genomic Alterations to Cancer Biology with Proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discovery* **3**, 1108, doi:10.1158/2159-8290.CD-13-0219 (2013).

8 Rodriguez, H., Zenklusen, J. C., Staudt, L. M., Doroshow, J. H. & Lowy, D. R. The next horizon in precision oncology: Proteogenomics to inform cancer diagnosis and treatment. *Cell* **184**, 1661-1670, doi:https://doi.org/10.1016/j.cell.2021.02.055 (2021).

9 Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians* **71**, 209-249, doi:https://doi.org/10.3322/caac.21660 (2021).

10 Beatson, G. T. The Treatment of Inoperable Carcinoma of the Female Mamma. *Glasgow medical journal* **76**, 81-87 (1911).

11 Lowenhaupt, E. & Steinbach, H. L. Clinical response of metastatic lesions of carcinoma of the female breast to hormonal therapy as related to histologic grade of malignancy. *Surg Gynecol Obstet* **88**, 291-294 (1949).

12 Carter, C. L., Allen, C. & Henson, D. E. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer* **63**, 181-187 (1989).

13 Cianfrocca, M. & Goldstein, L. J. Prognostic and predictive factors in early-stage breast cancer. *Oncologist* **9**, 606-616, doi:10.1634/theoncologist.9-6-606 (2004).

14 Elston, C. W. & Ellis, I. O. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* **19**, 403-410 (1991).

15 Hammond, M. E. *et al.* American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J Clin Oncol* **28**, 2784-2795, doi:10.1200/jco.2009.25.6529 (2010).

16 Wolff, A. C. *et al.* Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *Journal of Clinical Oncology* **36**, 2105-2122, doi:10.1200/JCO.2018.77.8738 (2018).

17 Penault-Llorca, F. & Radosevic-Robin, N. Ki67 assessment in breast cancer: an update. *Pathology* **49**, 166-171, doi:10.1016/j.pathol.2016.11.006 (2017).

18 Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747-752, doi:10.1038/35021093 (2000).

19 van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536, doi:10.1038/415530a (2002).

20 Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**, 8418-8423, doi:10.1073/pnas.0932692100 (2003).

21 Prat, A. *et al.* Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast* **24 Suppl 2**, S26-35, doi:10.1016/j.breast.2015.07.008 (2015).

22 Harris, L. N. *et al.* Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline. *Journal of Clinical Oncology* **34**, 1134-1150, doi:10.1200/jco.2015.65.2289 (2016).

23 Andre, F. *et al.* Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: ASCO Clinical Practice Guideline Update—Integration of Results From TAILORx. *Journal of Clinical Oncology* **37**, 1956-1964, doi:10.1200/JCO.19.00945 (2019).

24 Paik, S. *et al.* A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *New England Journal of Medicine* **351**, 2817-2826, doi:10.1056/NEJMoa041588 (2004).

25 Paik, S. *et al.* Gene Expression and Benefit of Chemotherapy in Women With Node-Negative, Estrogen Receptor–Positive Breast Cancer. *Journal of Clinical Oncology* **24**, 3726-3734, doi:10.1200/JCO.2005.04.7985 (2006).

26 Sparano, J. A. *et al.* Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. *The New England journal of medicine* **379**, 111-121, doi:10.1056/NEJMoa1804710 (2018).

27 Gluz, O. *et al.* West German Study Group Phase III PlanB Trial: First Prospective Outcome Data for the 21-Gene Recurrence Score Assay and Concordance of Prognostic Markers by Central and Local Pathology Assessment. *Journal of Clinical Oncology* **34**, 2341-2349, doi:10.1200/JCO.2015.63.5383 (2016).

28 Sparano, J. A. *et al.* Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. *New England Journal of Medicine* **373**, 2005-2014, doi:10.1056/NEJMoa1510764 (2015).

29 Henry, N. L. *et al.* Role of Patient and Disease Factors in Adjuvant Systemic Therapy Decision Making for Early-Stage, Operable Breast Cancer: Update of the ASCO Endorsement of the Cancer Care Ontario Guideline. *J Clin Oncol* **37**, 1965-1977, doi:10.1200/jco.19.00948 (2019).

30 Albain, K. S. *et al.* Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *The Lancet Oncology* **11**, 55-65, doi:10.1016/S1470-2045(09)70314-6 (2010).

31 Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

32 Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55-62, doi:10.1038/nature18003 (2016).

33 Krug, K. *et al.* Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. *Cell* **183**, 1436-1456 e1431, doi:10.1016/j.cell.2020.10.036 (2020).

34 Bouchal, P. *et al.* Breast Cancer Classification Based on Proteotypes Obtained by SWATH Mass Spectrometry. *Cell reports* **28**, 832-843.e837, doi:10.1016/j.celrep.2019.06.046 (2019).

35 Tyanova, S. *et al.* Proteomic maps of breast cancer subtypes. *Nature communications* **7**, 10259, doi:10.1038/ncomms10259 (2016).

36 Herbst, R. S., Morgensztern, D. & Boshoff, C. The biology and management of non-small cell lung cancer. *Nature* **553**, 446-454, doi:10.1038/nature25183 (2018).

37 Lung Cancer Survival Rates | 5-Year Survival Rates for Lung Cancer. (2021).

38 Govindan, R. *et al.* Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. *Cell* **150**, 1121-1134, doi:https://doi.org/10.1016/j.cell.2012.08.024 (2012).

39 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. *J Thorac Oncol* **10**, 1243-1260, doi:10.1097/jto.0000000000000630 (2015).

40 Hanna, N. *et al.* Systemic Therapy for Stage IV Non-Small-Cell Lung Cancer: American Society of Clinical Oncology Clinical Practice Guideline Update. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **35**, 3484-3515, doi:10.1200/jco.2017.74.6065 (2017).

41 Wakelee, H. A. *et al.* IMpower010: Primary results of a phase III global study of atezolizumab versus best supportive care after adjuvant chemotherapy in resected stage IB-IIIA non-small cell lung cancer (NSCLC). *Journal of Clinical Oncology* **39**, 8500-8500, doi:10.1200/JCO.2021.39.15_suppl.8500 (2021).

42 @US_FDA. FDA Approves First Adjuvant Therapy for Most Common Type of Lung Cancer | FDA. (2020).

43 Kalemkerian, G. P. *et al.* Molecular Testing Guideline for the Selection of Patients With Lung Cancer for Treatment With Targeted Tyrosine Kinase Inhibitors: American Society of Clinical Oncology Endorsement of the College of American Pathologists/International Association for the Study of Lung Cancer/Association for

Molecular Pathology Clinical Practice Guideline Update. *Journal of Clinical Oncology* **36**, 911-919, doi:10.1200/JCO.2017.76.7293 (2018).

44 Hanna, N. H. *et al.* Therapy for Stage IV Non–Small-Cell Lung Cancer With Driver Alterations: ASCO and OH (CCO) Joint Guideline Update. *Journal of Clinical Oncology* **39**, 1040-1091, doi:10.1200/JCO.20.03570 (2021).

45 Hanna, N. H. *et al.* Therapy for Stage IV Non–Small-Cell Lung Cancer Without Driver Alterations: ASCO and OH (CCO) Joint Guideline Update. *Journal of Clinical Oncology* **38**, 1608-1632, doi:10.1200/JCO.19.03022 (2020).

46 Sharma, S. V., Bell, D. W., Settleman, J. & Haber, D. A. Epidermal growth factor receptor mutations in lung cancer. *Nature reviews. Cancer* **7**, 169-181, doi:10.1038/nrc2088 (2007).

47 Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550, doi:10.1038/nature13385 (2014).

48 Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).

49 Pao, W. *et al.* KRAS mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib. *PLoS medicine* **2**, e17, doi:10.1371/journal.pmed.0020017 (2005).

50 Kwak, E. L. *et al.* Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *The New England journal of medicine* **363**, 1693-1703, doi:10.1056/NEJMoa1006448 (2010).

51 Shaw, A. T. *et al.* Crizotinib in ROS1-rearranged non-small-cell lung cancer. *The New England journal of medicine* **371**, 1963-1971, doi:10.1056/NEJMoa1406766 (2014).

52 Cardarella, S. *et al.* Clinical, pathologic, and biologic features associated with BRAF mutations in non-small cell lung cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **19**, 4532-4540, doi:10.1158/1078-0432.ccr-13-0657 (2013).

53 Frampton, G. M. *et al.* Activation of MET via diverse exon 14 splicing alterations occurs in multiple tumor types and confers clinical sensitivity to MET inhibitors. *Cancer discovery* **5**, 850-859, doi:10.1158/2159-8290.cd-15-0285 (2015).

54 Mazieres, J. *et al.* Lung cancer that harbors an HER2 mutation: epidemiologic characteristics and therapeutic perspectives. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **31**, 1997-2003, doi:10.1200/jco.2012.45.6095 (2013).

55 Kohno, T. *et al.* KIF5B-RET fusions in lung adenocarcinoma. *Nature medicine* **18**, 375-377, doi:10.1038/nm.2644 (2012).

56 Ricciuti, B. *et al.* Targeting NTRK fusion in non-small cell lung cancer: rationale and clinical evidence. *Medical oncology (Northwood, London, England)* **34**, 105, doi:10.1007/s12032-017-0967-5 (2017).

57 Reck, M. *et al.* Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *The New England journal of medicine* **375**, 1823-1833, doi:10.1056/NEJMoa1606774 (2016).

58 Herbst, R. S. *et al.* Atezolizumab for First-Line Treatment of PD-L1–Selected Patients with NSCLC. *New England Journal of Medicine* **383**, 1328-1339, doi:10.1056/NEJMoa1917346 (2020).

59 Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).

60 Campbell, J. D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics* **48**, 607-616, doi:10.1038/ng.3564 (2016).

61 Gillette, M. A. *et al.* Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell* **182**, 200-225.e235, doi:10.1016/j.cell.2020.06.013 (2020).

62 Satpathy, S. *et al.* A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184**, 4348-4371.e4340, doi:https://doi.org/10.1016/j.cell.2021.07.016 (2021).

63 Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82, doi:10.1126/science.aaf2403 (2016).

64 Vickovic, S. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* **16**, 987-990, doi:10.1038/s41592-019-0548-y (2019).

65 Giacomello, S. *et al.* Spatially resolved transcriptome profiling in model plant species. *Nature Plants* **3**, 17061, doi:10.1038/nplants.2017.61 (2017).

66 Ke, R. *et al.* In situ sequencing for RNA analysis in preserved tissue and cells. *Nature methods* **10**, 857-860, doi:10.1038/nmeth.2563 (2013).

67 Schulz, S., Becker, M., Groseclose, M. R., Schadt, S. & Hopf, C. Advanced MALDI mass spectrometry imaging in pharmaceutical research and drug development. *Current Opinion in Biotechnology* **55**, 51-59, doi:https://doi.org/10.1016/j.copbio.2018.08.003 (2019).

68 Piehowski, P. D. *et al.* Automated mass spectrometry imaging of over 2000 proteins from tissue sections at 100-µm spatial resolution. *Nature communications* **11**, 8, doi:10.1038/s41467-019-13858-z (2020).

69 Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968-981 e915, doi:10.1016/j.cell.2018.07.010 (2018).

70 Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nature medicine* **20**, 436-442, doi:10.1038/nm.3488 (2014).

71 Baharlou, H., Canete, N. P., Cunningham, A. L., Harman, A. N. & Patrick, E. Mass Cytometry Imaging for the Study of Human Diseases—Applications and Data Analysis Strategies. *Frontiers in Immunology* **10**, doi:10.3389/fimmu.2019.02657 (2019).

72 Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine* **25**, 1054-1056, doi:10.1038/s41591-019-0462-y (2019).

73 Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine* **24**, 1559-1567, doi:10.1038/s41591-018-0177-5 (2018).

74 Hong, R., Liu, W., DeLair, D., Razavian, N. & Fenyö, D. Predicting endometrial cancer subtypes and molecular features from histopathology images using multi-resolution deep learning models. *Cell Reports Medicine* **2**, doi:10.1016/j.xcrm.2021.100400 (2021).

75 Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer* **1**, 800-810, doi:10.1038/s43018-020-0085-8 (2020).

76 Kather, J. N. *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer* **1**, 789-799, doi:10.1038/s43018-020-0087-6 (2020).

77 van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nature medicine* **27**, 775-784, doi:10.1038/s41591-021-01343-4 (2021).

78 Jiang, Y., Yang, M., Wang, S., Li, X. & Sun, Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer communications (London, England)* **40**, 154-166, doi:10.1002/cac2.12012 (2020).

79 Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163**, 506-519, doi:10.1016/j.cell.2015.09.033 (2015).

80 McCart Reed, A. E., Kutasovic, J. R., Lakhani, S. R. & Simpson, P. T. Invasive lobular carcinoma of the breast: morphology, biomarkers and 'omics. *Breast Cancer Res* **17**, 12, doi:10.1186/s13058-015-0519-x (2015).

81 *Welcome to The Cancer Imaging Archive - The Cancer Imaging Archive (TCIA)*, <https://www.cancerimagingarchive.net/> (2021).

82 Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355, doi:10.1038/nature19949 (2016).

83 *Mass Spectrometry Data Analysis in Proteomics*. Vol. 2051 (Springer New York, New York, NY, 2020).

84 Gillet, L. C., Leitner, A. & Aebersold, R. Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annual review of analytical chemistry (Palo Alto, Calif.)* **9**, 449-472, doi:10.1146/annurev-anchem-071015-041535 (2016).

85 Murray, K. (Wikipedia, 2006).

86 Yamashita, M. & Fenn, J. B. Electrospray ion source. Another variation on the free-jet theme. *The Journal of Physical Chemistry* **88**, 4451-4459, doi:10.1021/j150664a002 (1984).

87 Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical Reviews* **113**, 2343-2394, doi:10.1021/cr3003533 (2013).

88 Griss, J. Spectral library searching in proteomics. *Proteomics* **16**, 729-740, doi:10.1002/pmic.201500296 (2016).

89 Matthiesen, R. *Mass Spectrometry Data Analysis in Proteomics*. 2nd ed. 2013 edn, Vol. 1007 (Totowa, NJ: Humana Press, 2013).

90 Ankney, J. A., Muneer, A. & Chen, X. Relative and Absolute Quantitation in Mass Spectrometry–Based Proteomics. *Annual Review of Analytical Chemistry* **11**, 49-77, doi:10.1146/annurev-anchem-061516-045357 (2016).

91 Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry* **389**, 1017-1031, doi:10.1007/s00216-007-1486-6 (2007).

92 Rauniyar, N. & Yates, J. R., 3rd. Isobaric labeling-based relative quantification in shotgun proteomics. *J Proteome Res* **13**, 5293-5309, doi:10.1021/pr500880b (2014).

93 Old, W. M. *et al.* Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics. *Molecular & Cellular Proteomics* **4**, 1487, doi:10.1074/mcp.M500084-MCP200 (2005).

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

95 Addona, T. A. *et al.* Multi-site assessment of the precision and reproducibility of multiple reaction monitoring–based measurements of proteins in plasma. *Nature Biotechnology* **27**, 633-641, doi:10.1038/nbt.1546 (2009).

96 Ludwig, C. *et al.* Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular Systems Biology* **14**, e8126, doi:10.15252/msb.20178126 (2018).

97 Ting, Y. S. *et al.* Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Molecular & Cellular Proteomics* **14**, 2301, doi:10.1074/mcp.O114.047035 (2015).

98 Gillet, L. C. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics* **11**, O111.016717, doi:10.1074/mcp.O111.016717 (2012).

99 Tetin, S. Y. & Stroupe, S. D. Antibodies in diagnostic applications. *Current pharmaceutical biotechnology* **5**, 9-16, doi:10.2174/1389201043489602 (2004).

100 Van Regenmortel, M. H. V. Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition. *Journal of Molecular Recognition* **27**, 627-639, doi:10.1002/jmr.2394 (2014).

101 Landsteiner, K. in *Die Spezifizität der Serologischen Reaktionen* 123 (Springer-Verlag, 1933).

102 Tarlinton, D. M. & Smith, K. G. Dissecting affinity maturation: a model explaining selection of antibody-forming cells and memory B cells in the germinal centre. *Immunology today* **21**, 436-441, doi:10.1016/s0167-5699(00)01687-x (2000).

103 Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. THE STRUCTURAL BASIS OF ANTIGEN-ANTIBODY RECOGNITION. *Annual Review of Biophysics and Biophysical Chemistry* **16**, 139-159, doi:10.1146/annurev.bb.16.060187.001035 (1987).

104 Dimitrov, J. D. *et al.* Antibodies use heme as a cofactor to extend their pathogen elimination activity and to acquire new effector functions. *The Journal of biological chemistry* **282**, 26696-26706, doi:10.1074/jbc.M702751200 (2007).

105 James, L. C. & Tawfik, D. S. The specificity of cross-reactivity: Promiscuous antibody binding involves specific hydrogen bonds rather than nonspecific

hydrophobic stickiness. *Protein Science* **12**, 2183-2193, doi:10.1110/ps.03172703 (2003).

106 Jain, D. & Salunke, D. M. Antibody specificity and promiscuity. *Biochemical Journal* **476**, 433-447, doi:10.1042/bcj20180670 (2019).

107 Manivel, V., Sahoo, N. C., Salunke, D. M. & Rao, K. V. Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site. *Immunity* **13**, 611-620, doi:10.1016/s1074-7613(00)00061-3 (2000).

108 Oberg, A. L. *et al.* Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J Proteome Res* **7**, 225-233, doi:10.1021/pr700734f (2008).

109 Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**, Article3, doi:10.2202/1544-6115.1027 (2004).

110 Daly, D. S. *et al.* Mixed-effects statistical model for comparative LC-MS proteomics studies. *J Proteome Res* **7**, 1209-1217, doi:10.1021/pr070441i (2008).

111 Zhu, Y. *et al.* DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis. *Molecular & cellular proteomics : MCP* **19**, 1047-1057, doi:10.1074/mcp.TIR119.001646 (2020).

112 Handler, D. C. *et al.* The Art of Validating Quantitative Proteomics Data. *Proteomics* **18**, e1800222, doi:10.1002/pmic.201800222 (2018).

113 Devita, V. T. J., Rosenberg, S. A. & Lawrence, T. S. *DeVita, Hellman, and Rosenberg's Cancer: Principles & Practice of Oncology*. 11th edn, (Lippincott Williams and Wilkins, 2018).

114 Valencia, O. M. *et al.* The Role of Genetic Testing in Patients With Breast Cancer: A Review. *JAMA surgery* **152**, 589-594, doi:10.1001/jamasurg.2017.0552 (2017).

115 Hertz, D. L. & Rae, J. Pharmacogenetics of Cancer Drugs. *Annual Review of Medicine* **66**, 65-81, doi:10.1146/annurev-med-053013-053944 (2015).

116 Chen, D. S. & Mellman, I. Elements of cancer immunity and the cancer-immune set point. *Nature* **541**, 321-330, doi:10.1038/nature21349 (2017).

117 Blackadar, C. B. Historical review of the causes of cancer. *World J Clin Oncol* **7**, 54-86, doi:10.5306/wjco.v7.i1.54 (2016).

118 Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics* **69**, 89-95, doi:10.1067/mcp.2001.113989 (2001).

119 Masucci, G. V. *et al.* Validation of biomarkers to predict response to immunotherapy in cancer: Volume I - pre-analytical and analytical validation. *Journal for immunotherapy of cancer* **4**, 76, doi:10.1186/s40425-016-0178-1 (2016).

120 Dobbin, K. K. *et al.* Validation of biomarkers to predict response to immunotherapy in cancer: Volume II - clinical validation and regulatory considerations. *Journal for immunotherapy of cancer* **4**, 77, doi:10.1186/s40425-016-0179-0 (2016).

121 Forshed, J. Experimental Design in Clinical 'Omics Biomarker Discovery. *J Proteome Res* **16**, 3954-3960, doi:10.1021/acs.jproteome.7b00418 (2017).

122 Lee, J. W. *et al.* Fit-for-purpose method development and validation for successful biomarker measurement. *Pharmaceutical research* **23**, 312-328, doi:10.1007/s11095-005-9045-3 (2006).

123 Mattes, W. B. & Goodsaid, F. Regulatory landscapes for biomarkers and diagnostic tests: Qualification, approval, and role in clinical practice. *Exp Biol Med (Maywood)* **243**, 256-261, doi:10.1177/1535370217739629 (2018).

124 Baker, M. in *Nature* Vol. 521 274-276 (2015).

125 Bordeaux, J. *et al.* Antibody validation. *BioTechniques* **48**, 197-209, doi:10.2144/000113382 (2010).

126 Uhlen, M. *et al.* A proposal for validation of antibodies. *Nature methods* **13**, 823-+, doi:10.1038/nmeth.3995 (2016).

127 Algenas, C. *et al.* Antibody performance in western blot applications is context-dependent. *Biotechnology Journal* **9**, 435-445, doi:10.1002/biot.201300341 (2014).

128 Stadler, C. *et al.* Systematic validation of antibody binding and protein subcellular localization using siRNA and confocal microscopy. *Journal of Proteomics* **75**, 2236-2251, doi:10.1016/j.jprot.2012.01.030 (2012).

129 Parker, C. E. & Borchers, C. H. Mass spectrometry based biomarker discovery, verification, and validation--quality assurance and control of protein biomarker assays. *Molecular oncology* **8**, 840-858, doi:10.1016/j.molonc.2014.03.006 (2014).

130 Marcon, E. *et al.* Assessment of a method to characterize antibody selectivity and specificity for use in immunoprecipitation. *Nature methods* **12**, 725-731, doi:10.1038/nmeth.3472 (2015).

131 Skogs, M. *et al.* Antibody Validation in Bioimaging Applications Based on Endogenous Expression of Tagged Proteins. *Journal of Proteome Research* **16**, 147-155, doi:10.1021/acs.jproteome.6b00821 (2017).

132 Orre, L. M. *et al.* SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. *Molecular cell* **73**, 166-182 e167, doi:10.1016/j.molcel.2018.11.035 (2019).

133 Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, doi:10.1126/science.aal3321 (2017).

134 Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535-550, doi:10.1016/j.cell.2016.03.014 (2016).

135 Edfors, F. *et al.* Enhanced validation of antibodies for research applications. *Nature communications* **9**, 4130, doi:10.1038/s41467-018-06642-y (2018).

136 Hoofnagle, A. N. *et al.* Multiple-Reaction Monitoring–Mass Spectrometric Assays Can Accurately Measure the Relative Protein Abundance in Complex Mixtures. *Clinical chemistry* **58**, 777, doi:10.1373/clinchem.2011.173856 (2012).

137 Hembrough, T. *et al.* Selected Reaction Monitoring (SRM) Analysis of Epidermal Growth Factor Receptor (EGFR) in Formalin Fixed Tumor Tissue. *Clinical proteomics* **9**, 5, doi:10.1186/1559-0275-9-5 (2012).

138 Hammond, T. G. *et al.* Development and characterization of a pseudo multiple reaction monitoring method for the quantification of human uromodulin in urine. *Bioanalysis* **8**, 1279-1296, doi:10.4155/bio-2016-0055 (2016).

139 Fu, Q. *et al.* An Empirical Approach to Signature Peptide Choice for Selected Reaction Monitoring: Quantification of Uromodulin in Urine. *Clinical chemistry* **62**, 198 (2016).

140 Andersson, A. *et al.* Development of parallel reaction monitoring assays for cerebrospinal fluid proteins associated with Alzheimer's disease. *Clinica Chimica Acta* **494**, 79-93, doi:https://doi.org/10.1016/j.cca.2019.03.243 (2019).

141 Végvári, Á. *et al.* Identification of a Novel Proteoform of Prostate Specific Antigen (SNP-L132I) in Clinical Samples by Multiple Reaction Monitoring. *Molecular & Cellular Proteomics* **12**, 2761, doi:10.1074/mcp.M113.028365 (2013).

142 Toki, M. I., Cecchi, F., Hembrough, T., Syrigos, K. N. & Rimm, D. L. Proof of the quantitative potential of immunofluorescence by mass spectrometry. *Laboratory Investigation* **97**, 329-334, doi:10.1038/labinvest.2016.148 (2017).

143 Calderón-Celis, F., Encinar, J. R. & Sanz-Medel, A. Standardization approaches in absolute quantitative proteomics with mass spectrometry. *Mass spectrometry reviews* **37**, 715-737, doi:10.1002/mas.21542 (2018).

144 World Medical, A. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA* **310**, 2191-2194, doi:10.1001/jama.2013.281053 (2013).

145 Johansson, I. & Lynøe, N. *Medicine and Phylosophy A Twenty-FIrst Century Introduction*. 475 (Ontos Verlag, 2008).

146 Aure, M. R. *et al.* Integrated analysis reveals microRNA networks coordinately expressed with key proteins in breast cancer. *Genome medicine* **7**, 21, doi:10.1186/s13073-015-0135-5 (2015).

147 Wiedswang, G. *et al.* Detection of Isolated Tumor Cells in Bone Marrow Is an Independent Prognostic Factor in Breast Cancer. *Journal of Clinical Oncology* **21**, 3469-3478, doi:10.1200/jco.2003.02.009 (2003).

148 Karlsson, A. *et al.* Gene Expression Profiling of Large Cell Lung Cancer Links Transcriptional Phenotypes to the New Histological WHO 2015 Classification. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **12**, 1257-1267, doi:10.1016/j.jtho.2017.05.008 (2017).

149 Halvorsen, A. R. *et al.* TP53 Mutation Spectrum in Smokers and Never Smoking Lung Cancer Patients. *Frontiers in Genetics* **7**, doi:10.3389/fgene.2016.00085 (2016).

150 Wiśniewski, J. R. Filter-Aided Sample Preparation for Proteome Analysis. *Methods in molecular biology (Clifton, N.J.)* **1841**, 3-10, doi:10.1007/978-1-4939-8695-8_1 (2018).

151 Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nature Protocols* **14**, 68-85, doi:10.1038/s41596-018-0082-x (2019).

152 Branca, R. M. M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature methods* **11**, 59, doi:10.1038/nmeth.2732

153 Taylor, C. R. Immunohistochemistry in surgical pathology: principles and practice. *Methods Mol Biol* **1180**, 81-109, doi:10.1007/978-1-4939-1050-2_5 (2014).

154 Dabbs, D. J. *Diagnostic Immunohistochemistry: Theranostic and Genomic Applications*. 5th edn, (Elsevier, 2018).

155 Hofman, F. M. & Taylor, C. R. Immunohistochemistry. *Curr Protoc Immunol* **103**, 21 24 21-21 24 26, doi:10.1002/0471142735.im2104s103 (2013).

156 Engel, K. B. & Moore, H. M. Effects of preanalytical variables on the detection of proteins by immunohistochemistry in formalin-fixed, paraffin-embedded tissue. *Arch Pathol Lab Med* **135**, 537-543, doi:10.1043/2010-0702-rair.1 (2011).

157 Shi, S. R., Shi, Y. & Taylor, C. R. Antigen retrieval immunohistochemistry: review and future prospects in research and diagnosis over two decades. *J Histochem Cytochem* **59**, 13-32, doi:10.1369/jhc.2010.957191 (2011).

158 Board, W. C. o. T. E. *Breast Tumours*. 5th edn, Vol. 2 (2019).

159 *WHO Classification of Tumours of the Lung, Pleura, Thymus and Heart*. Fourth edn, Vol. 7 (WHO Press, 2015).

160 Travis, W. D. *et al.* International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol* **6**, 244-285, doi:10.1097/JTO.0b013e318206a221 (2011).

161 Al-Shibli, K. I. *et al.* Prognostic effect of epithelial and stromal lymphocyte infiltration in non-small cell lung cancer. *Clin Cancer Res* **14**, 5220-5227, doi:10.1158/1078-0432.ccr-08-0133 (2008).

162 Dako, A. (ed Agilent Technologies) (Agilent, United States, 2018).

163 Lee, H. J. *et al.* Tertiary lymphoid structures: prognostic significance and relationship with tumour-infiltrating lymphocytes in triple-negative breast cancer. *Journal of Clinical Pathology* **69**, 422, doi:10.1136/jclinpath-2015-203089 (2016).

164 Yang, Z., Samanta, S., Yan, W., Yu, B. & Qu, J. in *Optical Imaging in Human Disease and Biological Research* (eds Xunbin Wei & Bobo Gu) 23-43 (Springer Singapore, 2021).

165 Göttfert, F. *et al.* Coaligned dual-channel STED nanoscopy and molecular diffusion analysis at 20 nm resolution. *Biophys J* **105**, L01-03, doi:10.1016/j.bpj.2013.05.029 (2013).

166 Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer cell* **33**, 690-705 e699, doi:10.1016/j.ccell.2018.03.014 (2018).

167 Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications* **4**, 2612, doi:10.1038/ncomms3612 (2013).

168 Charoentong, P. *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell reports* **18**, 248-262, doi:10.1016/j.celrep.2016.12.019 (2017).

169 Sautes-Fridman, C., Petitprez, F., Calderaro, J. & Fridman, W. H. Tertiary lymphoid structures in the era of cancer immunotherapy. *Nature reviews. Cancer* **19**, 307-325, doi:10.1038/s41568-019-0144-6 (2019).

170 Laflamme, C., Edwards, A. M., Bandrowski, A. E. & McPherson, P. S. Opinion: Independent third-party entities as a model for validation of commercial antibodies. *New biotechnology* **65**, 1-8, doi:10.1016/j.nbt.2021.07.001 (2021).

171 Roncador, G. *et al.* The European antibody network's practical guide to finding and validating suitable antibodies for research. *mAbs* **8**, 27-36, doi:10.1080/19420862.2015.1100787 (2016).