Detecting Protein Variants within Mass Spectrometry Datasets

by

Javier Antonio Alfaro

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy

Department of Medical Biophysics

University of Toronto

© Copyright by Javier Antonio Alfaro 2018
Detecting Protein Variants within Mass Spectrometry Datasets

Javier Antonio Alfaro

Doctor of Philosophy

Department of Medical Biophysics
University of Toronto

2018

Abstract

An important aim of cancer proteomics is to understand how the cancer proteome, with all its variation in both protein sequence and protein abundance, emerges from aberrations characterized within cancer genomic and transcriptomic datasets. A challenge in reaching this goal is that standard mass-spectrometry-based onco-proteomic strategies typically ignore variations in protein sequences, instead characterizing cancer proteomes in terms of changing abundances of standard reference proteins. In this thesis, I assess strategies for incorporating information about protein sequence variation within standard proteomics analysis strategies. I develop a pipeline for the characterization of variations within protein sequences and assess the best methodology to do so.
Acknowledgments

This PhD could not have been completed without the support of so many along the way. First and foremost I would like to thank Dr. Paul Boutros and Dr. Thomas Kislinger for their support and patience over the years. Through the training I received here I developed not just as a scientist but also the skills, organization and perfectionism needed to propel projects forward to completion. I would also like to thank my committee members Dr. Laurie Ailles and Dr. Geoffrey Liu as well as the external examiner Dr. Oliver Schilling and the examining committee for their helpful discussion along the way.

This thesis would not have been possible without the support of several collaborators who contributed significantly to the project. First and foremost I would like to acknowledge the efforts of Alex Ignatchenko and Vladimir Ignatchenko, who worked alongside me towards the development and completion of our proteogenomic pipeline. I would also like to acknowledge the many engaging discussions I had with Ankit Sinha and Lusia Sepiashvili, who helped contextualize the work presented here within the broader field of mass-spectrometry.

To the many friends and loved ones who have supported me throughout the thesis: Lusia Sepiashvili, Deborah Ng, Connie Li, Natalie Fox, Cindy Yang, Andrea Vargas and Rene Quevedo. You have been pillars of support throughout my time here at the University of Toronto. To my dearest Kiersten Kennedy: your love and energy has guided us both through graduate school. I am eternally grateful for your company. Generally I would like to thank all of the members of the Kislinger and Boutros labs who are too many to list here, but without whom I would not have developed intellectually as I have.
Table of Contents

Contents

Acknowledgments .......................................................................................................................... iii
Table of Contents ............................................................................................................................ iv
List of Tables .................................................................................................................................... vii
List of Figures ............................................................................................................................... viii
List of Abbreviations .................................................................................................................... x
Chapter 1 Introduction .................................................................................................................. 1
1 Cancer molecular profiling: integrating proteomics with genomics ..................................... 1
   1.1 Proteomics: applications in cancer research ..................................................................... 2
      1.1.1 Proteoforms .................................................................................................................. 4
      1.1.2 Proteotypes and molecular phenotypes ..................................................................... 4
      1.1.3 Relevance to cancer: PTEN proteoform complexity ................................................. 7
1.2 Studying the proteome using mass-spectrometry .............................................................. 7
   1.2.1 Mass-spectrometers in proteomics: a brief primer ....................................................... 8
   1.2.2 Characterizing proteomes using mass-spectrometry ................................................. 8
   1.2.3 Bottom up proteomics ................................................................................................. 10
   1.2.4 Computational strategies for inferring protein sequences from tandem mass
       spectrometry datasets ............................................................................................................. 17
1.3 Summary ................................................................................................................................. 27

Chapter 2 Onco-proteogenomics: cancer proteomics joins forces with genomics .................. 29
2 Onco-proteogenomics: cancer proteomics joins forces with genomics ............................. 30
   2.1 Introduction ......................................................................................................................... 30
      2.1.1 Proteogenomics technology primer ............................................................................ 33
## 2.1 Genome sequencing data

- 2.1.2 Genome sequencing data ................................................................. 33
- 2.1.3 Transcriptomic sequencing data ......................................................... 33
- 2.1.4 Proteogenomic database generation and searching ............................ 34

## 2.2 Proteomic detection of tumor-specific peptides

- 2.2.1 Onco-proteogenomics in the literature ........................................... 34

## 2.3 Challenges in onco-proteogenomics

- 2.3.1 Toward integrative cancer phenotyping ........................................... 38
- 2.3.2 Gauging the success of onco-proteogenomic techniques ..................... 38

## 2.4 Potential impact of onco-proteogenomics in cancer research

- 2.4.1 Reducing false negative exome calls .............................................. 39
- 2.4.2 Improving understanding of cellular networks and dysregulation in cancer ...... 41
- 2.4.3 Top-down proteomics, proteogenomics and the proteoform .................. 42
- 2.4.4 Targeted MS and translation to clinical research ................................ 43

## 2.5 Concluding remarks ............................................................................ 43

## Chapter 3 Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines

- 3.1 Methods ................................................................................................. 47
  - 3.1.1 Variant peptide database construction ........................................... 48
  - 3.1.2 Variant peptide detection ................................................................. 51
- 3.2 Results ................................................................................................... 56
  - 3.2.1 Characterizing reference and variant protein sequence databases ......... 56
  - 3.2.2 Scope of variant peptides identified ................................................. 62
  - 3.2.3 The relevance of proteogenomic peptides ....................................... 63
- 3.3 Discussion ............................................................................................. 73
- 3.4 Conclusions .......................................................................................... 75
Chapter 4 Discussion: Future prospects for the detection of protein variants .........................76

4 Discussion ........................................................................................................................................77

4.1 Peptide sequence coverage as a metric to further constrain proteogenomic search results. ........................................................................................................77

4.2 Hybrid approaches for variant peptide identification ...............................................................78

4.2.1 Proteogenomic semi-supervised search strategies .................................................................78

4.2.2 Other hybrid strategies ........................................................................................................79

4.2.3 Top-down and middle-down proteomics as alternative strategies for improving proteoform characterization ........................................................................81

4.2.4 Alternative strategies for global proteoform characterization .............................................82

4.2.5 Concluding remarks ............................................................................................................82

References ..........................................................................................................................................83
List of Tables

Table 1.1: Popular supervised proteomics search engines......................................................... 18

Table 2.1: A survey of onco-proteogenomic implementations in the literature ......................... 35
List of Figures

Figure 1.1: Proteoform diversity as a final frontier in molecular profiling. ........................................ 6

Figure 1.2: Primer to mass-spectrometry-based proteomics................................................................. 12

Figure 1.3: Strategies used in bottom up proteomics............................................................................ 14

Figure 1.4: There are three main strategies for identifying peptides from datasets generated by bottom-up proteomics experiments......................................................................................... 22

Figure 2.1: Detecting tumor-specific genetic aberrations in cancer proteomics. ....................... 32

Figure 2.2: The potential impact of onco-proteogenomics in cancer research............................. 40

Figure 3.1: Generation of proteogenomic databases............................................................................. 49

Figure 3.2: Proteogenomic search and filtering strategy ................................................................. 53

Figure 3.3: Comparison of reference proteomes................................................................................ 58

Figure 3.4: The detectable tryptic space of reference proteomes and the overlap to variant human proteins................................................................................................................................. 59

Figure 3.5: Comparison to other studies............................................................................................ 60

Figure 3.6: Overall database sizes and redundancies using prostate cancer cell-line PC3 as an example........................................................................................................................................ 61

Figure 3.7: Detection of variant proteins within the 9 deep proteomes with focus on prostate cancer cell-line PC3. ....................................................................................................................................... 65

Figure 3.8: Biophysical properties of detected variant peptides......................................................... 66

Figure 3.9: Tier 2 peptides offer significant peptide detections while mitigating potential false positives........................................................................................................................................ 68
Figure 3.10: Genomic distribution of Tier 2 variant peptides across the 59 shallow and 9 deep proteome datasets. .......................... 69

Figure 3.11: On the expression and druggability of genes containing detected variant peptides. 69

Figure 3.12: Focus on variants detected for CTNNB1 across cell lines. .......................... 71

Figure 3.13: MS2 spectra for FUS-CREB3L2 fusions. .......................... 73

Figure 4.1: Hybrid approaches for variant protein detection. .......................... 80
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CID</td>
<td>Collision induced dissociation</td>
</tr>
<tr>
<td>COSMIC</td>
<td>Catalogue of Somatic Mutations in Cancer</td>
</tr>
<tr>
<td>CPTAC</td>
<td>Clinical Proteomics Tumour Analysis Consortium</td>
</tr>
<tr>
<td>DDA</td>
<td>Data dependent acquisition</td>
</tr>
<tr>
<td>DIA</td>
<td>Data independent acquisition</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>EMBL</td>
<td>European Molecular Biology Laboratory</td>
</tr>
<tr>
<td>ESI</td>
<td>Electrospray ionization</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>HCD</td>
<td>High energy collision induced dissociation</td>
</tr>
<tr>
<td>iBAQ</td>
<td>Intensity based absolute quantification</td>
</tr>
<tr>
<td>indel</td>
<td>Insertion or deletion</td>
</tr>
<tr>
<td>LC</td>
<td>Liquid Chromatography</td>
</tr>
<tr>
<td>m/z</td>
<td>Mass to charge ratio</td>
</tr>
<tr>
<td>MALDI</td>
<td>Matrix-Assisted Laser Desorption Ionization</td>
</tr>
<tr>
<td>MRM</td>
<td>multiple reaction monitoring</td>
</tr>
<tr>
<td>MS</td>
<td>Mass spectrometry</td>
</tr>
<tr>
<td>MS(^1)</td>
<td>Precursor ion spectrum</td>
</tr>
<tr>
<td>MS(^2)</td>
<td>Fragment ion spectrum</td>
</tr>
<tr>
<td>MSI</td>
<td>Microsatellite instability</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>PSM</td>
<td>peptide spectrum match</td>
</tr>
<tr>
<td>PTEN</td>
<td>Protein phosphatase and tensin homologue deleted on chromosome 10</td>
</tr>
<tr>
<td>PTM</td>
<td>Post-translational modification</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>RPKM</td>
<td>Reads per kilobase per million mapped reads</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SNV</td>
<td>Single Nucleotide Variant</td>
</tr>
<tr>
<td>SRM</td>
<td>single reaction monitoring</td>
</tr>
<tr>
<td>SWATH</td>
<td>Sequential window acquisition of all theoretical fragment-ion spectra</td>
</tr>
<tr>
<td>UniProtKB</td>
<td>UniProt Knowledge Base</td>
</tr>
</tbody>
</table>
Chapter 1
Introduction

1 Cancer molecular profiling: integrating proteomics with genomics

The phenotype of the cancer cell is characterized by uncontrolled cell growth, invasion and metastasis, which often arises from a multi-step evolutionary process [1, 2]. This evolutionary process involves various environmental and genetic factors that support cancer ‘hallmarks’ including sustained proliferative signaling, evading growth suppressors, resisting cell death, establishing replicative immortality, inducing angiogenesis, and fostering invasion and metastasis [3]. Underlying the development of these hallmarks are genomic and epigenetic instabilities, which nurture phenotypic diversity in the underlying cellular population and expedite hallmark acquisition [4, 5]. While tremendous resources have been devoted to cancer research, the complexity of this disease has hampered efforts to establish a mechanistic understanding of and develop treatments for the disease.

The global characterization of analytes as they progress from genome through transcriptome and to the proteome is progressing quickly and contributing significantly to our understanding of cancer biology. The pace of genomic and transcriptomic characterization has accelerated from a few hundreds of cancer samples characterized in nearly every major cancer type [6-27] to ambitious attempts to sequence thousands of samples in the upcoming years. Simultaneously, the Clinical Proteomics Tumour Analysis Consortium (CPTAC) has emerged to complete global proteomics studies, with matched genomics, in several tumor types [28-30]. New analytical strategies are needed to transform the data in these integrative high throughput datasets into key results and biological insights for cancer.

This thesis consists of four Chapters illustrating my contributions to the development of these methodologies. In this first Chapter, I introduce proteomics, discuss how proteomics contributes beyond genomics and transcriptomics to our understanding of cancer, and describe the technological and computational innovations that have enabled the high throughput
characterization of cancer proteomes. The second and third Chapters introduce the main problem and novel contributions of this thesis: how genomic, transcriptomic and community derived mutation data can be leveraged to enable the detection of protein variants within cancer samples [31, 32]. The second Chapter focuses on the emergence of onco-proteogenomics, an exciting field gleaning valuable clinical insights from global DNA, RNA and protein measurements. This Chapter, which was previously published as a perspective [31], shows the ambitions and aspirations of a community embracing other high throughput profiling strategies. The third Chapter [32], systematically studies how genomically derived data can be used to enable the detection of protein variants within cancer samples. The fourth Chapter discusses perspectives on the future of this field and some problems that will need to be addressed in the future.

1.1 Proteomics: applications in cancer research

The question of how genomic information is expressed to determine cellular phenotypes that support cancer growth is of central importance to cancer research. The proteome consists of roughly 50% of the dry mass of the cell, with 2-4 million proteins within every microliter of cellular content [33]. Because of their overall stability and the chemical diversity of their consituant amino-acids, proteins have taken on diverse roles in the cell as enzyme catalysts, signalling molecules and important structural components. Yet, despite their importance, the way proteomes emerge from the cancer genome and epigenome remains a significant and understudied aspect of cell biology. Understanding the relevance of global proteomics studies in cancer research begins by understanding how modern studies of whole genomes, exomes, and transcriptomes have transformed our understanding of cancer.

Progress in human and cancer genomics, epigenomics and transcriptomics fueled by next generation sequencing has accelerated our understanding of how genes are altered in cancer. The cancer genome atlas, the international cancer genome consortium and the pediatric cancer genome project have emerged to complete sequencing projects for almost every known cancer type and report their findings in a pan-cancer and site-specific fashion. Pan-cancer analyses have helped to confirm or establish many fundamental concepts in cancer biology [34]. The landscape of different types of genomic aberrations have been described across many cancers [35-38]. These analyses have demonstrated that cancers tend to be exclusively driven by either copy-number variation or mutation [39] and that different environmental factors can induce specific mutational profiles in
cancer [40]. High throughput sequencing strategies have been applied to the analysis of liquid biopsies, where circulating tumour DNA can be used to identify earlier recurrence in cancer [41] or to study acquired resistance to cancer therapy [42]. Transcriptomics, enabled by RNA-seq has also experienced wide spread adoption and has contributed biomarkers like MammaPrint [43] and OncoTypeDX [44], which are now used routinely in clinical practice. More recently, the non-coding portions of the transcriptome have become important in understanding cancer development [45-47]. Findings enabled by technological developments in genomics are also transforming cancer treatment. For example, the landscape of DNA mismatch-repair deficiencies have been described in a range of cancers [48, 49] and are often indicative of response to immune checkpoint blockade [50]. The acceptance of microsatellite instability (an indicator of DNA-mismatch repair) as a pan-cancer phenomenon has prompted the first FDA approval of a cancer drug (pembrolizumab) based on the genetics of the disease as opposed to the tumour location [51]. Tumour-location-specific genomics, epigenomics and transcriptomics have also been pivotal in stratifying tumours into subclasses showing different molecular traits that may lead to future personalized treatments [6-22, 26, 52-61].

Cancer is a disease of the genome, but the selection pressures that support the development of cancer hallmarks act on the expressed phenotypic traits of the cancer cell within different microenvironment contexts. A modern view of cancer biology relies on building a comprehensive molecular portrait of the cancer cell, which must include the proteome and the metabolome that represent significant endpoints of gene processing. Inference of proteome content from genomic content or transcript abundance is a contentious topic [62-64]. Overall transcript abundance is an important indicator of protein abundance [65] but spatial and temporal fluctuations in mRNA abundance and the local availability of resources for protein biosynthesis strongly influence the relationship between protein levels and their coding transcripts [66]. In fact, fluctuations in mRNA abundances are often buffered at level of protein abundance, meaning transcript abundances alone are not sufficient to predict protein abundances in some scenarios [67-75]. Thus explaining genotype-phenotype relationships as they pertain to protein sequences requires high-quality data quantifying global protein abundances.

Even a complete understanding of protein abundance in different cellular contexts is just a first step in understanding proteome complexity. The proteome is dynamic in the sense that proteins translated from each gene can exist in a population of differing states defined by their various
isoforms, conformations, modifications, interactions and cellular localizations. These so-called proteoforms [76] and proteotypes [77] are vital to understanding aberrant molecular profiles, and simply cannot be effectively predicted by genomics and transcriptomics at this time. The large datasets required to adequately build these predictive models do not yet exist, although community efforts to elucidate their predictive potential were in progress at the time of writing this thesis [78]. However, until the proteomics community finds ways to adequately and comprehensively describe this aspect of proteome diversity, the capacity to predict them from other molecular datasets will be limited. Current technical efforts in proteomics, including the work in this thesis [31, 32], are aimed at resolving aspects of this problem.

1.1.1 Proteoforms

Despite twenty years of technological innovation, the proteomics community has only just scratched the surface of the complexity of the human proteome. The first in-depth tissue based maps of the human proteome were released in 2014 [62, 79] and covered 84% of human protein coding genes (~20,300 genes total) at a median sequence coverage of just 28%. However, of those 20,300 genes in the human proteome [80], each can exist in a multiplicity of different states (proteoforms) defined by post-translational modification (PTM), alternative splicing, and germline variants that may differ between proteins originating from the expression of the same gene [76]. The proteoforms present in a cancer sample are complex (Figure 1.1a). The bulk tumor contains many differing protein variants, resulting from non-canonical splicing, mutations, fusions and post-translational modification. A complete characterization of the cancer proteome must encompass the proteoforms of each protein, which remains a challenge to attain because of the combinatorial explosion of potential combinations of different varieties of site-specific modifications and the aforementioned low sequence coverage of current proteomic methods.

1.1.2 Proteotypes and molecular phenotypes

The state of a proteome that is associated with a specific phenotype is commonly called the proteotype [77]. It is characterized not only by the protein conformational and proteoform diversity
The proteome in general and the cancer proteome specifically is often underestimated in terms of complexity. (a) Cancer is a disease of the genome, and genomic and epi-genetic aberrations undoubtedly will have consequences as they propagate through the transcriptome and the proteome. However, with each transition comes an increasing amount of potential complexity as the 23,393 protein coding genes in the genome manifest in 104,763 putative transcripts that yield a proteoform diversity, estimated in the millions. (b) These proteoforms interact with one another as they shuttle between various cellular components, creating a diversity of proteotypes that may be characterized by proteomics experiments. (Adapted from [77]) (c) Proteoform complexity, which includes genetic and post-translational variants of a protein, are increasingly relevant in cancer. The possible post-translational modifications of PTEN are depicted on the left hand side, illustrating the varying consequences that genomic aberrations disrupting them could have on PTEN function. Mutations commonly observed for PTEN in a pan-cancer study have been depicted on the right hand side. While truncating mutations dominate this tumour suppressor, missense and in frame insertions and deletions occur throughout the protein sequence in cancer samples. Both mutations and aberrant gene expression through the central dogma can cause shifts in PTEN proteoform populations that have implications in cancer development. (Adapted from [81])

present for each gene in the cell, but also by the way that proteins interact with other cellular components and their subcellular localizations [77] (Figure 1.1b). This concept is also important in cancer, as even single amino acid variants can alter the protein interactome [38] or even globally change the proteome [82]. Note that the proteotype is a protein centric view of cellular dynamics that arises from within the proteomics community. Similar concepts exist for each molecule type resulting from gene processing and for interactions between these levels of organization. For example, long non-coding RNAs are known to fold, interact with both protein and DNA, and be sequestered and shuttled to different regions of the cell [83-85]. Yet another example is the genome, which undeniably has a dynamic structure [86] and complex interactions with both protein and transcript. There is a need in the integrative-omic community for a more comprehensive word than proteotype and in this thesis we use the molecular-phenotype to fill this
niche. Precisely defined, the molecular-phenotype or cellular state is characterized by the abundances, modifications, interactions and subcellular localities of measurable molecular entities including the genome, transcriptome, proteome and metabolome.

1.1.3 Relevance to cancer: PTEN proteoform complexity

To illustrate the complexity of the proteoform and proteotypes that cannot be observed from genomic and transcriptomic datasets, consider one of the most frequently disrupted tumour suppressors in cancer: protein phosphatase and tensin homologue deleted on chromosome 10 (PTEN) [87] (Figure 1.1c). PTEN has a lipid phosphatase activity that antagonizes the phosphatidylinositol 3-kinase (PI3K)/AKT/mTOR pathway, which represses tumor cell growth and survival [88]. When localized to the nucleus, PTEN also promotes chromosomal stability and supports DNA repair [88, 89]. In cancer, loss of PTEN function is often a source of genomic instability [88]. Loss of function is thought to arise from somatic mutations, epigenetic and transcriptional silencing, and from abnormal post-translational modification [81]. In fact, a recent study of over 10,000 metastatic patients revealed a complex mutational spectrum for PTEN [90]. Without specifically studying the proteome, the effect of genomic and epigenetic aberrations on PTEN proteoform distribution, abundance, interactions and subcellular localization cannot be precisely understood. The presence of a mutation identified by genomics does not guarantee the expression of the mutated protein product and the post-translational modification repertoire of PTEN is large and affects activity, cellular localization, stability and protein interactions [81]. In cancer, the complexity of the interaction network for PTEN, alongside the abundance of post-translational modifications that would be difficult to predict without proteomics, indicates that mutations in this protein can have a variety of consequences that are difficult to disentangle without a comprehensive characterization of proteoform complexity.

1.2 Studying the proteome using mass-spectrometry

The way whole proteomes are identified within cancer samples is very different from the way high throughput nucleotide-sequencing technologies are used to infer genomes and transcriptomes. The typical proteomics workflow involves digesting a protein sample using a protease and characterizing the resulting peptides by mass-spectrometry (MS). This technology (mass-spectrometry) ionizes chemical species and measures ions in terms of their mass-to-charge ratio
(m/z). In order to identify peptides by MS, peptides must be ionized and transferred into the gas-phase for analysis by a mass-analyzer. Since the first applications to proteomics, the depth of coverage of the proteome has increased steadily from hundreds or low thousands of proteins [91] to over ten thousand proteins [92, 93]. Behind each improvement in the depth of coverage of the human proteome lies technical advances in sample preparation, technological innovations in instrumentation and computational innovation in analysis. The source datasets analyzed in this thesis were generated using either an LTQ Orbitrap XL or an Orbitrap Elite mass spectrometer (Thermo Scientific, Bremen, Germany). We will discuss in sufficient detail the aspects of sample preparation and instrumentation required to understand these datasets.

1.2.1 Mass-spectrometers in proteomics: a brief primer

The fundamental measurement of interest in mass-spectrometry based proteomics is the mass-spectrum. The mass-spectrum is a measure of the mass-to-charge ratios (m/z) of ions present within the mass analyzer at an instant in time and their relative abundances as measured by intensity. There are four aspects in mass-spectrometry that help in acquiring this measurement in proteomics: the ion source, the mass analyzer, the detector, and the collision chamber [94] (Figure 1.2a). These aspects are often combined into different parts in the mass spectrometer but will be discussed separately here.

1.2.2 Characterizing proteomes using mass-spectrometry

As proteins and their fragments are large fragile molecules, mass-spectrometry based proteomics makes use of so-called “soft-ion” sources to transfer these ions into the gas phase (Figure 1.2b). The soft-ion source is chosen such that polypeptides are transferred into the gas phase with minimal fragmentation. These methods contrast with hard ionization techniques, which impart significant residual energy in the subject molecule resulting in substantial fragmentation. Electrospray ionization (ESI) [95, 96] and Matrix-Assisted Laser Desorption Ionization (MALDI) [97] have been the preferred ion source for creating gas-phase ions from proteins or their fragment peptides. Electrospray ionization is the most widely used technique for the analysis of protein samples, as it ionizes molecules directly from the liquid phase, making it compatible with chromatographic separation techniques used to simplify the peptide mixtures entering the mass-spectrometer. In ESI [95, 96], the protein sample at low pH is transported through a needle placed at high positive potential relative to the mass-spectrometer orifice nozzle. The high electric
potential between the needle and the nozzle causes the fluid to form a Taylor cone [98], which is enriched with positively charged polypeptides at the tip. A spray of positively charged droplets is then ejected from the Taylor cone by the electric field and the droplets shrink through evaporation. Ions are formed at atmospheric pressure before entering the vacuum of the mass-spectrometer. MALDI contrasts to electrospray ionization in that the sample is ionized from solid phase [97]. The protein sample is embedded in a matrix of light absorbing, low-mass aromatic molecules, which are resonantly excited by an ultraviolet or infrared laser pulse. The absorbed energy results in the ionization of the analyte molecules that then enter the mass-spectrometer. The datasets used in this thesis were acquired by electrospray ionization.

Applied to proteomics, the mass analyzer is a device that can separate polypeptides according to their mass-to-charge (m/z) ratios by exploiting their differential behavior when exposed to an electromagnetic field (Figure 1.2c; reviewed in [94]). For this reason, the ion sources must be coupled to a mass analyzer. There are several types of mass analyzers used in proteomics. Here we discuss quadrupole mass filters and the Orbitrap. The quadrupole mass-filter [99] is a device that can be set to let through ions within a very limited m/z range only. The paths of ions falling outside this specific m/z range are changed such that they cannot escape the filter. The mass spectrum from a quadrupole mass filter can be acquired by scanning through the whole m/z range of interest and detecting how many ions pass the filter at each m/z. The Orbitrap mass analyzer [100, 101] traps ions in a device by electrostatic fields that maintain ions oscillating around an inner electrode at frequencies related to their m/z. Oscillating ions around the central electrode produce a current, which can be de-convoluted to produce a mass-spectrum. The instruments used to produce the datasets analyzed in this thesis used Orbitrap mass analyzers to produce mass spectra.

Additional information regarding a peptide sequence is acquired by obtaining a mass-spectrum for its fragments, called an MS² spectrum. Fragmentation occurs within the collision cell in the mass-spectrometer (see [102, 103] for a review) (Figure 1.2d). The collision cell typically follows a quadrupole mass analyzer set to filter for a specific m/z range that maximally isolates a single peptide ion species. Once fragmentation occurs, the m/z ratios of the peptide fragments are measured in the mass analyzer. There are a range of options and fragmentation strategies, but we will focus on those pertinent to the instruments used to generate the data in this thesis: collision-induced dissociation (CID) and higher energy collision-induced dissociation (HCD). Both mass-
spectrometers used in this thesis achieve fragmentation through the use of an ion trap, which traps ions in an electrostatic field. In collision-induced dissociation (CID), precursor ions collide with gas atoms such as helium, and fragment. High-energy CID (HCD) involves the acceleration of the precursor ions to higher energies such that virtually all structurally possible fragmentations occur. The HCD MS\textsuperscript{2} spectra are typically more complex but more reproducible than those obtained using CID. Overall HCD spectra increase confidence in the identification of the precursor and make it more reproducible [104]. Because peptides are polymers of amino acids, peptide fragment ions along the peptide backbone occur in series. Peptide fragment ion series (Figure 1.2e) are indicated by $a_n$, $b_n$, or $c_n$ if the charge is retained on the N-terminus and by $x_n$, $y_n$, $z_n$ if the charge is maintained on the C-terminus [105, 106]. The subscript used indicates the number of amino-acid residues in the peptide fragment. The main fragment ions that dominate CID and HCD spectra are the $y$ and $b$ ion series.

1.2.3 Bottom up proteomics

A wide-spread strategy used to infer proteomes is ‘bottom-up’ proteomics [77, 107, 108]. Sample preparation involves protein extraction from the source material [109] and digestion using enzymes with known cleavage sites (trypsin is typically used) [110], followed by a reverse-phase liquid chromatography step that simplifies the resulting peptide mixture. Peptides eluting from the column are directly transferred into the gas phase by electrospray ionization before entering the mass-spectrometer (Figure 1.2a, Figure 1.3a). From this point, a number of data acquisition strategies are used; each with a specific purpose, performance profile and function. These strategies include data-dependent or data-independent acquisition, which are aimed at achieving unbiased and complete coverage of the proteome and targeted proteomics strategies including selected reaction monitoring, multiple reaction monitoring and parallel reaction monitoring. These targeted approaches are aimed at the reproducible and sensitive acquisition of a subset of known peptides of interest.
Figure 1.2: Primer to mass-spectrometry-based proteomics.

(a) Simplified schematic of an Orbitrap mass-spectrometer. Peptide ions, typically separated by reverse-phase liquid chromatography are ionized using a soft ionization source. The mass-filter, which is turned on only during the acquisition of MS² spectra, is used to select specific peptide ions for further sequencing. Ions are quantized, and each quantum is measured separately within an Orbitrap mass analyzer. Peptide ions selected for further sequencing are sent to the collision cell for fragmentation before the masses of the fragments are measured in the mass-analyzer.

(b) Soft ion sources are used to transfer peptide ions into the gas phase without fragmenting the peptide. Two popular strategies (ESI and MALDI) are depicted. (c) Mass-analyzers measure the mass-to-charge (m/z) ratios of ions being analyzed. Orbitrap mass-analyzers register changes in voltages produced as peptide ions orbit along specific paths related to their mass to charge. Peptide ions produce a complicated signal that is deconvoluted into the m/z ratios. (d) Peptides selected for sequencing are fragmented and the m/z ratios for each fragment provide additional information regarding the amino-acid sequence. Fragmentation is most-often accomplished through collision with a gas in a collision cell. (e) Collision-induced dissociation is optimized such that peptides preferentially fragment across the peptide bond. Different fragment ion series can be observed, and a nomenclature has been developed to annotate them depending on where the break occurred.
(a) Peptide separation (LC) and ionization

(b) DDA: 
MS\(^1\) followed by targeted MS\(^2\)

(c) DIA: 
wide MS\(^2\) window sweeps across m/z

(d) Targeted MS: 
Targeted MS\(^2\) acquired across peptide elution
**Figure 1.3:** Strategies used in bottom up proteomics.

(a) “Bottom-up” proteomics strategies involve the identification of peptides arising from the enzymatic digestion of the sample. Peptides identified in this way can then be used to infer protein presence and abundance. Data-dependent (DDA), data-independent (DIA), and targeted proteomics strategies all fall under the category of bottom-up proteomics (b-d). (b) DDA strategies involve the acquisition of both MS\(^1\) and MS\(^2\). In DDA, each MS\(^1\) is followed by several MS\(^2\) spectra, designed to facilitate sequencing peptides in the original MS\(^1\). (c) DIA involves the acquisition of only MS\(^2\) spectra, but uses a wider m/z window to select ions for fragmentation. The entire m/z region is screened in a repeated sequential manner as the sample passes through the mass spectrometer. (d) Targeted proteomics strategies include single reaction monitoring (SRM) and multiple reaction monitoring (MRM). These strategies specifically isolate and study specific ions according to their known m/z and elution times. As peptides of interest elute off the column, repeated MS\(^2\) spectra are collected across the elution peak to profile fragment abundances. Absolute quantification is obtained by comparing these elution profiles to those generated by a “heavy” stable isotope-labelled version of the same peptide at a known concentration within the same sample.

1.2.3.1 Data-dependent acquisition

The samples processed in this thesis were acquired using data-dependent acquisition. In data dependent acquisition (DDA) [77, 107, 108], two types of spectra are collected in an alternating fashion (**Figure 1.3b**). First, mass spectra of all the ion species co-eluting at a specific point in the elution gradient are recorded. This precursor-ion spectrum, also termed the MS\(^1\) spectrum, outlines the peptide ions present in the mass analyzer at that time. The next series of spectra produced are derived from individual peaks of the MS\(^1\) spectrum, which are used to select small m/z windows that isolate one or just a few peptides for further fragmentation. Peptide fragmentation is typically achieved by collision-induced dissociation or higher-energy collisional dissociation. Each of these fragmentation spectra, also known as MS\(^2\) or MS/MS spectra, are produced when the resulting peptide fragments are re-directed into the mass analyzer. Because the peptide content within the
ion beam is changing along the elution gradient, MS$^2$ acquisition is limited to a short time interval of 1-3 seconds before a subsequent MS$^1$ spectrum is collected. Computational methods, which we will discuss at length, are then used to match the MS$^2$ spectrum, along with the m/z for the source peptide in the MS$^1$ to an amino-acid sequence [111].

Overall, data-dependent acquisition has been the preferred choice for global proteomics pipelines. It has the advantage of specifically targeting and isolating peptides for fragmentation that will produce MS$^2$ spectra that are more easily sequenced. However, this advantage is also a disadvantage. Mass spectrometers have not yet reached the MS$^2$ scan rates required to sequence every peak of interest in the MS$^1$ spectrum, meaning that not all peaks that could be sequenced are sequenced [112]. Exacerbating this issue is that large proportions of peptide species eluting in a global proteomics experiment are of low abundance, requiring long MS$^2$ acquisition times in order to sequence [112]. Yet another significant issue affecting peptide sequencing is co-fragmentation, especially when dealing with complex peptide mixtures like those found in cell-lysates. It has been noted that more than 100,000 likely peptide features elute from the column in a global proteomics experiment of cell lysate [112]. These peptide features, while separated somewhat by liquid chromatography must fall into a tight m/z window of 350-1600 m/z. Overall, the density of the LC vs m/z space is too high to clearly target many individual MS$^1$ peaks for fragmentation, leading to frequent instances of co-fragmentation within mass-spectrometry datasets.

### 1.2.3.2 Data independent acquisition

Data independent acquisition (Figure 1.3c) is another strategy used in discovery proteomics [113, 114]. This strategy confronts some of the issues introduced for data-dependent acquisition, but has other problems that have limited its wide-spread adoption. These methods are exemplified by sequential window acquisition of all theoretical fragment-ion spectra (SWATH). Windowed ranges of m/z values, typically spanning 25 m/z units, are selected and peptides are fragmented before spectra are acquired. The mass-spectrometer is set to rapidly and seamlessly cycle through the entire m/z range within a few seconds. The resulting spectra collected result from complex mixtures of peptides present within each window, which must then be interpreted with regards to their peptide content. The methodology theoretically allows for the unbiased measurement of the entire proteome, but in practice the difficulty of sequencing convoluted co-fragmented spectra generated by this methodology has limited its utility.
1.2.3.3 Targeted proteomics

Targeted analysis of specific peptide ions follows discovery proteomics (Figure 1.3d). Peptides of interest discovered using DDA or DIA strategies are selected for further analysis [115, 116]. These peptides, having known m/z and elution times are specifically selected and monitored over time. Several to hundreds of these can be monitored within a single sample during the latter phases of biomarker discovery allowing for complete coverage of peptides of interest across samples.

1.2.3.4 Improvements in sample-preparation and instrumentation

The far-reaching goal in proteomics is to produce a sample by gene matrix that accurately quantifies the expression levels of each proteoform across all samples. The average tissue expresses between 11,000 and 13,000 proteins [117]. Technological improvements in the proteomics workflow now include online fractionation, which has been able to detect a median of 11,472 protein products within cancer cell lines [92]. The depth of proteome coverage using DDA methods employing significant sample fractionation is surely reaching the expected number of proteins expressed within the sample [118]. However, these sample by gene matrices are still plagued by missing values due to random sampling of MS\textsuperscript{1} peaks for fragmentation [119]. There is currently no one strategy for accumulating complete and accurate profiles for global proteomes. To deal with the lack of protein coverage several enrichment strategies are employed. In-gel fractionation of the protein sample prior to enzyme digestion can improve depth of proteome coverage [120]. Subcellular enrichments of organelles can improve proteome coverage for these regions [121]. Enrichment of post-translational modifications, like phosphorylation, can aid in the detection of specific low-abundance proteoforms [122]. Because of its limitations, DDA methodologies are typically used to discover proteins of interest that vary between differing conditions, but targeted strategies are then used to ensure breadth of coverage across samples [123]. Technological innovations in mass-spectrometry are focused on improving liquid chromatography [124], increasing MS\textsuperscript{1}/MS\textsuperscript{2} scan rates [125], increasing the sensitivity of mass-analyzers [125], and searching for more dimensions on which to separate complex peptide mixtures [126].
1.2.4 Computational strategies for inferring protein sequences from tandem mass spectrometry datasets.

In bottom up proteomics, the mass spectrometer is set to isolate and sequence as many peptide species as the technology will allow. The resulting dataset consists of a series of MS	extsuperscript{2} mass spectra, each tied to the m/z range chosen for fragmentation from a source MS	extsuperscript{1}. A multitude of algorithms (i.e. references [127-139]) (Table 1.1) can then assign protein sequences to these MS	extsuperscript{2} spectra given constraints on the original mass of the peptide from the MS	extsuperscript{1} and the fragment ion series observed in the MS	extsuperscript{2} [111]. This is a difficult process, as fragment ion series are not always complete, and each missing fragment ion increases the plausible alternative sequences for the peptide. Further, co-fragmented MS	extsuperscript{2} spectra that contain more than one peptide ion species can make resolving the peptide sequence a difficult task [112]. Given these two challenges, the proteomics community has opted for supervised approaches to generate these peptide-spectrum matches (PSMs). Supervised approaches assign sequences from a reference protein sequence database to experimental spectra. In doing so, these algorithms make assumptions regarding the protein sequences within the sample. While supervised approaches remain the method of choice, alternatives include semi-supervised and unsupervised strategies that allow PSMs to be assigned to sequence variants [140, 141] or be \textit{de novo} identified without the need for a reference database [142, 143].
Table 1.1: Popular supervised proteomics search engines

<table>
<thead>
<tr>
<th>Algorithm Name</th>
<th>Number of citations</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mascot</td>
<td>7,320</td>
<td>[129]</td>
</tr>
<tr>
<td>SEQUEST</td>
<td>5,642</td>
<td>[127]</td>
</tr>
<tr>
<td>Tandem</td>
<td>1,852</td>
<td>[130]</td>
</tr>
<tr>
<td>Andromeda</td>
<td>1,559</td>
<td>[132]</td>
</tr>
<tr>
<td>Protein Prospector</td>
<td>1,317</td>
<td>[128, 144]</td>
</tr>
<tr>
<td>OMSSA</td>
<td>1,232</td>
<td>[139]</td>
</tr>
<tr>
<td>MyriMatch</td>
<td>399</td>
<td>[131]</td>
</tr>
<tr>
<td>Comet</td>
<td>246</td>
<td>[134]</td>
</tr>
<tr>
<td>MS Amanda</td>
<td>80</td>
<td>[138]</td>
</tr>
<tr>
<td>MSGF+</td>
<td>75</td>
<td>[137]</td>
</tr>
<tr>
<td>Morpheus</td>
<td>59</td>
<td>[135]</td>
</tr>
<tr>
<td>Peppy</td>
<td>11</td>
<td>[145]</td>
</tr>
</tbody>
</table>

1.2.4.1 Supervised approaches for peptide spectrum matching

In supervised protein sequencing strategies (Figure 1.4a), the algorithm makes the assumption that the MS$^2$ spectra it encounters are well represented by a database of protein sequences. In standard proteomic pipelines the database is chosen to be as small as possible, but to cover the human proteome as comprehensively as possible [111, 146, 147]. Typically a reference human proteome database that contains a single canonical protein sequence per gene is chosen for this purpose. The underlying assumption is that enough peptides will be conserved in any given human sample to accurately quantify the expression of each canonical protein. Regardless of the database chosen, MS search algorithms attempt a peptide-spectrum-match for each MS$^2$ spectrum. Peptide-spectrum-matches are made by aligning each spectrum to in silico spectra generated from the sequence database. We discuss the details of reference human proteomes and a few commonly used database search algorithms that are used in this thesis.
Reference databases for use in proteomics

We discuss three different sources of human reference proteomes which we denote UniProt, Ensembl and RefSeq proteomes hereafter. The UniProt KnowledgeBase (UniProtKB) is a large resource of protein sequences and their annotation [80, 148]. UniProtKB is a central resource that combines UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. UniProtKB/Swiss-Prot contains more than 550,000 sequences carefully annotated by experts while UniProtKB/TrEMBL consists of more than 60 million sequences derived, for the most part, from high throughput sequencing technologies. UniProtKB has carefully annotated these proteins into 56,000 reference proteomes including the human reference (UP000005640), which is organized into 20,197 Swiss-Prot entries. Each Swiss-Prot entry corresponds to a gene and is associated with a canonical sequence. However, an additional 21,931 splice isoforms are curated within these entries bringing the total of expertly curated sequences to 42,128. With only 42,128 fully annotated sequences, the human proteome annotated by UniProtKB/Swiss-Prot remains incomplete. To complement this dataset, UniProtKB/TrEMBL, in collaboration with Ensembl, has developed >49,000 additional predicted alternative protein products.

The Ensembl gene annotation system [149], put forward by the European Molecular Biology Laboratory (EMBL), has been used to annotate over 70 different vertebrate species including human. Ensembl constructs a transcript model by aligning biological sequences, including cDNAs, proteins and RNA-seq reads, to the human genome in order to construct candidate transcript models. Careful assessment and filtering of these candidate transcripts precedes the generation of a final gene set and transcript model. Some portion of the transcript model can be translated to protein coding genes, which serves as the model proteome.

The RefSeq annotation pipeline is put forward by the National Center for Biotechnology Information (NCBI) is semi-automated and yields genomic, transcript, and protein RefSeq records [150]. There are a set of curated transcript and protein records in addition to a set of records generated computationally. RefSeq differs from Ensembl in that it uses evidence independent of a genome assembly to represent RNAs and proteins. RefSeq records are derived from sequences deposited to the International Nucleotide Sequence Database Collaboration (INSDC).
Algorithms for database searching

The algorithms used to perform the correlation between reference protein sequences and the set of tandem mass spectra obtained from an experimental protein digest mixture have many features in common. Reference protein sequence databases are used to create a set of potential peptide sequences, based on the known cleavage specificity of a digesting enzyme used in sample preparation. The resulting peptide sequence database, ProteinDB, contains peptide sequences $P = [P_1P_2P_3 \ldots P_n]$ expected to be found within an MS dataset. Each $P_i$ is of the form $p_1p_2p_3 \ldots p_n$ where $p_i$ are one of the 20 standard amino acids $\mathcal{A}$. When analyzing an $MS^2$ spectrum, a filter over ProteinDB is applied ensuring that the difference between the parent ion molecular mass and the experimental precursor ion mass is within a predetermined tolerance. The resulting set of peptides are selected for a detailed correlation between the peptide sequence and the tandem mass spectrum.
**Figure 1.4:** There are three main strategies for identifying peptides from datasets generated by bottom-up proteomics experiments.

(a) Supervised approaches require a reference protein sequence database containing proteins expected within the sample. These protein databases are *in silico* digested and peptides falling within a certain error tolerance of the m/z window used to acquire the MS² (the MS¹ search error tolerance) are chosen as candidate peptide assignments. Each candidate peptide is then scored against the spectrum, using an algorithm-specific methodology and the candidate with the highest score is assigned as the sequence of the MS² spectrum. (b) Semi-supervised or “open-search” approaches still require a reference protein sequence database to generate peptide spectrum matches. However, these strategies allow some reasonable flexibility to account for variant peptides arising from alternative proteoforms. MS-Fagger [140], simply widens the MS¹ search error tolerance in order to identify peptides that would have been missed due to the mass-shifts caused by mutations and post-translational modification. Maxquant dependent-peptides re-searches MS² spectra unassigned by a supervised strategy allowing for chemical modifications amongst those peptides originally identified [141]. These chemical modifications, once examined, amount to common mutations and post-translational modifications. Spectral networking approaches are similar in nature to Maxquant dependent-peptides. However, the search space is limited both to modifications of peptides already identified and to MS² spectra clustered according to their similarity. (c) Unsupervised or “*De novo* sequencing” approaches attempt to generate PSMs without a reference protein sequence database. Because fragment ion series are rarely complete and peptide co-fragmentation is common, each residue in the protein sequence assigned must be resolved probabilistically.

Algorithms differ in the specifics of how a peptide to spectrum match correlation is achieved. However, each algorithm must *in silico* fragment the candidate peptide into a series of peptide fragments of the original sequence $P' = [P'_1P'_2P'_3 ...]$ each with a known fragment ion mass $m(P'_i)$. *In silico* fragment ion series of this sort must be made for each ion type (i.e. a, b, c, x, y, z) to create a theoretical MS² spectrum. The calculated masses in $P'$ are then compared with the observed masses and associated intensities in the spectrum $s = [f_1 = (mz_1,I_1), f_2 = (mz_2,I_2), f_3 = (mz_3,I_3) ...]$ by using a scoring function $S(P,s,θ)$ to evaluate the peptide
spectrum match (PSM). A symmetric error tolerance for matching calculated and observed \( m/z \) values, called the MS\(^2 \) error tolerance, is used during the calculation. This scoring function may include additional factors (\( \theta \)), such as the intensities of individual peaks or the differences between observed and calculated masses and varies from algorithm to algorithm. Once the scores have been calculated for all of the peptides in the list of proteins, a ranked list of PSMs is produced. Further statistical analysis is performed to determine the set of PSMs that correspond to significant correlations. Chapter 3 in this thesis relies on the use of several different database searching algorithms, each of which we will discuss in the subsequent sections.

Dealing with non-canonical sequences and post translational modifications remains a significant challenge in database search strategies and is the topic of Chapter 2. However, typical mass spectrometry search engines do attempt to deal with this issue, so it is worth discussing here. These unexpected MS\(^2 \) spectra result in a mass shift deviating away from a reference sequence in much the same way that a post-translational or chemical modification does. As an example, the phosphorylation of serine in a reference peptide introduces a predictable mass-shift of 80 Da with respect to the unmodified peptide sequence. This mass-shift in both MS\(^1 \) and MS\(^2 \) due to a PTM is equivalent in principle to that caused by a germline mutation from glycine to lysine (71 Da). Since the function \( S(P, s, \theta) \) relies on accurately identifying the fragment ions of \( P \), and some subset of these fragments will be of different mass due to the PTM or mutations, adjustments must be made in order to detect them.

In order to deal with peptides having mass-shifts with respect to the reference sequence in both the MS\(^1 \) and MS\(^2 \) spectra, search algorithms commonly allow for a few mass shifts associated with common chemical modifications to amino acids \( \Delta = [\delta_1 \ldots \delta_l] \). Cysteine typically requires modification in order to be identified [151]. During sample preparation, reactive thiol groups are often blocked using iodoacetamide to produce an S-carboxymethyl derivative with a known mass offset with respect to the native amino-acid. Because the reaction is run to completion, this offset is often included as a static modification (i.e. replacing cysteine with its modified mass in calculations) in mass-spectrometry searches. Most modifications are variable meaning they can occur both modified and unmodified in the sample. As a typical example, Methionine is easily oxidized and can therefore be expected in both oxidized and native forms in the protein sample. Common post-translational modifications like glycosylation, acetylation and phosphorylation can also be expected to be present in the sample. A first step in dealing with these deviations is to
augment the amino-acid alphabet $\mathcal{A}$ to include chemical modifications and post-translational modifications of interest $\mathcal{A}^*$. The new alphabet increases the search space as the augmented database ($\text{ProteinDB}^*$) must include each peptide variant as defined by the modification rules. In practice, allowing for more than just a few chemical modifications, increases the size of the database to the point that spurious PSM identifications start to become a problem.

**Comet**

The COMET [134, 152] scoring function is derived from $xcorr$, which was first introduced by SEQUEST [127] in the context of proteomics. For each MS$^2$ spectrum in the dataset, this score is a closeness of fit measure between an acquired experimental tandem mass spectrum and a theoretical spectrum. The theoretical spectrum represents a candidate peptide sequence obtained from a sequence database. An *in silico* digestion of the candidate peptide produces $y$ and $b$ ions, amongst others, which must be aligned to the spectrum being evaluated. The $xcorr$ scoring function is a scalar dot product between the acquired and theoretical spectrum as shown below.

$$S(P,s) = R_0 - \frac{\left(\sum_{\tau=\pm75} R_{\tau}\right)}{151}$$

Where

$$R_{\tau} = \sum P'_{\tau} \cdot f_{\tau}$$

The variable $R_{\tau}$ is a correction factor which is allowed to vary across a 150 Da window. The $xcorr$ is calculated for every candidate peptide in the database search. The score histogram is then used to generate an expectation value or E-value. To generate the E-value, the cumulative distribution function of the score-histogram is log transformed. Following this, a linear least squares regression is fit to the transformed data and the E-value is extrapolated from where the top hit falls on this regression line.

**Tandem**

Tandem [130] offers an alternative approach for obtaining a peptide spectrum match. Using the simple dot product between theoretical and experimental spectrum implicitly assumes a linear relationship between the numbers of matches between the two spectra. Consider for example a
theoretical spectrum A and B having 11 amino-acids. If the actual spectrum contains 1 peak matching A and 10 peaks matching B, the algorithm described for comet above would consider B roughly 10 times stronger than A. However, intuitively, the number of peaks makes a significant difference to the confidence of the spectrum. Tandem employs a hyperscore $x_{hyperscore}$ which accentuates the number of a and b ions matched to the actual spectrum.

$$S(P, s) = x_{hyperscore} = \left(\sum_{i=0}^{n} f_i P'_i\right)(n_b! n_a!)$$

Similarly to COMET, $x_{hyperscore}$ can be converted into an E-value, which is used in practice. However E-value calculation differs somewhat from comet. The score histogram is restricted to the monotonically decreasing right half or higher scoring region. This region of the score histogram is then log-transformed and a linear model is fit. The E-value is then calculated by extrapolating the hyperscore for a particular PSM onto the resulting linear model.

**MS-GF+**

MS-GF+ [137] is more complex. This algorithm similarly uses a dot-product scoring $\text{Score}(P, s) = P' \cdot S^*$ but first converts the peptide P and spectrum s to a peptide vector $P^*$ and a spectral vector $S^*$. $S^*$ has a score at every nominal mass up to the parent mass of the spectrum. The spectral vector is of length $M$, where $M$ is the nominal precursor mass of the MS² spectrum. Conversion from the experimental spectrum to an experimental spectral vector proceeds as follows. First, noise is removed from the spectrum by retaining only the top k ion signals ($k = 150$ by default) and subsequently peaks in the resulting mass spectrum are used to create a ranked spectral vector in decreasing order in terms of intensity $\{(mz_i, rank_i) \cdots (mz_i, rank_i)\}$. For a specific sequence, each $(mz_i, rank_i)$ may correspond to a specific fragmentation ion type which can be denoted using the triplet $(\text{charge}, \text{offset}, \text{sign})$, where charge is the charge of the peptide fragment, offset positions the peak in $m/z$ and sign designates it as a prefix-ion (i.e. b-ions) or post-fix ion (i.e. y-ions). Using this nomenclature a b-ion can be assigned $(1,1,1)$ and the y-ion $(1,19,−1)$. The spectrum can be represented in an intermediate step as $S_{ion}$. 
$S_{ion} = \{(mass_1, rs_1) \cdots (mass_l, rs_l)\}$. Where

$$mass_j = \begin{cases} [mz_j \cdot \text{charge} \cdot 0.9995] - \text{offset} & \text{if } \text{sign} = +1 \\ \text{PrecursorMass}(S) - ([mz_j \cdot \text{charge} \cdot 0.9995] - \text{offset}) & \text{if } \text{sign} = -1 \end{cases}$$

and

$$rs_j = \text{RankScore}(ion, rank)$$

$\text{RankScore}(ion, rank)$ is an empirically determined function that returns a probabilistic log-likelihood score that the peptide from which the spectrum was derived contains a fragment ion of mass $m$ and negative peaks indicating that peaks at these masses are most likely generated by noise in the spectrum. The empirically derived function was computed across different instruments, fragmentation methods and takes into account the location of the observed peak in m/z space, the intensity (rank) of the fragment ion, presence of other ions supporting the fragment-ion, and the isotope patterns of the fragment ion in question. Note that $S_{ion}$ can have duplicate entries for some $mass_j$ as the algorithm attempts to assign ions in the spectrum to different ion types. The final spectral vector $S = \{S_1 \ldots S_m\}$ is determined as

$$S_i = \sum_{ion \in \{ion_{types}\}} \max(\{|rs|(mass_i, rs) \in S_{ion}\} \cup \text{RankScore}(ion, NA))$$

Where $\text{RankScore}(ion, NA)$ is the low score given to a missing ion.

**Target decoy approach for controlling false discoveries**

The algorithms discussed in the previous section identify the highest scoring peptide given an MS$^3$ spectrum and a list of candidate peptides, regardless of whether the correct peptide is in the list. In order to control the false discovery rate, a target-decoy approach is used to set a PSM score threshold for a successful peptide identification [153, 154]. In this approach, each protein sequence in the reference human proteome is reversed and appended to the reference. After searching using the algorithms above, a set of peptide spectrum matches is obtained containing sequences belonging to the human reference proteome (forward hits) and sequences belonging to the reversed reference proteome (reverse hits). PSM scores are affected by peptide length, and peptide length correlates with charge (data not shown). Hence, peptides are then binned according to charge or length and ordered from highest scoring to lowest scoring. A PSM score threshold is determined
for each bin by limiting the percentage of reverse hits present in the final list. This methodology is now a standard method by which false discovery is controlled in proteomics.

1.2.4.2 Semi-supervised ‘open’ searching peptide spectrum matching

Semi-supervised, or ‘open’ searches (Figure 1.4b), for peptide spectrum matching make fewer assumptions about the alterations to proteins expected within the sample [140, 141, 155]. These strategies are still supervised, in the sense that they rely on a FASTA database of protein sequences expected within the sample, but do not make assumptions about the protein variants that may occur. Instead, these strategies attempt a match to the reference sequence by either opening the MS\textsuperscript{1} search tolerance [140], allowing for mass shifts within the MS\textsuperscript{2} spectra that were not matched by an MS\textsuperscript{1} [141] or clustering assigned and unassigned spectra before identifying modified PSMs [155]. These approaches for peptide spectrum matching make assumptions about the proteins expected within the sample but account for the possibility for variation in the protein sequence.

1.2.4.3 Unsupervised approaches for peptide spectrum matching

Un-supervised approaches (Figure 1.4c), also known as ‘de-novo’ sequencing approaches [142, 143], make no assumptions about the alterations to proteins expected within the sample. Instead, only the MS\textsuperscript{2} spectrum, and its parent ion MS\textsuperscript{1} range are taken as input, and used to generate the protein sequence. While an attractive option, these strategies are prone to error arising from incomplete ion series. Since de-novo sequencing can only rely on the presence/absence of ions in order to determine the sequence, and the average PSM has only 85% peptide sequence coverage [156], it is easy to understand why the proteomics community has commonly opted to use supervised methodologies over these unsupervised strategies for PSM.

1.3 Summary

Cancer is a highly complex disease, and understanding how the molecular phenotype of the cancer cell emerges from genomic and epigenetic aberrations will require the integration of multiple molecular profiles across the central dogma. This thesis is focused on a specific problem in integrative –multi omics: Methods for the detection of variant proteins within mass-spectrometry based proteomics datasets. In this Chapter the identification of variant proteins was put in context of a much broader problem being investigated by the proteomics community: the complete
elucidation of cancer proteotypes, which includes a complete characterization of the proteoforms of each gene.

The subsequent Chapters push this idea further. As we will see in Chapter two, the identification of protein variants was suspected to benefit from their identification by genomics and transcriptomics. It is the subject of Chapter 3 to establish which methodology for incorporating genomic information will yield improved outcomes for variant protein detection. Here, I will rely on concepts introduced in this Chapter for the peptide-spectrum match algorithms. Finally, in Chapter 4, I discuss future directions for this work in the context of uncovering complete and accurate proteoforms from global proteomics. I will discuss the integration of our methodology with the semi-supervised strategies that can also be used to identify modified reference peptides within global proteomics mass spectrometry datasets.
Chapter 2
Onco-proteogenomics: cancer proteomics joins forces with genomics

Abstract

The complexities of tumor genomes are rapidly being uncovered, but how they are regulated into functional proteomes remains poorly understood. Standard proteomics workflows use databases of known proteins, but these databases do not capture the uniqueness of the cancer transcriptome, with its point mutations, unusual splice variants and gene fusions. Onco-proteogenomics integrates mass spectrometry–generated data with genomic information to identify tumor-specific peptides. Linking tumor-derived DNA, RNA and protein measurements into a central-dogma perspective has the potential to improve our understanding of cancer biology.

This Chapter was originally published in Nature Methods:

2 Onco-proteogenomics: cancer proteomics joins forces with genomics

2.1 Introduction

Large-scale cancer genome sequencing studies are beginning to reveal the complexity and heterogeneity of cancer genomes [34]. These studies have helped paint a picture of a dysfunctional transcriptome harboring large numbers of nonsynonymous single-nucleotide variants (SNVs) [157], insertions and deletions (indels), aberrant gene fusions [158], alternative splice variants [159] and copy-number aberrations [160].

Proteins are central to cellular function, and aberrant proteins drive tumor initiation, progression and response to treatment. Nevertheless, the use of proteomic data to model these phenotypes has been rare compared to the usage of other types of genomic data. This has in part been caused by the ubiquitous use of microarrays to estimate mRNA abundance and in part by the rapid progress in genomic profiling using high-throughput sequencing. As protein and mRNA abundances are only partially correlated [161, 162], a major unexplored question in cancer biology is how the information flow from genome to proteome is altered in tumors. High-throughput methods for investigating the cancer proteome are thus critical to facilitate the identification of changes in signaling pathways, of protein isoforms and of post-translational modifications (which themselves are potential therapeutic targets).

Onco-proteomic studies can take a variety of forms. Studies by The Cancer Genome Atlas and others have used reverse phase protein arrays to measure abundances of cancer-related proteins [163]. Although useful for this purpose, these arrays require a priori knowledge of the proteins of interest and availability of suitable antibodies, and they measure only a small fraction of the proteome. Similarly, ribosome profiling allows direct and global monitoring of mRNA translation through quantitative genome-wide assessment of ribosome occupancy [164]. However, protein abundance is ultimately a combination of both translation and degradation, limiting the interpretability of these studies. By contrast, mass spectrometry (MS) provides wide proteome coverage and is being rapidly improved by technological advances. Here, we focus on how MS-based proteomic studies can exploit cancer genomic information to better interrogate tumor-specific aspects of the proteome.
Recent technological and methodological advances in MS have greatly improved proteome coverage, reducing biases and increasing precision [118]. Characterizing a cancer proteome generally follows a bottom-up approach: samples are obtained as tissues, as cell lines or from peripheral fluids. Proteins extracted from each sample are digested using enzymes with a high specificity for known cleavage sites, such as trypsin. Finally, tandem mass spectra are generated for individual peptides using a mass spectrometer. These spectra can then be identified by searching for matches in generalized human proteome databases.

It is at this stage of database searching that proteomic approaches lose the capacity to identify many tumor-specific genetic aberrations. Sensitivity can be increased by searching against specialized databases such as OMIM [165], neXtProt [166], ECgene [167], ChimerDB [168] or COSMIC [169], which catalog genetic aberrations present in cancers (Figure 2.1). De novo sequencing algorithms can directly infer protein sequence from tandem MS (MS/MS) spectra without requiring database matching, although this approach suffers from several challenges, including ambiguous interpretations of MS/MS fragmentation spectra [170].

Thus, the first step toward 'personalized' onco-proteomics, wherein each sample is considered to have a unique universe of potential proteins, is the mass spectrometric detection of tumor-specific aberrant proteins by leveraging other -omic data to create custom search databases, an approach that can be classified under the broader term of proteogenomics. By representing aberrations specific for the tumor being analyzed, custom databases allow tumor-associated peptides to be assigned, thereby improving accuracy and proteome coverage. Here, we review the history of this approach and provide a perspective on how improvements in onco-proteogenomics can help refine the questions posed by cancer researchers.
Previously uncharacterized genetic aberrations cannot be identified by searching proteomic MS/MS spectra against a general protein database. For detection of these anomalies, modified databases must be generated by applying other -omic technologies to the tumor in question. Predicted proteins from genome sequencing, exome sequencing or transcriptional profiling may help to build these augmented databases and support 'genomically informed proteomics'. EST, expressed sequence tag; WGS, whole-genome sequencing.
2.1.1 Proteogenomics technology primer

In many ways, onco-proteogenomics is a specialized subclass of traditional proteogenomic approaches. Proteogenomics lies at the intersection of proteomics and genomics and has been essential in validating and extending well-established gene and proteome models for humans [171], other organisms [172, 173] and even subcellular organelles [174]. In the traditional proteogenomic approach, peptides identified by MS through searching against custom databases generated from genomic or transcriptomic data sets provide definitive proof of protein translation, yielding powerful evidence for the correctness of a gene model. Custom databases based on matched genomic or transcriptomic data can also be used to identify aberrant peptides not present in generalized protein databases: for example, those derived from fusion proteins, SNVs that change the primary protein sequence and aberrant splicing events.

2.1.2 Genome sequencing data

Whole-genome or exome sequencing can be used to generate tumor-specific peptide databases (Figure 2.1). Most commonly in proteogenomic applications, MS/MS spectra are assigned using a six-frame translation of whole genome sequences, allowing for the detection of novel (previously undiscovered) exons and open reading frames [175]. Further, existing exon definitions for the reference human genome can be used to generate junction sequences for all putative exons within a sample [172, 176]. Short tryptic peptides that cross the junction can then be added to the search database. Tumor-specific protein sequences from SNV data obtained through either DNA or RNA sequencing of tumors can also be appended to existing protein databases. Incorporating genome sequencing data has the downside of longer search times and elevated false discovery rates (FDRs) caused by the very large database search space.

2.1.3 Transcriptomic sequencing data

RNA-seq, expressed sequence tag and cDNA-based databases infer the expected proteome directly, avoiding the unnecessarily large databases that result from whole-genome sequencing or exhaustive exon splice junction analyses. RNA-seq based databases also allow for the proteomic validation of RNA editing, splice junctions and chimeric proteins. Notably, RNA editing can be difficult to detect from RNA-seq data alone, as demonstrated by controversies where its extent has been artificially amplified [177-181]. Chimeric proteins are also frequently called from RNA-seq data [182], but their existence is often questioned, demonstrating the need for proteomic analysis.
In cancer studies, tryptic peptides spanning splice junctions and putative chimera can also be compiled from RNA-seq data [183] or from existing databases of chimeras [182], without the need for the generation of exhaustive splice junction databases.

2.1.4 Proteogenomic database generation and searching

Approaches for generating and searching large proteogenomic databases were reviewed recently [184, 185]; here we will simply highlight a few available methods. Peppy [145] is a tool that generates a peptide database from a genome, tracks peptide loci, matches peptides to tandem spectra and returns identifications at the desired FDR threshold. Exon graph methods construct a compact graph representation of the database while simultaneously covering many theoretical splice variants and single-nucleotide polymorphisms (SNPs) in all genes [173, 186]. CustomProDB is a promising new tool for constructing customized protein sequences from RNA-seq data [187].

2.2 Proteomic detection of tumor-specific peptides

2.2.1 Onco-proteogenomics in the literature

The use of -omic data to improve human gene models is now well established, suggesting that the technologies for sample-specific proteomics are available and mature. The merits of a sample-specific approach for onco-proteomics are now just beginning to be explored, with a gradual transition from cell-line studies toward translational research (Table 1.1). The first large-scale analyses of primary tumors are now emerging [28] and under way.

The word 'onco-proteogenomics' was coined in a short article by Helmy and coworkers [188] that analyzed 15 shotgun proteomics runs of HeLa S3 cells. Resulting spectra were searched against four databases containing canonical human sequences corresponding to exon junction peptides, exonic peptides from cDNA and mRNA databases, and intragenic peptides (regions between known coding regions, based on current annotations) from whole genome–based databases. Unassigned spectra were then searched against a peptide database derived from the HeLa S3 transcriptome, leading to the identification of 25 cancer-specific peptides, including phosphorylated sites. In a similar, but more comprehensive, study of HeLa S3 cells, RNA-seq data were used to create a database that identified over 450 new peptides [189].
Table 2.1: A survey of onco-proteogenomic implementations in the literature

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Tissue type</th>
<th>Human general&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Cancer specific&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Sample proteogenomic&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Proteogenomic strategies used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myelogenous leukemia [190]</td>
<td>Cell line</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Six-frame genome translation, mRNA transcripts and UCSC hg19</td>
</tr>
<tr>
<td>Acute T-cell leukemia [183]</td>
<td>Cell line</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>RNA-seq to build junction database</td>
</tr>
<tr>
<td>Colorectal cancer [191]</td>
<td>Cell line</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>RNA-seq to predict putative SNVs and junctions</td>
</tr>
<tr>
<td>Colorectal cancer [192]</td>
<td>Cell line</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>RNA-seq to predict putative SNVs and junctions; six-frame translation of genes near microsatellite instability hotspots</td>
</tr>
<tr>
<td>Cervical cancer [189]</td>
<td>Cell line</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Cell-line RNA-seq data used to predict putative SNVs and junctions</td>
</tr>
<tr>
<td>Epidermoid carcinoma [193]</td>
<td>Cell line</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Proteomics obtained by fractionation based on isoelectric point (pI); cell-line RNA-seq data and six-frame translation of standard genome model used to predict proteome; resulting peptide search database fractionated by calculated pI</td>
</tr>
<tr>
<td>Pancreatic cancer [194]</td>
<td>Mouse model</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>ECgene-based database to identify and characterize tumor-associated splice isoforms for proteins of interest in cancer pathways</td>
</tr>
<tr>
<td>Colorectal cancer [28]</td>
<td>Mouse xenograft and tumor</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>dbSNP, COSMIC and RNA-seq data used to generate augmented database</td>
</tr>
<tr>
<td>Liver cancer [176]</td>
<td>Tumor</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Putative exon-exon database using genome model</td>
</tr>
<tr>
<td>Gastric cancer [196]</td>
<td>Tumor</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>neXtProt-based search database used to identify a previously unobserved protein isoform, gastrokine isoform 2, expressed within cancerous tissue</td>
</tr>
<tr>
<td>Lung cancer [197]</td>
<td>Cell line and tumor</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Fusion-peptide database generated from ChimerDB 2.0 and used to search non-small-cell lung cancer MS data</td>
</tr>
</tbody>
</table>

<sup>a</sup>General human proteome database used for searching.

<sup>b</sup>Specialized cancer aberration database used for searching.

<sup>c</sup>Sample-specific proteogenomic data set used for database searching.
Extensive proteogenomic analysis of colorectal cancer cell lines [191] was performed using RNA-seq data to generate a customized database in two steps. First, a threshold was used to remove low-abundance mRNAs according to the reads per kilobase per million mapped reads (RPKM, a measure of mRNA abundance). A threshold of 2 was set by studying the RPKM distribution for all proteins identified in the regular protein database and identifying an inflection point above which protein identification increased sharply. Second, high-quality nonsynonymous SNVs [198] were identified using the RNA-seq data. Proteins corresponding to these mutations were added to the database, leading to an additional 86 and 89 protein groups (a set of proteins that cannot be distinguished on the basis of MS-identified peptides) for the SW480 and RKO cell lines, respectively. This two-stage approach increased the total number of identified protein groups by 5.9% at a fixed FDR.

A subsequent study used proteogenomic analysis of transcriptomic and proteomic data on ten colorectal cancer cell lines to illuminate adaptations to deficiencies in the mismatch repair (MMR) machinery [199]. As deficiencies in MMR are commonly associated with microsatellite instability, a database was made from six-frame translation of genes predicted to be most vulnerable. This was supplemented with RNA-seq transcriptome data, leading to the identification of 571 variant sequences that would not have been identified using a general protein database. The study suggested that MMR deficiencies are accompanied with a stress response targeting misfolded proteins for degradation. A key insight was that detected variant peptides were enriched for previously characterized SNPs over cancer-specific SNVs; one possible explanation for SNP enrichment is decreased stability of proteins containing new sequence variants.

Although many onco-proteogenomic studies have used cell lines, several emerging studies are focusing on primary tumors. For example, a striking report from the Human Proteome Project's chromosome 15 team on gastric cancers [196] used a specialized neXtProt-based search database to identify a previously unobserved protein isoform, gastrokine isoform 2, expressed within cancerous tissue. Custom search databases for alternative splice variants have also been used [195] to identify tumor-associated splice isoforms related to cancer pathways. The Biology/Disease-driven Human Proteome Project is also an emerging consortium applying proteogenomics to several different diseases [200].
2.3 Challenges in onco-proteogenomics

Unlike proteogenomic studies that aim to better characterize the proteome and refine gene models, onco-proteogenomic studies revolve around detecting tumor-specific changes in the proteome. Such changes may lead to tumor initiation, progression and adaptation to treatment. As these approaches become more widely used, a number of challenges will need to be addressed, including some already commonly faced in large-scale genome-sequencing studies. For example, statistical models that discriminate between 'passenger' and 'driver' mutations need to be developed and applied, and data-deposition and reporting standards must be created and adopted [201].

A second significant challenge is that not all peptides in a sample can be detected. There are two reasons for this: the large dynamic range in protein abundance and the lack of selection of all parent ions for fragmentation \((i.e.,\) random sampling) [119]. As a result, there is often a clear mRNA abundance threshold below which no peptides can be identified [191]. Because many cancer-associated aberrations affect tumor suppressors, loss-of-function mutations are often of interest. Many loss-of-function aberrations cause incorrect protein folding, which increases the propensity for proteasomal degradation and, subsequently, low protein abundance. Thus, these mutated peptides may not be detectable by current mass spectrometers. Indeed, onco-proteogenomic studies of cancer cell lines are starting to reveal this trend [191, 199]. However, it should be noted that loss-of-function mutations do not always reduce protein abundance: mutations of the tumor suppressor p53 often increase protein abundance because of dominant-negative effects of the mutated protein [202].

Another challenge arises from the sheer size of putative protein sequence databases produced by genomic studies, which can dwarf the number of aberrations expected in a cancer proteome. To illustrate: both the commonly used six-frame translation and the three-frame translation of transcriptomes can generate more than \(2 \times 10^6\) sequences. This expands the size of typical protein databases by about two orders of magnitude. By contrast, a typical cancer genome may only harbor a few hundred mutations in the exome [203], of which many will be passenger mutations. To resolve this database size challenge, typical proteogenomic studies utilize stringent cutoffs. However, applying stringent thresholds minimizes the false positive rate at the expense of a high number of false negatives, and increasing database size increases the number of false positives expected by chance alone. One strategy to resolve this issue is to take a two-step search approach.
to improve the sensitivity of peptide-spectrum matching [204]. First, the augmented database is searched without multiple-testing control. Next, peptides identified in this way are used to generate a second search database. Final searching against this smaller database is then done with control for false discoveries using reversed sequences or other statistical methods [205]. Although this strategy may increase the number of peptides identified, it raises the possibility of error propagation without directly correcting the problem of false discovery. Thus, there is an urgent need to develop improved statistical methods and to critically analyze search results arising from augmented protein sequence databases.

A final obstacle to the widespread use of onco-proteogenomics involves challenges with integrating proteomic data with other types of genomic or metabolomic data. Such integration is of fundamental value for understanding the phenotype of cancer cells as well as for the development of multimodal biomarkers. But, because the proteome is spatially distributed into many different subproteomes (such as the N-glycoproteome or phosphoproteome), it is unclear which aspects of the proteome need to be studied to understand the systems flow of information ultimately leading to the cancer phenotype or to create effective biomarkers.

2.3.1 Toward integrative cancer phenotyping

The literature is abundant with publications that have studied the molecular phenotype of cancer cells using transcriptomic data or that have analyzed how genetic aberrations are associated with changes in mRNA abundance. Onco-proteogenomics can make such analyses more accurate by incorporating precise proteomic data and showing how the proteome is affected by the cancer genome. The techniques for developing this holistic view have only recently become available, as peptide and protein coverage by MS was previously not deep enough and tumor-specific peptides were often not detected. The characterization of tumor-specific MS-search databases will be critical to effectively merge cancer genomics and proteomics: for example, to illuminate the effects of genetic aberration on steady-state protein abundances and on the post-translationally modified proteome.

2.3.2 Gauging the success of onco-proteogenomic techniques

Many questions will need to be answered as onco-proteogenomic approaches are applied to human tumor tissues. The benefits of this technique will depend on the numbers and types of genetic aberrations that can be detected and on the extent of orthogonal information that onco-
proteogenomics will offer. One potential metric of success is much as in traditional proteogenomic studies the number of new peptides or proteins identified by onco-proteogenomics that are not found using standard approaches. However, this approach can lead to overfitting to maximize the number of tumor-specific peptides, elevating the false positive rate. This is especially true because the full repertoire of proteins is not known, even for normal human tissues. The detection of SNVs and indels by MS is a difficult problem, but searching against alternative 'decoy' search databases may provide a better estimate of the false positive rate; community competitions may play a key role in fostering improved statistical methodologies [206].

2.4 Potential impact of onco-proteogenomics in cancer research

Although we have focused on improving proteomic characterization by integrating genomic and transcriptomic data, the converse may also be true. Proteomic data are orthogonal to genomic data and can supplement integrative studies with other -omic data types. For example, proteomics can uniquely and directly observe post-translational modifications, changes in subcellular localization and protein-protein interactions.

2.4.1 Reducing false negative exome calls.

Detecting somatic aberrations from high-throughput sequencing data remains challenging [207]. Current genomic analysis methods are plagued by unknown false negative rates, particularly in repetitive regions of the genome, leading to potentially cancer-relevant somatic mutations being missed. Proteomics approaches can help validate nonsynonymous SNV calling or indel predictions. Further, by reducing current stringent cutoffs for SNV and indel detection, proteomic data could be used not just for validation but also to help improve parameterization of genomic analyses (Figure 2.2a). Such an approach would benefit cancer genomics as a whole by allowing for the detection and validation of aberrations otherwise lost as false negatives.
Figure 2.2: The potential impact of onco-proteogenomics in cancer research.

(a) False negative rates for aberrations called from whole-genome sequencing, RNA-seq and exome sequencing may be lowered by temporarily reducing the stringency used to call aberrations and then conducting a proteogenomic search to identify false negatives. TP, true positive; FP, false positive; TN, true negative; FN, false negative. (b) Proteogenomics has the benefit of being able to identify aberrations that are differentially expressed between tumor and normal pairs. Searching among these mutated proteins may reveal valuable drug targets. (c) Genetic aberrations validated through proteogenomic searches may have downstream effects on the proteome. For example, mutation of a kinase may substantially alter protein phosphorylation states. The identification of differentially phosphorylated proteins may amplify the signal of an aberrant kinase in the generation of a molecular signature for cancer. (d) Because proteogenomics can be a ‘personalized’ approach, it can be applied along with the other -omic technologies to improve our understanding of the information flow from gene to protein. In the end, protein abundances are the net result of transcription, translation, degradation and any regulation therein.
MS can identify only moderately to highly abundant peptides, although newer liquid chromatography (LC)-MS technologies are now providing impressive proteome depth of around 10,000 proteins [120, 208, 209]. Although there is still the caveat that many aberrations will be neutral passenger mutations, deep proteome profiling should help to focus on not only aberrations detected at the transcript level but also those apparently expressed within the proteome. Finding aberrant peptides corresponding to potentially actionable proteins is of vital therapeutic interest (Figure 2.2b). Reduced sets of aberrant peptides identified by proteomics will provide a significant push in the search for actionable targets among possible driver mutations.

2.4.2 Improving understanding of cellular networks and dysregulation in cancer.

The hallmarks of cancer define biological capabilities acquired during the development of human tumors [4]. A complete picture of these tumor phenotypes is difficult to measure; instead, the experimental focus in systems biology has been to characterize the molecular phenotype, in which the proteome plays an important role. Evidence is emerging that protein abundance may be under greater purifying selection than mRNA abundance, suggesting that the proteome more accurately reflects molecular alterations in cancers [210]. Incorporating the proteome into -omic studies will lead to a more complete understanding of how cellular networks and canonical signaling pathways are dysregulated. Along with the detection of mutant or aberrant proteins by MS, onco-proteogenomic characterizations of cancerous molecular phenotypes will need to include measurements of protein abundance, subcellular localization, protein interactions and post-translational modification.

Often, gene-specific aberrations can affect the proteome indirectly by altering enzymes involved in post-translational modifications, thus activating oncogenic pathways and deactivating tumor-suppressive pathways. The most prevalent examples are mutations in kinases and in chromatin remodeling factors [35], though enzymes involved in protein N-glycosylation [211] and proteolytic cleavage [212] are also implicated. Aberrations within pathways that regulate post-translational modification are commonly predicted by the other -omic technologies and used in biomarker design [213]. Complementarily, proteomics experiments commonly use differential abundances of post-translational modifications in the construction of clinically relevant biomarkers [214]. Proteogenomics provides an opportunity to bridge biomarkers generated by other -omic technologies and their downstream effects, direct or indirect, on the proteome (Figure
2.2c). Interesting observations will also emerge, as the detectability of certain genetic aberrations by MS is related to shifts in cancer-related signaling pathways observed through post-translation–specific proteomic strategies.

Steady-state protein abundances are dependent on the synthesis and degradation rates of both mRNA and protein. This heightens the need to understand dysregulation of both pathways commonly disrupted in cancer cellular networks. There is already much evidence, for example, that synonymous SNVs and genomic alterations in regulatory regions are important for changes in the overall proteome [215, 216]. Proteomics approaches are currently the best method to resolve protein steady-state levels, and recent innovations in protein MS allow quantification of as many as 10,000 proteins from whole tissue lysate [120, 209]. Although MS provides an accurate measure of protein abundance, inferring translation and degradation rates is more difficult [217]. A complete study of the proteome would include an analysis of protein synthesis rates through ribosome profiling followed by an analysis of turnover rates either by the abundance of ubiquitinated proteins or by more complex methodologies such as pulse-labeling or spatial proteomics to characterize abundance, localization, synthesis, degradation and turnover rates [218]. Employing a proteogenomic approach in such studies would highlight the intricacies of post-transcriptional regulation, completing our understanding of how the information flow from gene to protein is disrupted in cancer (Figure 2.2d).

2.4.3 Top-down proteomics, proteogenomics and the proteoform.
A recent paradigm shift has subtly changed the way the proteome is perceived, with important implications for proteogenomics. The focus of proteomics is shifting from simply characterizing the number of expressed proteins to defining the proteome as an amalgamation of 'proteoforms'. Each proteoform represents a different expressed variation of a gene (for example, splice variants or post-translational modifications). Understanding how proteoforms change during cancer initiation, progression and response to treatment is an important question that currently can only be addressed by top-down proteomics [76]. A more complete characterization of normal and aberrant proteoforms and their relationship to cancer pathways will undoubtedly improve our understanding of how genotypes relate to clinico-molecular phenotypes.
2.4.4 Targeted MS and translation to clinical research.
Some of the aberrations within the cancer genome eventually perpetuate the cancer-specific phenotype. For example, once diagnostic proteins have been detected using a proteogenomic strategy, MS could shift to using targeted techniques such as multiple reaction monitoring (MRM). Guided by a reference spectrum, predefined peptides are identified using only a few selected fragment ions (termed 'transitions'). Current MRM experiments can detect more than 100 peptides at a time, enabling rapid and targeted analysis of large biomarker panels or complete signaling pathways [219]. Detection of low-abundance peptides can still be difficult, but MRM assays can be particularly sensitive, selective and high throughput if sample complexity is simplified, either through biochemical enrichment [220-222] or through the use of analyte-specific antibodies [223]. This presents a challenge for onco-proteogenomics, as this lack of sensitivity is often seen with clinical material, especially when peptide antibodies cannot be used for enrichment.

Quantitative MRM assays have been developed for altered proteins. SDS-PAGE (polyacrylamide gel electrophoresis)-based fractionation of the proteome, followed by a targeted LC-MRM approach, was used to quantify tumor-derived mutant forms of KRAS [222] in cancer cell lines. Others have quantified levels of epidermal growth factor receptor (EGFR), often abnormally expressed in epithelial tumors, in formalin-fixed tumor samples from non–small-cell lung cancer patients using MRM [224]. MRM was shown to have greater quantitative resolution in determining abundance levels of EGFR than immunohistochemistry, the current clinical method to measure protein abundance. MRM therefore provides a more precise and multiplexable methodology to quantify the abundance of altered proteins in clinical samples.

2.5 Concluding remarks
Classic protein database searching methodologies cannot identify peptides corresponding to unique tumor-specific genetic aberrations. Tumor proteomes can profitably be examined in a similar fashion to proteogenomic approaches for nonmodel organisms, for which protein database searches are conducted against proteins predicted by other -omic technologies. Insights gained from using proteogenomic approaches will greatly improve our understanding of the cancer proteome and facilitate the union of proteomics and integrative genomics. Proteomic technologies have the further potential to provide alternative insights not easily obtained by upstream genomic technologies (i.e., protein abundance, proteoform distribution, post-translational modifications,
subcellular localization and protein-complex membership). In the end, bioinformatics integration of these complementary -omic data in conjunction with robust statistical interpretation will provide a more complete understanding of cancer biology.
Chapter 3
Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines.

Abstract

Background: Onco-proteogenomics aims to understand how changes in a cancer’s genome influence its proteome. One challenge in integrating these molecular data is the identification of aberrant protein products from mass-spectrometry (MS) datasets, as traditional proteomic analyses only identify proteins from a reference sequence database.

Methods: We established proteomic workflows to detect peptide variants within MS datasets. We used a combination of publically available population variants (dbSNP and UniProt) and somatic variations in cancer (COSMIC) along with sample-specific genomic and transcriptomic data to examine proteome variation within and across 59 cancer cell-lines.

Results: We developed a set of recommendations for the detection of variants using three search algorithms, a split target-decoy approach for FDR estimation, and multiple post-search filters. We examined 7.3 million unique variant tryptic peptides not found within any reference proteome and identified 4,771 mutations corresponding to somatic and germline deviations from reference proteomes in 2,200 genes amongst the NCI60 cell-line proteomes.

Conclusions: We discuss in detail the technical and computational challenges in identifying variant peptides by MS, and show that uncovering these variants allows the identification of druggable mutations within important cancer genes.

This Chapter was originally published in Genome Medicine. Supplementary data can be obtained from the online publication (https://doi.org/10.1186/s13073-017-0454-9).

3  Background

A global effort is underway by cancer researchers to annotate biobanks with molecular data captured across the genome, transcriptome and proteome. While the genomics and transcriptomics communities have established pipelines for the identification of disease variants, it remains difficult to elucidate the consequences of these variations on the proteome. There is a need for better methodologies to characterize all protein variants, formally defined as proteoforms [76], from global proteomics datasets. This includes germline, somatic and post-translational modifications, including all possible combinations, of any given protein. However, the identification of post-translational modifications and coding consequences of genomic variations are conceptually different, since genomic and transcriptomic studies can provide orthogonal evidence for the existence of such a variant.

A fundamental task in mass spectrometry-based proteomics is the assignment of collected spectra to the amino acid sequences that gave rise to them. Proteins are digested using enzymes with known cleavage sites to produce peptides, which are then analyzed by mass spectrometry (MS). These datasets consist of two types of measurements: (1) MS$^1$ spectra survey a set of peptides present in the mass-spectrometer at a given moment and (2) MS$^2$ spectra originate from an attempt to isolate and fragment a single peptide ion species identified in the MS$^1$. Peptide spectrum matches (PSMs) are assigned using search algorithms [130, 134, 137] that match MS$^2$ spectra to peptides originating from a database of reference protein sequences. Typically, a target-decoy approach [153, 154] is used to estimate the false discovery rate (FDR), allowing users to produce a final list of identifications at a selected confidence level.

Generally, the proteomics community has aimed to simplify these search databases by using canonical sequence representatives of each protein in the human proteome. The rationale has been to reduce the peptide search space in order to avoid spurious matching and extensive peptide inference (i.e. peptides matching to more than one database entry) [225]. Difficulties in assigning spectra originate from a variety of factors including low abundance, non-peptide molecules, modified peptides or mixtures of co-fragmenting peptides. The larger the search-database the higher the likelihood of a spurious match [226].

However, one limitation of using reference sequence databases is that it is unclear how the cancer genome, with all its mutations, structural variations and epigenetic modifications manifests in a
cancer proteome. Onco-proteogenomics expands search databases with protein sequences not found in reference human proteomes, such as germline variations, variants commonly found in cohorts of tumour samples or sample-specific variants identified in genomic or transcriptomic analyses. Global MS-based proteomic strategies, in combination with genomics and transcriptomics could resolve this gap in knowledge [28-31, 185, 227-231] with the goal of improving the characterization of the variant peptides (i.e. peptidoforms) present in the sample.

Two types of databases are commonly used to incorporate protein variants into MS searches: community-based databases include variations previously observed while sample-specific databases include variants identified by DNA or RNA sequencing of that sample [232]. Each approach has advantages and disadvantages. Large databases like dbSNP [233], COSMIC [234] and UniProt [148] contain millions of protein variants, which can increase the likelihood of spurious database hits due to the increased database size. By contrast, sample-specific databases may be smaller, but are prone to false negatives resulting from variants missed in DNA- or RNA-sequencing for experimental or computational reasons [235, 236]. Intratumoural heterogeneity adds yet another potential source of missed variant protein detection [237].

State-of-the-art MS is now reaching the resolution and sensitivity to interrogate protein variations [118]. In parallel, the computational developments needed to combine proteomics with DNA and RNA sequencing in cancer samples are already underway [145, 187, 229, 232, 238-243]. Here, using proteomic, transcriptomic and genomic characterization of the NCI60 cell lines, we systematically investigate how the choice of proteogenomic databases affects PSM assignment. We present a strategy for onco-proteogenomics to assess the scope of variant peptides identified and their potential impact to cancer biology.

3.1 Methods

We conducted our study within the NCI60 cell line panel with extensive genomic [244], transcriptomic [245] and proteomic [120] data available. The proteomics data consist of both a ‘deep’ proteome derived from extensive fractionation of cell lysate by electrophoresis into 24 gel pieces (9 cell-lines) and a ‘shallow’ proteome, which was generated using 12 gel pieces (59 cell-lines).
3.1.1 Variant peptide database construction

The first step in variant protein identification was the generation of protein sequence databases containing the modified amino acid sequences (Figure 3.1a). Briefly, protein-level outputs from variant effect predictor [246] were parsed to proteins containing single amino acid variants, insertions, deletions, frameshifts, stop-loss mutations and fusions. Variant peptides were filtered against a canonical human proteome from UniProt (20,187 non-redundant proteins) to remove peptides that also mapped to this reference database. Variant sequences longer than six amino acids and containing up to two missed tryptic cleavages on either side of the mutated site were produced and added to the FASTA file.

We explored variant-peptide detection with regards to proteogenomic database size and content. Variant proteins were obtained from 5 different sources: dbSNP [233], COSMIC [234], UniProt [148], exome-seq [244] and RNA-seq [245]. Augmented search databases were created in 23 different ways derived from combinations and subsets of these databases (Figure 3.1b; Supplementary file 1). We defined community-based databases to include dbSNP, COSMIC and variants annotated in UniProt. Four sub-databases of COSMIC and dbSNP were made to include single nucleotide variants, indels, variants affecting genes in the COSMIC cancer gene census and frameshifts or stop losses or fusions. For sample-specific database searches, all 59 NCI60 cell-lines containing exome-seq data and 41 cell-lines containing RNA-seq data were used. Three further databases restricted to subsets of variants were generated for a total of four sample-specific databases per cell-line and per analyte type. We combined sample-specific and community-based databases in two different ways: we used a sample specific approach and a general approach where all RNA-seq and exome-seq datasets were merged. In total the RNA-seq cell-line data characterized 675 cell lines, which were also included separately in their own database, as was all the exome-seq data. A total of 473 different database combinations (Supplementary file 2; Figure 3.1b) were explored across all available cancer cell-lines.
Figure 3.1: Generation of proteogenomic databases.

(a) Scheme for the generation of databases suitable for MS-based detection of protein variants (b) Overview of databases generated.
3.1.1.1 COSMIC, dbSNP and exome sequencing databases

COSMIC (v70), dbSNP (v141) and processed exome sequencing [120] datasets were downloaded in VCF format and parsed using Variant Effect Predictor [246] (VEP.v.77) from Ensembl tools release (v77) using the GRCh37 genome reference model. VEP output files were further parsed to introduce mutations by retrieving the described reference sequences from the Ensembl proteome (GRCh37.75) and applying described substitutions, insertions and deletions using a series of Bioconductor R scripts (R:v3.1.0; stringer:v0.6.2; cleaver:v1.2.0; Biostrings:2.32.1; Rsamtools:v1.16.1; GenomicFeatures:v1.16.2). Peptides were generated from these mutated sequences allowing for up to two missed cleavage sites. Duplicate peptides were collapsed and headers identifying each mutation merged together.

3.1.1.2 COSMIC Fusions

Gene fusions were obtained from those manually curated from peer reviewed publications by COSMIC curators [234]. Fusions lacking inversions were parsed from COSMIC HGVS format by extracting appropriate transcripts (from the GRCh37.75 Ensembl genome model) and merging the corresponding sequences. Tryptic peptides spanning a three-frame translation over the fusion were added to the FASTA database for proteogenomic searching. Note: inversions and more complex fusions were not included in our analysis.

3.1.1.3 RNA-seq

RNA-seq datasets were obtained from the authors [245] as tab delimited files with each mutation fully characterized within a RefSeq protein. Each line in the file was parsed using in-house R scripts to generate mutated protein sequences. Tryptic peptides with up to two missed cleavages were generated overlapping the mutation site. RNA-seq in-frame fusions were made by merging nucleotide sequences for the 5’ and 3’ regions of the fusion. All tryptic peptides spanning the fusion crossover were added to the database.

3.1.1.4 UniProt Variants

The UniProt database was downloaded in XML format (December 2015) and variants described therein were parsed, and corresponding UniProt reference sequences modified.
3.1.2 Variant peptide detection

Using these databases, variant peptides were identified from the NCI60 cell lines using a proteogenomic pipeline implementing a split target-decoy approach [185], three search algorithms [130, 134, 137] and several additional filters (Figure 3.2a-b; Supplementary files 3-4). These filters (1) removed sequences mapping to the human proteome as described above (RefSeq, Ensembl and UniProt) (2) removed peptide-spectrum-matches that could also be based on chemical or post-translational modifications of reference peptide sequences and (3) removed protein variants with no alternative evidence for their expression.

3.1.2.1 Target decoy database construction

For each FASTA file above, sequences were combined with reviewed canonical Swiss-Prot (v.2014.12.09) protein sequences and each combined sequence was reversed. These original and reversed sequences were merged together to create proteogenomics FASTA databases used for peptide-spectrum match assignment.

3.1.2.2 Target decoy database searching

MS RAW files were converted to mzXML format using ReAdW (http://tools.proteomecenter.org/software.php), and searched against the proteogenomics FASTA databases with X!Tandem [130] (v.13.09.01.1), Comet [134] (v.2014.02 r2), and MS-GF+ [137] (v.0.9949). The following search parameters were used for all searches: carbamidomethylation of cysteine as a static modification, oxidation of methionine as a dynamic modification, a ±10 ppm precursor mass tolerance, a ±0.4 Dalton fragment mass tolerance for CID, and ±10.0 ppm fragment mass tolerance for HCD. All searches were performed on a twenty-two node cluster with twelve cores and 64 GB RAM on each node. Output files were converted into tab-delimited files that standardized outputs from all search algorithms (Supplementary files 5-8). The search results were then subjected to a series filtration steps, described next.
**Figure 3.2:** Proteogenomic search and filtering strategy

(a) Search schema used to identify variant peptides within proteomic datasets. Databases were searched using a split target-decoy strategy with both sequences to the augmented and reference proteins reversed. Three search algorithms (Tandem, COMET, and MS-GF+) were used and results were combined. (b) Schematic representation of PTM filtering strategy. All MS2 spectra identified by both our proteogenomics pipeline and identified as having mass-shifts to canonical peptides by MaxQuant were collected. If there was disagreement between the reference peptide altered by either pipeline, the PSM was rejected. Conservatively, we also rejected peptides if there was further disagreement regarding the site of modification.

3.1.2.3 Spectral-level FDR cut-off

We calculated spectral-level FDR cut-offs using a split target-decoy approach as initially proposed in [185]. FDR was calculated separately for variant peptides and UniProt peptide-spectrum matches (PSMs) using decoys generated from each database respectively, although MS data was searched against one merged FASTA file. In each case, PSMs with different mass-to-charge ratios were treated separately. PSMs with less than 1% spectral FDR were retained for subsequent analyses.

3.1.2.4 Filtering of resulting peptide lists

Applying a stringent spectral-level FDR filter does not guarantee that every PSM represents a correct identification, especially when single peptide identifications are involved, as is the case in proteogenomics. A number of scenarios could result in false positive identifications. The detected peptide may be an adjacent tryptic peptide not-overlapping the variant, which can arise from FASTA sequences containing missed tryptic cleavage sites included within the database. A variant peptide could be correctly assigned to the spectrum, but inadvertently also match to or be isobaric with a sequence of a different reference protein. A peptide could be erroneously matched to a spectrum, because the mass shift caused by a substitution happens to coincide with the mass shift associated with a post translational modification (PTM) on the same or possibly different peptide.
Finally, when searching large databases, false positive rates can be harder to control because there is a higher probability of matching a high scoring peptide from amongst the larger number of sequences available. We developed a series of post-search filters to mitigate these potential caveats. In the future these approaches could be further refined, using either synthetic spectral libraries or more sophisticated statistical approaches.

### 3.1.2.5 Filtration against reference proteomes

A filter was required to deal with scenarios where detected peptides inadvertently matched or could not be distinguished from peptides in the reference proteome. Detected peptides were matched against reference proteomes including that of Ensembl (GRCh37.75), RefSeq (release 68), and UniProtKB/Swiss-Prot sequences. Isobaric leucine and isoleucine residues, which cannot be distinguished, were considered identical during this filtration process.

### 3.1.2.6 Chemical modification filter

Mass shifts in MS\(^2\) spectra could also be attributed to post-translational modifications (chemical or enzymatic) within some reference peptide sequence. To deal with the possibility that PTMs were being misidentified as mutations in our pipeline, all cell line proteomic data was re-searched with MaxQuant [247] against the reviewed canonical UniProtKB FASTA database in “dependent-peptide” mode. Dependent peptides are assigned to MS\(^2\) as possible modifications to already identified peptides within a sample (i.e. modifications could be classic PTMs or amino acid substitutions). A schematic detailing how MaxQuant dependent peptides were used to remove potentially misidentified mutants is in Figure 3.2b, representing a conservative way of dealing with this potential issue (i.e. preference was given to the MaxQuant results and discordant peptides were removed from our results).

Potential post-translational or chemical modifications that matched to filtered variant PSMs by scan header were examined as to the position of the proposed post-translational modification. For this analysis, dependent peptides were filtered such that the probability that the modification occurred at a specific site (the positional probability) was greater than 0.8. This relaxed threshold was used to ensure that MS\(^2\) spectra for proteogenomic peptides that could be assigned as chemically modified sequences from a differing starting peptide sequence or site of modification were removed. It was our observation that dependent peptides modified at the same site as
proteogenomic peptides nearly always described the same mutation (i.e. the modification simply resulted in a different amino acid that was also called by the proteogenomics search). An example table showing peptides removed by our approach is shown for the exome-seq data (Supplementary file 9).

We also generated a list of variant peptides derived from the dependent-peptide search. These were filtered from all modifications proposed by MaxQuant as follows. First, we identified the amino-acid residue in the canonical peptide sequence that was modified. Next, we assigned single amino-acid variants based on MS$^1$ mass-shift that was consistent with an amino-acid change from that starting amino acid. We used a positional probability threshold of equal to or greater than 0.95 to stringently threshold these dependent peptides and found 1031 unique single amino-acid variants (Supplementary file 10).

3.1.2.7 Protein abundance filter

To further reduce potential false-positives, we elected to remove all proteogenomics PSMs for which there was no additional evidence of protein abundance (i.e. identification of peptides mapping to canonical sequences of the same protein). Therefore, each mutated peptide included in our final list has additional evidence of being expressed within the same cell line.

The final list of PSMs from different search algorithms were then grouped based on the source RAW file and Scan ID and categorized into the following tiers:

**Tier 1:** All peptides identified after the above filtration process.

**Tier 2:** Peptides identified by at least two algorithms.

**Tier 3:** Peptides identified by all three algorithms.

**Tier 4:** Peptides identified by all three algorithms with two spectra or more.

Detailed information of search output results and filtration steps for all NCI60 cell lines is available in Supplementary file 3 and Supplementary file 4.
3.2 Results

3.2.1 Characterizing reference and variant protein sequence databases

Our aim was to describe protein sequence variation beyond what is already included in reference proteomes. We therefore began by examining the background of the reference human proteomes, with the aim to understand the differences between them in tryptic peptide space. We examined four commonly used reference proteomes: (1) a database of 20,187 canonical protein sequences from UniProt (Swiss-Prot), (2) a second UniProt database with 88,717 proteins including isoforms (Swiss-Prot + Trembl), (3) the reference proteome derived from the Ensembl genome model using GRCh37 (v75) with 104,763 protein sequences (henceforth denoted Ensembl) and (4) the reference proteome derived from the RefSeq annotation model (release 68) consisting of 72,128 proteins. We in silico digested each of these reference human proteomes to produce a total of 2.95 million distinct tryptic peptides within the range of 6-35 amino acids in length; peptides that are most commonly detected by MS (Figure 3.3, Figure 3.4a). Of these, 70% (2,064,452) showed 100% sequence identity between all reference proteomes. The remaining 30% (887,991) of tryptic peptides constituted a large number of potentially detectable tryptic peptides missing in at least one reference proteome.

Given this disagreement between reference proteomes at the peptide level, we recommend that variant peptides eventually reported by proteogenomics should be filtered against the Ensembl, RefSeq and UniProt derived proteomes. To illustrate why this is necessary, after filtering against the smallest human reference proteome “canonical protein sequences” from UniProt, 7.3 million distinct tryptic peptides remained within our proteogenomic databases. However, of these 35,446 overlapped with the other three reference human proteomes (Figure 3.4b), with 43% derived from Ensembl and RefSeq, and 57% were present within Uniprot + isoforms (Swiss-prot + Trembl).

Variants present in reference proteomes were all in community-derived databases, though 12% were also found in sample-specific exome sequencing. These peptides cannot be disambiguated from the reference and should not be included in the final set of variant peptides detected. Improper filtering of putative variant peptides is a critical and often overlooked issue in their detection. When we compare our methodology to other efforts [248], we find that while we start with nearly the same peptides, we are more conservative and exclude many variant peptides from our final lists (Figure 3.5). However, our filtration steps are conservative, aimed to rigorously reduce false
positive identifications, especially in the context of sample specific databases. If follow-up validation strategies using synthetic peptides and targeted peptide quantifications are applied, less stringent filters may be appropriate.

Taking the prostate cancer cell line PC3 as an example, the total number of unique protein variants contained within the major database types we generated is summarized in Figure 3.6a and Supplementary file 2. Millions of unique and distinct tryptic peptides (7.3 million) derived from our databases represent the tryptic-space of proteome variation explored in this study. Each peptide was included within at least one database, but there was much redundancy between databases (Figure 3.6 b-d). While thousands of peptides (12,043) with sample-specific genomic evidence were included (Figure 3.6 b-c), the vast majority of peptides (6.84 million) were exclusive to community-based databases (Figure 3.6d).
Figure 3.3: Comparison of reference proteomes.

Venn diagram comparing the four reference databases used in this study. Proteins in each reference proteome were in silico digested and filtered in the length range 6-35 amino-acids. The Venn diagram portrays the overlaps of unique peptides originating from each reference database.
Figure 3.4: The detectable tryptic space of reference proteomes and the overlap to variant human proteins

(a) Distribution of 2.9 Million reference proteome tryptic peptides (length 6-35 amino-acids; including two possible trypsin missed cleavages) derived from four commonly used reference proteomes. Counts are represented using a log_{10} scale. (b) Distribution of the 35,445 variant peptides that are also contained within at least one reference proteome. Y-axis covariate depicts the source of the variant. Color gradient indicates the percentage of the 35,446 variants that overlap with each reference using a log_{10} scale.
Figure 3.5: Comparison to other studies.

To illustrate the importance of peptide filters, we compared our results to a previous study [248]. In their study, both a cell-line specific database search and a database combining all exome sequencing data were used. The figure reports the number of peptides identified by [248] that remain after the various filtration approaches in our pipeline.
Figure 3.6: Overall database sizes and redundancies using prostate cancer cell-line PC3 as an example.

(a) Numbers of protein variants in the nine major database variants used to search PC-3 proteomics data. Counts are in a log10 scale. (b) Total number of exome-seq derived variant peptides and their membership in other databases. Counts are in a log10 scale. (c) Total number of RNA-seq derived variant peptides and their membership in other databases. Counts are in a log10 scale. (d) Total number of peptides derived from various community-based databases and their redundancy with each other. Counts are in a log10 scale.
3.2.2 Scope of variant peptides identified

In total, 13,302 unique variant peptides were identified within the deep NCI60 proteomic dataset (Supplementary file 5; Supplementary file 6). To understand how these peptides differed in terms of confidence of identification, we quantified the evidence for peptide identification using four tiers of stringency (Figure 3.7a). Tier 1 peptides were assigned by the union of the three search algorithms (13,302 peptides). Tier 2 and tier 3 peptides were identified by either two (3,071 peptides) or three algorithms (1,610 peptides) and tier 4 peptides were identified by three algorithms and more than one PSM (836 peptides). These overall trends were also representative for one cell line, as shown for PC3 (Figure 3.7b). The peptides identified in PC3 came from a diversity of databases and would often be present in smaller database searches as well as larger ones (Figure 3.7c). We further evaluated all PSMs to check for biases in hydrophobicity, charge and length (Figure 3.8). We found that variant peptides identified through our pipeline, tended to be larger and of higher charge than those identified using standard proteomic searches (see Discussion). The overall numbers of PSMs, unique peptides and mutations detected within the 9 deep proteomes has also been summarized (Figure 3.9a).

We focused on community-derived databases or sample-specific database searches (Figure 3.1b). Fewer peptides (272) were identified with genomic evidence than from the tryptic space of community-derived variants (11,761; Figure 3.9b). The proportion of peptides with genomic evidence increased from tier 1 to tier 4. This mild improvement for peptides with genomic evidence came at the cost of proteogenomic peptide identification (Figure 3.9a).

We evaluated how peptides with and without dataset-specific genomic evidence differed in their score distributions (Figure 3.9 c-d). We focused on those peptides that were derived from community-based databases, some of which also had genomic evidence. For each search, peptides were percentile ranked, with a percentile rank of 1% indicating a peptide in the top 1% of peptides in that search. At tier 1 there was only a slight bias showing better PSM scores if the peptide had sample-specific genomic evidence, supporting the validity of these community-based peptide identifications (Figure 3.9 c-d). MS-GF+ consistently identified more peptides than COMET and Tandem. The fraction of peptides with population variation evidence and the fraction of peptides with genomic evidence initially identified in tier 1 decreased relatively linearly with tier (Figure 3.9e). A 12% improvement in peptide median score occurred between tier 1 and tier 2 (Figure 3.9 c,f). This compared to a ~6% improvement from tier 2 to tier 3, indicating the benefit of
incorporating additional algorithms rapidly is exhausted. Similar trends for these score distributions were observed for a standard UniProt search (Figure 3.9c,d). We recommend using tier 2 as a balance between sensitivity and specificity, although we suggest that all proteogenomics PSMs should be closely examined (possibly using synthetic peptides) before subsequent analysis.

### 3.2.3 The relevance of proteogenomic peptides

Any proteogenomic pipeline must detect peptides in an unbiased manner across the entire genome as well as variations in relevant cancer genes and pathways. Peptide variants identified within the NCI60 dataset were broadly distributed across the genome (Figure 3.10), but clearly the detected variants are just a fraction of those theoretically detectable within the datasets searched. In total, we found 4,771 unique protein variations mapping to 2,200 genes at tier 2 (Supplementary file 11) across both the deep (1,511 HGNC gene ids) and the shallow (1,469 HGNC gene ids) proteomes. The median number of mutations per gene was just 1 in both proteome datasets. However, there were a few genes where an excess of variants were identified across cell-lines. 

**AHNAK**, a large 700 kDa structural scaffold nucleoprotein with known roles in cell-migration and metastasis topped the list with 91 variants identified across the 9 deep proteomes. In total, 211 COSMIC cancer gene census genes harbored detected variants, demonstrating the potential of proteogenomics for variant detection in cancer. These genes tended to be highly expressed within the 9 deep proteomes, as estimated using iBAQ scores from a standard UniProt search (Figure 3.11a).

Variants identified were assessed by the drug gene interaction database [249, 250] in order to identify those variants that could be targetable by a drug or affect targetable pathways. We tested whether the genes associated with variant peptides identified at tier 2 (3,071 unique peptides) were enriched in specific druggable gene categories when compared to equally sized random subsamples of unique peptides identified in a standard UniProt search against the 9 deep proteomes. As a null distribution, we took 100,000 subsamples of 3,071 peptides from a UniProt search and binned them into categories within the drug gene interaction database. Using this methodology, several druggable gene categories were statistically enriched (p<0.01) in variant peptide detections at tier 2 (Figure 3.11b). Statistically enriched categories included variants from various tumour suppressors, cell-surface proteins, proteins involved in drug resistance and proteins involved in transcription factor binding.
**Figure 3.7:** Detection of variant proteins within the 9 deep proteomes with focus on prostate cancer cell-line PC3.

(a) Numbers of unique variant peptides identified in tiers 1-4 using mass-spectrometry data from the 9 deep proteomes. (b) Unique variant peptides identified within the prostate cancer cell-line PC3 across tiers 1-4 (log_{10} scale). (c) Heatmaps depicting the percent contribution of each database towards the total number of peptides identified for that tier in PC3. The number of peptides overlapping each database pair is provided as well. Color scale is in log_{10}. (d) Total number of spectra, peptides and unique mutations identified by tier.
Figure 3.8: Biophysical properties of detected variant peptides.

Top row: Boxplots show the calculated Kyte-Doolittle hydrophobicity index of peptides and peptide variants identified by a PSM in the NCI60 ‘deep proteome’ dataset (p-value from Wilcoxon sum-rank test). Middle row: Barplots show the charge of peptides and peptide variants identified by a PSM in the NCI60 ‘deep proteome’ dataset. Variant peptides were observed to contain more charge +3 peptides than non-variant peptides which contained more charge +2 peptides. Bottom row: Boxplots show the length of peptides and peptide variants identified by a PSM in the NCI60 ‘deep-proteome dataset’ (p-value from chi-square test). Variant peptides were observed to contain longer peptides when compared to those peptides identified in a standard search against canonical UniProt.
**Figure 3.9:** Tier 2 peptides offer significant peptide detections while mitigating potential false positives.

(a) Total number of spectra, peptides and unique mutations identified by tier. (b) Summary of peptides identified within the 9 deep proteomes within sample-specific databases or within community-based databases (tiers 1-4). (c) Percentile score distribution summary by algorithm and tier. X-axis ranges from high scoring peptides (0’th percentile) to lower scoring peptides (100’th percentile). (d) A similar figure using original e-value scores is depicted. The distribution of peptide scores from a search against a standard UniProt database is shown in black. (e) Increasing the stringency of identifying a peptide influences the percentage of peptides present in community-based databases between tier 1 and tier 2 more than moving to subsequent tiers. (f) When compared, tier 2 peptides tend to be higher ranked by 12% than tier 1 peptides, this improvement in peptide rank drops off quickly from tier 2 to tier 3 (4%) and tier 3 to tier 4 (1%).

We mapped variant peptides back onto the canonical reference sequence for the oncogene beta-catenin (CTNNB1) (**Figure 3.12a**), revealing several mutations in both the deep and shallow proteomes in cell-lines derived from different cancers. While many variants were identified, they were only a small fraction of the possible variants for CTNNB1 (**Figure 3.12a bar plots**). As an example, we refer to a tier 2 PSM with both exome-seq and RNA-seq evidence for which we have identified a peptide sequence (**Figure 3.12b**).

We identified 111 fusion proteins in the 9 deep proteomes and 508 fusion proteins in the 59 shallow proteomes (**Supplementary files 7-8**). The gene encoding the RNA binding protein FUS is located at a common site of chromosomal translocations in human low grade fibromyxoid sarcomas and frequently forms chimeric fusions with one of several different genes [251]. We identified four different FUS-CREB3L2 fusions across seven cell lines, from a total of 101 FUS-CREB3L2 fusions present in COSMIC (**Figure 3.13**). These fusions were identified independently of RNA-seq, for which fusion calls from sample-specific transcriptomics (median three per cell line) were rare [245]. Based on our sample-specific RNA-seq searches, only three fusions were identified across the 9 deep proteomes and 33 across the 59 shallow proteomes.
Figure 3.10: Genomic distribution of Tier 2 variant peptides across the 59 shallow and 9 deep proteome datasets.

Genome coverage of potentially detectable proteogenomic peptides (6-35 amino acids) within the generated search databases (bottom). Variant proteins identified at tier 2 within 59 shallow and 9 deep proteomes have been summarized in black and gray, respectively (top). Black dots correspond to the locations of COSMIC cancer census genes and orange dots indicate those detected at tier 2.
Figure 3.11: On the expression and druggability of genes containing detected variant peptides.

(a) Variants identified for genes in the COSMIC cancer gene census tend to be highly expressed in the same cell-line. Protein abundances for the 9-deep proteomes (log_{10} iBAQ) were calculated from a MaxQuant search against a standard UniProt database and ranked from most abundant to least abundant. Proteins expressed in each cell line for which no variant was detected have been colored gray. Variant peptides identified but not in the COSMIC cancer gene census are colored in blue. Variant peptides identified in the cancer gene census are colored in red. Density plots are diagrammed to the right of the plot. (b) Variants identified were assessed by the drug gene interaction [250] database to identify variants that might potentially be targetable or affect related pathways. Counts relate to the number of variant peptides identified in each category for tier 2 peptides. Only categories significantly enriched at p<0.01 are depicted.
Figure 3.12: Focus on variants detected for *CTNNB1* across cell lines.

(a) Variant peptides detected for *CTNNB1*. Mutation locations have been depicted in orange. Identification of reference peptides for the same protein are shown in blue, with an alignment describing the peptides detected. Bar plots illustrate the variants that were present in genomics for this gene (top) and all mutations present in community-based databases (bottom). (b) A tier 3 peptide identified for *CTNNB1* showing clear coverage of y and b ions.
**Figure 3.13:** MS2 spectra for *FUS-CREB3L2* fusions.

*FUS-CREB3L2* fusions were repeatedly identified from searches including the COSMIC database. Here we present schematics of the fusions identified with lower case letters at the 5' end of the fusion site. In each case, peptide fragments are identified across the fusion.

### 3.3 Discussion

Proteogenomic approaches promise the personalized detection of genomic aberrations within protein samples and may represent an important untapped area in cancer biomarker discovery. We explored the limits of variant peptide detection using MS-based proteogenomics strategies. In general, there are three interrelated aspects of PSM assignment at play: (1) the capacity to separate peptides in chromatography and mass-to-charge space, (2) the sensitivity of the mass-spectrometer itself and (3) the overall sequence coverage of the tryptic peptidome. Proteomics search-algorithms must identify the amino-acid sequence with the highest likelihood to have produced a particular MS² spectrum, carefully taking these challenges into account. Algorithms must screen protein sequence databases and identify a set of putative peptides of the same mass (within error) of the peak in the MS¹ spectrum associated with the MS² in question. In variant peptide identification, as database size increases, the algorithm must choose from an increasingly large pool of potential peptides, which must be assigned to spectra that often may originate from more than one peptide molecule.

Interestingly, variant peptides identified through our pipeline, tended to be larger and of higher charge than those identified using standard proteomic searches. While the exact reason for this observation is currently not known, we speculate that for larger databases a better search score is required to pass a predefined 1% FDR (based on a target-decoy approach). Larger peptides, which in general are associated with a higher score could hence be favored in this process. However, as a caveat, longer peptides tend to have slightly lower overall y and b ion coverage, which could also lead to potential false positives.

We have developed a series of recommendations to serve as guidelines to better characterize variant proteoforms within cancer proteomics datasets using custom sequence databases and a target-decoy approach. (1) We recommend variant peptides be identified using more than one search algorithm using a split target-decoy approach [185]. (2) The use of several filters to reduce
sources of possible false positive identification not accounted for by commonly used proteomics approaches. This includes filters that remove variant peptides detected within standard reference proteomes, or that could be accounted for by post-translational modifications of a given peptide sequence. (3) We also recommend that identified protein variants be supported with additional evidence for the expression of their source protein.

Ultimately, generation of custom protein sequence databases and filtering of resulting data to balance the sensitivity and specificity of peptide detection will depend on the investigator and goal of the project. For example, it may be appropriate when using databases with sample-specific genomic evidence to keep peptides that match to reference proteomes for further investigation. Conversely, in the absence of sample specific data, variant peptides could be identified using large publically available databases, although with a higher risk of false positive identifications. As a final recommendation, we suggest that promising candidates be visually inspected and preferentially compared to spectra generated by synthetic peptides. This will provide additional validation and the possibility for the development of targeted proteomics assays.

Our study illustrates the need for further improvements in proteogenomics pipelines. With our stringent search criteria, we identified 4,771 protein variants corresponding to somatic and germline deviations from reference proteomes in 2,200 genes amongst the NCI60 cell-line proteomes. This is despite the tens of thousands of identifiable peptide variants with sample-specific genomic evidence present in our search databases. The detection of protein variants is particularly difficult as each can only be detected by six unique tryptic peptides after accounting for up to two missed cleavages. Proteins may be lost during protein extraction and peptide biases may be introduced during digestion, detection and PSM assignment. These technical challenges, as others have noted [241], lead to a lack of sequence coverage amongst all proteins identified and result in a lack of sensitivity for variant peptide identification. Compounding on a lack of sensitivity is the potential for false identification. As has been shown for post-translational modifications, it is plausible that the use of alternative proteases could increase the likelihood of detecting specific mutations by shotgun proteomics [252]. There are other strategies for detecting variants from mass-spectrometry datasets. The proteogenomic approach can easily be integrated with semi-supervised methods that search for variants of reference proteins present in standard search databases. The dependent peptide searches we used to filter out potential post-translational modifications allow for a comparison to these approaches. We collected 1,031 high confidence
single-amino-acid-variant dependent peptides (positional probability > 0.95) (Supplementary file 11). In Total, 97 variant peptides or 10.3% of dependent peptide variants overlapped with proteogenomic variants, highlighting the potential for these methodologies to expand our capacity for variant protein detection. Other semi-supervised or ‘open search algorithms’ such as the recently released MSFragger [140] and spectral network inference [155] could also be used as additional strategies for the parallel identification of post-translational modifications or proteoform variants. While beyond the scope of the current manuscript, head-to-head comparisons of open search algorithms, custom database proteogenomics searches and spectral libraries using massive synthetic peptide libraries [253] are now possible and will likely lead to the refinement of current proteogenomic strategies.

3.4 Conclusions

Proteogenomics can identify germline and somatic mutations within important cancer genes (Figure 3.11). While the underlying technology improves, the proteogenomics community can now focus on integrating alternative strategies for detecting protein variants. The proteogenomic approach described here can be integrated with semi-supervised methods that search for variants of canonical proteins and de novo sequencing (i.e. PEAKS [143]) based methodologies that could identify variants missed by genomics. Added sensitivity could be achieved by constructing spectral libraries from synthetic peptides derived from genomic evidence, which could help with the development of more statistically-refined proteogenomics pipelines.
Chapter 4

Discussion: Future prospects for the detection of protein variants
4 Discussion

This thesis has focused on one aspect of cancer proteotype elucidation using mass spectrometry: the characterization of protein variants using evidence from sample-specific or community annotated genomic variants. The onco-proteogenomic strategy proposed in Chapter 3 is very stringent. However, areas for improvement still remain. The pipeline does not directly take into account ion coverage when calling variant peptides, leaving variant calls with low peptide sequence coverage as a potential source of false positives. Further, the pipeline does not consider the possibility the sequences we are searching for are still missing from the database, meaning many variant peptides may still be missing. In this Chapter, I discuss these opportunities and potential proteogenomic strategies to address them.

4.1 Peptide sequence coverage as a metric to further constrain proteogenomic search results.

As mentioned in Chapter 1, the first in-depth tissue based maps of the human proteome were released in 2014 [62, 79] and covered 84% of human protein coding genes (~20,300 genes total) at a median sequence coverage of just 28%. This low sequence coverage is also reflected within the NCI-60 proteome dataset where the deep proteomes combined identify 80% of human protein coding genes at a median sequence coverage of just 15% [120]. However, peptides themselves are not identified at 100% sequence coverage, which is a potential source of false positive variant calls from mass spectrometry data.

To understand sequence coverage at the peptide level, consider that the fragment ion spectrum for a peptide consists of several fragment ion series of the form \((I_1, I_2, I_3, ..., I_n)\), where \(n\) is the length of the precursor peptide with sequence \(P = p_1p_2p_3...p_n\). Each fragment ion in a series \((I_i)\) can at best provide evidence for a single amino-acid in the sequence. Each fragment ion constrains the identity of amino-acids preceding that residue in the series by its mass. If a single fragment ion \((I_m; m \neq 1)\) in a series is detected, only the overall amino-acid content within the sequence of the fragment can be predicted. However, upon addition of the previous ion in the series \((I_{m-1}I_m)\), the identity of the amino-acid at the end of fragment \(p_m\) can be elucidated, and uncertainty remains with the amino acids \(p_1p_2p_3...p_{m-1}\). If instead an ion is measured far earlier in the series \((I_kI_m; k < m - 1)\), then direct sequence information is still missing, only the amino acid content up to residue \(k\) and then up to residue \(m\) can be inferred. I define fragment ion sequence coverage
as the percentage of residues for which there is ‘direct’ evidence through a fragment ion in any series. On median searches against a UniProt database of canonical human proteins against the nine deep proteomes showed that the median peptide was 12 amino-acids long and had fragment ion sequence coverage for 80% of residues in the protein sequence [120]. With half of identified peptides having a sequence coverage lower than 80%, it is clear that another filter may be needed in the pipeline.

The caveat in Chapter 3 that each peptide spectrum match be inspected manually was added because of the observation that sequence coverage over the variant site was not always complete. A sequence coverage filter would reduce dependency on manual inspection. After the introduction of this filter, further analysis would be needed to understand whether the tier system developed in Chapter 3 would be necessary for variant protein detection.

4.2 Hybrid approaches for variant peptide identification

4.2.1 Proteogenomic semi-supervised search strategies

There are other strategies for detecting variants in protein sequences that have not been considered here. For example, the proteogenomic approach can be integrated with semi-supervised methods that search for variants of reference proteins present in standard search databases. Such ‘hybrid’ approaches for variant peptide detection would formally incorporate onco-proteogenomics within the larger problem of proteotype and proteoform characterization. Hybrid approaches would combine complementary genomic evidence in variant peptide detection with identifications of variant sequences missing in the augmented database. This could be accomplished with MSFragger [140] and MaxQuant ‘dependent-peptides’ [141] and possibly result in the parallel identification of post-translational modifications or proteoform variants (Figure 4.1).

Regardless of the algorithm used, the approach involves standardizing the output of each algorithm such that the mass-shift of the modification and the approximate position of the modification in the database sequence (positional probabilities) can be identified. Next, MSFragger and MaxQuant dependent peptides would be separated into two groups based on whether a mass-shift could be attributable to a post-translational modification. Each group would then be filtered depending on whether a residue with high positional probability could sensibly undergo the post-translational modification or mutation. Those variants that pass this filter would move forward to be integrated
with variant proteogenomic peptides at Tier 1 from our pipeline. One final filter would ensure that all called variant peptides were obtained with high fragment ion sequence coverage, before producing a finalized list of variant peptide identifications. These variant peptides would now include mutations and post-translational modifications.

In our view, the complementary evidence provided by sample-specific proteogenomic identifications provide the highest level of evidence for variant protein identification. However, it will be interesting to see how many variant peptides are identified using these semi-supervised approaches. The identification of many germline and somatic mutations within important cancer genes using semi-supervised strategies could imply either a high false negative rate from genomics or a high false positive rate in proteomics. Either way, these results should be used to modify the way variants are called within both genomics and proteomics.

4.2.2 Other hybrid strategies

Another iterative improvement to these variant calling algorithms would be the inclusion of unsupervised de-novo peptide sequencing. Such algorithms have gained popularity in areas like antibody sequencing. As in 4.2.1, peptide fragment ion sequence coverage is not always high for these identifications. While denovo sequencing algorithms may claim complete sequence identity, they really identify the family of related sequences constrained by the fragment ion series in the MS² spectrum. Only PSMs with very high sequence coverage should be trusted as the alternative sequences for a given spectrum may overlap with reference protein sequences and not be true variants. Incorporating These denovo sequencing algorithms into variant protein detection should therefore include a filter that removes PSMs where reference sequences or their post-translational modifications could also account for the PSM.

Spectral library searching offers another means of identifying commonly occurring cancer protein variants. Experimental MS² spectra contain reproducible fragmentation patterns that may go beyond current in silico models of peptide fragmentation. Therefore, cross correlation between experimental spectra and libraries of recorded peptide-spectrum matches (spectral libraries) can offer strong lines of evidence for peptide assignment. For this reason, the construction of a spectral library containing commonly occurring variations to proteins in cancer would be beneficial. Such a library could be constructed for sequence variants from the COSMIC cancer gene census data.
Figure 4.1: Hybrid approaches for variant protein detection. A strategy for combining the proteogenomic identification algorithm put forward by this thesis and the open search algorithms implemented in MaxQuant and MS-Fragger.
Commonly occurring mutations in onco-genes, as well as peptides containing in-frame fusions between exons from different genes, could dramatically increase the ability to detect protein variants.

4.2.3 Top-down and middle-down proteomics as alternative strategies for improving proteoform characterization

The length of the peptide sequenced by mass spectrometry, is an important aspect to consider in global proteoform characterization. Using trypsin as the digestion enzyme produced a median peptide length of 12 amino-acids amongst the NCI-60 deep proteomes and typical peptide lengths are 6-35 amino-acids (< 2 kDa) in most bottom up experiments. These short reads of the protein sequence lead to a loss of combinatorial variant information (i.e. combinations of mutations and PTMS), and difficulties assigning peptides to individual proteins as opposed to protein groups. Top-down proteomics, which attempts to characterize intact proteins overcomes these issues but has other complications. The success of the technique declines in the high mass regions and suffers from sensitivity issues if the number of proteoforms of a specific protein are too large (which would result in too many low abundance ion species) [254].

Middle-down proteomics might combine the advantages of both strategies and improve the elucidation of protein variants [255]. In middle-down proteomics, proteins are digested with enzymes that cleave less frequently yielding peptides of larger length. OmpT has been proposed as an enzyme for use in middle down proteomics [255]. This enzyme primarily cleaves between dibasic sites, rather than single basic sites as with trypsin. These more stringent criteria produce peptides larger than 3 kDa. Practically, all cleavage products smaller than 15 kDa are used for middle-down proteomics. Using this strategy, 3,697 peptides corresponding to 1,038 different proteins were identified from Hela Cell whole cell lysate [255]. These low numbers of identifications in middle down and top-down proteomics will improve with the use of more powerful magnets that increase the resolution of these complex mixtures in the m/z dimension (i.e. [254, 256]).
4.2.4 Alternative strategies for global proteoform characterization

The importance of globally identifying and quantifying aberrant protein variants in cancer and disease has prompted the biophysics community to seek out alternatives for protein sequencing. Some methods involve parallelizing Edman degradation. Examples include fluorescently labelling amino-acids before sequencing peptides immobilized on a membrane based on changes in fluorescence during Edman-degradation [257]. Along the same line fluorescently labelled antibodies have been proposed to be used to recognize the N-terminal amino-acid of a peptide after each step of Edman-degradation [258]. The challenges affecting these strategies are whether fluorescence based fingerprints generated by the Edman degradation process will be sufficient to accurately identify peptides and sequences. Other next-generation protein sequencing strategies involve extending the success of nanopore sequencing technologies from DNA to protein sequences [259, 260]. The challenges in this field revolve around decreasing the transport time of proteins through the nanopore such that the signal blockade can be deconvoluted into an amino-acid sequence. The search space is large, and thus far signal blockades have only been resolved to overall amino-acid sizes. Regardless, these spectra may contain enough information about amino-acid size and order to uniquely identify many proteins.

4.2.5 Concluding remarks

Mass spectrometry is currently the preferred method for global proteome characterization and the field has spawned numerous Nobel laureates. Innovations in mass-spectrometry have continually improved the total number of proteins detected to the point that proteoform elucidation from global proteomic datasets is now within reach. The work in this thesis, presents data demonstrating that it is now possible to elucidate thousands of protein variants within mass-spectrometry datasets. As the technologies continue to improve, the methods developed herein will be of importance in helping to guide the detection of these protein variants.
References


51. Garber K: **In a major shift, cancer drugs go ‘tissue-agnostic’.** American Association for the Advancement of Science; 2017.


60. Verhaak RG, Cooper LA, Salama SS, Aldape K, Yung WA, Brat DJ: Comprehensive and integrative genomic characterization of diffuse lower grade gliomas. AACR; 2014.


71. Li JJ, Bickel PJ, Biggin MD: **System wide analyses have underestimated protein abundances and the importance of transcription in mammals.** *PeerJ* 2014, **2**:e270.

72. Li JJ, Biggin MD: **Statistics requantitates the central dogma.** *Science* 2015, **347**:1066-1067.


78. Costello J, Stolovitzky G: **Seeking the wisdom of crowds through challenge-based competitions in biomedical research.** *Training* 2013, **35**:18.


112. Michalski A, Cox J, Mann M: More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J Proteome Res 2011, 10:1785-1793.


