Algorithms and Methods for Robust Processing and Analysis of Mass Spectrometry Data

Eriksson, Jonatan

2021

Document Version:
Publisher's PDF, also known as Version of record

Algorithms and Methods for Robust Processing and Analysis of Mass Spectrometry Data

by Jonatan O. Eriksson

LUND
UNIVERSITY

Dissertation for the degree of Doctor of Philosophy in Biomedical Engineering.
Thesis advisors: Professor György Marko-Varga,
Associate Professor Peter Horvatovich,
Associate Professor Krzysztof Pawlowski
Faculty opponent: Associate Professor Liam McDonnell

To be presented, with the permission of the Faculty of Engineering of Lund University, for public criticism in Segerfalkssalen, BMC A10, Sölvegatan 17, Lund, the 11th of June 2021 at 13:00.

Organization
LUND UNIVERSITY

Department of Biomedical Engineering
Box 118
SE–221 00 Lund, Sweden

Document name
DOCTORAL DISSERTATION

Date of disputation
2021-06-11

Sponsoring organization

Author(s)
Jonatan O. Eriksson

Title and subtitle
Algorithms and Methods for Robust Processing and Analysis of Mass Spectrometry Data

Abstract
Liquid chromatography-mass spectrometry (LC-MS) and mass spectrometry imaging (MSI) are two techniques that are routinely used to study proteins, peptides, and metabolites at a large scale. Thousands of biological compounds can be identified and quantified in a single experiment with LC-MS, but many studies fail to convert this data to a better understanding of disease biology. One of the primary reasons for this is low reproducibility, which in turn is partially due to inaccurate and/or inconsistent data processing. Protein biomarkers and signatures for various types of cancer are frequently discovered with LC-MS, but their behavior in independent cohorts is often inconsistent to that in the discovery cohort. Biomarker candidates must be thoroughly validated in independent cohorts, which makes the ability to share data across different laboratories crucial to the future success of the MS-based research fields. The emergence and growth of public repositories for MSI data is a step in the right direction. Still, many of those data sets remain incompatible one another due to inaccurate or incompatible preprocessing strategies. Ensuring compatibility between data generated in different labs is therefore necessary to gain access to the full potential of MS-based research. In two of the studies that I present in this thesis, we used LC-MS to characterize lymph node metastases from individuals with melanoma. Furthermore, my thesis work has resulted in two novel preprocessing methods for MSI data sets. The first one is a peak detection method that achieves considerably higher sensitivity for faintly expressed compounds than existing methods, and the second one is a accurate, robust, and general approach to mass alignment. Both algorithms deliberately rely on centroid spectra, which makes them compatible with most shared data sets. I believe that the improvements demonstrated by these methods can lead to a higher reproducibility in the MS-based research fields, and, ultimately, to a better understanding of disease processes.

Key words
mass spectrometry, software, algorithms, signal processing, proteomics, metabolomics

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

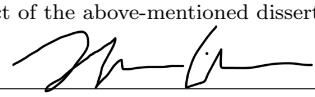
ISSN and key title

ISBN
978-91-7895-920-4 (print)
978-91-7895-919-8 (pdf)

Recipient's notes

Price

Security classification

Signature 

Date 2021-05-18

Algorithms and Methods for Robust Processing and Analysis of Mass Spectrometry Data

by Jonatan O. Eriksson

LUND
UNIVERSITY

Cover illustration front: Protein concept art. (Credits: Karin Yip).

Funding information: The thesis work was financially supported by Fru Berta Kamprads Stiftelse.

Faculty of Engineering, Department of Biomedical Engineering

Report No. 1/21
ISRN: LUTEDX/TEEM-1123-SE
ISBN: 978-91-7895-920-4(print)
ISBN: 978-91-7895-919-8(pdf)

Printed in Sweden by *E-husets Tryckeri*, Lund University, Lund 2021

Contents

List of publications . . . iii
Acknowledgements . . . v

Chapter 1: Introduction **1**
Biology and Medicine . . . 2
Biomarker Discovery . . . 4
Mass Spectrometry . . . 5
Aims and Contributions of This Thesis . . . 10

Chapter 2: Data Processing in LC-MS **11**
Flavors of LC-MS . . . 12
Processing LC-MS Single-Stage Spectra . . . 14
Searching for Matches in Sequence Databases . . . 15
Proteogenomics . . . 18

Chapter 3: Data Processing and Analysis in MSI **19**
Peak Detection . . . 21
Annotating Features . . . 24
Normalization and Quantification . . . 25

Chapter 4: Few Samples with Many Variables **27**
Dealing with High-Dimensional Data . . . 28
Statistical Hypothesis Testing . . . 29
Predictive Modeling . . . 32
Cross-Validation . . . 34
Survival Analysis . . . 35

Chapter 5: Summary of Papers **41**
Summary of Paper I . . . 41
Summary of Paper II . . . 42
Summary of Paper III . . . 44
Summary of Paper IV . . . 49

Conclusions and Future Perspectives 50
Populärvetenskaplig Sammanfattning 52

List of publications

This thesis is based on the following publications, referred to by their Roman numerals:

I **Improved survival prognostication of node-positive malignant melanoma patients utilizing shotgun proteomics guided by histopathological characterization and genomic data**
Lazaro Hiram Betancourt, Krzysztof Pawłowski, **Jonatan Eriksson**, A Marcell Szasz, Shamik Mitra, Indira Pla, Charlotte Welinder, Henrik Ekedahl, Per Broberg, Roger Appelqvist, Maria Yakovleva, Yutaka Sugihara, Kenichi Miharada, Christian Ingvar, Lotta Lundgren, Bo Baldetorp, Håkan Olsson, Melinda Rezeli, Elisabet Wieslander, Peter Horvatovich, Johan Malm, Göran Jönsson, György Marko-Varga
Shared first author, Scientific Reports, 2019, pp. 1–10

II **Clusterwise Peak Detection and Filtering Based on Spatial Distribution To Efficiently Mine Mass Spectrometry Imaging Data**
Jonatan O. Eriksson, Melinda Rezeli, Max Hefner, György Marko-Varga, and Peter Horvatovich
First author, Analytical Chemistry, 2019, pp. 2–30

III **MSIWarp: a general approach to mass alignment in mass spectrometry imaging**
Jonatan O. Eriksson, Alejando Sanchez Brotons, Melinda Rezeli, Frank Suits, György Marko-Varga, and Peter Horvatovich
First author, Analytical Chemistry, 2020, pp. 1–10

IV **Proteogenomic and Histopathologic Classification of Malignant Melanoma Reveal Molecular Heterogeneity Impacting Survival**
Magdalena Kuras, Lazaro Hiram Betancourt, Runyu Hong, Jimmy Rodriguez, Leticia Szadai, Peter Horvatovich, Indira Pla, **Jonatan Eriksson**, Beáta Szeitz, Bartek Deszcz, Yutaka Sugihara, Henrik Ekedahl, Bo Baldetorp, Christian Ingvar, Håkan Olsson, Lotta Lundgren, Göran Jönsson, Henrik Lindberg, Henriett Oskolas, Zsolt Horvath, Melinda Rezeli, Jeovanis Gil, Johan Malm, Aniel Sanchez, Marcell Szasz, Krzysztof Pawłowski, Elisabet Wieslander, David Fenyö, Istvan Nemeth, György Marko-Varga
Coauthor, manuscript

Acknowledgements

I want to thank my main supervisor, György Marko-Varga, for not being afraid of trying out new ideas, and my co-supervisor Krzysztof Pawlowski for having the patience and capacity to steer the more challenging projects toward their goals. I want to, in particular, thank my co-supervisor Peter Horvatovich for guiding me with expert knowledge and immense enthusiasm for science. Without your supervision this thesis would have been much thinner. Thank you Alex and Frank for enriching my PhD with your excellence. Finally, thank you Melinda for all the help with the experimental parts of my projects and for attempting to explain biology and chemistry to me.

In no particular order, I want to thank: Roger, for our casual chats that have brought valuable stress relief. Sugi-san, my first friend at BMC, for introducing me to curry rice and other tasty dishes. Past and present BME/BMC PhDs and Post Docs, Thomas, Joeri, Elin, Hannicka, Maria, Isabella, Moritz, and Billy for some pretty severe hangovers. My colleagues Magdalena, Barbara, Indira, Aniel, Lazaro, Henriette, Nicole, Kim, Boram, Andy, Doctor Wu (Who), for the friendliness and competence you bring to the group. Ping Li, Premkumar Siddhuraj, Naveen Ravi, and others for the spicy hot pots, barbeque sessions, and good times at the nerd gym.

Many thanks to my parents, step parents, siblings, and cousins for your support, especially to Bosse for introducing me to the field of mass spectrometry, to my friends for all work-unrelated adventures, and, of course, to my one and only Karin for your love, support, patience, and so many other things.

Chapter 1: Introduction

Biology is complex. Even simple life forms display intricate behaviors that are difficult to fully comprehend. The human body is an advanced organism composed of a vast number of molecules that interact with one another, form cells and tissue, and carry out various functions. The ability of such a biological system to remain reasonably stable is one of nature's true miracles. Medical research is tightly coupled to biology, and its objective is to provide understanding of, and ultimately control over, disease processes and other physiological phenomena. This is an incredibly difficult task.

Cancer is one of the most grief-causing diseases worldwide, and it takes many forms. In most cases it is thought to be driven by mutations that lead to the proliferation of disobedient cells, whose increasingly fast spread devastates the body unless stopped. Fortunately, a massive research effort has led to longer survival rates and an improved quality of life for many cancer-afflicted individuals. This has been enabled by the development of effective treatments that target specific cancer subtypes and diagnostic tools that enable an early detection of the disease. However, early detection is not yet guaranteed, and prognosis often remains poor when the disease is detected at a late stage.

Historically, biology has been studied at a component level; the behavior of a single compound is observed under various conditions, and the behavior of the compound's environment is extrapolated from those observations. The advent of DNA and RNA sequencing techniques brought on a new era; with these techniques, thousands of genes and transcripts can be measured in the same experiment, and the behavior of the system as a whole can be studied more directly.[1;2] Similarly, mass spectrometry (MS) is a technique that can measure thousands of peptides, proteins, metabolites, and other molecules in tissue samples.[3]

An MS experiment involves several nontrivial steps. These include carefully preparing the samples for MS analysis, calibrating and configuring the instrument, processing the mass spectra to identify and quantify compounds, statistical analysis to determine which compounds are related to the research question, and finally interpreting the results in a biological context. Errors

introduced at an early stage in the experimental pipeline are hard to remove or correct for in subsequent stages. Thus, it is critical that the experimental design is sound and that each step is carried out with great care.

In many ways, MS-based research resides at the interface between multiple disciplines; it is often used to answer biological or medical questions, but expert knowledge of chemistry, physics, and engineering is required to apply it successfully. During my thesis project I have focused on the development and evaluation of techniques that enable accurate preprocessing of MS data, and on the application of thorough statistical analysis to the final protein/peptide expression data in the context of clinical research. Before diving into all the details of MS-based biological research, however, I will briefly reiterate some of the fundamental concepts of biology.

Biology and Medicine

All life forms, as we define them, are different constellations of one or more cells. A cell is a living being in itself: it reproduces by replicating itself, maintains its genetic integrity through DNA repair, grows by metabolizing nutrients, and, importantly, synthesizes proteins. Proteins are instrumental to any organism since they carry out the majority of functions, and they are synthesized through two sequential processes: transcription and translation (Figure 1). During transcription, nucleotide sequences (genes) in the DNA are read and used to produce strands of RNA (transcripts), which in turn are converted to amino acid sequences during translation. Finally, the amino acid sequences are folded into three-dimensional structures to yield functional proteins. Some proteins are left either completely or partially unfolded, often due to post-translational modifications (PTMs), and these proteins were previously thought to be dysfunctional but are now known to have distinctive functions.[4;5] After synthesis, some proteins remain inside the cell and carry out intracellular functions while others are exported from the cell to perform extracellular functions.

The human body, an organ, and an individual cell can all be thought of as increasingly complex biological systems. The human body is a collection of organs with various functions and each organ is in turn composed of a vast number of cells grouped by function or type. Knowledge about an individual component, such as a protein complex, a cell type, or an enzyme, can potentially be used to diagnose or treat diseases. Indeed, some diseases are caused by a disturbance in the state of a single component that propagates to other components and ultimately affects many parts of the body. The traditional pathological model of Parkinson's disease represents such an example: misfolding of the protein alpha-synuclein causes it to attach to other alpha-synuclein, forming cytotoxic clumps

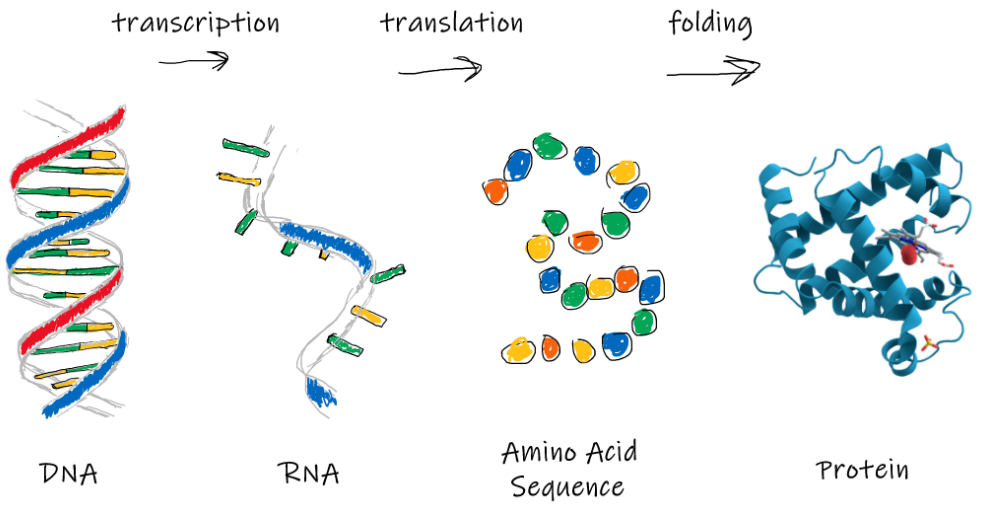


Figure 1: Protein synthesis.

called Lewy bodies. In the late stages of the disease, the damage caused by the Lewy bodies affect numerous parts of the brain, which severely degrades the cognitive ability and motor function of the afflicted individual. Traditionally, biological and medical research has been focused on studying small parts of biological systems in isolation. This approach is well suited to study diseases and other physiological states that are caused by modifications to a single compound. However, many diseases are caused by simultaneous disruptions in the function of multiple compounds, and the traditional approach to medical research is ill suited to study such diseases.

The development of DNA sequencing techniques brought on a new paradigm in biological research. These techniques (eventually) enabled the full set of genes, the *genome*, of an organism to be obtained in a single experiment. Related techniques that sequence RNA were developed simultaneously, and similarly they enable the full set of transcripts of an organism, its *transcriptome*, to be studied. At any given time, an organism contains a set of proteins in various quantities. These proteins, and their corresponding quantities, constitute the organism's *proteome*. The scientific field that aims to study organisms' whole set of genes is termed genomics. Similarly, the fields that study their complete set of transcripts and proteins are called transcriptomics and proteomics, respectively. The proteome is different from the genome and transcriptome in one important aspect: cells contain only a subset of the proteome and this subset varies between body sites, whereas nearly every cell in an organism contains the same genome. Moroever, during DNA sequencing, specific genes can be amplified to increase

their abundance, which makes them easier to measure. Proteins, however, can not be experimentally amplified in the same manner, which further complicates proteomic analysis.

The genome of an organism is mostly static. Although mutations to the DNA occur frequently, most of them are inconsequential due to countermeasures from the organism. The transcriptome is more dynamic: one gene can often be transcribed to multiple RNA sequences. An RNA transcript can produce multiple forms of a protein, called protein isoforms. Beyond this, modifications can be made to proteins during or after synthesis that change their ultimate function. Such modifications are referred to as Post-Translational Modifications (PTMs), and the most common type of PTM is phosphorylation. Other common types are acetylation, glycosylation, hydroxylation, and methylation.[3] Altogether, this makes the proteome of an organism highly diverse compared to its genome and transcriptome.

Biomarker Discovery

Genes, transcripts, proteins, and many other measurable biological compounds can all serve as potential biomarkers. A biomarker carries information regarding the physiological state of an organism and can be used to diagnose or grade disease; a mutated gene can indicate a particular cancer subtype, and the presence of an antibody in the blood can indicate a viral infection.[6;7] Much research effort is spent on finding biomarkers that can be used to detect various types of cancer at an early stage when curative treatment is still possible.

Generally, there are two types of biomarker studies: those that are hypothesis driven and those that are hypothesis free. In a hypothesis-driven study, researchers may suspect that a specific compound, a candidate biomarker, plays an important role in a particular disease, so they recruit a number of individuals with the disease, collect samples from the patients, and measure the expression of the compound in the samples. They also collect samples from healthy persons, perform the same measurement, and compare the expression of the compound between the diseased and the healthy samples. If the compound is systematically up- or down-regulated in the diseased samples compared to the healthy ones, it can be used to diagnose or grade the disease. In a hypothesis-free study, researchers may instead try to quantify every measurable compound, or a large fraction of them, in each sample (Figure 2). DNA sequencing, RNA sequencing, and liquid chromatography-mass spectrometry (LC-MS) are some techniques that enable such analyses. Differential Expression (DE) analysis can then be performed for each individual compound, which can lead to the simultaneous discovery of multiple novel biomarkers. Hypothesis-free and hypothesis-driven

approaches can also be used in conjunction: measuring a large number of compounds across a primary set of samples might yield a list of biomarker candidates that can subsequently be either validated or rejected by analyzing a secondary set of samples. This is commonly done in LC-MS studies by probing for biomarker candidates in one cohort with DDA or DIA and then validating the candidates, or rejecting them, in another cohort with high-accuracy techniques such as Targeted MS.[8;9] The validation step is crucial to ensure that the biomarkers found in the exploratory study are actually disease related and not the result of experimental or measurement errors.[10;11]

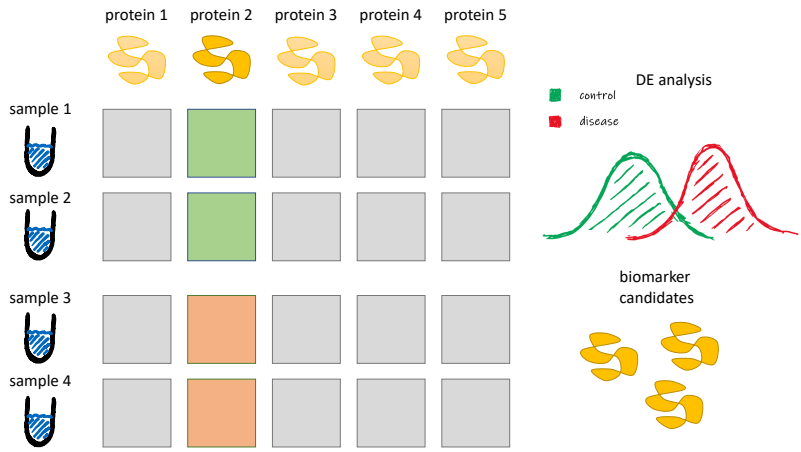


Figure 2: Exploratory studies often find biomarker candidate by investigating the expression of a large set of compounds in different sample groups. The candidates can then be validated, or rejected, in subsequent studies.

Mass Spectrometry

Proteins play an integral part in a vast number of functions in biological systems, and they often operate, and inter-operate, in highly complex ways. During this thesis project, I have focused on a technique that is commonly used to measure the proteome, namely mass spectrometry, and its utility in biological and medical research. The field that studies the proteome is called proteomics. MS-based proteomics relies heavily on the availability of genome sequence data and is therefore tightly coupled to genomics and transcriptomics. A mass spectrometer is an instrument whose ability to separate ionized molecules based

on their mass-to-charge ratio (m/z) makes it an invaluable tool for the analysis of complex biological samples.[12] To identify and quantify compounds from the data generated with mass spectrometry, substantial data processing is necessary. There are numerous ways to utilize mass spectrometry in biological and medical research and during my thesis I have dealt with two of the most common ones: liquid chromatography coupled to mass spectrometry and mass spectrometry imaging (MSI). The two techniques are complementary in many ways and can be used in conjunction to gain a deeper insight into the biology of a sample. LC-MS is a highly sensitive analytical technique that can resolve and quantify thousands of compounds in complex biological samples. MSI is not quite as sensitive as LC-MS but provides spatial information for each resolved compound. Although there are some fundamental differences between LC-MS and MSI, there are many shared aspects in how their data is processed.

There are three major components in a mass spectrometer: an ion source that ionizes molecules, a mass analyzer that separates molecule ions by their m/z, and a detector that counts the abundance of the ions. Compounds in both solid, liquid, and gas phases can be analyzed with mass spectrometry. The ionization technique depends on the phase of the compound; electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are the two primary techniques for analyzing liquid and solid biological samples.[13] There are multiple types of mass analyzers; time-of-flight (TOF) analyzers have a high acquisition speed but low resolution and precision, whereas Fourier transform ion cyclotron resonance (FT-ICR) and Orbitrap analyzers achieve excellent resolution and precision but are typically more expensive and have lower acquisition speeds.[14;15]

The mass spectrometer is sometimes coupled to a high-performance liquid chromatography (LC) system. The LC system physically separates molecules based on their hydrophobicity. The combined LC-MS system thus has the crucial property of separating molecules both by their hydrophobicity and by their molecular weight. The two-dimensional separation is needed since biological samples can contain more than 100,000 different compounds and many of them have similar or identical masses. During analysis with LC-MS, molecules continuously travel through the chromatographic column toward the mass spectrometer in which they are ionized, separated, and quantified. The travelling speed of a molecule depends on its hydrophobicity, and the time it takes to travel through the full length of the column is called its retention time (rt). The output of an LC-MS experiment is a data set consisting of a large number of mass spectra collected throughout the experiment. Figure 3 shows the distribution of intensities over the m/z and rt dimensions from an LC-MS data set. Gas Chromatography (GC) is an alternative to LC, and it can also be coupled to mass spectrometry. GC-MS is mainly used in metabolomic studies, and although I

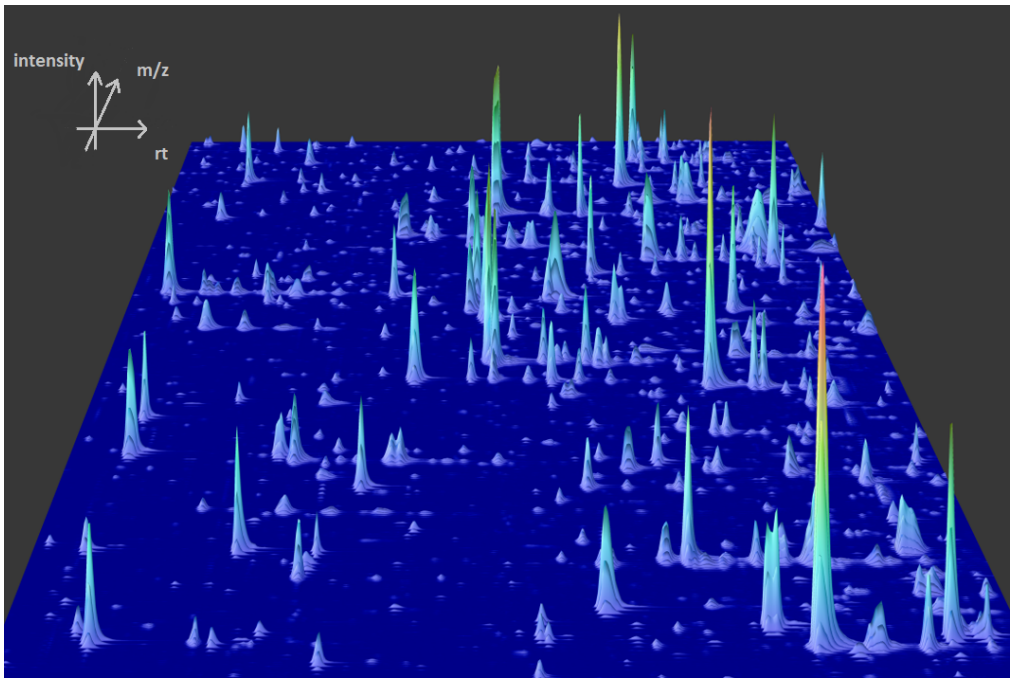


Figure 3: Compounds are separated in two dimensions, retention time and mass, with LC-MS. Isotopic envelopes appear as peak clusters on the retention time - m/z - intensity surface.

have not dealt with it during my thesis project, it is similar to LC-MS in many aspects.[16;17]

The LC-MS system provides excellent separation of compounds due to its two-dimensional nature, but it can generally not be used to uniquely identify those compounds unless an extra step is added. The reason for this is that many compounds have identical masses, and the retention time of a particular molecule can vary greatly between experiments and is therefore difficult to utilize for identification. To be able to uniquely identify a peak on the $m/z-rt-intensity$ surface, a second step is performed in the mass spectrometer. In this step, the mass spectrometer isolates molecule ions whose m/z is close to that of the peak of interest, and then it funnels the isolated molecule ions through a cell where collisions with high-energy particles cause them to break into fragment ions. The fragment ions are then sent to a secondary mass analyzer that collects another mass spectrum, a *fragment spectrum*. This fragmentation method is the most common one and is called Collision-Induced Dissociation (CID or CAD).[18] The fragment spectrum together with the mass of the intact molecule is often sufficient information for a unique identification. Multiple fragment spectra are

typically collected for various isolation windows across the m/z range. The first spectrum, that of the intact molecules, is called the MS1 spectrum and the fragment spectra are called the MS2 spectra. The process of collecting both MS1 and MS2 spectra is the standard approach for molecule identification with LC-MS and is sometimes called LC-MS/MS (or tandem MS) to explicitly state that mass spectra are collected in multiple stages. Figure 4 shows the conceptual structure of a peptide ion and an example MS2 spectrum and their b- and y-ions. The fragmentation techniques used in MS primarily result in y- and b-ions; however, a- and b-ions also occur to some extent.

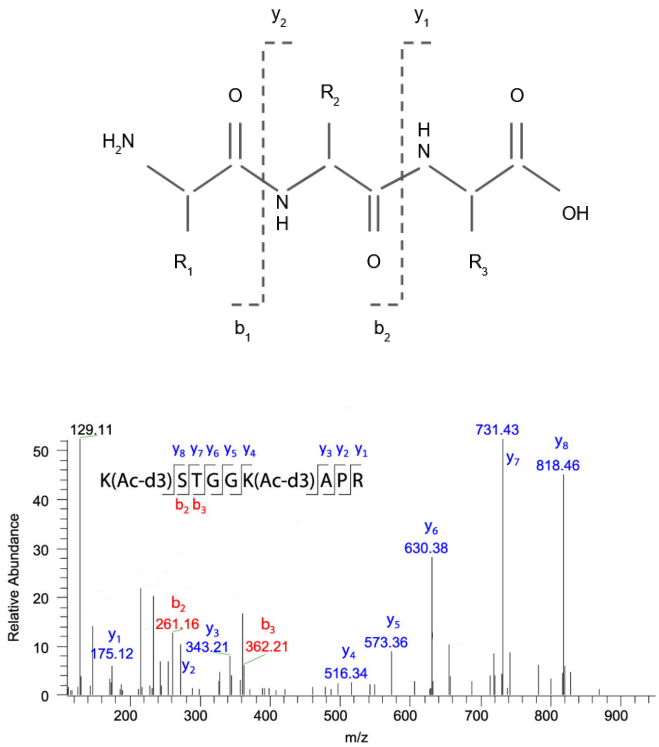


Figure 4: Peptide identification with fragment (MS2) spectra and data base matching. Top: an example peptide with three amino acid residues (R_1, R_2, and R_3). Bottom: an example MS2 spectrum of the peptide KSTGGKAPR.

Many tissue types can be analyzed with LC-MS. Each tissue type has some notable advantages and disadvantages regarding its informative value. Blood has the considerable advantages of being is easily collected and homogeneous throughout the body and improving the sensitivity of LC-MS blood analysis

is a prioritized and ongoing task.[19] Blood tissue is, however, more difficult to analyze for a number of reasons. The most important one is a technical limitation: mass spectrometers have limited dynamic range and the distribution of blood proteins is skewed toward a small number of high abundance proteins. Therefore, most blood proteins are invisible to mass spectrometer. This limitation can be overcome to some extent by depleting the blood of the high abundance proteins prior to LC-MS analysis. The compounds of interest can also be completely absent in the blood; for example, malignant cells are often localized to a single body site at the early stages of cancer and are thereby not measurable in the blood, irrespective of the analysis technique.

MSI is primarily used to visualize the spatial distribution of molecules in a tissue sample. During an MSI experiment, mass spectra are collected from different locations across the tissue section. This results in a data set containing at least one mass spectrum from each tissue location. An image of a molecule ion can be generated by isolating the peak corresponding to its m/z across all the spectra and mapping the resulting intensities to their locations on the tissue section (Figure 5). MSI is more commonly used in metabolomics than in proteomics. This is because larger molecules, such as proteins, are difficult to measure with MSI, and because comprehensive digestion of tissue molecules can not be performed without altering their spatial distributions. In matrix assisted desorption/ionization (MALDI) MSI, a matrix solution is sprayed across the tissue section. Molecules are ionized by firing a laser at the matrix-coated tissue section, after which they can be separated by the mass analyzer[13] There are other, less common, ionization methods in MSI such as secondary ion MS (SIMS), desorption electrospray ionization (DESI), and laser desorption/ionization (LDI).[20;21;22]

Like in LC-MS, fragment (MS2) spectra can be collected with MSI, but only a few from each location due to the limited amount of material at each spot. For the same reason, the spatial resolution is also limited, typically to a raster size between 30-100 μm, but more recent instrument setups have achieved raster sizes below 5 μm.[23] Fragment spectra are usually used to confirm the presence of a known substance rather than identifying an unknown one. A key difference between MSI and LC-MS is the lack of the rt dimension in MSI. This puts extra demands on the resolving power in the m/z dimension, and in **Paper III** we addressed these demands. It is important to note that MSI is typically not used to identify unknown compounds and that peaks can typically not be uniquely identified. Peaks can, however, still be annotated in an FDR-controlled manner.[24]

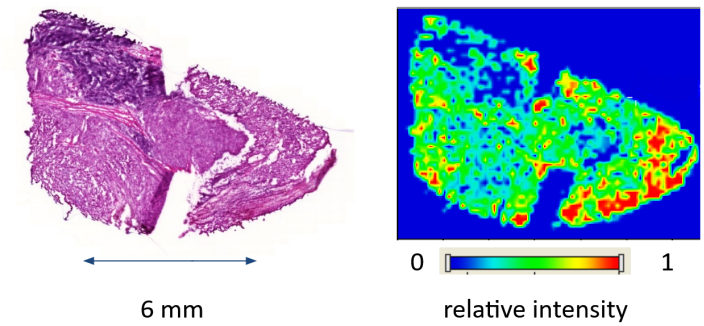


Figure 5: Left: Hematoxylin and Eosin (H&E) image of a tissue section. Right: ion image of the same tissue section. MSI can be used to record the spatial distribution of hundreds or thousands of molecules in a single experiment.

Aims and Contributions of This Thesis

In the study resulting in **Paper I**, we used LC-MS to characterize the proteomes of a set of tumor samples from a melanoma cohort. We then searched for candidate biomarkers related to patient survival. I carried out most of the statistical analysis of the data, co-wrote the manuscript, and participated in the interpretation of the results. In the project resulting in **Paper II**, we developed and evaluated a peak detection method for MSI data sets. I conceived the project, co-designed the experiments, developed the method, and wrote the manuscript with input from the co-authors. In the project resulting in **Paper III** we developed a method for mass alignment in MSI. I participated in the conception of the project, developed the method, analyzed the results, and wrote the manuscript with input from the co-authors. In **Paper IV**, we expanded on the work from **Paper I** by performing a more rigorous characterization of the samples with state-of-the-art LC-MS techniques and instruments, automatic and accurate histopathological assessment, and multi-omic data integration. We were unable to validate the biomarker candidates from **Paper I**, but were able to discover novel gene, protein, and phospho-proteomic biomarker candidates and investigate the predictive power of the different -omic data sets, both in terms of overall survival and survival after metastasis. I contributed to the work behind **Paper I** by performing a part of the survival analysis.

Chapter 2: Data Processing in LC-MS

With some of the fundamental concepts of MS-based proteomics established, we can start discussing some of the key challenges in the processing and analysis of MS data and how I have addressed them. In the following sections, I will use the term *-omics* when referring to the large-scale study of molecules of any type, e.g., genomics, transcriptomics, proteomics, or metabolomics. Moreover, LC-MS/MS can be run in different modes and I will use the abbreviated LC-MS to collectively refer to any or all of them.

For a long time, bottom-up proteomics has been the most popular approach to large-scale protein identification and quantification. Bottom-up proteomics indirectly measures proteins in biological samples by identifying and measuring peptides (cleaved proteins) and then inferring protein identities from the peptide measurements. A bottom-up MS experiment starts by preparing the sample for analysis with LC-MS, and the main steps of the sample preparation are extraction, denaturation, and digestion. Proteins are extracted with various lysis buffers, which destruct cells membrane and liberate single protein molecules. The proteins are then denatured by adding chaotropic reagents, such as urea, to collapse their 3-D structures. Specifically, the proteins' disulfide bonds are broken through the process of reduction/alkylation, and this causes them to lose their 3-D structure. Finally, digestion is performed by adding an enzyme, a *protease*, that cleaves the proteins into peptides to the sample mixture. Trypsin is the most used enzyme since it cleaves the proteins into peptides that are likely to have desirable properties such as high ionization probability. Tryptic peptides have charged basic amino acids, such as lysine and arginine, at the C peptide terminal, and this give them good ionization properties. The main benefit to analyzing peptides instead of proteins with LC-MS is that peptides are more uniform in size than proteins, which facilitates separation with LC. Proteins can also be measured in a top-down manner with MS. However, I have focused exclusively on bottom-up proteomics during my thesis project.

The complexity and size of an LC-MS data set demand sophisticated soft-

ware that can process the spectra automatically. After the sample has been analyzed by the LC-MS system, the peptides are identified and quantified by matching the experimental spectra against theoretical ones from a sequence database.[25;26;27;28] Finally, the proteins are inferred from the identified peptides. An obvious drawback of bottom-up proteomics is that it measures peptides instead of proteins. To obtain protein identities and quantities, the identified peptides belonging to the same protein must be aggregated. However, a peptide might be present in multiple proteins, which leads to the protein inference problem. This problem has no trivial solution.[29;30;31]

Flavors of LC-MS

LC-MS is a versatile technique that can be applied in multiple ways to measure peptides or proteins in biological samples. It is important to highlight that none of the modes of LC-MS achieves perfect identification or quantification accuracy; instead, each mode has unique strengths and weaknesses that make it suitable for certain types of experiments. The modes differ in sensitivity and specificity, quantitative accuracy, and reproducibility, and, naturally, the mode that best servers the objective of the experiment should be used.

Shotgun MS, or discovery MS, is a widely used mode of LC-MS whose primary purpose is to discover or identify proteins and peptides in biological samples. In this mode, the mass spectrometer decides which ions to fragment based on the intensity of the peaks in the MS1 spectra. Specifically, the instrument automatically selects between 10 and 100 of the most intense peaks in the MS1 spectra and collects an MS2 spectrum for each of these peaks. The number of selected peaks is limited by the acquisition speed of the instrument. Each selected peak will correspond to one or multiple intact molecule ions, and these ions are called the *precursor* ions. Since the isolation windows are chosen based on information in the MS1 spectra, Shotgun MS is often called Data-Dependent Acquisition (DDA) mode, and it was the mode we used in the study presented in **Paper I**.[32] The data dependency introduces a bias toward the most abundant peptides, which can lead to a decreased proteome coverage. Due to this bias, low abundance compounds are rarely selected for fragmentation, which leads to an overall low sensitivity for DDA MS. Moreover, the intensity of a compound in the MS1 spectrum is stochastic to some extent and therefore the set of fragmented peptides during a DDA experiment is also stochastic. This further reduces the reproducibility of DDA MS. For example, the overlap between the set of identified peptides in two replicates is typically 60-70 % but can be lower or higher depending on the sample preparation method and the instrument and its configuration. Although tens of thousands of MS1 and MS2

spectra can be generated during a DDA MS experiment, the number of detected peptides is considerably lower than the number of peptides actually present in the sample. Altogether, this means that reproducing the exact same set of identified peptides from the same sample is nearly impossible.

Targeted MS is used when the objective is to detect and accurately quantify a predetermined set of peptides in complex samples.[33] It requires a list of targeted peptides (precursors) and a corresponding fragment library prior to the experiment; the retention time, m/z, and high-intensity fragment ions of each precursor must be known. In a targeted MS experiment, the instrument is run in Selected Reaction Monitoring (SRM) mode (or, equivalently, Multiple Reaction Monitoring (MRM) mode). Unlike Shotgun MS, the instrument does not perform any MS1 scans. The targeted peptides are quantified by comparing their fragment ion intensities to the corresponding intensities of reference peptides. The references have amino acid sequences identical to those of their target counterparts but are isotopically labeled. Targeted MS is data-independent in the sense that the precursor ions are selected prior to the experiment rather than based on the data. Targeted MS have been used to quantify proteins from many different tissue and cell components with high accuracy, including those in mitochondrial pathways[34]

Although Shotgun MS can identify and quantify a large number of compounds in complex samples, it has some noteworthy weaknesses: low reproducibility, low sensitivity, and limited quantitative accuracy. These weaknesses mostly stem from the stochasticity in the selection of precursor ions. Targeted MS is in many ways complementary to Shotgun MS. It is reproducible, has a high quantitative accuracy, and is sensitive enough to detect most low abundance compounds. However, by definition, targeted MS is unable to discover unknown compounds outside the predefined isolation windows. Data-Independent Acquisition (DIA) MS is an alternative approach that attempts to to combine the principles behind DDA and Targeted MS to achieve both accurate identification and quantification.[35;36] In DIA mode, the mass spectrometer isolates and fragments all precursor ions within a relatively large isolation window (25 Da) at the low or high end of the m/z range. The isolation window is then shifted, and another fragment spectrum is collected. This process is repeated until the whole m/z range has been covered. Thereby, the whole m/z range is scanned in cycles, window by window, providing comprehensive fragmentation of all precursor ions in the sample. Since there is no bias in the selection of isolation windows, DIA experiments are significantly more reproducible than DDA experiments. The collection of MS2 spectra from each isolation window throughout the retention time dimension is sometimes called a swath (Figure 4), and *Swath MS* and *DIA Swath* are synonymous to DIA MS.

There is a trade-off between swath width and cycle time. On the one

hand, narrowing the swaths increases the cycle time because more fragment spectra must be collected during each cycle. If the cycle time is too long, some compounds may be missed or incorrectly quantified due to undersampling of the chromatographic peaks. On the other hand, a wide swath width leads to complex MS2 spectra that are the products of multiple concurrently fragmented precursor ions, which complicates identification and quantification.[37] Because of this, DIA demands high-performance instruments that are capable of collecting a large number of fragment spectra while keeping the cycle time low. The requirement of a fast instrument is high compared to Shotgun mode, where only a fixed number of the most intense precursor ions are fragmented, and targeted mode, where only a small number of narrow m/z windows are used at any given retention time.

To summarize, DIA MS generates a more complete and reproducible picture of the sample's molecular composition than shotgun MS or targeted MS, but puts greater demands on both the instrument and processing software. DIA experiments yield massive data sets that contain chromatograms of every fragment ion. This data sets can be mined *in silico* for any compound of interest; in other words, if a new peptide/compound becomes interesting for whatever reason, it can be searched for in the data set again without having to rerun the experiment.

Quantification accuracy can be improved by chemically labeling the peptides prior to analysis with LC-MS. We used Tandom Mass Tag (TMT-11) labeling for the MS experiments in **Paper IV**. A TMT-11 tag can be used to simultaneously analyze 2 to 11 different peptide samples prepared from cells, tissues or biological fluids.[38]

Processing LC-MS Single-Stage Spectra

DDA, DIA, and targeted experiments generate data sets that require different preprocessing strategies. Targeted MS is fundamentally different from DDA and DIA in the sense that the proteins of interest are known beforehand, so no identification is needed. Targeted data sets are therefore fairly simply to process: the extracted ion chromatograms for the predefined m/z windows are typically inspected manually but can be processed automatically, and each compound can be quantified by integrating the area under its chromatographic peak.[39] Processing DDA and DIA data sets is considerably harder, and typically involves two steps: (i) identify compounds by performing database searches with the MS2 spectra, and (ii) link the identification to precursor chromatograms at the MS1 level. DIA data requires more processing than DDA due to the multiplex nature of the MS2 spectra; because of the wide isolation windows, multiple

precursor ions are fragmented in each window. This results in MS2 spectra that contain fragment signals from multiple different compounds that must be separated somehow.[40] There are three general approaches to processing DIA data: those based on generating and querying spectral libraries, those based on deconvolution of fragment ions, and those based on machine learning.[36;41;42]

Searching for Matches in Sequence Databases

Peptide identification is a central part of MS-based proteomics, and much research effort has been spent on developing algorithms that make it as reliable as possible. The archetypal way of identifying peptides from LC-MS data is by matching experimental MS2 spectra against theoretical ones derived from a sequence database.[43] It is important to note that the traditional theoretical spectra are one dimensional: they are simply a list of mass values, one for each possible fragment ion. The mass of a fragment ion can easily be calculated from its amino acid sequence. A typical sequence database contains the amino acid sequences of all the known protein of some specific organism. The selection of the sequence database depends on the origin of the sample. Provided that the enzyme used to digest the protein is known and that it has a specific cleavage site, the peptide sequences can be derived from the protein sequences. Trypsin, for example, cuts amino acid sequences after lysine(K) and argenine (R).

To match an MS2 spectrum, a peptide spectrum match (PSM) score is calculated for all sequences whose intact mass is within the isolation window. Even a relatively narrow window (≈ 0.1 Da) can result in more than 100 candidate sequences, which makes it critical that the PSM score discriminates well between the correct sequence and the incorrect ones. The candidate sequence with the highest PSM score is then a potential match for the MS2 spectrum. There are numerous algorithms for scoring PSMs, but the factor that typically has the largest influence on the score is the number of b- and y-ions that are matched to the fragment spectrum. Figure 6 shows a schematic overview of sequence database matching. Provided a list of scores corresponding to candidate peptides for a specific MS2 spectrum (those with masses within distance D from the precursor ion), one must decide whether the highest score is the result of a true or false match. Fenyö and Beavis[44] use the distribution of the scores of the peptides whose masses fall within the accepted range and survival functions to calculate the probability that the highest scoring Peptide-to-Spectrum Match (PSM) corresponds to a true match.

Spectral libraries contain previously obtained spectra from known peptides, and they provide an alternative to sequence databases. Because the intensity dimension is considered as well, matching experimental spectra against those in

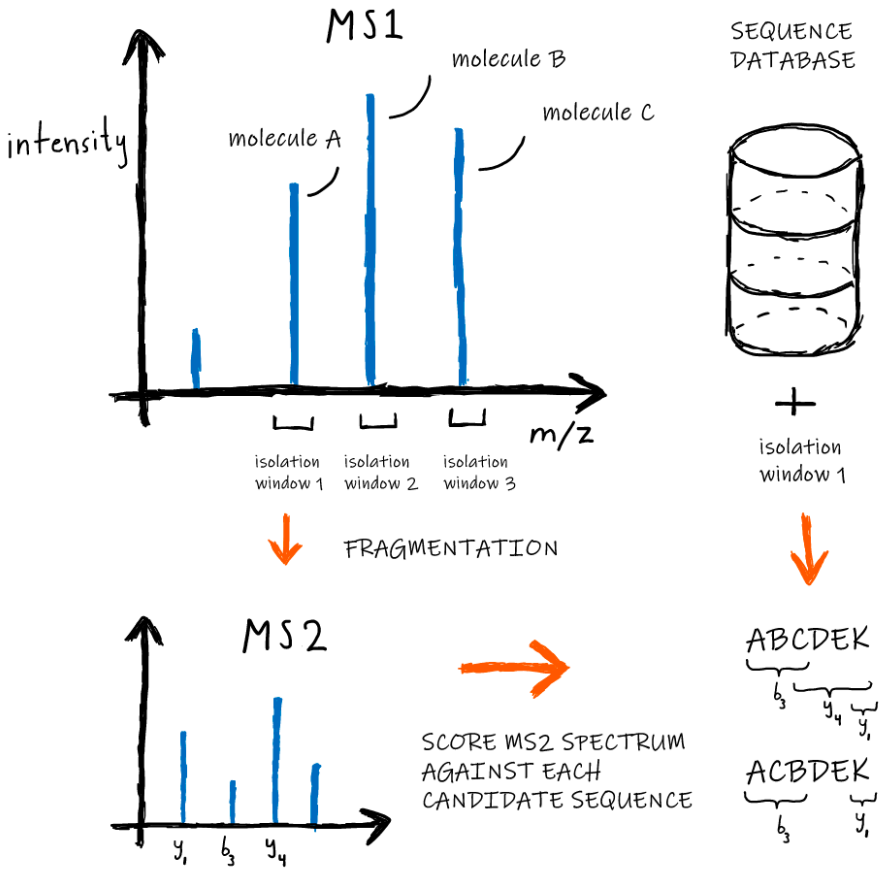


Figure 6: Conceptual description of peptide identification with LC-MS/MS. A fragment (MS2) spectrum is collected for each isolation window and then matched against candidate peptide sequences from the sequence database. In this example, the third peak in the MS2 spectrum matches y_4 in the first sequence (CDEK) but no fragment ion in the second sequence.

a spectral library provides better discrimination between true and false matches than matching experimental spectra against theoretical ones. Spectral libraries are often generated in the same laboratory, since technical variations may render libraries generated in different laboratories incomparable to each other. However, this limitation has been partially overcome lately due to the emergence of standardized sample preprocessing and analysis protocols.

A third option is to match experimental fragment spectra against predicted ones. In this approach, fragment spectra are predicted from peptide sequences. However, accurately predicting fragment spectra is generally difficult since the rules of CID-based fragmentation and ECS-based fragmentation are unknown, and it was deemed infeasible for a long time. Nevertheless, recent approaches based on neural networks have been shown to be able to accurately predict MS2 spectra from peptide sequences.[45;46] Like spectral libraries, this enables a direct comparison between the experimental MS2 spectrum and the predicted one. Deep learning has also recently been used to process DIA chromatograms and spectra.[42]

Even a successful scoring algorithm will sometimes assign the wrong sequence to a spectrum. Since a DDA (or DIA) experiment can produce more than 100,000 MS2 spectra, there are bound to be a considerable number of incorrect PSMs. In search engine terminology, correct and incorrect PSMs are called true and false discoveries, respectively, and the expected fraction of incorrect PSMs among all PSMs is called the false discovery rate (FDR). Search engines typically provide an FDR along with the set of peptide matches. The target-decoy approach is probably the most common one to estimating the FDR for a set of peptide identifications. It is based on searching for peptide matches both in the database containing correct peptide sequences (the target database) and in a database containing incorrect sequences (the decoy database). The simplest way to generate the decoy database is to reverse all sequences in the target database. If the *discoveries* are defined as the PSMs whose scores are above a specific threshold, then the FDR can be computed as the ratio between the number discoveries obtained from matching the MS2 spectra against the decoy database and that obtained from matching them against the target database.[47] The identification accuracy can be further improved by using the approach of Käll et al.[48]. By training a Support Vector Machine (SVM) classifier to separate true from false identifications, they were able to substantially improve identification accuracy. The highest scoring PSMs from the target database are used as examples of true identifications, and those from the decoy database are used as examples of false ones.

Peptides or other compounds identified by matching MS2 spectra against sequence databases can be quantified by linking them to the corresponding 3-dimensional peaks m/z-rt surface. The 3-D peaks are assembled by matching

MS1 peaks across spectra. Quantification using MS1 spectra is advantageous in the sense that it can be more stable than quantification based on MS2 spectra, which is often performed by counting the number of PSMs for each identified compound. Furthermore, 3-D peaks from different molecule isotopes can be connected to each other and thereby provide a robust means of obtaining the charge state of the corresponding molecule [49]

Proteogenomics

Peptide identification via database matching has one major disadvantage: peptides that are present in the sample but not in the database can not be identified. Sequence databases normally only contain the canonical sequences for each protein known to be expressed by a particular species. The canonical sequence is often the most common amino acid sequence for a specific gene but can be defined based on other criteria. However, the actual protein sequences can vary slightly between individual samples due to mutations and other factors. By pairing proteomic and genomic or transcriptomic experiments, the actual sequences can be determined. Thereby, mutated or otherwise modified protein sequences can be identified and quantified with LC-MS. The field that utilizes proteomics in conjunction with genomics and/or transcriptomics is called proteogenomics. A proteogenomic approach is especially appropriate when characterizing malignant tissue since cancer is known to be driven by mutations, and recent proteogenomic studies have brought new insights into cancer biology. [50;51;52;53;54] The process of generating a sequence data base for each individual sample is called generating sample-specific databases. [55;56;57]

At a first glance, adding all possible sequence variants to the database may seem like a viable alternative to performing an extra experiment for each sample. However, this is infeasible because it increases the size of the database exponentially, which leads to an exponential increase in the number of false PSMs. A larger number of false PSMs in turn leads to a lower number of true identifications for a specific FDR threshold. Generally, the most limiting factor when deciding whether to pair LC-MS with DNA or RNA sequencing is the cost in terms of reagents, time, and instrumentation. An alternative to generating paired proteomic and genomic/transcriptomic data sets is to use databases that contain known mutated sequences specific to certain types of cancer. Such data bases can be created from DNA or RNA sequence data collected across multiple studies. [58] This approach requires less resources in terms of instrumentation and reagents compared to generating sample-specific sequence databases but is less sensitive and specific.

Chapter 3: Data Processing and Analysis in MSI

LC-MS is a technique whose primary strength is its ability to identify and quantify a large number of molecules in the same sample. MSI is a related technique that is can be used to investigate the spatial location of molecules within a tissue sample. A key difference between LC-MS and MSI is how well they can distinguish different molecules from one another. In contrast to LC-MS, which separates compounds both in the m/z and retention time dimensions, MSI separates molecules only in the mass dimension. Therefore, MSI is unable to distinguish between molecules with the same mass. Furthermore, different molecules often have the same spatial distribution, which makes it hard to utilize the spatial dimensions to improve identification. In an MSI experiment, mass spectra are typically collected from tens of thousands, or hundreds of thousands, of positions across the tissue section. Figure 7 summarizes images of molecule ions are generated with MSI. In LC-MS, fragment spectra have a crucial role in compound identification, and multiple fragment spectra are typically collected at every time point. In MSI, however, the sampling locations are small (approx. 10-200 square micrometers) and contain only a limited amount of tissue material. Consequently, only a small number of spectra can be collected from the same location, which makes it impossible to collect fragment spectra for more than a small number of precursor peaks. However, to confirm the presence and spatial distribution of a single compound of interest, such as a drug metabolite, a small number of fragment spectra is sufficient.

There are different approaches to analyzing MSI data and the appropriate one depends on the design and objective of the experiment. These approaches can be roughly divided into two groups: those that aim to discover unknown compounds in the data set and those that try to relate the spatial distribution of a known compound to tissue structures or to the spatial distributions of other compounds. MSI is commonly used to investigate the spatial distribution of drugs and their metabolites. Figure 8 summarizes MSI data analysis. Features in MSI data sets are typically peaks or isotope clusters that are present in a

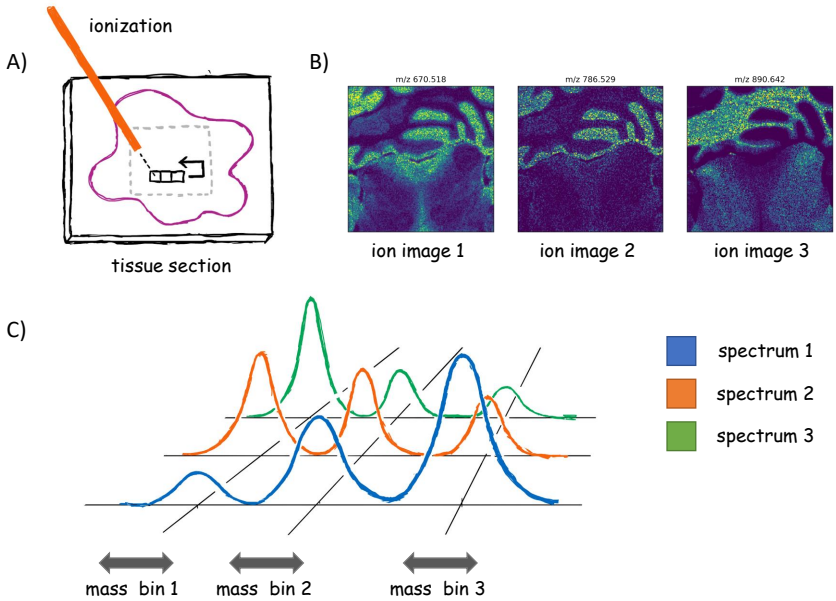


Figure 7: Conceptual description of MSI. (A): mass spectra are collected from different locations across the tissue section. (B): example ion images of three different compounds from a MetaSpace data set. Ion images visualize the spatial distribution of ions and (C) are generated by isolating peaks across the data set mass spectra.

sufficiently large fraction of the mass spectra. A typical data analysis workflow begins by detecting common peaks in the data set that represent tissue molecules. Peaks that are co-localized with that from the drug are then of potential interest and may be identified with subsequent experiments. Co-localized peaks are typically found by computing the correlation coefficient between the target peak (e.g., the drug peak) and all data set peaks. Peaks that are co-localized with specific tissue structures can be searched for in a similar manner.

Before extracting peaks from an MSI data set, the mass spectra are typically processed in a series of steps. The steps include baseline correction, smoothing, mass alignment, and peak picking. Mass spectra from TOF instruments are noisy and generally require substantial preprocessing, whereas spectra from high performance FT instruments, e.g., those from instruments with Orbitrap or FT-ICR analyzers, are much cleaner and require less processing. Baseline correction is performed to remove the baseline signal from mass spectra generated by TOF instrument and mass alignment is performed to reduce shifts in the mass dimension between different spectra. Peak picking, now often called centroiding, is performed to find the location and height of peaks in the mass spectra. A fully processed mass spectrum, a centroid spectrum, is represented by a set of m/z-intensity pairs. Although previously a popular research topic in the MSI field, many processing steps are now performed by instrument hardware and/or vendor software, and the primary focus of data processing is instead on developing methods for peak annotation and/or identification.[59]

Peak Detection

Finding a common set of molecule peaks across the data set spectra is a critical step when processing MSI data sets. This step is sometimes called peak picking in the literature, but I will use the term peak detection here since peak picking also often refers to extracting peaks from individual spectra. After the molecule peaks have been found, ion images are generated by extracting the intensities around the m/z locations of the peaks from all spectra. The molecule peaks therefore correspond to a set of mass channels or *mass bins*. Ideally, a mass bin should capture the peak of a compound in every spectrum where it is present without capturing peaks from any other compound.[60] Carefully selecting the locations and widths of the mass bins is thus essential to MSI data processing, and in **Paper II** we proposed a novel method for sensitive and specific MSI peak detection. Figure 9 highlights how the placement of the mass bin can lead to fragmented or mixed ion images in peak-crowded m/z regions.[61] It is important to note that a molecule peak does not have to be present in all

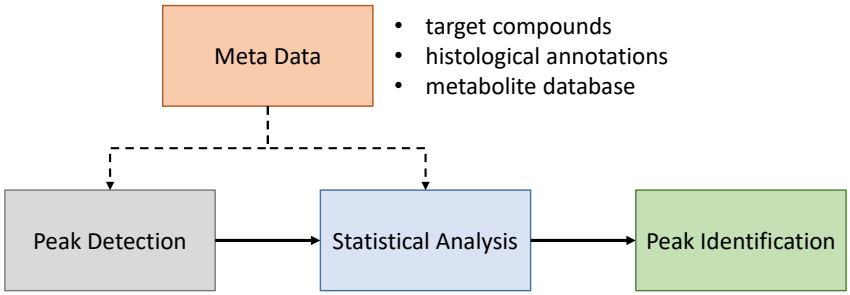


Figure 8: Summary of MSI data analysis. Peaks are detected across the data set spectra. The meta data may include histological annotations (such as tissue structures or cell types) and/or the masses of predefined target compounds (e.g, drug compounds). Statistical analysis includes searching for peaks that are spatially correlated to specific tissue regions or target compounds. After MSI analysis, peaks of interest may be identified with LC-MS.

spectra, or even in most of them, since the molecule may be localized to a small area of the tissue.

A common way to set the mass bins automatically is to compute an average data set spectrum, a mean spectrum, and place the mass bins at the m/z locations of its peaks. Averaging multiple spectra has the desired effect of attenuating noise but also the undesired effect of attenuating faint compound signals. This behavior is reflected in the mean spectrum approach, which often leads to concise lists of high-quality ion images but tends to miss faint signals, especially those that are localized to small regions of the tissue.

Data set peaks and ion images can also be obtained in a more hypothesis-free manner by *slicing* the mass range into uniform mass bins.[62;63;64] In the slicing approach, ion images are generated by extracting the maximum intensity value for each spectrum and mass bin. The slicing approach has no bias toward high-intensity peaks/compounds and can therefore be more sensitive than the mean spectrum approach. However, many of the bins will be placed in non-informative regions of the mass range, i.e., regions that contain no compound peaks or other peaks of interest. This can make slicing especially unsuitable for HRMS (high-resolution mass spectrometry) since an impractically large number of mass bins must be used to match the resolution. The peak width at 400 m/z of a modern FT instrument can be below 0.5 ppm; to match that resolution with the slicing

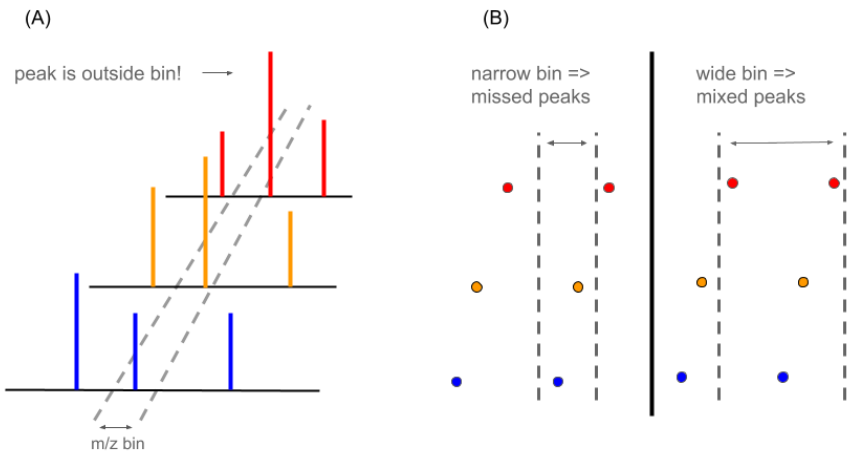


Figure 9: Sensitivity-specificity trade off. If the mass bin is narrow (A, B), peaks may be missed in some spectra; but, if it is too wide, different compounds may be mixed in the same mass bin (B). The resolving power and mass precision of the instrument are the factors that determine the severity of this problem.

approach, hundreds of thousands of mass bins must be used. Peak splitting is another disadvantage; since mass bins are placed without regard for the m/z locations of the peaks, some peaks may overlap multiple bins simultaneously (at most two if large bins are used). Thus, some compounds result in duplicated ion images that may be fragmented or mixed.

Annotating Features

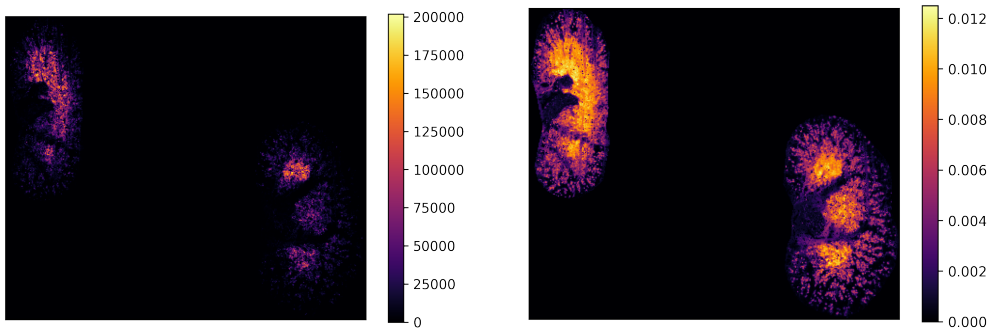
An MSI experiment often results in a list of peaks whose spatial distribution is related to biologically relevant tissue structures or to the spatial distribution of a compound of interest. The compounds that correspond to these peaks can generally not be identified from the MSI spectra, and they must therefore be identified with another technique. Recently, however, Palmer et al. [24] proposed a method that enables FDR-controlled annotation of metabolites from MSI spectra. Like some algorithms for LC-MS data processing, they identify features as isotopic envelopes at the MS1 level. However, since it is impossible to collect MS2 spectra for a large number of MS1 peaks in MSI, they instead base their peak annotation on knowledge about which adducts are likely and unlikely to be attached to the molecule ions. They define a metabolite-signal match (MSM) score:

$$MSM = p_{chaos} \cdot p_{spatial} \cdot p_{spectral}. \tag{1}$$

For a given compound, the subscore $p_{spatial}$ accounts for the (average) spatial similarity between its isotope peaks, the spectral similarity score, $p_{spectral}$, reflects the similarity between its experimental isotope pattern and the expected one, and the measure of spatial structure, p_{chaos}, reflects the level of structure in the ion image of its monoisotopic peak.

Provided a database of known metabolite molecular formulae, or sum formulae, for a particular species, the MSM score is computed for every combination of sum formula and plausible adduct. The set of MSM scores for these combinations is analogous to the set of PSM scores from the target database in FDR-controlled peptide identification with LC-MS. The decoy distribution is obtained by computing MSM scores for the same sum formulae but with implausible adducts instead of plausible ones. The decoy distribution can then be used to set a threshold on the MSM score so that a desired FDR is obtained. For positive mode MALDI MS, H^+, Na^+, and K^+ are likely adducts. Since many metabolites have identical sum formulae, this approach does not generally yield unique identifications for annotated peaks. Instead, it provides a set of possible molecules that share the same sum formula for each annotated peak.

It should be noted that FDR-controlled identification/annotation with MSI is far less sensitive than FDR-controlled peptide/protein identification with LC-



(a) Before normalization.

(b) After normalization.

Figure 10: Ion image of the lipid PI (40:7) before and after TIC normalization.

MS. The number of annotated features ranges between 2 and 200 for most data sets uploaded to MetaSpace (https://metaspace2020.eu), whereas more than 10,000 peptides are routinely identified in an FDR-controlled manner with LC-MS.[65] For the purpose of compound identification, the gain of spatial information in MSI does not make up for its limited fragmentation capability and its lack of the retention-time dimension.

Normalization and Quantification

Label-free quantification with MSI remains an issue for multiple reasons. Firstly, the tissue topography may affect overall ionization, and this can lead to large variations in the total ion count (TIC) throughout the measured m/z range between pixels/spectra. Secondly, a single peak intensity of a compound is typically not sufficiently stable to be used as the quantitative metric of a compound. Finally, the ionization yield can differ between molecules, which complicates relative quantification.[66] Figure 10 shows the effect of TIC normalization on the ion image of the lipid PI (40:7) in the mouse kidney data set originally published by Noh et al.[67]. TIC normalization is performed by dividing each intensity value in a spectrum by its total ion count. There are many other normalization methods, such as median or root mean square (RMS) normalization, yet no consensus whether one method should be preferred over the others.

Chapter 4: Few Samples with Many Variables

One of the primary objectives of exploratory -omic studies is to find molecular signatures that can be related to clinical outcomes. A particular expression of a set of genes or proteins might indicate that an individual is expected to respond well to some treatment or be at a high risk of recurring disease. A well-known example of such a signature is the MammaPrint test, which predicts the risk of metastasis for women with early-stage breast cancer.[68] The MammaPrint test is based on a 70-gene signature that was initially derived in 2002 and then validated later the same year.[68;69] Another example is the PAM50 gene signature, which is known to accurately reflect the subtypes of breast cancer and is routinely used as a prognostic tool.[70;71] However, deriving such signatures from complex LC-MS or gene sequencing data is no trivial task. This is partially due to the uncertainty in the data generated with LC-MS and other -omic techniques, but mostly due to the difficulty in obtaining the actual biological material. The reason for the latter is somewhat obvious: the number of individuals suffering from a particular disease is limited, and, therefore, as are the number of available tissue samples. It can be even harder to obtain control samples from healthy individuals (or from healthy tissue) because doing so may cause unnecessary harm. Furthermore, analysis with high-throughput techniques yields measurements of a large number of molecules from each sample. The combination of this and the scarcity of the samples results in data sets that are composed of a small number of samples with many variables.

The samples can be thought of as existing in a high-dimensional space with the same number of dimensions as the number of measured molecules. The position of a sample in this space is then defined by its expression of the molecules. Formally, a data set is high dimensional when $p \gg N$, where N is the number of samples and p the number of variables. In data sets generated with LC-MS and other high-throughput technologies, the variables frequently outnumber the samples by a ratio of 10-to-1 or larger. As an example, this ratio was approximately 80-to-1 in the TMT data set we generated for the study

presented in **Paper IV** (we quantified 9800 proteins in 120 samples). From a statistical perspective, this scenario is highly unfavorable, and many statistical methods require the inverse scenario: that there are more samples than there are features, or at least as many.[72;73] To reiterate, increasing the number of samples to the extent that they match the number of features is rarely an option, but, fortunately, there are strategies for dealing with high-dimensional data and I will introduce and discuss some of them in this section.

Dealing with High-Dimensional Data

Being aware of the problems that come with high-dimensional data, and dealing with them appropriately, is crucial when carrying out -omic studies, irrespective of the research question. One of the most common objectives in an exploratory study is to find compounds whose expression levels differ between two or more sample groups. This is frequently done by performing a series of hypothesis tests, one for each compound. However, the risk of observing spurious differences increases with the number of tests, and this risk should be estimated somehow. Another objective might be to predict the outcome of new patients from their molecular expressions. But most predictive models generalize poorly when the number of variables is too large compared to the number of samples, and reducing the variable space is therefore often necessary. This can be done "outside" the model or be directly incorporated into the model. Dimension reduction can be performed through some feature selection procedure or by projecting the data onto a lower-dimensional space. Both feature selection and data projection can be done in either a supervised or unsupervised manner. Latent variable (LV) projection is one of the most popular supervised approaches while principal Component Analysis (PCA) is probably the most common unsupervised approach. By design, the first principal component (PC) explains most the of the variance in the data and the second one explains the second most variance and so on. For -omic data, the first PCs tend to capture large sources of variation such as batch effects, differences between body sites, or differences in the cellular composition of the tissue while later PCs mostly correspond to noise. This is undesirable since subtle, but often clinically relevant, features of the data are missed. Independent component analysis (ICA) is an alternative to PCA that does not order the different sources of variation based on their magnitudes.[74] The statistical analysis in **Paper IV** relies heavily on ICA, and we were able to relate multiple independent components of the multi-omic data to clinical data.

A common type of unsupervised analysis in -omic studies is hierarchical clustering during which samples are grouped in a stepwise manner based on

their similarity to each other. The Euclidean distance and Pearson correlation are two common measures that can be used to assess the pairwise similarity between two samples. Hierarchical clustering can be performed both with and without dimension reduction. If it is performed without dimension reduction, the clustering may be driven by unrelated biological processes. In fact, it is reasonable that the most dominant processes in the tissue are unrelated to disease. The processes of interest may be faint in comparison, and, therefore, supervised approaches are often more appropriate. Dimension reduction with PCA also carries the same risk: the most dominant patterns in the data can be related processes that are medically uninteresting.

Statistical Hypothesis Testing

Hypothesis testing is probably one of the most, if not the the most, common application of statistics. Statistical hypothesis testing relies on stating a hypothesis and then testing it using observed data. For example, we might hypothesize that there is no association between sex and height, but when we look at the distributions of height among females and males, we conclude that it would be highly unlikely to observe such data if our hypothesis was true. Consequently, we reject our hypothesis, and, by extension, indirectly accept the alternative to our hypothesis: that there is a difference in height between females and males. We can take a similar approach when evaluating whether there is any association between the expression of a particular compound and a disease. In this case our default hypothesis, or *null hypothesis*, is that there is no link between the disease and the expression of the compound. To test our hypothesis, we start by measuring the compound in sick and healthy individuals. We then record the difference in expression between the two groups, estimate the probability of the observed difference under the null hypothesis, and reject the null hypothesis if this probability is sufficiently low.

The probability of making an observation at least as extreme as the one observed, on the condition that the null hypothesis is true, is perhaps the most well-known statistical measure: the p-value. The p-value can be defined as

$$p = \Pr(observation|H_0), \tag{2}$$

and the way it is estimated depends on the type of hypothesis test. The process of identifying differentially expressed compounds in -omic data sets is called differential expression analysis, and one of the most common approaches to DE analysis is to perform an independent Student's t-test for each individual compound. The t-test looks at how large the difference in mean expression between the two groups is compared to the standard deviation of the expressions. When

the variance of the expressions is similar in the two groups, the independent t-test is defined as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{1/n_1 + 1/n_2}}, \tag{3}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the mean expression in the first group and second group, respectively, and s_p is an estimation of the pooled standard deviation. Two frequently used alternatives to t-tests are the Linear Models for Microarray Data (limma) and Significance Analysis of Microarray data (SAM) methods, which are specifically designed for omic data.[75;76] The limma approach is based on modeling the expression of a specific compound and sample, y_i, as a linear function of a set of variables $\boldsymbol{x_i} = [1, x_{i1}, ..., x_{ip}]$:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i = \boldsymbol{x_i}\boldsymbol{\beta} + \epsilon_i, \tag{4}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients, one for each variable, and ϵ_i a random error. The sample group is represented by a dummy variable that is either 0 or 1 depending on whether the sample is a member of the group or not. Multiple dummy variables are used when there are more than two groups. Covariates are easily incorporated directly into the model, which is an advantage of using linear models compared to t-tests.

It is important to note that the p-value is not the same as the probability that the null hypothesis is true: $\Pr(H_0|observation)$. Bayes's theorem gives the relationship between the two: let $P(X) = \Pr(observation)$, then

$$P(H_0|X) = \frac{P(X|H_0) \cdot P(H_0)}{P(X)} = \frac{P(X|H_0) \cdot P(H_0)}{P(X|H_0) \cdot P(H_0) + P(H_1|X) \cdot P(H_1)}. \tag{5}$$

Note that $P(H_0)$ and $P(H_1)$ (the prior probabilities of the null and alternative hypotheses) are generally unknown, and when they are, the posterior probability of the null hypothesis is also unknown. Furthermore, it is important to highlight that when the prior probability for H_0 is high, the posterior probability can be high as well, even when the p-value is very low. In other words, a low p-value, on its own, does not necessarily indicate that the null hypothesis is unlikely.

Two types of errors can be made when performing hypotheses tests in this manner. Firstly, the null hypothesis can be falsely rejected: we incorrectly conclude that there is an association between the disease and the compound when there in truth is none. Secondly, we can fail to reject the null hypothesis when there is an actual association. The first type of error is called a false positive, or a type I error, and the second type is called a false negative, or a type II error. Reducing the risk of a type I error is generally not possible without increasing the risk of a type II error, and vice versa (Figure 11). The probability of a type I error, given that the null hypothesis is true, is conventionally denoted

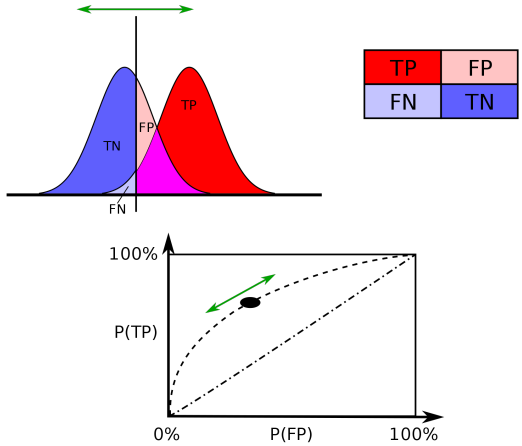


Figure 11: ROC curve for hypothesis testing. The test statistic distribution for positive samples overlaps with that for negative samples, leading to a trade-off between specificity and sensitivity, or type I and type II errors.

by α. Similarly, the probability of a type II error, given that the alternative hypothesis is true, is denoted by β. The statistical power of a test $(1 - \beta)$ is defined as the probability that the test will correctly reject the null hypothesis when the alternative hypothesis is true.

The p-value and the type I error rate, α, are not the same, even though they are closely related. For example, we might want to perform a set of tests and ensure that the type I error rate is at most $\alpha = 0.05$. We compute a p-value for each test and reject the null hypothesis in the tests whose p-value is below α. The p-values from these tests will range between 0 and α. Each individual p-value below the threshold will thus be different from α. To summarize, the p-value is a random variable and by bounding it, we can control α.

This brings us to the issue of multiple testing and why it matters in the context of exploratory -omic studies. Consider the scenario where we have measured a large number of compounds, say 1000, in two groups of individuals. We perform a set of hypothesis tests, one for each compound, and "discover" 20 compounds whose p-value is below our significance threshold 0.01. Then we might be curious about how probable it is that at least one of these discoveries is false: that there is no actual difference in expression between the two groups, and that the unlikely observation is purely by chance. This probability is known as the family-wise error rate (FWER). The Bonferroni method controls the FWER by rejecting the null hypothesis only when $p \leq \alpha/m$.[77] However, this threshold is too strict for -omic data since m is often very large; to obtain a

FWER of 0.01 in our example data set with we would have to use a p-value threshold of $0.01/1000 = 10^{-5}$. Most true discoveries would be filtered out by such a strict threshold. The Bonferroni approach also assumes all compounds are independent, but many compounds are highly correlated. Instead of ensuring that the FWER is below some threshold, we could ensure that we discover as many true associations as possible, while at the same time restricting the number of false discoveries. Indeed, accepting some false discoveries in an exploratory study is reasonable because those discoveries can be detected and filtered out with subsequent experiments or in follow-up studies. Then it is more appropriate to control the false discovery rate than the FWER. Let Q be the fraction of false positives among all positives:

$$Q = \frac{FP}{FP + TP} = \frac{V}{V + S} = \frac{V}{R}, \tag{6}$$

where V and S are the number of false discoveries and true discoveries, respectively. The FDR is then defined as

$$FDR = E[Q]. \tag{7}$$

The approach of Benjamini and Hochberg[78] is among the most popular ones for controlling FDR. Consider once more our example with 1000 compounds and 20 discoveries: how many of these discoveries are expected to be false? We used an α of 0.01, meaning that we expect to incorrectly reject the null hypothesis in 1 out of 100 tests (on the condition that the null hypothesis is correct in all cases). We have performed 1000 tests, so with our p-value threshold we expect to incorrectly reject $1000 \times 0.01 = 10$ null hypotheses. With our significance threshold, we get $R = 20$ (20 discovered compounds) and $V = 10$ resulting in an FDR of at most 0.5. Blindly choosing a FDR threshold makes as little sense as blindly setting the type I error rate, α, to 0.05 or 0.01. There is no correct threshold; the threshold should reflect the goal of the study. If the goal is to discover as many biomarker candidates as possible, a threshold considerably larger than 0.1 can be warranted.

Predictive Modeling

Predictive modeling is another major area in statistics, and it has many applications in biology and medicine. Conventionally, predictive models are divided into two groups: those that make categorical predictions and those that make continuous predictions. Models of the former type are called *classifiers* and those of the latter type are called *regression models*. In the context of medicine and the -omic fields, predictive modeling can be used to diagnose a disease or

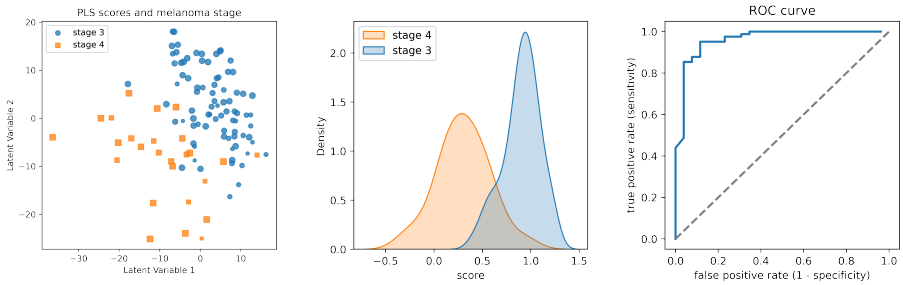


Figure 12: TMT protein expression and disease stage from **Paper IV**. Left: projection of 8572 protein features onto the first two latent variables found with PLS regression. Center: the PLS prediction largely separates the two sample groups. Right: the Receiver Operator Curve (ROC) provides a summary of the trade-off between sensitivity and specificity.

to give a disease prognosis based on some test results. For example, Esteva et al.[79] trained an Artificial Neural Network (ANN) to diagnosed melanoma from (regular camera) images of suspicious moles. Their ANN model was able to automatically diagnose melanoma with at least as high accuracy as expert pathologists. Yuan et al.[80] predicted patient survival for various types of cancer using different types of molecular -omic data. They found that the predictive power of the molecular data was modest in most cases. Nevertheless, their results highlight an important aspect of predictive modeling in the context of medicine: even a modestly powerful predictive model can reveal subtle links between a disease and the expression of proteins, transcripts, or other biomolecules, which may lead to a deeper understanding of the underlying pathological process.

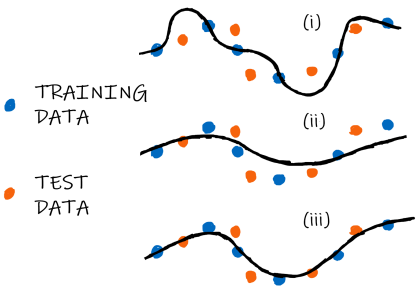
For high-dimensional data with continuous features, Partial Least Square (PLS) regression and Nearest Shrunken Centroid (NSC) are fast and high-performing algorithms.[81;82;83] We used PLS to predict survival in **Paper I**, and Figure 12 shows how PLS can be used to predict disease stage in melanoma from protein expression data. Christin et al.[84] showed that multivariate models can outperform univariate approaches in DE analysis. A multivariate model can be trained on a classification or regression task, e.g., disease stage prediction, and then be used for feature selection. Pure -omic data sets typically only contain continuous features, but when they are combined with clinical data, e.g, the age, sex, or disease stage of the individuals, a predictive model that accepts a mix of continuous and categorical features should be used. Random Forests (RF),

and Support Vector Machine (SVM) algorithms can handle data with a mix of continuous and categorical features and are suitable for purely continuous data as well. PLS was initial developed for high-dimensional chemical data, such as spectra, while NSC was developed for the analysis of high-dimensional gene expression data.

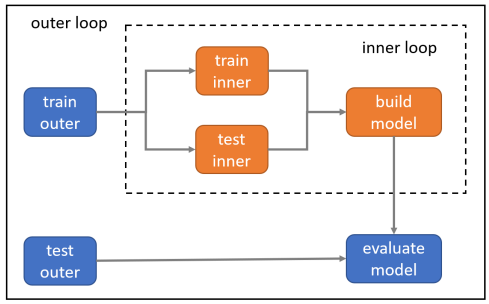
Cross-Validation

The more features we include in a predictive model, the better the model will be able to fit the data. This might lead us to add as many features as possible, but once we try to make predictions on new samples whose outcome is unknown, we will notice that our model performs much worse than expected. Including too many features in the model makes it *overfit* to the data. To get a better estimate of the model's accuracy, a set of *test* samples is typically set aside before fitting the model. The left-out samples can then be used to get a correct estimate of the predictive accuracy of the model. This approach is appropriate when there is an abundance of samples; but, when samples are scarce, it produces unreliable results. On the one hand, if a small number of samples are set aside, say 5 or 10, the randomness in the choice of those samples is high and they may be a poor representation of the whole sample set. On the other hand, if a larger fraction of the samples is set aside, there might too few samples left for training. One approach to circumvent this unfavorable trade-off is *repeated cross-validation*. The objective of repeated CV is to reduce the randomness in the selection of training and test samples by repeatedly generating many splits, and it is one of the most common approaches to estimating the expected prediction accuracy when the sample size is small. Figure 13 illustrates how a model with too many variables can overfit to training data, and how repeated cross-validation is typically performed. As an example of how cross validation has been used in -omic research, the 70 genes defining the MammaPrint signature were initially derived by maximizing the cross-validated prediction accuracy of a survival classifier.

It is important that all steps of the model building is performed using only the training data. For example, if we apply PLS in a cross-validation loop, the direction of the latent variables should be re-computed in each iteration using only the training data. If we compute them once using all samples, the test samples are no longer completely unseen, and our prediction error will be underestimated. In fact, it can be considerably underestimated; in an example from Hastie et al.[85], a k-nearest neighbor (KNN) classifier achieved a 90% cross-validated classification accuracy when the class labels were assigned at random, and the true accuracy should have been around 50%. By selecting the



(a) Model fitting.



(b) Repeated cross-validation.

Figure 13: (a): Example of a model with (i) too many features, (ii) too few features, and (iii) an appropriate number of features. If a model has too many features, it will perform poorly on new data even though it performs well on training data. (b): Nested cross-validation where the inner loop is used to select the optimal parameters. The performance of the model when the optimal parameters are used is estimated in the outer loop.

features using all data, the classifier found spurious relationships between some variables and the response variable, which led to the overestimated accuracy.

Survival Analysis

In survival analysis, the survival function is defined as the probability that a subject will not experience an event within a specific time period. In other words, the subject "survives" as long as it does not experience an event. The subject may be an individual and the event defined as his or her death from a particular disease, but the subject and event can also be defined in many other ways. Conventionally, the survival function of a subject is defined as

$$S(t) = Pr(T > t), \tag{8}$$

where T is the elapsed time, the *survival time*, between the beginning of the observation period and the event. A key property of survival analysis is that it considers the possibility that the survival times of some subjects are only partially observed; if a subject does not experience an event within its observation period, then it is censored because we only know that its survival time was at least as long as the observation period. This partial knowledge is still informative and is utilized in survival modeling. In a medical context, the

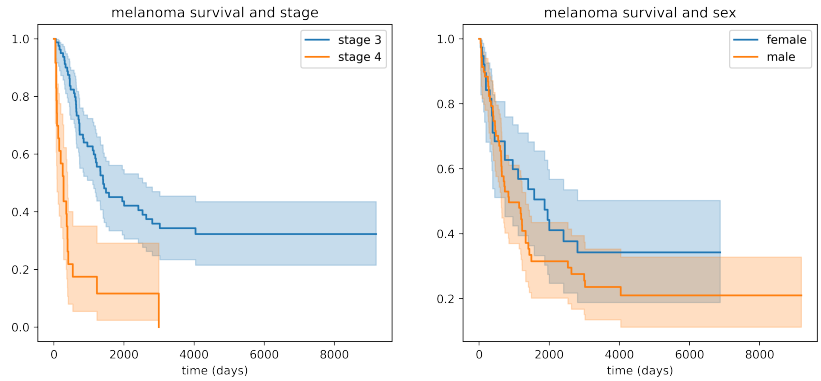


Figure 14: Kaplan-Meier estimates of the survival functions of melanoma patients of **Paper I** and **IV**. Individuals with stage 3 melanoma clearly survive longer than those with stage 4 melanoma, and there also appears to be minor difference in survival between males and females.

observation period of an individual often begins when he or she is diagnosed with a particular disease and ends when he or she dies from the disease.

While there is an abundance of algorithms for classification and regression tasks, there are only a few approaches to survival analysis. Perhaps the most popular one is the Kaplan-Meier (KM) estimator, which is used to estimate the survival function of members of a specific group. Figure 14 shows the KM estimates for individuals in the Lund Melanoma cohort with and without distant metastases. When comparing survival times between different groups the log-rank test is frequently used. The KM estimator and log-rank test are often used together to provide a visualization that is easy to interpret and an accompanying p-value. The simplicity of this approach is treacherous, however, since researchers are tempted to use it to answer all survival-related questions, even those for which it is unsuitable. One example is when investigating the relationship between a quantitative value, such as the expression of a gene or protein, and survival. Because the log-rank test compares survival between groups, the quantitative value must somehow be used to divide the samples into groups. A common way to do this is to split the samples into two groups based on whether their value is below or above some cut point. The median value is frequently used as the cut point. Information is, however, lost when categorizing continuous variables in this manner, and therefore KM analysis is rarely the best choice for survival analysis with continuous variables/features. [86]

A better choice for investigating the effect of continuous variables on survival

is the Cox Proportional Hazards model.[87] The Cox model describes how a subject's instantaneous risk of experiencing an event (the hazard of the subject) depends on a set of variables. The hazard of a subject at a specific time point, t, is defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt | T \geq t)}{dt}. \tag{9}$$

In the Cox model, the hazard is relative to a baseline hazard, $\lambda_0(t)$, and is defined as

$$\lambda(t|x) = \lambda_0(t) \cdot \exp\left(\beta_1 x_1 + ... + \beta_p x_p\right) = \lambda_0(t) \exp(x \cdot \beta). \tag{10}$$

For example, if the Cox model is used to describe a person's risk of dying at a specific time point, the baseline hazard would likely be related to his or her age, and the relative risk would depend on various lifestyle factors. The Cox model is a multivariate model that accepts both continuous and categorical variables. Like in linear regression models, categorical covariates are modeled with dummy variables. The hazard ratio for a particular covariate is $h_j = \exp(\beta_j)$. For continuous covariates, the Cox model provides a higher statistical power than KM analysis with the log-rank test.

Unfortunately, the Cox model is not directly applicable in high-dimensional settings. It requires a sufficiently large event-per-variable (EPV) ratio; a ratio of 10 is often recommended but one between 5-9 can also be sufficient for certain types of analyses.[88] By definition, this requirement can not be met by high-dimensional data sets. Therefore, some strategy for reducing the number of variables must be employed when performing survival analysis with -omic datasets. A straightforward approach is to perform clustering of the samples using the expression data and then look for survival differences between the clusters. We did this as a part of the analysis for **Paper I**, but there is no guarantee that survival related features will drive the clustering and the clusters may thus be unrelated to survival. The same strategy as that used when performing FDR-controlled DE analysis can also be used. Survival analysis with SAM is such an example; the univariate Cox coefficients are computed for each individual compound and are compared to a null distribution of coefficients to obtain those that are differentially expressed. The null distribution is generated by shuffling, or permuting, the samples.[76]) Alternatively, the feature-space of the data can be reduced prior to Cox analysis somehow, for example with PCA or PLS.[89;90]. Survival analysis can be performed with PLS in the following manner:

1. Compute the first m first latent variables with PLS. Use the response variable $y_i = \min(T_i, C_i)$ where T_i is the survival time and C_i the censoring time.

2. Compute the values of the latent variables for each observation in the test set and insert them into the fitted Cox model.

Note that the first step is only reasonable when most of the samples are uncensored. If this is the case, the PLS approach is similar to the Supervised Principal Components (SPCA) approach. While the simulation study of Bair et al.[89] indicated that SPCA performs slightly better than PLS, the computational demands of PLS are much smaller than those of SPCA. Low computational demands can be important since extensive cross-validation is often needed to obtain reliable estimations of the performance of these methods. The scores on the latent variable for new data, u', can be calculated by first scaling the new data with the stored means and standard deviations from previous samples: $X'_s = (X' - m)/s$, and then projecting X_s onto the latent variable: $u' = X'_s w$. Finally, u' can be used in the Cox model that was fitted to u.

Rank products were initially developed to detect differentially expressed genes in noisy data collected from a set of replicated microarray experiments. In each experiment, the genes are ranked based on their fold change between sample groups; the gene with the smallest (largest negative) fold change is given rank 1 and that with the second smallest one is given rank 2 and so on. The rank product, RP, for a specific gene is then defined as the geometric mean of its rank, r, over the set of experiments:

$$RP = (\prod_{i=0}^{k} r_i)^{1/k}. \tag{11}$$

The rank products can also be used with a cross-validation procedure if there is natural way to rank features based on their contribution to the prediction. Each iteration in the cross-validation loop will then correspond to a repetition of the experiment.

The objective of survival analysis in proteomic or other -omic studies is rarely to maximize the predictive power of some statistical model, but rather to determine which compounds, if any, impact survival time. Nevertheless, predictive modeling can be used as a means of detecting those survival-related compounds. Like previously mentioned, the features that contribute the most to the prediction can be extracted from a regression or classification model. As a small example of how PLS can be used to extract survival-related features I a generated a simulated omic data set. This data consisted of 100 samples that each had 5000 continuous predictors. The survival times of the first 50 samples were drawn from an exponential distribution with ratio 0.5 and those of remaining samples were drawn from an exponential distribution with ratio 1.0. All values for the predictors were initially drawn from a standard normal

distribution. I then 'upregulated' features 1-50 in the 50 first samples by redrawing them from a normal distribution with mean of 1.0. Finally, I upregulated features 50-300 in randomly selected samples by redrawing them from a normal distribution with a mean of 2.0. As a result, features 1-50 were related to survival and features 51-300 were differentially expressed between some samples, but in a manner unrelated to survival.

I then fit a univariate Cox model for each feature independently. Around two thirds (32) of features 1-50 were among the features with the 50 lowest p-values. The 18 remaining discoveries were thus false, resulting in an FDR of 0.36. I also performed a cross-validated PLS procedure, and in each iteration ranked the features based on their contribution to the first latent variable. After 100 iterations I looked at the 50 features with the lowest rank products, and 44 out of features 1-50 were among these, resulting in an FDR of 0.12. Figure 15 summarizes the univariate Cox and PLS-Cox survival analyses on the simulated data. Finally, I ran the same analysis on a slightly more difficult data set. The results were similar: univariate cox found 5 of the true positives whereas PLS with rank products found 25 of them, resulting in FDRs of 0.9 and 0.5, respectively. In real applications, the number of survival-related features is unknown, and the FDR must be estimated. Like shown before, this is trivial for an approach based on p-values because the p-value is uniformly distributed under the null hypothesis, which makes the number of false discoveries easy to estimate. For the PLS approach, however, the null distribution of the rank products depends on the data and must be estimated. The authors of the original rank products paper suggested to do this by shuffling the sequence 1 to m in each iteration. But this approach underestimates the FDR considerably when the number of features is large. A alternative approach is to re-compute the rank products using permuted samples. The samples are assigned a random outcome, and they are thereby 'disconnected' from their expression values. The resulting distribution of rank products then represent the null distribution. This is the approach we used in **Paper I**. In this simulated example, however, the permutation-based approach seems to overestimate the FDR considerably.

Because the Cox model deals with partially censored data, it does not directly predict the survival time of subjects. Instead, it predicts their hazard scores; but, since the true hazards of the subjects are unknown, the error of the prediction can not be trivially estimated. However, one measure that can be used to evaluate the predictive power of the Cox-Model is the *Concordance Index* (C-index). The C-index is defined as the fraction of concordant pairs among all valid pairs of subjects. Specifically, a pair is concordant if the subject that experienced the event first has a higher hazard score than the other subject. Conversely, if the subject with the lower hazard score has a shorter (uncensored) survival time than the other subject (whose survival time may be censored), the

pair is disconcordant. Finally, if the survival time of the subject with the lower hazard score is censored *and* shorter than that of the other subject, the pair is considered invalid. The value of the C-index will range between 0 and 1 and a value of 0.5 represents the worst possibly model, i.e., one that makes completely random predictions.

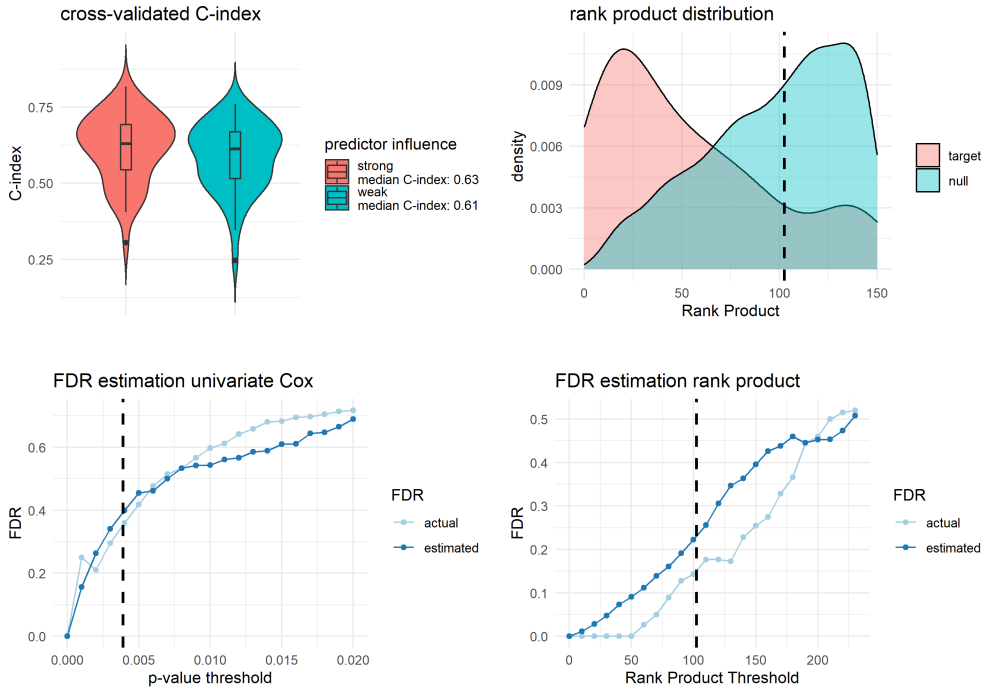


Figure 15: Feature selection via PLS and univariate Cox in a simulated data set with 100 samples and 5000 features. The PLS approach found more survival-related features than the univariate Cox approach. Top left: C-index distributions for two levels of impact for the 50 suvival-related features. Top right: rank product distributions for target and null distributions. Bottom: actual and estimated FDR for the two approaches.

Chapter 5: Summary of Papers

Two of the four works that form my thesis attempt to relate molecular signatures to survival in metastasized melanoma. Both studies are based on the same cohort. The first study (**Paper I**) focuses on peptide and protein expression, but also includes histopathological information. In the second study (**Paper IV**), we characterized the samples more deeply by performing TMT-labelled LC-MS and phospho proteomics. In this section, I will briefly summarize the methodology of these studies and the conclusions we drew from them.

In addition to my contributions to the melanoma studies, I have developed two preprocessing methods for MSI data. The first one (**Paper II)** describes a sensitive peak detection approach. The second one (**Paper III**) describes a general and accurate mass alignment algorithm.

Summary of Paper I

The focus of the study presented in **Paper I** was on the relationship between the protein and peptide expression of tumor tissue and survival in melanoma. Specifically, we characterized tumor tissue from lymph node metastases that had previously been surgically removed from individuals with melanoma. To measure the peptide and protein expression in these tissues, we used DDA MS. The survival time was defined as the duration between the removal (and freezing) of the lymph-node metastasis and the death of the individual. The survival times of individuals whose deaths were unrelated to melanoma were considered censored, as were those who were still alive at the end of the study. The individuals who only had lymph-node metastases at the time of surgery were classified as having stage 3 melanoma and those who also had distant metastases were classified as having stage 4 melanoma.

We used two primary approaches to investigating the relationship between protein expression and survival. Firstly, we clustered the samples in an un-

supervised manner (Hierarchical Clustering with the ConsensusClusterPlus R package[91]) and compared survival between the resulting groups. Secondly, we selected proteins that were strongly connected to survival with a PLS-Cox procedure, and then clustered the samples based on their expression of the selected proteins. Both approaches resulted in clusters with significant survival differences, but as expected, the the clusters obtained with the supervised approach had larger survival differences than those obtained with the unsupervised approach.

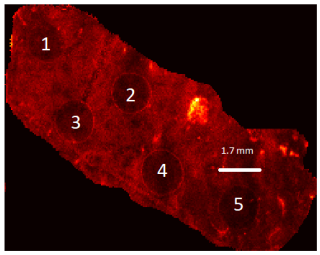
We found the proteins with the strongest link to survival with a repeated cross-validation procedure. In each iteration of the procedure, we generated a training fold by randomly selecting two thirds of the samples. The remaining one third of the samples formed the test set. With the training samples, we computed the first PLS latent variable using survival time as the response variable and protein expression as input. To determine which proteins contributed the most to the latent variable, and by extension to the prediction of the hazard ratios, we computed the inner product between the expression of each protein and the scores on the latent variable. The proteins were ranked by the absolute values of their inner products. We also fit a Cox model to the survival times and scores on the latent variable. Finally, we used the test samples to evaluate the model by projecting their protein expression onto the latent variable and then predicting their hazard ratios using the Cox model. In other words, the direction of the latent variable and the coefficients of the Cox model were determined from the training samples and subsequently evaluated with the test samples. We computed the rank product the proteins as the product of their ranks in each iteration. To estimate the FDR for the rank products, we generated a null distribution of rank products. The null distribution was generated by shuffling/permuting the samples. An FDR threshold of 0.1 gave us 27 proteins whose expressions were associated to survival.

Summary of Paper II

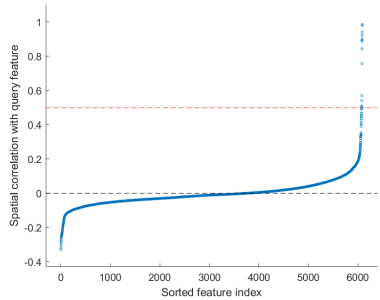
In **Paper II** we assessed the ability of some of the most popular MSI software to detect compound-related peaks. We evaluated three approaches: peak detection based using the mean spectrum with MALDIQuant, Cardinal's unknown peak detection algorithm, and the *slicing* approach. We also developed a novel method based on the distribution of all data set peak masses. Our method clusters data set peaks using an approach similar to that of Tibshirani et al. They group peak masses with a hierarchical clustering method, but since the complexity of hierarchical clustering grows exponentially with number of data points, their approach is unsuitable for high-resolution MSI data sets. Therefore,

we instead proposed a simple graph-based clustering method: sort the data-set peaks by m/z in ascending order, add edges between peaks whose inter-distance is below a small distance threshold proportional to their width, and then find peak clusters by extracting the connected components from the resulting graph. This method has $O(n \log n)$ complexity (due to the sorting step) and is thereby compatible with large data sets.

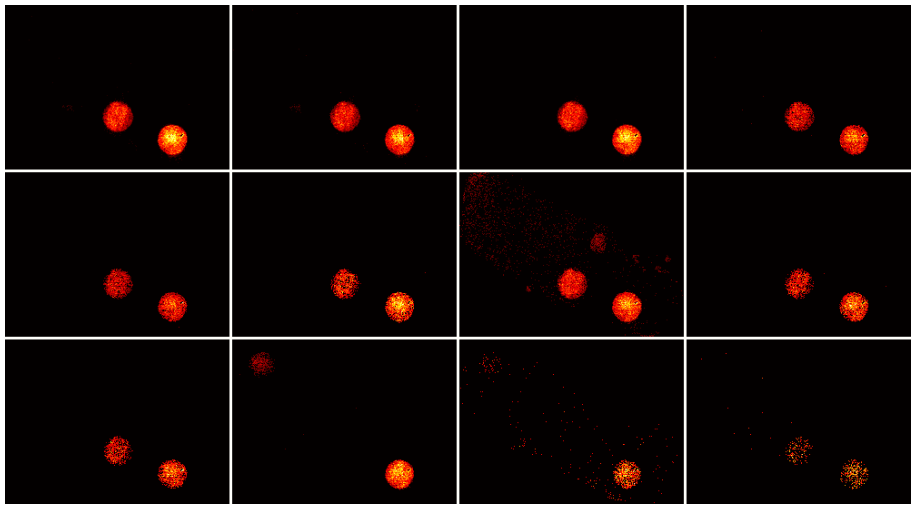
To prove that the mean spectrum and slicing approaches lack in sensitivity, we generated a ground-truth data set. We deposited mixtures of known compounds at various concentration levels on the tissue section. The spots were deliberately made small to limit the number of spectra/pixels for each compound. This gave us a data set containing known compounds whose concentrations spanned three orders of magnitude and spatial distributions were highly localized. We then tried to recall the spiked-in compounds with the peak detection algorithms of two popular MSI software (Cardinal and MALDIQuant), the slicing approach, and our novel cluster-wise KDE approach. The peak detection of MALDIQuant is based on the mean spectrum approach, and that of Cardinal unknown (we were unable to find any documentation). MALDIQuant was only able to recall one spiked-in compound; it was expected to perform poorly since the mean spectrum approach is especially bad at detecting highly localized compounds. Cardinal recalled 9, and the slicing approach recalled 10. However, the monoisotopic peaks of two and three of the compounds MALDIQuant and Cardinal recalled, respectively, were mixed with peaks from the background and other compounds. Our cluster-wise KDE approach recalled all 12 compounds with high mass accuracy (an average of 2.6 ppm). The high accuracy allowed us to correctly separate all known compound peaks from other peaks close in m/z, and to generate ion images that agreed well with the spotting patterns of the spiked-in compounds. In contrast, the average mass error of Cardinal was 13.03 ppm, which was insufficient for separating all the compounds' peaks from matrix peaks or peaks from other compounds. For each spiked-in compound, we searched for its fragment and isotope peaks in addition to its monoisotopic peak. Figure 16 shows the ion images of some fragments and isotopes of Dasatinib (one of the spiked-in compounds).



(a) Spiked-in locations.



(b) Spatial correlation Dasatinib.



(c) Most correlated ion images.

Figure 16: Ion images of the 12 most correlated peaks to that of the monoisotopic peak of Dasatinib (which we spotted at location 4 and 5). Except for two images (second and third from the left, bottom row), the images capture isotope or fragment ions of Dasatanib with minimal contamination from other ions.

Summary of Paper III

The spatial resolution of the mass spectrometer and its resolving power (RP) in the m/z dimension are critical to an MSI experiment. High spatial resolution enables molecular distributions to be related to fine tissue structures, and high resolving power is needed to distinguish different compounds with similar masses from one another. Like previously mentioned, there are many factors that limit

the spatial resolution, such as the amount of material at each tissue spot and the low ionization efficiency of the commonly used MSI setups. The resolving power depends almost exclusively on the instrument: a high-performance FT instrument can achieve an RP of 500,000 while TOF instruments rarely achieve an RP of more than 50,000. The low RP of TOF instruments can is often decreased even further by a low mass precision; systematic shifts in the measured masses of peaks over the experiment are known to be common with TOF instruments. In **Paper III**, we investigated the effect of these shifts and showed that the effective resolving power can be improved considerably by performing mass alignment.

Our mass alignment algorithm is based on the Correlation Optimized Warping (COW) algorithm [92] and relies on modeling peaks as Gaussian variations in intensity. Our peak model takes into account the peak broadening that occurs with increasing m/z for most instrument types. The exact relationship between m/z and peak width depends on the instrument type and is described by the ion separation equation for the instrument's mass analyzer. Mass alignment of an MSI data set is performed by warping the mass axis of each spectrum so that its similarity to a common reference spectrum is maximized. If the reference spectrum is calibrated prior to alignment, the overall mass accuracy of the data set can be improved as well. Calibration typically involves computing the mass shifts of a small number identified peaks in the reference spectrum.

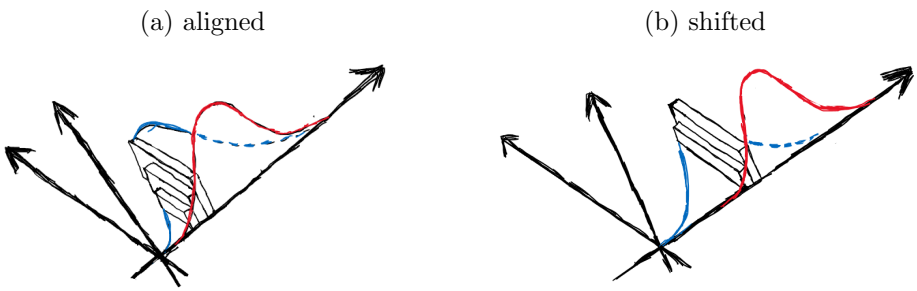


Figure 17: Visualization of peak overlap as a measure of m/z alignment. The overlap (volume blocks) is maximized when the peaks are aligned perfectly and approaches zero as the peaks are shifted relative one another.

The intensity of a peak, p_i, varies with m/z according to

$$p_i(mz) = H_i \cdot \exp(-\frac{1}{2} \cdot \frac{(\mu_i - mz)^2}{\sigma_i^2}), \tag{12}$$

where μ_i is the peak's m/z centroid location, H_i its centroid height, and σ_i its width. The overlap between two peaks, p_i and p_j, is defined as

$$I(p_i, p_j) = \int_{-\infty}^{\infty} \Big(p_i(mz) \cdot p_j(mz)\Big) dmz. \tag{13}$$

The integral in Equation 13 can be solved analytically, which is important since it must be computed repeatedly when aligning two spectra. Figure 17 illustrates how a mass shift between two peaks is reflected in their overlap; the overlap has its maximum value when the peaks are aligned perfectly, and it approaches zero as the peaks are shifted relative to each other in either direction. We also define a similarity score, B, between two spectra, S_1 and S_2 in the following manner:

$$B(S_1, S_2) = \sum_{|\mu_i - \mu_j| < \epsilon} I(p_i, p_j), \tag{14}$$

where ϵ depends on the peak width, σ. The purpose of the criterion in Equation 14 is to reduce the number of pairwise overlap computations in B. A value between 4σ and 6σ for the threshold ϵ is reasonable since the overlap is negligible at larger distances. The similarity score is a general measure of similarity between two centroid spectra, and it can be used for multiple purposes.

Aligning two spectra in the mass dimension is equivalent to maximizing their similarity score. We do this by warping the mass axis of one of the spectra so that it matches that of the other. We split the mass axis into segments and allow each segment to be stretched, compressed, or shifted either upward or downward in m/z. We refer to the points between two segments as the warping nodes. The set of possible warpings is defined by all combinations of warping node shifts. To find the optimal alignment, we evaluate B for each segment individually and then find the optimal combination of shifts with *Dynamic Programming*.

The combination of our pairwise similarity score and the segment-based warping from COW results in a flexible, yet robust, alignment algorithm. Another virtue of our method is its compatibility with centroid spectra. A centroided spectrum is a list of m/z-intensity pairs, and its data size is much smaller than that of a continuous spectrum. Public repositories such as MetaSpace therefore often store MSI data sets in centroid mode. These repositories contain hundreds of data sets, many of which are generated in different laboratories and/or with different instruments. Mass alignment facilitates direct comparisons between such data sets, which is highly valuable because it enables public data sets to be used for validation purposes in biomarker studies.

We applied our mass alignment method, called MSIWarp, to four publicly available data sets and were able to demonstrate improvements of up to 95% in mass precision. The data sets were generated with different mass analyzers and ionization techniques. Our results thereby indicate that our method performs

well for data sets from multiple instrument setups, which makes it especially suitable when comparing data sets from different laboratories. Figure 18 shows the mass precision of a peak from one of the TOF data sets before and after alignment. The improvement in mass precision after alignment is striking, and it enabled us to separate peaks that were initially indistinguishable.

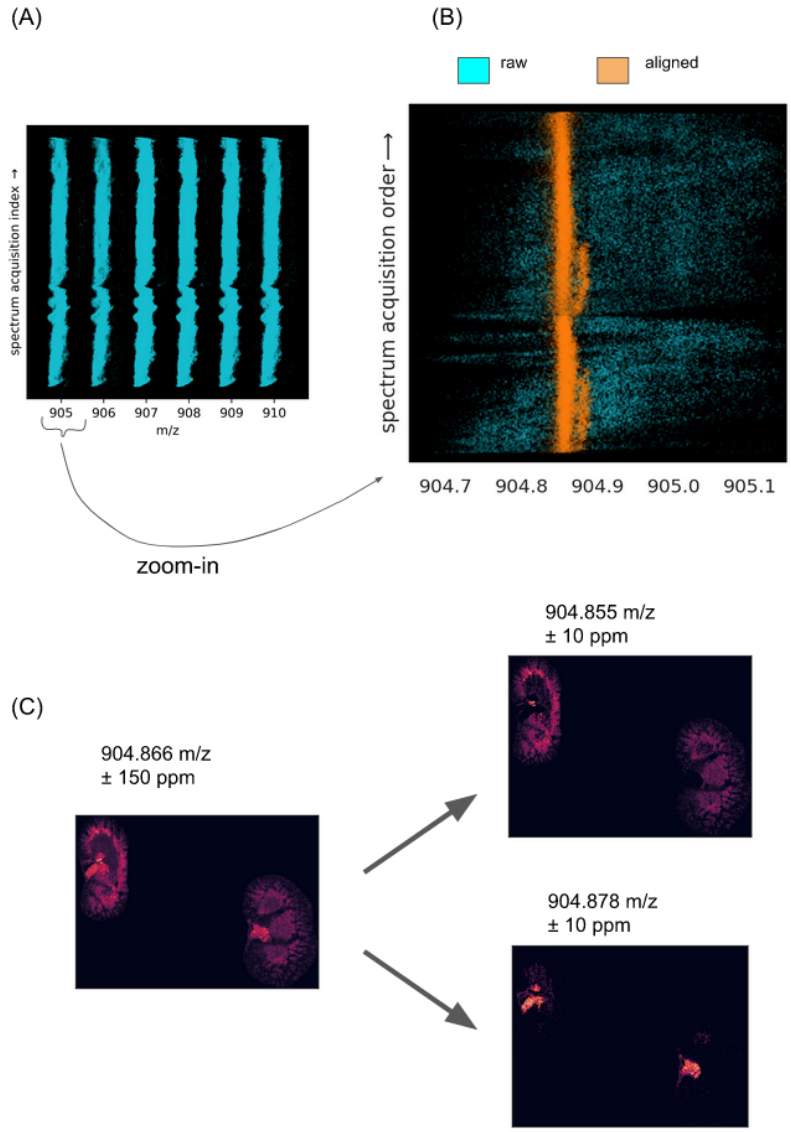


Figure 18: (A): mass scatter compound isotopes. (B): zoom-in on one mass scatter before and after alignment. (C): alignment allows the compounds to be separated.

Summary of Paper IV

The focus of the study presented in **Paper IV** was again on the relationship between the protein and peptide expression of tumor tissue and survival in melanoma. This time, however, we went further with the molecular characterization of the tissue. We aimed to unify the proteomic, phosphoproteomic, transcriptomic expressions with in-depth histopathology analysis, and relate these data to clinical variables. Survival analysis was performed with two different approaches: Cox analysis and outlier analysis (OLA). Instead of using a PLS-Cox model to select survival-related compounds like we did in **Paper I**, we used the regularized reformulation of the Cox model described by Simon et al.[93]. Feature selection was performed in two steps: the features whose univariate Cox coefficient was above a specific threshold were initially used as input to the regularized Cox model. The features that survived the regularization step were then selected as the final features. This procedure was repeated 100 times inside a cross-validation loop, and the features that were selected in at least 50 of the 100 repetitions were defined as related to survival. Aggregation of the results from the outlier analysis and Cox analysis yielded a total of 298 survival-related proteins. Out of these, 9 were selected to be validated in an independent cohort with immunohistochemical (IHC) characterization. The independent IHC validation cohort consisted of primary melanomas from 42 patients. Some of these patients developed locoregional or distant metastases during the follow-up period. Nine candidate biomarkers were studied by immunohistochemical analysis.

We also searched for independent components in the different -omic data sets with ICA and then investigated the association between the independent components and clinical features. We found that multiple independent components were significantly related to several clinical features, including survival. We performed the same analysis with PCA instead of ICA and found that the principal components generally exhibited a weaker relationship to the clinical features than the independent components. This suggests that the multi-omic data sets are better represented by additive subsets of independent non-Gaussian sources rather than by pieces of uncorrelated information.

The fact that we were unable to validate the protein signature derived in the first study (**Paper I**) has many possible explanations. Firstly, melanoma is known to be a highly heterogeneous disease. Secondly, we used DDA MS in the first study, which may have introduced uncertainties in the data that led to spurious findings. Regardless, these results highlight the importance of validating biomarker candidates.

Conclusions and Future Perspectives

Mass Spectrometry and other high-throughput techniques have changed how we approach many complex and challenging questions in cancer research. Still, the full potential of mass spectrometry is yet to be realized. Low reproducibility, largely due to the randomness of DDA MS, has been a longstanding obstacle for research based on LC-MS. Improving the quality of the preprocessing of LC-MS data is essential to achieving higher reproducibility, and new algorithms and software MS data processing are published at a high rate. The MSI field is similarly dependent on reproducible results, and, during my thesis, I have focused on improving some of the most essential steps of MSI data preprocessing. The results from **Paper II** indicate that routinely used software packages miss a substantial fraction of compounds, particularly the faintly expressed ones. Sensitivity is essential for reproducibility, and the method we proposed highlights that there are still considerable improvements to be made.

Shareable data is critical to the success of the research fields related to mass spectrometry. The first requirement for easily shareable data is a common data format. For LC-MS data the common format is "mzML", while that for MSI data is "imzML".[94;95] Instrument vendors have gradually improved their support for these formats throughout recent years, yet some compatibility issues remain. Public data repositories that are convenient to use is a second requirement, and notable examples of such repositories include ProteomeXchange, MetaSpace, and MetaboLights.[96;97] The key to maximizing the utility of these repositories is the availability of software packages that can process data sets from different instruments types and vendors. The software we published together with **Paper III**, MSIWarp, is such an example.[98] Together with the peak detection method presented in **Paper II**, it can hopefully help improve reproducibility and facilitate data sharing in the MSI field.

It is important to remember what a proteomic data set represents: a snapshot of the proteome at the time of sample collection. On its own, a single snapshot is insufficient to fully understand how disease processes develop within the tissue, how and when the tissue responds to treatment, and how the tissue affects and is affected by neighboring tissue. This limitation is hardly unique to MS proteomics; every study based on in-vitro experiments is limited in the same sense. A complete understanding of the disease process can only be gained from continuous measurements of the same tissue, but collecting a sample is always invasive to some degree, especially when collecting a large amount of tissue. At the same time, each sample must contain enough biological material to reflect the state of the tissue, regardless of the sensitivity of the analytical technique. Frequent and systematic sample collection requires substantial dedication from individuals participating in disease studies, particularly when the disease is

cancer. Collecting tumor tissue from the same individual over a long time is rarely an option because it is critical to remove all the cancerous tissue as soon as possible to minimize the risk of metastasis. If an individual is unfortunate enough to develop metastases, additional tissue material may be collected from subsequent surgeries, but the previous principle is still true: no cancerous tissue should be left after surgery. Therefore, it is hard to get more than a couple of samples from the same individual.

For most cancers, more and better biomarkers are needed to paint a more complete picture than the one we currently have, and a necessary step toward obtaining them is developing better analytical techniques. Perhaps even more important is to facilitate data sharing. The lack of validation samples were a limitation to the study described in **Paper I**, and although the validation cohort we used in the follow-up study (**Paper IV**) adds confidence to its conclusions, more samples are still needed to fully validate the biomarkers it proposes. This is especially true due to the heterogeneity of melanoma.

Beyond having access to data from other research groups is having permission and incentive to share our own. Encouraging data sharing is a political rather than a scientific task; clinical samples are a valuable commodity, and research groups often compete for the same grant money. This has the unfortunate consequence that many groups protect their data, even after publishing their studies. The scarcity of tissue samples remains a major bottleneck in cancer research, and in addition to ensuring a high experimental quality and consistency, sharing data must become a top priority for any research organization, be it a global, national, or regional one.

Populärvetenskaplig Sammanfattning

Biologiska system förefaller vara nästan oändligt komplexa. Människans kropp sägs innehålla flera biljarder celler som utför en mängd olika uppgifter och som utgör olika typer av vävnad. En enskild cell är i sin tur en komplicerad organism som är uppbyggd av proteiner, lipider och andra biomolekyler. Att fullständigt förstå ett sjukdomsförlopp är därför ingen enkel uppgift. Att dessutom kunna styra det för att bota sjukdomen är ännu svårare. Trots det utvecklas det ständigt nya läkemedel och behandlingsmetoder som förbättrar våra chanser att bli botade från svåra sjukdomer och att leva hälsosamma liv.

Utveckling inom gensekvensering har möjliggjort genetisk karaktärisering av vävnadsprover. Detta har i sin tur lett till länken mellan genetik och sjukdom studerats i stor utsträckning. En organisms genetik kan säga mycket om hur den troligtvis kommer bete sig i olika sammanhang. Proteinerna är dock de molekyler som faktiskt utför många av de funktioner som krävs för att upprättahålla organismen, så som energiproduktion och replikation. Att studera proteiner och hur de påverkar vid och påverkas av sjukdom är därför naturligt. Den tekniken som på senaste år visat störst potential för att mäta proteiner i stor skala är masspektrometri. Att studerna proteiner med hjälp av masspektrometri är dock långt ifrån trivialt: det krävs en noggrann förberedelse av vävnadsprovet innan det kan analyseras av instrumentet och sofistikerade algoritmer och datorprogram för att analysera mätdatan.

En masspektrometer joniserar molekyler i ett prov och separerar dem sedan baserat på deras molekylvikt delat på laddning. Utdatan efter mätning av ett prov är ett eller flera masspektra. Ett masspektrum är en fördelning av molekylvikter. Masspektrometrar kan analysera flera typer av molekyler, men de som oftast studeras i medicinska sammanhang är proteiner/peptider eller metaboliter.

I min avhandling har jag fokuserat på tillämpningen av masspektrometri inom biologisk och medicinsk forskning. Arbetet som ligger till grund för **Artikel I** bestod av en retroaktiv studie av patienter med metastaserat malignt melanom. Vi analyserade tumörvävnad med masspektrometri och länkade därefter uppmätt proteindata till patientöverlevnad. Jag har också utvecklat två metoder för att förfina instrumentdata med målet att i slutändan kunna få så hög kvalitet på mätdatan som möjligt. Den första metoden (**Artikel II**) ökar sensitiviteten i MSI. Den andra metoden, som vi beskriver i **Artikel III**, korrigerar små förskjutningar i mass-dimensionen mellan masspektra. Om förskjutningarna inte korrigeras kan det leda till att somliga molekyler skuggas av andra och därmed blir osynliga i masspektran. Slutligen har vi även genomfört en fortsättningsstudie till studien som beskrivs i **Artikel I**. I fortsättningsstudien (**Artikel IV**) tillämpade vi kemisk ”labeling” för att kvantifiera fler proteiner

med större mätsäkerhet. Vi utökade även våran statistiska analys med flera andra metoder som är komplementära till de vi använde i den första studien.

Bibliography

[1] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. Dna sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, 2017.

[2] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[3] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[4] Zimei Bu and David JE Callaway. Proteins move! protein dynamics and long-range allostery in cell signaling. *Advances in protein chemistry and structural biology*, 83:163–221, 2011.

[5] Shina CL Kamerlin and Arieh Warshel. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins: Structure, Function, and Bioinformatics*, 78(6):1339–1375, 2010.

[6] Biomarkers Definitions Working Group, Arthur J Atkinson Jr, Wayne A Colburn, Victor G DeGruttola, David L DeMets, Gregory J Downing, Daniel F Hoth, John A Oates, Carl C Peck, Robert T Schooley, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology & therapeutics*, 69(3):89–95, 2001.

[7] Kyle Strimbu and Jorge A Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463, 2010.

[8] Mathias Uhlén, Linn Fagerberg, Björ M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, Ing Marie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan

Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle Von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar Von Heijne, Jens Nielsen, and Fredrik Pontén. Tissue-based map of the human proteome. *Science*, 347(6220), 2015. ISSN 10959203. doi: 10.1126/science.1260419.

[9] Mathias Uhlen, Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, Muhammad Arif, Zhengtao Liu, Fredrik Edfors, Kemal Sanli, Kalle Von Feilitzen, Per Oksvold, Emma Lundberg, Sophia Hober, Peter Nilsson, Johanna Mattsson, Jochen M. Schwenk, Hans Brunnström, Bengt Glimelius, Tobias Sjöblom, Per Henrik Edqvist, Dijana Djureinovic, Patrick Micke, Cecilia Lindskog, Adil Mardinoglu, and Fredrik Ponten. A pathology atlas of the human cancer transcriptome. *Science*, 357(6352), 2017. ISSN 10959203. doi: 10.1126/science.aan2507.

[10] Andreas Scherer. *Batch effects and noise in microarray experiments: sources and solutions*, volume 868. John Wiley & Sons, 2009.

[11] Frederic Chibon. Cancer gene expression signatures–the rise and fall? *European journal of cancer*, 49(8):2000–2009, 2013.

[12] Mike S. Lee and Edward H. Kerns. LC/MS applications in drug development. *Mass Spectrometry Reviews*, 18(3-4):187–279, 1999. ISSN 02777037. doi: 10.1002/(SICI)1098-2787(1999)18:3/4¡187::AID-MAS2¿3.0.CO;2-K.

[13] Franz Hillenkamp, Michael Karas, Ronald C Beavis, and Brian T Chait. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical chemistry*, 63(24):1193A–1203A, 1991.

[14] Melvin B Comisarow and Alan G Marshall. Fourier transform ion cyclotron resonance spectroscopy. *Chemical physics letters*, 25(2):282–283, 1974.

[15] Alexander Makarov. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical chemistry*, 72(6): 1156–1162, 2000.

[16] Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006. ISSN 00032700. doi: 10.1021/ac051437y.

[17] Alisdair R. Fernie, Richard N. Trethewey, Arno J. Krotzky, and Lothar Willmitzer. Metabolite profiling: From diagnostics to systems biology. *Nature Reviews Molecular Cell Biology*, 5(9):763–769, 2004. ISSN 14710072. doi: 10.1038/nrm1451.

[18] RB Cody and BS Freiser. Collision-induced dissociation in a fourier-transform mass spectrometer. *International Journal of Mass Spectrometry and Ion Physics*, 41(3):199–204, 1982.

[19] Philipp E. Geyer, Nils A. Kulak, Garwin Pichler, Lesca M. Holdt, Daniel Teupser, and Matthias Mann. Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Systems*, 2(3):185–195, 2016. ISSN 24054720. doi: 10.1016/j.cels.2016.02.015. URL `http://dx.doi.org/10.1016/j.cels.2016.02.015`.

[20] Ian S Gilmore, Sven Heiles, and Cornelius L Pieterse. Metabolic imaging at the single-cell scale: recent advances in mass spectrometry imaging. *Annual Review of Analytical Chemistry*, 12:201–224, 2019.

[21] Jiaying Han, Hjalmar Permentier, Rainer Bischoff, Geny Groothuis, Angela Casini, and Péter Horvatovich. Imaging of protein distribution in tissues using mass spectrometry: An interdisciplinary challenge. *TrAC Trends in Analytical Chemistry*, 112:13–28, 2019.

[22] Amanda Rae Buchberger, Kellen DeLaney, Jillian Johnson, and Lingjun Li. Mass spectrometry imaging: a review of emerging advancements and future insights. *Analytical chemistry*, 90(1):240–265, 2018.

[23] Bernhard Spengler. Mass spectrometry imaging of biomolecular information. *Analytical chemistry*, 87(1):64–82, 2015.

[24] Andrew Palmer, Prasad Phapale, Ilya Chernyavsky, Regis Lavigne, Dominik Fay, Artem Tarasov, Vitaly Kovalev, Jens Fuchser, Sergey Nikolenko, Charles Pineau, et al. Fdr-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature methods*, 14(1):57–60, 2017.

[25] Jimmy K Eng, Ashley L McCormack, and John R Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the american society for mass spectrometry*, 5(11):976–989, 1994.

[26] Andrew J Link, Jimmy Eng, David M Schieltz, Edwin Carmack, Gregory J Mize, David R Morris, Barbara M Garvik, and John R Yates. Direct analysis of protein complexes using mass spectrometry. *Nature biotechnology*, 17(7):676–682, 1999.

[27] David N. Perkins, Darryl J.C. Pappin, David M. Creasy, and John S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20 (18):3551–3567, 1999. ISSN 01730835. doi: 10.1002/(SICI)1522-2683(19991201)20:18¡3551::AID-ELPS3551¿3.0.CO;2-2.

[28] Robertson Craig and Ronald C. Beavis. TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, 2004. ISSN 13674803. doi: 10.1093/bioinformatics/bth092.

[29] Andrew D. Keller, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical statistical model to estimate the accuracy of protein identifications made by MS/MS and database search. *Proceedings 50th ASMS Conference on Mass Spectrometry and Allied Topics*, 74(20):37–38, 2002.

[30] Alexey I Nesvizhskii, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry*, 75(17):4646–58, 2003. ISSN 0003-2700. doi: 10.1021/ac0341261. URL `http://www.ncbi.nlm.nih.gov/pubmed/14632076`.

[31] Alexey I. Nesvizhskii and Ruedi Aebersold. Interpretation of Shotgun Proteomic Data: The Protein Inference Problem. *Molecular & Cellular Proteomics*, pages 1419–1440, 2005. doi: 10.1074/mcp.R500012-MCP200.

[32] Lazaro Hiram Betancourt, Krzysztof Pawłowski, Jonatan Eriksson, A Marcell Szasz, Shamik Mitra, Indira Pla, Charlotte Welinder, Henrik Ekedahl, Per Broberg, Roger Appelqvist, et al. Improved survival prognostication of node-positive malignant melanoma patients utilizing shotgun proteomics guided by histopathological characterization and genomic data. *Scientific reports*, 9(1):1–14, 2019.

[33] Michael A Gillette and Steven A Carr. Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. *Nature methods*, 10(1):28, 2013.

[34] Justina C. Wolters, Jolita Ciapaite, Karen Van Eunen, Klary E. Niezen-Koning, Alix Matton, Robert J. Porte, Peter Horvatovich, Barbara M. Bakker, Rainer Bischoff, and Hjalmar P. Permentier. Translational Targeted Proteomics Profiling of Mitochondrial Energy Metabolic Pathways in Mouse and Human Samples. *Journal of Proteome Research*, 15(9):3204–3213, 2016. ISSN 15353907. doi: 10.1021/acs.jproteome.6b00419.

[35] Ludovic C. Gillet, Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, and Ruedi Aebersold. Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *2*, 11(6):O111.016717, jun 2012. ISSN 1535-9476. doi: 10.1074/mcp.O111.016717. URL `http://www.mcponline.org/lookup/doi/10.1074/mcp.O111.016717`.

[36] Hannes L Röst, George Rosenberger, Pedro Navarro, Ludovic Gillet, Saša M Miladinović, Olga T Schubert, Witold Wolski, Ben C Collins, Johan Malmström, Lars Malmström, et al. Openswath enables automated, targeted analysis of data-independent acquisition ms data. *Nature biotechnology*, 32 (3):219–223, 2014.

[37] Stephane Houel, Robert Abernathy, Kutralanathan Renganathan, Karen Meyer-Arendt, Natalie G. Ahn, and William M. Old. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *Journal of Proteome Research*, 9(8):4152–4160, 2010. ISSN 15353893. doi: 10.1021/pr1003856.

[38] Andrew Thompson, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms. *Analytical chemistry*, 75(8):1895–1904, 2003.

[39] Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, and Michael J MacCoss. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7):966–968, 2010.

[40] Hendrik Weisser, Sven Nahnsen, Jonas Grossmann, Lars Nilse, Andreas Quandt, Hendrik Brauer, Marc Sturm, Erhan Kenar, Oliver Kohlbacher, Ruedi Aebersold, and Lars Malmström. An automated pipeline for high-throughput label-free quantitative proteomics. *Journal of Proteome Research*, 12(4):1628–1644, 2013. ISSN 15353893. doi: 10.1021/pr300992u.

[41] Chih-Chiang Tsou, Dmitry Avtonomov, Brett Larsen, Monika Tucholska, Hyungwon Choi, Anne-Claude Gingras, and Alexey I Nesvizhskii. Diaumpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods*, 12(3):258–264, 2015.

[42] Vadim Demichev, Christoph B Messner, Spyros I Vernardis, Kathryn S Lilley, and Markus Ralser. Dia-nn: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature methods*, 17(1):41–44, 2020.

[43] Viktoria Dorfer, Peter Pichler, Thomas Stranzl, Johannes Stadlmann, Thomas Taus, Stephan Winkler, and Karl Mechtler. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of Proteome Research*, 13(8):3679–3684, 2014. ISSN 15353907. doi: 10.1021/pr500202e.

[44] David Fenyö and Ronald C Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry*, 75(4):768–774, 2003.

[45] Shivani Tiwary, Roie Levy, Petra Gutenbrunner, Favio Salinas Soto, Krishnan K Palaniappan, Laura Deming, Marc Berndl, Arthur Brant, Peter Cimermancic, and Jürgen Cox. High-quality ms/ms spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature methods*, 16(6):519–525, 2019.

[46] Siegfried Gessulat, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Julia Rechenberger, Bernard Delanghe, Andreas Huhmer, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods*, 16(6):509–518, 2019.

[47] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for mass spectrometry-based proteomics. In *Proteome bioinformatics*, pages 55–71. Springer, 2010.

[48] Lukas Käll, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, 2007. ISSN 15487091. doi: 10.1038/nmeth1113.

[49] Frank Suits, Berend Hoekman, Therese Rosenling, Rainer Bischoff, and Peter Horvatovich. Threshold-avoiding proteomics pipeline. *Analytical chemistry*, 83(20):7786–7794, 2011.

[50] Hui Zhang, Tao Liu, Zhen Zhang, Samuel H Payne, Bai Zhang, Jason E McDermott, Jian-Ying Zhou, Vladislav A Petyuk, Li Chen, Debjit Ray, et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell*, 166(3):755–765, 2016.

[51] David J Clark, Saravana M Dhanasekaran, Francesca Petralia, Jianbo Pan, Xiaoyu Song, Yingwei Hu, Felipe da Veiga Leprevost, Boris Reva, Tung-Shing M Lih, Hui-Yin Chang, et al. Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell*, 179(4):964–983, 2019.

[52] Yongchao Dou, Emily A Kawaler, Daniel Cui Zhou, Marina A Gritsenko, Chen Huang, Lili Blumenberg, Alla Karpova, Vladislav A Petyuk, Sara R Savage, Shankha Satpathy, et al. Proteogenomic characterization of endometrial carcinoma. *Cell*, 180(4):729–748, 2020.

[53] Philipp Mertins, DR Mani, Kelly V Ruggles, Michael A Gillette, Karl R Clauser, Pei Wang, Xianlong Wang, Jana W Qiao, Song Cao, Francesca Petralia, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605):55–62, 2016.

[54] Suhas Vasaikar, Chen Huang, Xiaojing Wang, Vladislav A Petyuk, Sara R Savage, Bo Wen, Yongchao Dou, Yun Zhang, Zhiao Shi, Osama A Arshad, et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell*, 177(4):1035–1049, 2019.

[55] Charles Ansong, Samuel O Purvine, Joshua N Adkins, Mary S Lipton, and Richard D Smith. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Briefings in Functional Genomics and Proteomics*, 7(1):50–62, 2008.

[56] Nitin Gupta, Stephen Tanner, Navdeep Jaitly, Joshua N Adkins, Mary Lipton, Robert Edwards, Margaret Romine, Andrei Osterman, Vineet Bafna, Richard D Smith, et al. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome research*, 17(9):1362–1377, 2007.

[57] Alexey I Nesvizhskii. Proteogenomics: concepts, applications and computational strategies. *Nature methods*, 11(11):1114, 2014.

[58] Andy T. Kong, Felipe V. Leprevost, Dmitry M. Avtonomov, Dattatreya Mellacheruvu, and Alexey I. Nesvizhskii. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, 14(5):513–520, 2017. ISSN 15487105. doi: 10.1038/nmeth.4256.

[59] Theodore Alexandrov. Spatial metabolomics and imaging mass spectrometry in the age of artificial intelligence. *Annual Review of Biomedical Data Science*, 3, 2020.

[60] Shaji Krishnan, Jack T.W.E. Vogels, Leon Coulier, Richard C. Bas, Margriet W.B. Hendriks, Thomas Hankemeier, and Uwe Thissen. Instrument and process independent binning and baseline correction methods for liquid chromatography-high resolution-mass spectrometry deconvolution. *Analytica Chimica Acta*, 740:12–19, 2012. ISSN 00032670. doi: 10.1016/j.aca.2012.06.014. URL `http://dx.doi.org/10.1016/j.aca.2012.06.014`.

[61] Jonatan O Eriksson, Melinda Rezeli, Max Hefner, Gyorgy Marko-Varga, and Peter Horvatovich. Clusterwise peak detection and filtering based on spatial distribution to efficiently mine mass spectrometry imaging data. *Analytical chemistry*, 91(18):11888–11896, 2019.

[62] Judith M Fonville, Claire Carter, Olivier Cloarec, Jeremy K Nicholson, John C Lindon, Josephine Bunch, and Elaine Holmes. Robust data processing and normalization strategy for maldi mass spectrometric imaging. *Analytical chemistry*, 84(3):1310–1319, 2012.

[63] Theodore Alexandrov and Andreas Bartels. Testing for presence of known and unknown molecules in imaging mass spectrometry. *Bioinformatics*, 29 (18):2335–2342, 2013.

[64] Frank Suits, Thomas E Fehniger, Akos Vegvari, Gyorgy Marko-Varga, and Peter Horvatovich. Correlation queries for mass spectrometry imaging. *Analytical chemistry*, 85(9):4398–4404, 2013.

[65] Theodore Alexandrov, Katja Ovchnnikova, Andrew Palmer, Vitaly Kovalev, Artem Tarasov, Lachlan Stuart, Renat Nigmetzianov, Dominik Fay, and Key METASPACE Contributors. Metaspace: A community-populated knowledge base of spatial metabolomes in health and disease. *BioRxiv*, page 539478, 2019.

[66] Patrik Källback, Mohammadreza Shariatgorji, Anna Nilsson, and Per E Andrén. Novel mass spectrometry imaging software assisting labeled normalization and quantitation of drugs and neuropeptides directly in tissue sections. *Journal of proteomics*, 75(16):4941–4951, 2012.

[67] Sue Ah Noh, Su-Mi Kim, Seon Hwa Park, Dong-Jin Kim, Joon Won Lee, Yang Gyun Kim, Ju-Young Moon, Sung-Jig Lim, Sang-Ho Lee, and Kwang Pyo Kim. Alterations in lipid profile of the aging kidney identified by maldi imaging mass spectrometry. *Journal of proteome research*, 18(7): 2803–2812, 2019.

[68] Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin Van Der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

[69] Marc J Van De Vijver, Yudong D He, Laura J Van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

[70] Therese Sørlie, Charles M Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001.

[71] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160, 2009.

[72] Carlos M Carvalho, Jeffrey Chang, Joseph E Lucas, Joseph R Nevins, Quanli Wang, and Mike West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

[73] Robert Clarke, Habtom W Ressom, Antai Wang, Jianhua Xuan, Minetta C Liu, Edmund A Gehan, and Yue Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature reviews cancer*, 8(1):37–49, 2008.

[74] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[75] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[76] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[77] Charles W Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.

[78] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[79] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[80] Yuan Yuan, Eliezer M Van Allen, Larsson Omberg, Nikhil Wagle, Ali Amin-Mansour, Artem Sokolov, Lauren A Byers, Yanxun Xu, Kenneth R Hess, Lixia Diao, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature biotechnology*, 32(7):644–652, 2014.

[81] Svante Wold, Arnold Ruhe, Herman Wold, and WJ Dunn, Iii. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984.

[82] Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

[83] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, pages 104–117, 2003.

[84] Christin Christin, Huub CJ Hoefsloot, Age K Smilde, Berend Hoekman, Frank Suits, Rainer Bischoff, and Peter Horvatovich. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Molecular & Cellular Proteomics*, 12(1):263–276, 2013.

[85] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[86] Douglas G Altman and Patrick Royston. The cost of dichotomising continuous variables. *Bmj*, 332(7549):1080, 2006.

[87] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[88] Eric Vittinghoff and Charles E McCulloch. Relaxing the rule of ten events per variable in logistic and cox regression. *American journal of epidemiology*, 165(6):710–718, 2007.

[89] Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

[90] Danh V Nguyen and David M Rocke. Partial least squares proportional hazard regression for application to dna microarray survival data. *Bioinformatics*, 18(12):1625–1632, 2002.

[91] Matthew D Wilkerson and D Neil Hayes. Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12):1572–1573, 2010.

[92] Niels-Peter Vest Nielsen, Jens Michael Carstensen, and Jørn Smedsgaard. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of chromatography A*, 805(1-2):17–35, 1998.

[93] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.

[94] Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H. Tang, Andreas Römpp, Steffen Neumann, Angel D. Pizarro, Luisa Montecchi-Palazzi, Natalie Tasman, Mike Coleman, Florian Reisinger, Puneet Souda, Henning Hermjakob, Pierre-Alain Binz, and Eric W. Deutsch. mzML—a Community Standard for Mass Spectrometry Data. *Molecular & Cellular Proteomics*, 10(1):R110.000133, jan 2011. ISSN 1535-9476. doi: 10.1074/mcp.R110.000133. URL `http://www.mcponline.org/lookup/doi/10.1074/mcp.R110.000133`.

[95] Thorsten Schramm, Zoë Hester, Ivo Klinkert, Jean-Pierre Both, Ron MA Heeren, Alain Brunelle, Olivier Laprévote, Nicolas Desbenoit, Marie-France Robbe, Markus Stoeckli, et al. imzml—a common data format for the flexible exchange and processing of mass spectrometry imaging data. *Journal of proteomics*, 75(16):5106–5110, 2012.

[96] Juan A Vizcaíno, Eric W Deutsch, Rui Wang, Attila Csordas, Florian Reisinger, Daniel Ríos, José A Dianes, Zhi Sun, Terry Farrah, Nuno Bandeira, et al. Proteomexchange provides globally coordinated proteomics

data submission and dissemination. *Nature biotechnology*, 32(3):223–226, 2014.

[97] Kenneth Haug, Reza M Salek, Pablo Conesa, Janna Hastings, Paula De Matos, Mark Rijnbeek, Tejasvi Mahendraker, Mark Williams, Steffen Neumann, Philippe Rocca-Serra, et al. Metabolights—an open-access general-purpose repository for metabolomics studies and associated metadata. *Nucleic acids research*, 41(D1):D781–D786, 2013.

[98] Jonatan O Eriksson, Alejandro Sanchez Brotons, Melinda Rezeli, Frank Suits, Gyorgy Marko-Varga, and Peter Horvatovich. Msiwarp: A general approach to mass alignment in mass spectrometry imaging. *Analytical Chemistry*, 2020.

SCIENTIFIC REPORTS

OPEN

Improved survival prognostication of node-positive malignant melanoma patients utilizing shotgun proteomics guided by histopathological characterization and genomic data

Received: 20 August 2018
Accepted: 13 March 2019
Published online: 26 March 2019

Lazaro Hiram Betancourt[1], Krzysztof Pawłowski[1,2], Jonatan Eriksson[1], A. Marcell Szasz[1,4,5], Shamik Mitra[1], Indira Pla[1], Charlotte Welinder[1], Henrik Ekedahl[1], Per Broberg[1], Roger Appelqvist[1], Maria Yakovleva[1], Yutaka Sugihara[1], Kenichi Miharada[1], Christian Ingvar[1], Lotta Lundgren[1], Bo Baldetorp[1], Håkan Olsson[1], Melinda Rezeli[1], Elisabet Wieslander[1], Peter Horvatovich[1,3], Johan Malm[1], Göran Jönsson[1] & György Marko-Varga[1,6]

Metastatic melanoma is one of the most common deadly cancers, and robust biomarkers are still needed, e.g. to predict survival and treatment efficiency. Here, protein expression analysis of one hundred eleven melanoma lymph node metastases using high resolution mass spectrometry is coupled with in-depth histopathology analysis, clinical data and genomics profiles. This broad view of protein expression allowed to identify novel candidate protein markers that improved prediction of survival in melanoma patients. Some of the prognostic proteins have not been reported in the context of melanoma before, and few of them exhibit unexpected relationship to survival, which likely reflects the limitations of current knowledge on melanoma and shows the potential of proteomics in clinical cancer research.

The incidence of malignant melanoma is increasing worldwide, particularly in Western countries, and survival does not seem to improve substantially[1]. Primary surgery is curative in most patients but around 10–15% of tumors are showing progression. Thus, it is important to early identify those patients who carry a skin tumor with progressive pathobiology. Currently, Breslow thickness is the most accurate tool for predicting the disease outcome of primary melanoma[2]. To improve the prediction of disease outcome, more fine-tuned molecular profiling and data integration tools and efforts are needed to search for alternative biomarkers[3].

Metastatic melanoma (MM) still remains a tumor with poor outcome[4,5] despite interventions with targeted therapy and antibody-driven modulation of the immune response[6–11].

Recent technological developments utilizing both genomic and proteomic analysis provide the opportunity to identify better predictive markers of melanomas[12–16]. It is possible to monitor the expression of certain genes and also gain understanding how these genes are expressed and regulated as functional proteins. Accordingly, detailed, personalized information on gene and protein expression and regulation, as well as data on specific mutations that may guide the treatment, can be monitored. Another cornerstone of prognostic predictions is

[1]Lund University, Lund, Sweden. [2]Warsaw University of Life Sciences SGGW, Warszawa, Poland. [3]University of Groningen, Groningen, The Netherlands. [4]National Koranyi Institute of Pulmonology, Budapest, Hungary. [5]Semmelweis University, Budapest, Hungary. [6]Tokyo Medical University, Tokyo, Japan. Lazaro Hiram Betancourt, Krzysztof Pawłowski, Jonatan Eriksson and A. Marcell Szasz contributed equally. Correspondence and requests for materials should be addressed to K.P. (email: krzysztof_pawlowski@sggw.pl) or G.M.-V. (email: gyorgy.marko-varga@bme.lth.se)

Clinicopathological properties		n	% of total
Gender	*Female*	43	39
	Male	68	61
Location	*trunk*	47	42
	head/neck	1	1
	upper extremity	12	11
	lower extremity	27	24
	other	7	6
Histological type	*SSM*	27	24
	NM	35	32
	ALM	4	4
	LMM	1	1
	Mucosal	1	1
	Other	1	1
	Unknown	13	12
Clark level	*1*	1	1
	2	4	4
	3	25	23
	4	43	39
	5	5	5
Breslow scale mm	*<1.00*	11	10
	<2.00	26	23
	<3.00	23	21
	<4.00	27	24
BRAF status	*V600E mut*	38	34
	V600K mut	3	3
	V600A mut	1	1
	WT	64	58

Table 1. Clinicopathological information about the patients and patient samples. Histological types: ALM - acral lentiginous melanoma, SSM - superficial spreading melanoma, NM - nodular melanoma, LMM - lentigo maligna melanoma.

clinicopathological characterization based on high quality pathological and clinical information. Equally important is to investigate the cellular composition of the tissue, to morphologically assess in detail the quality of tumor samples submitted for analysis and the identification of features important for disease progression.

In this study, we combine in depth histopathology analysis of melanoma lymph node metastases with deep-mining protein expression analysis using high-resolution mass spectrometry and a complex bioinformatics workflow to integrate clinical data with protein and genomics profile information. The protein data is matched to genomic analysis of the same tumor tissue. This information coupled with extensive clinical information on each subject provides an excellent opportunity to identify novel protein markers to predict progression and survival of melanoma.

Results and Discussion

Clinical data. A total of 111 patients diagnosed with melanoma metastasis between 1975 and 2011 were evaluated in the study (Table 1). There were 68 men and 43 women among the investigated cases. Average age ± standard deviation (range) at diagnosis of lymph node metastasis was 62.4 ± 13.7 (25–89) years. The time elapsed to progression from primary tumor to lymph node metastasis was 5.0 ± 5.6 (0–18.0) years and overall survival was 7.9 ± 6.8 (0.2–43.0) years. The dominant histotypes of primary tumors were Superficial Spreading Melanoma (SSN) and Nodular Melanoma (NM) (see Table 1). The cohort included 59% of patients with wild type BRAF status and 36% of patients with V600E mutation in the BRAF gene (4% had V600A or V600K mutation).

Histopathological data. Frozen specimens (snap frozen immediately after surgery) were subjected to this evaluation. In order to relate protein expression data to the tumor cellular composition, histological analyses were performed on the frozen tissue sections adjacent to sections used for mass spectrometry (see Methods). Parameters such as tumor content, surrounding lymph node area, necrosis and connective tissue percentages and lymphocytic infiltration were examined by a certified pathologist (Table 2).

The range of tumor content was 0 to 100%, and for most downstream analyses the inclusion criterion was to have at least 15% neoplasm of the tissue. The pieces for this analysis were removed from the surgically resected sample at macroscopic examination (grossing), thus, their content cannot represent the whole material excised from the patient. Nevertheless, assuming that histopathological properties in lymph node metastases display relatively low variation[17] we correlated the information with clinicopathological and proteomic data. The samples

Samples' properties:	mean	sd	min	max
tumor %	66	33	0	99
necrosis %	5	11	0	63
lymph node %	12	23	0	97
connective tissue %	17	26	0	100
Tumor properties		**n**	**%**	
tumor cell size	*<20 microns*	98	88	
	20–25 microns	2	2	
	>25 microns	1	1	
tumor cell shape	*epithelioid*	82	74	
	mixed epithelioid and spindle	17	15	
	spindle	2	2	
Tumor cell pigmentation	*0*	48	43	
	1	20	18	
	2	13	12	
	3	20	18	
Lymphocyte density	*0*	17	15	
	1	37	33	
	2	33	30	
	3	11	10	
Lymphocyte distribution	*0*	17	15	
	1	35	32	
	2	25	23	
	3	21	19	
Immunoscore, = sum of lymphocyte density and distribution	*0*	15	14	
	1	3	3	
	2	24	22	
	3	18	16	
	4	16	14	
	5	16	14	
	6	6	5	

Table 2. Tumor and tumor samples properties. Tumor cell pigmentation (0 = absent: no melanin pigment discernible even at high power magnification, 1 = slight: melanin pigmentation hardly visible at low power, at high power, melanocytes show a faint diffuse hue or a few scattered melanin pigment granules, 2 = moderate: pigmentation visible at low power, the cytoplasm is translucent and appears significantly lighter than the hematoxylin stained nuclei, 3 = high: pigmentation is easily visible at low power, the cytoplasmic pigmentation reaches an intensity approximating that of the nucleus). Lymphocyte distribution (0 = no lymphocytes within the tissue, 1 = lymphocytes present involving <25% of the tissue cross sectional area, 2 = lymphocytes present in 25 to 50% of the tissue, 3 = lymphocytes present in >50% of tissue). Lymphocyte density (0 = absent, 1 = mild, 2 = moderate, 3 = severe).

were mostly composed of epithelioid shaped melanocytes infiltrating the lymph nodes, displaying necrosis to various extent, the background was lymphocytic sheets of otherwise normally appearing lymph nodes, in most cases with connective tissue present (Fig. 1A,B, Table 2).

Proteomics data. Samples were analysed by high-resolution tandem mass spectrometry. Label-free LC-MS/MS analysis allowed the quantitation of 4963 proteins, and more than one third of them was quantified in more than 50% of samples (see Suppl. Fig. S1). Most analyses of protein expression data, e.g. correlation with tumor content/percentage and patient overall survival, were restricted to 1306 proteins, i.e. those quantified in at least 70% of the samples.

Relationship of protein expression to tumor content. In this relatively heterogeneous sample set, many proteins exhibited significant correlation to histopathological features. Two hundred and five proteins were significantly positively correlated to sample tumor cell content (using unadjusted p-value < 0.0001) and a smaller number, 29 proteins, were negatively correlated. As expected, the proteins correlated to tumor cell content usually showed inverse correlation to connective tissue content. In principle, correlation p-values should be adjusted for multiple testing using Benjamini-Hochberg (BH) approach. Approximately, the conservative raw p-value of 0.0001 used here corresponds to the value of 0.006 after the BH correction (Suppl. Table ST7).

Positive and negative correlation of protein expression to tumor cell content was connected to particular molecular and biological functions. A Panther[18] analysis of tumor cell-correlated proteins yielded molecular functions such as tRNA ligases and glycogen phosphorylases for the positively correlated set, while complement

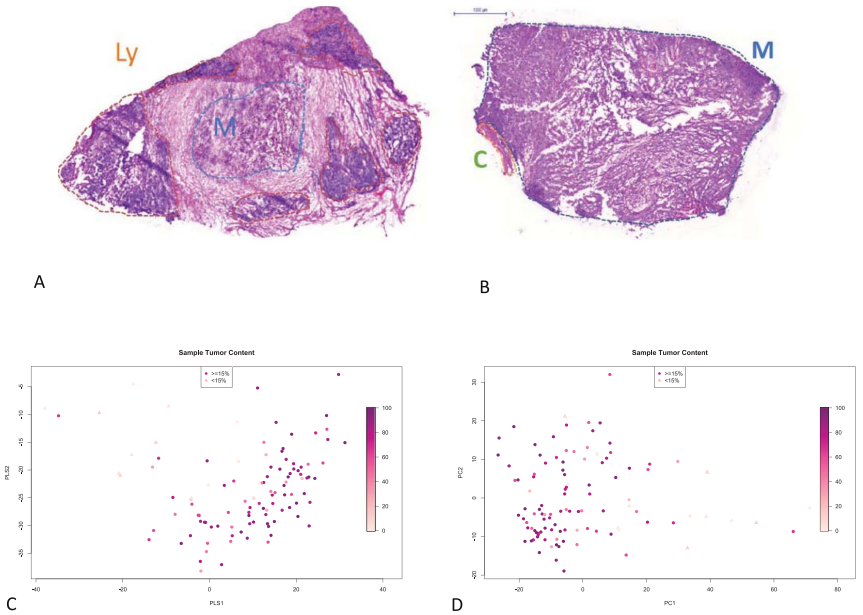


Figure 1. Variability of the tumor samples. (**A**,**B**) Representative histopathology images of the tumor samples. (**A**) Low tumor content sample. Ly – lymphatic cells, M – tumor. (**B**) High tumor content sample. C – connective tissue. (**C**,**D**) Unsupervised multidimensional analysis of the proteomics data. Colouring by tumor content (dark: high content). Samples with <15% tumor shown as triangles, others – as circles. (**C**) Partial Least Squares (PLS) analysis. (**D**) Principal Component Analysis (PCA).

activation, structural constituent of cytoskeleton and actin-binding characterized the proteins negatively correlated to tumor content. Similarly, an Ingenuity Pathway Analysis (IPA) performed for the tumor-correlated proteins provided relationship networks enriched in proteins related to transcription, translation, glycolysis, tRNA charging, ubiquitination, tubulins, and splicing (See Suppl. Fig. S2 and Suppl. Table ST1). Similar functional themes were found to be associated with tumor cell content in a smaller subset of the current cohort analysed previously[19], thus, confirming our earlier pilot findings. These functions are in line with well-known features of malignant tumors and connective tissues, and suggest that proteomics data could be used for tissue discrimination and quality assessment of the sample with respect to tumor content[20].

Unsupervised view of the data - PCA. A non-supervised multivariate analysis of proteome profile allows to explore the main components of variability between the melanoma samples. Here, a principal component analysis (PCA) of protein expression data did not show obvious separation with respect to clinical or histopathological parameters (e.g. BRAF mutation status, survival, see Suppl. Fig. S7A,B). The only exception was tumor cell content, where a clear trend was visible (see Fig. 1C,D) indicating that sample heterogeneity in terms of tumor cell content was a major source of variability in the proteomics data.

Relating proteomics data to survival. In order to relate protein expression in lymph node metastatic melanomas to patient survival, we attempted an unsupervised classification based on consensus clustering[21]. This approach, applied to the whole sample set (111 patients) produced clusters that did not differ significantly in survival (Suppl Fig. S3A,B). Thus, for subsequent analyses only the 96 samples with tumor content of at least 15% were considered (choosing higher thresholds did not improve survival prediction while obviously lowered the number of available samples). Here, we investigated the predictive power of the protein expression data from metastatic melanoma using two approaches. The unsupervised approach involved hierarchical consensus clustering. The supervised approach consisted of Partial Least Square (PLS) regression in combination with Cox Proportional Hazards modeling (PLS-Cox). Both approaches were able to produce patient clusters with significant differences in survival. Applying unsupervised clustering to the proteomic data produced three patient clusters which show distinct differences in survival (log-rank test p-value = 0.0028, see Fig. 2A).

The PLS-Cox model reduces the expression of the whole feature-set (~1300 proteins) to a single latent (inferred) variable, which explains the main part of the variability with respect of patient survival and which is then used in a Cox Proportional Hazards model. A high score on this latent variable is linked to a low hazard score, i.e. better prognosis. Furthermore, we used rank products to extract the features (proteins) which contribute most to the latent variable[22]. After cross-validation and FDR testing, we obtained 27 proteins which were strong contributors

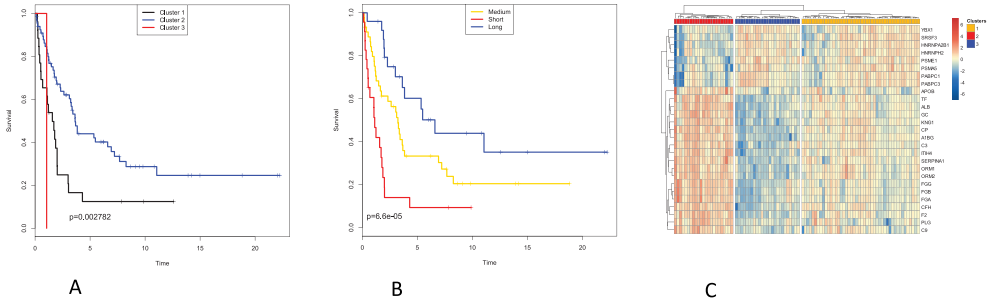


Figure 2. Proteomics data is related to patient survival. (**A**,**B**) 2A. Kaplan Meier plots for patient clusters obtained by (**A**) consensus clustering using 1306 proteins quantified in at least 70% of the samples (shown in Suppl. Fig. S3C) (**B**) consensus clustering using only the 27 survival-related proteins, with significant Cox scores (shown in Suppl. Fig. S3D). (**C**) Two-way hierarchical clustering of the 27 survival-related proteins and the patient samples. Red: high expression. Blue: low expression. Patient clusters coloured as in (**B**).

to the latent variable (see Suppl. Table ST2). Of these, 9 were positively correlated (thus high expression is linked to long survival) and 18 negatively correlated (overexpression of these is linked to short survival).

When applied to only the 27 proteins obtained from the PLS-Cox model, the same hierarchical clustering algorithm gave us three patient clusters, even more distinct in terms of survival (log-rank test p-value = 0.000066, see Fig. 2B). One of the clusters corresponded to poor survival and was characterized by downregulation of the 9 proteins positively correlated and by upregulation of the 18 proteins negatively correlated to survival. A second cluster had expression profiles opposite to those of the first one and corresponded to a more favourable survival. A third cluster corresponded to intermediate survival and an intermediate expression pattern (see Fig. 2C).

Analogous analyses were performed using peptide quantitation data. Here, unsupervised consensus and supervised PLS-Cox clustering also produced clusters significantly differing in survival, albeit with weaker effect.

In order to ascertain that the 15% tumor content cutoff was not too subjective, several other cutoffs were tested (0, 25, 50 and 75% tumor) and the PLS-Cox survival analysis was repeated for each. The 15, 25 and 50% cutoffs produced very similar results in terms of candidate survival biomarker sets (Suppl. Table ST6), albeit the 15% threshold provided the largest number of significant candidates (twenty seven). Also, the 15% cutoff provided the most significant statistical model while the 25% cutoff resulted in a model of similar significance (Suppl. Table ST9).

Further, the Cox survival analysis was performed using several histological features of the samples instead of protein expression data (see Suppl. Table ST8). While some such features (related to cytoplasm features) did show a weak relationship to survival (univariate Cox regression model p-values 0.003–0.03), protein expression clearly outperformed these features in terms of survival prediction. All univariate Cox models built for the 27 candidate proteins were significant and most had p-values below 0.003 (minimum $3*10^{-6}$, see Suppl. Table ST10). Of note, tumor content was not a significant survival predictor (see Suppl. Table ST8).

The PLS-Cox based supervised clustering built on protein expression was compared with two genomics-based sample classifications applied previously to the same tumor samples. The four-category classification of Jönsson *et al.* (high immune, normal, pigmentation and proliferative[23]) and TCGA classification (immune, keratin, MITF-low[16]) were not in perfect accordance with the three survival clusters elucidated herein (see Fig. 3 and Suppl. Fig. S4A,B). However, there were clear differences between the longer and shorter survival clusters in terms of composition of the genomics categories. Interestingly, the short survival cluster 2 had largest proportion of proliferative-type tumors (Jönsson's classification[23]) while the long survival cluster 3 had approx. 75% samples of the pigmentation type. In terms of TCGA classification[16], short survival cluster 2 had largest proportion of MITF-low tumors while the long and medium survival clusters 1 and 3 had largest proportion of immune-type tumors (Suppl. Fig. S4). The long survival clusters obtained by two approaches (unsupervised and supervised) using protein data agreed well - they were composed mostly of the same patient samples (90% agreement, i.e.: 90% of the samples from the supervised good prognosis cluster belonged also to the unsupervised good prognosis cluster). The same applies to the short survival clusters (78% agreement, see Suppl. Fig. S4C). The chi-squared test comparing the unsupervised and supervised patient sample clustering supports their consistency (p-value $< 10^{-5}$). Interestingly, the short survival cluster (supervised) had significantly higher necrosis content than other clusters (Kruskal-Wallis p-value $< 10^{-6}$, see Suppl. Fig. S6).

Although for the survival prediction model there was no independent proteomics validation cohort available, we performed a tentative validation of the candidate proteomic survival biomarkers found in our study by using a large transcriptomic dataset of melanoma lymph node metastases (TCGA, N = 336, see Materials and Methods). Several of the 27 candidate biomarkers could be validated in this independent cohort, including those positively related to survival (high expression in long survival): PSME1, HNRNPA2B1 and SRSF3, and those negatively related to survival (high expression in short survival): APOB and ORM1 (see Suppl. Table ST5). This result is encouraging, bearing in mind the fact that on the average the corresponding signals for mRNA and protein expression correlate moderately.

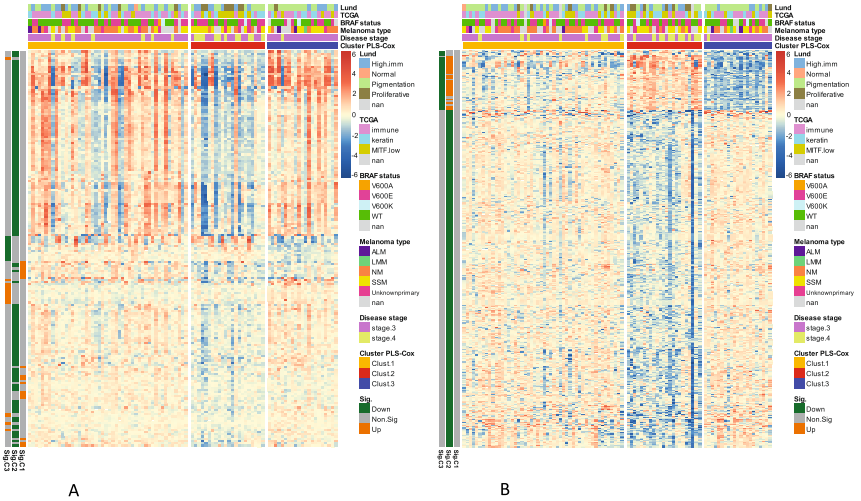


Figure 3. Proteins and mRNA exhibit differential expression among the survival-related patient clusters. Two-way hierarchical clustering of the transcripts (**A**) and proteins (**B**) differentially expressed between the survival-related patient clusters as per SAM analysis. Only highly significant transcripts and proteins shown (q value below 0.0005). Red: high expression. Blue: low expression. Patient clusters coloured as in Fig. 2B. Additional annotations (coloured bars at top) indicate selected patient/sample parameters: Lund genomics cluster[23], TCGA genomics cluster, BRAF status, Melanoma type, disease stage. Additional annotations (coloured bars on the left, orange or green) indicate that a given transcript or protein is significantly up- or down regulated for a given cluster.

Functional analysis of the survival-related clusters. The three clusters obtained by supervised PLS-Cox analysis[24] of proteomics data and significantly differing in survival were explored in order to understand the molecular differences. To this end, the current proteomic data and mRNA expression data obtained previously for the same melanoma samples[23], were subjected to SAM analysis (a technique conceptually similar to ANOVA[25]) to obtain genes and proteins differentially expressed between sample clusters. The analyses included more than 1300 proteins and more than 11000 genes. At significance level of FDR < 0.0005, 419 proteins and 177 genes were found to be differentially expressed between the three clusters (1368 proteins and 777 genes at more relaxed significance level of FDR < 0.05). The heatmaps in Fig. 3A,B show the genes/proteins with cluster-specific expression patterns. Within the three clusters, cluster 3 (long survival) clearly had underrepresentation of melanomas that were stage 4 while cluster 2 (poor survival) clearly had overrepresentation of stage 4 melanomas (see Fig. 3A,B).

The sets of proteins and genes significantly differing between the three survival-related patient clusters were rather different (for FDR < 0.0005, overlap between the 419 proteins and the 177 genes was only 8, while for FDR < 0.05, the gene/protein list overlap was 68, see Suppl. Table ST3). This clearly shows that proteomics and genomics analyses capture to some extent complementary aspects of melanoma biology. Using mRNA profiling data of the same patient cohort (the same tumor samples, but different sections) as previously published[23], one can correlate mRNA and protein expression signals. For these, a median correlation of 0.306 is obtained (Suppl. Fig. S8). This is generally in agreement with the previous studies, however, since mRNA and protein data were obtained from different tissue sections of the same samples, the actual correlation is probably slightly underestimated.

The differential expression analysis of genes and proteins provides tumor- and survival-related functions in short and long survival sample clusters. Although, the differentially expressed sets of genes and proteins were by large different, the biological functions related to the patient clusters were to a certain extent similar (see Suppl. Table ST4). For the short survival cluster, the significantly downregulated genes and proteins alike were enriched in functions such as antigen processing and presentation, TCR and interferon signalling. The three survival-related patient clusters did not differ in terms of mutation burden in an analysis of genes known to often harbor mutations in melanoma (See Suppl. Fig. S5).

Functional analysis of the 27 proteins obtained from the PLS-Cox model. Ingenuity Pathway Analysis (IPA) split most of the 27 proteins that were guiding the three survival clusters into two functional relationship networks. The first network was mostly extracellular and included proteins negatively correlated to survival (low expression in tumors from patients with good prognosis, i.e. long survival). The second network was a nuclear/cytoplasmic one, and included proteins positively correlated to survival (high in tumors from patients with long survival, Fig. 4A,B). A complementary IPA analysis was executed using an extended set of 160 top

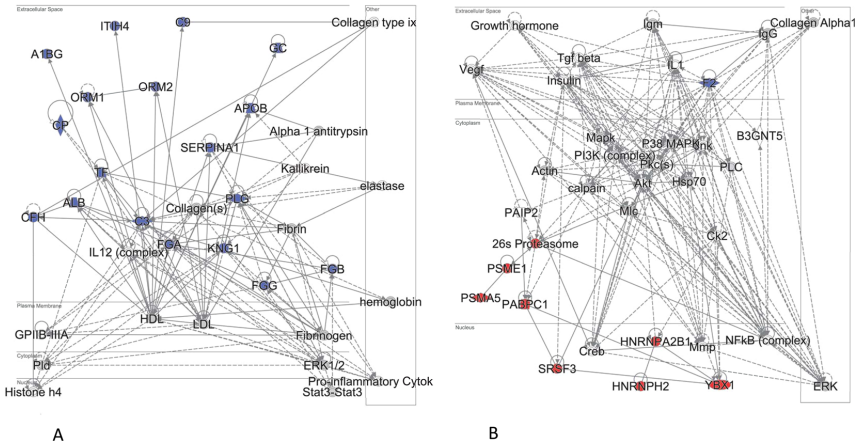


Figure 4. Pathway analysis for 27 survival-related proteins. Ingenuity Pathway Analysis (IPA) for the proteins identified by the PLS-Cox analysis as significantly related to survival (Cox score FDR < 0.1). Protein-protein relationship subnetworks shown that are enriched in the 27 query proteins. (**A**) First subnetwork, (**B**) Second subnetwork. Blue – proteins with expression negatively correlated to survival. Red – positively correlated to survival. Data were analyzed through the use of IPA (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis)[109].

proteins most strongly related to survival albeit not all strictly significant. Of these, 80 were negatively correlated to survival and 80 - positively correlated. Here, the proteins positively related to survival as per Cox analysis were enriched in functions such as RNA post-transcriptional modifications, protein synthesis and cell death. The proteins negatively correlated to survival were enriched in cell-to-cell signalling and cell movement proteins.

Proteins negatively correlated to survival (high expression in short survival). Interestingly, many of the 18 proteins showing negative significant correlation to survival are high-abundance plasma proteins. This may reflect the vascularisation aspect of melanoma metastases as well as immune component of tumor development. One might speculate that the lymph nodes are thought to be filters of the circulating lymph which contains enriched fractions of the proteins and lipids of the blood which may show in the results. Alternatively, the tumor cells might be "hiding" while metastasizing and covering themselves with platelets, thus exhibiting expression of platelet proteins (all but one of the 18 proteins are present in platelets[26]). Also, the negative correlation to survival of coagulation-related proteins (F2, PLG, FGB, FGG, FGA, KNG1) likely reflects the well-known relationship between cancer and thrombosis[27].

The role of the copper and iron transport protein ceruloplasmin (CP) in cancer has been reported[28] and it was found elevated in plasma of melanoma patients[29], hence a negative correlation to survival could be expected. Human serum transferrin is a glycoprotein which is involved in iron transport. Since neoplastic cells have a high requirement of iron related to their rate of proliferation[30], it seems logical that we found high level of transferrin in the poor survival cluster.

More than 5-fold higher level of the protease inhibitor ITIH4 was reported previously in sera from patients with hepatocellular carcinoma with good prognosis compared to patients with poor prognosis[31]. The ITIH4 gene expression was lost in multiple human solid tumors[32]. However, in a rat model for colon cancer, ITIH4 was one of four proteins that was upregulated in sera compared to wild-type rats[33]. The serine protease inhibitor, SERPINA1, has been reported to modulate invasive and metastatic capacity in lung cancer, gastric cancer, and CRC[34–36]. Elevated expression of SERPINA1 was previously correlated with advanced stage, lymph node metastasis, and poor prognosis[37], which is in accordance to our current findings.

Complement factor H (CFH) is the main actor inhibiting complement responses by regulating the Complement Alternative Pathway[38]. CFH binds to "*self marker*" structures on matrix and the cell surface, e.g. GAG chains and sialylated sugars, and prevents further activation/attack by the complement system[39]. CFH may have dual roles in cancer, either promoting tumor progression (by immune evasion) or supporting tumor suppression (by inducing an anti-inflammatory microenvironment[38]). Tumor cells may "hijack" the complement system by expressing, releasing or recruiting CFH and other complement inhibitors in high amounts, thus evading complement attack. This has been described in ovarian, lung, glioma and colon cancer cells[40–43]. In addition, CFH has been suggested as a marker in lung adenocarcinoma[44], where shorter survival time of patients with adenocarcinoma was associated with increased CFH staining. Data from the TCGA cohort suggest that increased mRNA levels of CFH are significantly related to poor prognosis in kidney carcinoma[45] and urine levels of a closely related protein CFHR1 were negatively related to bladder cancer survival[46]. To our knowledge, negative relation of CFH protein to survival in metastatic melanoma tissue has not been reported.

A role of Vitamin D signaling and the activity of Vitamin D binding protein GC (VDBP) in melanoma is known[47,48] and vitamin D deficiency is associated with worse prognosis[49]. VDBP is responsible for transporting Vitamin D analogues in plasma. While SNPs in VDBP were reported not to influence melanoma survival in a case-control study[50], meta-analysis of VDBP polymorphisms suggested that VDBP rs12512631 TT genotype was linked to a poorer survival compared with those with TC and CC genotypes[47]. The involvement of VDBP in cancer has a complex mechanism: on one hand, VDBP enhances epithelial ovarian cancer progression[51], on the other hand, higher circulating VDBP levels were observed in healthier melanoma patients[52]. Also, a meta-analysis including 28 studies of 12 different cancers, and analyzing VDBP protein levels vs. cancer risk found trends toward significance (lower risk related to high expression), suggesting a role of VDBP in cancer etiology[53]. The negative relation of VDBP expression to melanoma survival observed by us is not in agreement with some previous reports, whereas promising results were obtained by using VDBP in cancer immunotherapy[54,55]. However, these results cannot be compared directly with ours since serum levels of VDBP need not be correlated to levels in tumor tissue.

APG1 and 2 (Orosomucoid 1 and 2) are heavily glycosylated acute phase reactants, mainly expressed in the liver but also extrahepatically[56,57] and increased in the circulation during acute inflammation as well as in several cancers including melanoma[58–60]. APG1 seems to be the primary acute phase responder while the proportion of APG1 to APG2 changes significantly in cancer[59]. The APGs display a multitude of biological activities such as acute-phase reactants, modulating immunity, and maintaining the barrier function of capillaries[56,57]. In addition, APGs are involved in binding synthetic drugs which has been described in cancer patients[61–63]. Aberrant glucosylation of the APGs is related to pathophysiological situations including cancer[64]. Overall, the negative relation to melanoma survival of APGs detected in metastatic melanoma tissue in the current study would be in agreement with previous literature describing circulating levels in cancer patients.

A recent study[65] related serum albumin levels to melanoma stage in a large patient cohort showing a significant reduction in circulating levels in stage 4 and in older patients. Albumin is a negative acute phase protein, e.g. levels are reduced during inflammation. The reduced levels in cancer and several other illnesses may be due to decreased synthesis, increased catabolism and other mechanisms[66,67]. In the current study, serum albumin level in melanoma tissue is negatively related to survival (high in patients with poor survival) which appears not in accordance with most other studies. However, most studies look at circulating levels and not metastatic tumor tissue.

Apolipoprotein B-100 (APOB) is a receptor for cholesterol which has been shown to increase melanogenesis[68] and targeting cholesterol transport in melanoma CTCs was shown to retard metastasis development[69]. This may be in line with current results of increased APOB expression in poor survival.

Alpha-1B-glycoprotein is a secreted glycoprotein with some similarity to the immunoglobulin family and basically very few known functions[70]. Interestingly, it has been described in proteomic studies of several cancer types like breast cancer[71], oral squamous carcinoma[72], in the serum of non-small cell lung cancer[73], and in pancreatic ductal adenocarcinoma[74]. Here we describe for the first time a negative correlation of alpha-1B-glycoprotein tissue expression to melanoma survival.

Proteins positively correlated to survival (high expression in longer survival). The splicing factor SRSF3 has been reported as an oncogenic factor in several types of cancer[75–79]. However, in colorectal cancer, loss of SRSF3 was significantly associated with poor survival and shorter disease-free survival in early cancer stages[80]. It was also shown that loss of SRSF3 was necessary for metastatic cells to colonize the liver microenvironment in mice[80]. Loss of SRSF3 has also been shown to predispose to hepatocellular carcinoma[81] and myeloid leukemia[82]. In this study, higher expression of SRSF3 was also found in the better prognosis cluster.

The transcription factor YBX1 is positively associated with a proliferative cellular state and might therefore be reported to be overexpressed in a variety of human cancers[83–86]. However, the YBX1 expression seems to be tightly regulated by a feedback mechanism ensuring optimal proliferation and survival of melanoma cells. The levels of YBX1 are also critical in melanoma cells for proliferation. High levels inhibit cell cycle progression and low levels induce apoptosis[87]. The YBX1 has been reported to correlate with bad prognosis in liver cancer[88,89] while here YBX1 is upregulated in melanoma patients with good prognosis.

Among the proteins positively correlated to survival, there are two proteasome related proteins PSMA5 and PSME1. The role of immunoproteasome in cancer is known[90], however high expression in better prognosis patients is not an obvious result. In a recent meta-analysis, PSMAs were generally found to be upregulated in cancers, including melanoma. Expression of some members of the PSMA family correlated with poor prognosis[91], however no melanoma prognosis data was available for the PSMA5 gene/protein that we find correlated with better prognosis. The Proteasome activator PSME1 (PA28alpha) that has been reported to regulate presentation of T lymphocyte epitopes on melanoma cells[92] is found here to be upregulated in good prognosis melanoma patients, similarly to a previous proteomics study[15]. Interestingly, quite to the opposite, in oral squamous cell carcinoma PSME1 expression has been reported to be related to poor prognosis[93].

The Poly A binding proteins PABPC1 and PABPC3 function in post-transcriptional control of mRNA and regulate cell proliferation[94]. PABPC1 expression was previously reported positively correlated to survival in esophageal cancer[95], but this protein has also been found to be oncogenic in gastric carcinoma[96].

The splicing factor HNRNPA2B1 has been reported as a candidate biomarker in lung cancer and regulator of epithelial-mesenchymal transition in pancreatic cancer (PDAC)[97–99]. Another splicing factor, HNRNPH2, was shown to drive anticancer drug resistance[100] and to drive hepatocellular carcinoma development[101]. Hence, higher expression in good prognosis of these two factors is an unexpected result.

Conclusion

We present a comprehensive proteomic, histopathological and genomic evaluation of malignant melanoma lymph node metastases. Our study is unique in applying in-depth histopathological characterisation to individual tumor samples. This, combined with detailed clinical information, allows elucidation of an efficient set of proteomic prognostic biomarkers. Since many of these candidate biomarkers are known to be relatively common plasma proteins, they present a possible opportunity for development of prognostic blood-based biomarker panel. This work builds on our own exploratory studies[19,102] as well as work by other groups[15,103] but differs from the previous work also by a much larger study cohort. By analysing the protein data alongside the genomic data obtained of the same tumor tissue, we highlight the complementarity of proteomic and transcriptomic molecular images of melanoma.

The fact that some of the prognostic proteins have not been reported in melanoma context before, and the fact that some exhibit unexpected relationship to survival, only exemplifies the complexity of melanoma progression mechanisms.

Materials and Methods

Reagents and solutions. All chemical reagents were purchased from Sigma Aldrich (St. Louis, MO) unless otherwise specified. Water and organic solvents were of LC–MS quality and supplied by Merck (Darmstadt, Germany). All solutions were degassed by sonication before use.

Tissue samples and sample preparation. 111 lymph node metastasis samples from patients with malignant melanoma (Stage 3 and 4), archived in the local malignant melanoma biobank were obtained from Skåne University Hospital, Sweden. Each sample was marked as 'MM' followed by identification number. Ethical approval was granted by Central Ethical Review board at Lund University; approval number: DNR 191/2007, 101/2013. All patients within the study provided a written informed consent. All experiments were performed in accordance with relevant guidelines and regulations. The malignant melanoma biobank "Tissue bank for research on tumour diseases" (BD20)" is located at Barngatan 2B, 221 85 Lund, Sweden. The samples were originally snap frozen immediately after surgery. Frozen tissue samples from BD20 were sectioned on a cryostat into 10 µm thick slices (approximately 6 × 6 mm), placed into a 96 well plate and stored at −80 °C until further use. From each tissue, 15 to 20 slices were withdrawn for sample preparation. Patient characteristics are summarised in Table 1. Clinical and histopathological parameters were retrieved from patients' clinical records, pathology reports and the Swedish National Population Registry. Survival was defined as time (days) from lymph node excision to patient's death or censoring date.

Histopathological evaluation. Frozen sections of all lymph node metastases stained with HE were evaluated by a certified pathologist. Serial sections were taken of each tumor, and at least seven slices per sample were examined. The tissue was assessed for its content regarding tumor, normal lymph node, necrosis, and background of any further component (e.g. fat or connective tissue). As previously described[16,19], the tumor was then evaluated for its histological characteristics containing epithelioid or spindle or mixed architecture, the tumor cell average size (scale 1–3) and pigmentation (scale 1–3). The tumor infiltrating lymphocytes were also assessed for their distribution (scale 1–3) and intensity (scale 1–3) in the tumor - only those which directly infiltrated the metastases were taken into account. The sum of distribution and density was then summarized in a 0–6 score considered as immunoscore.

cDNA synthesis and BRAF DNA sequencing. Two cell lines, SK-MEL-2 and SK-MEL-28 (ATCC®, Manassas, USA), were used as reference BRAF wild type and V600E respectively. Total RNA was extracted from the cell lines or frozen tissues from the malignant melanoma patients using RNeasy mini kit (Qiagen, Venlo, The Netherlands). The extracted RNA were reverse transcribed to cDNA by using Superscript III First Strand Synthesis System kit (ThermoFisher, Waltham, MA) according to the manufacturer's instructions. The cDNA was amplified with a set of primers that produced a PCR product including BRAF mutation at the position V600; 5′-(AGCCTTACAGAAATCTCCAGGACC)-3′ and 5′-(TTGGGGAAAGAGTGGTCTCTCATC)-3′. The PCR conditions were 95 °C for 5 min, followed by 36 cycles of 95 °C for 30 sec, 62 °C for 30 sec, and 72 °C for 2.5 min with a final incubation of 72 °C for 7 min. A portion of the PCR product was amplified a second time using the same condition as the first PCR, and the amplification was 24 cycles, instead of 36 cycles. The PCR products were run on a 1% agarose gel, and DNA was extracted from the gel using a QIAquick Gel Extraction kit (Qiagen) according to the manufacturer's instruction. The purified PCR products were sequenced using a primer 5′-(TTCCACAAAGCCACAACTGG)-3′ by Eurofins Genomics (Ebersberg, Germany).

Mutation data. Mutational information for selected 1697 cancer-associated genes were obtained by targeted deep sequencing of the patient tumor samples with matched blood, as described previously[23,104]. Visualization of mutational information was obtained using the oncoprinter function from R package ComplexHeatmap[105].

Tissue lysis and protein extraction. Frozen tissue slices were lysed with 6 M urea in 50 mM ammonium bicarbonate buffer (AmBic) for 30 min on ice bath. Samples were additionally vortexed for 10 min in order to promote protein extraction. After incubation with urea the lysate was sonicated for 5 min and centrifuged at 10 000 g at room temperature for 10 minutes. Supernatant was transferred into a new tube and the pellet was discarded. Protein concentration was measured using a bicinchoninic acid protein assay according to the manufacturer's instructions (Micro BCA kit, Pierce/Thermo Scientific, Rockford, IL). The samples were spiked with 0.1 mg of internal standard – chicken lysozyme (CL, Swiss-Prot accession no. P00698).

In-solution digestion with trypsin. A fixed amount (80 μg) of protein were reduced with 10 mM DTT for 1 h at 37 °C, then it was alkylated using 50 mM iodoacetamide for 30 min and kept in dark at room temperature. Urea was removed from the samples using Amicon Ultra centrifugal filters (0.5 mL, 10 kDa, Millipore, Ireland) according to the manufacturer's instructions. Briefly, the protein samples were mixed with 200 μL of 50 mM AmBic, then centrifuged at 14 000 g at room temperature for 20 minutes and the eluates were discarded. These steps were repeated two more times. The samples were transferred to an Eppendorf tube and digested with sequencing grade trypsin (Promega, Madison, WI) in a ratio 1:100 w/w (trypsin:protein) overnight at 37 °C. The digestion was stopped by adding formic acid till 1% as final concentration. The samples were dried using a centrifugal evaporator and resuspended in 80 μL of 0.1% formic acid and centrifuged for 5 min at 10 000 g. The supernatants were stored at −80 °C until further use. Prior to injection onto LC–MS/MS, 20 μL of samples were mixed with 20 μL of peptide retention time calibration mixture (PRTC, Pierce/Thermo Scientific, Rockford, IL, 20 fmoL/mL).

LC-MS/MS Analysis of the tumor lysate digests. Online chromatography was performed with a Thermo Easy nLC 1000 system (Thermo Fisher Scientific) coupled online to a Q-Exactive Plus mass spectrometer (Thermo Scientific, San José, CA). The peptides were first loaded onto a trap column (Acclaim1 PepMap 100 pre-column, 75 μm, 2 cm, C18, 3 mm, 100 Å, Thermo Scientific, San José, CA) and then separated on an analytical column (EASY-Spray column, 25 cm, 75 μm ID, PepMap RSLC C18, 2 mm, 100 Å, Thermo Scientific, San José, CA). Flow rate of 300 nL/min and a column temperature of 35 °C were utilised. A gradient was applied, using solvent A (0.1% formic acid) and solvent B (0.1% formic acid in acetonitrile). The gradient went from 5% to 40% B in the first 120 min, followed by raise to 90% B in the next 5 min, which was maintained for 10 min. To avoid carryover, each sample analysis was followed by a blank injection (water containing 0.1% formic acid). Mass spectrometry data were measured using a data-dependent top-15 method. Full MS scans were acquired over m/z 350–1800 range with resolution of 70 000 (at m/z 200), target AGC value of $1{\cdot}10^6$ and maximum injection time of 100 ms. Selected ions were fragmented in the HCD collision cell with normalised collision energy of 30%, and tandem mass spectra were acquired in the Orbitrap mass analyzer with resolution of 17 500 (at m/z 200), target AGC value of $1{\cdot}10^6$ and maximum injection time of 120 ms. The ion selection threshold was set to $4.2{\cdot}10^4$ and dynamic exclusion was 20 s.

Proteomics data analysis. The LC-MS/MS raw files were analyzed with Proteome Discoverer 2.1 (Thermo Scientific, San José, CA) for protein identification and quantitation. The files were searched against the UniProtKB human database (released May 2016) excluding isoforms. The search was performed with the following parameters: carbamidomethylation as static modification, oxidation of methionine as dynamic modification, 20 ppm precursor tolerance and 0.02 Da fragment tolerance. Up to two missed cleavages for tryptic peptides was allowed. Filters: high confidence at peptides and protein levels were applied (FDR 0.01). Protein intensities were $\log_2$ transformed, followed by sample median subtraction using R (version 2.41–3).

Multivariate survival analysis. We have used unsupervised and supervised approaches to linking proteomic data to survival. The unsupervised method was performed using consensus clustering in R with ConsensusClusterPlus library (version 1.42.0). The supervised approach is based on PLS-Cox regression similar to that of Nguyen and Rocke[24]. The PLS-step of the model is used to reduce the high dimensionality of the proteomic data, while Cox regression was used on the first PLS component. We use a similar approach as Bair *et al.*[22] to assess the performance of this model. For cross-validation, the dataset is split into two subsets; the first is used to fit the model, the second to evaluate its performance. This process is repeated 100 times and the results of all iterations are averaged. Simultaneously, we extract the most important features, i.e. proteins, using rank products[24,106] of the PLS loadings. Correction for multiple testing with Benjamini-Hochberg approach results in 9 proteins which are significantly positively correlated to long survival and 18 which are significantly negatively correlated at adjusted significance level of 0.05. We performed this analysis both for the full sample set (N = 111) as well as for a subset (N = 94) wherein all samples contain at least 15% tumor. The supervised survival analysis was performed using peptide data as well, but the identified sample clusters showed less significant relationships to survival.

For the Kaplan-Meier survival analysis, the *survdiff* function in R (version 2.41–3) was used, which implements the log-rank test.

Differentially expressed genes and proteins for the survival-related patient clusters were elucidated using the SAM method[25], applying multiple testing correction as described[107]. Gene expression data for the patient samples analysed in the current study were obtained in a previous study using the same sample set but different tissue sections[23].

By using the pheatmap library in R, two clustered heatmaps were built for the differentially expressed proteins and genes obtained from SAM analysis (FDR < 0.0005). Melanoma type, disease stage, BRAF status, TCGA classification and four-category classification of Jönsson *et al.*[23] were used as annotation terms. Comparison of clinical and histopathological parameters between the sample clusters was performed by chi-squared test (categorical variables) and by Kruskal-Wallis test (quantitative nonparametric variables). Differences were considered significant when p-value < 0.05 (without multiple testing adjustment).

The transcriptomic dataset of melanoma lymph node metastases from the TCGA database[16] was used for validation of the candidate proteomic survival biomarkers found in our study. The SurvExpress tool[108] was applied to assess if query transcripts were promising predictors of survival.

Protein set functional analysis. Functional analysis of the protein sets identified with PLS-Cox regression and correlation analysis with tumor content was conducted using IPA, Ingenuity Pathway Analysis (Qiagen,

Redwood City, CA, USA)[109], in particular by generating networks of protein-protein functional relationships. As background, the set of proteins detected in >70% of the samples was used.

Functional analysis of lists of proteins mentioned above was also performed using the Panther server[110]. Overrepresentation of specific functional annotations within the protein lists was determined by Fisher's exact test, the background protein set consisted of all proteins detected. Gene Ontology annotations, SwissProt keywords, and Reactome and KEGG pathways were used as annotation terms for the enrichment analysis.

Data Availability

The proteomics dataset associated with the current article is publicly available in ProteomeXchange (http://www.proteomexchange.org/), dataset identifier: PXD009630.

References

1. Glazer, A. M., Winkelmann, R. R., Farberg, A. S. & Rigel, D. S. Analysis of Trends in US Melanoma Incidence and Mortality. *JAMA dermatology*, https://doi.org/10.1001/jamadermatol.2016.4512 (2016).
2. Gershenwald, J. E. *et al.* Melanoma staging: Evidence-based changes in the American Joint Committee on Cancereighth edition cancer staging manual. *CA Cancer J Clin* **67**, 472–492, https://doi.org/10.3322/caac.21409 (2017).
3. Alegre, E., Sammamed, M., Fernandez-Landazuri, S., Zubiri, L. & Gonzalez, A. Circulating biomarkers in malignant melanoma. *Adv Clin Chem* **69**, 47–89, https://doi.org/10.1016/bs.acc.2014.12.002 (2015).
4. Merrill, R. M. & Bateman, S. Conditional Melanoma Cancer Survival in the United States. *Cancers* **8**, https://doi.org/10.3390/cancers8020020 (2016).
5. Thiam, A., Zhao, Z., Quinn, C. & Barber, B. Years of life lost due to metastatic melanoma in 12 countries. *Journal of medical economics* **19**, 259–264, https://doi.org/10.3111/13696998.2015.1115764 (2016).
6. Ascierto, P. A. *et al.* The role of BRAF V600 mutation in melanoma. *Journal of translational medicine* **10**, 85, https://doi.org/10.1186/1479-5876-10-85 (2012).
7. Bollag, G. *et al.* Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. *Nature* **467**, 596–599, https://doi.org/10.1038/nature09454 (2010).
8. Chakraborty, R., Wieland, C. N. & Comfere, N. I. Molecular targeted therapies in metastatic melanoma. *Pharmacogenomics and personalized medicine* **6**, 49–56, https://doi.org/10.2147/pgpm.s44800 (2013).
9. Rizos, H. *et al.* BRAF inhibitor resistance mechanisms in metastatic melanoma: spectrum and clinical impact. *Clin Cancer Res* **20**, 1965–1977, https://doi.org/10.1158/1078-0432.CCR-13-3122 (2014).
10. Chapman, P. B. *et al.* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* **364**, 2507–2516, https://doi.org/10.1056/NEJMoa1103782 (2011).
11. Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med* **363**, 711–723, https://doi.org/10.1056/NEJMoa1003466 (2010).
12. Hua, L., Zheng, W. Y., Xia, H. & Zhou, P. Detecting the potential cancer association or metastasis by multi-omics data analysis. *Genetics and molecular research: GMR* **15**, https://doi.org/10.4238/gmr.15038987 (2016).
13. Hayward, N. K. *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature*, https://doi.org/10.1038/nature22071 (2017).
14. Rodriguez-Cerdeira, C., Molares-Vila, A., Carnero-Gregorio, M. & Corbalan-Rivas, A. Recent advances in melanoma research via "omics" platforms. *J Proteomics*, https://doi.org/10.1016/j.jprot.2017.11.005 (2017).
15. Mactier, S. *et al.* Protein signatures correspond to survival outcomes of AJCC stage III melanoma patients. *Pigment Cell Melanoma Res* **27**, 1106–1116, https://doi.org/10.1111/pcmr.12290 (2014).
16. Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681–1696, https://doi.org/10.1016/j.cell.2015.05.044 (2015).
17. Murray, C. A., Leong, W. L., McCready, D. R. & Ghazarian, D. M. Histopathological patterns of melanoma metastases in sentinel lymph nodes. *Journal of clinical pathology* **57**, 64–67 (2004).
18. Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* **45**, D183–d189, https://doi.org/10.1093/nar/gkw1138 (2017).
19. Welinder, C. *et al.* Correlation of histopathologic characteristics to protein expression and function in malignant melanoma. *PLoS One* **12**, e0176167, https://doi.org/10.1371/journal.pone.0176167 (2017).
20. Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* **47**, 1168–1178, https://doi.org/10.1038/ng.3398 (2015).
21. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573, https://doi.org/10.1093/bioinformatics/btq170 (2010).
22. Bair, E. & Tibshirani, R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* **2**, E108, https://doi.org/10.1371/journal.pbio.0020108 (2004).
23. Cirenajwis, H. *et al.* Molecular stratification of metastatic melanoma using gene expression profiling: Prediction of survival outcome and benefit from molecular targeted therapy. *Oncotarget* **6**, 12297–12309, https://doi.org/10.18632/oncotarget.3655 (2015).
24. Nguyen, D. V. & Rocke, D. M. Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics* **18**, 1625–1632 (2002).
25. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* **98**, 5116–5121, https://doi.org/10.1073/pnas.091062498 (2001).
26. Boyanova, D., Nilla, S., Birschmann, I., Dandekar, T. & Dittrich, M. PlateletWeb: a systems biologic analysis of signaling networks in human platelets. *Blood* **119**, e22–34, https://doi.org/10.1182/blood-2011-10-387308 (2012).
27. Falanga, A., Russo, L., Milesi, V. & Vignoli, A. Mechanisms and risk factors of thrombosis in cancer. *Critical reviews in oncology/hematology* **118**, 79–83, https://doi.org/10.1016/j.critrevonc.2017.08.003 (2017).
28. Babich, P. S. *et al.* Non-hepatic tumors change the activity of genes encoding copper trafficking proteins in the liver. *Cancer biology & therapy* **14**, 614–624, https://doi.org/10.4161/cbt.24594 (2013).
29. Ros-Bullon, M. R., Sanchez-Pedreno, P. & Martinez-Liarte, J. H. Serum ceruloplasmin in melanoma patients. *Anticancer research* **21**, 629–632 (2001).
30. Richardson, D. R. & Baker, E. The uptake of iron and transferrin by the human malignant melanoma cell. *Biochim Biophys Acta* **1053**, 1–12 (1990).
31. Lee, E. J. *et al.* Inter-Alpha Inhibitor H4 as a Potential Biomarker Predicting the Treatment Outcomes in Patients with Hepatocellular Carcinoma. *Cancer research and treatment: official journal of Korean Cancer Association*. https://doi.org/10.4143/crt.2016.550 (2017).
32. Hamm, A. *et al.* Frequent expression loss of Inter-alpha-trypsin inhibitor heavy chain (ITIH) genes in multiple human solid tumors: a systematic expression analysis. *BMC Cancer* **8**, 25, https://doi.org/10.1186/1471-2407-8-25 (2008).

33. Ivancic, M. M., Irving, A. A., Jonakin, K. G., Dove, W. F. & Sussman, M. R. The concentrations of EGFR, LRG1, ITIH4, and F5 in serum correlate with the number of colonic adenomas in ApcPirc/+ rats. *Cancer Prev Res (Phila)* **7**, 1160–1169, https://doi.org/10.1158/1940-6207.capr-14-0056 (2014).
34. Higashiyama, M., Doi, O., Kodama, K., Yokouchi, H. & Tateishi, R. An evaluation of the prognostic significance of alpha-1-antitrypsin expression in adenocarcinomas of the lung: an immunohistochemical analysis. *Br J Cancer* **65**, 300–302 (1992).
35. Karashima, S., Kataoka, H., Itoh, H., Maruyama, R. & Koono, M. Prognostic significance of alpha-1-antitrypsin in early stage of colorectal carcinomas. *International journal of cancer* **45**, 244–250 (1990).
36. Tahara, E., Ito, H., Taniyama, K., Yokozaki, H. & Hata, J. Alpha 1-antitrypsin, alpha 1-antichymotrypsin, and alpha 2-macroglobulin in human gastric carcinomas: a retrospective immunohistochemical study. *Human pathology* **15**, 957–964 (1984).
37. Kwon, C. H. *et al*. Snail and serpinA1 promote tumor progression and predict prognosis in colorectal cancer. *Oncotarget* **6**, 20312–20326, https://doi.org/10.18632/oncotarget.3964 (2015).
38. Parente, R., Clark, S. J., Inforzato, A. & Day, A. J. Complement factor H in host defense and immune evasion. *Cell Mol Life Sci* **74**, 1605–1624, https://doi.org/10.1007/s00018-016-2418-4 (2017).
39. Blaum, B. S. *et al*. Structural basis for sialic acid-mediated self-recognition by complement factor H. *Nat Chem Biol* **11**, 77–82, https://doi.org/10.1038/nchembio.1696 (2015).
40. Junnikkala, S. *et al*. Secretion of soluble complement inhibitors factor H and factor H-like protein (FHL-1) by ovarian tumour cells. *Br J Cancer* **87**, 1119–1127, https://doi.org/10.1038/sj.bjc.6600614 (2002).
41. Wilczek, E. *et al*. The possible role of factor H in colon cancer resistance to complement attack. *International journal of cancer* **122**, 2030–2037, https://doi.org/10.1002/ijc.23238 (2008).
42. Ajona, D. *et al*. Expression of complement factor H by lung cancer cells: effects on the activation of the alternative pathway of complement. *Cancer Res* **64**, 6310–6318, https://doi.org/10.1158/0008-5472.can-03-2328 (2004).
43. Junnikkala, S. *et al*. Exceptional resistance of human H2 glioblastoma cells to complement-mediated killing by expression and utilization of factor H and factor H-like protein 1. *J Immunol* **164**, 6075–6081 (2000).
44. Cui, T. *et al*. Human complement factor H is a novel diagnostic marker for lung adenocarcinoma. *International journal of oncology* **39**, 161–168, https://doi.org/10.3892/ijo.2011.1010 (2011).
45. Uhlen, M. *et al*. A pathology atlas of the human cancer transcriptome. *Science* **357**, https://doi.org/10.1126/science.aan2507 (2017).
46. Raitanen, M. P. *et al*. Prognostic utility of human complement factor H related protein test (the BTA stat Test). *Br J Cancer* **85**, 552–556, https://doi.org/10.1054/bjoc.2001.1938 (2001).
47. Yin, J. *et al*. Genetic variants in the vitamin D pathway genes VDBP and RXRA modulate cutaneous melanoma disease-specific survival. *Pigment Cell Melanoma Res* **29**, 176–185, https://doi.org/10.1111/pcmr.12437 (2016).
48. Slominski, A. T. *et al*. Vitamin D signaling and melanoma: role of vitamin D and its receptors in melanoma progression and management. *Lab Invest* **97**, 706–724, https://doi.org/10.1038/labinvest.2017.3 (2017).
49. Timerman, D. *et al*. Vitamin D deficiency is associated with a worse prognosis in metastatic melanoma. *Oncotarget* **8**, 6873–6882, https://doi.org/10.18632/oncotarget.14316 (2017).
50. Schafer, A. *et al*. No association of vitamin D metabolism-related polymorphisms and melanoma risk as well as melanoma prognosis: a case-control study. *Archives of dermatological research* **304**, 353–361, https://doi.org/10.1007/s00403-012-1243-3 (2012).
51. Huang, Y. F. *et al*. Vitamin D-Binding Protein Enhances Epithelial Ovarian Cancer Progression by Regulating the Insulin-like Growth Factor-1/Akt Pathway and Vitamin D Receptor Transcription. *Clin Cancer Res*, https://doi.org/10.1158/1078-0432.ccr-17-2943 (2018).
52. Navarrete-Dechent, C. *et al*. Circulating vitamin D-binding protein and free 25-hydroxyvitamin D concentrations in patients with melanoma: A case-control study. *J Am Acad Dermatol* **77**, 575–577, https://doi.org/10.1016/j.jaad.2017.03.035 (2017).
53. Tagliabue, E., Raimondi, S. & Gandini, S. Meta-analysis of vitamin D-binding protein and cancer risk. *Cancer Epidemiol Biomarkers Prev* **24**, 1758–1765, https://doi.org/10.1158/1055-9965.epi-15-0262 (2015).
54. Thyer, L. *et al*. GC protein-derived macrophage-activating factor decreases alpha-N-acetylgalactosaminidase levels in advanced cancer patients. *Oncoimmunology* **2**, e25769, https://doi.org/10.4161/onci.25769 (2013).
55. Saburi, E., Saburi, A. & Ghanei, M. Promising role for Gc-MAF in cancer immunotherapy: from bench to bedside. *Caspian J Intern Med* **8**, 228–238, https://doi.org/10.22088/cjim.8.4.228 (2017).
56. Luo, Z., Lei, H., Sun, Y., Liu, X. & Su, D. F. Orosomucoid, an acute response protein with multiple modulating activities. *Journal of physiology and biochemistry* **71**, 329–340, https://doi.org/10.1007/s13105-015-0389-9 (2015).
57. Fournier, T., Medjoubi, N. N. & Porquet, D. Alpha-1-acid glycoprotein. *Biochim Biophys Acta* **1482**, 157–171 (2000).
58. Silver, H. K., Karim, K. A. & Salinas, F. A. Relationship of total serum sialic acid to sialylglycoprotein acute-phase reactants in malignant melanoma. *Br J Cancer* **41**, 745–750 (1980).
59. Budai, L. *et al*. Investigation of genetic variants of alpha-1 acid glycoprotein by ultra-performance liquid chromatography-mass spectrometry. *Analytical and bioanalytical chemistry* **393**, 991–998, https://doi.org/10.1007/s00216-008-2518-6 (2009).
60. Ayyub, A. *et al*. Glycosylated Alpha-1-acid glycoprotein 1 as a potential lung cancer serum biomarker. *Int J Biochem Cell Biol* **70**, 68–75, https://doi.org/10.1016/j.biocel.2015.11.006 (2016).
61. Zsila, F., Fitos, I., Bencze, G., Keri, G. & Orfi, L. Determination of human serum alpha1-acid glycoprotein and albumin binding of various marketed and preclinical kinase inhibitors. *Curr Med Chem* **16**, 1964–1977 (2009).
62. Ohbatake, Y. *et al*. Elevated alpha1-acid glycoprotein in gastric cancer patients inhibits the anticancer effects of paclitaxel, effects restored by co-administration of erythromycin. *Clinical and experimental medicine* **16**, 585–592, https://doi.org/10.1007/s10238-015-0387-9 (2016).
63. Kremer, J. M., Wilting, J. & Janssen, L. H. Drug binding to human alpha-1-acid glycoprotein in health and disease. *Pharmacol Rev* **40**, 1–47 (1988).
64. Hashimoto, S. *et al*. alpha1-acid glycoprotein fucosylation as a marker of carcinoma progression and prognosis. *Cancer* **101**, 2825–2836, https://doi.org/10.1002/cncr.20713 (2004).
65. Datta, M., Savage, P., Lovato, J. & Schwartz, G. G. Serum calcium, albumin and tumor stage in cutaneous malignant melanoma. *Future oncology (London, England)* **12**, 2205–2214, https://doi.org/10.2217/fon-2016-0046 (2016).
66. Margarson, M. P. & Soni, N. Serum albumin: touchstone or totem? *Anaesthesia* **53**, 789–803 (1998).
67. Kim, J. E. *et al*. Serum albumin level is a significant prognostic factor reflecting disease severity in symptomatic multiple myeloma. *Annals of hematology* **89**, 391–397, https://doi.org/10.1007/s00277-009-0841-4 (2010).
68. Schallreuter, K. U. *et al*. Cholesterol regulates melanogenesis in human epidermal melanocytes and melanoma cells. *Exp Dermatol* **18**, 680–688, https://doi.org/10.1111/j.1600-0625.2009.00850.x (2009).
69. Chen, Y. C. *et al*. Targeting cholesterol transport in circulating melanoma cells to inhibit metastasis. *Pigment Cell Melanoma Res* **30**, 541–552, https://doi.org/10.1111/pcmr.12614 (2017).
70. Ishioka, N., Takahashi, N. & Putnam, F. W. Amino acid sequence of human plasma alpha 1B-glycoprotein: homology to the immunoglobulin supergene family. *Proc Natl Acad Sci USA* **83**, 2363–2367 (1986).
71. Zeng, Z. *et al*. A proteomics platform combining depletion, multi-lectin affinity chromatography (M-LAC), and isoelectric focusing to study the breast cancer proteome. *Anal Chem* **83**, 4845–4854, https://doi.org/10.1021/ac2002802 (2011).
72. Jessie, K. *et al*. Aberrant proteins in the saliva of patients with oral squamous cell carcinoma. *Electrophoresis* **34**, 2495–2502, https://doi.org/10.1002/elps.201300107 (2013).

73. Liu, Y. *et al*. Integrative proteomics and tissue microarray profiling indicate the association between overexpressed serum proteins and non-small cell lung cancer. *PLoS One* **7**, e51748, https://doi.org/10.1371/journal.pone.0051748 (2012).
74. Tian, M. *et al*. Proteomic analysis identifies MMP-9, DJ-1 and A1BG as overexpressed proteins in pancreatic juice from pancreatic ductal adenocarcinoma patients. *BMC Cancer* **8**, 241, https://doi.org/10.1186/1471-2407-8-241 (2008).
75. Peiqi, L. *et al*. Expression of SRSF3 is Correlated with Carcinogenesis and Progression of Oral Squamous Cell Carcinoma. *International journal of medical sciences* **13**, 533–539, https://doi.org/10.7150/ijms.14871 (2016).
76. He, X. *et al*. Knockdown of splicing factor SRp20 causes apoptosis in ovarian cancer cells and its expression is associated with malignancy of epithelial ovarian cancer. *Oncogene* **30**, 356–365, https://doi.org/10.1038/onc.2010.426 (2011).
77. Jia, R., Li, C., McCoy, J. P., Deng, C. X. & Zheng, Z. M. SRp20 is a proto-oncogene critical for cell proliferation and tumor induction and maintenance. *International journal of biological sciences* **6**, 806–826 (2010).
78. Lin, J. C. *et al*. RBM4-SRSF3-MAP4K4 splicing cascade modulates the metastatic signature of colorectal cancer cell. *Biochim Biophys Acta* **1865**, 259–272, https://doi.org/10.1016/j.bbamcr.2017.11.005 (2018).
79. Kim, H. R. *et al*. MicroRNA-1908-5p contributes to the oncogenic function of the splicing factor SRSF3. *Oncotarget* **8**, 8342–8355, https://doi.org/10.18632/oncotarget.14184 (2017).
80. Torres, S. *et al*. Proteomic Characterization of Transcription and Splicing Factors Associated with a Metastatic Phenotype in Colorectal Cancer. *J Proteome Res* **17**, 252–264, https://doi.org/10.1021/acs.jproteome.7b00548 (2018).
81. Sen, S., Langiewicz, M., Jumaa, H. & Webster, N. J. Deletion of serine/arginine-rich splicing factor 3 in hepatocytes predisposes to hepatocellular carcinoma in mice. *Hepatology* **61**, 171–183, https://doi.org/10.1002/hep.27380 (2015).
82. Liu, J. *et al*. Aberrant expression of splicing factors in newly diagnosed acute myeloid leukemia. *Onkologie* **35**, 335–340, https://doi.org/10.1159/000338941 (2012).
83. Evdokimova, V., Tognon, C., Ng, T. & Sorensen, P. H. Reduced proliferation and enhanced migration: two sides of the same coin? Molecular mechanisms of metastatic progression by YB-1. *Cell Cycle* **8**, 2901–2906, https://doi.org/10.4161/cc.8.18.9537 (2009).
84. Evdokimova, V. *et al*. Translational activation of snail1 and other developmentally regulated transcription factors by YB-1 promotes an epithelial-mesenchymal transition. *Cancer Cell* **15**, 402–415, https://doi.org/10.1016/j.ccr.2009.03.017 (2009).
85. Wu, J. *et al*. Disruption of the Y-box binding protein-1 results in suppression of the epidermal growth factor receptor and HER-2. *Cancer Res* **66**, 4872–4879, https://doi.org/10.1158/0008-5472.can-05-3561 (2006).
86. Kohno, K., Izumi, H., Uchiumi, T., Ashizuka, M. & Kuwano, M. The pleiotropic functions of the Y-box-binding protein, YB-1. *Bioessays* **25**, 691–698, https://doi.org/10.1002/bies.10300 (2003).
87. Sinnberg, T. *et al*. MAPK and PI3K/AKT mediated YB-1 activation promotes melanoma cell proliferation which is counteracted by an autoregulatory loop. *Exp Dermatol* **21**, 265–270, https://doi.org/10.1111/j.1600-0625.2012.01448.x (2012).
88. Xu, L., Li, H., Wu, L. & Huang, S. YBX1 promotes tumor growth by elevating glycolysis in human bladder cancer. *Oncotarget* **8**, 65946–65956, https://doi.org/10.18632/oncotarget.19583 (2017).
89. Zhou, L. L. *et al*. High YBX1 expression indicates poor prognosis and promotes cell migration and invasion in nasopharyngeal carcinoma. *Exp Cell Res* **361**, 126–134, https://doi.org/10.1016/j.yexcr.2017.10.009 (2017).
90. Joyce, S. Immunoproteasomes edit tumors, which then escapes immune recognition. *Eur J Immunol* **45**, 3241–3245, https://doi.org/10.1002/eji.201546100 (2015).
91. Li, Y. *et al*. The transcription levels and prognostic values of seven proteasome alpha subunits in human cancers. *Oncotarget* **8**, 4501–4519, https://doi.org/10.18632/oncotarget.13885 (2017).
92. Sun, Y. *et al*. Expression of the proteasome activator PA28 rescues the presentation of a cytotoxic T lymphocyte epitope on melanoma cells. *Cancer Res* **62**, 2875–2882 (2002).
93. Feng, X. *et al*. Overexpression of proteasomal activator PA28alpha serves as a prognostic factor in oral squamous cell carcinoma. *J Exp Clin Cancer Res* **35**, 35, https://doi.org/10.1186/s13046-016-0309-z (2016).
94. Stupfler, B., Birck, C., Seraphin, B. & Mauxion, F. BTG2 bridges PABPC1 RNA-binding domains and CAF1 deadenylase to control cell proliferation. *Nat Commun* **7**, 10811, https://doi.org/10.1038/ncomms10811 (2016).
95. Takashima, N. *et al*. Expression and prognostic roles of PABPC1 in esophageal cancer: correlation with tumor progression and postoperative survival. *Oncology reports* **15**, 667–671 (2006).
96. Zhu, J., Ding, H., Wang, X. & Lu, Q. PABPC1 exerts carcinogenesis in gastric carcinoma by targeting miR-34c. *International journal of clinical and experimental pathology* **8**, 3794–3802 (2015).
97. Hu, Y. *et al*. Splicing factor hnRNPA2B1 contributes to tumorigenic potential of breast cancer cells through STAT3 and ERK1/2 signaling pathway. *Tumour Biol* **39**, 1010428317694318, https://doi.org/10.1177/1010428317694318 (2017).
98. Dai, S. *et al*. HNRNPA2B1 regulates the epithelial-mesenchymal transition in pancreatic cancer cells through the ERK/snail signalling pathway. *Cancer cell international* **17**, 12, https://doi.org/10.1186/s12935-016-0368-4 (2017).
99. Dai, L. *et al*. Identification of autoantibodies to ECH1 and HNRNPA2B1 as potential biomarkers in the early detection of lung cancer. *Oncoimmunology* **6**, e1310359, https://doi.org/10.1080/2162402x.2017.1310359 (2017).
100. Stark, M., Bram, E. E., Akerman, M., Mandel-Gutfreund, Y. & Assaraf, Y. G. Heterogeneous nuclear ribonucleoprotein H1/H2-dependent unsplicing of thymidine phosphorylase results in anticancer drug resistance. *J Biol Chem* **286**, 3741–3754, https://doi.org/10.1074/jbc.M110.163444 (2011).
101. Li, X. *et al*. A splicing switch from ketohexokinase-C to ketohexokinase-A drives hepatocellular carcinoma formation. *Nat Cell Biol* **18**, 561–571, https://doi.org/10.1038/ncb3338 (2016).
102. Welinder, C. *et al*. A protein deep sequencing evaluation of metastatic melanoma tissues. *PLoS One* **10**, e0123661, https://doi.org/10.1371/journal.pone.0123661 (2015).
103. Byrum, S. D. *et al*. Quantitative Proteomics Identifies Activation of Hallmark Pathways of Cancer in Patient Melanoma. *J Proteomics Bioinform* **6**, 43–50, https://doi.org/10.4172/jpb.1000260 (2013).
104. Harbst, K. *et al*. Molecular and genetic diversity in the metastatic process of melanoma. *The Journal of pathology* **233**, 39–50, https://doi.org/10.1002/path.4318 (2014).
105. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849, https://doi.org/10.1093/bioinformatics/btw313 (2016).
106. Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* **573**, 83–92, https://doi.org/10.1016/j.febslet.2004.07.055 (2004).
107. Storey, J. D. A direct approach to false discovery rates. *J. R. Statist. Soc. B* **64**, 479–498 (2002).
108. Aguirre-Gamboa, R. *et al*. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One* **8**, e74250, https://doi.org/10.1371/journal.pone.0074250 (2013).
109. Kramer, A., Green, J., Pollard, J. Jr. & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523–530, https://doi.org/10.1093/bioinformatics/btt703 (2014).
110. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nature protocols* **8**, 1551–1566, https://doi.org/10.1038/nprot.2013.092 (2013).

Acknowledgements

This study was supported by the Mrs Berta Kamprad Foundation, ThermoFisher Scientific, and Liconic with Biobank technology, and was also supported by grants from the National Research Foundation of Korea, funded by the Korean government (2015K1A1A2028365 and 2016K2A9A1A03904900) and Brain Korea 21 Plus Project, Republic of Korea, as well as the NIH/NCI International Cancer Proteogenome Consortium (MOU NIH LUND 2017-01), and the European Cancer Moonshot Lund Center scientific and outreach activities, through which data will be made public. G.J. was supported by the Swedish Cancer Society and the Swedish Research Council. A.M.S. was supported by the KNN121510 and NVKP_16-1-2016-0042 grants by the National Research, Development and Innovation Office of Hungary and Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Author Contributions

G.M.V. conceived and supervised the project. L.H.B., C.W., M.Y., M.R. performed the mass spectrometry experiments. A.M.S., Y.S. performed the histopathology work and analyses. K.M., Y.S. performed the BRAF genotyping. L.L., C.I., H.E., C.W., A.M.S. and E.W. collated and analysed the patient clinical data. J.E., I.P., S.M., P.H., P.B., G.J., K.P. performed the informatics and bioinformatics analyses. B.B., H.O., J.M., R.A. contributed to data analysis and manuscript writing. J.E., I.P., S.M., A.M.S. and K.P. prepared the figures. K.P., E.W., C.W., A.M.S., P.H., G.J., J.E., L.H.B., J.M., G.M.V. wrote the manuscript with input from all authors.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-41625-z.

Competing Interests: The authors declare no competing interests.

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

analytical chemistry

Article

Cite This: Anal. Chem. 2019, 91, 11888−11896

pubs.acs.org/ac

Clusterwise Peak Detection and Filtering Based on Spatial Distribution To Efficiently Mine Mass Spectrometry Imaging Data

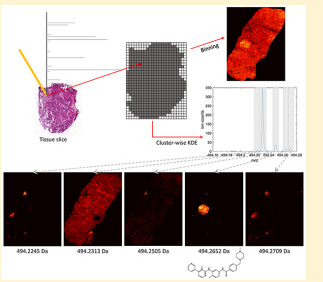
Jonatan O. Eriksson,[†] Melinda Rezeli,[†] Max Hefner,[†] Gyorgy Marko-Varga,[†] and Peter Horvatovich*[,‡,†]

[†]Lund University, Department of Biomedical Engineering, Lund, Sweden

[‡]University of Groningen, Department of Analytical Biochemistry, Groningen Research Institute of Pharmacy, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands

Supporting Information

ABSTRACT: Mass spectrometry imaging (MSI) has the potential to reveal the localization of thousands of biomolecules such as metabolites and lipids in tissue sections. The increase in both mass and spatial resolution of today's instruments brings on considerable challenges in terms of data processing; accurately extracting meaningful signals from the large data sets generated by MSI without losing information that could be clinically relevant is one of the most fundamental tasks of analysis software. Ion images of the biomolecules are generated by visualizing their intensities in 2-D space using mass spectra collected across the tissue section. The intensities are often calculated by summing each compound's signal between predefined sets of borders (bins) in the m/z dimension. This approach, however, can result in mixed signals from different compounds in the same bin or splitting the signal from one compound between two adjacent bins, leading to low quality ion images. To remedy this problem, we propose a novel data processing approach. Our approach consists of a sensitive peak detection method able to discover both faint and localized signals by utilizing clusterwise kernel density estimates (KDEs) of peak distributions. We show that our method can recall more ground-truth molecules, molecule fragments, and isotopes than existing methods based on binning. Furthermore, it automatically detects previously reported molecular ions of lipids, including those close in m/z, in an experimental data set.



Mass spectrometry imaging (MSI) is a technique often used to study the localization of known and unknown biomolecules such as lipids, metabolites, or peptides in tissue. Today's instruments can scan samples with both high spatial and mass spectral resolution and, consequently, generate massive data sets that require highly efficient and accurate processing. Thus, one of the key components of MSI data processing is data-reduction, which typically involves detection and extraction of signals originating from tissue or drug compounds while discarding noise.[1,2] The peaks of each spectrum are mapped onto a common reference, and by visualizing the intensities of individual peaks as images the spatial distribution of biomolecules can be revealed. The reference spectrum is generated by detecting peaks which are common to multiple spectra. Accurate peak detection facilitates the isolation of signals from individual compounds which is necessary to obtain high quality images.

Many existing MSI software, such as Cardinal[3] and MALDIquant,[4] detect isotopic peaks of compounds on a data set mean spectrum and subsequently rank them based on the frequency of their presence in ion image pixels. This method is fast and produces concise peak lists but has limited performance for low-intensity peaks and those localized to small regions in the analyzed tissue section.[1] Many tools generate ion images by binning around each peak of interest; the intensity value for each pixel is calculated by summing ion intensities between predefined m/z borders (bins). When doing this, however, it is crucial to use narrow bins to avoid mixing signals from multiple compounds in one image and to ensure that the mass of the peak around which binning is performed is accurate.

Suits et al.[5] showed that *slicing* the entire m/z range into ion images of fixed mass widths enables MSI practitioners to explore MSI data sets in a hypothesis-free manner. This approach sets no threshold on either peak intensity or presence in a minimum number of pixels and is thus not biased toward large or high intensity molecules in the tissue. Choosing bin width is a specificity-sensitivity trade off. A small bin width results in higher sensitivity but increases the risk of peak splitting and a higher number of empty or noninformative ion images. Larger bin widths on the other hand result in fewer noninformative images but are unable to discriminate between compounds that are close in mass, resulting in ion images containing signals from multiple compounds. Unfortunately, even when using relatively large bin widths, slicing leads to impractically large sets of ion-images unless the experimentalist is guided by known ion masses. However, previous studies have demonstrated that

Received: June 9, 2019
Accepted: August 12, 2019
Published: August 12, 2019

DOI: 10.1021/acs.analchem.9b02637
Anal. Chem. 2019, 91, 11888−11896

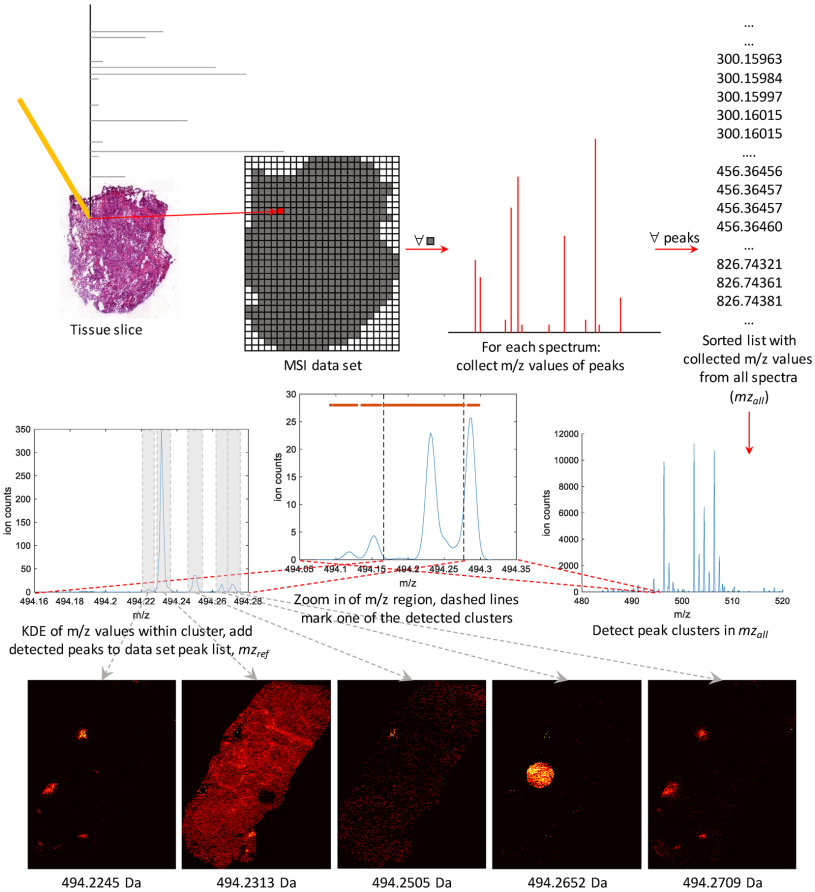


Figure 1. Flowchart of our peak picking algorithm. m/z values of peaks from each individual spectrum are collected and sorted in mz_{all}. We then identify clusters in mz_{all} as connected components in a directional graph. For each cluster we fit an optimized KDE to the distribution of m/z values. Data set peaks are obtained as local maxima on the resulting KDE curve. Finally, the level of structure in the ion images corresponding to the data set peaks is estimated and used to filter out noise peaks. The peak corresponding to the center ion image, at $m/z = 494.2505$, is an example of one filtered out in the last step.

incorporating information about the ion-images' spatial structure in MSI data analysis pipelines is an effective way to automatically separate high and low quality images in these large image sets.[6−9]

In this paper, we present a peak detection method that enables automatic detection of faint and localized signals as well as high intensity and/or abundant signals. We show that our peak detection can serve as a part of an MSI data analysis pipeline that is both sensitive and specific by combining it with established methods that filter peaks based on their spatial arrangement. A sensitive peak detection algorithm is not only essential for exploratory analysis but also for discovering molecules spatially colocalized with those expected to be present, e.g., drug compounds and metabolites. This is highly relevant in both scientific and clinical settings where drug−tissue interaction and tissue composition are often investigated. To assess and compare the performance of our method to existing MSI data processing tools, we used a rat liver section spiked with several drugs, most of which are anticancer drugs, where the masses of the spiked drugs are used as ground-truth. Using this data set, we show that we are able to detect drug peaks as well as fragment and isotopic peaks, including those that are close in m/z to more intensive and/or abundant peaks. We also used the MSI data set from a mouse bladder section originally presented by Römpp et al.[10] to further assess our method.

MATERIALS AND METHODS

Drug Compounds and Matrix Composition. For the MALDI-MSI experiment, we selected 12 different drugs (see chart in Supporting Information). The drugs were purchased from the LC Laboratories (Woburn, MA; CAS numbers: dabrafenib: 1195765-45-7, dasatinib: 302962-49-8, erlotinib: 183321-74-6, gefitinib: 184475-35-2, imatinib: 152459-95-5, lapatinib: 388082-78-8, pazopanib: 444731-52-6, sorafenib:

DOI: 10.1021/acs.analchem.9b02637
Anal. Chem. 2019, 91, 11888−11896

284461-73-0, sunitinib: 557795-19-4, trametinib: 871700-17-3, vatalanib: 212141-54-3) and from SelleckChem (Munich, Germany; CAS numbers: ipratropium: 60205-81-4) with >99% purity and were dissolved in methanol (MeOH, (Chromasolv Plus for HPLC) (Sigma-Aldrich, Steinheim, Germany) at 10 mg/mL concentration. These stock solutions were further diluted with 50% MeOH and five mixtures were generated, each containing four different drug compounds. The spreadsheet in Supporting Information summarizes the composition of the five drug mixtures. A 5 mg/mL solution of α-cyano-4-hydroxycinnamic acid (CHCA, Sigma-Aldrich) dissolved in 50% MeOH containing 0.1% trifluoroacetic acid (TFA, Sigma-Aldrich, Steinheim, Germany) was used as matrix solution.

Sample Preparation. For MALDI-MSI, a 10 μm section was cut from frozen rat liver tissue using a cryotome and placed on a glass slide. Then 0.3 μL from each drug mixture was pipetted on the tissue section at predefined positions. After drying of the tissue, CHCA matrix solution was deposited on the tissue surface by an automated pneumatic sprayer (TM-Sprayer, HTX Technologies). The nozzle distance was 46 mm, and the spraying temperature was set to 35 °C, the matrix was sprayed (19 passes) over the tissue section at a linear velocity of 750 mm/min with a flow rate set to 0.1 mL/min and a nitrogen pressure set at 10 psi. After each pass, a drying time of 30 s was set on the spraying machine to give time for the sample to dry completely before the next pass. The frozen rat liver tissue was provided by Prof. Roland Andersson (Dept. Clinical Sciences Lund (Surgery), Skane University Hospital, Lund University). Animals were housed and bred according to regulations for the protection of laboratory animals.

MALDI MSI. MSI data was collected by sampling the tissue section with 50 μm raster arrays without laser movement within each measuring position. The dimensions of the measured liver tissue section was approximately 0.9 by 1.2 cm in x, y sampling coordinates. A total of 23 823 sampling positions (x = 247, y = 181) were collected. Full mass spectra were collected using a MALDI LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, Bremen, Germany), equipped with a 60 Hz 337 nm nitrogen pulse laser (LTB Lasertechnik Berlin, Berlin, Germany). This instrument was operated at 60 000 resolution (at m/z 400) collecting spectral data in the mass range of 150−1000 m/z in profile mode generated by 20 laser shots at 10 μJ with automatic gain control switched off. Data were acquired using Xcalibur v 2.0.7. software (Thermo Fisher Scientific, San Jose, CA). The MSI raw data contains mass spectra from all measurement points together with their x, y coordinates.

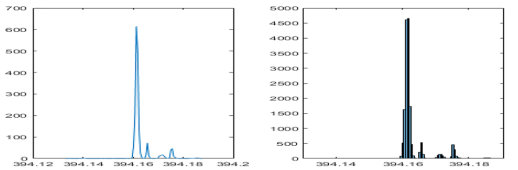
The Thermo Scientific raw files were first converted to *mzML* using *msconvert* and then to *imzML*[11] format using *imzmlConverter*. Finally, the *imzML* data was loaded into MATLAB and analyzed with custom scripts. The mouse bladder data set with PXD001283 ID was downloaded from ProteomeXchange in *imzML* format.

Peak Picking. We propose a two-step peak picking scheme: in the first step, candidate peaks are detected on clusters of peak m/z values from all spectra, and in the second, the spatial distribution of the candidate peaks is evaluated and we select those that display a coherent structure. For the first step, we have devised a novel method that relies on clusterwise kernel density estimates (KDEs) of spectral peaks. KDEs are smooth histograms and we use them to estimate the distribution of the peak m/z values within clusters along the m/z axis. The level of smoothness is adapted to each cluster independently. Candidates of data set peaks are then detected as local maxima on the resulting KDE curves. For the second step, we use two established ways to automatically estimate the quality of the images corresponding to peaks obtained in the first step as a means to filter out noninformative peaks. Figure 1 summarizes all parts of our peak picking scheme.

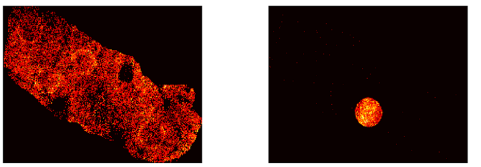
Peak Detection. First, we collect the peak masses from every spectrum in one list, mz_{all}, which is then sorted in ascending order. Centroided spectra are taken as input and peaks with heights below a very low intensity threshold are discarded to reduce the impact of background noise. Consequently, mz_{all} will contain most peak masses from the data set. Depending on data set size and RAM availability mz_{all} is processed either in segments or in its entirety. Second, peak clusters in the m/z dimension are identified using a one-dimensional directional graph. If the distance between an m/z value, m_i, and the next, m_{i+1}, is smaller than d_c, an edge connecting the two is added to the graph. The connected components in the resulting graph represent the m/z clusters. We let d_c increase with m/z to account for the peak broadening described by the known theoretical relationship between peak width (at half-maximum) and m/z: $d_c = f(m/z)$ where f depends on instrument type.[12] Suits et al.[13] summarized the relationship between peak width and instrument type. To reduce processing time, we discard clusters containing fewer than a minimum number of peaks. The threshold should be set sufficiently low to retain peaks representing meaningful anatomical structures in the tissue and is therefore dependent on the spatial resolution of the experiment. Finally, to test whether a cluster contains one or more peaks, a KDE is fitted to the distribution of m/z values within the cluster. The kernel bandwidth is optimized for each cluster individually using the normal optimal smoothing method described by Bowman and Azzalini.[14] Peaks are detected on the KDE curve in an iterative fashion: first the local maxima are detected and added together with their corresponding heights to a cluster-specific peak list, p_{kde}. The m/z corresponding to the highest peak in this list, mz_{max}, is added to the global peak list, mz_{ref}, and all surrounding peaks in p_{kde}, that fall within d_{kde} including mz_{max}, are removed. This step is repeated until p_{kde} is empty. The parameter d_{kde} is proportional to the expected peak width of the instrument in the same manner as d_c. The ion images are then generated by aligning each centroided spectrum to the resulting reference spectrum mz_{ref}, using a nearest neighbor method with maximum drift threshold dependent on the expected theoretical peak width (at half-maximum), similarly to the threshold used when generating edges between peaks in the clustering step.

Peak Selection. Although our method is more directed than slicing the spectra across the m/z range (since it only considers a selection of the m/z regions), it still generates many peaks representing noise in addition to those correlated with actual tissue structures, making it essential to separate the former from the latter. We use the spatial chaos[8] (SC) and the principal component analysis (PCA)-based variance explained[15] (VE) measures to automatically estimate the level of structure in the ion images. The spatial chaos counts the number of connected objects in an ion image. More structured ion images are expected to have fewer disconnected (separate) objects than unstructured ones. The VE measure is the percentage of total variance explained by the first pair of singular vectors of each ion image. This corresponds to how much of the variation in intensity along one axis of the image is explained by the intensities along the other. The first principal component inherently explains the

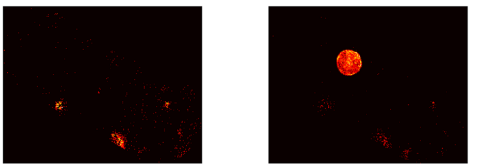
DOI: 10.1021/acs.analchem.9b02637
Anal. Chem. 2019, 91, 11888−11896



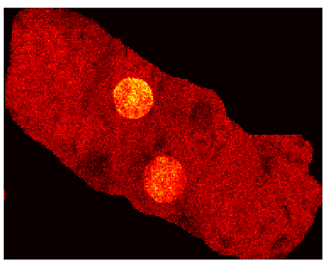
(a) **Left**: Optimized KDE curve. **Right**: Histogram of m/z distribution.



(b) 394.161 m/z (c) 394.166 m/z



(d) 394.172 m/z (e) 394.176 m/z



(f) Ion-image obtained by binning spectra between 394.15 and 394.20 m/z.

Figure 2. (a) The distribution of m/z peak values within the cluster containing erlotinib (m/z 394.176). (b−e) The ion images that correspond to the four peaks on the KDE curve. (f) The ion image obtained by binning the spectra between 394.15 and 394.20 m/z; this image demonstrates how four signals can be mixed in the same ion image and even when a relatively narrow m/z window is used.

most variance and, thus, if it explains very little, so will all others. In structured images there is typically an intensity relationship between the axes and therefore their VE is expected to be higher than that of images with randomly distributed intensities, i.e., unstructured images, in which this relationship is unlikely to exist.

RESULTS AND DISCUSSION

Two data sets were used to assess the performance of our novel MSI data preprocessing algorithm based on clusterwise peak detection. The first MALDI-MSI data set (referred to as the "spiked data set") was generated by spiking a rat liver section with 5 mixtures of 4 ground-truth drugs (12 different compounds in total) in various concentrations. These mixtures were spotted on a rat liver tissue section at five different locations in circular areas of the same size (Figure S1) and, after matrix deposition, the whole tissue section was analyzed by MALDI-MSI using 50 μm spatial resolution. The concentrations of the drug compounds covered an intensity range of 3 orders of magnitude between trametinib (1.70×10^4) and ipratropium (1.49×10^7). Furthermore, some of the ground-truth drugs such as erlotinib and dasatinib, were spotted at multiple locations in different concentrations. The second data set, originally from Römpp et al.,[10] comes from a mouse bladder section and was downloaded from ProteomeXchange (XD001283). This MSI data set was generated by a LTQ Orbitrap instrument with an ion source built in-house used to scan the mouse bladder section with 10 μm spatial resolution. The authors of this study presented the ion images of 11 compounds. These images were generated with a narrow bin width of 0.01 Da. For this data set, we use the mass of these compounds as ground truth, i.e., peaks known to be present.

DOI: 10.1021/acs.analchem.9b02637
Anal. Chem. 2019, 91, 11888−11896

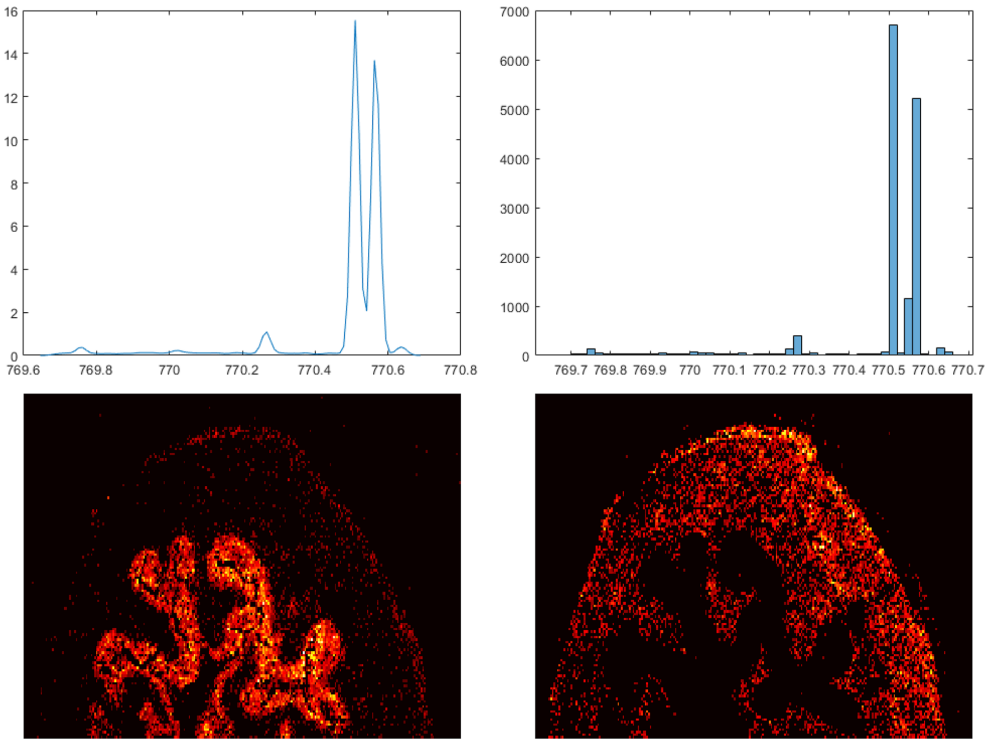


Figure 3. Distribution of peak m/z values within the cluster containing PC (32:1) (770.5109 m/z) and SM(18:0) (770.5609 m/z). The ion images corresponding to the two highest peaks on the KDE curve are shown in the bottom left and bottom right.

Recall of Known Compounds. We applied Cardinal, MALDIquant, slicing the spectra into 0.05 Da bins, and our clusterwise peak detection method to the spiked data set to compare their ability to recall compounds. The difference between the known mass of each ground-truth drug and the mass of the closest detected peak is used as the measure of accuracy for Cardinal and our method. The ion images corresponding to the monoisotopic peak of the ground-truth drugs were manually evaluated to confirm that a compound had been correctly found. First, we ran Cardinal and detected 4751 peaks; we did not filter out those with too low pixel frequency. The corresponding ion images were generated by binning around each peak. Eight of the 12 compounds were detected with a mass deviation ranging between 4.23 and 198.85 ppm (mean 83.983 ppm). Figure S2 shows the ion images of the drug compounds generated by Cardinal. The ion images of erlotinib (394.176 Da) and geftinib (447.160 Da) are contaminated with signal from other compounds while sunitinib (399.220 Da), imatinib (494.267 Da), and trametinib (616.086 Da) are completely missed. Second, we used MALDIquant to compute a mean spectrum on which we detected 521 peaks. Only the peak from the drug with the highest measured intensity, ipratropium, was found with a mass deviation of 4.7145 ppm. The ion image corresponding to the monoisotopic peak of iptratropium indicates that this compound has diffused from the spotting location and because of this covers a significantly larger region of the tissue than the other compounds; this might contribute to its presence in the mean spectrum which favors signals that have high intensity and/or pixel frequency. Third, we sliced the spectra with a bin width of 0.05 Da across the 150−1000 m/z range resulting in 17 000 slices. To asses the sensitivity of the slicing approach we manually examined the ion images corresponding to the slices containing the m/z of the spiked-in drug compounds (Figure S3). The signal from trametinib (616.086) is missed and those from erlotinib (394.176 Da) and imatinib (494.267 Da) are mixed with others, resulting in contaminated ion images. Finally, when applying our method, we identified 3148 m/z clusters in the data set peak list and on the KDEs of these we detected 6088 peaks. We used a value of 0.2 times the theoretical peak width at half-maximum for d_c, the parameter controlling the maximum distance between connected points that form the m/z clusters. Decreasing or increasing d_c between 0.1 and 0.5 results in a higher or lower number of clusters, respectively, but ultimately has little impact on the final peak list. All of the 12 spiked-in compounds are detected with mass deviations ranging between 1.00 and 4.29 ppm (mean 2.598 ppm). Figure S4 shows the ion images corresponding to the monoisotopic peaks of the drug compounds generated by our method. The signal from trametinib is weak but detected nevertheless; it had the lowest measured intensity which can explain its absence in some of the spectra. Generally, the quality of images generated with our

DOI: 10.1021/acs.analchem.9b02637
Anal. Chem. 2019, 91, 11888−11896

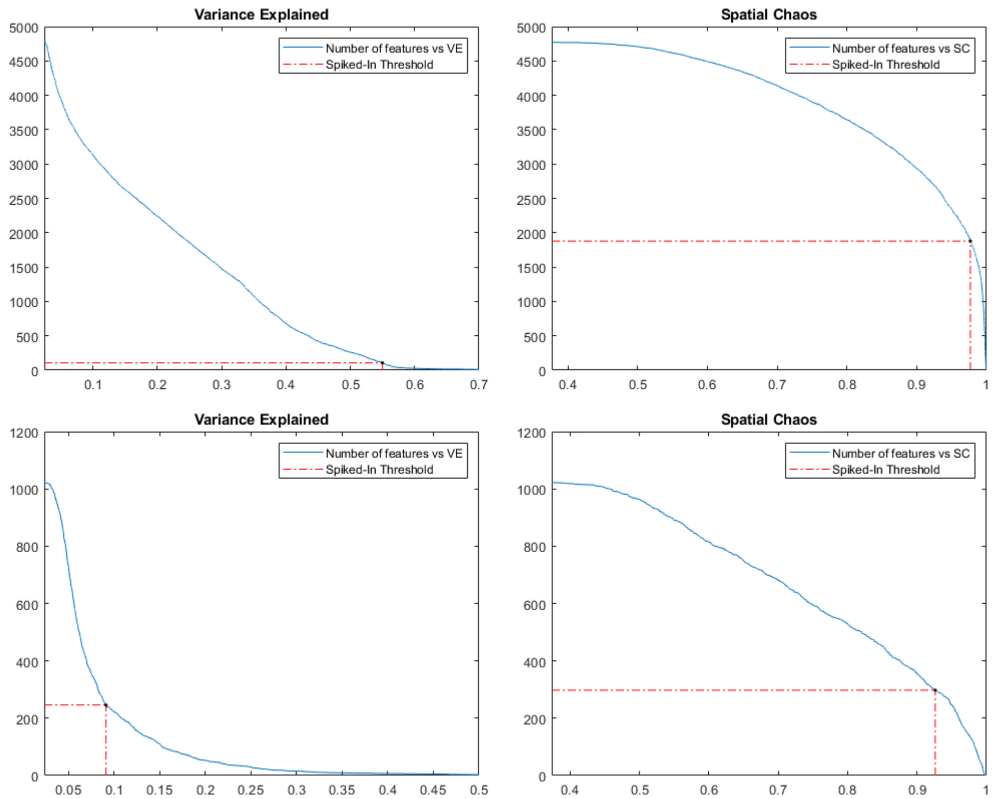


Figure 4. Number of ion images surviving varying thresholds on the VE and SC scores in the two data sets. Dashed lines mark the lowest scores (excluding the low quality image for m/z 616.127) of the ion images corresponding to the drugs in the spiked data set (top) and known compounds in the mouse bladder data set (bottom).

approach is higher than that of the images generated with Cardinal or by slicing. The drug signals are clearly visible against the background, and there is no contamination with signals from other compounds, background, or matrix. Table S1 shows the mass deviations of the detected peaks corresponding to the spiked-in drugs obtained with Cardinal and our algorithm. The corresponding ion images are shown in Figure S2 and Figure S4, respectively.

An example of a cluster with densely located molecule signals is that containing erlotinib (394.176 Da) (Figure 2a). There are four distinctive signals within this relatively narrow m/z window (0.04 Da) at 394.161, 394.166, 394.172, and 394.176 m/z with interpeak distances of 13, 15, and 10 ppm. The peak at 394.161 m/z is tissue-derived while those at 394.166 m/z and 394.172 come from a fragment molecule of imatinib and the matrix, respectively. Using our method we are able separate the four peaks and generate a clean image for each of them. Figure 2b–e shows the ion images related to these peaks. If the spectra are binned between 394.150 and 394.200 m/z instead, the signals from three of the four compounds appear in the same ion image, i.e., they are incorrectly combined into one ion-image while that from the peak at 394.172 m/z is invisible (Figure 2f) due to its low intensity compared to the other three. We found that a value between 0.25–0.5 times the theoretical peak width at half-maximum is a good choice for d_{kde}, the parameter controlling the minimum distance between two adjacent peaks on the KDE curve. Using a higher value results in fewer noise peaks, however, we lose true peaks, e.g., those from imatinib and erlotinib. Because of this, we recommend using a small d_{kde} to delay filtering out noise peaks until after alignment by using one of the spatial distribution based peak selection methods. The kernel bandwidth used when generating the cluster KDEs is optimized for each cluster individually to account for the variability in peak density. This parameter determines the level of smoothing when estimating the distribution of the peak masses within the clusters. Similarly to d_{kde}, using a higher bandwidth results in less noisy data, however, may lead to losing true peaks or mixing signals from multiple compounds.

We also applied our cluster-based peak detection method to the high spatial resolution mouse bladder data set. In this data set we detected 1702 m/z clusters and 6482 peaks. We then filtered out peaks which were present in fewer than 200 of the 33 000 spectra, resulting in a final list of 1024 data set peaks. The original paper reported 11 ion images that were manually generated by binning around peaks with known m/z using a very narrow bin width of 0.01 Da. All peaks corresponding to these

DOI: 10.1021/acs.analchem.9b02637
Anal. Chem. 2019, 91, 11888–11896

Table 1. VE and SC Scores of the Ion Images Corresponding to the Spiked-in Drug Compound in the Spiked Data Set and Their Corresponding Rank among the 4771 Ion Images That Remain after Removing Those with Fewer Than 400 Nonzero Pixels

compound	mass	VE	percentile	rank (VE)	SC	percentile	rank (SC)
ipratropium	332.223	0.5997	99.43	27	0.9997	99.94	3
vatalanib	347.107	0.7183	99.79	10	0.9952	79.29	988
erlotinib	394.177	0.7837	99.85	7	0.9775	61.04	1859
sunitinib	399.220	0.6845	99.73	13	0.9921	72.23	1325
pazopanib	438.171	0.8853	99.98	1	0.9837	64.60	1689
gefitinib	447.160	0.8362	99.92	4	0.9948	78.22	1039
sorafenib	465.094	0.8328	99.90	5	0.9951	79.04	1000
dasatinib	488.164	0.6400	99.62	18	0.9980	92.10	377
imatinib	494.267	0.7611	99.81	9	0.9766	60.64	1878
dabrafinib	520.109	0.5499	97.78	106	0.9964	83.29	797
lapatinib	581.143	0.6715	99.69	15	0.9775	60.97	1862
trametinib	616.086	0.1696	70.72	1397	0.9038	53.07	2239

Table 2. VE and SC Scores of the Ion Images Corresponding to the 11 Compounds Reported by Römpp et al.[10] and Their Corresponding Rank among the 1053 Candidate Ion Images That Remain after Removing Those with Fewer Than 200 Nonzero Pixels

compound	mass	VE	percentile	rank (VE)	SC	percentile	rank (SC)
LPC (16:0), $[M + K]^+$	535.296	0.1770	92.76	74	0.9897	94.52	56
LPC (18:0), $[M + K]^+$	562.327	0.2732	98.14	19	0.9964	99.12	9
heme b, M^+	616.177	0.2385	96.67	34	0.9261	70.84	298
unknown	713.452	0.0911	75.93	246	0.9444	73.68	269
SM (16:0)	742.531	0.2140	95.50	46	0.9953	98.24	18
unknown	743.548	0.1921	94.42	57	0.9691	84.34	160
PC(32:1), $[M + K]$	770.507	0.2688	97.95	21	0.9814	88.85	114
SM(18:0), $[M + K]$	770.565	0.1439	87.87	124	0.9849	90.90	93
PC (32:0),$[M + K]^+$	772.525	0.3177	98.83	12	0.9975	99.80	2
PC (34:1), $[M + K]^+$	798.541	0.3383	99.02	10	0.9979	99.90	1
PE(38:1)	812.557	0.1623	91.39	88	0.9909	95.21	49

ion images are found by our peak detection method in an unsupervised fashion, including the two densely located peaks at 770.5097 and 770.5698 m/z originating from the K^+ adduct of PC(32:1) [phosphatidylcholine] and an isotope of the K^+ adduct of SM(36:1), [sphingosylphosphorylcholine], respectively (Figure 3). Figure S5 shows the ion images related to the 11 detected peaks.

Peak Selection. As previously mentioned, we find more than 6000 peaks in the rat liver data set with our cluster-based peak detection, resulting in an equal number of ion images. Manually evaluating each image is impractically slow, but by computing the spatial chaos (SC) and the variance explained (VE) for all ion images, including those of the compounds known to be present, we can estimate how much we can reduce the number of images without losing relevant information. For each data set, we took the VE and SC scores of the ion images corresponding to the known compounds and used their mean scores minus two standard deviations as low-end thresholds. The number of peaks whose images had scores above these thresholds indicates how many of the detected peaks should be kept and how many can be rejected as noise. In the spiked data set this filtering resulted in a final list of 843 and 2170 peaks when we filtered based on VE and SC scores, respectively. The numbers of peaks obtained for the mouse bladder data set are 418 and 288 for VE and SC, respectively. The number of ion images whose VE or SC score is above various thresholds is shown in Figure 4. The number of peaks can potentially be further reduced if off-tissue regions are available; biologically irrelevant peaks, such as those coming from solvents or the matrix, can be filtered out since their signal often is stronger in these regions.[15]

Despite its simplicity, the VE score proved to be very effective in ranking the quality of the ion images generated from both the spiked and mouse bladder data sets. Specifically, VE favors images which have intensities localized to small regions, e.g., all of the spiked-in compounds in the spiked data set and heme b, M^+ at $m/z = 616$ (Figure S5c) in the mouse bladder data set. In contrast, ion images with high levels of structure across the entire scanned region tend to be rewarded with the highest SC scores, making it suitable as a general measure of image quality but less effective than the VE score in identifying ion images with localized structured intensity patterns. The two scores appeared to be partially complementary to each other; the Pearson correlation between the VE and SC scores in the spiked and mouse bladder data sets were 0.6158 and 0.4821, respectively. Tables 1 and 2 show the VE and SC scores of the ion images corresponding to the ground truth compounds in the spiked and mouse bladder data sets, respectively.

Detection of Fragments and Isotopes. MALDI-MSI is an important tool often used to investigate the distribution of drugs and drug metabolites in tissue during pharmaceutical research, and obtaining comprehensive lists of interacting molecules is crucial during their development. To this end, we further assessed the performance of our peak detection method by searching for molecules colocalized with the drugs in the spiked data set. Colocalization analysis can be performed by computing the Pearson correlation coefficient between the ion image of a peak of interest and all other images.[5,16,17] For each

DOI: 10.1021/acs.analchem.9b02637
Anal. Chem. 2019, 91, 11888−11896

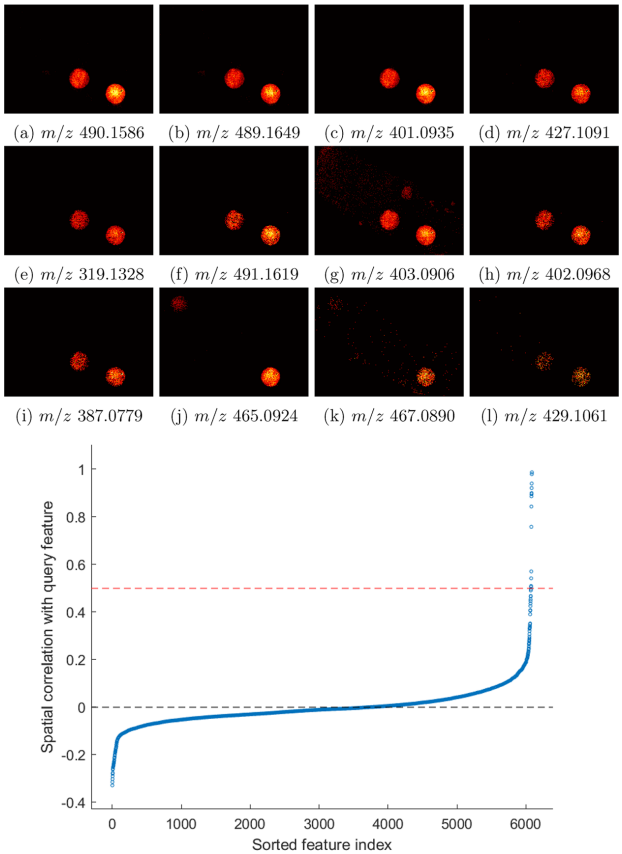


Figure 5. Top: The ion images of the 12 most correlated peaks to dasatinib's monoisotopic peak. Panels a−i and l are isotopes or fragments of dasatinib while panels j and k are related to sorafenib. Bottom: Sorted Pearson correlation between all ion images and that of the monoisotopic peak of dasatinib.

drug compound, we computed the correlation coefficient between the ion image corresponding to its monoisotopic peak and every ion image from the full image sets generated using the peaks found with our clusterwise peak detection method and that generated by slicing, without performing peak filtering based on spatial distribution. We manually assessed images whose correlation coefficient was ≥0.5 to search for candidate fragments and isotopes with spatial intensity distributions matching those of the drugs. The m/z of the matching images and existing knowledge about the theoretical fragmentation pattern of the drugs were then used to identify the fragments. This resulted in the identification of 46 isotopes and fragments in the ion image set generated by our method and 32 in the set generated by slicing. We gain an additional 14 fragments and isotopes when using our peak detection approach compared to when slicing the spectra with a bin width of 0.05 Da.

The correlation analysis result of dasatinib is shown in Figure 5. In total, 12 ion images have a correlation coefficient ≥0.5. The nine most correlated images (≥0.75) consist of three isotopes of dasatinib with an m/z of 489.165, 490.159, and 491.162, and six fragments with an m/z of 319.133, 387.078, 401.094, 402.097, 403.091, and 427.110. The fragments' and isotopes' ion images show minimal signal mixing with other compounds as shown in Figure 5. The remaining three consist of another fragment of dasatinib with an m/z of 429.106 and a correlation coefficient of 0.5422 and two ion images related to sorafinib. The indentified fragments and results of the correlation analysis are presented in Supporting Information spreadsheet and Figures S6−S16. We also assessed the most anticorrelated images to investigate whether there was evidence of ion suppression from any of the ground-truth drugs. However, no images uniquely anticorrelated to any one of the spiking spots were found. Instead, these images were anticorrelated to all spiking spots simultaneously, indicating that they are the result of washing or ion suppression from the solvent used in the drug mixtures.

CONCLUSIONS

In this paper we have presented an efficient peak picking approach combining a novel peak detection algorithm with filtering based on spatial information to automatically identify ion images corresponding to isotopic peaks of both endogenous

DOI: 10.1021/acs.analchem.9b02637
Anal. Chem. 2019, 91, 11888−11896

and drug compounds in high-resolution MSI data sets. It should be noted that these data sets were generated using high-resolution Orbitrap MSI, which is low-pass-filtered during acquisition by default. Applying our method to noisier data such as that generated by QTOF MSI would require additional preprocessing such as baseline removal and smoothing. Our KDE clusterwise peak detection algorithm enables us to find low intensity and localized peaks with minimal contamination from other peaks close in m/z, resulting in high ion image quality. We believe that implementing our MSI preprocessing algorithm in an interactive tool would be valuable to experimentalists who aim to identify a priori unknown endogenous compounds, reveal drug distributions in tissue, or find compounds that spatially correlate to known ones. Such a tool could help users gain deeper insight into the effect of drugs in tissue and considerably reduce the number of ion images that have to be examined manually.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b02637.

Methods and figures (PDF)

Tables of correlating peaks for each spiked-in compound with structures and annotations (isotopes, fragments) and the description of the 5 drug mixtures (XLSX)

Structures of spiked-in drugs (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: p.l.horvatovich@rug.nl.

ORCID

Melinda Rezeli: 0000-0003-4373-5616

Peter Horvatovich: 0000-0003-2218-1140

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Frank Suits for his support and insightful discussions throughout the project and we kindly acknowledge the support from Fru Berta Kamprads Stiftelse.

■ REFERENCES

(1) Jones, E. A.; Deininger, S.-O.; Hogendoorn, P. C.; Deelder, A. M.; McDonnell, L. A. *J. Proteomics* **2012**, *75*, 4962−4989.

(2) Gessel, M. M.; Norris, J. L.; Caprioli, R. M. *J. Proteomics* **2014**, *107*, 71−82.

(3) Bemis, K. D.; Harry, A.; Eberlin, L. S.; Ferreira, C.; van de Ven, S. M.; Mallick, P.; Stolowitz, M.; Vitek, O. *Bioinformatics* **2015**, *31*, 2418−2420.

(4) Gibb, S.; Strimmer, K. *Bioinformatics* **2012**, *28*, 2270−2271.

(5) Suits, F.; Fehniger, T. E.; Végvári, Á.; Marko-Varga, G.; Horvatovich, P. *Anal. Chem.* **2013**, *85*, 4398−4404.

(6) Alexandrov, T.; Bartels, A. *Bioinformatics* **2013**, *29*, 2335−2342.

(7) Wijetunge, C. D.; Saeed, I.; Boughton, B. A.; Spraggins, J. M.; Caprioli, R. M.; Bacic, A.; Roessner, U.; Halgamuge, S. K. *Bioinformatics* **2015**, *31*, 3198−3206.

(8) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. *Nat. Methods* **2017**, *14*, 57.

(9) Inglese, P.; Correia, G.; Takats, Z.; Nicholson, J. K.; Glen, R. C. *Bioinformatics* **2019**, *35*, 178−180.

(10) Römpp, A.; Guenther, S.; Schober, Y.; Schulz, O.; Takats, Z.; Kummer, W.; Spengler, B. *Angew. Chem., Int. Ed.* **2010**, *49*, 3834−3838.

(11) Schramm, T.; Hester, A.; Klinkert, I.; Both, J.-P.; Heeren, R. M.; Brunelle, A.; Laprévote, O.; Desbenoit, N.; Robbe, M.-F.; Stoeckli, M.; Spengler, B.; Römpp, A. *J. Proteomics* **2012**, *75*, 5106−5110.

(12) Hoffman, E. D.; Stroobant, V. *West Sussex*; John Wiley & Sons, Bruxellas, Bélgica, 2007, *1*, 85.

(13) Suits, F.; Hoekman, B.; Rosenling, T.; Bischoff, R.; Horvatovich, P. *Anal. Chem.* **2011**, *83*, 7786−7794.

(14) Bowman, A. W.; Azzalini, A. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*; OUP Oxford, 1997; Vol. *18*.

(15) Fonville, J. M.; Carter, C.; Cloarec, O.; Nicholson, J. K.; Lindon, J. C.; Bunch, J.; Holmes, E. *Anal. Chem.* **2012**, *84*, 1310−1319.

(16) Nemes, P.; Woods, A. S.; Vertes, A. *Anal. Chem.* **2010**, *82*, 982−988.

(17) Fehniger, T. E.; Suits, F.; Végvári, Á.; Horvatovich, P.; Foster, M.; Marko-Varga, G. *Proteomics* **2014**, *14*, 862−871.

DOI: 10.1021/acs.analchem.9b02637
Anal. Chem. 2019, 91, 11888−11896

analytical chemistry

pubs.acs.org/ac

Article

MSIWarp: A General Approach to Mass Alignment in Mass Spectrometry Imaging

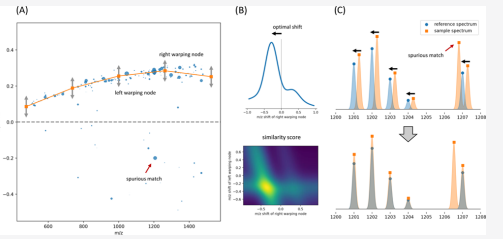
Jonatan O. Eriksson, Alejandro Sánchez Brotons, Melinda Rezeli, Frank Suits, György Markó-Varga, and Peter Horvatovich*

Cite This: *Anal. Chem.* 2020, 92, 16138−16148

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Mass spectrometry imaging (MSI) is a technique that provides comprehensive molecular information with high spatial resolution from tissue. Today, there is a strong push toward sharing data sets through public repositories in many research fields where MSI is commonly applied; yet, there is no standardized protocol for analyzing these data sets in a reproducible manner. Shifts in the mass-to-charge ratio (m/z) of molecular peaks present a major obstacle that can make it impossible to distinguish one compound from another. Here, we present a label-free m/z alignment approach that is compatible with multiple instrument types and makes no assumptions on the sample's molecular composition. Our approach, MSIWarp (https://github.com/horvatovichlab/MSIWarp), finds an m/z recalibration function by maximizing a similarity score that considers both the intensity and m/z position of peaks matched between two spectra. MSIWarp requires only centroid spectra to find the recalibration function and is thereby readily applicable to almost any MSI data set. To deal with particularly misaligned or peak-sparse spectra, we provide an option to detect and exclude spurious peak matches with a tailored random sample consensus (RANSAC) procedure. We evaluate our approach with four publicly available data sets from both time-of-flight (TOF) and Orbitrap instruments and demonstrate up to 88% improvement in m/z alignment.



INTRODUCTION

Mass spectrometry (MS) is a widespread analytical technique used to detect and quantify ionized molecules, and it has many applications in biology, chemistry, and medicine. In MS imaging (MSI), molecular ions are sampled from different locations on a surface area, such as a tissue section, allowing the mass spectrometer to serve as a molecular imaging device. The ability to determine the spatial distribution of thousands of biological compounds in a single experiment makes MSI a powerful tool for tissue characterization. There have been extensive developments in the MSI field during the last decades, resulting in new experimental workflows, improved ionization and sampling methods, and advances in both the spatial and mass resolutions of instruments.[1−3] The availability of a wide variety of ionization techniques such as secondary-ion MS (SIMS), desorption electrospray ionization (DESI), and matrix-assisted laser desorption/ionization (MALDI) allows ionization of many compound classes with both targeted and untargeted approaches.[1] High-performance Orbitrap and Fourier transform ion cyclotron resonance (FT-ICR) mass analyzers can scan tissue sections with a subcellular spatial resolution and a mass resolution exceeding 500 000. Low sensitivity has been a longstanding obstacle for high-spatial-resolution MSI but has recently been improved by optimizing the laser wavelength in MALDI to increase ionization efficiency, or by post-ionizing neutral molecules to increase ion yield.[2] Novel sample preparation workflows have led to enhanced quantification and identification of metabolites, peptides, and proteins. These include protein extraction methods, in situ protease digestion of proteins, and the use of chemical derivatization such as labeling with photocleavable mass tags to enhance low-intensity molecule signals.[3,4] Altogether, this leads to highly complex data sets that demand accurate preprocessing and sophisticated bioinformatic analysis to maximize their utility in biological and clinical research.[5]

A persisting issue in MSI is systematic mass misalignment, leading to slight shifts in the m/z ratio of molecule peaks across spectra. These shifts can result in misidentified peaks or an increased risk of mixing peaks from different molecules with similar masses in the same ion image. Mass misalignment is typically more severe for time-of-flight (TOF) instruments than for FT-ICR, Orbitrap, or other Fourier transform (FT) instruments.[6] Variations in temperature throughout the

Received: September 9, 2020
Accepted: November 16, 2020
Published: December 2, 2020

experiment, contractions and dilatations of the ion tube, contamination of the ion source, and tissue topography are some factors that are related to mass shifts in spectra generated by TOF instruments. For FT instruments, the most common source of mass misalignment is the space-charge effect, which causes mass shifts that increase with the number of ions in the trap.[7,8] The mass shifts in FT instruments can often be limited using automatic gain control (AGC), but can be considerable if suboptimal AGC settings are used.

When discussing mass misalignment and mass shifts, it is important to distinguish between relative mass alignment and absolute mass accuracy. Here, the former refers to how tightly peak masses are distributed across spectra, while the latter refers to the difference between a molecule's theoretical peak mass and its observed peak mass. A common approach to correct mass shifts is to perform either external or internal calibration by comparing the measured masses of predefined peaks to their expected theoretical masses. External calibration is performed by depositing a calibration standard outside the tissue region and comparing the measured peak masses to the known masses of the calibrants. The same calibration function is then applied to all tissue spectra. While this approach is simple, it does not reduce the relative misalignment or correct mass shifts introduced during the experiment. Internal calibration, on the other hand, relies on identifying known peaks in each tissue spectrum. The known peaks can be either prominent molecule peaks intrinsic to the tissue or peaks corresponding to calibrants sprayed on the whole tissue area. Internal calibration can improve both the relative alignment between spectra and absolute mass accuracy but performs poorly for spectra in which the calibration peaks are missing or incorrectly assigned.[9]

An alternative approach is to align all spectra to a common reference spectrum. Given that the absolute mass accuracy of the reference spectrum is high, this can reduce relative misalignment and improve absolute mass accuracy simultaneously. A simple approach to aligning one spectrum to another is fitting a polynomial recalibration function to the mass difference of their shared peaks and then using this recalibration function to recalibrate all peak masses. While this approach is flexible, it is highly sensitive to spurious peak matches. Bocker and Makinen[10] introduced a linear curve-fitting approach that is robust to spurious peak matches, and Kulkarni et al.[11] later used this approach, together with spatial information, to further improve mass alignment. More recently, there has been increased focus on mass alignment for TOF instruments. Ráfols et al.[6] proposed an alignment algorithm that uses the cross-correlation between two spectra in the upper and lower parts of the mass range to estimate mass shift. They then use the upper and lower shifts to recalibrate the mass axis of one of the spectra to that of the other. Boskamp et al.[9] elegantly exploit statistical properties of the peptide background signal to improve mass alignment and absolute mass accuracy. They estimate the mass shift across the m/z range by comparing observed to theoretical peak masses on the Kendrick mass scale. Unlike Ráfols et al.'s method, their method can correct nonlinear mass shifts, but the dependency on the peptide background limits its generalizability.

Another aspect of mass alignment is at which stage in the processing pipeline it is performed. It can be performed in the time domain for TOF spectra, in the frequency domain for FT spectra, using profile mass spectra, or using centroided mass spectra.[6,10,12,13] In principle, aligning spectra at an early stage is advantageous in the sense that errors are not accumulated in subsequent processing steps. In practice, however, time or frequency spectra are often inaccessible as vendor software typically only provide mass spectra, and the majority of data sets uploaded to repositories are processed to some extent. FT spectra, in particular, are often centroided to reduce their otherwise impractical size.[14] It is impossible to recover a raw spectrum from one that has been processed. Hence, an approach must be able to perform alignment with processed spectra to be compatible with most MSI data sets. This compatibility is essential; data sets generated independently in other labs are frequently used to validate biological findings or novel methods. A public data set may be generated with any instrument and additional processing is sometimes required to ensure its quality. This compatibility requirement, together with the growing popularity of public MSI data set repositories such as MetaboLights[15] or METASPACE,[16] creates a demand for algorithms that can perform accurate mass alignment on spectra acquired with multiple instrument types and regardless of whether they are already partially processed.

In this work, we adapt the correlation optimized warping (COW)[17] algorithm to perform label-free MSI mass alignment using a custom benefit function and show that we can greatly reduce variation in peak masses between spectra. COW aligns a pair of signals by performing local warpings on one signal so that the global similarity relative to the other is maximized. We have previously shown that COW is effective in reducing misalignment in the time dimension between liquid chromatography−mass spectrometry (LC-MS) sample runs.[18,19] Here, we instead use COW to reduce mass misalignment between spectra by warping the mass dimension.

Crucially, our method finds the optimal warping of one spectrum relative to another using only a list of centroided peaks from each spectrum. We model centroided peaks as Gaussians whose widths vary with m/z, and define the similarity between two spectra as the total overlap of their shared peaks. To assess and further improve COW's robustness for particularly peak-sparse or misaligned spectra, we include an optional outlier detection step in the form of a tailored random sample consensus (RANSAC)[20] procedure. The concept of our method is similar to that of Ráfols et al.,[6] but differs in two key aspects. First, we find the mass recalibration function by maximizing the product (overlap) of centroided peaks instead of the cross-correlation of continuous spectra. Second, our method can, since it is derived from COW, correct nonlinear mass shifts. Thereby, our method utilizes information about peak height and width (not simply mass location) while remaining compatible with most MSI data sets. We demonstrate the effectiveness and generalizability of our method, named MSIWarp, by applying it to four publicly available data sets. MSIWarp performs accurate and robust alignment with centroided spectra, is compatible with multiple instrument types, and makes no assumptions on the molecular composition of the sample. We provide a fast C++ implementation of MSIWarp together with a Python binding at https://github.com/horvatovichlab/MSIWarp.

THEORY

The core of MSIWarp is a spectrum-to-spectrum similarity score that is used to find an optimal warping function between the mass axis of one spectrum and that of another. To align an entire MSI data set, all spectra, the sample spectra, are warped to a common reference spectrum that can be selected from the

https://dx.doi.org/10.1021/acs.analchem.0c03833
Anal. Chem. 2020, 92, 16138−16148

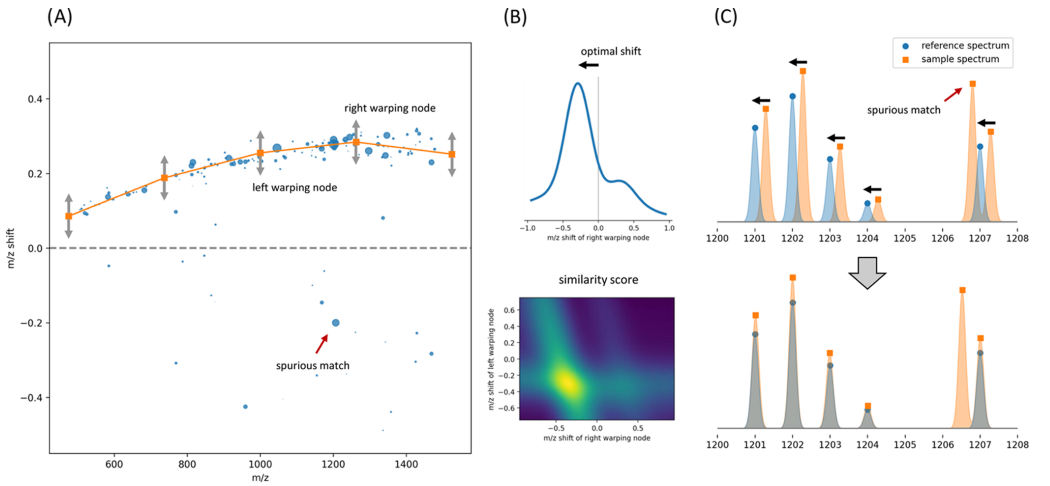


Figure 1. Conceptual description of MSIWarp. After matching the peaks in the sample spectrum against those in the reference spectrum, the sample spectrum is warped so that its similarity to the reference spectrum is maximized. (A) Scatter of the mass shift between matched peaks across the m/z range. The shifted warping nodes are marked with vertical arrows (the actual search space is centered around zero and extends beyond the arrows). The orange curve shows the estimated mass shift based on our similarity score. (B) The similarity score is evaluated for the set of candidate warpings, and the warping resulting in the highest score is used to align the sample spectrum. (C) Zoom-in of the spectra with the centroided peaks modeled as Gaussians. Two sample spectrum peaks are matched to the reference spectrum peak at 1207 m/z, but the spurious match (also marked in (A)) has no effect on the warping.

data set, or be a composite spectrum constructed from multiple spectra. Similar to our previous work,[19] MSIWarp deliberately relies on centroided, i.e., peak-picked, spectra to perform alignment. Peak-picking can improve alignment by retaining most compound-related signals while removing background noise that degrades performance. More importantly, however, an alignment method that takes centroided spectra as input can trivially be used to align profile spectra, but not vice versa. MSIWarp relies on centroid spectra instead of profile spectra and is thereby readily applicable to almost any MSI data set.

MSIWarp can be summarized in three steps (Figure 1): first, we match the peaks of the sample spectrum against those of the reference spectrum. Second, based on the peak matches from step one, we split the m/z range in a manner that ensures there is sufficient shared information in all segments. Finally, we find an optimal warping function with the peak matches and the partitioning of the m/z range obtained in steps one and two, respectively, and use this function to recalibrate the peak masses of the sample spectrum.

Mass Alignment. Our method aligns a pair of spectra by warping one in the mass dimension so that its similarity to the other is maximized. Provided that the type of mass spectrometer used to generate the spectra is known, it requires only a list of peak heights and m/z locations for each spectrum. We model peak intensity as a Gaussian function of m/z with centroid mass μ and height H. With this peak model, we can then compute the similarity of two centroided spectra as the sum of all pairwise peak overlaps, and use this similarity score as a measure of alignment quality. To model peak width, σ, we use known theoretical relationships between peak width and m/z, together with the mass resolution of the data set. The theoretical relationships depend on instrument type and are summarized in Suits et al.[21] If the mass resolution is unknown, a good estimate of a single peak's full width at half-maximum (FWHM) is enough to model the width of all other peaks in the data set. While not a true representation of a mass spectrum peak, the Gaussian peak model is a sufficient approximation for the purpose of alignment. Similarly, the modeled peak width does not have to match the true width exactly, since its main purpose is to provide some freedom when matching and aligning peaks.

Equations 1−4 formally define our Gaussian peak model p, the overlap between two peaks I, and the similarity between two spectra, B. The intensity of a peak varies with m/z according to

$$p(mz) = H \exp\left(-\frac{(mz-\mu)^2}{2\sigma^2}\right) \tag{1}$$

The overlap of two peaks is

$$I(p_i, p_j) = \int_{-\infty}^{\infty} H_i \exp\left(-\frac{(mz-\mu_i)^2}{2\sigma_i^2}\right) \times H_j \exp\left(-\frac{(mz-\mu_j)^2}{2\sigma_j^2}\right) dmz \tag{2}$$

The integral in eq 2 can be solved analytically to yield

$$I(p_i, p_j) = H_i H_j \cdot \sqrt{2\pi} \cdot \frac{\sigma_i \sigma_j}{\sqrt{\sigma_i^2 + \sigma_j^2}} \exp\left(\frac{1}{2\sigma_i^2 \sigma_j^2}(\alpha-\beta)\right) \tag{3}$$

where

https://dx.doi.org/10.1021/acs.analchem.0c03833
Anal. Chem. 2020, 92, 16138−16148

$$\alpha = \frac{(\sigma_j^2 \mu_i + \sigma_i^2 \mu_j)^2}{\sigma_j^2 + \sigma_i^2}$$

and

$$\beta = \sigma_j^2 \mu_i^2 + \sigma_j^2 \mu_j^2$$

The similarity score, B, between two spectra, S_i and S_j, is the sum overlap of all matched peaks

$$B(S_i, S_j) = \sum_{|\mu_i - \mu_j| < \epsilon} I(p_i, p_j) \tag{4}$$

To compute the similarity score between a pair of spectra, the set of peak pairs that satisfy the condition in eq 4 must be found. A peak in the first spectrum is matched to one in the second spectrum if their m/z locations are within a small distance threshold, ϵ, of each other. Note that a peak in one spectrum can be matched to multiple peaks in the other spectrum. The threshold ϵ in eq 4 is proportional to the modeled peak width and therefore increases with m/z.

With our definition of pairwise similarity between two centroided spectra, we can search for an optimal warping from a set of candidate warpings in a similar way as the original COW implementation. This involves dividing the m/z range into segments, evaluating B for all candidate warpings for each segment independently, and then finding the optimal combination of segment warpings. The analytical form of the integral in eq 2 enables fast computation of the similarity score, which is critical since it must be repeated for each candidate warping. Here, we denote the lower and upper edges of an m/z segment n_l and n_r, respectively, and refer to them as the warping nodes of the segment. Note that each warping node, except those at the lowest and highest ends of the m/z range, are shared by two segments. To generate the set of candidate warpings for an m/z segment, the warping nodes at its edges are shifted a fixed number of steps upward and downward in m/z. The set of warpings for that segment then corresponds to all possible combinations of shifts of n_l and n_r. Computing B for a particular warping and segment is then performed by warping the peaks in the segment and then computing B with the warped peaks and the peaks from the corresponding segment in the reference spectrum. Peaks are warped by updating their mass with linear interpolation according to

$$\mu' = x \cdot (n_r' - n_l') + n_l' \tag{5}$$

where

$$x = (\mu - n_l)/(n_r - n_l)$$

μ is the original peak mass, μ' is the warped peak mass, and n_l' and n_r' are the shifted positions of n_l and n_r, respectively. Note that while segments are compressed, stretched, and/or shifted, a peak is never warped out of its segment and its width and height are left untouched. Finally, like in the original COW implementation,[17] the optimal combination of warping node moves is found with dynamic programming. The shift of a warping node is bounded by the slack parameter: $|n' - n| \leq m_n$. The slack is reflected by the amplitude of the warping function and should be sufficiently large to capture the largest shifts. Like the peak matching threshold (ϵ in eq 4), m is proportional to FWHM and is computed for each warping node individually.

By maximizing our similarity score, we find a piecewise linear mass recalibration function with a degree of freedom determined by the number of warping nodes. A large number of short segments gives a flexible warping function that can correct local m/z shifts, whereas a small number of long segments generally results in a more stable, but less flexible, warping function. The risk of overfitting the warping function to noise or random variations in peak mass is smaller with segments that have many matched peaks. Due to this, we prefer long segments that accommodate a sufficient number of peak matches (at least 10−20) to short segments that are potentially more flexible.

Placement of Warping Nodes. In many MSI data sets, there is a large variation in peak density across the m/z range. For such data sets, the placement of the warping nodes can greatly influence the warping quality. The same warping nodes can be used for all spectra, or they can be placed uniquely for each spectrum. The goal of the warping node placement is to have a segment length that is adapted to the amount of shared information, i.e., peak matches, between the sample and reference spectra in all parts of the m/z range. This can be achieved by generating a density estimate (smooth histogram) of the peak matches between the sample and reference spectra over the m/z range and then placing the warping nodes between the peaks of the density curve. If the warping nodes are uniquely placed for each spectrum, they can alternatively be placed so that the number of peak matches is the same in all segments. A third option is to use segments with uniform lengths. This may work well for spectra with a high peak density throughout the m/z range but can result in segments without peak matches for peak-sparse spectra.

RANSAC Outlier Detection. Unlike alignment methods that rely solely on the difference in mass between matched peaks, MSIWarp is naturally robust to spurious matches, as long as there are sufficiently many true matches. To confirm this, we use a custom RANSAC procedure to detect spurious matches, perform alignment both with the full set of peak matches and with that obtained after having removed spurious matches, and compare the results to evaluate whether spurious matches degrade the alignment quality of MSIWarp in practice. Generally, RANSAC fits a model to a minimal subset of data points that may contain outliers. The subset is resampled numerous times and the model is fit to each subset. The best model, given some criteria, is then selected and all data points that fit the model are included in the "inlier" set. We combine RANSAC with our method to separate true matches (inliers) from spurious matches in the following way:

(i) Generate a list of preliminary peak matches with a permissive distance threshold proportional to peak FWHM.
(ii) Randomly sample two matches from the preliminary set of matches for each segment, and fit a trial warping model to the sampled matches. Warp all other preliminary matches according to the trial model.
(iii) Add peak matches whose mass distance after alignment is below a strict threshold to the inlier set.
(iv) Repeat steps (i)−(iii) n times and return the largest set of inliers. Given an estimate of the fraction of inliers among the peak matches, n can be set to obtain a desired probability of an outlier-free candidate model.

We use ϵ from eq 4 as the threshold in (i) and 0.3 times peak FWHM as that in (iii). When using a large number of

https://dx.doi.org/10.1021/acs.analchem.0c03833
Anal. Chem. 2020, 92, 16138−16148

segments, the warping function found using the subset of peak matches in (ii) is highly unstable and can sometimes fit a large number of spurious matches by chance. Therefore, we use only one or two warping segments in the RANSAC step. After removing the spurious matches, more warping segments can be added in parts of the m/z range that are supported by the number of true matches. The final warping is then searched for using all, or a large fraction, of the true matches.

MATERIALS AND METHODS

Data Sets. To evaluate MSIWarp, we applied it to four publicly available data sets that together represent the most common MSI experimental setups. The first data set was generated from two mouse kidney sections with a rapifleX MALDI TOF/TOF instrument (Bruker Daltonics).[22] The second data set was generated from human cancer spheroids with an ultrafleXtreme MALDI-TOF/TOF instrument (Bruker Daltonics).[23] The third data set was generated from a rat liver section with a MALDI LTQ Orbitrap XL instrument (Thermo Fischer Scientific).[24] The fourth data set was generated from a colorectal adenocarcinoma sample using a home-built motorized DESI ion source and an LTQ XL Orbitrap Discovery instrument (Thermo Fischer Scientific).[25] The data sets, referred to as the TOF kidney, TOF spheroids, Orbitrap liver, and Orbitrap DESI data sets, are summarized in Table 1. A more detailed description of the data sets is available in the Supporting Information, and total ion current (TIC) images for each data set are shown in Figure S1. Before performing mass alignment, we filtered out peaks whose intensity was below a signal-to-noise ratio (SNR) of 2.5 from the mouse kidney TOF data set and centroided all data sets, except for the Orbitrap DESI, with the parabolic centroiding algorithm by Robichaud et al.[26] We downloaded the Orbitrap DESI data set in centroid mode from the MetaboLights repository. The data sets were preprocessed with in-house Python scripts.

Table 1. Summary of the Data Sets[a]

	TOF kidney	TOF spheroids	Orbitrap liver	Orbitrap DESI
ionization technique	MALDI	MALDI	MALDI	DESI
species	mouse	human	rat	human
no. spectra	33 242	1114	23 823	20 286
raster size (μm)	100	50	50	100
m/z range (Da)	500−2500	800−4500	150−1000	200−1000
resolution	2600	17 500	60 000	60 000
avg. no. peaks	244	57	730	701

[a]Both TOF data sets were uploaded to ProteomeXchange smoothed and with their baseline removed. Resolutions for the Orbitrap data sets were calculated at 400 m/z.

Data Analysis. To measure the effect of alignment in each data set, we calculated the mass dispersion around a set of reference peaks. We obtained the mass dispersion of a reference peak by binning all spectra around its m/z and then calculating the standard deviation of peak masses within the resulting mass bin. By binning we mean isolating peaks across spectra at predefined m/z locations, and we refer to the isolation windows as mass bins. We used a bin width two times the FWHM of the reference peak. As reference peaks, we used the most intense peaks of the mean spectrum (100 for the TOF kidney, Orbitrap liver, and Orbitrap DESI data sets and 50 for the TOF spheroid data set), some matrix peaks, and peaks that were identified in the papers that originally published the data sets. The mean spectrum was generated after alignment with MSIWarp, and we performed the binning and calculated mass dispersion both before and after alignment.

To further assess the quality of the alignment, we generated scatter plots of peak mass and spectrum acquisition time for the mass bin of each reference peak (Figures S3−S13). The scatter plots provide a clear view of the mass shift before and after alignment and serve as valuable quality control for the alignment of specific peaks. Despite the previous intensity filtering of the TOF kidney data set based on SNR, some mass bins were still contaminated with faint background peaks. To reduce the influence of these, we applied an intensity threshold to each mass bin. The threshold was defined as the lower intensity quartile of all of the peaks in the mass bin, and peaks whose intensity was below this threshold were excluded when calculating mass dispersion.

RESULTS AND DISCUSSION

To reiterate, MSIWarp aligns a data set by maximizing the similarity between each spectrum and a common reference spectrum. Like any method that performs pairwise alignment, it relies on shared information. In the ideal case, all spectra share numerous peaks with the reference spectrum throughout the m/z range. In a more challenging case, there are few shared peaks overall and/or wide gaps in the m/z range without any shared peaks. Figure 2 shows a pair of spectra from the Orbitrap data set, another from the TOF kidney data set, and the m/z difference between preliminary matched peaks. The Orbitrap spectra are homogeneous, with shared peaks throughout the m/z range, and the mass shifts are small (<1.5 ppm). In contrast, the TOF spectra are heterogeneous, the mass shifts are significantly larger (>200 ppm), and there is a part of the m/z range with almost no shared peaks (1000−1400 m/z). Note that the mass differences between shared peaks for a pair of aligned spectra are expected to be distributed around zero throughout the m/z range. However, in these examples, the misalignment is apparent; the mass shift of matched peaks consistently increases with m/z for both pairs. To accommodate large mass shifts, we used a peak matching threshold (ϵ in eq 4) of approximately two times the FWHM when matching peaks. With the low mass resolution in the TOF kidney data set, this meant matching peaks within a window of $\pm$0.76 m/z at 1000 m/z. A wide matching window increases the number of spurious matches; the scatter in Figure 2b contains numerous examples, most notably those clustered around 850 and 1050 m/z. Finally, we chose to place the warping nodes based on the density estimate of peak matches, since it generally resulted in a smoother partitioning of the m/z range than when placing them so that all segments contained the same number of peak matches. We generated the density estimate by performing a Kernel density estimation (KDE) of peak m/z and then placed the warping nodes between the peaks of the density curve, resulting in 4−10 warping segments for the four data sets. Increasing and decreasing the bandwidth of the KDE is a flexible way to adjust the number of warping segments and the number of peak matches in each segment. We used a bandwidth of 15 Da for the TOF kidney and Orbitrap data sets, and a bandwidth of 100 Da for the TOF

https://dx.doi.org/10.1021/acs.analchem.0c03833
Anal. Chem. 2020, 92, 16138−16148

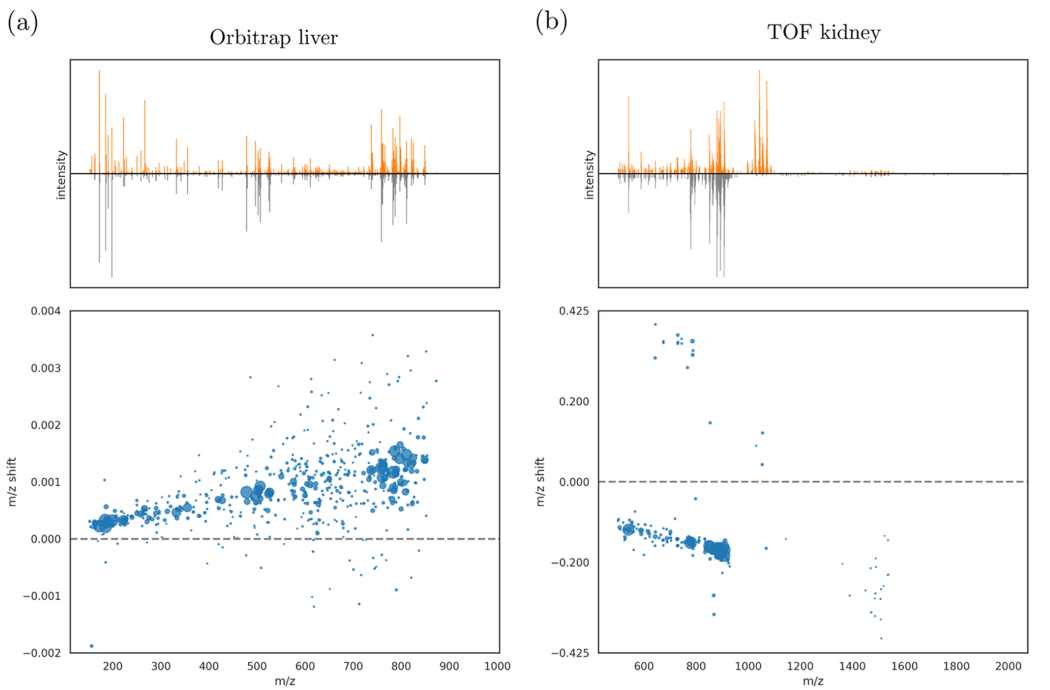


Figure 2. Top: pair of raw spectra from the Orbitrap liver data set (a) and another from the TOF kidney (b). Bottom: scatters of m/z difference between matched peaks with point size scaled by intensity. The Orbitrap spectra share peaks across the entire m/z range. In contrast, the TOF spectra share no peaks in a large part of the m/z range (1000−1400 m/z). This example also highlights the severity of the mass shift in the TOF kidney data set: at 800 m/z, the shift is close to 0.18 m/z (219 ppm) between the TOF pair compared to 0.001 m/z (1.25 ppm) between the Orbitrap pair.

spheroid data set since it was significantly less peak-dense than the other data sets.

Spurious peak matches are largely inconsequential to MSIWarp as long as spectra are reasonably peak-dense; in fact, most spectra are aligned accurately and reliably without RANSAC, even those in the TOF data sets. We hypothesized that the importance of identifying true matches increases when aligning peak-sparse spectra. For this reason, we aligned the TOF data sets both with and without RANSAC. When combining RANSAC with COW, it is important to use a small number of warping segments in this step to avoid overfitting the candidate models to spurious peak matches; here, we used two segments for both TOF data sets in the RANSAC step. Manual inspection of scatter plots like those in Figure 3 suggests that RANSAC confidently filters out spurious peak matches in almost all spectra across both TOF data sets. We used RANSAC to filter out spurious peak matches before searching for the optimal warping. Then, we added more warping nodes in the peak-dense parts of the m/z range to gain more flexibility. Thereby, we obtained a flexibility in the warping function that was adapted to the number of shared peaks throughout the m/z range. We provide some examples of how RANSAC finds the true peak matches (Figure S2) along with an animation in the Supporting Information.

When aligning a data set by aligning each spectrum to a common reference spectrum, the quality of that reference spectrum is essential. We tried an approach similar to that of Kulkarni et al.,[11] where the reference spectrum is continuously updated with each aligned sample spectrum, but observed no significant improvement over aligning against a constant reference spectrum of high quality. The spectrum with the highest TIC was sufficient in terms of peak coverage for all four data sets, and we therefore chose to use it as a reference. Although the spectra with the highest TIC were appropriate references for the alignment of the data sets that we discuss here, a composite spectrum may be needed to fully cover the m/z range in other data sets.

How we dealt with segments without shared peaks also deserves mention. We chose to interpolate the warping function in empty regions, as is evident in Figure 3b at around 1200 m/z. This is reasonable under the assumption that some relevant data set peaks are missing in the reference spectrum, which is often the case, so they can be present in a sample spectrum without being shared with the reference. An alternative approach is to leave the empty parts of the m/z region unaligned, which can be more appropriate if the reference spectrum has a very high peak coverage throughout the m/z range. When this is the case, the sample spectrum likely has little or no information in regions where it does not share peaks with the reference spectrum, and aligning those regions is unnecessary.

https://dx.doi.org/10.1021/acs.analchem.0c03833
Anal. Chem. 2020, 92, 16138−16148

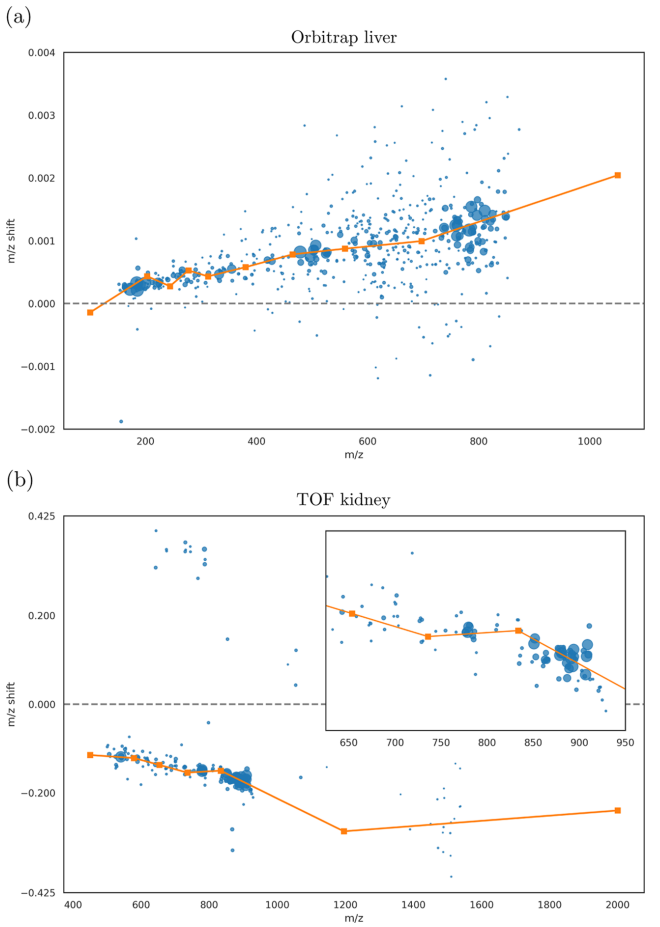


Figure 3. (a, b) Mass shift estimated by MSIWarp (orange line) overlaid on the peak match scatter from Figure 2. (b) We use more warping nodes in the peak-dense part of the m/z range than in the peak-sparse part; the zoom-in shows that the warping function closely follows the local shifts between 700 and 900 m/z.

Reduction in Mass Misalignment. After alignment with MSIWarp, mass dispersion is reduced considerably in all four data sets. In Table 2, the mass dispersions of the mean spectrum peaks before and after alignment are reported in both ppm and FWHM. The full list of dispersions of the mean spectrum peaks is available in the Supporting Information spreadsheet. Interestingly, we observed no significant improvement when aligning the TOF data sets with RANSAC compared to aligning it without RANSAC. This suggests that the inherent robustness of COW is sufficient in dealing with spurious peak matches for the majority of spectra. To assess the sensitivity of MSIWarp to the modeled peak width, σ, and the peak matching threshold, ϵ, we reran the analysis of the Orbitrap liver and TOF kidney data sets for various values of these parameters (Table S1 and S2). This parameter sensitivity analysis suggests that MSIWarp can perform well even when σ over- or underestimates the experimental peak width by a factor of up to 2. It also suggests that a large peak matching threshold is better than a small one, which is further evidence that MSIWarp is robust to spurious peak matches and that the most important factor is that true matches are captured

Table 2. Median Mass Dispersion of Mean Spectrum Peaks before (Raw) and after Alignment (Warped) for the Four Data Sets

	TOF kidney	TOF spheroids	Orbitrap liver	Orbitrap DESI
raw (ppm)	106.39	35.63	0.63	0.52
warped (ppm)	12.73	6.48	0.18	0.17
raw (FWHM)	0.27	0.63	0.03	0.03
warped (FWHM)	0.03	0.11	0.01	0.01
reduction (%)	88.03	81.82	72.03	68.00

https://dx.doi.org/10.1021/acs.analchem.0c03833
Anal. Chem. 2020, 92, 16138−16148

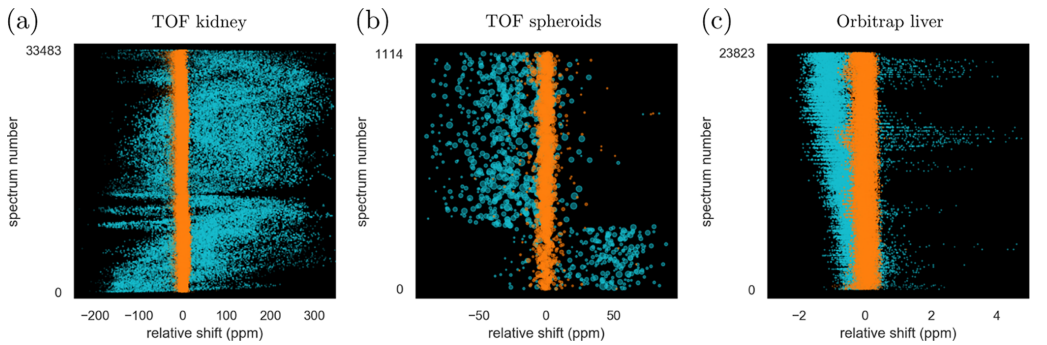


Figure 4. Scatter plots of mass shift relative to the reference peak (*y*-axis) and spectrum index ordered according to acquisition time (*y*-axis) before (cyan) and after alignment (orange). (a−c) Scatters around reference peaks at m/z 850.80 (PI 36:7), 1403.10 (unknown), and 172.04 (matrix) in the TOF kidney, TOF spheroids, and Orbitrap liver data sets, respectively.

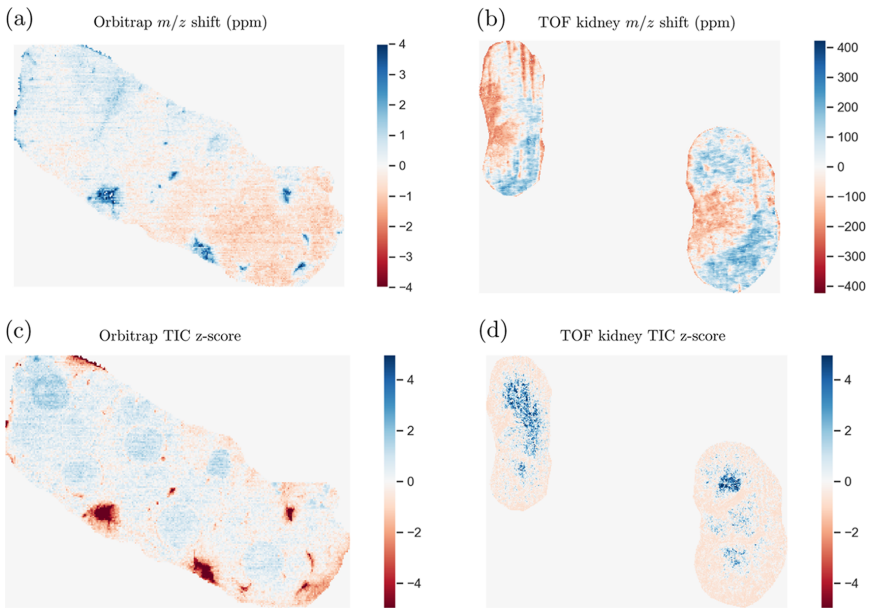


Figure 5. Mass shift and TIC images from the Orbitrap liver and TOF kidney data sets. Left: the mass shift (a) of the matrix peak at 172.04 m/z in the Orbitrap data set is correlated to TIC (c). Right: the mass shift (b) of the lipid peak (PI 38:2) in the TOF kidney data set appears to be related to tissue structures rather than to TIC (d) or acquisition time.

throughout the m/z range. In addition to using the spectrum with the highest TIC as a reference, we also aligned the TOF kidney data set to each of the 11 spectra with TICs closest to the data set median, resulting in median mass dispersions of mean spectrum peaks between 11.64 and 16.10 ppm (avg. 12.53). For 9 out of 11 spectra, the median mass dispersion was lower than when using the spectrum with the highest TIC as a reference (12.73 ppm). In comparison, using the spectrum with the lowest TIC as a reference resulted in a median mass dispersion of 73.45 ppm. Altogether, this indicates that a minimum TIC threshold can be used to filter out unsuitable reference spectra, but otherwise, the TIC provides little or no information about a spectrum's suitability as a reference.

Visualization of how the peak mass varies across the experiment gives a good overview of alignment; Figure 4a−4c shows the scatter of the relative peak mass before and after alignment with MSIWarp and spectrum acquisition time for three example peaks. The same pattern is seen for the three peaks: aligning with MSIWarp visibly tightens the peak scatter and removes the systematic shifts related to spectrum acquisition time. These scatter plots are effective in visualizing time-dependent mass shift but do not provide a clear picture of how mass shift relates to tissue location. To better visualize this

https://dx.doi.org/10.1021/acs.analchem.0c03833
Anal. Chem. 2020, 92, 16138−16148

relationship, we show the relative mass shift of a matrix peak ($[M - H_2O + H]^+$, 172.04 m/z) from the Orbitrap data set as a function of tissue location in Figure 5a. In this plot, it is evident that the peak masses decrease with time (the tissue was scanned from top to bottom), but also that some shifts are related to tissue location. The blue spots in the bottom half of the section break the trend related to acquisition time; in these spots, peak masses are increased by more than 4 ppm, compared to an average decrease of approximately 1 ppm in the bottom half of the section. The blue spots appear to be correlated to the TIC of the spectra, which could be due to the space-charge effect. Figure 5b shows the mass shift image of the peak at 890.8 m/z (PI 38:2) from the TOF kidney data set. In contrast to the matrix peak from the Orbitrap data set, the mass shift of this peak appears to be unrelated to both spectrum acquisition time and TIC (Figures 4a and 5b). Instead, there is a strong relationship to tissue location: in both kidney sections, peak masses appear to be increased in the cortex and decreased in the medulla. The "stripes" at the top part of the left section are likely an experimental artifact (due to tissue folding or damage) rather than being related to any tissue structure.

The mass dispersion of some identified compounds and matrix peaks in the Orbitrap liver and TOF kidney data sets are shown in Tables 3 and 4, respectively. The monoisotopic peak of the phosphatidylcholine head group (184.07 m/z) in the Orbitrap data set has a notably lower dispersion after alignment (0.036 ppm) than those of the other compounds (0.086−0.435 ppm). The intensity of this peak is several orders of magnitude larger than that of all other peaks. As a consequence, it dominates the similarity score and effectively acts as a lock mass for the warping function. Importantly, this does not appear to increase dispersion for lower intensity peaks close in m/z; the matrix peaks at 172.04 and 190.05 m/z contribute insignificantly to the similarity score in comparison, but still have lower dispersion than most other peaks after alignment. This suggests that the shift in low-intensity peaks can be estimated accurately with the shift of nearby high-intensity peaks. In the TOF kidney data set, dispersion is reduced consistently by more than 80 percent, except for the peak at m/z 919.90 (PI 40:1), whose dispersion remains relatively high after alignment (32.69 ppm). By looking at the mass scatter of this peak, however, it is evident that it has been mixed with another peak in the same mass bin and that this causes the relatively high dispersion after alignment (Figure S5, Supporting Information).

Table 3. Mass Dispersion (ppm) of Five Matrix (α-Cyano-4-hydroxycinnamic Acid) Peaks, the Monoisotopic Peak of Two Spiked-In Compounds, and the Peak of Phosphocholine before (Raw) and after Alignment (Warped) in the Orbitrap Liver Data Set

compound	m/z	disp. raw (ppm)	disp. warped (ppm)
$[M - H_2O + H]^+$	172.039	0.644	0.159
$[M + H]^+$	190.050	0.743	0.086
$[M + Na]^+$	212.032	0.724	0.222
$[M - H + 2Na]^+$	234.014	0.998	0.435
$[2M + H]^+$	379.092	0.699	0.343
phosphocholine	184.073	0.706	0.036
ipratropium	332.222	0.675	0.271
dasatinib	488.164	0.475	0.364

Table 4. Mass Dispersion (ppm) of Some Identified Lipid Peaks before (Raw) and after Alignment (Warped) in the TOF Kidney Data Set[a]

compound	m/z	disp. raw (ppm)	disp. warped (ppm)
LPS 18:0	524.19	126.83	15.24
PA 34:1	673.68	128.19	24.99
PA 36:1	701.76	123.15	16.80
PS 36:3	784.38	118.82	16.13
PI 36:7	850.80	103.38	5.70
PS 42:5	864.82	105.18	7.23
PI 40:6	907.87	103.49	6.51
PI 40:1	919.90	87.83	32.69

[a]All lipids are represented as $[M - H]^-$ ions.

An interesting example of the utility of MSIWarp is shown in Figure 6. Like the mass bin at 919.90 m/z, other mass bins in the TOF kidney data set contain mixed peaks as well, including that of the unidentified reference peak at 904.90 m/z. Before alignment, the two peaks in this bin are indistinguishable, but after alignment, the peaks are separated sufficiently to enable us to generate a distinct ion image for each. The distribution of peak mass within the bin is considerably narrower after alignment (Figure 6b). Crucially, the density curve of the warped peak masses reveals, albeit just barely, the second peak as distinct from the first one. By re-binning the spectra with two tighter windows (±10 ppm), whose respective centers are at the main and shoulder peaks of the density estimate, two distinct ion images can be generated (Figure 6d,6e) instead of one in which the two compounds are mixed (Figure 6c). Despite the low mass resolution of the TOF kidney data (the FWHM is approximately 380 ppm), these two peaks, which are 25 ppm apart, can still be separated by looking at the mass locations of their centroids. Note that while we separate them on a data set level, we do not separate them in a single spectrum; when the compounds corresponding to the peaks are present in the same spectrum, i.e., tissue spot, the peak of the more intensive compound masks that of the other. This is evident in the ion images of the two peaks (Figure 6d,6e): only the more intense peak (904.878 m/z) is visible in the pixels where the two peaks overlap.

Implementation and Processing Time. Although our method involves repeating the similarity score computation for a large number of candidate warpings to align a single spectrum, we keep the processing time for a whole data set low by implementing the core part of MSIWarp in efficient C++. Aligning the TOF kidney data set, consisting of 33 242 spectra with 244 peaks on average, took approximately 150 s when MSIWarp was run without RANSAC and in parallel mode on a laptop with an Intel i7-6700HQ CPU (2.6 GHz) and 16 GB RAM. Aligning the Orbitrap liver data set with the same settings took approximately 300 s. The parameter that has the largest impact on processing time is the number of steps by which the warping nodes are shifted when searching for the optimal warping. Given that the slack has been set to capture the mass shift for all spectra, the number of steps corresponds to the resolution of the warping function. We found that a step size of 0.05−0.10 times the peak FWHM gives a sufficient alignment resolution. If the search space of the warping function is ±2 FWHM, this results in 40−80 steps for each warping node. The source code of MSIWarp along with Python bindings are available at https://github.com/horvatovichlab/MSIWarp. Our goal is to make it possible to

https://dx.doi.org/10.1021/acs.analchem.0c03833
Anal. Chem. 2020, 92, 16138−16148

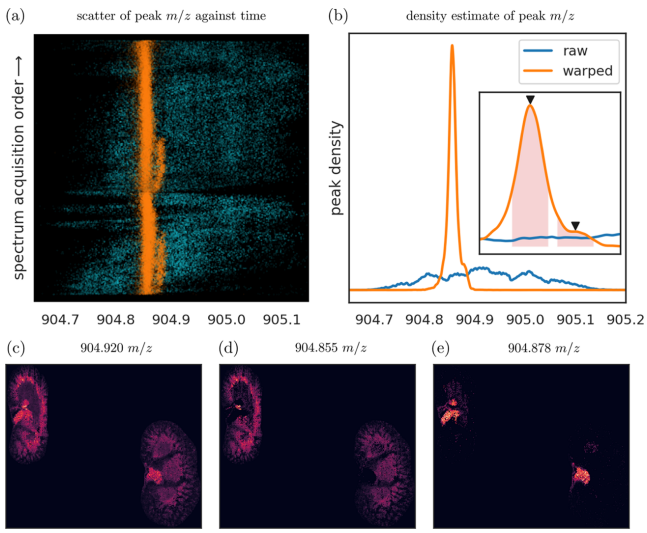


Figure 6. Mass bin from the TOF kidney data set exemplifies how severe misalignment leads to mixed ion images. The scatter of peak mass and acquisition time in (a) reveals two peaks after alignment (orange) that are indistinguishable before alignment (cyan). Alignment similarly reveals the two peaks in the density estimate of peak mass within the bin (b): before alignment (raw), the variation in peak mass across the spectra is too large to separate the two peaks, but after alignment (warped), the peaks appear as the main and shoulder peaks of the density curve. The left ion image (c) was generated from the raw (unaligned) spectra with a wide bin window (±200 ppm), while the center (d) and right (e) ion images were generated by binning the warped spectra with a narrow window (±10 ppm) around 904.855 and 904.878 m/z, respectively. The median mass of the two peaks and the narrow windows are marked in the zoom-in on the density curve in (b). The ion images in (c)−(e) were generated by summing peak intensities in the bin for each spectrum/pixel.

interface MSIWarp with existing MSI analysis packages such as MALDIquant,[27] Cardinal,[28] and rMSIproc.[29]

CONCLUSIONS

We have presented an approach, MSIWarp, that readily improves relative alignment in both TOF and Orbitrap data sets that together represent a large variety of MSI experimental setups. Even the severe misalignment in the TOF kidney data set is brought down to a level that enables separation of peaks close in m/z. With a median mass dispersion of 6.48 ppm (and below 5 ppm for more than half of the peaks) in the TOF spheroid data set, MSIWarp matches the alignment performance of methods that rely on profile spectra using only centroided spectra. While the largest improvements in relative mass alignment can be gained for TOF spectra, our results suggest that MSIWarp can further improve the already high mass alignment in Orbitrap data sets. By investigating the effect of spurious peak matches with RANSAC, we have also shown that MSIWarp is robust and performs well even for most peak-sparse spectra. Finally, we believe that a careful assessment of mass alignment is critical when analyzing MSI data sets. Tools such as scatter plots of peak mass and acquisition time, scatter plots of mass shift between individual pairs of spectra, and images of mass shift as a function of tissue location for individual peaks provide a way to do this in a simple manner.

Although MSIWarp has demonstrated significant benefits in analyzing MSI data sets, there are several research paths that could yield additional improvements. Currently, the output of MSIWarp is an m/z recalibration function for each data set spectrum. It is important to highlight that even though this function is found by searching for the optimal alignment between a pair of centroided spectra, it can also be used to align the corresponding profile spectra. Conceptually, the recalibration function could also be found by directly computing the correlation integral between the profile spectra, if available, instead of using our analytical expression based on the overlap of centroided peaks. This would allow MSIWarp to utilize subtle features in the profile data, such as peak shape, that may be lost in the peak-picking step. Another possible enhancement is to create a hybrid approach by combining MSIWarp with existing calibration strategies to improve absolute mass accuracy in addition to relative mass alignment. To do this, a spectrum with accurate masses should be used as a reference, and if no such spectrum is available, the reference spectrum can be chosen based on other criteria and calibrated prior to alignment. Together, these approaches represent a rich area of research that would allow interesting comparisons of related techniques and potential for even further improvements in performance.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c03833.

Description of the TOF kidney, TOF spheroids, Orbitrap liver, and Orbitrap DESI data sets. Tables S1 and S2, median mass dispersions of mean spectrum peaks from the TOF kidney and Orbitrap liver data sets after alignment with various values of σ and ϵ. Figure S1,

https://dx.doi.org/10.1021/acs.analchem.0c03833
Anal. Chem. 2020, 92, 16138−16148

TIC images for the four MSI data sets. Figure S2, RANSAC outlier detection results for three example spectra from the TOF kidney data set. Figures S3−S13, scatter plots of peak mass and acquisition time for mean spectrum peaks from the top 100, 50, 100, and 100 mean spectrum peaks' TOF kidney, TOF spheroids, Orbitrap liver and Orbitrap DESI data sets (PDF)
Ransac outlier detection animation (ZIP)
Mean spectrum dispersions (XLSX)

AUTHOR INFORMATION

Corresponding Author

Peter Horvatovich − *Department of Biomedical Engineering, Lund University, Lund 221 00, Sweden; Department of Analytical Biochemistry, Groningen Research Institute of Pharmacy, University of Groningen, 9713 AV Groningen, The Netherlands;* orcid.org/0000-0003-2218-1140; Email: p.l.horvatovich@rug.nl

Authors

Jonatan O. Eriksson − *Department of Biomedical Engineering, Lund University, Lund 221 00, Sweden*

Alejandro Sánchez Brotons − *Department of Analytical Biochemistry, Groningen Research Institute of Pharmacy, University of Groningen, 9713 AV Groningen, The Netherlands*

Melinda Rezeli − *Department of Biomedical Engineering, Lund University, Lund 221 00, Sweden;* orcid.org/0000-0003-4373-5616

Frank Suits − *IBM Research - Australia, Southbank, VIC 3006, Australia*

György Markó-Varga − *Department of Biomedical Engineering, Lund University, Lund 221 00, Sweden*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c03833

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We acknowledge the support from Fru Berta Kamprads Stiftelse. This research was part of the Netherlands X-omics Initiative and partially funded by NWO, project 184.034.019. We also thank the reviewers for their valuable feedback.

REFERENCES

(1) Buchberger, A. R.; DeLaney, K.; Johnson, J.; Li, L. *Anal. Chem.* **2018**, *90*, 240−265.
(2) Gilmore, I. S.; Heiles, S.; Pieterse, C. L. *Annu. Rev. Anal. Chem.* **2019**, *12*, 201−224.
(3) Han, J.; Permentier, H.; Bischoff, R.; Groothuis, G.; Casini, A.; Horvatovich, P. *TrAC, Trends Anal. Chem.* **2019**, *112*, 13−28.
(4) Ryan, D. J.; Spraggins, J. M.; Caprioli, R. M. *Curr. Opin. Chem. Biol.* **2019**, *48*, 64−72.
(5) Verbeeck, N.; Caprioli, R. M.; Van de Plas, R. *Mass Spectrom. Rev.* **2020**, *39*, 245−291.
(6) Ràfols, P.; del Castillo, E.; Yanes, O.; Brezmes, J.; Correig, X. *Anal. Chim. Acta* **2018**, *1022*, 61−69.
(7) Gorshkov, M. V.; Good, D. M.; Lyutvinskiy, Y.; Yang, H.; Zubarev, R. A. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1846−1851.
(8) Kharchenko, A.; Vladimirov, G.; Heeren, R. M.; Nikolaev, E. N. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 977−987.
(9) Boskamp, T.; Lachmund, D.; Casadonte, R.; Hauberg-Lotte, L.; Kobarg, J. H.; Kriegsmann, J.; Maass, P. *Anal. Chem.* **2020**, 1301.
(10) Bocker, S.; Makinen, V. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2008**, *5*, 91−100.
(11) Kulkarni, P.; Kaftan, F.; Kynast, P.; Svatoš, A.; Böcker, S. *Anal. Bioanal. Chem.* **2015**, *407*, 7603−7613.
(12) Tracy, M. B.; Chen, H.; Weaver, D. M.; Malyarenko, D. I.; Sasinowski, M.; Cazares, L. H.; Drake, R. R.; Semmes, O. J.; Tracy, E. R.; Cooke, W. E. *Proteomics* **2008**, *8*, 1530−1538.
(13) Barry, J. A.; Robichaud, G.; Muddiman, D. C. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 1137−1145.
(14) Alexandrov, T. *Annu. Rev. Biomed. Data Sci.* **2020**, *3*, 61.
(15) Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; De Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; et al. *Nucleic Acids Res.* **2013**, *41*, D781−D786.
(16) Alexandrov, T.; Ovchnnikova, K.; Palmer, A.; Kovalev, V.; Tarasov, A.; Stuart, L.; Nigmetzianov, R.; Fay, D.; Contributors, K. M. *BioRxiv* **2019**, No. 539478.
(17) Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard, J. *J. Chromatogr. A* **1998**, *805*, 17−35.
(18) Christin, C.; Smilde, A. K.; Hoefsloot, H. C.; Suits, F.; Bischoff, R.; Horvatovich, P. L. *Anal. Chem.* **2008**, *80*, 7012−7021.
(19) Suits, F.; Lepre, J.; Du, P.; Bischoff, R.; Horvatovich, P. *Anal. Chem.* **2008**, *80*, 3095−3104.
(20) Fischler, M. A.; Bolles, R. C. *Commun. ACM* **1981**, *24*, 381−395.
(21) Suits, F.; Hoekman, B.; Rosenling, T.; Bischoff, R.; Horvatovich, P. *Anal. Chem.* **2011**, *83*, 7786−7794.
(22) Noh, S. A.; Kim, S.-M.; Park, S. H.; Kim, D.-J.; Lee, J. W.; Kim, Y. G.; Moon, J.-Y.; Lim, S.-J.; Lee, S.-H.; Kim, K. P. *J. Proteome Res.* **2019**, *18*, 2803−2812.
(23) Mittal, P.; Price, Z. K.; Lokman, N. A.; Ricciardelli, C.; Oehler, M. K.; Klingler-Hoffmann, M.; Hoffmann, P. *Proteomics* **2019**, *19*, No. 1900146.
(24) Eriksson, J. O.; Rezeli, M.; Hefner, M.; Marko-Varga, G.; Horvatovich, P. *Anal. Chem.* **2019**, *91*, 11888−11896.
(25) Gerbig, S.; Golf, O.; Balog, J.; Denes, J.; Baranyai, Z.; Zarand, A.; Raso, E.; Timar, J.; Takats, Z. *Anal. Bioanal. Chem.* **2012**, *403*, 2315−2325.
(26) Robichaud, G.; Garrard, K. P.; Barry, J. A.; Muddiman, D. C. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 718−721.
(27) Gibb, S.; Strimmer, K. *Bioinformatics* **2012**, *28*, 2270−2271.
(28) Bemis, K. D.; Harry, A.; Eberlin, L. S.; Ferreira, C.; van de Ven, S. M.; Mallick, P.; Stolowitz, M.; Vitek, O. *Bioinformatics* **2015**, *31*, 2418−2420.
(29) Ràfols, P.; Heijs, B.; del Castillo, E.; Yanes, O.; McDonnell, L. A.; Brezmes, J.; Pérez-Taboada, I.; Vallejo, M.; García-Altares, M.; Correig, X. *Bioinformatics* **2020**, *36*, 3618−3619.

Proteogenomic and Histopathologic Classification of Malignant Melanoma Reveal Molecular Heterogeneity Impacting Survival

Magdalena Kuras[1*], Lazaro Hiram Betancourt[2*], Runyu Hong[3*], Jimmy Rodriguez[4], Leticia Szadai[5], Peter Horvatovich[6], Indira Pla[1], Jonatan Eriksson[7], Beáta Szeitz[8], Bartłomiej Deszcz[9], Charlotte Welinder [2?], Yutaka Sugihara[7], Henrik Ekedahl[2],[10], Bo Baldetorp[2], Christian Ingvar[10,11], Håkan Olsson[2,10], Lotta Lundgren[2,10], Göran Jönsson[2], Henrik Lindberg[7], Henriett Oskolas[2], Zsolt Horvath[7], Melinda Rezeli[7], Jeovanis Gil[7], Roger Appelqvist[7], Johan Malm[1], Aniel Sanchez[1], Marcell Szasz[12.] Krzysztof Pawłowski[1,9,13] Elisabet Wieslander[2**], David Fenyö[3**], Istvan Nemeth[5**], György Marko-Varga[7,14,15**]

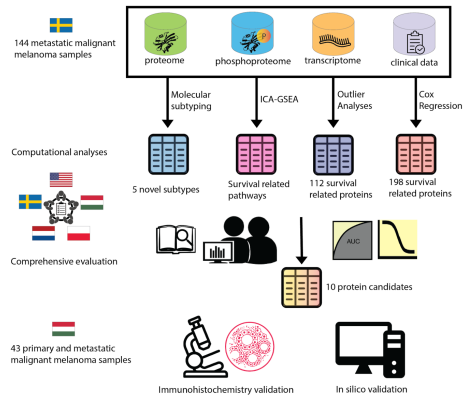
*=First shared authorship
**=Last shared authorship

[1] Section for Clinical Chemistry, Department of Translational Medicine, Lund University, Skåne University Hospital Malmö, Malmö, Sweden
[2] Division of Oncology, Department of Clinical Sciences Lund, Lund University, Lund, Sweden
[3] Institute for Systems Genetics, New York University Grossman School of Medicine, New York, USA.
[4] Department of Biochemistry and Biophysics, Karolinska Institute, Stockholm, Sweden
[5] Department of Dermatology and Allergology, University of Szeged, Szeged, Hungary
[6] Department of Analytical Biochemistry, Faculty of Science and Engineering, University of Groningen, Groningen, The Netherlands
[7] Clinical Protein Science & Imaging, Biomedical Centre, Department of Biomedical Engineering, Lund University, Lund, Sweden
[8] Division of Oncology, Department of Internal Medicine and Oncology, Semmelweis University, Budapest, Hungary
[9] Department of Biochemistry and Microbiology, Warsaw University of Life Sciences, Warszawa, Poland
[10] SUS University Hospital Lund, Lund, Sweden
[11] Department of Surgery, Clinical Sciences, Lund University, SUS, Lund, Sweden
[12] Department of Bioinformatics, Semmelweis University, Budapest, Hungary
[13] Department of Molecular Biology, University of Texas Southwestern Medical Center, Texas, USA
[14] Chemical Genomics Global Research Lab, Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, Seoul, Republic of Korea
[15] 1st Department of Surgery, Tokyo Medical University, Tokyo, Japan

Highlights

- Proteomics classification of metastatic melanoma defines subtypes linked to survival
- Composition of tumor microenvironment is linked to patient outcome
- Proteogenomics of metastatic melanoma reveals biomarkers related to patient survival
- Disease association confirmed by immunohistochemistry in an independent cohort

SUMMARY



Study aims

The study aimed to unify melanoma proteomics, phosphoproteomics, transcriptomics with in-depth histopathology analysis, and relate these to clinical variables, in particular survival parameters. Thus, molecular features were sought that allow prediction of patient outcome.

INTRODUCTION

Malignant melanoma is the most aggressive type of skin cancers and it has high metastatic potential (Erdmann et al., 2013; Grzywa et al., 2017). The worldwide incidence of melanoma has increased rapidly during the last 50 years and it is anticipated to increase profoundly in the next 20 years according to the GLOBOCAN database (Erdmann et al., 2013). UV-radiation seems to be the main cause of malignant melanoma development and its significance is even more apparent in those countries where the malignant melanoma occurrence is above the world average, such as Australia, US, Canada and Scandinavia. Due to its heterogeneous nature and rapid metastatic progression, malignant melanoma poses a major challenge to the healthcare system (Grzywa et al., 2017). There has been a huge development of treatment options from the early less effective practices to the current kinase and immune checkpoint inhibitor therapies (Hugo et al., 2016;

Jannin et al., 2019; Leonardi et al., 2018; Sullivan et al., 2019). The latest generation of drugs, with new mechanisms of action, has had a profound impact on modern healthcare resulting in nearly doubling the five-year survival rates for patients with disseminated disease. Preventing tumor recurrence is, however, still a struggle (Hamid et al., 2019; Robert et al., 2019).

In 2015, The Cancer Genome Atlas (TCGA) published a study combining genomics and transcriptomics for improving clinical management based on discriminating features within mutation and transcriptomic subtypes (Cancer Genome Atlas Network, 2015). The genomic subtypes alone were not able to predict patient outcome, however the immune transcriptomic subtype with higher amount of infiltrating immune cells was associated with a more favorable prognosis and a high overall response rate to inhibitors of the PD-1/PD-L1 pathway. Nevertheless, these results alone are insufficient to describe the alterations responsible for melanoma recurrence and response to treatment. Another transcriptomics-based classification was proposed by Jonsson and co-workers (Cirenajwis et al., 2015). To increase the understanding of how changes in the molecular landscape are influencing patient survival, we have performed comprehensive protein profiling of malignant melanoma metastases from 142 well-characterized patient samples (with support from the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC)). Integration of quantitative transcriptomic and proteomic data (including posttranslational modifications) was used to explore changes in the proteomic landscape during disease progression, and to identify novel protein markers for predicting progression and survival of melanoma patients.

RESULTS

Establishment of a high-resolution proteogenomic map of treatment-naïve lymph node melanoma metastases

In this work, 144 metastases of lymph nodes (128), cutaneous (1), subcutaneous (7), visceral (3) and uncharacterized (3) origins were obtained from metastatic melanoma patients. The clinical and histopathological features of the tumors are summarized in the Supplement. Each sample was submitted to global proteomic and phosphoproteomic analysis. In addition, relative levels of mRNA were extracted from a previous study which involved a larger set of melanoma samples, including those used in this study (Cirenajwis et al., 2015).

The proteomic and phosphoproteomic analyses identified a total of 12,695 proteins and 45,356 phosphosites with an average of 10,705 proteins and 18,722 phosphosites identified per sample. Across all samples 8,124 proteins and 4,644 phosphosites were commonly quantified (STAR Methods).

Classification of metastatic melanoma subtypes by proteogenomics

Two independent studies have classified melanomas based on transcript expression (Jonsson 2010), (Cancer Genome Atlas Network, 2015). Due to the limited correlation between mRNA and protein levels, we investigated whether proteomic data can provide a further improved melanoma classification, in closer association with clinico-histopathological features.

We performed unsupervised hierarchical consensus clustering analysis of the 3000 proteins with the most variable expression levels, and commonly quantified across 118 samples with tumor content >30%. We identified five major melanoma subtypes that were named extracellular space/region (EC, n=23), extracellular space/region-immune (EC-Im, n=26), mitochondria (Mit, n=30), mitochondria-immune (Mit-Im, n=23), and extracellular space/region-mitochondria (EC-Mit, n=16) according to the top enriched terms revealed after Gene ontology and KEGG pathway analysis performed on the five subtypes.

A previous transcriptomic study performed on a sample set partly overlapping with the set used here (Lund transcriptomics study) (Cirenajwis et al., 2015) classified 110 (93%) of these metastases in four subgroups: high-immune (n=38), pigmentation (n=53), proliferative (n=18) and normal-like (n=1). In addition, TCGA transcriptomic classification has also been proposed for melanoma tumors (Cancer Genome Atlas Network, 2015). Applying this classifier (see Materials and methods) we assigned 93% of the tumors to the three transcriptomics subtypes: immune (n=50), MITF-low (n=32) and keratin (n=28).

The Fisher's exact test showed significant association between the proteomics subtypes and the transcriptomics classifications (Figure 1). Moreover, the proteomic classification displayed a more complex picture of melanoma tumors by revealing subjacent heterogeneities within transcriptomic subtypes. We also found significant associations between both transcriptomics classifications, which provided independent validation of these classification systems.

Proteogenomic signature of MM subtypes

To gain further information about the underlying biology of this melanoma classification, we studied the corresponding differential expression of proteins, phosphosites, and transcripts and performed Gene Ontology enrichment analysis on differentiated molecular clusters between the subtypes (Figure 1).

Association between subtypes and histological and clinical features

Associations between the proteomic subtypes and clinico-histopathological features were also determined using Fisher's exact test. There were no significant differences among the five subtypes regarding tumors with BRAF mutation at V600. In contrast, we found enrichment of the Mit (p-value=0.0410) and Mit-Immune (p-value= 0.0213) subtypes in tumors with NRAS Q61K/R mutations and tumors without mutations in NRAS or BRAF (WT), respectively.

The proteomic subtypes were also distinguished by cellular composition and histological characteristics. The EC-Mit subtype was associated with a lack of melanin pigmentation. The Mit-Immune subtype was associated with higher lymphocyte distribution and higher density. The opposite situation was observed for the Mit subtype which was associated with absence of lymphocytes in the tissue. Overall, both immune subtypes displayed a significantly higher lymphatic score (ANOVA FDR<0.05, Post Hoc Tukey's FDR 0.05) than the rest group of metastases. The tumor cell content was significantly higher in the Mit-immune, Mit and EC-Mit metastases when compared to the EC-Immune and EC subtypes. The adjacent lymph node and the necrosis contents were higher (ANOVA FDR<0.05, Post Hoc Tukey's FDR 0.05) in the EC-Immune and EC subtypes, respectively, compared to the rest of metastasis subtypes, while the content of connective tissue was significantly higher only in EC subtypes compared to the Mit-Immune.

The Mit metastases were more frequently found in patients older than 60 years, while tumors from younger patients were associated with the immune subtypes. Tumors from females were associated with the EC and Mit-Immune subtypes, whereas tumors from males were linked to the EC-Immune and Mit subtypes. Interestingly, we found significant associations between this melanoma classification of metastases and features of the primary tumors. The Mit subtype was associated with primary tumors of NM clinical class. Both immune subtypes were associated with Breslow classes 2 and 3, while Mit subtype was associated with Breslow class 4. Also, the Mit and EC-Immune subtypes were associated with primary tumors localized in the trunk, while the EC-Mit, EC and Mit-Immune subtypes were more associated with primary tumors detected in the extremities. These connections mirrored the association of subtypes and gender described above and pointed to a relationship between primary tumor localization and patient gender as previous studies have suggested.

A statistically significant association was found between the tumors of the Mit-Immune and EC-Immune subtypes and patients at stage 3 of melanoma. On the other hand, the subtypes EC-Mit and Mit were significantly correlated with patients at disease stage 4. Accordingly, clinical features closely dependent on the disease stage were also significantly associated with the melanoma molecular classification. Patients with tumors classified as Mit, EC-Mit or EC had an increased risk of developing distant metastases before 5 months from the time of sample collection (surgery) than patients with tumors belonging to the immune subtypes. The Mit and EC-Mit subtypes were associated with an increased risk of patient death within 3 years from the detection of first metastases or within one year from the first distant metastasis. In contrast, patients with tumors classified as Mit-Immune and EC-Immune were associated with longer times (>1year) from the first distant metastasis to the death of the patient.

An increased risk of death from melanoma (3-year disease specific survival, DSS) was observed for patients with metastases classified as Mit and EC-Mit, compared to both immune subtypes. Particularly, the Mit-Immune subtype was significantly associated with >5 years DSS. The corresponding Kaplan-Meier analysis showed that patients with melanoma metastases of subtypes Mit-Immune and EC-immune had better prognosis while the melanoma subtypes Mit, EC-Mit, and EC are associated with worse patient outcome.

In addition, we determined the association between the different subtypes and the patient's overall survival. An increased risk of death from melanoma (overall survival <5 years) was observed for patients with metastases in the EC-Mit subtype, while the risk significantly decreased for patients with tumors classified as Mit-Immune. These results suggest a connection between early-stage melanoma and patient outcome, where features of primary tumors are transmitted to and also expressed by the metastases, and dependent on the acquired subtype, the disease may become more aggressive, with poor prognosis.

Relating the proteogenomic map to clinical and histological variables

Independent component analysis (ICA) is the optimal feature extraction method

Due to the high dimensionality nature of the multi-omics data, unsupervised feature extraction methods are required to efficiently explore the cohort in order to gain novel systematic understanding of melanoma at the molecular level. Two analyses were conducted for the proteomics, transcriptomics, and phosphoproteomics datasets, applying the two most commonly used unsupervised feature extraction methods: principal component analysis (PCA) and independent component analysis (ICA). To keep the comparison fair, the number of principal components and the number of independent components were the same, which was the number of samples in each dataset. Then, we used the dimensionally reduced representations from both methods to perform association tests with clinical and histopathological features. In general, we observed that more ICA-extracted signatures were significantly associated with various clinical and histopathological features and oftentimes showed higher significance level than PCA-extracted features (Figure 2A).

This comparison demonstrated that ICA could extract larger quantities and higher quality of representational components that correlated with more specific clinical and histopathological features than PCA, suggesting that ICA was the preferred feature extraction and dimensionality reduction method. The outcome that ICA allows to link basis vectors (independent components) with more clinical variables than PCA basis vectors (principal components), which also supports the idea that the multi-omics datasets are better explained as additive subsets of independent non-Gaussian sources rather than pieces of uncorrelated information. The ICA results laid ground for our downstream data analyses.

ICA-based gene set enrichment analyses provided pathway-level understanding of melanoma

Gene set enrichment analysis (GSEA) is a widely used bioinformatics method for finding within a set of genes or proteins significant enrichment in elements belonging to specific biological pathways. In

this study, we used independent components (ICs) associated with clinical and histopathological features as the basis to identify the relationships between clinical features and pathways across our omics datasets (Figure 2B, Supplementary Figure 2A-B). For each of these ICs, unlike in the conventional GSEA, the IC centroids were used to rank the genes/proteins/phospho-sites. Searching against the Reactome pathway database, the enrichment scores (ES) were calculated. The statistical significance of ES of biological pathways were determined by comparing with the null distribution using permutation tests. Significant pathways associated with the ICs were found by using an adjusted p-value threshold at 0.05. Since the ICs were significantly associated with specific clinical or histopathological features, it can be concluded that these pathways were related to the corresponding features. To ensure the consistency of ICs, ICA was performed 100 times for each omics dataset, which was also used as a criterion to evaluate the reliability of associations between the ICs and clinical features.

Using the ICA-GSEA analysis workflow, a variety of pathways were found to be correlated with histopathological and clinical features. The average “tumor cell” percentage of samples was significantly related to 207 Reactome pathways, out of which 27 were supported simultaneously by proteomics and phosphoproteomics data while none was supported by transcriptomics data. Other notable pathways related to “tumor cell” content included Hemostasis, Signal Transduction and Signaling by Rho GTPases that were supported by the largest numbers of ICs, as well as Metabolism of proteins, Post-translational protein modification and Transcription. The “average adjacent lymph node” percentage was significantly related to 147 pathways, out of which eight were supported simultaneously by proteomics, phosphoproteomics, and transcriptomics data. The variable “average necrosis" was related to 150 pathways, among them Extracellular Matrix Organization was supported simultaneously by proteomics, phosphoproteomics and transcriptomics data. When using the strict significance criterion (p-value < 0.00001; at least 50 times in the repeated ICA trials with the same IC) for relationship between omics data and clinical data, no ICs were linked to survival-related variables. This likely reflects the heterogeneity of the melanoma patient cohort and their tumors. Thus, a relaxed criterion (0.005 for at least 30 times in the repeated ICA trials with the same IC) was used to explore omics - survival relationships. The complex relationships between clinical and histopathological variables and biological pathways arising from the combined IC-GSEA analysis were visualised (Figure 2C, Supplementary Figure 2C-D). The graphs show, perhaps not surprisingly, close proximity between certain histopathological variables, such as “average adjacent lymph node” percentage and “average lymphocyte density”, which are known to be closely related, oftentimes by definition. Also, the graph highlights a number of pathways clearly related to a number of histopathology variables, such as TCR signaling and cytokine signaling in the immune system. Clearly, these ”hub” pathways are important features of tumor heterogeneity.

Contribution of proteins to the proteogenomics subtypes revealed by ICA

The ICA on proteomics data also harvested several independent components that were significantly correlated with the 5 molecular subtypes defined by consensus clustering of the proteomics dataset (Figure 2B). Ranking the proteins with these ICs’ centroids, the top 10 contributing proteins for each of these ICs were obtained (Figure 2D). Overall, these proteins were differentially expressed in the molecular subtypes correlating with the ICs, which indicated that ICA were able to capture critical proteomics features that could distinguish the molecular subtypes. Furthermore, this finding also served as a piece of evidence to support this novel molecular subtyping in melanoma.

Relationships between phosphosites related to clinical and histological parameters via ICA

The large numbers of phosphosites related to clinical and histopathological variables via the IC analysis begged the question as to which kinases are likely responsible for the generation of the significant phosphosites and hence involved in melanoma-related processes. The Netphorest / NetworKIN approach (Linding et al., 2008; Miller et al., 2008) uses consensus sequences of known phosphorylation sites and protein-protein interaction networks to predict likely culprit kinases. For example, the variables “survival 6 months” and “average lymphatic score” are significantly related to 3 and 6 phophosite ICs, respectively. For each of these ICs, fifty most strongly related features (phosphosites) were selected, yielding 85 and 86 phosphosites, respectively for the two clinical/histology parameters. Then, each set of phosphosites was subjected to the Netphorest / NetworKIN analysis, which resulted in 57 significant associations of phosphorylation sites with kinases for “survival 6 months" and 136 for “average lymphatic score” variable. Here, the major kinases appearing to be responsible for the phosposites related to histopathological features of the samples, and to survival, are isoforms of CK2 and CDK1 kinases (Supplementary Figure 3E). CK2 is a well-known drug target and a factor affecting drug resistance in melanoma (Zhou et al., 2016). CDK1 belongs to the group of major regulators of the cell cycle and is being investigated for its usefulness in targeted therapy in breast cancer (García-Gutiérrez et al., 2019; Kang et al., 2014). Although there were no very strong sequence preferences in the phosphosites related to survival and histological parameters, the most common feature of these sites was the presence of a proline immediately following the phosphosite. Also, often polar and acidic amino acid residues appeared in the phosphosite motifs, e.g. at position (phospho+2) (Supplementary Figure 3F-G).

Identification of subgroups of BRAF V600E mutated metastases based on proteomic signature and patient survival

In a previous study, different levels of BRAF mutation were associated with different survival rates (Betancourt et al., 2019). The BRAF V600E mutated protein was quantified in sixteen patients with BRAF

mutation (DNA-based detection) and high expression level of this protein was associated with a more aggressive tumor progression and short survival. The protein profile associated with different expression levels of BRAF V600E was linked to biological pathways related, for instance, to the immune system and cell proliferation. Although we could identify this protein in a small set of samples, it is well known that the identification and quantitation by mass spectrometry of BRAF V600E protein is challenging and is an understudied topic for most cancerous tissues, including melanoma.

In the present study, we intended to identify subgroups of patients with different mortality risk rates within a cohort of 49 patients with BRAF mutation (DNA-based detection) but with missing values of the BRAF V600E protein in 33 of them. To carry out this analysis, we focused on the expression levels of proteins that belong to pathways previously linked to tumors with different expression levels of BRAF V600E (Betancourt et al., 2019). We also included the expression of proteins that belong to pathways significantly enriched (Reactome, adj.p < 0.05) among proteins previously described by Betancourt et. al., as differentially expressed between groups of patients with low and high B-raf V600E. The R package 'InGRiD' was utilized to perform the analysis (Wei et al., 2019). This package provides a pathway-guided identification of patient subgroups based on protein expression while utilizing patient survival information as the outcome variable.

From the cohort of 49 BRAF mutated patients, three subgroups with different mortality risk rates were identified (**Figure 3A**). The median survival times for the low, medium and high risk groups were 5.1 years, 2.3 years and 0.5 years, respectively. We joined the medium and high risk groups to evaluate the specificity and sensitivity of the model using different survival times; the best prediction was at 3 year where the specificity and sensitivity are 0.82 and 0.88 respectively.

A total of 192 proteins were identified as involved in the activation of seven risk pathways (Transcription (hazard ratio = 0.03); Developmental Biology (hazard ratio = 0.53); Metabolism (hazard ratio = 0.72); Ub-specific processing proteases (hazard ratio = 0.98); DNA Repair (hazard ratio = 1.03); Signaling by TGF-beta Receptor Complex (hazard ratio = 1.11); [Immune System, Signal transduction and Vesicle-mediated transport] (hazard ratio = 1.52)) responsible for the grouping of the patients. The most enriched pathways were those related to Metabolism, Immune System and Signal transduction. **Figure 3B** shows the distribution of the proteins involved in the activation of these pathways. Most of the proteins belonging to Immune system sub-pathways such as Neutrophil degranulation and Complement cascade are positively related to mortality risk while proteins involved in sub-pathways related to Signal transduction and Vesicle-mediate transport are negatively related to mortality risk. On the other hand, among sub-pathways related to Metabolism, proteins belonging to Metabolism of nucleotides seem to be negatively related to risk of mortality. Proteins from Transcription, Signaling by TGF-beta Receptor Complex, DNA Repair and Developmental Biology pathways increase their expression level in the group of patients with high mortality risk.

A concert of Single Amino Acid Variants (SAAVs) in known dysregulated pathways in melanoma is observed in metastases

To identify single amino acid variants (SAAVs) in our sample cohort, the non-assigned MS/MS spectra were submitted to a second search using a protein database which included somatic mutations and variants found in 369 cases of skin cutaneous melanoma from TCGA, in 7 melanoma cell lines from the NCI-60 cancer panel, and some others collected from the COSMIC database (STAR Methods, Study of SAAVs 1.). The identified spectra assigned to SAAVs were validated using the SpectrumAI tool (STAR Methods, Study of SAAVs 2.). This resulted in the identification of 1015 unique SAAVs from 828 proteins, out of which 727 SAAVs were validated with at least 2 PSMs. To the best of our knowledge this represents the largest number of SAAVs identified in melanoma samples. Comparison between the number of SAAV peptide and wild-type peptide PSMs demonstrated that about half of the variants are present at comparable levels or are more abundant than the wild-type peptides in our melanoma cohort. Online resources (PeptideAtlas, CanProVar, UniProt, NCBI) were also utilized to retrieve more information on the SAAVs (STAR Methods, Study of SAAVs 3.). 191 SAAVs were not previously identified in another proteomic study according to PeptideAtlas, and consistently, these SAAVs were validated generally with a few PSMs only. Only 27 SAAVs were found previously to be cancer-related based on CanProVar, including also key melanoma mutations (BRAF-V600E, NRAS-Q61R/K). We were also able to retrieve the alternative allele frequency in the European population for 950 variants, mainly using data from the ALFA project. Detailed description of the SAAVs can be found in Table S4 and an overview is provided on Figure S3A-B.

Gene ontology (GO) and KEGG pathway enrichment analyses of the proteins identified with SAAVs were performed to portrait their functional annotations (Figure 3C). Over-represented terms for biological process, cellular compartment and molecular function were mainly distributed in three general categories including the extracellular matrix (ECM), cellular metabolism and the immune system. Interestingly, the biological terms presented here are also enriched for the 500 proteins that show the largest variation between the proteome-based melanoma subtypes (see Figure 1A). In fact, 71 SAAVs corresponding to 57 proteins were found among these signature proteins (11.4%). As we have 828 proteins with SAAVs among all identified 12695 proteins (6.5%), this is a notable enrichment according to a Chi-square test (χ^2 (df=1) = 17.52, p < 0.001). The data thus suggests that the proteins showing differential expression between melanoma subtypes are more frequently affected by SAAVs.

In addition, we identified some protein variants as members of known signaling pathways playing a functional role in melanoma (Chamcheu et al., 2019; Dantonio et al., 2018; Lopez-Bergami et al., 2008; Paluncic et al., 2016). Seven pathways were outlined in Figure 3D,

including at least 3 SAAVs. Interestingly, the data revealed the largest amount of SAAVs can be detected in the members of the PI3K/AKT signaling pathway (37 variants from 22 proteins). Besides the well-known role of mutations in members of MAPK pathways such as BRAF and NRAS, an increasing number of studies have associated gene polymorphism and genetic variants of the members of the PI3K/AKT signaling pathway as an indicator of susceptibility or risk to develop cancer (Li et al., 2013; Lin et al., 2010; Qi et al.).

Investigating the relationship between the alternative allele frequency in the European population (AAF) and the PSM ratio of SAAVs (PSMr, defined as: $(n_{SAAV\ PSM})/(n_{SAAV\ PSM} + n_{wild\text{-}type\ PSM})$) revealed that PSMr is a suitable indicator for the AAF (Spearman's rank correlation, rs = 0.75, $p < 0.001$, see Figure S3B). This relation allowed us to identify SAAVs which deviate clearly from the expected PSMr/AAF trend. Briefly, the ratios between PSMr and AAF were calculated (SAAVr), followed by Johnson transformation to normality, and then SAAVs with extreme values in the left and right tail were noted. The selection process is described in more detail in the experimental section (STAR Methods, Study of SAAVs 4.). We found 19 over-represented and 6 under-represented variants (i.e., PSMr is significantly higher or lower than what is expected based on their AAF, respectively) in our patient cohort (Figure 3E). As this deviation from what is observable in the normal population might indicate a key functional process for melanoma, the proteins of these SAAVs were searched in the literature for their potential role in cancer. The most outstanding over-represented variants were the well-known driver mutations of melanoma (NRAS-Q61K, NRAS-Q61R and BRAF-V600E). Additional SAAVs with unexpectedly high occurrence included proteins with (proven/potential) involvement in cell growth, adhesion, proliferation, invasion, migration of melanoma cells, tumor angiogenesis and metastasis (**ACTN4** (Shao et al., 2014), **MAGEC1** (Melanoma-associated antigen 1) (Koh et al., 2012), **AHNAK** (Sheppard et al., 2016), **TKT** (Kamenisch et al., 2016), **ELN** (Timar et al., 1991), **ANXA3** (Xu et al., 2021), **COL4A2** (Chelberg et al., 1989). Additionally, a few proteins promote the resistance to radio/drug therapy, or are involved in the DNA damage response pathway (**MX1** (Khodarev et al., 2009), **ANK1** (Hall et al., 2016), **NSUN2** (Delaunay and Frye, 2019)). In the list we also find **LZTS1**, a tumor suppressor in uveal melanoma, and 4 proteins (**ARL6IP6** (Xu et al., 2016), **H1-1**, **RBM19** (Wang et al., 2021), **GSTM1** and **LSM14A)** with no clear relevance to cancer found until now, to the best of our knowledge.

Six SAAVs showed higher frequency within the normal population (AAF > 0.50) but were detected with a significantly lower PSMr than expected in our melanoma samples. Most interesting proteins in this list are C6 and EXD2. **C6** is part of the complement system, which has both cancer-promoting activities and anticancer cytotoxic activity (Fishelson and Kirschfink, 2019). **EXD2** has been suggested as a potential target for cancer therapy due to its regulatory function in mitochondrial translation and replication stress response (Nieminuszczy et al., 2019; Stracker, 2018). The other proteins in this list are **MYH3** with no clear relationship to melanoma (Vivancos et al., 2016), and another muscle-related protein, **TTN**, that has been reported to be mutated frequently in several tumor types (Jia et al., 2019). **ERO1B** is an oxidoreductase, showing a significant overexpression in different metastatic sites compared to primary tissue in breast cancer (Li et al., 2020) and was identified as prognostic marker in pancreatic cancer (low ERO1LB expression was associated with poorer overall survival, (Zhu et al., 2017)). Additionally, **AKAP13** was also shown to be involved in cancer on a few occasions but no clear role in the disease has been established (Bentin Toaldo et al., 2015; Fehér et al., 2012; Molee et al., 2015). Variants with no corresponding wild-type peptide can also be viewed as potentially over-represented variants. Out of such 189 SAAVs, 3 were shown to have very high reference allele frequency in the European population (> 0.89) and were verified with more than 20 PSMs. The proteins of these SAAV peptides are ITGB2, ICAM1 and APOL2. **ITGB2** is a protein involved in cell adhesion and cell-surface mediated signalling was found under-expressed in BRAF positive tumors (Trilla-Fuertes et al., 2019), and correlated with survival in renal and colorectal tumors (Boguslawska et al., 2016; Cavalieri et al., 2007). **ICAM1** plays important roles in cell adhesion, in inflammation processes and in the activation of lymphocytes (Figenschau et al., 2018; Yang et al., 2005), and a study revealed its expression in aggressive subtypes of breast cancer (Figenschau et al., 2018). Lastly, **APOL2** is suggested to be involved in apoptosis (Vanhollebeke and Pays, 2006).

Analysis of tumor microenvironment

Principal Component Analysis (PCA) of the 18 samples with low tumor content (<30% tumor cells as per histopathology assessment) (Figure 4A), showed marked differences closely associated with the cellular composition of the tumor microenvironment (TME). The samples clustered in three main groups: one with the highest adjacent lymphocyte content (6 samples, left side of the PCA), a second group with highest connective tissue content (3 samples, right side of the PCA) and a third group (9 samples) which was located in the middle of the PCA map and shared characteristics with both above mentioned groups but with a significantly higher component of connective tissue (Figure 4B). Hence, subsequent analyses were performed considering two groups of samples (high and low, 12 and 6 samples, respectively) in relation to the connective tissue content of the TME.

The Kaplan-Meier analysis (Figure 4C) showed that the composition of the TME was associated with patient outcome (p-value= 0.031). Patients within the group of tumors with TME with low connective tissue content had better prognosis (mean DSS 1474 days, where more than 70% (5) of patients survived more than 2 years and 57% survived more than 9 years after sample collection). The opposite situation was observed for patients corresponding to the group of tumors with TME with higher connective tissue content (mean DSS 265 days, where less than 17% (2) of patients survived more than 2 years).

T-test analysis performed with the global proteomic data from these two groups of samples found 2,213 (FDR<0.05) proteins significantly differentially expressed (Figure 4D). Samples with low connective tissue content were enriched in pathways such as cell adhesion molecules, spliceosome and ribosome, but also other pathways intimately related to the immune response including natural killer cell mediated cytotoxicity, B and T cell receptor signaling pathways, which reflected the expected activity of lymph node cells. In contrast, samples of the group with high connective tissue content were enriched in proteins belonging to the complement and coagulation cascades, PPAR as well as the ECM-receptor interaction pathways.

Strong relationship between proteogenomics profiles and survival time reveal candidate biomarkers and disease pathways

Two complementary supervised approaches to relate omics data to survival were applied. First, outlier analysis, treats survival as a binary variable. Second, Cox analysis considers survival as a continuous variable. 111 samples with larger than 30% tumor cell content were included.

Outlier analyses discovered critical genes related to melanoma patient survival and stage in multi-omics datasets

To find biomarkers related to the survival of melanoma patients, we stratified our omics datasets based on whether the patients, at the time of censoring, had survived over 6 months, 1 year, 3 years, or over 5 years from the sample collection (surgery) date. Outlier tables were then calculated for proteomics for each protein isoform, transcriptomics for each gene, and phosphoproteomics for each modified sequence (phosphosite). This determines if it is an outlier for each patient sample based on extending the interquartile ranges by a factor of 1.5. For each feature, we conducted group-wise (e.g. survived over 6 months vs. dead before 6 months) Fisher's exact test on counts of patients that are outliers in each group. If a feature was an outlier in less than 30% of patients' data in one group, it was skipped to make sure that our results were not driven by a small subset of samples. Compared to other differential expression analysis methods, outlier analysis does not assume specific underlying distribution of the features of interest, which makes outlier analysis more robust. In addition, outlier analysis is capable of detecting extreme values related to increase of protein expression variability that could imply the loss of biological functions, which may sometimes be ignored by other methods. On the other hand, it also embeds a dedicated coefficient to prevent a small subset of samples from driving the whole analysis.

The FDR values of putative biomarkers found by outlier analysis of multi-omics data were shown in Figure 5A. We found 82 proteins significantly enriched as outliers in proteomics data of patients who survived less than 6 months from sample collection (surgery) (Supplementary Figure 5A). 19 genes were found in transcriptomics data and 3 phosphosites were found in phospho-proteomics data (Supplementary Figure 5B-C). In this 6 month survival analysis, HMOX1 was significant in both transcriptomics and phospho-proteomics data while XYLB was significant in both transcriptomics and proteomics data. 6 proteins were found to be significantly enriched in proteomics data in patients who lived less than 1 year (Supplementary Figure 5A). Also for 1-year survival, one gene was found in transcriptomics and one phosphosite was found in phospho-proteomics (Supplementary Figure 5B-C). Apart from the survival, we also conducted outlier analyses between stage 3 and stage 4 patients, between NRAS mutant and wild-type patients, and for the time from primary diagnosis to metastasis. Four proteins (PDZ11, HEATR5A, SIL1, MAN1B1) were found significantly enriched in stage 4 vs stage 3 patients in proteomics data and these proteins were also significantly enriched in patients who survived less than 6 months from diagnosis (Supplementary Figure 5A). NRAS mutant patients had FAM45A and TOLLIP enriched in proteomics data as compared to NRAS WT patients (Supplementary Figure 5A). However, outlier analysis did not discover any significant proteins or genes differentiating between groups of patients with different genders and BRAF mutation status.

Cox analysis discovered further features related to melanoma survival

An alternative approach to link survival to omics features, the Cox survival analysis, is widely used in biomedical studies and uses survival information as a continuous variable. Here, the features (proteins, phosphosites, transcripts) that were selected by the Cox model in at least 30 of the 100 repetitions were selected for further analysis. Figure 5B shows the result of the Cox analysis for the three omics datasets. The predictive power of the molecular data was moderate, the concordance indices (C-indices) varied between 0.538 and 0.601. A C-index of 1.0 means that the model "ranks" the samples perfectly, i.e. patients with a higher risk score (hazard score) died earlier than those with lower scores. A C-index of 0.5 is the expected performance of a nonsense/random model. The C-index parameter is analogous to the AUC parameter (Area Under Curve) used for a binary classifier. A C-index of 0.6 is considerably better than 0.5, indicating that our molecular data can predict survival to some extent, but not completely. Cox analysis showed a strong relationship to disease-specific survival (time from surgery/sample collection to death or censoring) for 12 genes (transcriptomics) and 8 phosphosites. No significant relationships between protein expression and disease-specific survival were identified in the Cox analysis, although fifteen proteins were related to overall survival, whereas Cyclin-dependent kinase 4 (CDK4) was most significant. In phospho-proteomics data, the features (phosphosites) most

significantly related to disease-specific survival included: AKAP-12 (A-kinase anchor protein 12), MARCKS (Myristoylated alanine-rich C-kinase substrate), PTPRC (Receptor-type tyrosine-protein phosphatase C), ADAM10, FGA and HMOX1. In transcriptomics data, the features most significantly related to survival were: SYTL2 (Synaptotagmin-like protein 2), CCND1 (G1/S-specific cyclin-D1), LQK1 (Putative uncharacterized protein LQK1), LATS2 (Serine/threonine-protein kinase LATS2).

Biological significance of the proteins related to survival

Biological relationships within a set of omics features of interest, e.g. proteins related to survival, can be explored by diverse algorithms that map gene/protein sets onto networks of relationships extracted from sources such as literature mining, interaction databases or pathway databases.

Here, Ingenuity Pathway Analysis (IPA) was performed for the proteins related to survival as per the outlier and COX analyses. The IPA algorithm provided tight biological relationship networks for the query protein sets. For proteins related to survival as per outlier analysis, the top three IPA relationship subnetworks produced by the algorithm focused on proteasome, MAPK kinase signaling and MYC-FOS signaling, respectively. Notably, the analysis highlighted a number of targets of existing drugs among the survival-related proteins, including the PI3 kinase PIK3CB, TXNRD1, MMP12 and MME proteases and PSMD2. Also, for proteins related to survival as per Cox analysis, the top IPA relationship subnetwork produced by the algorithm focused on MAPK kinase signaling (Figure 5C-D).

Additionally, several functional themes were notable examining the survival related proteins. Among the 82 proteins significantly enriched (outliers in proteomics data) when comparing patients who survived less than 6 months from sample collection (surgery) to the rest, there are as many as eleven mitochondrial proteins whereas mitochondria are known to have critical roles in cancer (Yuan et al., 2020). Also, there are two members (AIMP1, IARS) of the tRNA synthetase complex known for involvement of translation but also for cancer-related functions (Hyeon et al., 2019). Other proteins with known roles in cancer include atypical protein kinases involved in cancer progression, PIK3CB (Phosphatidylinositol 3-kinase beta), RIOK1, CSNK1G3 (Casein kinase I gamma-3), FASTKD2 (Berto et al., 2019; Fang et al., 2020).

Interestingly, among the survival-related proteins, there are several poorly studied, potentially druggable targets. The orphan receptor GPR126/ADGRG6 of the Adhesion family of G protein-coupled receptors is a member of the large group of poorly understood emerging cancer drug targets (Gad and Balenga, 2020) that couple cell-cell signalling to intracellular signal transduction and are implicated in migration, proliferation, and survival of tumor cells (Garinet et al., 2019; Musa et al., 2019). COG4 and COG3 are components of the Conserved Oligomeric Golgi complex (COG) which is a vesicle tethering complex that regulates retrograde vesicle traffic within the Golgi (Blackburn et al., 2018; D'Souza et al., 2020). CARNMT1/C9orf41, a poorly studied histidine methyltransferase, methylates the dipeptide carnosine producing anserine, which is believed to be proton buffer and radical scavenger (Drozak et al., 2015). Further, xylulose kinase (XYLB), a protein whose relation to survival is supported by two omics datasets (transcriptomics and proteomics) is an important enzyme in carbohydrate metabolism and as such may play an important role in metabolic disease (Bunker et al., 2013).

Survival-related protein candidates for IHC validation

By aggregating the results from outlier analyses and Cox regression analyses, 247 unique survival-related proteins were identified, where 112 proteins were found in outlier analyses and 198 proteins were found in Cox regression. To narrow the list for further immunohistochemistry validation, the survival times (DSS) were stratified at 6 months, 1 year, and 5 years, and an attempt was made to predict survival using expression of each candidate protein. For each protein, the expression level that results in the maximum sum of sensitivity and specificity was used as the cutoff point. The areas under the receiver operating characteristic (AUROC) curves of each protein at different survival thresholds were used as the major indication of its relative importance (Supplementary Table 5D). In addition, we considered each protein's involvement in the critical pathways found by ICA-GSEA. Based on the results of the outlier analyses, Cox regression, AUROC, and the involvement in melanoma-related pathways, a comprehensive evaluation and literature search was conducted to generate a top list of 10 proteins (ADAM10, HMOX1, FGA, DDX11, SCAI, CTNND1, CDK4, PAEP, PIK3cB, TEX30) (Table 5E). Notably, one of these proteins, Heme oxygenase1 (HMOX1) was supported by outlier analysis in two omics datasets: transcriptomics and phosphoproteomics. ADAM10 (A disintegrin and metalloproteinase 10) and FGA (fibrinogen alpha chain) were supported by COX analysis of phosphoproteomics. FGA was also supported by outlier analysis in proteomics. These ten proteins were selected for IHC validation in a separate cohort.

ADAM10, FGA and HMOX1 were highlighted in the present study based on specific phosphosites linked to survival. For FGA (Ser364) and HMOX1 (Ser229), the phosphosites were linked to poor survival while the ADAM10 phosphosite Thr719 was enriched in tumors from patients surviving longer. By ICA, CTNND1 is linked to IC066 "Primary Breslow" by more than 24 phosphosites. Furthermore, the protein expression of CDK4, CTNND1, PAEP, TEX30, DDX11, SCAI and PIK3CB was related to a poor prognosis.

Of the three proteins highlighted by specific phosphosites, several links to melanoma in the literature were established. ADAM10 (a disintegrin and metalloproteinase 10) is a member of the ADAM family, which are endopeptidases with broad specificity, and involved in membrane shedding of several proteins. Interestingly, ADAM10 together with ADAM17 may promote membrane shedding of immunosuppression proteins such as PDL1 and LAG3 (Andrews et al., 2015; Lambrecht et al., 2018; Orme et al., 2020). Fibrinogen alpha chain (FGA), is a part of the glycoprotein fibrinogen which has major

functions in hemostasis, wound healing and also immune responses. High plasma levels of fibrinogen have been attributed to poor prognosis in lung cancer and also in melanoma. The PTM state of fibrinogen has earlier been linked to the disease state of cancer (Cardinali et al., 1990; Ciereszko et al., 2019; Guida et al., 2003; Gunji and Gorelik, 1988; Nagel et al., 2018; Ogata et al., 2006; Palumbo et al., 2000; Ryu et al., 2015). The third selected protein, HMOX1 (Heme oxygenase), is an antioxidant and anti-inflammatory enzyme involved in generating biliverdin and bilirubin. HMOX1 may promote cancer cell growth, tumor cell survival and resistance to treatment. HMOX1 has also been linked to a poor outcome of melanoma (Hjortsø and Andersen, 2014), (Furfaro et al., 2020),

Next, seven proteins were selected based on regulation at the protein level. Also for these, literature links to melanoma and/or cancer were found. The tumor suppressor SCAI (suppressor of cancer cell invasion) is a highly conserved protein that acts on the RhoA–Dia1 pathway to regulate invasive cell migration. SCAI is downregulated in many human tumors and high expression of SCAI correlates with better survival in patients with breast and lung cancers (Brandt et al., 2009; Gasparics et al., 2018). There is however little information about SCAI in melanoma. CDK4 (Cyclin-dependent kinase 4) is a well-known cancer target that regulates cell cycle and proliferation. In melanoma, mutations and dysregulations are commonly seen in CDK4 and proteins in the CDK4 pathways. Several candidate drugs are in the clinical phase (Freedberg et al., 2008; Guo et al., 2020; Hocker and Tsao, 2007). CDK4 was therefore considered a candidate for further validation.

CTNND1 (catenin delta-1, isoform 1A) is a key regulator of cell-cell adhesion. Several studies suggest a link to melanoma. The long isoform (1A), is highlighted in our study. The longer isoforms, often enriched in tumors, have been reported as pro-tumorigenic and playing a role in EMT (Aho et al., 2002; Aslund-Ostberg et al., 1992; van Hengel and van Roy, 2007; Kourtidis et al., 2015; Pieters et al., 2012; Yanagisawa et al., 2008). The helicase DDX11 (ATP-dependent DNA helicase DDX11) has a role in chromatid cohesion. DDX11 has been reported upregulated with progression from noninvasive to invasive melanoma, and expressed at high levels in advanced melanoma (Bhattacharya et al., 2012; Li et al., 2019; Mahtab et al., 2021; Marchese et al., 2016).

Glycodelin or PAEP (progestagen associated endometrial protein), is a secreted glycoprotein that regulates critical steps during fertilization and also has immunomodulatory effects. PAEP is expressed in melanoma and involved in tumor proliferation, migration and may promote development of immune tolerance in the tumor. PAEP expression is regulated in part by MITF (Liu, 2011; Luke et al., 2017; Ren et al., 2010, 2011, 2015). The novel hydrolase TEX30 (Testis expressed 30), is likely a lipid metabolizing enzyme (Lord et al., 2013). It was upregulated at the gene level in a study of melanoma compared to normal skin (Huang et al., 2019), however the role and mechanism in melanoma remain to be established. The tenth candidate protein, PIK3CB (Phosphatidylinositol 4,5-bisphosphate 3-kinase), is part of the PI3K–AKT cascade which is one of the most studied pathways in cancer. This pathway has a role in cell survival, migration and also in oncogenic transformation. In melanoma, the PI3K–AKT pathway may be activated by mutations in NRAS or loss of the suppressor PTEN (Gao et al., 2020; Kwong and Davies, 2013; Yuan and Cantley, 2008).

Taken together, the ten candidate proteins in the present study can be linked to several aspects of melanoma and/or cancer in the literature. CDK4 and PIK3CB are well known melanoma targets and may serve as “positive controls” for further validation. For the other proteins, different mechanisms in melanoma development may be considered by expression analysis in tumors from a separate cohort of melanoma in different stages.

Clinical disease presentation of candidate biomarkers provides evidence of tumor compartments with molecular implications for metastasis.

In the current clinical practice of the malignant melanoma, immunohistopathology has a crucial role in the confirmation of the tissue-based protein characteristics of the disease. Combining the results from immunohistochemical (IHC) validation with proteomic/phosphoproteomic data provides a good opportunity to identify the potential prognostic, and predictive biomarkers, for the determination of the survival possibilities as well as for responder state of melanoma patients.

Based on the independent IHC validation cohort of 42 patients of primary melanomas causing locoregional or distant metastatic disease (*STAR Methods*) in accordance with the proteomic results, nine candidate biomarkers (ADAM10, SCAI, HMOX1, DDX11, FGA, CTNND1, CDK4, PAEP, PIK3cB) served as the basis for further immunohistochemical analysis (the IHC validation of the TEX30 marker was not carried out due to inadequate antibody performance). The IHC expression value was determined by the application of densitometry in the representative areas in melanoma cells and the stromal part.

To demonstrate the prognostic/predictive impact of the listed markers and their link to melanoma progression, first, we divided the mean values of the expression of each protein in the melanoma cells and in the stromal parts, then ROC curve and Kaplan-Meier survival analyses were conducted based on the cutoff scores of the related proteins. The tissue validation of the chosen markers on the primary melanomas showed notable differences between the expression values for the two groups (who survived until censoring date and those that did not) of the patients. CDK4 (independent t-test, $p < 0.001$), ADAM10 (independent t-test, $p < 0.05$), and SCAI (independent t-test, $p < 0.05$) proteins showed significant differences in melanoma cells, based on the two groups of the patients in the IHC validation cohort. These findings are correlated to the production of the indicated biomarkers in the primary melanoma phase predicting the progression of the melanoma. Furthermore, the three above mentioned candidates, namely CDK4 (ROC curve, AUC= 0.646, $p < 0.000$), ADAM10 (ROC curve, AUC= 0.579, $p = 0.055$), and SCAI

(ROC curve, AUC= 0.591, $p < 0.05$) displayed the ROC curves with the most significant differences regarding expression in melanoma cells. Consequently, low expression (below the estimated cutoff value) of CDK4, SCAI, ADAM10 proteins observed in melanoma cells, predicted poor survival rate of patients in the IHC validation cohort.
Kaplan-Meier analyses elucidated the survival of the patients based on cut off-low or high protein expression in the primary melanoma. Although tumoral CDK4 showed almost significant association with the DFS (early prognostic marker) together with the lower stromal CDK4 (DFS, PFS, OS, p=0.000), it may predict an aggressive behavior of the disease. For ADAM10 and SCAI, lower protein expression in the melanoma cells indicated significantly poor prognosis: DFS (ADAM10, $p<0.05$) (SCAI, $p<0.001$), PFS (ADAM10, $p<0.05$) (SCAI, $p<0.001$), and OS (ADAM10, $p<0.05$) (SCAI, $p<0.001$). Similar results were obtained with the low stromal SCAI expression (PFS p=0.001; OS, p=0.002).

To complement the understanding of the candidate biomarkers, IHC validation was also conducted in 9 cases of primary melanomas, followed by a comparison with their corresponding metastases.
All proteins, with the exception of ADAM10, exhibited increased protein expression in melanoma cells of the metastases compared to primary melanoma. The protein expression was particularly higher in the metastatic melanoma cells, in the case of six proteins: CDK4 (Wilcoxon signed-rank test, $p < 0.001$), FGA (Wilcoxon signed-rank test, $p < 0.05$), PAEP (Wilcoxon signed-rank test, $p < 0.001$), SCAI (Wilcoxon signed-rank test, $p < 0.001$), HMOX (Wilcoxon signed-rank test, $p < 0.001$), and DDX11 (Wilcoxon signed-rank test, $p < 0.001$). In contrast, ADAM10 expression did not show a significant difference in the metastases compared to its corresponding primary melanomas, which was also supported by the proteomic data of the Cox regression analysis. Contrary to previous reports (Lee et al., 2010), the low expression of ADAM10 in metastases may highlight an immunological role in the melanoma cell recognition during progression (Lambrecht et al., 2018), therefore further investigation is needed.

Cox Regression on Quantified IHC Features

We built a multivariate Cox regression survival model based on quantified melanoma cell and stroma cell scores from the IHC slides of 9 toplist proteins (CDK4, PAEP, SCAI, CTNND1, FGA, HMOX1, PIK3CB, ADAM10, DDX11) from the validation cohort. Another Cox regression model was also built using the proteomics data of the same 9 proteins from the discovery cohort. By comparing the corresponding Cox regression coefficients, z-scores between the melanoma cell in IHC-based model and the proteomics-based model, the Cox coefficients of 6 toplist proteins (CDK4, PAEP, CTNND1, FGA, PIK3CB, ADAM10) showed same directionality in both models, suggesting their consistent effects on patients' survival in both cohorts (Figure 6 B and C). Taken together, the validation by IHC in a separate cohort with both primary and metastatic melanoma combined with the proteomic data (**Figure 6**) pointed to the role of ADAM10, SCAI and CDK4 as candidate biomarkers of survival in melanoma and suggests further clinical evaluation of their role in melanoma disease progression and therapy.

Strategies for translating the patient survival biomarkers into the clinic for guiding treatment and therapeutic opportunities

The high mortality of the melanoma patients is mainly due to the rapidly disseminated visceral and cerebral metastases highlighting the importance of the metastatic cascade of the usually embarrassingly small primary melanoma. As metastases develop after a certain time of wide marginal resection of the primary melanoma with negative staging characterizes the presence of latent clinical phase of early dissemination and forthcoming late tumorigenic phase with the apparent lymphatic and visceral metastases. Although some of the important drivers as BRAFV600, NRAS, NF1 mutations are revealed, it is still unknown, how the drivers and the passenger mutations lead the primary melanoma to provide dormant tumorigenic information into the surrounding niche. Indeed, little is known on tissue level about the tumorigenic shift from the dormant phase to the fatal progressive disease. Nevertheless, more data gained about the role of the tumor microenvironment (TME) in the formation of melanoma-supporting stroma as well as in the premetastatic niche. The revolutionary improvement of immune checkpoint blockade by PD1 inhibitors targeting the antitumorigenic TME has shown up a marked improvement compared to the disappointing response rate of the conventional chemotherapy for metastatic melanoma patients.

As recently the two main therapeutical arms of the metastatic melanoma are targeted (i) on the tumor cells themself aiming the direct elimination of the tumor by BRAFV600E target therapy or (ii) on the antitumoral TME by the indirect antitumoral action of the immune checkpoint therapy. However, approximately 50-60% of metastatic melanoma patients could have a benefit for these treatments necessitating the unmet need for tissue biomarker search by precision protein-based platforms. Compared to PCR data gained on the summarized mutational state of the complex tumor counterparts together, the tissue-based protein research platforms highlight not just the quantity and quality but the tissue heterogeneity of the mutated proteins prone to be targeted by small molecule drugs as well as by biologics.

The present high-throughput proteomic analysis resulted in proposed candidate tissue biomarkers which were further validated in a metastatic primary melanoma cohort on the tissue level, compared to the individual standardized follow up data. Each tissue array exhibited the personalized topographic expression data endowed with the colorimetric density and distribution of the examined biomarker protein allowing to test its prognostic and/or predictive value. Thus, the validated markers, in particular CDK4, SCAI and ADAM10, are promising starting points for developing clinically useful diagnostic and prognostic tools, as well as for understanding the biology of melanoma. The similarities and differences in validation results between melanoma cells and tumor environment highlights the importance of tumor-stroma relationship.

METHODS

A total of 142 patients (48 females and 94 males) diagnosed with metastatic melanoma between 1975 and 2011, were included in the study. The average age ± standard deviation (range) at diagnosis of metastases was 62.3 ± 13.7 (25–89) years. The cohort comprised metastatic tissues from lymph node (126), subcutaneous (7), cutaneous (1), visceral (3), while for five the origin could not be established. The mutational status was determined for 124 of the tumors, with 50 found mutated at BRAF (92% V600E), 37 with NRAS Q61K/R, and 37 tumors had wild type variants for both mutations. Only four patients received targeted B-raf treatment with vemurafenib. The study was approved by the local ethical committees, including the Regional Ethical Committee at Lund University, Southern Sweden (DNR 191/2007, 101/2013 (BioMEL biobank), 2015/266 and 2015/618). All patients provided written informed consent.

Sample preparation for mass spectrometry and data acquisition

Protein extraction was performed on sectioned (3x10 µm), fresh-frozen melanoma metastatic tissues using the Bioruptor plus, model UCD-300 (Dieagenode). A total of 142 MM tissue samples were lysed in 100 µL lysis buffer containing 4 M urea and 100 mM ammonium bicarbonate. After a brief vortex, samples were sonicated in the Bioruptor for 40 cycles at 4°C. Each cycle consisted of 15 s at high power and 15 s without sonication. The samples were then centrifuged at 10,000 ×g for 10 min at 4°C. The protein content in the supernatant was determined using a colorimetric micro BCA Protein Assay kit (Thermo Fisher Scientific, Rockford, IL).

Protein digestion was performed on the AssayMAP Bravo (Agilent Technologies) platform using the digestion v2.0 protocol. Protein concentrations were adjusted to 2.5 µg/µl and 100 µg of protein from each sample were reduced with 10 mM DTT for 1 h at room temperature (RT) and sequentially alkylated with 20 mM iodoacetamide for 30 min in the dark at RT. To decrease the urea concentration, the samples were then diluted approximately seven times with 100 mM ammonium bicarbonate. Digestion was performed in two steps. Proteins were first incubated with Lys-C at a 1:50 (w/w) ratio (enzyme:protein) for 5 h and then trypsin was then added at a 1:50 (w/w) ratio and the mixture incubated overnight at RT. The reaction was quenched by adding 20% TFA to a final concentration of ~1%. Peptides were desalted on the AssayMAP Bravo platform using the peptide cleanup v2.0 protocol. C18 cartridges (Agilent, 5 µL bed volume) were primed with 100 µL 90% acetonitrile (ACN) and equilibrated with 70 µL 0.1% TFA at a flow rate of 10 µL/min. The samples were loaded at 5 µL/min, followed by an internal cartridge wash with 0.1% TFA at a flow rate of 10 µL/min. Peptides were eluted with 30 µL 80% ACN, 0.1% TFA and dried in a Speed-Vac (Eppendorf) prior to TMT labelling.

For the phosphoproteomic analysis, protein digestion and C18 peptide cleanup was repeated on protein lysates. After Speed-Vac, peptides were resuspended in 80% ACN, 0.1% TFA prior to phosphopeptide enrichment.

The peptide amount in each sample was estimated using a quantitative colorimetric peptide assay kit (Thermo Fisher Scientific, Rockford, IL). Within each batch, equal amounts of peptides were labelled with TMT 11-plex reagents. The TMT labelling was performed according to manufacturer's instructions. Peptides were resuspended in 100 µL of 200 mM TEAB and individual TMT 11-plex reagents were dissolved in 41 µL of anhydrous acetonitrile and mixed with the peptide solution. The internal reference sample, a pool consisting of aliquots from protein lysates from 40 melanoma patient samples, was labelled in channel 126 in each batch. After one hour of incubation, the reaction was quenched by adding 8 µL of 5% hydroxylamine and incubated at room temperature for 15 minutes. The labelled peptides were mixed in a single tube, the volume was reduced in a Speed-Vac and then the peptides were cleaned up using a C-18 Sep-Pak cartridge (Waters). The eluted peptides were dried in a Speed-Vac and finally resuspended in water prior to high pH RP-HPLC fractionation. The samples were distributed among 15 different batches, as described in the Supplement, and in each of them the internal reference sample was included (TMT tag 126).

The TMT-11 batches were fractionated using a Phenomenex Aeris Widepore XB-C8 (3.6 µm, 2.1 × 100 mm) column on an 1100 Series HPLC (Agilent) operating at 80 µL/min. The mobile phases were solvent A: 20 mM ammonium formate and solvent B: 80% ACN - 20% water containing 20 mM ammonium formate. Both solvents were adjusted to pH 10 with ammonium hydroxide. An estimated amount of 200 µg was separated using the following gradient: 0 min 5% B; 1 min 20% B; 60 min 40% B; 90 min 90% B; 120 min 90% B. The column was operated at RT and the detection wavelength was 220 nm. 96 fractions were collected at 1 min intervals and further concatenated to 24 or 25 fractions (by combining 4 fractions that were 24 fractions apart so that #1, #25, #49, and #73; and so forth, were concatenated), and dried in a Speed-Vac.

The Phospho Enrichment v2.0 protocol on the AssayMAP Bravo platform was used to enrich phosphorylated peptides using 5 µL Fe(III)-NTA cartridges. The cartridges were primed with 100 µL 50% ACN, 0.1% TFA at a flow rate of 300 µL/min and equilibrated with 50 µL loading buffer (80% ACN, 0.1% TFA) at 10 µL/min. samples were loaded onto the cartridge at 3.5 µL/min. The samples were washed with 50 µL loading buffer and the phosphorylated peptides were eluted with 25 µL 5% NaOH directly into 10 µL 50% formic acid (FA). Samples were dried in a Speed-Vac and stored at -80°C until analysis by LC-MS/MS.

nLC-MS/MS analysis was performed on an Ultimate 3000 HPLC coupled to a Q Exactive HF-X mass spectrometer (Thermo Scientific, San Jose, CA). TMT 11 labelled peptides from each fraction (1 µg) were loaded onto a trap column (Acclaim1 PepMap 100 pre-column, 75 µm, 2 cm, C18, 3 mm, 100 Å, Thermo Scientific, San José, CA) and then separated on an analytical column (EASY-Spray column, 25 cm, 75 µm i.d., PepMap RSLC C18, 2 mm, 100Å, Thermo Scientific, San José, CA) using a 120 min ACN gradient with 0.1% formic acid at a flow rate of 300 nL/min and a column temperature of 45°C. Q

Exactive HF-X mass spectrometer was set using the TMT node as follows: full MS scans at m/z 350-1400 with a resolution of 120000 at m/z 200, a target AGC value of 3×106 and IT of 50 ms, DDA selection of the 20 most intense ions for fragmentation in HCD collision cell with an NCE of 34 and MS/MS spectra acquisition in the Orbitrap analyzer at a resolution of 45000 (at m/z 200) with a maximum IT of 86 ms, fixed first mass of 110 m/z, isolation window of 0.7 Da and dynamic exclusion of 30 s.

Data-dependent analysis (DDA) of phosphopeptides for spectral library DDA building was carried out in the same LC-MS/MS system as for the global proteome analysis. Peptides were dissolved 2% ACN, 0.1% TFA and spiked in with iRT peptides (Biognosis AG) in a 1:10 dilution (iRT:peptides). The peptides were loaded onto a trap column (Acclaim1 PepMap 100 pre-column, 75 µm, 2 cm, C18, 3 mm, 100 Å, Thermo Scientific, San José, CA) and then separated on an analytical column (EASY-Spray column, 25 cm, 75 µm i.d., PepMap RSLC C18, 2 mm, 100Å, Thermo Scientific, San José, CA) using a 120 min ACN gradient in 0.1% formic acid at a flow rate of 300 nL/min and a column temperature of 45°C. Mass spectrometer parameters were set as follows: selection of the 15 most intense ions for fragmentation, full MS scans at m/z 375-1,750 with a resolution of 120,000 at m/z 200, a target AGC value of 3×106 and IT of 100 ms, fragmentation in HCD collision cell with a normalized collision energy (NCE) of 25 and MS/MS spectra acquisition in the Orbitrap analyzer at a resolution of 60,000 (at m/z 200) with a maximum IT of 120 ms and dynamic exclusion of 30 s.

For DIA-MS, the phosphopeptides were separated using the same gradient and the MS system as for the DDA analysis and the iRT mix was also added to the individual samples. The full scans were processed in the Orbitrap analyzer with resolution of 120,000 at (200 m/z), injection time of 50 ms and 3×106 of AGC target in a range of 350 to 1410 m/z. Fragmentation was set in 54 variable isolation windows based on the density distribution of m/z precursors in the spectral library (see below, Supplementary information). MS2 scans were acquired through the windows with a resolution of 30,000 at 200 m/z, NCE of 25, 1×106 of AGC and 200 m/z as fixed first mass.

Quality control measurements were introduced to assess the performance of LC-MS/MS systems. A protein digest from Hela cells (Pierce HeLa Protein Digest Standard, Thermo Fisher Scientific) mixed with a standard peptide mixture (Pierce Peptide Retention Time Calibration Mixture) was used as a QC sample and measured every tenth LC-MS/MS analysis. This allowed monitoring peak width, retention time, base peak intensity, number of MS/MS, PSMs, number and of peptides and proteins identified, among others QC metrics.

QUANTIFICATION AND STATISTICAL ANALYSIS

Proteomic data processing and MS data interpretation

The global proteomics experiment generated a total of 375 raw files that were processed with Proteome Discoverer 2.3 (Thermo Fisher Scientific, San José, CA, USA) using the Sequest HT search engine. The search was performed against the Homo sapiens UniProt revised database (downloaded 2018-10-01) with isoforms. Cysteine carbamidomethylation (+57.0215 Da) and TMT 6plex (+229.1629 Da) at peptide N-terminus and lysine were set as fixed modifications while methionine oxidation (+15.9949 Da), N-terminal acetylation (+42.0105 Da) and were set as variable modifications; peptide mass tolerance for the precursor ions and MS/MS spectra were 10 ppm and 0.02 Da, respectively. A maximum of two missed cleavage sites was accepted and a maximum false discovery rate (FDR) of 1% was used for identification at peptide and protein levels. The Proteome Discoverer software allowed the introduction of reporter ion interferences for each batch of TMT 11-plex reagents as isotope correction factors in the quantification method. The database search resulted in the identification of 12,695 proteins, corresponding to 11,468 genes. Only the peptides that could be uniquely mapped to a protein were used for relative protein abundance calculations.

These search results were imported into Perseus software v. 1.6.6.0 (Tyanova et al., 2016) where a filtering was applied to merely include proteins with quantification values for all reporter ions. To correct for experimental differences related to sample handling and other biases such as column changes, the protein intensities were log2 transformed and centered around zero by subtracting the median intensity in each sample. To allow for the comparison of relative protein abundances between the different batches of TMT11 the protein intensities from the pooled references sample (in channel 126 in each batch) was subtracted from each channel in the corresponding batch to obtain the final relative protein abundance values (log2 transformed and zero centered).

The TMT-based proteomic analyses of 142 metastatic malignant melanoma samples resulted in the identification of 12,695 proteins (11,468 genes) with an average of 10,705 proteins identified per tumor sample. The data displayed 15.5% missing values for the protein abundances and 8124 proteins were present across all samples (Table SX). Long term reproducibility of the digestion workflow was previously shown (Kuras et al., 2018). The reliability of the TMT workflow was evaluated by repeating the entire experiment of batch one (B1). Although factors such as sample aging and change of RP-high pH fractionation column and MS-instrument influenced the analysis, the overall agreement and correlation of protein abundances between the experiments was good (Figure S-1A). In addition, good longitudinal performance across the 15 batches was demonstrated by the rather constant sequence coverage (Figure S-1B).

Principal component analysis, using 8124 proteins commonly quantified among the 142 melanoma samples could separate between high- (>70%) and low-containing (<30%) tumor samples based on protein abundance (Figure S-1C). In addition, no batch effects were observed for the global proteomic data. Furthermore, Student's t-test was performed to look at the expression of known melanoma protein markers and drivers in the high tumor-containing samples compared to the samples containing mostly tumor

microenvironment (Figure S-1E, Table SX). Known melanoma protein markers and drivers such as TYR, S100A1, MLNA, RB1, BRAF and WDR12, were upregulated in the high tumor-containing samples. For example the S100A1 and TYR were highly overexpressed in the sample containing a high percentage of tumor cells. Furthermore, m-RNA and protein abundances showed strong positive correlation (median 0.408), and 84% showed significant correlation ($p<0.05$) for the 6101 overlapping genes across 104 patient samples (Figure S-F). The average correlation was in the middle of previously reported CPTAC colorectal ($r = 0.23$), breast ($r = 0.39$), ovarian ($r = 0.45$) and endometrial ($r = 0.48$) mRNA-protein correlations.

Phosphoproteomic data processing and MS data interpretation

The phosphoproteomic spectral library was generated from 45 data dependent (DDA) raw files in the Spectronaut X platform (Biognosis AG) against the Homo sapiens database from Uniprot (downloaded 2019-01-15). The following parameters were used: cysteine carbamidomethylation (+57.0215 Da) as fixed modification and methionine oxidation (+15.9949 Da), N-terminal acetylation (+42.0105 Da) and phosphorylation (+79.9663 Da) on serine, threonine and tyrosine were selected as variable modifications. Maximum of two missed cleavages were accepted. Precursor mass tolerance was set to 10 ppm and for the MS/MS fragments it was set to 0.02 Da. Between 3 and 25 fragments were collected per peptide. The phosphosite localization algorithm was set so that phosphosites with a score that was equal or higher than 0.75 were considered as Class I. Filtering was performed at a 1% false discovery rate (FDR) for all the peptides and proteins that were used to construct the spectral library.

The 122 DIA raw files were analyzed in Spectronaut X. In the transition settings, charges +2 and +3 were set for precursor ions, and +1, +2 and +3 for b- and y- ion products with a mass tolerance of 0.02 Da. Both precursor and protein Q value cutoffs were set to 0.01 and the peptides were quantified based on the intensity of the MS1 signal precursor. In all samples, the retention time alignment was performed with spiked-in iRT peptides.

From the database search, a total of 45,356 phosphosites in 29,484 phosphopeptides were identified with an average of 18,722 phosphosites per sample (Table SX). The data displayed 58.7% missing values in the phosphosite abundance. The data were exported into Perseus software v. 1.6.2.3. Valid value filtering was applied and all phosphosites with more than 5% missing values were removed. The data were then log2-transformed and centered around zero by subtracting the median intensity in each sample. For those phosphosites with less than 5% missing values, the phosphosite abundance values were imputed by applying the K Nearest Neighbor (KNN) method, resulting in 4644 phosphosites corresponding to 1613 proteins in each patient that could be used for further analyses.

The sample preparation workflow was previously assessed for its reliability to produce data from malignant melanoma tissue samples (Murillo et al., 2018). Furthermore, principal component analysis, using 1,267 phosphosites commonly quantified among 118 patient samples showed similar separation as for the global proteomic dataset, separating the high- (>70%) and low-containing (<30%) tumor samples based on phosphosite abundance (Figure S-1D). Protein and phosphoprotein abundances showed strong positive correlation (median 0.506), and 94% significant correlation ($p<0.05$) for the 809 overlapping proteins across 94 patient samples (Figure S-1G).

IHC Staining Analysis

For the immunohistochemical study primary metastatic (n=42) melanoma tissues were used. Representative tissue areas from paraffin-embedded blocks were selected based on the HE-stained slides, then 5 mm circumferential columns were put into the tissue microarrays (TMAs) in an ordered manner. From TMAs 3.5µm sections were put into an automated immunostainer (Leica Bond Max, Leica Biosystems, US, IL) for standardized deparaffinization, rehydrating and staining protocol. Antibodies against ADAM10, CDK4, CTNND1, DDX11, FGA, HMOX1,NBP1, PAEP, PIK3c, and TEX30 were applied in a dilution and antigen retrieval according to ***Table 1***. For visualization, high affinity polymer-based, AF-linked secondary was used, with a chromogen substrate fast red. For negative controls open containers were filled with only primary antibody diluent without primary antibody. Before, coverslipping slides were counterstained with hematoxylin.

The colorimetric immunostained slides were scanned by 3D Histech slide scanner. The digitized high resolution pictures served as the basis for the densitometry using Image Pro Plus software. Multicolor pictures were converted into grayscale spectrum, then 5-5 representative areas for cell cytoplasm and/or nucleus of both melanoma and stromal cells were measured separately. The gained continuous scale variables were collected in an Excel file for statistical analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Study of SAAVs

1. Custom database construction and SAAV peptide identification

Custom protein sequence database was built by downloading protein mutation data from the Cancer Mutant Proteome Database (CMPD, http://cgbc.cgu.edu.tw/cmpd/, download date November 2018). This included the skin cutaneous melanoma data of TCGA (369 cases) and 7 melanoma cell lines from the NCI-60 panel. Additional data of melanoma samples was retrieved from COSMIC. Protein IDs, mutation positions and mutated protein sequences were extracted, and a UniProt ID was assigned to the proteins. Then *in silico* mutation was applied to the UniProt canonical protein sequences to verify the validity of the mutated sequences. Using the matched protein sequence a peptide that carried the mutation site was generated by performing an in silico tryptic digestion of the protein, and allowing for one additional missed cleavage at both sides of the mutation site. Redundant mutations were then removed and entries with the same

mutated peptide sequence were grouped into one single entry. The resulting database comprised 57,134 entries.

Raw files were processed with Proteome Discoverer 2.3 (Thermo Scientific) using the Sequest HT search engine in a two-step search. The first search was performed against the *Homo Sapiens* Swissprot database, and unassigned MS/MS spectra were searched against the above described in-house built database. Cysteine carbamidomethylation was set as fixed modification while methionine oxidation and TMT 11plex at peptide N-terminus and lysine were set as variable modifications; peptide mass tolerance for the precursor ions and MS/MS spectra were set to 10 ppm and 0.02 Da, respectively. A maximum of two missed cleavage sites were accepted and FDR were set at 0.01 for identification at peptide level.

2. Validation of search results and data cleanup

SAAV peptides were validated using SpectrumAI quality control tool available as an R script (Zhu et al., 2018). SpectrumAI automatically inspects the MS/MS spectra of the peptide sequences. In order to pass the quality control, the matching MS2 peaks needed to be present (within 0.02 Da fragment ion mass accuracy) for both the b and y ions which confirm the change in the amino acid sequence. The presence of only b or only y ions were sufficient when the peptide had a proline residue adjacent to the substituted amino acid on its N-terminal side, due to the thermodynamically unfavored fragmentation on the C-terminal side of a proline residue. Additional criterion was set for the ion intensity. The sum intensity of the flanking MS2 ions was required to be larger than the median intensity of all fragmentation ions. SAAVs for which the substitution occurred on the 0th position also passed the quality control if the sum intensity of the supporting b ions was larger than the median intensity of all fragmentation ions.

A custom R script was used for data cleanup and post processing. The verified SAAVs pointing at the same mutation position on a protein were merged into one entry. The reason for multiple entries included missed cleavages as well as complementary peptides pointing at the same mutation. The latter occurred if the amino acid change generated a new trypsin cleavage site leading to a peptide that cannot be predicted from the original canonical sequence. For peptides which were assigned to an isoform of the master protein, the mutation positions were corrected to reflect the position in the canonical Uniprot sequence. This was performed by using a customized script analyzing the UniProtKB isoform sequences (accessed on 21 August 2019). Both the corrected and uncorrected position was used for online database searches, to ensure that we would not miss matching results due to the position disparity caused by isoform sequences. Additionally, the matching wild-type peptide PSMs originating from the normal database search were linked to the corresponding SAAVs, which allowed to assess the ratio of wild type and SAAV peptide PSMs.

3. Annotation of validated SAAVs

Merging the results of various database searches and cleaning the data was performed with inhouse custom R scripts. First the SAAV peptides were searched in PeptideAtlas database (Desiere, 2006) to find out if the SAAV peptides were already observed previously in another study. The search was performed on the webpage https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetPeptides using the “Human 2020-01” Atlas Build and only keeping the canonical and isoform protein accessions for which SAAVs were identified in our study. The resulting peptide sequences were downloaded in text format and custom R script was used to retrieve exact and partial matches.

Validated coding SNPs and cancer-related mutations were downloaded from the CanProVar database (Li et al., 2010) (http://canprovar2.zhang-lab.org/datadownload.php, version 2.0). The UniProt IDs were first converted to Ensembl IDs using the biomaRt (version 2.42.1) R package (Durinck et al., 2005, 2009), and then the CanProVar database was used to retrieve the variant’s reference SNP ID (rs#) and any cancer related variation ID of CanProVar. Additionally, the “Index of human polymorphisms and disease mutation” document was downloaded from UniProt (https://www.uniprot.org/docs/humsavar) and was also used to retrieve reference SNP IDs, as CanProVar database has not been updated since 2012.

Aggregated Allele Frequency (ALFA frequency) frequencies were accessed using the NCBI Variation Service API as described in https://github.com/ncbi/dbsnp/blob/master/tutorials/Variation%20Services/Jupyter_Notebook/by_rsid.ipynb. For this analysis a custom Python script was used. Additional resources such as ExAc (Lek et al., 2016), 1000Genomes (2015), HapMap (2003) were used to manually retrieve allele frequency information when this information was missing from the ALFA frequency query.

4. Selection of over- and under-represented SAAVs

Firstly, we calculated PSM ratios (PSMr) using the following formula:

$$PSMr = (n_{SAAV\ PSM}) / (n_{SAAV\ PSM} + n_{wild\text{-}type\ PSM}),$$

where $n_{SAAV\ PSM}$ is the number of PSMs supporting the SAAV peptide, while $n_{wild\text{-}type\ PSM}$ is the number of PSMs supporting the wild-type peptide.

The enrichment factor of SAAVs (SAAVr) was then defined as

$$SAAVr = PSMr / AAF,$$

where PSMr is PSM ratio (see above) and AAF is alternative allele frequency in the European population.

Log_2-transformed SAAVr values (n=760) were subjected to Johnson transformation using Minitab (vs 17) to achieve values following the normal distribution. Significance level was set to $\propto = 0.1$, and so, values outside of the range [-1.627; 1.654] were considered outliers (i.e., were considered under- or over-represented variants). The obtained list was further filtered to select over-represented variants with highly confident identification (number of PSMs supporting the SAAV > 2, SAAVr > 4) and with AAF < 0.11. For variants where no wild-type PSM was identified, the number of verified PSMs had to exceed 10 and AAF < 0.11. Under-represented SAAVr variants were not further filtered and their PSMs ranged between 1-4.

Independent Component Analysis to Connect Pathway-level Features with Clinical Variable

Pre-processed and normalized proteomics, transcriptomics, and phosphoproteomics data, were dimensionally reduced by independent component analysis (ICA) separately. Unlike principle component analysis (PCA), which assumes Gaussian distribution of the data, ICA can be used for non-Gaussian distributed data sets. To ensure the quality of the ICA, we only included omics data of samples with tumor content larger than 30%, this resulted in 111 samples in the proteomics dataset, 118 for phosphoproteomics and 134 for transcriptomics. ICA was carried out at per-accession level for proteomics data and per-modified-sequence (phosphosite) level for phosphoproteomics data. An R-based package, “fastICA”, was used for implementation. The ICA was performed 100 times for each omics dataset to make sure that the ICs were consistent. The extracted independent components (ICs) mixing scores of the omics data were then passed through association tests with the joint table of clinical features of patients in our cohort. If the clinical variable is binary, a logistic regression model was built for association tests. Otherwise, a linear regression model was built. The association tests were conducted for all the 100 ICA analyses for each omics dataset and its ICs, and the ICs showing significant correlations (p-value<0.00001) with a clinical feature for at least 50 out of 100 ICA runs were picked as significant ICs (these conditions were called here the strict criteria). Alternatively, applying “relaxed criteria”, the ICs showing p-values smaller than 0.005 for at least 30 ICA runs per 100 were picked as significant. For each of these significant ICs, the centroid of IC coefficients were used to rank the omics data where accessions in proteomics and modified sequences in phosphosites were matched to their corresponding gene names. We then used these rankings to conduct Gene Set Enrichment Analysis (GSEA) and significant pathways were found (adjusted p-value<0.01). The GSEA was implemented by an R-based package, “fgsea”, and searched against the “Reactome” database. The ICs served as links between clinical features and pathways at this point. In order to have a systematic understanding of the multi-omics datasets, we gathered significant pathways from proteomics, transcriptomics, and phosphoproteomics data that their ICs were associated with the same clinical variables (Liu et al., 2019).

Outlier Analysis

Outlier analyses were performed for different variables of interests including survival, BRAF mutation, NRAS mutation, gender, and tumor stage. For the survival related variables, the dataset was divided into 2 groups and binary variables were created based on whether the patients lived longer than certain cutoff times (5 years, 3 years, 1 year, 6 months) from their sample collection (surgery) date or not. Outlier analyses were conducted to find which genes (proteins) were significantly enriched in one of two groups but not in another group. These analyses were carried out at per-accession level for proteomics data and per-modified-sequence level for phosphoproteomics data. We used a python-based package, “BlackSheep”, to implement these analyses on proteomics, transcriptomics, and phosphoproteomics data separately with default median and interquartile range (IQR) of 1.5. Significant genes were picked by FDR cutoff at 0.05 in the group-wise comparisons. Genes (protein isoforms) labeled as outliers in less than 30% of patient samples in one group were excluded from the group-wise comparisons. Picked outlier accessions in proteomics and modified sequences in phosphoproteomics were matched to their corresponding gene names (Blumenberg et al., 2019).

Cox’s proportional hazards survival analysis

We also performed survival analysis using regularized Cox regression in a similar manner as (Yuan et al., 2014). The samples were randomly split into a training and a test set (80-20 training-test set). Using the training samples, a univariate Cox model was fitted for each feature individually and the 30 features with the lowest univariate p-values were selected and used as input to an elastic-net Cox model. The C-index was computed on the left out test samples. This procedure was repeated 100 times for each omics dataset. We then considered the features that were selected by the Cox model in at least 50 of the 100 repetitions as significant and investigated these further (Yuan et al., 2014).

ROC curve analysis

Survival data of the samples were stratified at 6 months, 1 year, 3 years, and 5 years into binary variables. Univariate receiver operating characteristic (ROC) curves of each binary survival variable and each protein expression were constructed by the ‘pROC’ package in R. Area under the ROC curve (AUC) was used as a measurement to determine the correlation between survival and the expression of specific proteins. For each protein, the cutoff point of expression that gave the maximum sum of sensitivity and specificity was used to divide the samples into a high expression group and a low expression group. Using an R package called ‘survival’, Kaplan-Meier curves were then introduced to explicitly reveal the survival differences between samples in these groups. The Kaplan-Meier (log rank) test p-values were also calculated with ‘survminer’ in R, which were used as another statistical value to evaluate the relationship between survival and the expression of specific proteins.

Clinical data of samples in the IHC validation cohort (independent cohort)

The melanoma samples involved in the IHC validation cohort were collected from the Department of Dermatology and Immunology of the University of Szeged. 42 patients were selected from 2001 to 2020 whose primary melanoma is archived in paraffin-embedded tissue blocks. All primary tumors resulted loco-regional and/or disseminated disease. The tissue microarrays (TMAs) were made from formalin-fixed, paraffin-embedded (FFPE) blocks, represented 42 primary melanomas.

A total of 42 samples were collected with the clinical information including gender (male = 24, female = 19), age at primary tumor (mean = 61.35 yrs, SD ± 10.158, n = 43), localization of primary tumor (trunk = 24, lower limbs = 8, upper limbs = 7, head and neck region = 3, acral region = 1) and metastases, disease-free survival interval (mean = 23.16, SD ± 38.77, n = 43), progression-free survival interval

(mean = 54.11, SD ± 49.73, n = 43), overall survival (mean = 58.62, SD ± 50.29, n = 43), live status (marked by 0, meaning “dead” patients, n = 27; marked by 1, meaning “alive” patients, n = 16), histological subtypes (SSM, NM, ALM, LMM etc.), pathological TNM staging (according to AJCC cancer staging system, 8th edition), histological parameters of the tumor (Clark level, Breslow, presence of regression and ulceration), BRAF status, and long term follow up data (DFS, PFS, OS). The clinicopathological data of the samples were collected in an Excel file for statistical analysis.

All patient samples were obtained with the approval of the Research Ethics Committee in the University of Szeged with written informed consent provided by all participants. Ethical authorization number is MEL-PROTEO-001, 4463-6/2018/EÜIG.

P-M correlation analysis

Based on the IHC validation cohort, nine additional metastases were observed for the purpose of protein-based correlation between the primary melanoma and the metastasis. First, a one-sample Kolmogorov-Smirnow test was conducted to determine the normality of the protein expression data. For the comparison, Wilcoxon signed-rank test was used to examine whether proteins were differentially produced between the tumors and the matched metastases. P-values of One-Sample Kolmogorov-Smirnow test and Wilcoxon signed-rank test were calculated by the IBM SPSS statistics package (26.0 version) software. $P < 0.05$ was considered statistically significant.

Survival analysis - Roc curve and Kaplan-Meier analysis in the IHC validation cohort

First, to show the predictive impact of the identified 9 markers in progression, we performed independent T-test to assess the differences of the production of each protein using two categorical variables, “alive” (live status, marked by 1) and “dead” (live status, marked by 0) patientgroups, and using one continuous dependent variable, the means of the protein expression values. The assumption of homogeneity of variances was tested by Levene's Test of Equality of Variances.

The receiver operating characteristic curve (Roc curve) was used for the graphical illustration of the expression of related proteins based on their diagnostic ability on survival rates (binary classification). The Roc curve was generated by plotting the True Positive Rate (TPR) (on the y-axis) against the False Positive Rate (FPR) (on the x axis) and was calculated based on the optimal cutpoints (coordinates of Roc curve were measured) of each protein.

Based on the cutoff points of the indicated proteins, the area under the curve (AUC) measures the degree of separability between two patients’ group according to their survival rate.

Kaplan-Meier survival analyses were conducted with the DFS, PFS and OS intervals (measured with months) and they were calculated (KM, log-rank test) based on models generated by the optimal cutpoint of each protein.

The independent T-test, Roc curve, Kaplan-Meier survival analysis and figures including box plots showing p-values, quartile values, mean values and 95% confidence intervals (CI) were produced by IBM SPSS statistics package (26.0 version) software. $P < 0.05$ was considered statistically significant.

Identification of mortality risk subgroups of BRAF V600E mutated patients

The R package ‘InGRiD’ (Wei et al., 2019) was utilized to identify subgroups of patients with different mortality risk rates within a cohort of 49 patients with BRAF mutation. This package provides a pathway-guided identification of patient subgroups based on protein expression while utilizing patient survival information as the outcome variable. The analysis was done using the expression of proteins that belong to pathways previously linked to tumors with different expression levels of BRAF V600E (Betancourt et al., 2019).

DATA AND CODE AVAILABILITY

Github repository: https://github.com/rhong3/Segundo_Melanoma

FIGURES

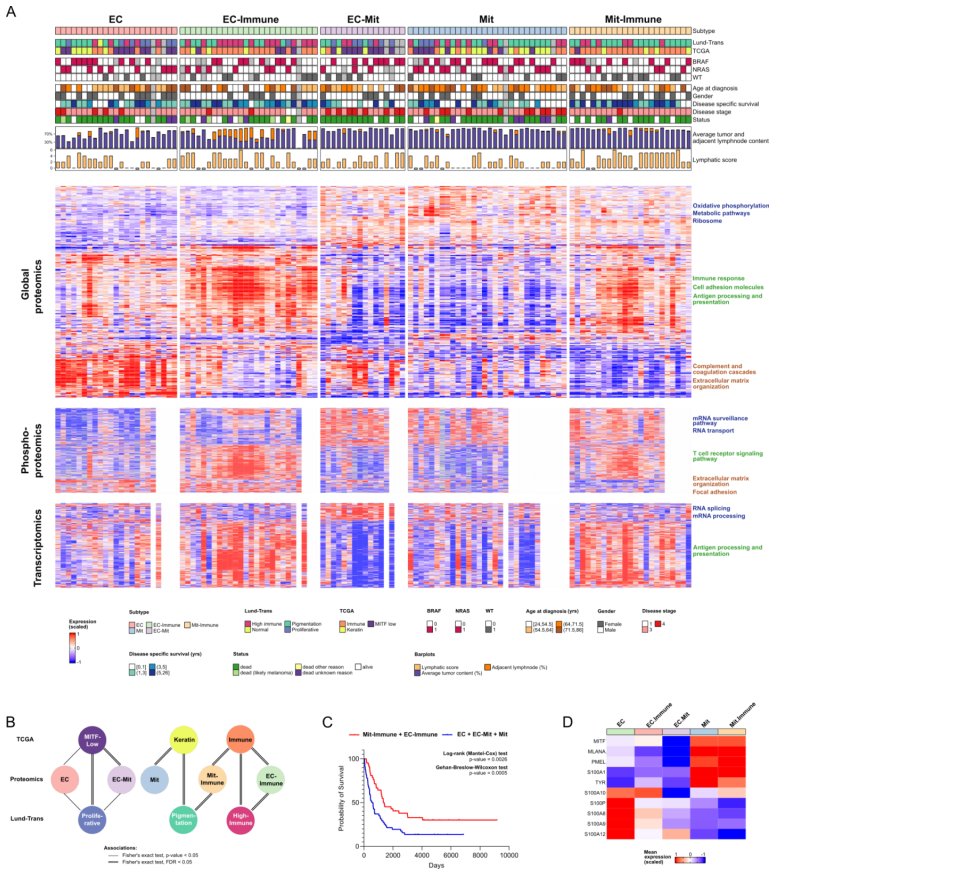


Figure 1. Proteomic classification of Metastatic Melanoma. A. Proteogenomic profiles of our melanoma sample cohort. Samples are grouped based on the proteomic subtypes established in this study (EC, EC-Im, EC-Mit, Mit, Mit-Im). Transcriptomic classifications (Lund, TCGA) as well as their important clinical and histological data are displayed for each sample. The heatmaps show the most differentially regulated proteins (ANOVA top500, FDR<0.005), phosphosites (ANOVA top1000, FDR<0.05) and transcripts (ANOVA top500, FDR<0.05) among the five proteomic subtypes, and the molecular clusters are annotated with representative pathways. The subtypes exhibit similar pathway-level features on all molecular levels. **B**. Networks representing the association between subtypes defined by the transcriptomic and proteomic classification systems. The subtypes are denoted by nodes and the significance of the association was computed by Fisher's exact test and represented by double and single lines for FDR<0.05 and unadjusted p-value<0.05, respectively. **C**. Kaplan-Meier survival plots displaying the disease-specific survival probability for patients with tumors assigned to proteomics subtypes and distributed between long survival (EC-Immune and Mit-Immune) and short survival (EC, EC-Mit, and Mit) subgroups.The significance of comparisons is shown by the p-values derived from log-rank (Mantel-Cox) and Gehan-Breslow-Wilcoxon tests. **D**. Mean scaled expression of melanoma markers in each subtype.

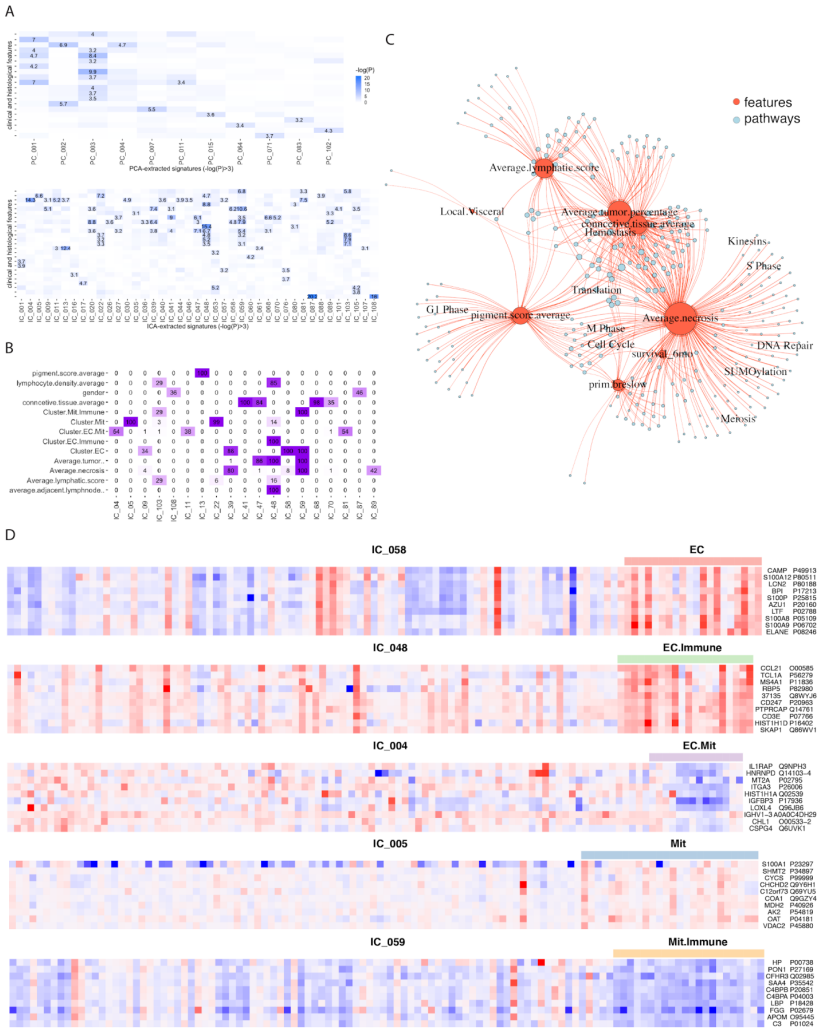


Figure 2. Independent Component Analysis (ICA) extracted high-level features from proteomics data. A. Correlations of extracted proteomics signatures from PCA and ICA respectively with clinical and histological features with P-value in -log scale threshold of 3. **B**. Counts of significant correlations between proteomics extracted independent components (ICs) and clinical and histological features (P-value < 0.00001), ICA repeated 100 times. ICs with significant correlations observed fewer than 30 times are excluded. **C**. Interconnections between pathways and clinical and histological features based on proteomics ICA-GSEA (P-value < 0.0005). Clinical and histological features are shown as red nodes while pathways are shown as blue nodes. The sizes of vertices are proportional to numbers of incoming and outgoing edges. **D**. Top 10 proteins contributing to the ICs correlated with each of the 5 subtypes.

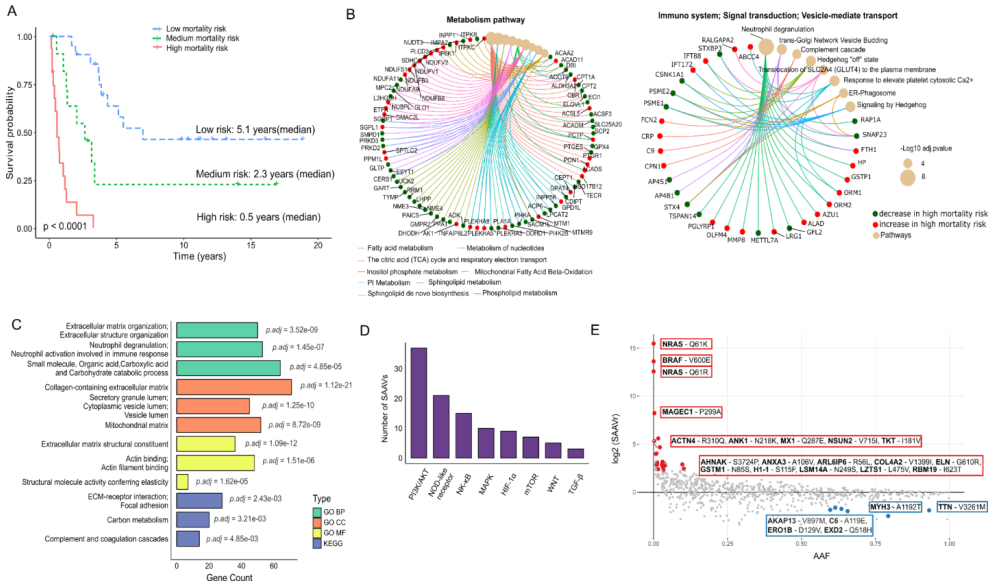


Figure 3. Insights from studying BRAF V600E mutated metastases and single amino acid variants in melanoma. A. Kaplan-Meier curves of the molecularly-defined subgroups of BRAF mutated patients. Patient subgroups are color-coded according to survival probabilities: red (high-risk of mortality, n=16), green (medium-risk of mortality, n=12) and blue (low-risk of mortality, n=21). Median survival times for the three patient groups is shown. **B**. Distribution of the proteins that belong to the most enriched significant Pathways involved in the identification of mortality risk subgroups of patients with BRAF mutation. **C**. KEGG and GO enrichment analysis of the 828 proteins with SpectrumAI validated SAAVs. The 3 most significantly (Benjamini-Hochberg adjusted p-value < 0.05) enriched GO terms in biological process, molecular function, cellular component and KEGG pathways are presented. Similar GO terms were collapsed into one entry and adjusted p-values were then averaged. The numbers to the right of each bar represent the p-values associated with each term. **D**. Representation of SAAVs in our melanoma cohort linked to signaling pathways. The number of SAAVs are indicated for each pathway. **E.** Relationship between the AAF (alternative allele frequency in the European population) and $\log_2$ SAAVr, where SAAVr is the ratio of PSMr to AAF (PSMr = $(n_{SAAV\ PSM})/(n_{SAAV\ PSM} + n_{wild\text{-}type\ PSM})$). Out of all the 760 SAAVs with information on AAF and PSMr, 19 were found to be over-represented (red) and 6 were found to be under-represented (blue). The corresponding gene symbols for these variants and the amino acid changes are shown.

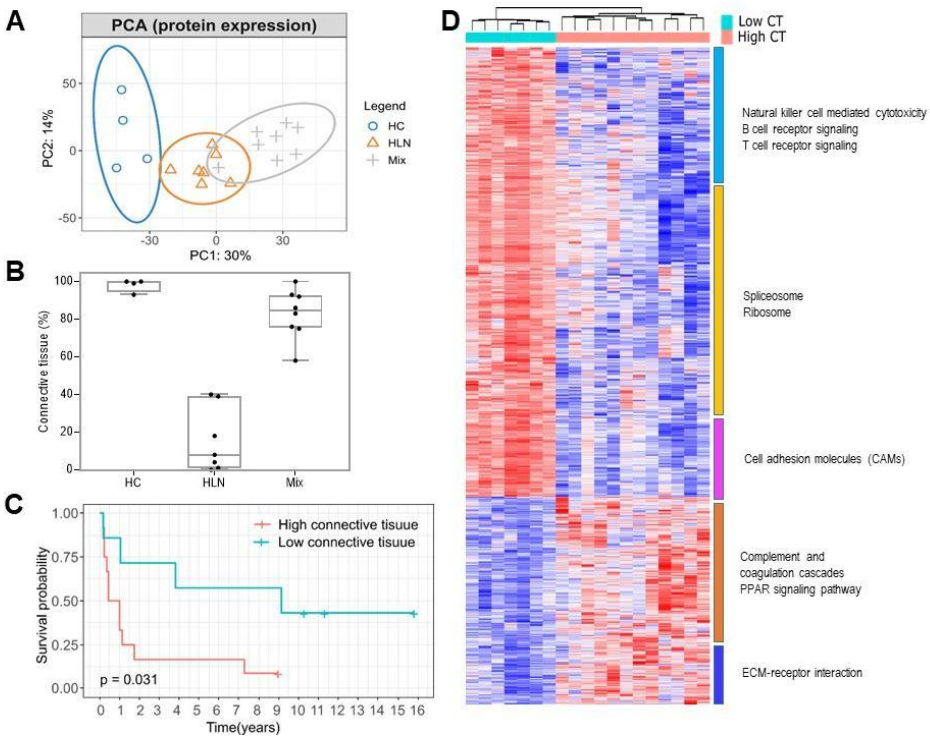
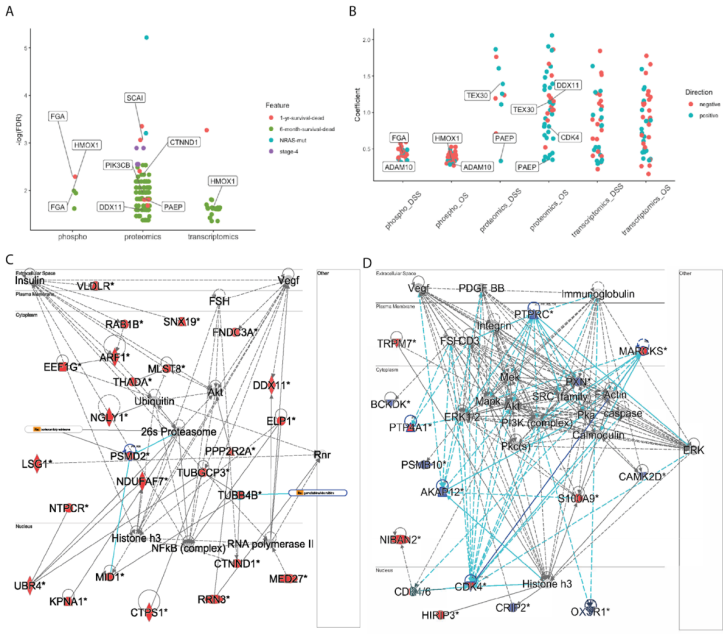


Figure 4. The composition and proteome profiles of surrounding TME of melanoma samples with <30% tumor cell content are associated with patient survival. **A.** PCA of commonly quantified proteins shows clusters of samples associated with tissue composition. HC, HLN, and Mix denote the clusters of samples with the highest connective tissue content (³92%), highest adjacent lymph node content (³60%), and samples with intermediate values of these features, respectively. **B.** Box plot representation of the connective tissue content for the three sample clusters. The Anova analysis showed significant differences (p-value<0.0007) between the means of the HC and Mix compared with the HLN cluster. This result defined the subgroups of high and low connective tissue (CT) for subsequent analyses. **C**. Kaplan-Meier survival plots displaying the disease-specific survival probability for patients with tumors of high and low CT. **D.** Heatmap for the 2,213 significantly differentially expressed proteins (t-test, FDR<0.05) between the groups of tumor samples of high and low CT in the TME.



E

Accession (Uniprot)	P11802	O60716-3	Q96FC9	P09466	P42338	Q8N9R8	Q5JUR7	O14672	P09601	P02671
Protein Symbol	CDK4	CTNND1	DDX11	PAEP	PKI3CB	SCAI	TEX30	ADAM10	HMOX1	FGA
Protein Name	Cyclin-dependent kinase 4	Catenin Delta-1, isoform 1A	ATP-dependent DNA helicase DDX11	Glycodelin	Phosphatidylinositol 4,5bisphosphate 3-kinase	Suppressor of cancer cell invasion protein	Testis-expressed protein 30	A Disintegrin and metalloproteinase 10	Heme oxygenase (decycling) 1	Fibrinogen alpha chain
Drug Target	Yes	No	No	No	Yes	No	No	Yes	No	No
Approach	COX Regression	Outlier Analysis	Both	Both	Outlier Analysis	Outlier Analysis	COX Regression	COX Regression	Both	Both

Figure 5 Survival-related biomarkers discovered by outlier analysis and cox regression analysis.

A. FDR in -log scale for significant putative biomarkers associated with short survival, tumor stage, and NRAS mutation found by outlier analysis in multi-omics data. Candidate proteins selected for validation are labeled. **B**. Cox coefficients of significant putative biomarkers associated with survival in Cox regression analysis of multi-omics data. Candidate proteins selected for validation are labeled. The direction represents whether the expression positively or negatively affects the survival. **C**. Ingenuity Pathway Analysis (IPA) for proteins related to survival in the outlier analysis (red), first top relationship subnetwork. Drug targets are labeled. **D**. IPA for proteins and phosphoproteins related to survival in the Cox analysis, top relationship subnetwork. Entities with expression correlated to high hazard shown in red. Those with expression correlated to low hazard shown in blue. **E**. The ten proteins selected for validation by IHC (green: significant in proteomics data, blue: significant in phosphoproteomics data).

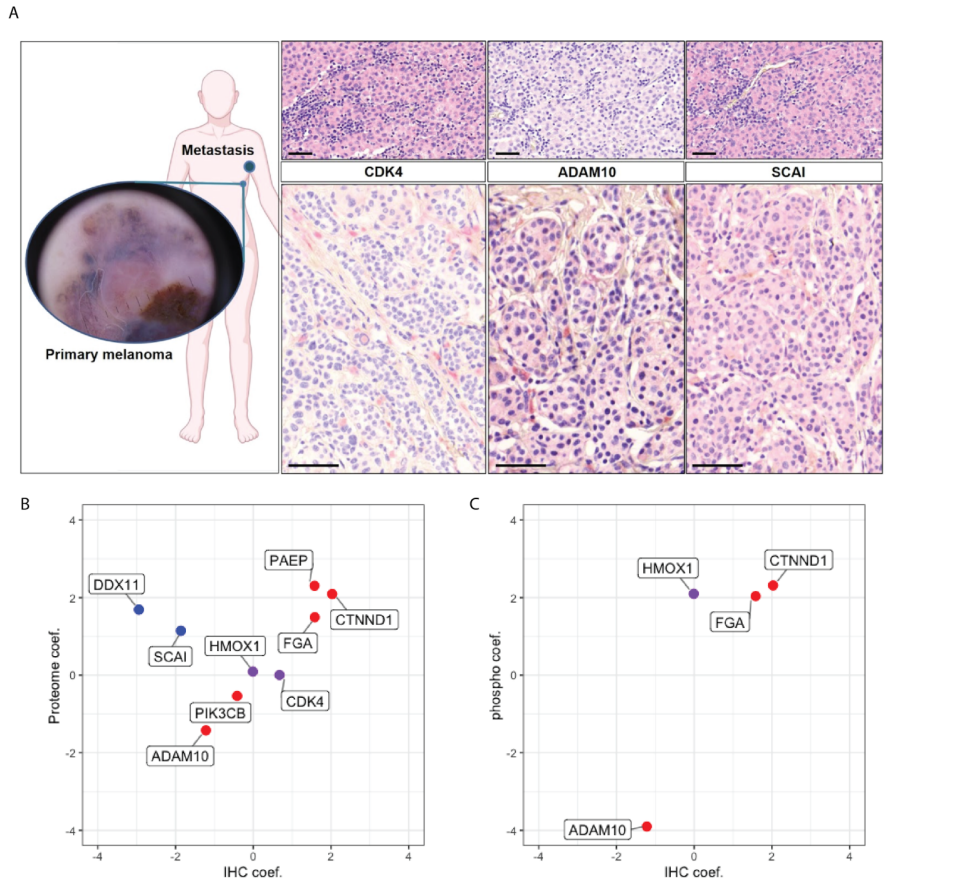


Figure 6. Case presentation and Cox Regression Validation Models. A. left: dermoscopic image of the primary breast skin melanoma with atypical features, focal regression; right tissue marker expression in the primary tumor (downstream) and its lymph node metastasis (upstream); OM 112x; scale bar 50μm. Forty-two year old female melanoma patient was referred to the radiology because of a palpable axillary lump. Its cytology revealed malignancy. Behind the metastatic diseases a suspicious mole was noted on the breast skin. Histology and clinical staging showed AJCC –IIIB disease (primary: pT3a; lymph node: pN2b). BRAFV600E mutational state was positive. During the induction of adjuvant PD1 blocker therapy the patient showed a metastatic cerebral disease, therefore topical irradiation was induced followed by BRAF and MEK target inhibitor therapy. Her metastatic disease rapidly progressed and she died (DFS=0, PFS=4, OS=14 months). **B, C**. Horizontal axis shows the z-score of Cox coefficients of protein melanoma cell expression in IHC-based Cox model for the validation cohort while vertical axis shows the Cox coefficients of proteins in proteomics (**B**) and phosphoproteomics (**C**) Cox models for the discovery cohort. Proteins with consensus of coefficient directionality are colored in red while those with different directionality are colored in blue. Proteins with minimal z-score (between -0.5 and 0.5) in either cohort are colored in purple.

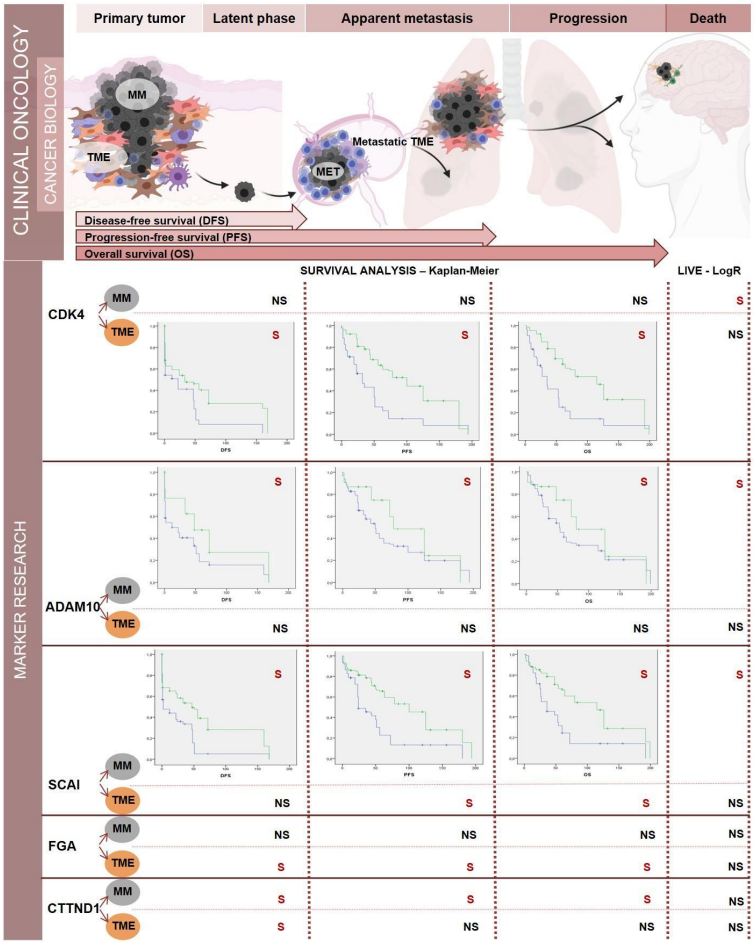


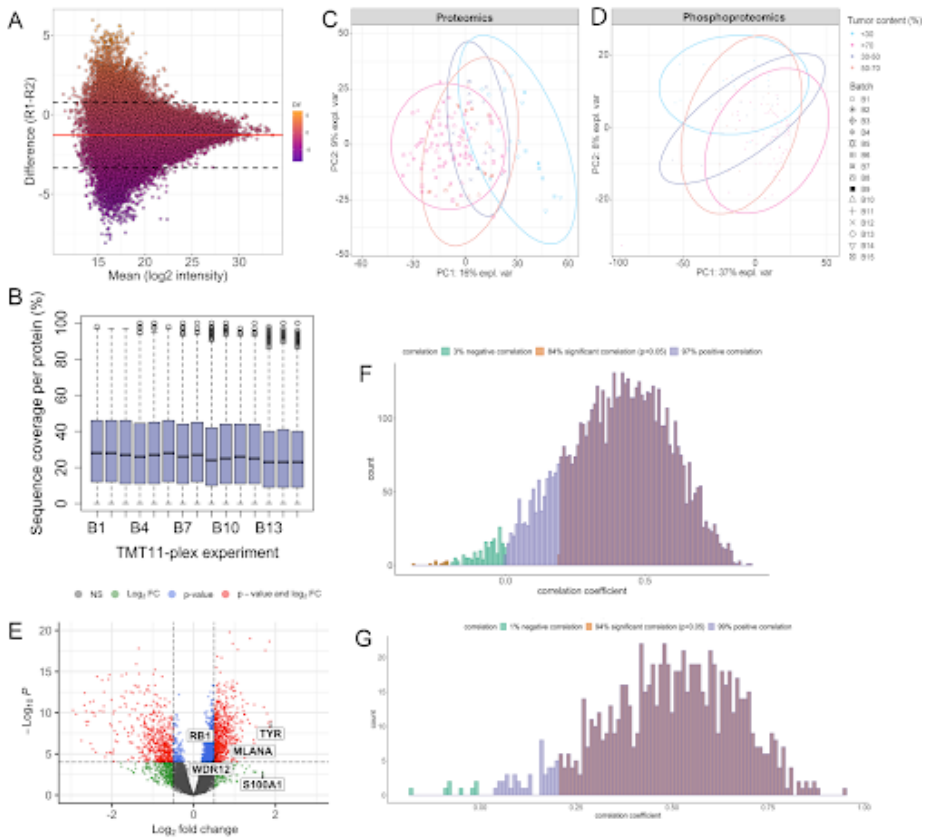
Figure 7. The complexity and layers of melanoma as a biological disease. Primary melanoma on the skin shows a marked heterogeneity both on melanoma cellular level but also in its microenvironmetal counterpart (TME) including lymphoid-, histiocytic- and fibroblastic elements. The detached but still dormant derivates of the primary melanoma may be disseminated in the body in an undetectable way forming the minimal residual disease of the latent clinical phase. For the apparent metastatic disease the surrounding metastatic niche of TME is necessary for the progression into visceral and cerebral dissemination leading to death. Although the steps are usually sequential, the time courses largely differ among patients highlighted by DFS, PFS and OS values in the clinical oncology. As the clinical behaviour of each melanoma is identical, wide-range of proteomic biomarker research was called to life by the personalized follow up and treatment strategies providing new prognostic and predictive tissue biomarkers.

ANTIBODY	HIER-BUFFER	DILUTION	INCUBATION TIME (min)
ADAM10	pH=9	1:300	60
CDK4	pH=9	1:100	20
CTTND1	pH=9	1:50	20
DDX11	pH=9	1:150	20
FGA	pH=9	1:100	20
HMOX1	pH=9	1:100	20
NBP1	pH=9	1:200	20
PAEP	pH=9	1:150	20
PIK3c	pH=9	1:150	20
TEX30	pH=9	1:100	60

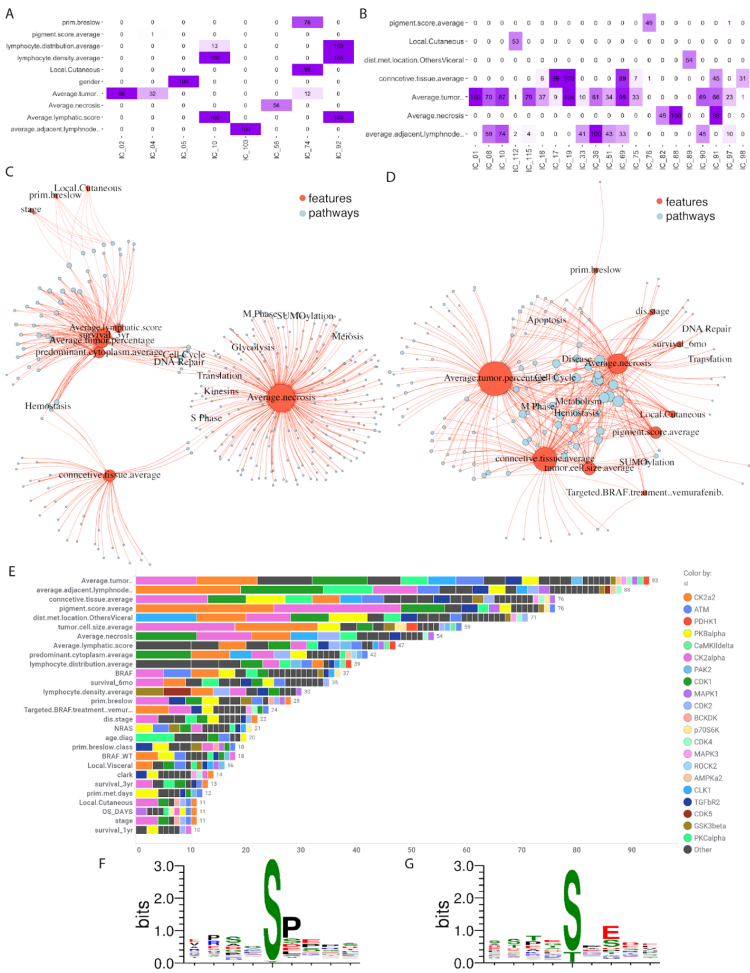
Marker	Antibodies	Source	Identifier
CDK4	Polyclonal Rabbit anti-Human CDK4 Antibody	LSBio	Cat# LS-C99873
CTNND1	PathPlusTM Polyclonal Rabbit anti-Human CTNND1	LSBio	Cat# LS-B14421
FGA	IHCPlus™ Polyclonal Rabbit anti-Human FGA	LSBio	Cat# LS-B11024
PIKc3	Monoclonal Mouse anti-Human PIK3CB	LSBio	Cat# LS-C105010
PAEP	IHCPlus™Polyclonal Rabbit anti-Human PAEP	LSBio	Cat# LS-B10557
ADAM10	Polyclonal Rabbit anti-Human ADAM10 Antibody	LifeSpan Biosciences	Cat# LS-C289
HMOX1	IHCPlus™ Monoclonal Mouse anti-Human HMOX1	LSBio	Cat# LS-B3407
SCAI	SCAI antibody	Novus Biologicals	Cat# NBP1-86711
DDX11	Monoclonal Mouse anti-Human DDX11	LSBio	Cat# LS-C133879

Table 1 represents the applied primary antibodies and their work package.

SUPPLEMENTARY MATERIALS

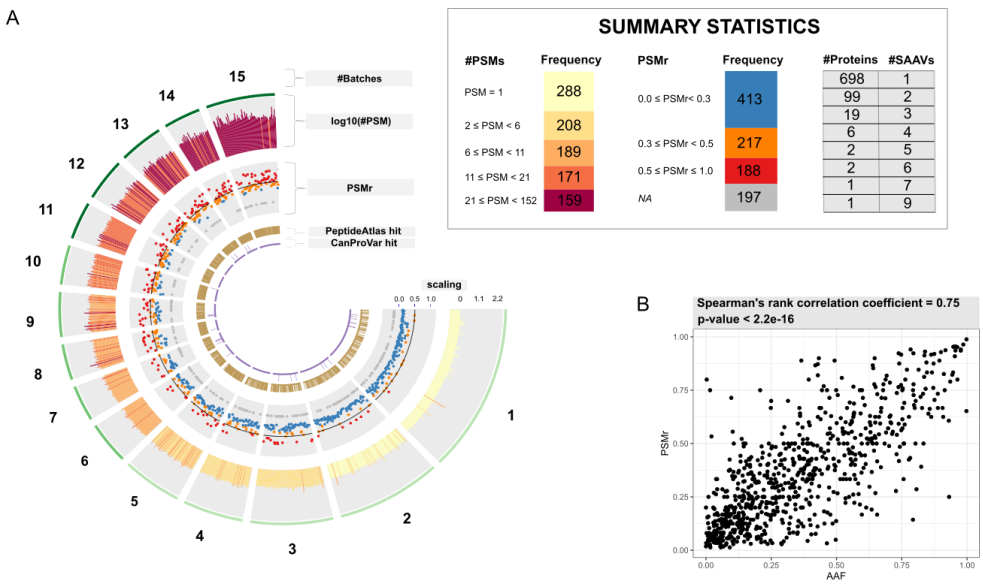


Supplementary Figure 1. A. Bland-Altman plot depicting the agreement across repeated experiments of batch one (B1), upper limit: 0.770, lower limit: -3.363 and mean difference: -1.297. **B**. Distribution of sequence coverage of the identified proteins by MS/MS across the fifteen TMT11 plex batches, (whiskers show the 5–95 percentiles and the dots represents the outliers). **C**. Principal component analysis of the TMT global proteome data after normalization and ratio calculation. The ellipses represent the 95% confidence interval per group based on 8125 proteins. **D**. Principal component analysis of 1267 commonly quantified phosphosites. The ellipses represent the 95% confidence interval per group. **E**. Volcano plot showing known melanoma protein markers and drivers in the high tumor-containing samples (>70%) compared to the samples containing mostly tumor microenvironment (<30%) (p-value 0.05, log2 fold change |0.5|). **F**. mRNA and protein abundance correlation (median 0.408), 84% of the mRNA and protein pairs (6101) showed significant correlation (p-value 0.05) across 104 patient samples. **G**. Protein and phosphoprotein abundance correlation (median 0.506), 94% of the proteins and phosphoprotein pairs (809) showed significant correlation (p-value 0.05) across 94 patient samples.



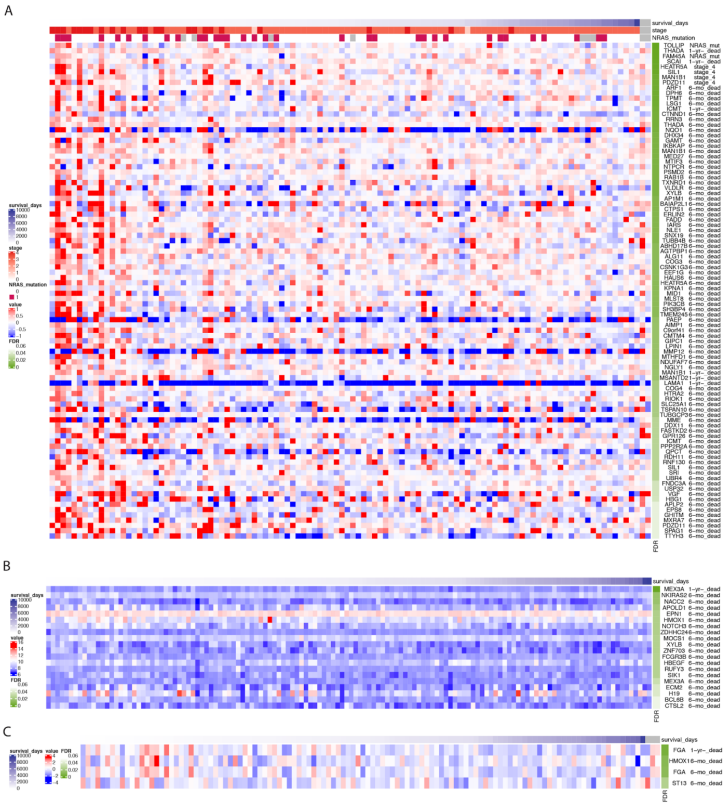
Supplementary Figure 2. ICA and ICA-GSEA on transcriptomics and phospho-proteomics data. A, **B**. Counts of significant correlations (P-value < 0.00001) in 100 times ICA between transcriptomics (**A**) and phospho-proteomics (**B**) extracted independent components (ICs) and clinical and histological features. ICs with less than 30 times of significant correlations are excluded. **C, D**. Interconnections between pathways and clinical and histological features based on transcriptomics (**C**) and phospho-proteomics (**D**)

ICA-GSEA (P-value < 0.0005). Clinical and histological features are shown as red nodes while pathways are shown as blue nodes. The size of vertices is proportional to how many edges are pointed to and from them. **E**. Relationships between phosphosites related to clinical and histological parameters via ICA and kinases predicted to generate the phosphosites. Graph shows the phosphosite counts, colored by the kinases generating them belonging for the set of all analyzed parameters selected based on the ICA methodounts of phosphosite-kinase relationships predicted by Netphorest/Networking for different clinical and histological parameters. For each phosphosite, only the top predicted kinase is considered (the values of the predicted kinase refer to the Netphorest score >0.42 or NetworKIN score >5). **F, G**. Graph shows the amino acid motifs for sets of the most important phosphosites of the ICA components associated with tumor stage (**F**) and 3-year survival (**G**) (P-value < 0.0005).



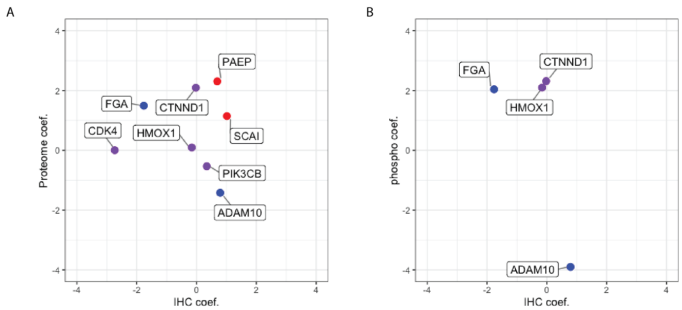
Supplementary Figure 3. Overview of single amino acid variants validated in melanoma. A. Summary of the 1015 validated SAAVs. On the circular plot, each SAAV is depicted as a separate entry. The variants are grouped by the ∑(number of batches) in which they were verified. From outside to inside: Track 1 depicts the number of PSMs for each SAAV on a logarithmic scale. Track 2 illustrates the PSM ratio (PSMr) of the variant. Track 3, and 4 represent the results from database searches (PeptideAtlas and CanProVar). On each track of the latter two, full lines represent “hits”, which in PeptideAtlas is interpreted as “Found in PeptideAtlas” (824 SAAVs) and in CanProVar as “Cancer-related SAAV” (27 SAAVs). On the embedded Summary Statistics table, the bar charts show the number of PSMs (#PSMs), the frequency of the PSMr and the distribution of the number of SAAVs per protein. Coloring scheme corresponds to the colors on the circular plot. The table on the right summarizes the data on protein level and demonstrates, in how many proteins were 2 or more mutated sites detectable. More than 70% of the SAAVs were detected with 2 or more PSMs which added confidence to the identification. Additionally, 2 or more SAAVs were found for 130 canonical proteins. The PSMr could be calculated for 80% of the SAAVs, and from these, ca. 50% (PSMr = 0.3 - 1.0) are suggested to be at comparable levels with the wild-type or to be the predominant proteoform in the analysed melanoma sample cohort. Circular plot was generated by the R package OmicCircos (vs. 1.28.0). **B**. The observed significant correlation between PSMr and AAF. The scatter plot entails 760 pairwise complete observations out of all the 1015. The result indicates that the PSMr is a valuable proxy/indicator of the abundance

of the SAAVs in comparison with the canonical sequence. AAFs were mainly extracted from the Aggregated Allele Frequency project through the NCBI Variation Service API (STAR Methods, Study of SAAVs 3.).

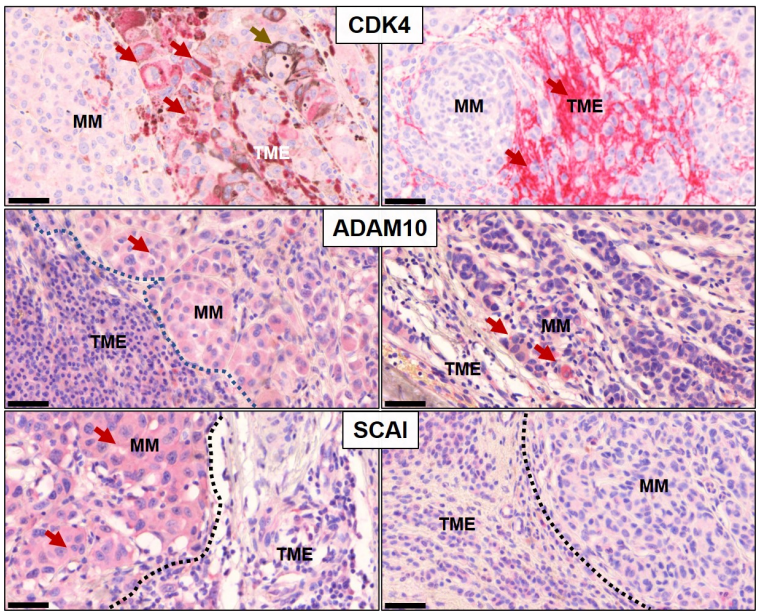


Protein Symbol	CDK4	CTNND1	DDX11	PAEP	PKI3CB	SCAI	TEX30	ADAM10	HMOX1	FGA
Best Expression Cut-off	-0.34	0.20	0.15	-0.60	0.15	0.15	-0.55	-0.21	0.18	-0.50
Best AUROC	0.70	0.68	0.57	0.66	0.64	0.56	0.64	0.63	0.64	0.60
Best Survival Stratification	6 month	6 month	6 month	6 month	6 month	1 year	6 month	6 month	5 year	6 month

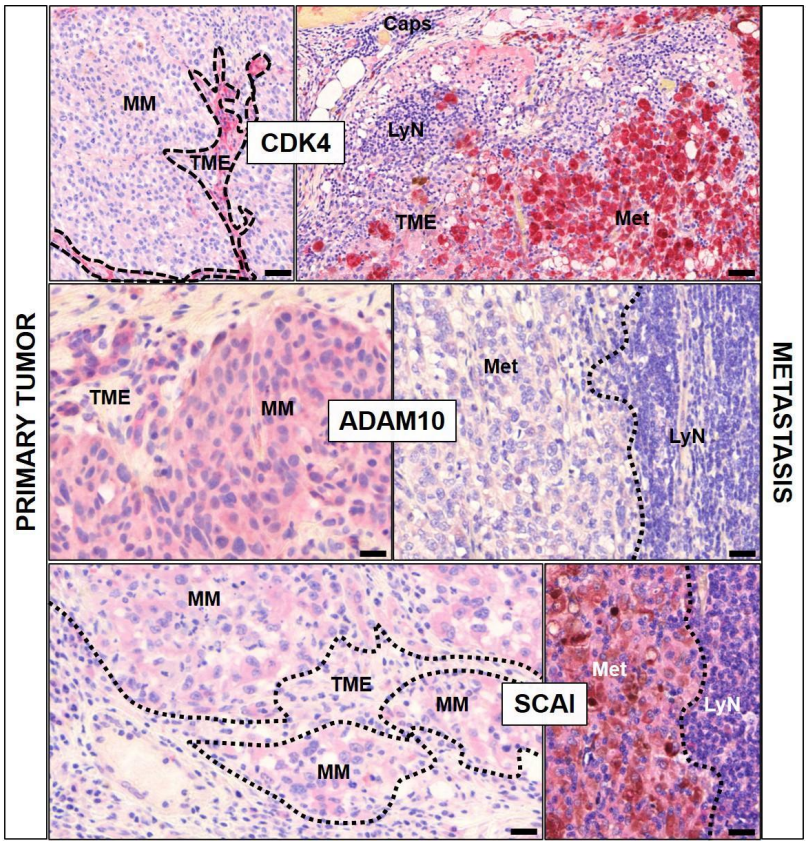
Supplementary Figure 5. Outlier analysis and Cox regression analysis. A. Significant proteins associated with survival, tumor stage, and NRAS mutation in proteomics data. **B**. Significant genes associated with survival in transcriptomics data. **C**. Significant phosphosites associated with survival in phosphoproteomics data. **D**. Best AUROC and corresponding stratification criteria of the ten proteins selected for validation by IHC (green: significant in proteomics data, blue: significant in phosphoproteomics data).



Supplementary Figure 6. Coefficients z-scores of Cox Regression Models. A, B. Horizontal axis shows the z-score of coefficients of genes' stromal cell quantity in IHC model of validation cohort while vertical axis shows the coefficients of genes in proteomics (**A**) and phosphoproteomics (**B**) models of discovery cohort. Genes with consensus of coefficients' directionality are colored in red while those with different directionality are colored in blue. Genes with minimal z-score (between -0.5 and 0.5) in either cohort are colored in purple.



Supplementary Figure 7a. Different expressional features of the candidate biomarkers. Tissue heterogeneity of CDK4 protein expression was shown with low positivity on the upper left side (MM), with stronger focal expression (red arrows) in the melanoma cells as well as in the TME highlighted by the red colorimetric (fast red) reaction. Note the native brown dyscoloration by melanin pigment accumulation. On the upper right side a more pronounced CDK4 stromal expression was seen in the TME. Similar heterogeneity was also noted at ADAM10 and SCAI markers.



Supplementary Figure 7b. Expressional differences of the candidate biomarkers in the primary - metastasis relation. Metastatic melanoma patients with a changed tissue protein expression. The primary melanomas compared to their metastases, exhibiting pronounced CDK4 and SCAI, but less ADAM10 colorimetric positivity in the lymph node metastases.

Bibliography

Aho, S., Levänsuo, L., Montonen, O., Kari, C., Rodeck, U., and Uitto, J. (2002). Specific sequences in p120ctn determine subcellular distribution of its multiple isoforms involved in cellular adhesion of normal and malignant epithelial cells. J. Cell Sci. *115*, 1391–1402.

Andrews, L.P., Szymczak-Workman, A.L., Workman, C.J., and Vignali, D.A. (2015). The extent of metalloproteinase-mediated LAG3 cleavage limits the efficacy of PD1 blockade. J. Immunother. Cancer *3*, P216.

Aslund-Ostberg, A.M., Marklund, B., and Hegen, C. (1992). [Outpatient clinics, open during evening hours for consultation on skin changes, attracted many visitors]. Lakartidningen *89*, 3923–3924.

Bentin Toaldo, C., Alexi, X., Beelen, K., Kok, M., Hauptmann, M., Jansen, M., Berns, E., Neefjes, J., Linn, S., Michalides, R., et al. (2015). Protein Kinase A-induced tamoxifen resistance is mediated by anchoring protein AKAP13. BMC Cancer *15*, 588.

Berto, G., Ferreira-Cerca, S., and De Wulf, P. (2019). The Rio1 protein kinases/ATPases: conserved regulators of growth, division, and genomic stability. Curr. Genet. *65*, 457–466.

Betancourt, L.H., Szasz, A.M., Kuras, M., Rodriguez Murillo, J., Sugihara, Y., Pla, I., Horvath, Z., Pawłowski, K., Rezeli, M., Miharada, K., et al. (2019). The Hidden Story of Heterogeneous B-raf V600E Mutation Quantitative Protein Expression in Metastatic Melanoma-Association with Clinical Outcome and Tumor Phenotypes. Cancers (Basel). *11*.

Bhattacharya, C., Wang, X., and Becker, D. (2012). The DEAD/DEAH box helicase, DDX11, is essential for the survival of advanced melanomas. Mol. Cancer *11*, 82.

Blackburn, J.B., Kudlyk, T., Pokrovskaya, I., and Lupashin, V. V (2018). More than just sugars: Conserved oligomeric Golgi complex deficiency causes glycosylation-independent cellular defects. Traffic *19*, 463–480.

Blumenberg, L., Kawaler, E., Cornwell, M., Smith, S., Ruggles, K., and Fenyö, D. (2019). BlackSheep: A Bioconductor and Bioconda package for differential extreme value analysis. BioRxiv 825067.

Boguslawska, J., Kedzierska, H., Poplawski, P., Rybicka, B., Tanski, Z., and Piekielko-Witkowska, A. (2016). Expression of Genes Involved in Cellular Adhesion and Extracellular Matrix Remodeling Correlates with Poor Survival of Patients with Renal Cancer. J. Urol. *195*, 1892–1902.

Brandt, D.T., Baarlink, C., Kitzing, T.M., Kremmer, E., Ivaska, J., Nollau, P., and Grosse, R. (2009). SCAI acts as a suppressor of cancer cell invasion through the transcriptional control of beta1-integrin. Nat. Cell Biol. *11*, 557–568.

Bunker, R.D., Bulloch, E.M.M., Dickson, J.M.J., Loomes, K.M., and Baker, E.N. (2013). Structure and function of human xylulokinase, an enzyme with important roles in carbohydrate metabolism. J. Biol. Chem. *288*, 1643–1652.

Cancer Genome Atlas Network (2015). Genomic Classification of Cutaneous Melanoma. Cell *161*, 1681–1696.

Cardinali, M., Uchino, R., and Chung, S.I. (1990). Interaction of fibrinogen with murine melanoma cells: covalent association with cell membranes and protection against recognition by lymphokine-activated killer cells. Cancer Res. *50*, 8010–8016.

Cavalieri, D., Dolara, P., Mini, E., Luceri, C., Castagnini, C., Toti, S., Maciag, K., De Filippo, C., Nobili, S., Morganti, M., et al. (2007). Analysis of Gene Expression Profiles Reveals Novel Correlations With the Clinical Course of Colorectal Cancer. Oncol. Res. Featur. Preclin. Clin. Cancer Ther. *16*, 535–548.

Chamcheu, J., Roy, T., Uddin, M., Banang-Mbeumi, S., Chamcheu, R.-C., Walker, A., Liu, Y.-Y., and Huang, S. (2019). Role and Therapeutic Targeting of the PI3K/Akt/mTOR Signaling Pathway in Skin Cancer: A Review of Current Status and Future Trends on Natural and Synthetic Agents Therapy. Cells *8*, 803.

Chelberg, M.K., Tsilibary, E.C., Hauser, A.R., and McCarthy, J.B. (1989). Type IV collagen-mediated melanoma cell adhesion and migration: involvement of multiple, distinct domains of the collagen molecule. Cancer Res. *49*, 4796–4802.

Ciereszko, A., Dietrich, M.A., Słowińska, M., Nynca, J., Ciborowski, M., Kisluk, J., Michalska-Falkowska, A., Reszec, J., Sierko, E., and Nikliński, J. (2019). Identification of protein changes in the blood plasma of lung cancer patients subjected to chemotherapy using a 2D-DIGE approach. PLoS One *14*, e0223840.

Cirenajwis, H., Ekedahl, H., Lauss, M., Harbst, K., Carneiro, A., Enoksson, J., Rosengren, F., Werner-Hartman, L., Törngren, T., Kvist, A., et al. (2015). Molecular stratification of metastatic melanoma using gene expression profiling: Prediction of survival outcome and benefit from molecular targeted therapy. Oncotarget *6*, 12297–12309.

D'Souza, Z., Taher, F.S., and Lupashin, V. V (2020). Golgi inCOGnito: From vesicle tethering to human disease. Biochim. Biophys. Acta. Gen. Subj. *1864*, 129694.

Dantonio, P.M., Klein, M.O., Freire, M.R.V.B., Araujo, C.N., Chiacetti, A.C., and Correa, R.G. (2018). Exploring major signaling cascades in melanomagenesis: a rationale route for targetted skin cancer therapy. Biosci. Rep. *38*.

Delaunay, S., and Frye, M. (2019). RNA modifications regulating cell fate in cancer. Nat. Cell Biol. *21*, 552–559.

Desiere, F. (2006). The PeptideAtlas project. Nucleic Acids Res. *34*, D655–D658.

Drozak, J., Piecuch, M., Poleszak, O., Kozlowski, P., Chrobok, L., Baelde, H.J., and de Heer, E. (2015). UPF0586 Protein C9orf41 Homolog Is Anserine-producing Methyltransferase. J. Biol. Chem. *290*, 17190–17205.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Erdmann, F., Lortet-Tieulent, J., Schüz, J., Zeeb, H., Greinert, R., Breitbart, E.W., and Bray, F. (2013). International trends in the incidence of malignant melanoma 1953-2008--are recent generations at higher or lower risk? Int. J. Cancer *132*, 385–400.

Fang, R., Zhang, B., Lu, X., Jin, X., and Liu, T. (2020). FASTKD2 promotes cancer cell progression through upregulating Myc expression in pancreatic ductal adenocarcinoma. J. Cell. Biochem. *121*, 2458–2466.

Fehér, L.Z., Pocsay, G., Krenács, L., Zvara, Á., Bagdi, E., Pocsay, R., Lukács, G., Győry, F., Gazdag, A., Tarkó, E., et al. (2012). Amplification of Thymosin Beta 10 and AKAP13 Genes in Metastatic and Aggressive Papillary Thyroid Carcinomas. Pathol. Oncol. Res. *18*, 449–458.

Figenschau, S.L., Knutsen, E., Urbarova, I., Fenton, C., Elston, B., Perander, M., Mortensen, E.S., and Fenton, K.A. (2018). ICAM1 expression is induced by proinflammatory cytokines and associated with TLS formation in aggressive breast cancer subtypes. Sci. Rep. *8*, 11720.

Fishelson, Z., and Kirschfink, M. (2019). Complement C5b-9 and Cancer: Mechanisms of Cell Damage, Cancer Counteractions, and Approaches for Intervention. Front. Immunol. *10*.

Freedberg, D.E., Rigas, S.H., Russak, J., Gai, W., Kaplow, M., Osman, I., Turner, F., Randerson-Moor, J.A., Houghton, A., Busam, K., et al. (2008). Frequent p16-independent inactivation of p14ARF in human melanoma. J. Natl. Cancer Inst. *100*, 784–795.

Furfaro, A.L., Ottonello, S., Loi, G., Cossu, I., Piras, S., Spagnolo, F., Queirolo, P., Marinari, U.M., Moretta, L., Pronzato, M.A., et al. (2020). HO-1 downregulation favors BRAFV600 melanoma cell death induced by Vemurafenib/PLX4032 and increases NK recognition. Int. J. Cancer *146*, 1950–1962.

Gad, A.A., and Balenga, N. (2020). The Emerging Role of Adhesion GPCRs in Cancer. ACS Pharmacol. Transl. Sci. *3*, 29–42.

Gao, X., Leone, G.W., and Wang, H. (2020). Cyclin D-CDK4/6 functions in cancer. Adv. Cancer Res. *148*, 147–169.

García-Gutiérrez, L., Bretones, G., Molina, E., Arechaga, I., Symonds, C., Acosta, J.C., Blanco, R., Fernández, A., Alonso, L., Sicinski, P., et al. (2019). Myc stimulates cell cycle progression through the activation of Cdk1 and phosphorylation of p27. Sci. Rep. *9*, 18693.

Garinet, S., Pignot, G., Vacher, S., Le Goux, C., Schnitzler, A., Chemlali, W., Sirab, N., Barry Delongchamps, N., Zerbib, M., Sibony, M., et al. (2019). High Prevalence of a Hotspot of Noncoding Somatic Mutations in Intron 6 of GPR126 in Bladder Cancer. Mol. Cancer Res. *17*, 469–475.

Gasparics, Á., Kökény, G., Fintha, A., Bencs, R., Mózes, M.M., Ágoston, E.I., Buday, A., Ivics, Z., Hamar, P., Győrffy, B., et al. (2018). Alterations in SCAI Expression during Cell Plasticity, Fibrosis and Cancer. Pathol. Oncol. Res. *24*, 641–651.

Grzywa, T.M., Paskal, W., and Włodarski, P.K. (2017). Intratumor and Intertumor Heterogeneity in Melanoma. Transl. Oncol. *10*, 956–975.

Guida, M., Ravaioli, A., Sileni, V.C., Romanini, A., Labianca, R., Freschi, A., Brugnara, S., Casamassima, A., Lorusso, V., Nanni, O., et al. (2003). Fibrinogen: a novel predictor of responsiveness in metastatic melanoma patients treated with bio-chemotherapy: IMI (italian melanoma inter-group) trial. J. Transl. Med. *1*, 13.

Gunji, Y., and Gorelik, E. (1988). Role of fibrin coagulation in protection of murine tumor cells from destruction by cytotoxic cells. Cancer Res. *48*, 5216–5221.

Guo, L., Qi, J., Wang, H., Jiang, X., and Liu, Y. (2020). Getting under the skin: The role of CDK4/6 in melanomas. Eur. J. Med. Chem. *204*, 112531.

Hall, A.E., Lu, W.-T., Godfrey, J.D., Antonov, A. V, Paicu, C., Moxon, S., Dalmay, T., Wilczynska, A., Muller, P.A.J., and Bushell, M. (2016). The cytoskeleton adaptor protein ankyrin-1 is upregulated by p53 following DNA damage and alters cell migration. Cell Death Dis. *7*, e2184–e2184.

Hamid, O., Robert, C., Daud, A., Hodi, F.S., Hwu, W.J., Kefford, R., Wolchok, J.D., Hersey, P., Joseph, R., Weber, J.S., et al. (2019). Five-year survival outcomes for patients with advanced melanoma treated with pembrolizumab in KEYNOTE-001. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. *30*, 582–588.

van Hengel, J., and van Roy, F. (2007). Diverse functions of p120ctn in tumors. Biochim. Biophys. Acta *1773*, 78–88.

Hjortsø, M.D., and Andersen, M.H. (2014). The expression, function and targeting of haem oxygenase-1 in cancer. Curr. Cancer Drug Targets *14*, 337–347.

Hocker, T., and Tsao, H. (2007). Ultraviolet radiation and melanoma: a systematic review and analysis of reported sequence variants. Hum. Mutat. *28*, 578–588.

Huang, L., Chen, J., Zhao, Y., Gu, L., Shao, X., Li, J., Xu, Y., Liu, Z., and Xu, Q. (2019). Key candidate genes of STAT1 and CXCL10 in melanoma identified by integrated bioinformatical analysis. IUBMB Life *71*, 1634–1644.

Hugo, W., Zaretsky, J.M., Sun, L., Song, C., Moreno, B.H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., et al. (2016). Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell *165*, 35–44.

Hyeon, D.Y., Kim, J.H., Ahn, T.J., Cho, Y., Hwang, D., and Kim, S. (2019). Evolution of the multi-tRNA synthetase complex and its role in cancer. J. Biol. Chem. *294*, 5340–5351.

Jannin, A., Penel, N., Ladsous, M., Vantyghem, M.C., and Do Cao, C. (2019). Tyrosine kinase inhibitors and immune checkpoint inhibitors-induced thyroid disorders. Crit. Rev. Oncol. Hematol. *141*, 23–35.

Jia, Q., Wang, J., He, N., He, J., and Zhu, B. (2019). Titin mutation associated with responsiveness to checkpoint blockades in solid tumors. JCI Insight *4*.

Kamenisch, Y., Baban, T.S.A., Schuller, W., von Thaler, A.-K., Sinnberg, T., Metzler, G., Bauer, J., Schittek, B., Garbe, C., Rocken, M., et al. (2016). UVA-Irradiation Induces Melanoma Invasion via the Enhanced Warburg Effect. J. Invest. Dermatol. *136*, 1866–1875.

Kang, J., Sergio, C.M., Sutherland, R.L., and Musgrove, E.A. (2014). Targeting cyclin-dependent kinase 1 (CDK1) but not CDK4/6 or CDK2 is selectively lethal to MYC-dependent human breast cancer cells. BMC Cancer *14*, 32.

Khodarev, N.N., Roach, P., Pitroda, S.P., Golden, D.W., Bhayani, M., Shao, M.Y., Darga, T.E., Beveridge, M.G., Sood, R.F., Sutton, H.G., et al. (2009). STAT1 Pathway Mediates Amplification of Metastatic Potential and Resistance to Therapy. PLoS One *4*, e5821.

Koh, S.S., Wei, J.-P.J., Li, X., Huang, R.R., Doan, N.B., Scolyer, R.A., Cochran, A.J., and Binder, S.W. (2012). Differential gene expression profiling of primary cutaneous melanoma and sentinel lymph node metastases. Mod. Pathol. *25*, 828–837.

Kourtidis, A., Yanagisawa, M., Huveldt, D., Copland, J.A., and Anastasiadis, P.Z. (2015). Pro-Tumorigenic Phosphorylation of p120 Catenin in Renal and Breast Cancer. PLoS One *10*, e0129964.

Kuras, M., Betancourt, L.H., Rezeli, M., Rodriguez, J., Szasz, M., Zhou, Q., Miliotis, T., Andersson, R., and Marko-Varga, G. (2018). Assessing Automated Sample Preparation Technologies for High-Throughput Proteomics of Frozen Well Characterized Tissues from Swedish Biobanks. J. Proteome Res. acs.jproteome.8b00792.

Kwong, L.N., and Davies, M.A. (2013). Navigating the therapeutic complexity of PI3K pathway inhibition in melanoma. Clin. Cancer Res. *19*, 5310–5319.

Lambrecht, B.N., Vanderkerken, M., and Hammad, H. (2018). The emerging role of ADAM metalloproteinases in immunity. Nat. Rev. Immunol. *18*, 745–758.

Lee, S.B., Schramme, A., Doberstein, K., Dummer, R., Abdel-Bakky, M.S., Keller, S., Altevogt, P., Oh, S.T., Reichrath, J., Oxmann, D., et al. (2010). ADAM10 is upregulated in melanoma metastasis compared with primary melanoma. J. Invest. Dermatol. *130*, 763–773.

Lek, M., Karczewski, K.J., Minikel, E. V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Leonardi, G.C., Falzone, L., Salemi, R., Zanghì, A., Spandidos, D.A., Mccubrey, J.A., Candido, S., and Libra, M. (2018). Cutaneous melanoma: From pathogenesis to therapy (Review). Int. J. Oncol. *52*, 1071–1080.

Li, J., Duncan, D.T., and Zhang, B. (2010). CanProVar: a human cancer proteome variation database. Hum. Mutat. *31*, 219–228.

Li, J., Liu, L., Liu, X., Xu, P., Hu, Q., and Yu, Y. (2019). The Role of Upregulated DDX11 as A Potential Prognostic and Diagnostic Biomarker in Lung Adenocarcinoma. J. Cancer *10*, 4208–4216.

Li, Q., Yang, J., Yu, Q., Wu, H., Liu, B., Xiong, H., Hu, G., Zhao, J., Yuan, X., and Liao, Z. (2013). Associations between Single-Nucleotide Polymorphisms in the PI3K–PTEN–AKT–mTOR Pathway and Increased Risk of Brain Metastasis in Patients with Non–Small Cell Lung Cancer. Clin. Cancer Res. *19*, 6252–6260.

Li, W., Liu, J., Zhang, B., Bie, Q., Qian, H., and Xu, W. (2020). Transcriptome Analysis Reveals Key Genes and Pathways Associated with Metastasis in Breast Cancer. Onco. Targets. Ther. *Volume 13*, 323–335.

Lin, J., Wang, J., Greisinger, A.J., Grossman, H.B., Forman, M.R., Dinney, C.P., Hawk, E.T., and Wu, X. (2010). Energy Balance, the PI3K-AKT-mTOR Pathway Genes, and the Risk of Bladder Cancer. Cancer Prev. Res. *3*, 505–517.

Linding, R., Jensen, L.J., Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M.B., and Pawson, T. (2008). NetworKIN: a resource for exploring cellular phosphorylation networks. Nucleic Acids Res. *36*, D695-9.

Liu, S. (2011). Editorial - a potential new target gene of the master-regulator microphthalmia-associated transcription factor in melanoma. Ochsner J. *11*, 210–211.

Liu, W., Payne, S.H., Ma, S., and Fenyö, D. (2019). Extracting Pathway-level Signatures from Proteogenomic Data in Breast Cancer Using Independent Component Analysis. Mol. Cell. Proteomics *18*, S169–S182.

Lopez-Bergami, P., Fitchman, B., and Ronai, Z. (2008). Understanding Signaling Cascades in Melanoma. Photochem. Photobiol. *84*, 289–306.

Lord, C.C., Thomas, G., and Brown, J.M. (2013). Mammalian alpha beta hydrolase domain (ABHD) proteins: Lipid metabolizing enzymes at the interface of cell signaling and energy metabolism. Biochim. Biophys. Acta - Mol. Cell Biol. Lipids *1831*, 792–802.

Luke, J.J., Flaherty, K.T., Ribas, A., and Long, G. V (2017). Targeted agents and immunotherapies: optimizing outcomes in melanoma. Nat. Rev. Clin. Oncol. *14*, 463–482.

Mahtab, M., Boavida, A., Santos, D., and Pisani, F.M. (2021). The Genome Stability Maintenance DNA Helicase DDX11 and Its Role

in Cancer. Genes (Basel). *12*.

Marchese, F.P., Grossi, E., Marín-Béjar, O., Bharti, S.K., Raimondi, I., González, J., Martínez-Herrera, D.J., Athie, A., Amadoz, A., Brosh, R.M., et al. (2016). A Long Noncoding RNA Regulates Sister Chromatid Cohesion. Mol. Cell *63*, 397–407.

Miller, M.L., Jensen, L.J., Diella, F., Jørgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S.A., Bordeaux, J., Sicheritz-Ponten, T., et al. (2008). Linear motif atlas for phosphorylation-dependent signaling. Sci. Signal. *1*, ra2.

Molee, P., Adisakwattana, P., Reamtong, O., Petmitr, S., Sricharunrat, T., Suwandittakul, N., and Chaisri, U. (2015). Up-regulation of AKAP13 and MAGT1 on cytoplasmic membrane in progressive hepatocellular carcinoma: a novel target for prognosis. Int. J. Clin. Exp. Pathol. *8*, 9796–9811.

Murillo, J.R., Kuras, M., Rezeli, M., Miliotis, T., Betancourt, L., and Marko-Varga, G. (2018). Correction: Automated phosphopeptide enrichment from minute quantities of frozen malignant melanoma tissue. PLoS One *13*, e0210234.

Musa, G., Cazorla-Vázquez, S., van Amerongen, M.J., Stemmler, M.P., Eckstein, M., Hartmann, A., Braun, T., Brabletz, T., and Engel, F.B. (2019). Gpr126 (Adgrg6) is expressed in cell types known to be exposed to mechanical stimuli. Ann. N. Y. Acad. Sci. *1456*, 96–108.

Nagel, T., Klaus, F., Ibanez, I.G., Wege, H., Lohse, A., and Meyer, B. (2018). Fast and facile analysis of glycosylation and phosphorylation of fibrinogen from human plasma-correlation with liver cancer and liver cirrhosis. Anal. Bioanal. Chem. *410*, 7965–7977.

Nieminuszczy, J., Broderick, R., Bellani, M.A., Smethurst, E., Schwab, R.A., Cherdyntseva, V., Evmorfopoulou, T., Lin, Y.-L., Minczuk, M., Pasero, P., et al. (2019). EXD2 Protects Stressed Replication Forks and Is Required for Cell Viability in the Absence of BRCA1/2. Mol. Cell *75*, 605-619.e6.

Ogata, Y., Heppelmann, C.J., Charlesworth, M.C., Madden, B.J., Miller, M.N., Kalli, K.R., Cliby, W.A., Bergen, H.R., Saggese, D.A., and Muddiman, D.C. (2006). Elevated Levels of Phosphorylated Fibrinogen-α-Isoforms and Differential Expression of Other Post-Translationally Modified Proteins in the Plasma of Ovarian Cancer Patients. J. Proteome Res.

Orme, J.J., Jazieh, K.A., Xie, T., Harrington, S., Liu, X., Ball, M., Madden, B., Charlesworth, M.C., Azam, T.U., Lucien, F., et al. (2020). ADAM10 and ADAM17 cleave PD-L1 to mediate PD-(L)1 inhibitor resistance. Oncoimmunology *9*, 1744980.

Palumbo, J.S., Kombrinck, K.W., Drew, A.F., Grimes, T.S., Kiser, J.H., Degen, J.L., and Bugge, T.H. (2000). Fibrinogen is an important determinant of the metastatic potential of circulating tumor cells. Blood *96*, 3302–3309.

Paluncic, J., Kovacevic, Z., Jansson, P.J., Kalinowski, D., Merlot, A.M., Huang, M.L.-H., Lok, H.C., Sahni, S., Lane, D.J.R., and Richardson, D.R. (2016). Roads to melanoma: Key pathways and emerging players in melanoma progression and oncogenic signaling. Biochim. Biophys. Acta - Mol. Cell Res. *1863*, 770–784.

Pieters, T., van Roy, F., and van Hengel, J. (2012). Functions of p120ctn isoforms in cell-cell adhesion and intracellular signaling. Front. Biosci. (Landmark Ed. *17*, 1669–1694.

Qi, L., Sun, K., Zhuang, Y., Yang, J., and Chen, J. Study on the association between PI3K/AKT/mTOR signaling pathway gene polymorphism and susceptibility to gastric cancer. J. BUON. *22*, 1488–1493.

Ren, S., Liu, S., Howell, P.M., Zhang, G., Pannell, L., Samant, R., Shevde-Samant, L., Tucker, J.A., Fodstad, O., and Riker, A.I. (2010). Functional characterization of the progestagen-associated endometrial protein gene in human melanoma. J. Cell. Mol. Med. *14*, 1432–1442.

Ren, S., Howell, P.M., Han, Y., Wang, J., Liu, M., Wang, Y., Quan, G., Du, W., Fang, L., and Riker, A.I. (2011). Overexpression of the progestagen-associated endometrial protein gene is associated with microphthalmia-associated transcription factor in human melanoma. Ochsner J. *11*, 212–219.

Ren, S., Chai, L., Wang, C., Li, C., Ren, Q., Yang, L., Wang, F., Qiao, Z., Li, W., He, M., et al. (2015). Human malignant melanoma-derived progestagen-associated endometrial protein immunosuppresses T lymphocytes in vitro. PLoS One *10*, e0119038.

Robert, C., Grob, J.J., Stroyakovskiy, D., Karaszewska, B., Hauschild, A., Levchenko, E., Chiarion Sileni, V., Schachter, J., Garbe, C., Bondarenko, I., et al. (2019). Five-Year Outcomes with Dabrafenib plus Trametinib in Metastatic Melanoma. N. Engl. J. Med. *381*, 626–636.

Ryu, J.K., Petersen, M.A., Murray, S.G., Baeten, K.M., Meyer-Franke, A., Chan, J.P., Vagena, E., Bedard, C., Machado, M.R., Rios Coronado, P.E., et al. (2015). Blood coagulation protein fibrinogen promotes autoimmunity and demyelination via chemokine release and antigen presentation. Nat. Commun. *6*, 8164.

Shao, H., Li, S., Watkins, S.C., and Wells, A. (2014). α-Actinin-4 Is Required for Amoeboid-type Invasiveness of Melanoma Cells. J. Biol. Chem. *289*, 32717–32728.

Sheppard, H.M., Feisst, V., Chen, J., Print, C., and Dunbar, P.R. (2016). AHNAK is downregulated in melanoma, predicts poor outcome, and may be required for the expression of functional cadherin-1. Melanoma Res. *26*, 108–116.

Stracker, T.H. (2018). EXD2: A new regulator of mitochondrial translation and potential target for cancer therapy. Mol. Cell. Oncol. *5*, e1445943.

Sullivan, R.J., Hamid, O., Gonzalez, R., Infante, J.R., Patel, M.R., Hodi, F.S., Lewis, K.D., Tawbi, H.A., Hernandez, G., Wongchenko, M.J., et al. (2019). Atezolizumab plus cobimetinib and vemurafenib in BRAF-mutated melanoma patients. Nat. Med. *25*, 929–935.

Timar, J., Lapis, K., Fulop, T., Varga, Z.S., Tixier, J.M., Robert, L., and Hornebeck, W. (1991). Interaction between elastin and tumor cell lines with different metastatic potential; in vitro and in vivo studies. J. Cancer Res. Clin. Oncol. *117*, 232–238.

Trilla-Fuertes, L., Gámez-Pozo, A., Prado-Vázquez, G., Zapater-Moros, A., Díaz-Almirón, M., Fortes, C., Ferrer-Gómez, M., López-Vacas, R., Parra Blanco, V., Márquez-Rodas, I., et al. (2019). Melanoma proteomics suggests functional differences related to mutational status. Sci. Rep. *9*, 7217.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods *13*, 731–740.

Vanhollebeke, B., and Pays, E. (2006). The function of apolipoproteins L. Cell. Mol. Life Sci. *63*, 1937–1944.

Vivancos, A., Caratú, G., Matito, J., Muñoz, E., Ferrer, B., Hernández-Losa, J., Bodet, D., Pérez-Alea, M., Cortés, J., Garcia-Patos, V., et al. (2016). Genetic evolution of nevus of Ota reveals clonal heterogeneity acquiring BAP1 and TP53 mutations. Pigment Cell Melanoma Res. *29*, 247–253.

Wang, L., Zhang, Z., Li, Y., Wan, Y., and Xing, B. (2021). Integrated bioinformatic analysis of RNA binding proteins in hepatocellular carcinoma. Aging (Albany. NY). *13*, 2480–2505.

Wei, W., Sun, Z., da Silveira, W.A., Yu, Z., Lawson, A., Hardiman, G., Kelemen, L.E., and Chung, D. (2019). Semi-supervised identification of cancer subgroups using survival outcomes and overlapping grouping information. Stat. Methods Med. Res. *28*, 2137–2149.

Xu, B., Zhang, X., Gao, Y., Song, J., and Shi, B. (2021). Microglial Annexin A3 promoted the development of melanoma via activation of hypoxia-inducible factor-1α/vascular endothelial growth factor signaling pathway. J. Clin. Lab. Anal. *35*.

Xu, X., Huang, L., Chan, C.H., Yu, T., Miao, R., and Liu, C. (2016). Assessing the clinical utility of genomic expression data across human cancers. Oncotarget *7*, 45926–45936.

Yanagisawa, M., Huveldt, D., Kreinest, P., Lohse, C.M., Cheville, J.C., Parker, A.S., Copland, J.A., and Anastasiadis, P.Z. (2008). A p120 catenin isoform switch affects Rho activity, induces tumor cell invasion, and predicts metastatic disease. J. Biol. Chem. *283*, 18344–18354.

Yang, L., Froio, R.M., Sciuto, T.E., Dvorak, A.M., Alon, R., and Luscinskas, F.W. (2005). ICAM-1 regulates neutrophil adhesion and transcellular migration of TNF-α-activated vascular endothelium under flow. Blood *106*, 584–592.

Yuan, T.L., and Cantley, L.C. (2008). PI3K pathway alterations in cancer: variations on a theme. Oncogene *27*, 5497–5510.

Yuan, Y., Van Allen, E.M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., Byers, L.A., Xu, Y., Hess, K.R., Diao, L., et al. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat. Biotechnol. *32*, 644–652.

Yuan, Y., Ju, Y.S., Kim, Y., Li, J., Wang, Y., Yoon, C.J., Yang, Y., Martincorena, I., Creighton, C.J., Weinstein, J.N., et al. (2020). Comprehensive molecular characterization of mitochondrial genomes in human cancers. Nat. Genet. *52*, 342–352.

Zhou, B., Ritt, D.A., Morrison, D.K., Der, C.J., and Cox, A.D. (2016). Protein Kinase CK2α Maintains Extracellular Signal-regulated Kinase (ERK) Activity in a CK2α Kinase-independent Manner to Promote Resistance to Inhibitors of RAF and MEK but Not ERK in BRAF Mutant Melanoma. J. Biol. Chem. *291*, 17804–17815.

Zhu, T., Gao, Y.-F., Chen, Y.-X., Wang, Z.-B., Yin, J.-Y., Mao, X.-Y., Li, X., Zhang, W., Zhou, H.-H., and Liu, Z.-Q. (2017). Genome-scale analysis identifies GJB2 and ERO1LB as prognosis markers in patients with pancreatic cancer. Oncotarget *8*, 21281–21289.

Zhu, Y., Orre, L.M., Johansson, H.J., Huss, M., Boekel, J., Vesterlund, M., Fernandez-Woodbridge, A., Branca, R.M.M., and Lehtiö, J. (2018). Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat. Commun. *9*, 903.

(2003). The International HapMap Project. Nature *426*, 789–796.

(2015). A global reference for human genetic variation. Nature *526*, 68–74.

Betancourt LH, Sanchez A, Pla I, et al. Quantitative Assessment of Urea In-Solution Lys C/Trypsin Digestions Reveals Superior Performance at Room Temperature over Traditional Proteolysis at 37 °C. J Proteome Res. 2018;17(7):2556–2561.

Jönsson J, Busch C, Knappskog S, Geisler J, Miletic H, Ringnér M, Lillehaug JR, Borg A, Eystein Lønning P. Clin Cancer Res. 2010;16(13) 3356-67.

Liu, Wenke, et al. "Extracting pathway-level signatures from proteogenomic data in breast cancer using independent component analysis." *Molecular & Cellular Proteomics* 18.8 suppl 1 (2019): S169-S182.