Computational Approaches to Immunopeptidomics

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Dipl.-Inform. Mathias Walzer
aus Villingen-Schwenningen

Tübingen
2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	10.12.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Nico Pfeifer

Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

Computational Approaches to Immunopeptidomics

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Abstract

Life science has a wide range of available technologies for gathering and analysing qualitative and quantitative insights into biological systems. These can enable the development of new clinical applications with novel combined approaches, for example to personalised cancer immunotherapy. Among these technologies, mass spectrometry is a complex but versatile analysis method for proteomics and beyond. It is also a prolific source of experimental data in these fields. Improved instrumentation, technological advances, and high-throughput experiments result in a substantial growth in data volume calling for automation in analysis and quality control. Combination with other technologies is essential for advances in clinical applications, like the development of personalised cancer immunotherapies, inherently multidisciplinary. Here we describe how integration of data generation and analysis tools can have synergistic effects, unlock novel analysis designs, and overall create a more comprehensive picture of the underlying biology. First, we describe the development of standard file formats for proteomics and their role in data integration. Many experimental proteomics techniques share core aspects that are reflected in their data and the analysis steps necessary for successful interpretation. The formats cover mass spectrometry-based proteomics identification and quantification data (mzIdentML, mzTab), limited support for small molecules (mzQuantML, mzTab), and QC data from both acquisition and analysis (qcML). In the second part of this work, we describe the integration of analysis tool frameworks (OpenMS, FRED2) and their accompanying analysis tools in workflow orchestration solutions (KNIME). Compatible tools (with standard format in- and output) allow us to create complete data analysis workflows with data from multiple domains (here genomics, transcriptomics, proteomics, and HLA peptidomics) and explore the benefits of automated analysis with a closer look at the utility of add-on workflows for quality control in parallel to a main analysis. The final part of the thesis concludes this work with the application of a combined workflow in the research and development of personalised immunotherapies against cancer. We show how our workflow is comparably sensitive in detecting neoepitopes applied to a melanoma dataset with previously published detected neoepitopes. We also explore how cancer type can influence the prospect of detecting actionable immunotherapy targets with the example dataset from hepatocellular carcinoma patients.

Kurzfassung

Die Lebenswissenschaften verfügen über eine breite Palette von Technologien zur Erfassung qualitativer und quantitativer Erkenntnisse über biologische Systeme. Die Massenspektrometrie ist ein komplexes, aber vielseitiges Analysesystem für die Anwendung in der Proteomik und darüber hinaus. Verbesserte Instrumentierung, technologische Fortschritte und Hochdurchsatzexperimente führen zu einem erheblichen Wachstum des Datenvolumens, das eine Automatisierung bei der Analyse und Qualitätskontrolle erfordert. Die kombinierte Analyse mit Daten anderer Technologien ist für Fortschritte in der klinischen Anwendung, von sich aus multidisziplinär, unerlässlich. Wir zeigen hier, wie die koordinierte Integration von Daten und Analysewerkzeuge synergistische Effekte für die Forschungsfelder der Lebenswissenschaften haben kann, neuartige Analyseansätze eröffnet und ein umfassenderes Bild biologische Systeme ermöglicht. Zunächst beschreiben wir die Entwicklung standardisierter Dateiformate für die Proteomik und ihre Rolle in der Datenintegration. Viele experimentelle Proteomik-Techniken weisen gemeinsame Kernaspekte auf, die sich in den Messdaten und erforderlichen Analyseschritten widerspiegeln. Die Formate umfassen massenspektrometriebasierte Proteomik-Identifikations- und Quantifizierungsdaten (mzIdentML, mzTab), begrenzte Unterstützung für “small molecules” (mzQuantML, mzTab) sowie Qualitätskontrolldaten von den Messungen selbst als auch von den Teilergebnissen der Analyse (qcML). Im zweiten Teil dieser Arbeit beschreiben wir die Integration von Analyse-Frameworks (OpenMS, FRED2) in Softwarelösungen zur automatisierten Analysyeorchestrierung (z.B. KNIME). Wir untersuchen, wie kompatible Analysewerkzeuge (mit Standardformaten für die Daten Ein- und Ausgabe) es uns ermöglichen, komplette Analyseabläufe für komibinierte Daten aus der Genomik, Transkriptomik, Proteomik, und Immunoinformatik flexibel zu erstellen. Dabei werfen wir einen genaueren Blick auf die Möglichkeit, mit einem “add-on workflow” zur Hauptanalyse die Zuverlässigkeit der Messungen und Analyseresultaten einzuschätzen. Der letzte Teil schließt die Arbeit mit der Anwendung eines kombinierten Analyseablaufs in der Entwicklungsforschung personalisierter Immuntherapien gegen Krebs ab. Wir zeigen, wie unser Workflow vergleichsweise sensitiv bei der Erkennung von Neoepitopen ist, welche essentiell für gezielte und personalisierte Therapien sind. Außerdem untersuchen wir, wie, anhand der Analyseresultate im Vergleich, der Krebstyp die Aussicht auf Erkennung behandlungsrelevanter Immuntherapiepeptide beeinflussen kann.

Acknowledgments

Preface

With a long view to the past, it might appear that an ungainly large proportion of day-to-day work in bioinformatics was comprised of writing (custom) data parsers and gluing together data for analysis, even when working in only one of the diverse fields of omics. With the ongoing fusion of fields to gain a sufficiently comprehensive insight into the true workings of the biology and biochemistry involved in the life sciences, this artisanal aspect of the bioinformagician's work will hopefully belong to the past sooner, rather than later. A big part in that play standardised formats, as they afford the analyst with a comprehensive view of the data and its connections, and do so in a consistent way. This saves development time through commonality and run time by enabling automation. While the topic of standard format development and implementation might be boring at best to most audiences, I think it is necessary to be at least informed about the goals and cornerstones of design and implementation. After all, these are the keystones in making seamless tool integration, and a structured archival of our research data for truly automated, high-throughput workflows possible. Only with firmly integrated methods of data acquisition and analysis can we profit from the new wealth of data, faster collected and recorded in more detailed than ever. Like so, the task of knowledge integration and data mining with new and exciting methods like deep learning is about the discovery of new insights and not about massaging data. This work contains more boring details about how to alleviate the drudging but necessary meticulousness of data collecting that is mandatory to form an understanding from complex systems bordering on the chaotic, like proteomics and the human immune system. This is a story how to bring together the pieces, not only for one particular goal (the development of personalised immunotherapies), but align closer all involved steps for better data analysis in general.

I would like to thank all my friends and (present and former) colleagues who helped me with suggestions, motivation, and a constructive, critical view on this work. And of course all my fellow developers in the OpenMS, PSI, FRED2, and WorkflowConversion teams; it wouldn't have worked out without you!
In memory of A. Bertsch.

In accordance with the standard scientific protocol, I will use the personal pronoun *we* to indicate the reader and the writer or my scientific collaborators and myself.

J

Contents

1 Introduction **1**
1.1 Motivation 1
1.2 Structure of this Thesis 3

2 Background **5**
2.1 Protein Biochemistry 5
2.1.1 The Central Dogma: From DNA to RNA to Proteins 6
2.2 Immunobiology 10
2.2.1 The HLA Molecule and Its Ligands 12
2.2.2 The Immune System in Health and Disease (Cancer) 15
2.2.3 Isolation of Naturally Processed HLA Peptides 17
2.3 Mass Spectrometry-based Proteomics and Immunopeptidomics 18
2.3.1 Sample complexity and Liquid Chromatography 18
2.3.2 Mass Spectrometry 20
2.4 Computational Mass Spectrometry, Immunoinformatics, and Integrative Bioinformatics 30
2.4.1 MS Data 30
2.4.2 Quantification 30
2.4.3 Identification 31
2.4.4 Analysis Tools, Frameworks, and Workflows 34

3 Data Integration for Automated Workflows **39**
3.1 Introduction 40
3.2 The mzIdentML Format for MS Identifications 42
3.2.1 Methods 42
3.2.2 Results 47
3.2.3 Discussion 49
3.3 The mzQuantML Format for MS Quantifications 50
3.3.1 Methods 51
3.3.2 Results 59

3.3.3 Discussion . . . 62
3.4 The mzTab Format for MS Identifications and Quantifications . . . 63
3.4.1 Methods . . . 63
3.4.2 Results . . . 71
3.4.3 Discussion . . . 74

4 Reporting Quality Control in Mass Spectrometry 77
4.1 Introduction . . . 77
4.2 Methods . . . 79
4.3 Results . . . 81

5 A Framework for Immunopeptidomics 87
5.1 Introduction . . . 87
5.2 Methods . . . 90
5.3 Results . . . 95
5.4 Discussion . . . 96

6 Automated Workflows for Quality Control and Immunopeptidomics 99
6.1 Introduction . . . 99
6.2 Methods . . . 102
6.2.1 OpenMS in KNIME . . . 102
6.2.2 FRED2 in KNIME: ImmunoNodes . . . 104
6.2.3 Automated QC for HLA-ligandomics workflows . . . 105
6.3 Results . . . 106
6.4 Discussion . . . 117

7 Applications - Personalised Immunopeptidomics 121
7.1 Introduction . . . 121
7.2 Methods . . . 123
7.3 Materials and Data . . . 124
7.4 Results . . . 126
7.4.1 A Multi-Omics Approach to Detect Mutated HLA Ligands in HCC 126
7.4.2 Discovery of Mutation-Derived HLA Ligands on Different Omics Levels . . . 128
7.5 Discussion . . . 132

8 Conclusion 135

Bibliography 139

Appendices 179

Abbreviations 179

Appendix A Permissions and Contributions 185

Appendix B Background 189

Appendix C Formats 193

Appendix D Automated Workflows for Quality Control and Immunopeptidomics 197

List of Figures

2.1 Amino Acid Code Sun 7
2.2 HLA:Peptide Ligand:T-Cell Receptor Complex 10
2.3 HLA:Peptide Ligand Complex and Sequence Motifs 13
2.4 Official HLA Nomenclature 13
2.5 Tissue Preparation for HLA-Peptide Analysis 17
2.6 LC and ESI Schematics 20
2.7 Mass Spectrometer Schematic 23
2.8 MS1 Peak Map 25
2.9 Data Dependent Acquisition 26
2.10 Peptide Fragmentation Schema 28
2.11 Database-Driven Peptide Identification 32

3.1 mzIdentML Top-Level Structure 44
3.2 mzIdentML Structural Relations 45
3.3 mzQuantML Top-Level Structure 54
3.4 mzQuantML Structural Relations 56
3.5 Experimental Design Represented in mzQuantML 58
3.6 mzTab Table Sections 65
3.7 Experimental Design Represented in mzTab 67
3.8 mzTab Use in Scripting Environments (Example R) 72

4.1 qcML Top-Level Structure 80
4.2 QC Report rendered from qcML 83
4.3 Automatic QC Report Flagging with qcML 85

5.1 FRED 2 Module Overview 91
5.2 FRED 2 Usage Example 94

6.1 KNIME Integration of OpenMS 104
6.2 KNIME Integration of FRED2 105
6.3 Proteomics MS Identification Workflow with QC Report 107

6.4 HLA-Peptide Identification Workflow 109
6.5 QC TIC Plot 111
6.6 QC Mass Error Plot 111
6.7 QC Hydrophobicity Plot 112
6.8 QC Sample Timeseries Plot 114
6.9 Outlier Detection from QC Metrics 115

7.1 Variant-Count Variations Between Samples of the HCC Cohort 128
7.2 Variation Bottleneck for Potential Neoepitopes 129
7.3 Evidence for Mutated Proteins in HCC 130

B.1 HLA Polymorphy and Worldwide Population Diversity 189

C.1 Annotated Visualisation of an Identified Cross-Linked Spectrum 195

D.1 Outlier Detection from QC Metrics - PC 1,3 197
D.2 Outlier Detection from QC Metrics - PC 2,3 198

List of Tables

2.1 Mouse-Monoclonal Antibodies with HLA Specificity 18

3.1 mzTab Example with Multiple Search Engines 69
3.2 Identifications at a Glance with mzTab 73

4.1 QC Tools in OpenMS 82

5.1 Tools and Methods Available in FRED2 93

Chapter 1

Introduction

1.1 Motivation

Computational analysis is indispensable for the processing of modern life-science datasets, that have shifted research from a solely hypothesis-driven to a data-driven science approach with their unprecedented size, coverage, and depth of detail. Prime examples are projects that led us to our current understanding of the human genome and the human proteome. The human genome project, in particular, included experiments of enormous proportions, stretching in time over a decade and the results fundamentally expanded our understanding of the human genetic makeup and their functions. Beginning with the experiments' size and scope, the coverage of the human genome laid the basis to many hypotheses for the complex interplay of mechanisms underpinning life. Constant technological advancements, born out of large-scale efforts, like sequencing of the human genome, let us create experimental measurements in ever higher resolution and speeds, high-throughput fashion. High-throughput experimentation allows for better coverage but also necessitates a degree of automation and quality control by virtue of scale alone. Next-generation sequencing (NGS), for example, enables researchers to sequence a specific sample's genome and transcriptome as just another part of an experiment in day's time.
Domains of research that carry the '-omics' suffix usually aim for large-scale characterization and quantification of the biological molecules of their domain as they are present in cells, organs, or organisms. Large parts of the analysis in any omics experiment is computational, as the data measured needs usually multi-step algorithmic processing. The first "complete" human genome also created more questions, that can only be attempted to be answered by drawing on knowledge from other domains. In the post-genomics era, it became obvious that complex systems like cells are more than a list of their genes. Mere knowledge of the genes will not be sufficient to understand the properties of such a system. But the extreme complexity of the proteome, unfolding from the genome and transcriptome, or rather some of its analytical complexity is reduced by a known genome. The proteins of a sample can be characterized by

systematically profiling protein-coding genes as a start. In turn, proteomics can contribute aspects that are not evident from genomics alone, evidence of the successful synthesis of proteins from genes via transcripts and their abundance for one. This is why life-science research is often spanning multiple domains and technologies: to get a more complete picture and derive better testable hypotheses.
Immunology, for example, greatly relies on proteomics techniques to discover the cellular machinery and regulatory mechanisms behind a targeted immune response. Nevertheless, some of the complexity is rooted in human leukocyte antigen (HLA) locus polygeny and polymorphy. Research on the immune system has been influenced greatly by advances in proteomics, as proteins and peptides, small parts of proteins, play a crucial role in directing the immune systems specificity and function. One set of techniques in proteomics is particularly influential, mass spectrometry (MS) and its enabling technologies. Proteomics has its own generational turnover of technology, with sophisticated separation techniques and advanced MS instrumentation, etc., extending the coverage and sensitivity of measurements. The MS is a highly complex, and highly flexible, high-throughput instrument, for the identification and quantification of proteins and peptides. Here, too, a degree of automation and quality control is necessary by virtue of scale but also by virtue of complexity and diversity in which MS data comes. The dynamic nature and diversity of the samples measured (e.g., the immune systems peptides) introduces another layer of data complexity that often requires custom-designed computational analysis. Similarly, with the flexibility and diversity of MS techniques, follows the necessity for equally flexible analysis software. Workflow orchestration systems provide a flexible basis for automated analysis. With a sufficient amount of compatible tools integrated into such systems, they can become powerful tools for the analysis of multi-omics data. A great way to boost compatibility is through standardised formats for in- and output.
Complex workflows can be built for automated analysis of clinical data, to keep up with the increasing demand of clinical timelines to return actionable results. The quality assessment of data and (intermediate) results along the analysis can help researchers to use their available instruments and tools at best possible performance and meet the demands for supporting clinical decisions.

1.2 Structure of this Thesis

The thesis starts with a description of the background information on topics the following chapters center around. Proteinbiology, forming the biochemical basis for the interactions behind many cellular processes and more complex systems, like the immune system, is introduced. As one of the main sources of motivation to the research presented, our current understanding of immune system, in particular its role in cancer research, is characterized to highlight the implications to experimental and data analysis. Proteinbiology is also defining the technology producing measurement data used in this work, so MS and the techniques involved in successfully measuring proteins and peptides are described. The resulting data itself is detailed for its role in the analysis, and how the different experimental techniques and analysis task shape the results and tools used to arrive there.

The next chapter will pick up the topic of data and introduce the development of data standards that help with the handling and integration of data from various MS analysis task. Different emphasis on data handling, archival, and use in automated analysis, is put into the design of their structure, depending on the researchers' requirements. The community-driven approach for standards will be described for three formats, mzIdentML, mzQuantML, and mzTab. The first, mzIdentML, is accommodating new experimental techniques to an established standard for extensive documentation of identification results. Quantitative results at similar depth can be documented and archived in mzQuantML, which is described next. A combined standard for identification and quantification results is focused on emergent development of novel techniques and easy access as summary report on an analysis is mzTab.
A versatile format to accommodate quality control data at all stages of the analysis is described in the next chapter. The development of qcML took the best fitting aspects of the previously developed standards to fulfil multiple roles of quality report, archival, and hand-over format in a multi-step analysis.

The following chapters introduce the open-source frameworks Open Mass Spectrometry (OpenMS) and Framework for Epitope Detection (FRED)2 that provide a comprehensive set of data analysis tools for computational MS and immunology. By integration into

a common workflow orchestration system, their synergy through compatibility (not least because of the use of data standards) will be highlighted in complex workflow designs, aimed to provide data analysis to personalised cancer therapy. One workflow aspect will focus on the identification of low-abundance immunopeptides, measured from specialised experimental setup for immunological research. The other will focus on the integration of quality control along existing workflows, to provide best possible results with complex experimental systems, such as MS and immunopeptidomics.

We build on the last chapter's workflows to illustrate their application in personalised cancer therapy development in the next chapter. The data analysis for such personalised therapy development needs to draw on the previously described intercompatibility, flexibility, in a multi-omics setting to arrive at results that can be applied in leading-edge therapy development. This chapter follows the study of personalised target discovery from the genomics, transcriptomics, proteomics, and immunopeptidomics data of a cohort of hepatocellular cancer patients.
The thesis finishes with a conclusion in Chapter 8.

Chapter 2

Background

Proteins play the most important roles in the majority of cellular processes. They constitute the means for the many complex tasks of sustaining a cell's function and proliferation[1]. Unsurprisingly, most of a cell's dry matter consists of proteins[2,3].
Proteins are also subject to degradation and renewed synthesis, in essence reflecting the present state of a cell. Hence, analysis of a tissue's proteins can provide insight into the function of microbiological processes and give indications of disease. The following chapter will introduce the life science background to computational and mass-spectrometry based proteomics in general and in particular the immunological background behind cancer immunotherapy design and the approach to personalised medicine.
The analytical techniques applied will be introduced together with their computational challenges and algorithmic solutions.

2.1 Protein Biochemistry

Proteins are built from a sequence of amino acids, chemically chained by linking the α carboxy group of an amino acid with the α amino group of another amino acid under elimination of water.
A protein is not only defined by its amino acid sequence but also by its spatial structure. This structure is dependent on the sequence of amino acids but also on the chemical environment (e.g., acidity or the presence of chaperons, proteins that influence structural formation without permanent integration into the resulting structure itself). There are four levels of protein structure, with the first or primary being the sequence itself. The secondary structure is defining the form a stretch of consecutive amino acids is taking. Common forms are alpha helices and beta sheets[1]. The tertiary structure is defining how the secondary structures are arranged in space (in relation to each other). As functional proteins can be comprised of several amino acid substructures unconnected by peptide bonds but noncovalent bonds, e.g., hydrogen bonds, van der Waals interactions, ionic interactions, and hydrophobic interactions. This aggregation

is defined as the quaternary structure. For example, β_2 microglobulin of the major histocompatibility complex (MHC) is non-covalently linked with the α-chain (Fig. 2.3). The order in which amino acids are chained together is ultimately determined by the genetic makeup of an organism.

2.1.1 The Central Dogma: From DNA to RNA to Proteins

Genetic information is expressed into protein functionality by

1. copying the protein encoding section (gene) of the deoxyribonucleic acid (DNA) molecule (chromosome) into ribonucleic acid (RNA) (transcription)
2. trimming non-coding intervening sequences (introns) off the transcript (splicing) and apply post-transcriptional controls
3. using the spliced transcript (mRNA) as templates for the protein synthesis (translation)

As all cells use this chain of processes in some form[1] to create their protein machinery, it is called the central dogma of molecular biology.

Building Molecular Machinery: Proteins

The genetic material in its most basic form is a double helix of complementary nucleotide chains (DNA). There are four nucleotides in DNA, the deoxyribonucleotides adenine (A), guanine (G), cytosine (C), and thymine (T) that can be covalently linked at the 3' and 5' ends of their carbon backbone. They are complementary, adenine pairing with thymine, cytosine with guanine, meaning single strands of complementary DNA will bond together and form a double helix. This also means that information encoded through a sequence of A/C/G/T is stored redundantly in a double helix of DNA. In a first step to convert genetic instructions into working proteins, the gene is transcribed into a single stranded RNA nucleotide sequence. RNA differs from DNA, by the presence of an additional hydroxyl group and the use of the nucleotide uracil (U) instead of thymine, similar enough that pairing with adenosine is unimpaired. Among other processes, eucaryotic transcribed RNA is capped with a 5'-methyl cap, designating it as mRNA. The ribosome, a complex of RNAs and proteins, associates with the mRNA and catalyses the synthesis of protein. From four different code characters (A/C/G/U) in triplets (codons), the translation process can theoretically form sequences of 4^3 different characters. This is sufficient for the 20 proteinogenic amino acids (not counting selenocysteine) with partial redundancy (Fig. 2.1).

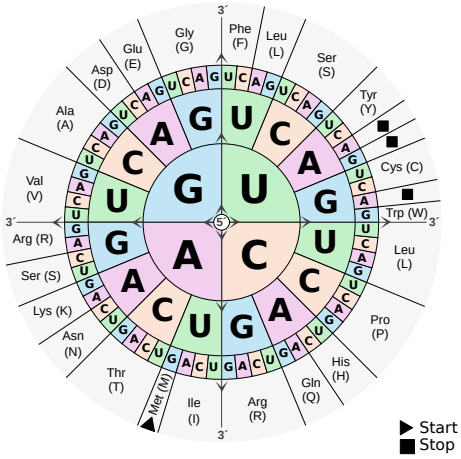


Figure 2.1: Representation of the base sequence encoding of different amino acids

Small RNA-aminoacid adapters (aminoacyl-tRNA) with the fitting complementary code to the mRNA attach to the ribosome and the RNA (Fig. 2.1). Adapters are produced by coupling tRNA to a fitting activated amino acid by specific aminoacyl-tRNA synthetase enzymes. The ribosome catalyses the formation of a peptide bond between the coupled amino acid C-terminus sitting closer or at the start of the mRNA and the N-terminus of the next adapter's amino acid. The ribosome moves from the mRNA 5' start to the 3' end and a polypeptide chain grows by the stepwise addition of amino acids. The ribosome also facilitates the correct start of translation by first attaching to the initiation codon of the mRNA. Translation is moderately fast ($\sim$ 200 amino acids per minute[4]) but is sped-up in scale through the formation of polyribosomes. Here, multiple ribosomes initiate translation successively on one mRNA molecule, translating in offset parallel, which provides an amplification effect in the production of protein molecules. To complete the protein production, the completed polypeptide chain must fold correctly, either with the help of assembly catalysts (chaperones) or through self-assembly, into its three-dimensional conformation.

DNA/RNA sequencing technology

With great initial effort, the human genome had been sequenced and revealed the genes to the proteins known by research[5–8]. Done with automated Sanger sequencing, it took over a decade to assemble, as the technique was still based on the determination of DNA sequences by primed synthesis with DNA polymerase and concurrent size fractionation of the products by electrophoresis, developed by Sanger and Coulson[9]. With the advent of low-cost, high-throughput methods for genome sequencing (DNA-seq) and transcriptome profiling (RNA-Seq) through NGS, the development of individualised treatments to diseases that have their origin in the aberration of the genetic code, like cancer, became feasible. For example, the study and comparison of an individual's tumour and healthy tissue reveals the genomic changes described in the next section and can provide targets for the development of individualised therapies (see section 2.2.2). NGS comprises several different techniques but can largely be summarised as

a three-step process: sample and library preparation, sequencing, followed by data analysis and bioinformatics. For individualised sequencing of a tissue sample, the target nucleic acid is isolated from the sample with established protocols. RNA samples are converted to cDNA by reverse transcription for stability during the measurement. The sample is then fragmented and common sequences (adaptors) added to either end. This 'library' of fragments is attached to a flow cell, a microfluidic device to sequence the library. Before sequencing, each fragment is expanded by polymerase chain reaction (PCR), creating clusters of the same fragment in the flow cell, to amplify the measured signal. With the cycle reversible terminator technique, modified nucleotides with fluorophores are added to the flow cell and pair with the first complementing and open library nucleotide, the fluorophores are excited and the library imaged. The terminators are removed, and the previous step is repeated, until the library nucleotides are all complemented. The images colours for each cluster are translated into a sequence, one read per cluster and reported in a plain text FASTQ file. Details, benefits, and drawbacks of the existing technologies are beyond the scope of this work, but are covered in many reviews[10–12]. The reads must then be mapped to the reference genome, allowing mismatches, insertion-deletions (INDELs) and clipping, to assemble the individualised sequence of the sample. Starting with heuristic placement of the reads, coverage is corrected and extended through alignment algorithms. State of the art read mapping software BWA[13,14], Bowtie[15,16], and TopHat[17,18], for example, use implementations of Burrows-Wheeler transform (BWT)-based alignment algorithms. The read alignment is usually reported in Sequence Alignment Map (SAM) files or their binary representation (BAM) to conserve disk space.

Genetic Mutations and Protein Modifications

Due to the codon redundancy, proteins can have the same amino acid sequence despite genetic mutations –silent mutations– and are of no immediate consequence for individualised targets. Non-silent (ns) mutations will change the protein sequence, and so too, their mass as observed by MS will change. Small changes can affect the protein little to not at all, if the changes are outside functional site coding regions or otherwise regions with impact on the protein structure. If they are, proteins can yet retain the same or still functional shape and fulfil the same tasks, though sometimes to varying degrees of efficiency in different environments[19,20]. These are called isoforms, proteins of different shape but from the same gene, and can have stable occurrence in the population. This emphasises the need for healthy and tumour tissue comparison in individualised targeting. Isoforms can also arise from alternate splicing or variable

promotor use.
Genetic mutations can be categorised as follows:

- A single alteration is called a single nucleotide variation (SNV)
- If the sequence alteration is common and can be considered established in the gene pool, it is called single nucleotide polymorphism (SNP)
- Insertions are sequence changes that introduce more than one new nucleotide by either mis-repair/mis-replication or injection by mobile genetic elements such as transposons and viral DNA.
- Deletions define the event of a loss of nucleotides, the replacement of one part of the sequence with another is an INDEL.

Insertions and deletions are potentially shifting the reading frame of the gene (frameshift, fs) if their size is not equal to multiples of three. Their occurrence usually leads to a defective protein, which still can be a valid target for individualised therapy depending on its HLA presentation potential[21–23].
To distinguish between misalignment, allelic variants, and true mutations, the read alignments must be further analysed. For cancer research, the variants of most interest are those only present in somatic cells, which have been acquired not inherited. Calling somatic variants, for example for the discovery of tissue specific therapy targets, read alignments from different sequenced tissues need to be compared. Subtractive methods (GATK, SAMtools) first detect variants from the samples individually, then split by variants present in all samples (germline) and tissue specific variants (somatic). Alternatively, variant detection can be performed on the combined read alignments, using statistical modelling to separate germline and somatic variants, which needs additional information on allelic frequencies[24] or genotype likelihoods[25]. Detected variants can be further annotated with their transcript-relative coordinates, functional roles, occurrence in disease variant databases[26], and cancer association[27], for better prioritisation. Variants and annotations can be exchanged with a common format, the variant call format (VCF)[28].
Protein variations can also arise after the synthesis, through post-translational modifications (PTMs), the family of proteins sometimes referred to as proteoforms[29]. PTM can be reversible or irreversible, affecting the structure and function of proteins, and generally demonstrate the dynamic interaction of proteins in a cell. Their functions range wide, from protein localisation indication, signal transduction, gene regulation, to enzyme function regulation, and influencing protein–protein interaction (PPI). They are generally events that change the structure, mass, and chemical properties of the

protein by proteolytic cleavage, or adding a modifying group (e.g., phosphoryl, methyl, acetyl, glycosyl) to one or more amino acids. Most studied are the types of phosphorylation, methylation, acetylation, ubiquitylation, SUMOylation, and different types of glycosylation. However, PTM are a field of active research and many PTM functions have yet to be discovered[30,31].

2.2 Immunobiology

Figure 2.2: T-cell receptor (purple, ochre) bound to HLA-A2 (emerald, pink) presenting a peptide ligand (IMDQVPFSV, grey); rendered from PDB:6VM8[32] with Mol*[33].

The immune system is commonly defined as a collection of biological systems that protects its host from disease. More precisely, many of its functions are directed against invading pathogens, aberrant host cells and in cases of an aberrant immune system, the host cells in general.

The part of importance for this work is the adaptive immune system, which employs pathogen and antigen-specific responses to counteract the presence of foreign molecules in the host. Specificity is achieved through highly adaptive types of cells, the lymphocytes. These bear variable cell-surface receptors for antigens with great diversity, enabling the immune system to specifically recognise (and remember) a vast variety of antigens. The cell-mediated immunity is provided by lymphocytes called T cells, their eponymous cell surface receptors are T-cell receptors (TCRs). The TCR consists of $\alpha : \beta$ heterodimeric receptors associated with the proteins of the CD3 complex. The V domains of the extracellular polypeptide chains of the TCR (α, β) are encoded in sets of gene segments that undergo somatic recombination to form a complete V-domain exon during T-cell development in the thymus. These domains form the centre of the highly variable antigen-binding site, whereas

the TCR periphery will be subject to little variation (Fig. 2.2, purple, ochre). The recombination contributes to the huge diversity of the TCR repertoire, in the order of $1 * 10^{15}$ before thymic selection, and $1 * 10^{13}$ after[34]. Thymic selection is the process that ensures that the resulting TCR are self-tolerant and MHC-restricted. The MHC (, or HLA in humans) molecule makes up a major part of the ligand for the TCR, and fulfils two critical roles in the detection:

1. as the molecule facilitating the antigen presentation and
2. as part of the control measure restricting T-cell activation preventing general host targeting

Since most pathogen activity is intercellular, i.e., pathogens invading the host's cells (e.g., by viruses hijacking the host's protein production facilities) the immune system needs a window into the host's cellular activity to act against infected sites. This is achieved through the presentation of peptide fragments derived from (self and foreign) proteins, bound to MHC and transported to the cell surface of most host cells. These peptides have been captured and bound to the MHC by specialised molecules in the host cell that sample the protein degradation products within the host cell. Cytosolic proteins are degraded to peptide fragments by the proteasome[35], as part of the ubiquitin-dependent degradation pathway for cytosolic proteins, maintaining protein homeostasis[36]. Damaged, misfolded, short-lived, over-expressed, including non-self proteins, are tagged with ubiquitin by a sequential cascade referred to as ubiquitylation. Ubiquitylation is one of the most prevalent reversible PTMs[37]. The proteasome is a large multicatalytic protease, with caspase-, trypsin-, and chymotrypsin-like activities[37], and three ubiquitin receptors (Rpn1, Rpn10, and Rpn13), components of a proteasome subunit[38], targeting the tagged proteins.
Transporters associated with antigen processing (TAP) will transport some of the degradation products, peptides, from the cytosol into the Endoplasmic Reticulum (ER) which then bind to partially folded MHC class I molecules already present in the ER, bound to TAP via tapasin and held stable via a complex of chaperones (calreticulin and Erp57)[39]. The TAP transporters prefer peptides of eight or more amino acids with hydrophobic or basic residues at the carboxy terminus. Peptide binding to the MHC class I molecules completes the folding of the then stable MHC class I molecules, which are then released from TAP and exported to the cell's surface. The loading of MHC class II molecules differs[39] but is not described here in detail.
The MHC molecules' part in the control measure restricting T-cell activation is one of positive selection. Since only the smallest and highly variable domain of the TCR contacts the bound peptide, the rest of the TCR-binding domains must be able to make

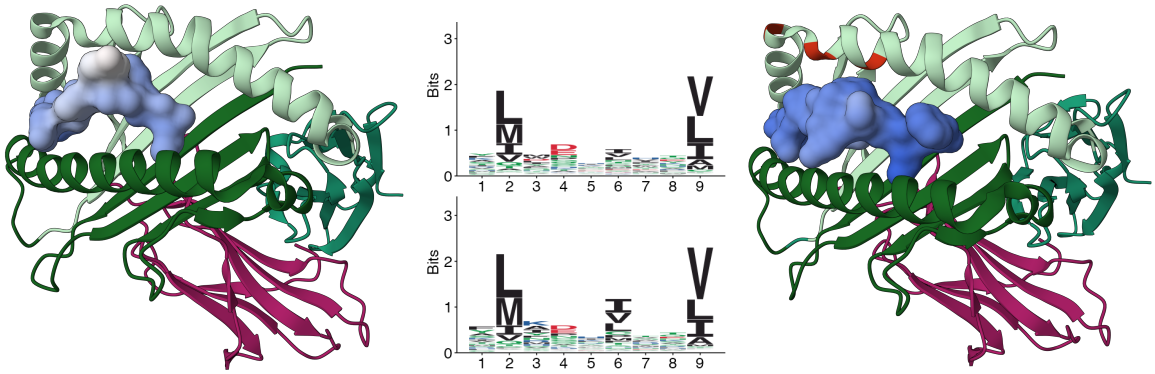
contact with the surrounding polymorphic surface of the MHC molecule itself. Only T-cell precursors whose receptors can interact with the MHC molecules (i.e., they are accepting MHC:self-peptide complexes[39] as their ligands) are allowed to survive and mature and the surviving T-cell population is thus MHC-restricted. Additionally, the developing T-cell population undergoes negative selection. This happens by clonal deletion if a TCR interacts with ubiquitous self-antigens, and by clonal inactivation through tissue-specific antigens presented in the absence of co-stimulatory signals. The surviving T-cell population is thus self-tolerant[40].

2.2.1 The HLA Molecule and Its Ligands

The human homologues of the MHC are most often referred to as HLA molecules. The HLA locus consists of three regions, designated class I, class II, and class III based on their proximity and function of gene products. The main function of HLA class I gene products is to present endogenous peptides (self, viral and tumour-derived peptides), while class II processes exogenous peptides on antigen-presenting cells (APC). In the previous section, the HLA molecule plays an important role in mediating the immune response. A part of this role is the presentation of antigens on the cell surface. Due to the huge diversity of the TCR repertoire, the recognition of pathogen evidence or other non-self peptides is highly adaptive. This must be matched with a capability to present a huge diversity of peptides to be recognised. The HLA molecules achieve this with a number of mechanisms, all based on the structure and origin of the HLA molecules, explained in the following sections. The HLA class I products are the relevant molecules for the core of this work, and only their function will be explained in more detail.

Structure of the HLA class I molecule

HLA class I molecules are expressed on almost all nucleated cells and thrombocytes. Expression levels however vary, with liver cells (hepatocytes) expressing relatively low levels[39]. Functional HLA class I molecules consist of two peptide chains and a presented peptide. The light β_2-microglobulin chain ($\sim$ 12 kDa), encoded on chromosome 15 outside the HLA locus, is bound non-covalent to the heavy α chain ($\sim$ 43 kDa)[41]. The α chain ($\sim$ 365 aa) has three extracellular domains, of which the α3 domain anchors the molecule with a transmembrane region to the cell membrane and α1 and α2 form the binding cleft for the presented peptide[42] (Fig. 2.3 (a) & (c)). The cleft is formed from eight anti-parallel β sheets, forming a base, and two α helices, forming a rim, which limits the length of the presented peptide (see Fig. 2.3).



(a) (b) (c)

Figure 2.3: The figure shows a) HLA-A*02:01 and c) HLA-A*02:03 with a peptide ligand bound, the amino acid sequence differences for HLA-A*02:03 highlighted in red. b) shows the binding motifs, and visualisation of sequence patterns of 9-mer peptides binding HLA-A*02:01 (top), HLA-A*02:03 (bottom). The HLA molecules are illustrated with their β_2-microglobulin chain at the bottom (pink), the peptide ligand (blue) sitting in the binding cleft formed by the $\alpha 1$ and $\alpha 2$ domains of the α chain (emerald, light green); rendered from PDB:1DUY[43] and PDB:3OX8[44] with Mol*.

HLA Locus Polygeny and Polymorphy Make A High Diversity of Presented Peptides Possible

The HLA variability to match the TCR comes from two sources. First, the polygeny of the HLA locus, which has three genes for the HLA class I molecule (HLA-A, -B,-C) and three genes for the HLA class II (HLA-DR, -DQ,-DP). The other comes from the polymorphy of the HLA genes. Especially the regions in contact with the presented peptide are highly polymorphic. Given an inheritance without recombination events, HLA genes are transmitted on a single chromosome from each parent, and a codominant expression results in potentially six different HLA class I molecules expressed on the descendant's cells[45].

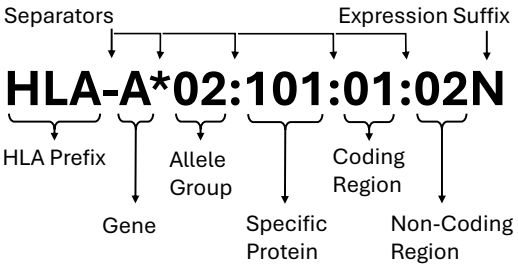


Figure 2.4: Official HLA-type nomenclature

A unified nomenclature (Fig 2.4) for the different alleles discovered in the human population was first introduced by the WHO Nomenclature Committee (1968)[46] and has been regularly updated[47].

The `https://hla.alleles.org` definition of the nomenclature reads[48]: "Each HLA allele name has a unique number corresponding to up to four sets of digits separated by colons. The length of the allele designation is dependent on the sequence of the allele and that of its nearest relative. All alleles receive at least a four digit name, which corresponds to the first two sets of digits, longer names are only assigned when

necessary. The digits before the first colon describe the type, which often corresponds to the serological antigen carried by an allotype. The next set of digits are used to list the subtypes, numbers being assigned in the order in which DNA sequences have been determined. Alleles whose numbers differ in the two sets of digits must differ in one or more nucleotide substitutions that change the amino acid sequence of the encoded protein. Alleles that differ only by synonymous nucleotide substitutions (also called silent or non-coding substitutions) within the coding sequence are distinguished by the use of the third set of digits. Alleles that only differ by sequence polymorphisms in the introns, or in the 5' or 3' untranslated regions that flank the exons and introns, are distinguished by the use of the fourth set of digits."
Polymorphy, Mendelian inheritance, and codominant expression result in a huge diversity of HLA types found in the worldwide population. However, certain Alleles dominate in certain populations[49–51]. An example can be found in Appendix B.

Structure and Polymorphy Result in Different Binding Motifs

A function of HLA molecules is to bind peptide fragments derived from pathogens and display them on the cell surface for recognition by the appropriate T cells. The polymorphy in the gene regions for the HLA protein that form the peptide binding cleft yields allele-specific binding motifs, which are dominated by anchor residues at position 2 and 9[52–54]. HLA class I molecules typically present intracellular peptide antigens of 8 to 13 amino acids[55–57]. The sequence of peptides presented has to mirror the pockets formed by the environment of biochemical attracting and repulsing forces formed by the binding-cleft facing amino acids of the HLA molecule and can be measured by large-scale observation of the peptides found binding to the HLA molecule and then expressed with a sequence motif. The sequence motif[58] represents the magnitude of statistical occurrence of an amino acid at each position of a binding peptide by relative height of the stacked amino acid representations in one-letter code[59]. Fig. 2.3 shows (b) the ligand motif for HLA-A*02:01 from the MHC Motif Atlas[60] on top of the ligand motif for HLA-A*02:03 and (a) the schematic representation of a ligand, fitting into the binding cleft of HLA-A*02:01, with the amino acid side chains forming protrusions of the otherwise more or less linear peptide ligand fitting into the binding pockets of HLA-A*02:01. (c) shows peptide ligand binding of HLA-A*02:03, the three amino acid differences to HLA-A*02:01 highlighted in red.

HLA Typing

Using serology[61] was the first method to determine an individual's HLA allele types. Molecular typing has since replaced serologic typing. First solutions involved DNA-matched sequence-specific primers for the HLA region and PCR. With the introduction of NGS, the specific genetic HLA allele makeup of a sample can be read at unprecedented precision. Several algorithms have been developed to help with the typing from NGS data[62–65], as the substantial sequence similarity within the HLA region but high polymorphism of the HLA loci renders HLA genotype deduction from sequence difficult. The specific preferences of peptides presented in different HLA molecules and a documented[66] diverse HLA genotype in the world population make HLA typing essential for the development of vaccines, and especially important for the development of personalised immunotherapies.

2.2.2 The Immune System in Health and Disease (Cancer)

Cancerous tumours represent the growth of an abnormal cell population, with a changed expression profile, mutated, inappropriately expressed, or overexpressed proteins, and following, a different profile of HLA-presented peptides. In 1909, Paul Ehrlich envisioned what is now called immunosurveillance, the repression of developing tumours[67] by the immune system. Since then, it has been shown that some tumours elicit specific immune responses that suppress their growth[68,69]. The concept of immunosurveillance was later broadened into cancer immunoediting, consisting of three phases: elimination (i.e., cancer immunosurveillance), equilibrium, and escape[70].
With growing insight into cancer development, it was possible to define the (six) characteristic hallmarks of cancer[71], common traits that mark alterations in cell physiology and drive the transformation from normal to cancer cells. These are: sustaining proliferative signalling, evading growth suppression, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis. While the three-step progression of immune system elimination, equilibrium, and escape, still holds true, Hannahan and Weinberg[71] describe a much more diverse picture of cancer development, later adding four more emerging hallmarks[72]: reprogramming of energy metabolism, genome instability and mutation, tumour-promoting inflammation, and immune destruction avoidance. These emerging hallmarks effectually create a tumour microenvironment, where the paradoxical roles of the immune system during cancer development[73] play out. The cytokine TGF-β, playing an integral role in regulating immune responses[74] (effector and regulatory CD4 (+) T cell responses), can be tumour suppressing at early stages of tumourigenesis and tumour promoting later

on[72]. Tumour necrosis, for example, is a synergistic consequence of metabolic stress and inflammation, and is a common histological feature. Necrotic cell death releases proinflammatory signals into the surrounding tissue. As a consequence, a tumour microenvironment can recruit inflammatory cells of the immune system[75,76]. Those immune inflammatory cells are major sources of the angiogenic, epithelial, and stromal growth factors, which can be actively tumour promoting, helping angiogenesis, cancer cell proliferation, and invasiveness, three of the other hallmarks. Similarly, subclasses of B and T lymphocytes may facilitate the recruitment of those tumour-promoting macrophages and neutrophils[77–79]. Our picture of tumourigenesis and the progression in different types of cancer is still very much incomplete. But with i.a. radiotherapy and certain cytotoxic drugs showing double-edged effects in the tumour microenvironment (metabolic stress, autophagy and necrosis), conflicting inflammatory responses (TGF-β signalling), and the overall role of the immune system, come good arguments for research into targeted immunotherapy to (re)direct the immune cells toward tumour destruction.

Cancer Vaccines

The facts that cancers have a changed protein expression profile (reflected by the HLA peptides presented on the cell surface), that tumours do elicit specific immune responses, that a hallmark of cancer development is immune evasion, all bear the prospect of helping the immune system to better fight tumours with targeted immunotherapy. Active immunotherapy treatments revolve around vaccines designed to (re)activate tumour cell recognition. Since it has been shown that CD8- und CD4-positive T cells can recognise tumour associated antigen (TAA) to induce a specific immune response[80–83], the first found and tested TAA were those over-expressed in tumours, such as tyrosinase-related protein (TRP) 2, gp100, the cancer-testis antigens NY-ESO-1 and the MAGE family[84–88]. A drawback of many of those antigens is that though they are able to elicit T-cell response, they are expressed in malignant tissues with restricted normal expression (e.g., cancer-testis antigens). A more targeted approach was pioneered by Rammensee et al. with the identification of naturally processed peptides[89–91]. The approach also greatly helped the increasingly detailed definition of the MHC ligandome and allele-specific binding preferences. A peptide vaccine based on naturally processed peptides has the advantage of faster production and an easier-to-handle product than a similarly specific approach like the adoptive T-cell transfer. It also scales better, allowing to formulate vaccines with more than one target to avoid tumour evasion[92]. Even with more than one peptide in a vaccine cocktail, specific reactions can be documented

for the individual peptide through immunomonitoring[93]. The specific targeting also makes it a preferable method in personalised medicine[94].

2.2.3 Isolation of Naturally Processed HLA Peptides

The method to identify naturally processed HLA peptides can start directly from tissue samples or cell culture (Fig. 2.5). Tissue is lysed and solubilised proteins are isolated

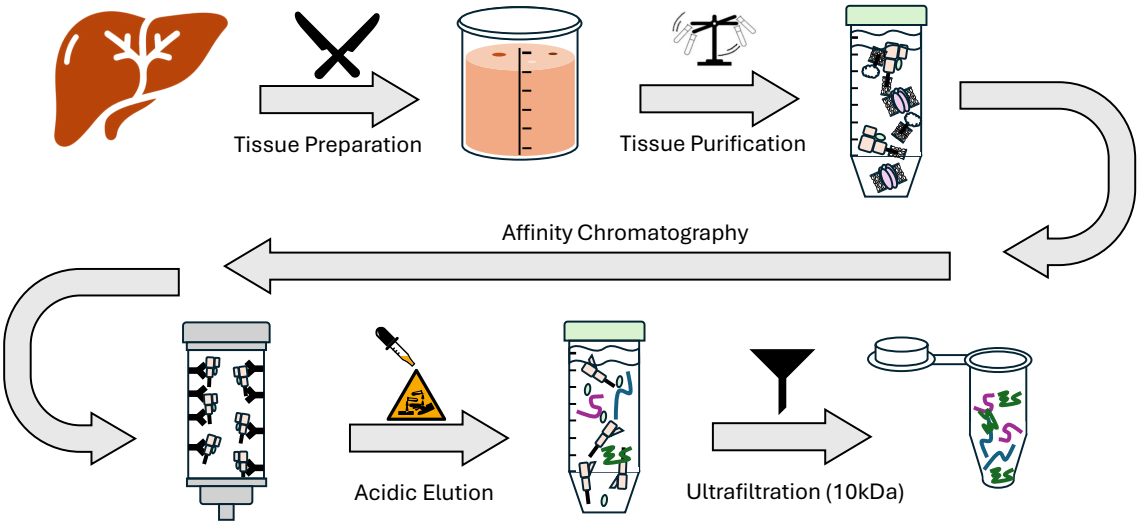


Figure 2.5: Process of tissue preparation for HLA-peptide analysis

through centrifugation of the lysate. To control and standardise the preparation, protein concentration can be measured through spectrophotometry. Sample UV absorbance is measured at 280 nm, the maximum absorbance wavelength for aromatic amino acid side chains, like in tryptophan and tyrosine residues. An approximate protein concentration is then calculated given the concentration of a light-absorbing molecule species is proportional to its absorbance (Beer-Lambert Law), and the concentration in solution can be adjusted to standard.

The HLA proteins are filtered from the solution through targeted affinity chromatography. Highly specific antibodies are available for different HLA molecule types (Tab. 2.1). The antibodies are bound to the solid phase of a chromatography column and used to precipitate the targeted HLA molecules. Dissociation of peptides and HLA molecules bound to the column through the antibody is achieved by acidic elution and results in a solution of free HLA α chains, β_2-microglobulin, and formerly HLA-bound peptides. An ultrafiltration step at 10 kDa isolates the peptides for further analysis.

Antigen	Clone	HLA Class
HLA-A, HLA-B, HLA-C	W6/32[95]	I
HLA-A*02	BB7.2[96]	I
HLA-B, HLA-C	B1.23.2[97]	I
HLA-DR, DP and most DQ	Tue39	II

Table 2.1: Mouse-monoclonal antibodies with HLA specificity

2.3 Mass Spectrometry-based Proteomics and Immunopeptidomics

The further analysis of the HLA peptides from the previous section, the identification of the peptide sequences in particular, can be achieved by MS measurements. As the name implies, MS is an analytical technique to determine the mass of molecules. In general, MS is a valuable tool for the life sciences, in particular proteomics, immunopeptidomics, and metabolomics, because of its great accuracy and precision for the identification and quantification of molecular compositions from complex samples.

2.3.1 Sample complexity and Liquid Chromatography

Though MS measurements are fast (20-200 Hz[98–100]) and can be used at very high resolution (on the atomic scale, resolving isotopic peaks[101]), the number of components in complex samples usually found in the life sciences, poses a problem. This comes as no surprise considering the size of the human proteome. Of the 20,435 human genes curated in UniprotKB/Swiss-Prot[102] (release 2024_3) mass spectrometric evidence has been found for 18,097 proteins[103]. The amount of different compounds in biological samples (thousands for proteins[104]), the combination of resulting peptides, and the presence of the same compounds in multiple isotopic variances, with or without a variety of small mass chemical modifications, multiple charge states, can easily overwhelm signal separation for reliable data analysis[105].
To overcome the sample complexity issue, liquid chromatography (LC) can be coupled to the MS inlet, effectively spreading the sample by physicochemical properties over time. The liquid sample, most often already present in liquid form for proteomics and immunopeptidomics, is pumped as the mobile phase through a separation column containing an absorbent material (stationary phase). The most widely used technique, reversed-phase (RP), uses non-specific hydrophobic interactions between the polar sample and an apolar stationary phase (e.g., a nonpolar carbon chain (usually C18)) bonded to a silica base. A polar organic solvent (e.g., acetonitrile) competes with the sample molecules for sites on the apolar stationary phase. The more polar a sample molecule is, the higher the displacement through the solvent, the faster it

moves through the column. Increasing the solvent concentration on a gradient will affect less polar sample molecules later and thus effectively separate the sample by hydrophobicity.
Proteome complexity is however still challenging with modest separation efficiency of intact protein analysis through LC[106]. Surfactants used for protein solubilisation have ion suppressing effects, making them MS-incompatible. A solution is to digest the proteins into peptides, e.g., with the sequence-specific protease trypsin. The resulting peptide lengths are comparable to naturally presented HLA-peptides introduced in previous sections, albeit on the larger side (~ 14[107]). Surfactants can be washed after digestion, and LC of peptides generates a much smoother and equal signal distribution. In addition, large proteins generate unnecessarily complex MS spectra, harder to identify. As Washburn et al.[108] established in 2001, peptides can identify proteins and allow for many protein identifications in a single experiment. This bottom-up sample preparation technique is sometimes also called shotgun-proteomics in reference to the shotgun-approach of nucleotide sequencing found in NGS.
LC is also a major contributor to experiment variation. The retention time (RT) of the sample molecules is highly dependent on the gradient and the instrument status. High pressure is used to pump the analyte through the column to achieve a high degree of separation. Here, flow rate fluctuations from inconsistent pump performance or system leaks, contribute to different RT-points and elution lengths between the same sample molecules between experiments. Similarly, subtle gradient changes between experiments can affect the RT of the sample at different stages of the gradient between experiments. Some sample contents or contaminants may be strongly retained and result in 'carryover' into the next experiment, depending on column wash protocols in place and the use of guard columns. Finally, the age of the column contributes to the measurement performance. Over time, the stationary phase may degrade, microbial growth in the porous stationary phase can contaminate the measurement, small (nonbiological) particulates from the sample preparation or shed from the LC/Autosampler system can clog the column, sample buffer salts can precipitate in reaction to the organic solvent and also clog the column. This creates hard-to-precisely-recreate measurements, which needs to be considered during data analysis. Some issues, like the RT differences, are inconvenient but can be compensated with computational methods (e.g., RT peak alignment methods). Others are more serious and can lead to intermittent or complete sample loss, data misinterpretation, and require close Quality Control.

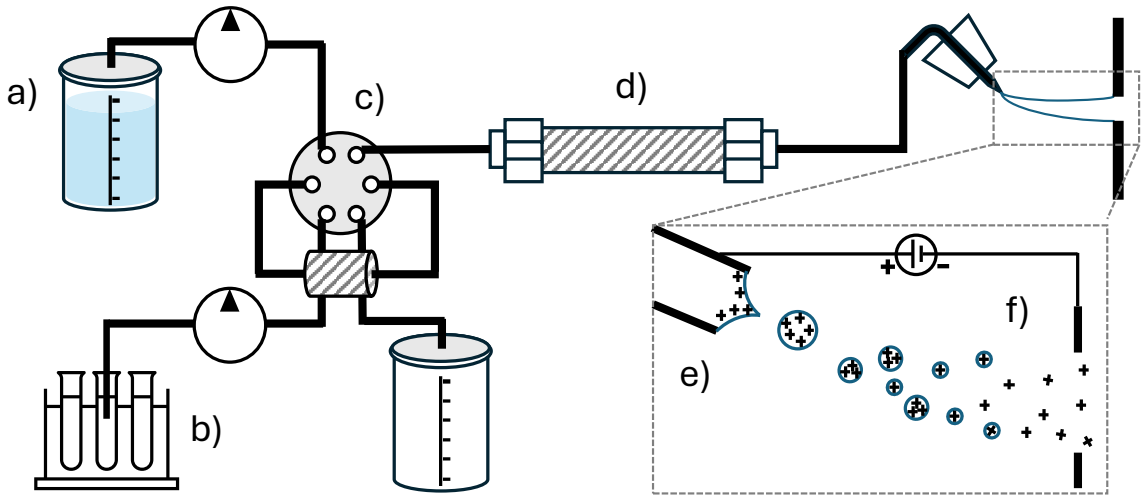


Figure 2.6: LC and ESI setup commonly coupled to a mass spectrometer. a) a solvent (mobile phase) is pumped into an c) injection valve where it can carry the b) sample material injected from an autosampler under high pressure through d) a column with chromatographic packing material (stationary phase). The eluting analytes form a Taylor cone shape at the tip of the electrospray needle, from which charged droplets emerge. The spray plume droplets get accelerated in an electric field (between the tip and MS), and evaporate in the depressurised space around the f) MS inlet, resulting in (almost) solvent free ions entering.

The LC/Autosampler system can be directly coupled to the MS (Fig. 2.6), which allows for automated and high-throughput sample measurements. The instrumentation setup is usually summarised as liquid chromatography mass spectrometry (LC-MS).

2.3.2 Mass Spectrometry

In order to achieve highly accurate and sensitive measurements MS manipulates charged particles (ions) using magnetic and electrostatic fields. Mass is indirectly measured by separating ions and measuring the mass-to-charge ratio of individual ions by observation of their movement through an electric field in vacuum. The first mass spectrometers, then called parabola spectrographs, were built at a time when the atomic theory was not yet settled, electrons newly discovered, and the existence of isotopes unconfirmed. This was the beginning of the 20th century, and the results of experimental use of these mass spectrometers contributed in part to our modern understanding of matter (e.g., the existence of isotopes). The key principle behind this method of analysis is the defined behaviour of an ion of certain mass and certain charge state in an electric and magnetic field, a vacuum preventing influences from ambient molecules on the ion movement and neutralisation. The ionised molecules have to be directed into the analysis instrumentation like the mass analyser or detector.

As the molecules are charged, precisely applied forces of electric fields will steer the direction. Modern setups involve several instrumentation stages through which the ion stream has to be directed. Each MS consists of three basic components, the ion source, the mass analyser, and the detector.

Ion Source

Only ions can be analysed with a MS, i.e., each molecule must carry at least one charge. The ions also need to be in gas phase, so the sample in liquid form from the previous LC separation step needs to be vaporised. The simplest techniques to create ions are processes which impart high quantities of residual energy to the samples' components, also called hard ionisation techniques. The energies involved in hard ionisation techniques would cause however unwanted degradation in biological samples. Most samples are thermally labile, and their primary structure (e.g., amino acid sequence) is needed intact –at least at first– for successful analysis. Methods that can accomplish this task are called soft ionisation methods and their development for protein ionisation was awarded the Nobel Prize in Chemistry, 2002. The nowadays predominant method in proteomics is electrospray ionisation (ESI) and was introduced in 1985 by John Fenn and co-workers. Here, the liquid sample solution has to have a proton or electron donor mixed in. Depending on the donor type, positive mode (protons) and negative mode (electrons) are distinguished. And depending on the pI of the sample molecules, they will be ionised more easily in one or the other mode, with positive mode dominating proteomics. The organic solvent acetonitrile used in the LC step has no proton donor capabilities, thus formic acid can be added to enhance protonation. The conductive mixture is pushed through a needle of very low diameter into an electric field between needle tip and mass spectrometer. The electric field is induced through the application of a high voltage between the tip of the needle as the positive pole in positive mode and the MS ion inlet as negative pole. The resulting aerosol is formed of charged droplets accelerated in the electric field (Fig. 2.6). The electric field also competes with droplet surface tension, forming a cone shape from the droplets named after the discoverer of the effect, Sir Geoffrey Ingram Taylor. The solvent in the aerosol droplets evaporates, aided by heat and an applied pressure differential. Through the volume reduction in the droplets, the droplets reach their maximal number of equally oriented charges, the Rayleigh limit. This results in their fission, which marks the final release of the now ionised sample molecules into gas phase. The resulting ions, carrying potentially multiple charges, then enter the high vacuum of the MS.
As with the LC, the ion source is a potential contributor to experiment variability.

The main factors are the spray stability, impacted by instrumentation performance, delivering an uninterrupted at constant, suitable pressure, and a suitable voltage for a stable electric field. This is also where ambient ions can enter the MS as contaminants. The sample consistency and the available proton donor can impact the ion yield, some peptide species not ionising well in positive mode, but carefully supplemented mobile phase of the LC (e.g., DMSO) can boost ion recovery[109,110].

Mass Analyser

The mass analyser takes over the core part of MS, measuring the mass and charge of the sample ions by separating the ions according to their mass-to-charge ratio. Mass analysers can work as mass filters, letting pass only ions of a given mass-to-charge ratio for detection. The quadrupole mass filter for example creates an oscillating electric field created by four parallel rods, the particular ratio of voltages used to create the field oscillation defining the mass-to-charge filtered. Through the application of a combination of static and oscillating electric fields to the quadrupole, the axial motion of the ions can be confined, forming an ion trap (Fig. 2.7 c). This is useful for the analysis of different ion species in sequence with the same detector.
The orbitrap mass analyser electrostatically traps ions in an orbit around a central spindle electrode. The electrostatic attraction of the central spindle electrode is balanced by the ion's velocity, hence they enter an elliptical orbit. A barrel-shaped electrode confines the ions along the spindle axis (Fig. 2.7 f). Injected from the spindle equator, the ions also oscillate along the spindle axis. The axial motion of the ions is a solely mass-to-charge dependent harmonic oscillation. The orbitrap can therefore also be used as a high-resolution detector. The difference in image currents from orbiting ions between the electrodes can be detected by a differential amplifier. The recorded time-domain signal can be converted with fast Fourier transform (FFT) into a mass-to-charge ratio spectrum[100]. This also allows for the detection of multiple ion species at once.

Detector

The final task in MS is to detect how many ions are passing the mass analyser for a given ion species, representing the abundance of that ion species in the measured sample. As already described, the orbitrap can also be used as a high-resolution detector (Fig. 2.7 f). A more simple type of detector is the Faraday cup. Here, incoming ions collide with a detector plate, imparting their charge, inducing a current proportional to their abundance. Increasing sensitivity, electron multipliers amplify the signal intensity of the detected currents. Amplification is achieved through the use of

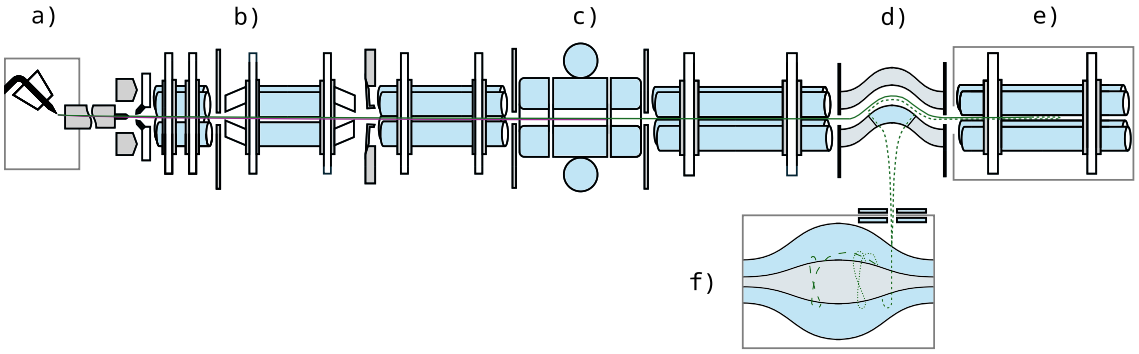


Figure 2.7: Schematic of a mass spectrometer with Orbitrap mass analyser. Ions from a) ESI travel through b) ion optics for beam focus. A initial mass selection of ions can be made in c) linear quadrupole ion trap, from which the ions can be transferred with d) C-trap into either e) collision cell for peptide ion fragmentation and back, or f) Orbitrap mass analyser.

secondary-emissive material, creating a snowball effect when used in a multi-reflection configuration. In contrast, ions in the orbitrap move in discrete orbits, i.e., they produce measurable signal not only on impact but over a period of time (seconds[100]). The extra measurement time and great precision possible via oscillation frequency measurement and FFT allow for higher mass accuracy and better resolution of signal. The signal for any detector is a pair of values per measured ion species, reflecting the mass-to-charge ratio of the ion and its signal intensity. The measurements of multiple ion species, as for example all ion species eluting from the LC at a given time, result in a mass spectrum for that time point, each ion species represented as a 'peak'. The FFT produces a continuous signal of intensity over m/z, with each ion measurement creating a signal intensity distribution centred around the m/z of the respective ion. Only after peak picking, that is the algorithmic determination of signal-peak centres and intensity, the spectrum becomes a list of discrete peaks, each of which is considered the representation of a measured ion. This also has the advantage of compressing the data to a computationally manageable set of peaks as the original spectral data still is represented as individual data points. Modern setups allow for high throughput analyses at high accuracy.

Resolution, accuracy, dynamic range, sensitivity, and speed

The resolution of a mass analyser is defined by the minimal distance at which the peaks of two intensity distributions can still be distinguished. This distance is usually defined as the full width at half maximum (FWHM) intensity of the peak. Commonly, the mass resolving power of an instrument varies with the mass range observed, so

it is given by $R = M/\delta M_{50\%}$ where $\delta M_{50\%}$ is the minimal FWHM at a certain mass (order of magnitude) M, which itself is defined by the frequency resolving power of the instrument[111]. The higher the resolving power of the instrument the narrower each signal intensity distribution gets and the more defined signal peaks of close masses are resolved. This also means better mass accuracy when the signal intensity distribution (also known as signal profile) is getting 'peak picked'[112], i.e., transformed into discrete values by picking the most likely peak centroids to represent the ions observed. Linear ion trap mass analysers are considered to produce lower resolution spectra (unit resolving power and 20-40 ppm accuracy for the quadrupole ion trap[113], where unit refers to 1 Da, the peak distance at which each isotopic peak for every singly charged ion is nominally separated) than orbitrap mass analyser (2-5 ppm), but also require about an order of magnitude fewer ions for an analytically useful spectrum[100].
The mass accuracy can be defined as the ratio of the m/z measurement error to the theoretical m/z and can be given in parts per million (ppm). In practice, identification algorithms can take advantage of precursor ions measured with high accuracy (e.g., survey scans measured in the orbitrap, to improve the identification). Over the course of measurements the mass accuracy can drift, likely to be caused by external measurement parameter changes like ambient temperature changes[114]. Internal calibration can help to keep the accuracy of measurements in a run at tolerable levels. Here, ubiquitous contaminants like polysiloxanes, deliver lock masses to which a measurement can be adjusted[114]. When necessary, external calibration with standard analytes, or software lock masses from unambiguously identified peptides[115], can also be used to reset the measurement error tolerance levels. The intensities with which the m/z values for ions are recorded in modern MS are growing linear in relation to the ion abundance only in a given dynamic range, up to four orders of magnitude[113] for orbitrap instruments. This is important when comparing abundances of substances with significant mass difference. To ensure a linear response, robust system suitability testing is key to improving reproducibility and ensuring transferable science among laboratories. Internal calibration (IC) methods use internal standards (IS) such as synthesised or recombinant proteins or peptides to calibrate MS measurements by comparing endogenous analyte signal to the signal from known IS concentrations spiked into the same sample, but single-point external calibration strategies have been reported[116]. The peptide abundances in biological samples can span several orders of magnitude, and some peptides of interest may be present in very low abundance. Parameters from the setup environment and instrument type dominate the sensitivity with which faint signals can be confidently distinguished. One set of parameters influencing sensitivity is noise; either stochastic from technical instrument interference or chemically from ubiquitous

and temporarily present contaminants. Known sources of stochastic noise are signal amplifiers susceptible to thermal noise[100] and small variations of the high-voltage power supply[117]. Chemical noise can come from sample matrix ions and ambient environment ions reaching the MS ion path. A measure of this is the signal-to-noise ratio (S/N), which can be measured by the comparison of a sample measurement to a blank measurement or estimated by taking the ratio of the mean intensity to the standard deviation of the intensities or by splitting the signal from the noise by the median intensity. On the other hand, ion transfer efficiency influences the amount of target ions reaching the detector and thus influences the S/N from the signal side. The measure of the S/N will determine the minimum signal an analyte must present to be differentiated from the noise, i.e., the lower the S/N, the higher the measurement sensitivity.

Overall Signal from One Peptide Species

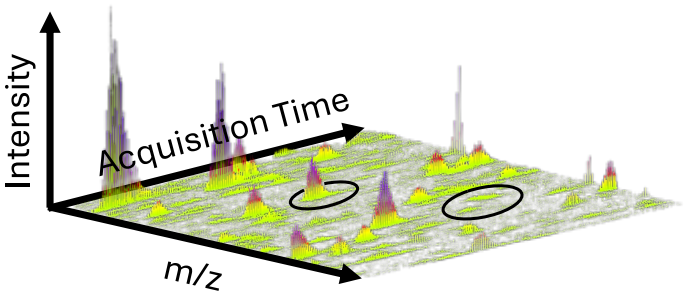


Figure 2.8: A small section of the peaks of the collective MS1 spectra. All peaks form a map, where spread elution and measurement of peptides and their isotopes produces a notorious pattern of peaks, or features of the peak map. Higher intensity is colour coded with a darker colour. Circled left is a fairly abundant peptide ion feature, circled right are two potential very low abundance features, as are common in HLA-peptide runs.

Scanning the whole (preset) mass range is called a full scan or survey scan (MS1), as it provides a picture of all ions present and measurable at the time the separated sample peptides arrive from the LC and enter the MS as ions through ESI. After peak picking, the signal peaks will still form signal distribution approximations, however on a wider scale over all MS1 combined. The isotopic variants of a peptide are generally less abundant and form a distribution around the monoisotopic peptide abundance, the shape of which can be described by the averagine model[118] in absence of the peptide's identification. Peptides will also elute from the LC over a period of time, starting with a sharp rise in abundance and trailing off, forming a peak in another dimension, the shape dependent on the gradient. These are also known as chromatographic mass traces. The two combined signal distributions of a given peptide species form a distinctive peak shape, i.e., a peptide feature on the peak map, spanned by the coordinate system of RT, mass-to-charge ratio, and intensity (Fig.

2.8). Feature detection algorithms will detect the different mass traces of a peptide's different isotopic compositions and assemble them into peak map features. Integrating the area held up by the peaks is the most comprehensive method of calculating a single representative intensity from the feature peaks, there are however alternatives, such as the extracted ion chromatogram (XIC) representation from all peaks from the monoisotopic mass trace. Detection of low-abundance peptide ions, as are common in MS runs from isolated naturally-processed HLA peptides, is challenging because of their shallow intensity distribution and most isotopic mass traces not registering above the background-noise level (Fig. 2.8, right circle highlight).

Tandem Mass Spectrometry

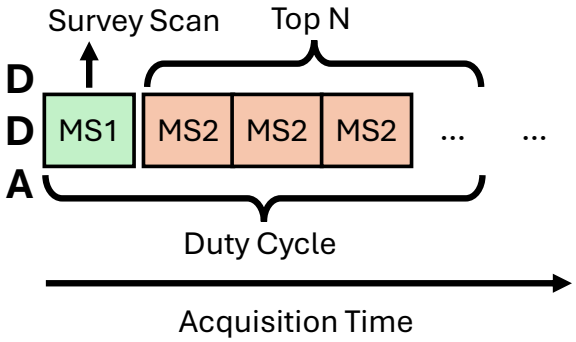


Figure 2.9: Data dependent acquisition mode schematic

Measuring the m/z of the peptide ions, yet with recent instruments very accurate, is insufficient to unambiguously identify a peptide. Tandem mass spectrometers can isolate a subset of ions from a narrow m/z window, as selected from the previous survey scan, for deeper analysis. The window size is usually chosen to select only ions from one MS1 peak, around unit size (2-4 Da) [119]. To extract more information about the peptide, or precursor, for which the ions are selected, a previously described drawback of alternative ionisation methods can be repurposed. The sample degrading properties of fast atom bombardment are unwanted at the initial ionisation stage. Their imparted energy can however be carefully controlled to break up the peptide ions selected, preferably at the peptide backbone. The selected ions are transferred into a collision cell and fragmented at a preset collision energy and time. The resulting fragment ions are then measured in a subsequent mass analysis and detection, their spectra recorded in tandem with the survey scan and thus called fragment or tandem mass spectrum (MS2). The most widely used fragmentation techniques are collision-induced dissociation (CID) [120] and higher-energy collisional dissociation (HCD) [121], but other methods such as electron-transfer dissociation (ETD) [122] are used to complement the resulting fragment ion picture by their different fragmentation behaviour. The selection of ions from the MS1 defines the tandem mode.
An untargeted selection, usually with wider isolation windows, adjacent in subsequent

scans, is used to create a fragment map of the whole m/z range of the survey scan. This mode is called data-independent acquisition (DIA) and is predominantly used for quantitative experiments, for its broad protein coverage and robust reproducibility. Due to the usually large amount of ions present in any MS1, DIA methods still segment the mass-to-charge space covered by the survey scan into windows ($\sim 25Da$[123]). The peptides' fragment ions are thus (almost) fully time-resolved, enabling DIA to quantify sample peptides with high reproducibility at the high throughput of shotgun proteomics. For the identification of peptides, still, fragment peak libraries of data-dependent acquisition (DDA)-identified peptides from the same or very similar instrument and measurement setup need to be prepared. Newer machine learning methods have been developed[124], however, instrument and data analysis requirements, as well as identification sensitivity of DIA[125] leave DDA as the method of choice for this work. For deeper insight to liquid chromatography coupled tandem mass spectrometry (LC-MS/MS) and all its applications, the reader is referred to the extensive body of review literature[126–133]. The DDA mode (Fig. 2.9) selects the n most abundant ions of a MS1 for fragmentation analysis. This obviously introduces bias to the analysis, which must be considered for the sample, instrument, and experiment type. The higher n, the more fragmentations have to be conducted, equalling more time consumed, in which newly eluting analytes might be missed. The lower n, the more low abundance ions might remain unanalysed. There are however mediating methods available (e.g., dynamic exclusion) where for some time after fragmentation analysis of a selected m/z, that mass is excluded from the top-n list. Ions can also be selected or excluded according to their charge state. As common ambient contaminants like plasticiser ions tend to be singly charged, as opposed to usually doubly or higher charged peptides, excluding singly charged from collection can prevent their fragmentation and reserve scan time for potentially better targets. Also, an operator-selected inclusion list of masses can be forced to override the top-n list if present. With DDA, broad coverage of peptide ions on MS2 level can be accomplished, though reproducibility may also depend on less controllable factors like slight abundance differences by sample preparation[134,135] introducing top-n competition[136–139]. This is prominent in isolated naturally-processed HLA peptide runs, where many low-abundant ions can be present.

The fragmentation of the ions reveals additional information on their composition. During collision, covalent bonds are broken and the ion dissociates typically into two fragments. Multiple backbone breaks give rise to internal cleavage ions. The charge of the precursor ion is distributed on the resulting ions (uncharged molecules are undetectable). The accepted nomenclature for fragment ions[140,141] is demonstrated in Fig. 2.10. N terminal fragment ions are either a, b or c; C terminal fragment ions either x, y or z. A subscript indicates the number of residues in the fragment, a superscript indicates the charge state of the ion. Where peptides fragment largely depends on the dissociation method, but may also depend on the primary sequence, the amount of internal energy, and the charge state. For CID, the majority of fragmentations occur at the backbone of the peptide, generating predominantly b and y ions; ETD generates primarily c, y, and z ions[142]. The resulting ions correspond to prefixes or suffixes of the precursor ion peptide, forming ion series, from which the sequence of the peptide can be inferred.

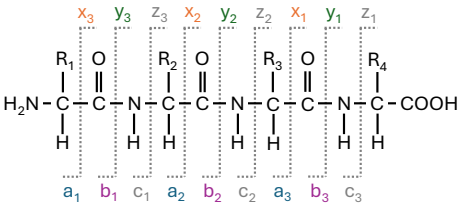


Figure 2.10: Peptides tend fragment at the peptide-bond backbone, creating characteristic ion patterns. Shown is the fragment ion nomenclature for a peptide of four generic amino acids with residues R_{1-4}. The a/b/c ions are enumerated starting from the N-terminus, the x/y/z ions from the C-terminus.

Applications to the annotation of identified spectra can be found in Chapter 3 (Fig. C.1) and Chapter 7 (Fig. 7.3).

Quantitative Methods

Quantification of proteins in a biological sample is, next to the identification of said proteins, one of the key tasks in proteomics that can be accomplished with MS. Due to the necessary sample preparation, only peptides can be measured reliably in complex samples. Unique peptides can stand for the parent protein quantities, however, the assignment of peptides to proteins is not always uniquely possible, and peptides may need to be aggregated at protein-group level by protein inference[143,144]. A common method is the relative quantification across samples. Here, quantification techniques come as labelled and label-free approaches.

The labelled approaches include the labelling of the samples, either in vivo through metabolomic labels, or in vitro, with chemical labels, attached to the prepared peptide samples before MS. The chemical labels are usually designed as isobaric tags, attaching to the peptides. Upon fragmentation as preparation of a tandem spectrum measurement

of a labelled peptide, the label dissociates into a neutral padding mass and a reporter ion with a mass specific to the label type. The reporter ion peak intensities represent the relative abundance of the labelled peptides. Because each sample gets a different label type applied, the samples can be mixed for one single measurement. This method is called multiplexing and can avoid instrument-related run-to-run measurement variations, improving the precision. Another benefit is that the quantification is done on exactly the same spectra as the identification, reducing data analysis complexity. The ion suppression properties of the chemical labels can however affect the accuracy of the quantification.

Multiplexing is also possible with metabolomic labels, where stable isotopes are introduced to the growth medium of a particular sample group (e.g., for an experimental condition). More precisely, the amino acids in the growth medium are synthesised with stable isotopes, such as ^{13}C or ^{15}N, which results in a computable mass shift in the peptides of the labelled sample. This technique is economically feasible only for cell cultures, and in extreme cases small animals[145], due to the amount of necessary stable isotopes and purity requirements of the synthesised amino acids, and fittingly called stable isotope labeling by amino acids in cell culture (SILAC). Peptides from different samples are thus distinguishable by mass, otherwise chemically identical, i.e., are expected to elute at roughly the same time and display the same ionisation response. They can therefore be multiplexed into one measurement and compared without the assumption of a label bias to the quantification. SILAC is an MS1-level quantitation technique, as opposed to the previously described techniques, where the quantities were derived from the MS2 level. For best quantitative measurement fidelity on the MS1 level, the peptide features must be detected and their constituent peak intensities included in the quantification. They can then be linked with MS2s identifications, further linked into feature tuples of the same RT, identification and expected mass shift. From there, it is straightforward to calculate their peptides' fold change between conditions or save them for further analysis.

Without labelling, no multiplexing is possible, but comparing peptide quantities between several biological samples is still possible. Label-free quantification is an MS1-level based quantitation method, and therefore the peptide features need to be detected and identified, however now also need to be linked between the individual MS runs for each sample. An advantage of this is that it scales well to large numbers of samples and experiments, compared with labelled methods, where the number of different samples is limited by the complexity of the label, e.g., 8-16plex for TMT, and limited in the number of experiments by the costs of the label. Successful label-free feature linking relies on the algorithms to correctly link corresponding peptide features. Algorithms have to

compensate for chromatographic RT variation between runs and other variations such as in instrument performance and measurement conditions subtly shifting over instrument operation time (n.b. runs over a study are not necessarily measured successively).

2.4 Computational Mass Spectrometry, Immunoinformatics, and Integrative Bioinformatics

Running long gradients and acquisition of high-resolution mass spectra yields an enormous amount of data that needs to be processed efficiently. Various instruments and experimental techniques are employed in current research in order to answer complex biological questions. This variety in experimental setups raises the need for highly flexible and efficient computational approaches to analyse experimental data.

2.4.1 MS Data

The variety of experimental setups is reflected in the data. Mass spectra are commonly recorded in continuous profile mode and can be converted as described to discrete peak data before storage. Depending on the instrument settings various contextualising metadata can accompany each spectrum, e.g., trap fill times or fragmentation energy used. Modern instruments create thousands of spectra per run, themselves containing hundreds to thousands of data points (peaks), and this abundance is reflected in the analysis data for identification and quantification. Scientific and commercial software vendors for proteomics data analysis, as well as the data producers themselves, quickly realised the benefits of common data standards for compatibility. With the mission to facilitate community-driven standardisation in proteomics data reporting, the proteomics standards initiative (PSI)[146,147] was formed from within the Human Proteome Organization (HUPO). PSI has created data standards and reporting recommendations for consistent and concise data publication, sharing, and archival. Widely recognised, the mzML[148] format for storing measurements from a MS experiment, builds the basis for the common exchange and storage formats the PSI has developed. As part of this work, more data standards for quality control, identification, and quantification analysis will be described in more detail in Chapter 3.

2.4.2 Quantification

The mass spectrometer, sometimes called a molecular scale for its ability to measure the mass-to-charge ratio of ions is best suited for quantitative proteomics. The quantitative

methods described previously depend on the correct detection of peptide features, and a representative abundance estimation from the peak intensities involved. This can be done by assembling the peaks of consecutive MS1 spectra that can be attributed to an isotopic variant of a peptide. Should no identifying information be available at time of analysis or kept independent on purpose, features can still be detected. As described, the general feature shape is defined by the elution profile and the natural isotope abundance in the amino acids. Without internal or external standards to calculate a calibration curve over intensity and mass ranges observed, conversion into concentrations is not feasible. But relative quantitation between different conditions can be established, if samples of the conditions are measured, and fold-ratios for the identified quantitative entities matched between the runs are calculated.

2.4.3 Identification

The quantification of proteins is a key task for mass spectrometry-based proteomics, but depends on another task, the identification of proteins, for the correct interpretation of the acquired quantitative values. The described fragmentation of precursor peptide ions in MS2 reveals the subtractive masses of the amino acids in sequence per fragment ion type (a/b/c, x/y/z) against the peptide precursor mass. With complete ion-type mass ladders, the peptide sequence can be inferred without further input of information. Algorithms for de novo identification have to deal in practice however with incomplete fragmentation, noise peaks, internal fragmentation peaks, and the combinatorial burden that comes along with many types of peaks that have to be considered.
A more robust method for the identification requires additional information in the form of a sequence database.

Database Search

Sequence database search engines for fragment spectra leverage the information gained from creating theoretical spectra from database sequences to match against the experimental spectra. The protein sequences in the given database are in silico digested using the specific cutting rules for the protease used in an experiment. Unspecific digestion is used if no clear-cutting rules can be applied (e.g., for HLA immunopeptides). For each experimental spectrum, the sequences can be filtered for theoretical masses matching the precursor ion at given tolerance. Further mass-influencing conditions of the samples' peptides need to be considered in the theoretical spectrum generation, like the presence of PTM, simple oxidations, or chemical modification introduced by

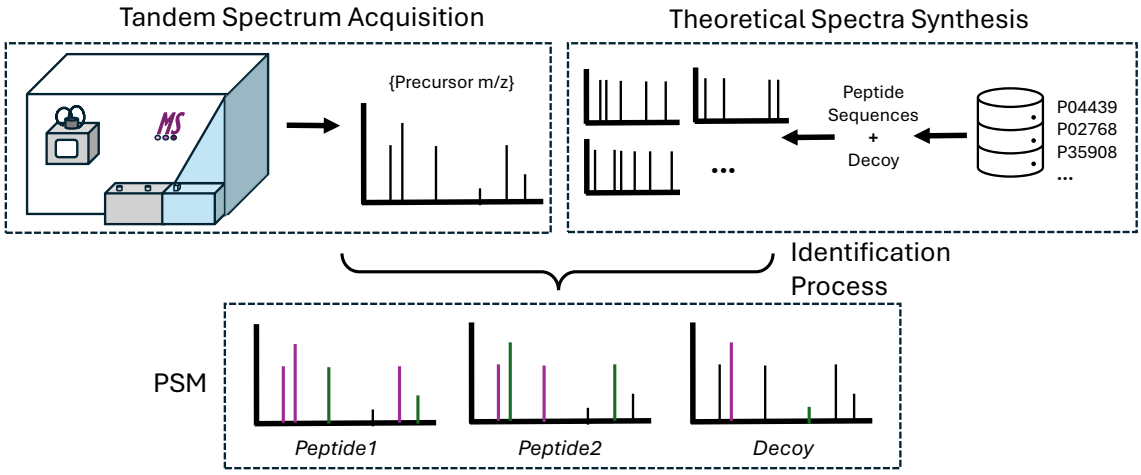


Figure 2.11: Database-driven peptide identification process from fragment spectra

the sample preparation. Theoretical spectrum generation can also be tailored to the mode and energy settings of the fragmentation used. The match between candidate theoretical spectra and experimental spectra is scored with a search engine-specific score[149–152] and reported as peptide-spectrum match (PSM). One obvious drawback is that sequences have to be known and included in the database to be identified, which can be resolved with additional genomic sample analysis. A further drawback is the statistical error rate when making many comparisons against the many fragment spectra included in a common MS proteomics run, further described in the next section. Another approach is to use databases of identified spectra, so-called spectral libraries, to transfer identifications to new spectra. Difficulties for the approach lie with the transferability of spectral libraries between MS instrument types, and mainly with the incapability to detect novel peptides, indispensable for individualised proteomics and immunopeptidomics.

False Discovery Rate

The problem with the sequence database search approach is that it compares multiple candidate theoretical spectra to each fragment spectrum, and will do so for each spectrum in the run. This makes the approach a case of the statistical multiple-testing problem. Conducting a statistical test, one can choose a significance level (α) to set a threshold for the risk of making a type I error (false positive conclusion). The test statistic would then be calculated from the observed result and the value distribution under the null hypothesis (e.g., no relationship between the peaks observed and the sequence matched). The resulting p-value is the probability that the test statistic under the null hypothesis produces values at least as extreme as the observed value.

Given the chosen α, the null hypothesis is then either accepted ($p > \alpha$) or rejected ($p \leq \alpha$). Further, because the p-value is the probability to observe a result at least as extreme in one test, with more independent tests (e.g., different sequences matched to a spectrum) conducted, the likelihood of random matches (false positives) increases. Several methods to control this error rate for multiple tests exist. Controlling the family-wise error rate (FWER) is to control the probability of making one or more type I errors, among all tests. The most conservative method for FWER control is the Bonferroni method, choosing an acceptable FWER over all tests, and dividing it by the number of tests yields the significance level for each individual test to keep the FWER below the chosen value. Choosing the right value distribution for a PSM however is not simple, given that:

- the search space from the database does not necessarily cover all peptide species in the sample measured
- spectra derived from nonpeptide-ion species may be assigned database matches
- no perfect match scoring is available, hence incorrect peptide sequences may outscore correct sequences

Also, controlling the FWER comes at the expense of statistical power, guarding against the occurrence of false positives leads to many missed findings, which for identification purposes may be unwanted. The false discovery rate (FDR) was developed[153] to only control the expected proportion of falsely rejected null hypotheses (false positives). The FDR is the expectation of an unobserved (unknown) random variable, the proportion of the rejected null hypotheses which are erroneously rejected. The positive FDR (or pFDR) can be expressed as a Bayesian posterior probability

$$FDR = E\left[\frac{FP}{P}|P > 0\right]$$

where FP is the number of false positives and P the total number of positives, given there is at least one rejected hypothesis. A measure of significance can be assigned to each statistic, the q-value[154], which is similar to the p-value, but incorporates a multiple testing correction under the pFDR. It is also much easier to calculate for PSM with the 'target-decoy' approach[155], avoiding a direct estimation of the value distribution under the null hypothesis. In the target-decoy approach, it is assumed that random matches (false positives) should occur with similar likelihood in a target database as in a decoy (reversed, shuffled, or otherwise randomised) version of the same database. Thus, the number of random matches in the target database can be estimated by the number

of matches in the decoy database. For reversed-decoy databases, it can generally be assumed that there is a 1:1 correlation between target and decoy sequences. Cutoff values from the ranked PSM (using an arbitrary identification score) can thus be chosen for a targeted FDR (e.g., "a score threshold of 4.14 yields 4 decoy PSMs and 919 target PSMs, implying an FDR of 0.35%"[156]). The FDR as a function of the scoring will not likely be strictly monotonous as there will be a range of scores for a given number of decoy PSMs. Empirical q-values can be calculated from the ranked PSMs best to worst (from both target and decoy databases, only the highest scoring match per spectrum), reflecting the minimal FDR with a certain score or worse. Each q-value is assigned by dividing the cumulative number of decoy matches by the number of cumulative target matches in order from best to worst. Controlling the FDR at a given percentage threshold $\hat{\alpha}$ is then cutting of the list at q-value $> \frac{\hat{\alpha}}{100}$.

2.4.4 Analysis Tools, Frameworks, and Workflows

Given the number of tasks described that need computational analysis with dedicated algorithms, there is a thriving and diverse ecosystem of software, providing solutions. As seen in the previous descriptions of different methods and analysis approaches, many challenges share common features or data input needs. It is for this shared basis of tasks that common frameworks were built to reduce development redundancy, pool development and maintenance efforts in scientific software, and enable the development of new algorithms with code infrastructure for common tasks already in place. Here, we introduce two open-source frameworks, one for computational MS and one for immunoinformatics, that were crucial for the work presented in this thesis.

OpenMS

As seen in the previous descriptions of mass-spectrometry methods, there are numerous common core tasks, like the discovery of peptide features in peak maps or the control of the FDR in a set of identifications, for any particular MS method's analysis. From a data analysis development standpoint, too, there are many common structures that need to be handled efficiently, like reading spectra and handling large amounts of peaks. To allow flexibility in data analysis, it is advantageous to split computation into manageable steps, for parallelisation of analysis for large datasets with many individual runs, and the combination of different methods to design an analysis workflow tailored to a specific type of experiment.
OpenMS is an open-source development framework that unifies these three mainstays

of computational MS. Reflecting the described conceptual aspects, the framework operates on three layers:

1. The OpenMS core library is a C++ library that provides functionality for the development of new applications through available implementations of established analysis algorithms and for the development of novel algorithms through common data structures and a testing framework.

2. The OpenMS Proteomics Pipeline (TOPP) provides ready-to-use applications for common analysis steps and makes it easy to compose new applications using the core library to create specialist applications for data analysis from new MS methodologies.

3. The OpenMS workflow integrations allow the chaining and orchestration of analysis steps for the design of custom data analysis on large datasets.

OpenMS[157] saw its start in 2003 as an academic initiative led by Prof. Knut Reinert (FU Berlin) and Prof. Oliver Kohlbacher (EKU Tübingen). As the name implies, OpenMS is based on the open-source software model. It is licensed under the permissive 3-clause BSD license, which allows to freely use OpenMS in both academic and commercial contexts. The openness is also reflected through the broad support of operating systems (all major operating systems Windows, MacOS, and Linux).
As a C++ library, the OpenMS core offers a broad suite of functionality and data structures to computational biologists and bioinformaticians with sufficient programming skills to efficiently develop new algorithms. The implemented data structures are designed to work flexibly with MS data (proteomics, metabolomics, chemistry) and adhere to open data standards for reading and writing mass spectrometric data, and analysis results. The OpenMS core library makes use of other external open-source libraries to leverage specialist solutions to computational problems such as machine learning (libSVM), optimisation (Coin-OR), and visualisation (Qt). Providing over 1,300 classes, the library supports an extensive set of mass-spectrometry related representations (peaks, spectra, mass traces, chromatograms) and tasks (algorithm implementations, network and file I/O, GUI). With the pyOpenMS bindings for Python, OpenMS can speed up application development and integration with other software (i.e., software that can interface with Python such as R). Efficient algorithms for signal processing, identification, quantification, data filtering and visualisation provide the basis for TOPP tools. These are command-line applications for well-defined steps within the analysis, like feature-finding, the composition of a decoy database, or the filtering of identification data at a given FDR level. MS is a highly dynamic field, with new instruments

frequently introduced by manufacturers and new experimental methods published in rapid succession. Adapting data analysis sometimes means changing the sequence and type of analysis steps, sometimes the subtle changes to some analysis functionality, and sometimes the implementation of novel algorithms to solve conceptually new analyses. OpenMS is designed in a modular fashion, which allows flexible analysis designs with a high degree of customisation for every occasion. One of the main aspects that afford OpenMS the analysis design flexibility with the split of functionality into TOPP tools is their ability to efficiently hand over data through the adoption of data standards throughout the framework. Another aspect is the TOPP tool integration into workflow orchestration systems like Konstanz information miner (KNIME) and Galaxy, which will be described in more detail in Chapter 6. With these it is possible to rapidly design new workflows, flexibly add post-processing scripts and visualisation, and automate analysis on a large scale.
A focus on reusable code and a breakdown into common tasks has given the large base of OpenMS developers some ease of maintenance over two decades of development.

FRED2

With the increased understanding of the complexities of the immune system and more experimental data available, over the last decades, the availability of software modelling aspects of the immune system, e.g., HLA-epitope prediction, antigen processing, and vaccine design has increased. To enable large-scale analyses, the integration and evaluation of methods and models was necessary. To this effect, the first iteration of the framework for epitope detection (FRED) was developed. Written in Python, it had a modular design for the integration of external methods. To leverage the full advantage of a tool development framework on the basis of FRED, as seen with OpenMS, FRED2 was developed. Immunological and biomedical research requires multiple tools and data sources for effective data analysis workflows. FRED2, as an open-source immunoinformatics framework, offers access to data and immunoinformatics applications under a unified framework. With Python as its implementation language, it provides a low entry threshold for developers and supports the development of complex applications through a massive developer base and choice of 3rd party open-source modules. Through a focus on common data structures and function interfaces, the interoperability of tools in FRED2 is much improved and makes the implementation of new algorithms less complex. The permissive three-clause BSD license allows for it to be freely used in both academic and commercial contexts. File and database adapters allow for a rich variety of data sources as input to data analysis. The integration of

external applications and workflow orchestration integration make it a powerful tool to design and adapt complex analysis workflows for a wide variety of applications. As part of this work, FRED2 will be described in more detail in Chapter 5.

Workflow Orchestration Systems

The utility of combining tools for specific tasks in a data analysis was already described in the OpenMS section and is similarly true for FRED2, but applies to data analysis in general. The requirement to make inputs and outputs compatible between tools of a workflow is a tall order, but the benefits are as tall in return. With workflow orchestration systems, not only is a flexible and uncomplicated design of workflows possible, but the automated analysis of datasets at unprecedented scale. Galaxy[158] provides user-side graphical construction of a workflow of genomics tools in a web browser, and server-side data access and tool execution. Availability of powerful compute infrastructure was an early requirement for genomics data analysis, as many of the first processing methods as well as the size of single run datasets would overwhelm most personal computers at the time. The Konstanz Information Miner[159] (KNIME) is a data analytics platform for visual design of workflows in a modular fashion. Workflows can be run on local hardware as well as on remote compute clusters. An open source community behind KNIME offers a rich environment for data analysis, from statistics and visualisation to machine learning, and domain-specific data analysis tools beyond proteomics. A unifying layer to different workflow systems' tool access can be provided by the Common Tool Descriptors (CTD)[160] to increase tool compatibility. As part of this work, KNIME and its integrations will be described in more detail in Chapter 6. Other workflow orchestration systems place less preference on visual workflow design aids and address a more code-oriented audience. Nextflow[161] workflows are implemented through a declarative language, and the input and output connections between tools have to be handled explicitly, therefore a robust knowledge of all tools involved in a workflow is required. Instead, a deeper focus is placed on the integration of, and easier access to different (cloud) compute infrastructures and software provision through containers.

Chapter 3

Data Integration for Automated Workflows

This chapter includes partially identical or adapted content with permission from:

The mzIdentML Data Standard Version 1.2, Supporting Advances in Proteome Informatics

Vizcaíno JA, Mayer G, Perkins S, Barsnes H, Vaudel M, Perez-Riverol Y, Ternent T, Uszkoreit J, Eisenacher M, Fischer L, Rappsilber J, Netz E, Walzer M, Kohlbacher O, Leitner A, Chalkley RJ, Ghali F, Martínez-Bartolomé S, Deutsch EW, Jones AR.

The mzIdentML data standard version 1.2, supporting advances in proteome informatics. Mol. Cell. Proteomics 16, 1275–1285 (2017); `https://doi.org/10.1074/mcp.M117.068429`

The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics.

Walzer M, Qi D, Mayer G, Uszkoreit J, Eisenacher M, Sachsenberg T, Gonzalez-Galarza FF, Fan J, Bessant C, Deutsch EW, Reisinger F, Vizcaíno JA, Medina-Aunon JA, Albar JP, Kohlbacher O, Jones AR.

The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. Mol Cell Proteomics. 2013 Aug;12(8):2332-40.; `https://doi.org/10.1074/mcp.O113.028506`

The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience.

Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, Hartler J, Thallinger GG, Salek RM, Steinbeck C, Neuhauser N, Cox J, Neumann S, Fan J, Reisinger F, Xu QW, Del Toro N, Pérez-Riverol Y, Ghali F, Bandeira N, Xenarios I, Kohlbacher O, Vizcaíno JA, Hermjakob H.

The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol Cell Proteomics. 2014 Oct;13(10):2765-75.;
`https://doi.org/10.1074/mcp.O113.036681`

3.1 Introduction

Motivation

Life sciences, and computational MS, in particular, are data-intensive sciences. And the growth of data has so far followed an exponential trajectory as evidenced by the amount of experimental mass spectrometry-based proteomics data submitted to the proteomics identifications database (PRIDE) archive[162] since 2012, now amounting to several thousands of terabytes of data. Clearly, automated data analysis is necessary to leverage the insights possible from the data accumulated. Experiment data analysis is dependent on stable data in- and output to run automated and reproducible in different computational environments or workflow setups. To any workflow, both primary inputs and intermediate result files, the use of open standards to represent the MS (analysis) data provides great benefits to all parties. Open community standards offer data representation to facilitate exchange, comparison, and verification. They simplify, or at least stratify implementation efforts for analysis tool vendors to reach a wider audience and maintain compatibility with other tools. Fortunately, for MS and proteomics data, there are usually various alternatives of analysis tools to accomplish a given task within a data analysis workflow. The decoupling of raw MS data from vendor-specific formats with mzML was a first step to unlocking many data analysis possibilities for data producers. The next steps for a robust and thriving data analysis environment are to integrate identification and quantification data analysis based on the raw MS data into the standards ecosystem and unlock data analysis possibilities in a similar amount. However, the data derived from identification and quantification analysis is more diverse than the raw MS data. Identifications are dependent on external data input, there is a multitude of different quantification methodologies and they depend at least to some extent on prior identification, and both identification and quantification include possibly multiple levels of statistical control that need careful documentation. The diversity is what actually makes data standardisation so appealing. When using the data in standard format as a handover medium between specialist tools, the user does not have to rely on one tool to provide all (e.g., using different methods of decoy generation, applying FDR control on different levels, or applying statistical post-processing), so the data analysis workflows can become more powerful as their possibilities grow with tools integrated. Bioinformatics deals with data from all the life sciences and has matured to a point where for most steps in any established analysis, a specialised tool or function library exists. In the right circumstances, these tools can be combined to conduct the complete analysis without the need for case-by-case software

development. Data standards provide these right circumstances.
Reisinger et al.[163] assessed in 2008, that during the last decades, various fields of molecular biology have seen the adoption of high-throughput methods and to this end, data analysis was faced with a growing collection of very heterogeneous types of data. Since then, a dedicated community has formed from within the HUPO addressing the topics of data format standards, and controlled vocabularies for mass spectrometry-based proteomics.

Background

The mission of the PSI[146,147] of the HUPO is to facilitate the community-driven standardization in proteomics data reporting. One of the groups' earliest open data standards, the mzML[148] format for storing a MS experiment's information, is implemented in the vendor-neutral extensible markup language (XML) data format. Its cross-community acceptance has shown wide-reaching success, not the least because of its capability to flexibly capture important metadata in an easily accessible fashion. Building on the proven structures used to design mzML, and incorporating components derived from the Functional Genomics Experiment data model[164] (FuGE), mzIdentML[165] and mzQuantML[166] have been developed to facilitate the convergence of data standards for high-throughput, comprehensive analyses in biology.
Other preceding open formats for MS identification data also influenced the mzIdentML standard design. For example, the pepXML format[167] was developed at the SPC/Institute for Systems Biology for the common storage, exchange, and processing of different MS/MS search engines and subsequent peptide-level analyses from software tools mainly from within the Trans-Proteomic Pipeline[168] ecosystem. OpenMS'[169] XML representations of data structures for the TOPP tool internal handover of identification data, idXML, also contributed conceptual inspiration to the mzIdentML design. Likewise did the other OpenMS XML representations of data structures for quantitative data, featureXML and consensusXML, to the design of mzQuantML. For quantitative data, text-based analysis-specific formats, dominate the mode of data exchange. Like the quantitation report of mascot server, which provides quantitative information on a high level without evidence trace and focuses mainly on labelled quantification methods. Also, existing XML-based formats like OpenMS' featureXML, had their final destination as a text-based export (via the TOPP tool `TextExporter`) to feed into statistical software for final, high-level analysis. The technically very different needs of reporting and exchanging quantitative data versus archival with exhaustive evidence trail and step-by-step analysis documentation gave rise to the more condensed design of mzTab. Here,

the dominant role of identifications to the quantification data inspired the design of mzTab to function as a compact format for both identification and quantification data.

3.2 The mzIdentML Format for MS Identifications

The mzIdentML format has also proven effective in avoiding considerable wasted effort in writing custom data adaptors to keep data analysis workflows compatible with rapidly changing proprietary formats through widespread adoption in many MS identification software and analysis workflows. To keep the format compatible with the latest identification methods and to improve usability, its format specification has been updated. This section describes our work, based on the initial mzIdentML format specifications (v1.0 and v1.1), implementing its use within OpenMS, and to refine general usability of the format, adding support of additional use cases with mzIdentML format specification (v1.2)[170].

3.2.1 Methods

The methods section is split into two subsections of design rationale and structure (of mzIdentML) to better reflect the substance of developing, discussing, refining, and implementing file formats.

Design Rationale

Two previous versions of the mzIdentML format have been released, with each iteration, improving existing, or introducing new use cases. Its main design goal is to provide "systematic descriptions of polypeptide identification and characterisation based upon MS"[i] (v1.0), i.e., to represent the outputs of proteomics search engines.
Several design rationales guided the specification processes behind its main goal:

1. To provide a format for sharing identification data and to serve as a unifying data representation layer for the many search engines available, the structure needs to be flexible enough to accommodate the identification results from a wide variety of identification approaches.

2. To provide sufficient result detail depth to serve as archival and data-sharing format, an evidence trail to the input data of the identification process needs to be provided, therefore the format needs to work closely with mzML, but also vendor-specific formats.

[i] mzIdentML specification documents: `https://www.psidev.info/mzidentml`

3. The details of identification need to be represented for spectrum identification and subsequent identification processes (e.g., protein inference). The format needs to hold the results from an identification data analysis with the primary results being the scored spectrum identifications, but also groupings formed thereof as done in protein inference.

4. Multiple search protocols may be applied to a particular spectra file, and the identification of multiple files may be considered together (e.g., with a pre-fractionated sample), the format thus needs to capture both application of different search protocols, and the application to several spectra files.

5. To enable result evaluation and improvement, the format needs to hold multiple scores for an identification process, represent combined identifications with meta-scoring, and report cut-off values and selection criteria for a final report, and software configuration details, therefore also acting as a format to share best practices.

From the start of the design process, it was clear that not all use cases involving MS identification data could be covered in a first version[i] (see v1.0, v1.1 for initially developed use cases), and extensions to the format must be made to accommodate important emerging use cases and new methodological developments. Following, new support scenarios have been added with mzIdentML (v1.2)[i].

- Represent the identification results from different experimental techniques, including for example different labelling techniques or cross-linking.

- Store the results of MS/MS cross-linking approaches, whereby two peptides cross-linked using chemical reagents or biologically occurring modifications have been identified.

- At a basic level of detail, also store the molecular interaction data that can be inferred from cross-linking approaches.

- It should be possible to represent statistical values or scores associated with the positions of modifications on a polypeptide's chain of amino acids.

- It should be possible to represent the statistical values or scores associated with peptide identifications, formed from groups of redundant PSMs reporting on the same peptide.

Structure

Following the design rationales, information in mzIdentML has been divided into multiple elements under the document root. These can be roughly categorized into four types (Fig. 3.1). The first is a collection of metadata for provider details, bibliographic

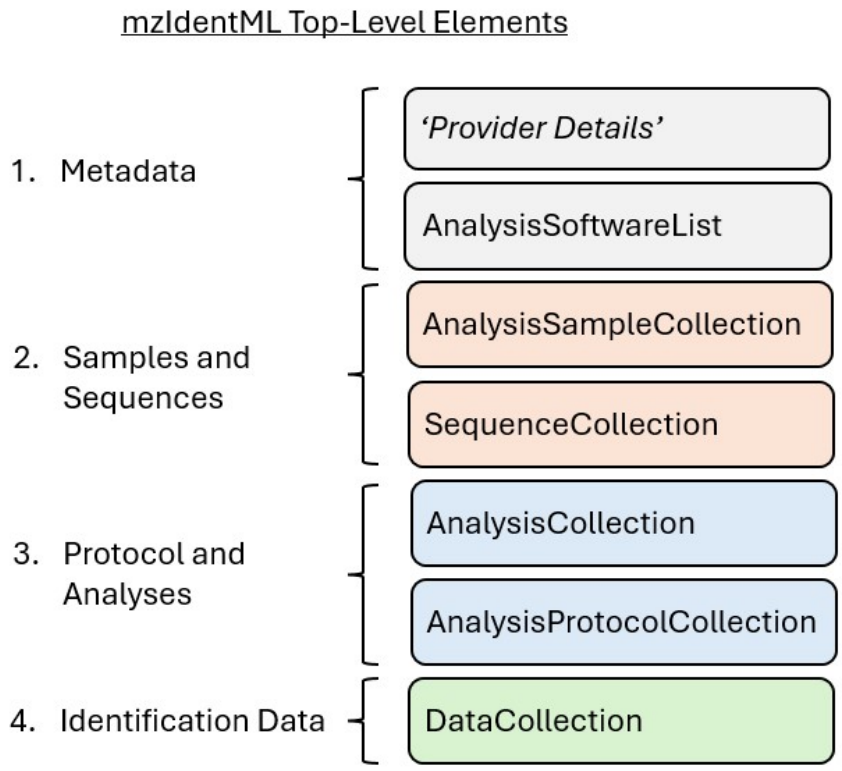


Figure 3.1: Overview of the top-level elements of mzIdentML with categorical grouping. The meta-data section also includes top-level XML attributes, indicated here as provider details.

references, controlled vocabularies of terms used in the document, and software used in the analysis. In the next category are lists of sample descriptions that were measured and identified, and amino acid sequences plus their sequence database entries used for identification for later referencing in the next categories. The elements of the third protocol-and-analyses category connect input data and search engine results within the context of the applied search protocol, and finally, a fourth identification-data category holds the details of input data and search engine results (i.e., spectrum matches).
From the rationales above, it becomes evident that the format must handle a high level of interconnectedness, both within the data of one file and between different files (of potentially different PSI formats). As a result, the elements within the structure of mzIdentML have a mandatory W3C XML[171] Schema Definition `@xsd:id` attribute,

giving each element a referenceable anchor, unique within each file. To allow for efficient storage of multiple identifications, individual elements of the four categories (1. metadata and software, 2. samples, and sequences, 3. protocol and analyses, 4. identification data) have subordinate collection elements according to their type and are then only referenced throughout an mzIdentML instance to reduce information redundancy.

The bulk of identification results is kept within the fourth category, which is comprised

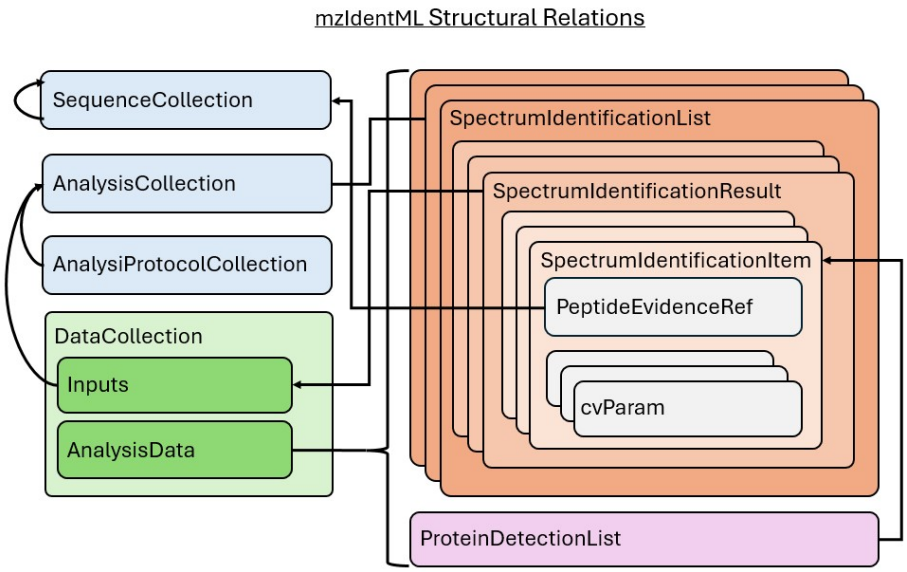


Figure 3.2: The mzIdentML format stores individual aspects of the analysis in distinct collection elements and uses extensive referencing to minimise data redundancy. The figure shows the elements involved in reflecting the main results of the identification analysis and their position within the formats XML schema. Arrows indicate references in format elements to the details of other elements. Nested element representations indicate possible multiple occurrences within a mzIdentML instance. PSM are recorded via `<SpectrumIdentificationItems>`, referencing `<PeptideEvidence>` elements and scoring information via `<cvParam>` elements, and are contained within `<SpectrumIdentificationResult>` items, referencing external files' (preferably from mzML) spectra. The `<SpectrumIdentificationResult>` elements themselves are collected in a `<SpectrumIdentificationList>`, representing one 'identification run', the details of which are collected by the `<AnalysisCollection>` elements, referencing to the respective `<SpectrumIdentificationList>`. Protein identifications are stored in one `<ProteinDetectionList>` and handle referencing the evidence trail in an equivalent fashion.

under a collective parent element, `<DataCollection>` (Fig. 3.2). Within, the `<Inputs>` element holds subordinate elements for detailing input files contributing to an analysis, like search database and spectra input, or the original search engine result file. The latter is then regarded as input to the creation of the mzIdentML instance if mzIdentML

is not natively supported by the search engine and made compatible through an adapter. However, only spectrum input and search database specifications are mandatory.

On the same level as <Inputs>, the <AnalysisData> element keeps the spectrum identifications as one or more <SpectrumIdentificationList> elements, containing <SpectrumIdentificationResult> elements, which reference a specific spectrum and, in turn, contain the amino acid sequence matches to its referenced spectrum (or PSM), represented as <SpectrumIdentificationItem> elements. These are referring with a @peptide_ref attribute to the sequence and are adding corresponding confidence scores represented as subordinate <cvParam> elements (defined by controlled vocabulary (CV) terms). Likewise, protein detections are recorded within a <ProteinDetectionList> of <ProteinAmbiguityGroup> elements, which will contain any protein inference results as <ProteinDetectionHypothesis> elements. These give evidence for any one database sequence listed in the <SequenceCollection>. They reference the sequence through the @dBSequence_ref attribute directly and any <SpectrumIdentificationItem> in the file implicitly through a list of subordinate <SpectrumIdentificationItemRef> referencing these (Fig. 3.2).

The <SpectrumIdentificationResult> element references its spectrum with both a reference to the spectra data from <Inputs> and a reference to the source peak file's spectrum itself by a referencing index. The type of index is defined in the element for the spectra data by a CV term (within the <SpectraData> subordinate <SpectrumIDFormat>). The 22 types of encoding of the individual references for different types of spectra data are defined in the specification document.

The <Peptide> elements referenced by a <SpectrumIdentificationItem> @peptide_ref attribute are collected in <SequenceCollection> (second category) and define a specific amino acid sequence (with modifications if applicable). Also collected in <SequenceCollection> are <PeptideEvidence> elements, which link the peptide to a <DBSequence>, also within <SequenceCollection>, via references. As the name implies, <DBSequence> elements can reflect the search engine's search database input, or more condensed, only those items with search matches. The database items optionally include the complete amino acid sequences of the protein. With this construct of references, the input spectra, identified peptides, and proteins can be linked together, maintaining any possible 1-to-many relationships.

The <SpectrumIdentificationList> and <ProteinDetectionList> elements are connected with the respective <SpectrumIdentificationProtocol> or <ProteinDetectionProtocol> as paired references contained in one <SpectrumIdentification> or <ProteinDetection> element, located under <AnalysisCollection> (third category). As such, they represent a concrete application of a spectrum identification to a particular set of spectra. The

protocol elements in the <AnalysisProtocolCollection> hold detailed information about the identification process: the parameter values used in the identification process, and reference to the respective <AnalysisSoftware> element under <AnalysisSoftwareList>. Further details necessary for reproduction can be specified in the <AnalysisSampleCollection> element. Here, a description of the biological samples used in the analysis can be given with CV terms, albeit optional. Finally, the format needs further important metadata elements to be complete (first category). These are the list of controlled vocabularies, which terms are used within a standard format file, the list of analysis software and versions used to produce the included results as the <AnalysisSoftware> elements, and root element attributes describing provider details. This design allows for the versatile combination of input data and analysis method combinations to be recorded and enables for faithful reproduction of identification analyses.

3.2.2 Results

Added to the mzIdentML specification (v1.2) were several specialist use-case descriptions. These, and the refined specification, were developed over multiple HUPO-PSI meetings and regular teleconferences of the mzIdentML task-group members of PSI. Two use cases are described in the following, explaining how, without intrusive changes to the original schema, these can be accommodated in mzIdentML. Software consuming mzIdentML must be made aware of the presence of any of the specialist use cases, the files must be marked by a mandatory occurrence of a labelling CV term in the <SpectrumIdenticationProtocol> section.

OpenMS Implementation

We implemented the mzIdentML format within OpenMS as XML stream handler to fit into the OpenMS ecosystem of computational analysis (de)serialisation. The design updates for the cross-linking and modification location use cases, described in the next section, were included in the implementation. With this, mzIdentML files can be loaded into the internal data model of identifications in OpenMS. The implementation overcomes the frequent cross-section referencing in mzIdentML through delayed intermediate construction of OpenMS identification object structure. Conversely, identifications from the internal model can be stored in mzIdentML format through the implemented handler, essentially building a converter from and to other OpenMS-supported identification formats. With the derivation of the mzIdentML handler from the OpenMS core XML handler also comes schema validation functionality, whereas

the check for semantic validity was implemented in a separate class derived from the OpenMS `SemanticValidator` to complete mzIdentML file validation. TOPP tool developers can choose between different identification object (de-)serialisation interchangeably, including mzIdentML, to facilitate algorithmic in- and output with compatible interfaces within OpenMS.

Modification Location

One design update made was the inclusion of reporting modification localisation scoring to the specification. For that, a CV term "modification localization scoring" (MS:1002491) needs to be used as a use-case label in the protocol section.
The specification extension allows description of the score for each position of a modification in `<SpectrumIdentificationItem>`. Within, the modification scores are added as CV terms as other identification scoring terms would be (Listing 3.1), with its `@value` attribute encoding following the schema of `{modification index}:{score value}:{position within the peptide}:{threshold filter passed}`. The position can include a logical "OR" as `"|"` if the score relates to multiple positions. For example:

Listing 3.1: A modification location score CV term encoding example

```
1 <cvParam accession="MS:1002380" cvRef="PSI-MS" value="1:0.03:2|3:
    ↪ true" name="modification␣rescored␣by␣false␣localization␣rate
    ↪ "/>
```

The score type depends on the specific type of CV term (child terms of "modification position score" MS:1002506) used, the position of the modification is indexed the same as modifications before the update in `<Peptide>` elements (N-terminus = 0, C-terminus = peptide length + 1), and lastly a simple boolean argument (true | false) for a chosen threshold passed, true by default if no threshold was specified.
The `<Peptide>` elements, too, must be supplied with another CV term, "modification index" (MS:1002504) if variable modifications are present in the respective peptide. This is the same modification index found in the encoding of the modification position score CV term of the `<SpectrumIdentificationItem>`. To remain backwards compatible, the `<Peptide>` element attributes for the modification residue and location can still be used and are recommended to reflect the most likely modification position.

Cross-linking Experiment Spectra Identifications

Another design update was the addition of a new mechanism to encode the identification of cross-linked peptides. With the present cross-linking techniques, one spectrum

may identify more than one peptide, which are either the cross-linked peptides or fragments of the peptides and crosslinker. In case these values are reported, the protocol section needs to include the CV term "cross-linking search" (MS:1002494) as a use-case label.
The structure of the previous mzIdentML standard definitions restricted the reference of one spectrum identification to one peptide. This was solved with the formation of peptide pairs, annotating the respective `<Peptide>` elements with the cross-link modification and a pair-specific identifier. The implementation of the cross-linking use case, available in OpenMS (v2.5), can be utilised to write the results from OpenPepXL[172] in mzIdentML or to read cross-linking identification results for visualisation in `TOPPView`. An example visualisation and more cross-linking details can be found in Appendix C.

Support for other specific use cases was also added to the specification, namely sample pre-fractionation, de novo search, proteogenomics search, and spectral library search. Further, scoring possibilities for peptide-level scoring and consensus scoring were also built into the schema. The complete specification document of the mzIdentML standard can be found on the PSI website[ii] or in the PSI document repository[iii].

3.2.3 Discussion

The peptide and protein identification from LC-MS/MS data and subsequent representation thereof is a fundamental part of the analytical process in MS-based proteomics. Representing this data in an open data standard achieves compatibility with other (PSI) formats and more importantly unlocks cross-functionality with all analytical software tools that support open data formats in a compound effect, as for example with MS-GF+[173], Comet[174], and Percolator[175].
The continuing development and support of the mzIdentML format ensures that new or specialised analytical developments of emerging importance, like the described cross-linking support, can be integrated and used in established analytical workflow environments. Implementation efforts are likely to be more efficient due to code reuse. Thus, the MS data community can benefit from better integration of advanced analytical methods more easily.
The integration of cross-linking support in mzIdentML through the described OpenMS implementation brings additional benefits through the network effect of a mature development framework such as OpenMS. These are the enrolment in continuous

[ii] `https://www.psidev.info/mzidentml/`
[iii] `https://github.com/HUPO-PSI/`

integration and testing, and support for a wide variety of platforms, and an active user base and regular training and teaching for the scientific community.
More prominently, the OpenMS integration unlocks the use of mzIdentML within workflow environments. This enables users to build automated workflows and (quality) reports that are explained in more detail in dedicated Chapters 4 on Quality Control, and Chapter 6 on Workflows.
Finally, the mzIdentML file format serves as standardised format for storing and exchanging proteomics identification data. The described use case extensions' search protocols (i.e., the parameters applied, and the method used for the identification) can be described in sufficient detail to reproduce an analysis result. This, and its design as an open data format, supports the ProteomeXchange[176][iv] consortium to provide globally coordinated standard data submission and dissemination pipelines involving the main proteomics repositories and to encourage open data policies. This is important for evaluation and re-use, and to effectively integrate more data into every (proteomics) researcher's search and discovery process. It promotes transparency, reproducibility, and collaboration among researchers. In turn, this encourages data producers to even further improve meta-data deposition to repositories, to improve the visibility of their work and the wider scientific community's knowledge base.

3.3 The mzQuantML Format for MS Quantifications

A heterogeneous range of approaches is available for quantifying peptide or protein abundance via MS. The mzQuantML[166] standard format addresses the issue of systematic description of quantification for those different approaches for the purpose of integration, exchange, archival, and reuse. Due to the current practices for publicising quantitative results being even more heterogeneous as available quantification protocols, the format has the potential to improve the community's insight into quantitative data. And, like other standard formats, it can improve tool compatibility through shared implementations and reanalysis/reproduction efforts through a structured documentation of the experimental design and computational analysis setting.
Before mzQuantML was developed within the context of the PSI, there had been no widely accepted format for capturing quantitative information available. Good practice for data analysis documentation requires such a format to be able to accommodate the quantitative data at each step of the analysis workflow and preferably keep evidence trails. This encompasses data from both the level of the measured analyte (i.e., peptides or small molecules) and their aggregates from experimental groupings as well as

[iv] `https://www.proteomexchange.org/`

metadata like the experimental design itself.

3.3.1 Methods

The mzQuantML specification was developed over multiple HUPO-PSI meetings and regular teleconferences of the mzQuantML task-group members of PSI, with the author taking a leading role in consensus building and early implementations for use-case testing and refinement. The methods section is split into two subsections of design rationale and structure (of mzQuantML) to better reflect the substance of developing, discussing, refining, and implementing file formats.

Design Rationale

MzQuantML was designed to reflect quantitative information flexibly and to accommodate the many different possible experimental designs and quantification methodologies to derive basic quantitative information from MS proteomics experiments and higher-level quantitative results between different groups of measurements.
The rationales guiding the development were as follows:

1. Due to the nature of proteomics analysis with MS, direct quantitative data comes from the signal intensity attributed to identified polypeptides. To be useful as a general format for quantitative data, the format must be able to also represent quantitative data about proteins and protein groups. Here, ambiguity and confidence levels from the peptide-to-protein inference must be accounted for.

2. Quantitative information can come from different parts of the MS run, or multiple runs, depending on the quantitative method employed. Some methods also require additional data normalisation steps to make the data comparable. The quantitative data representation in the format should be impartial to the source level without loss of fidelity.

3. In the field of metabolomics, similar methods for quantification with MS were established. The format should offer, albeit limited, support for small molecule quantification from metabolomic MS data.

4. Because of the varying nature of studies conducted, quantification can be undertaken on multiple groups of samples with multiple steps of replication. Hence, it must be possible to represent replicate MS runs, groupings of replicates, and different study branches or treatment groups (for example, as study variables). Therefore, the experimental design must be reflected.

5. Many methods for statistical analysis may be used and combined to derive high-level quantitative information, the order of which is mutable. A general representation of the quantitative data in order of analysis steps should be possible with the format.

6. Quantitative information should be well integrated, without necessarily duplicating the whole spectral and identification information from other formats. From within the format, standards such as mzML and mzIdentML, should be referenceable.

In addition, there are general principles and roles a modern format for quantitative data should fulfil. To improve common communication on research employing quantitative methods with MS, the format should be suitable for journals' and data repositories' request for submission of supporting evidence to the primary findings of a study. For this, it should also be possible to base MIAPE[177] (Minimal Information About a Proteomics Experiment) reports on the format. Data exchange should be possible between different software tools within an analytical pipeline. With a full mzQuantML file, it should be possible to reprocess or recreate the analysis, given no undocumented steps have been taken.
With the latter lists of goals and guiding rationales, a concrete collection of use cases that mzQuantML should support was defined in the mzQuantML specification document[v] (v1.0).

- Represent final abundance values (relative or absolute) for peptides, proteins and protein groups where protein inference cannot be performed in an unambiguous manner.

- Record quantification values about peptide/protein modifications, such as post-translational modifications.

- Provide abundance values at the level of a single run (called an assay in this context) and logical groupings of runs (called study variables in this context), which the user, for example, wishes to report relative values for.

- Provide an evidence trail for how final abundance values were calculated, such as the features used for quantifying peptides and proteins.

- Relationships between features either on different regions of the same LC-MS run or on different spectra that report on the same peptide or small molecule. These are particularly required for relative quantification approaches.

[v] `https://www.psidev.info/mzquantml/`

- Details about pre-fractionation, sufficient to describe the combination of multiple input data files (e.g., raw files) into a single assay where this has been performed.

The structural design defined by these rationales and use cases is described in the following.

Forms of Quantitative Data

As per item 2) of the design rationales, quantitative information can come from different parts of the MS run(s). The data can be derived from peak profiles of eluting analytes. These are regions of peaks on the two-dimensional space of retention time versus mass/charge, that are considered to be derived from one analyte entity, so-called features. The features of polypeptides can be identified by their characteristic shape owed to by the average isotopic abundance of their constituent amino acids. Further sequence identification can be associated trough the evidence from fragment spectra within the features' retention time versus mass/charge envelope. If additional sample-protein labelling methods are employed and measured against another (un)labelled sample in the same experiment, so-called multiplexing quantification will result in pairs or tuples of quantification features. Since those labels are non-isobaric, it is important to include the recorded mass differences in the paired features.
Through chemical labelling after the protein digestion phase of the sample preparation in the commonly employed bottom-up measurement technique, multiple samples can be multiplexed to be measured in the same experiment. Here, the quantitative information is contained in the fragment spectra of selected ions. The individual samples' label is revealed through the fragmentation process, producing distinctive reporter ions per label included in one fragment spectrum, with intensities directly related to the amount of peptide ions with the respective label.
Although not as widely practised as other label-free methods, due to the underlying selection bias of DDA MS, spectral counting can be employed to infer basic quantitative information. Here, the observational count of identified peptides needs to be recorded as a representation of the abundance of the respective peptide. Similarly, the precursor intensity of identified fragment spectra's peptide ions can be used as quantitative values. These methods, and label-free methods in general, have the additional disadvantage of requiring multiple experiments to be compared, where differences in loaded sample per run can influence the peak intensity, and normalisation methods need to be applied to compare multiple experiments.
Single- or multiple-reaction monitoring experiments focus on measuring characteristic

fragment ions produced by one or more selected peptides. These peptides are observed through narrow mass and retention time windows as determined with previous experiments, during which also characteristic fragment ions of a particular m/z are selected. All ions in the respective window get collected and fragmented, and the characteristic fragment ion peaks selected to represent the abundance of that peptide. Support for these types of experiments was not included in the first version of mzQuantML, however introduced in a follow-up specification.

Structure

From the design rationales it is clear that, next to the quantitative information itself, the format needs to keep additional information about the type of quantification data contained, provenance of that data and the file itself, and external definitions used throughout the file, or collectively: metadata. The format specification includes elements for experimental metadata (Fig. 3.3, 1.), like `<BibliographicReference>`, `<Provider>` information, `<AuditCollection>` with information on all involved contacts, and the experiment type in `<AnalysisSummary>`. While the former three elements are

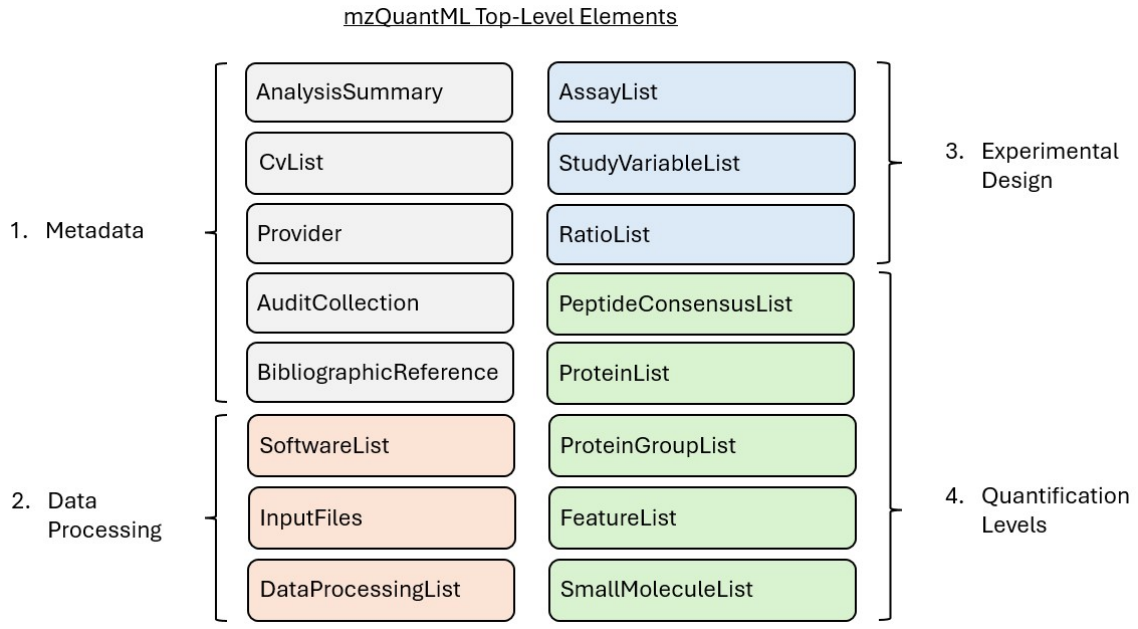


Figure 3.3: Overview of the top-level elements of mzQuantML with categorical grouping.

optional and contribute more to the completeness of an experiment's documentation and archival purposes, the latter is essential for the correct consumption of the rest of the elements and thus required. A `<cvParam>` element (or CV term) subordinate to

the <AnalysisSummary> element defines how to interpret the quantitative values in a file, specifically whether the data came from a:

- "MS1 label-based analysis" Experiment (MS:1002018),
- "LC-MS label-free quantitation analysis" Experiment (MS:1001834),
- "MS2 tag-based analysis" Experiment (MS:1002023),
- "Spectral counting quantitation analysis" Experiment (MS:1001836)

A <CVList> element keeps track of controlled vocabularies as source references for terms used in a mzQuantML file.
In a similar manner, the format accommodates a number of elements for the documentation of analytical parameters, connecting the various files of input with analysis software and parameters used. The data processing involved (Fig. 3.3, 2.). The <SoftwareList> element tracks the software and versions used in the analysis. <InputFiles> tracks the name, location, format, and type of inputs that enter the analysis, and the <
↪ DataProcessingList> element tracks the analysis steps taken with a combination of the steps' processing method description and references to the steps' input and output objects.
The experimental design (Fig. 3.3, 3.) is reflected through three types of elements. The <AssayList> elements define the context of each input file with an assay, or multiple assays in case of multiplexed experiments. In the latter case, an assay will also combine a specific (labelling) modification with the input file. The <StudyVariableList> elements record the logical combination between an assay and a certain experimental condition (usually disease, treatment group, time point, or a combination of conditions). The <RatioList> elements define the pairwise comparisons between study variables, forming ratios. The quantitative information itself is kept in elements of a layered design defined as follows.

A Layered Structure

The quantitative information in mzQuantML is organised in layers, for analysis on every different level of quantitative analysis: on feature level, peptide level, small molecule level, protein or protein group level. (Fig. 3.4)
Each level of features (<FeatureList> element), peptides identified and quantified over different LC-MS runs and samples (<PeptideConsensusList> element), proteins (<ProteinList>) and protein groups (<ProteinGroupList> element) have a set of quantification layers for the different types of values. Additionally, due to the generic

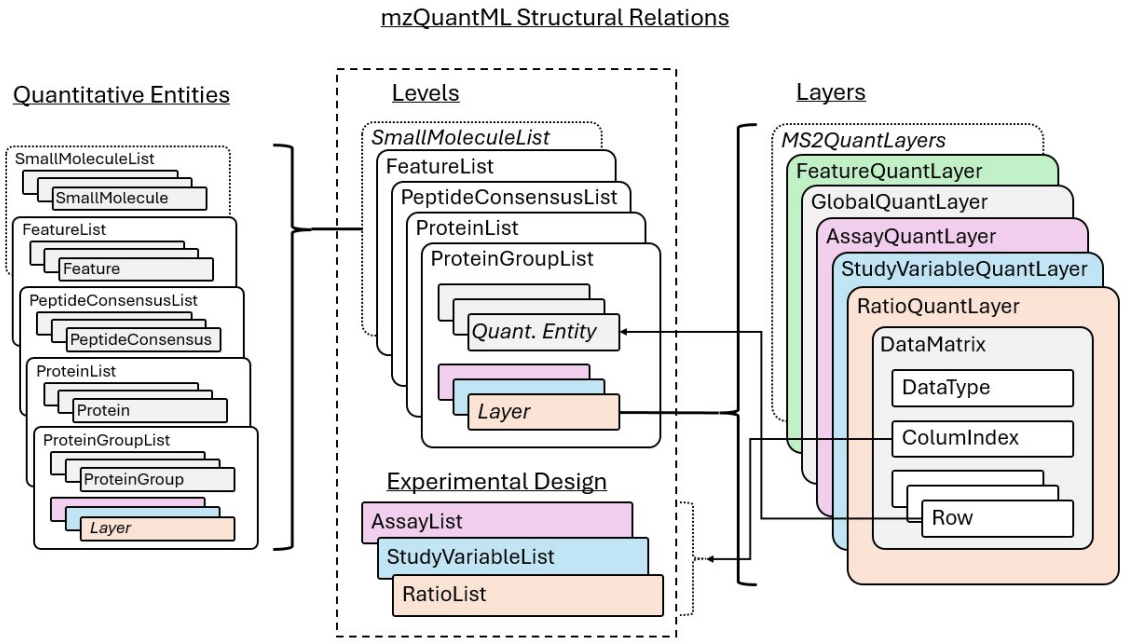


Figure 3.4: The mzQuantML format is structured by levels of quantification (feature, peptide, protein, or small molecule level) and layers on which that quantification is undertaken (assay, study variable, or a comparative ratio). The quantification data is stored in table-like `<DataMatrix>` elements, whose' column element types can be custom-defined, the rows need to refer to the elements in the respective level element, e.g., `<Proteins>` of the `<ProteinList>` element. The quantification of these elements can be captured on different layers, each with its own `<DataMatrix>`.

definition of LC-MS features, small molecule quantification can be covered without loss of generality.

The set of layers is designed to accommodate the quantitative values at different stages of the analysis, on a per-run basis (features only), on assay level, study variable level, or a ratio level comparing different study variables.

The layers are all designed as tables of values in `<DataMatrix>` elements. The row index of the table references a specific feature/peptide/protein/protein group. The column index (or header) references specific assays/study variables/ratios, corresponding to the type of layer. The data type for values within each `<DataMatrix>` must be defined with a CV term (e.g., intensity, raw abundance, normalised abundance, etc.), and multiple data matrices of the same layer can be included with different value types, depending on the necessary report detail.

The `<AssayQuantLayer>` element is for reporting quantitative values related to different assays, the values relating to the objects within the parent list type (e.g., `<PeptideConsensus ↪ >`, `<Protein>` or `<ProteinGroup>`), one per row. The column index must refer to assays defined in the file. In the same way, the `<StudyVariableQuantLayer>` reports data values

about peptides related to different study variables. The <RatioQuantLayer> holds values for predefined ratios of assays or study variables (a collection of biological or technical replicates over which averaging of quantitative values can be performed).
The <GlobalQuantLayer> and <FeatureQuantLayer> are special cases in that they report values relating to the objects within the parent list type but are defined by CV terms for the column types. For the <FeatureQuantLayer> this will be details like full-width-at-half-maximum (FWHM) of the features identified on a single LC-MS run or LC-MS run group. In a <GlobalQuantLayer> these will be global values corresponding to the peptide/protein/protein group such as the total intensity in all assays.
For MS2 based approaches, there are also designated MS2 quant layers: <MS2AssayQuantLayer ↪ >, <MS2StudyVariableQuantLayer>, and <MS2RatioQuantLayer>.
The values in <DataMatrix> elements of the different QuantLayers are described by CV terms and implicitly by the type of their QuantLayer, however, it is possible that different types of values arise as special cases during data analysis. These are described by the standard together with their encoding:

- Null ("null") values arise where a feature, peptide, or protein has not been measured,
- Zero ("0.0") values are to be placed where an entity has been measured to have zero value,
- Infinity values ("INF") arise for example in ratios where the denominator is zero,
- Calculation errors result in the "not a number" ("NaN") type.

Representation of the Experimental Design

The experimental design is reflected by the <AssayList>, <StudyVariableList>, and <↪ RatioList>, which group files or assays into referenceable groups (of <Assay>, <↪ StudyVariable>, or <Ratio> elements) from which differential quantification is to be calculated. An assay captures the concept of one sample analysed and can constitute multiple files. In a label-free scenario, such a sample will usually be a specific technical or biological replicate, and the assay hence references the peak files from that replicate. The reference is the identifier of the <RawFilesGroup> element within the <InputFiles> element, which may contain more than one LC-MS run file in case pre-fractionation is applied. Similarly, a list of <IdentificationFile> identifiers, also from the <InputFiles> element, can be referenced within <IdentificationFile_refs>.
In a label-based scenario, the labelled samples are multiplexed before injection. Here, each assay will reference a specific label or 'channel' within the corresponding peak

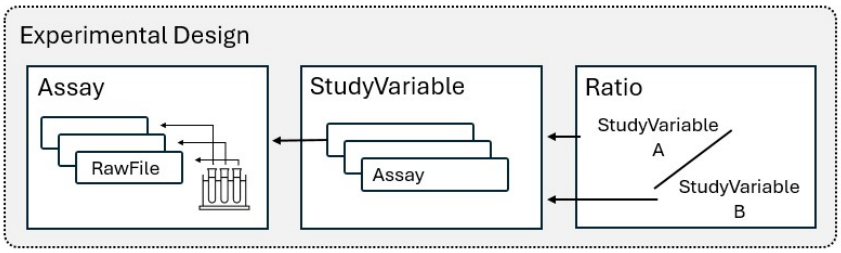


Figure 3.5: Experimental design represented with mzQuantML elements

file(s). Each `<Assay>` element will then specify the label modification with a CV term from UniMod or PSI-MOD, defining a mass shift, and reference a `<RawFilesGroup>` element.
A study variable will group assays by condition as given by the experimental design and can describe the experimental condition, also known as factor, with CV terms. The `<StudyVariable>`s are collected in the `<StudyVariableList>` element. This way, even complex designs, with randomized blocking, batches, and common references can be represented.
In relative quantification experiments, the quantitative values are often reported as quantitative ratios between conditions. In mzQuantML, the `<Ratio>` elements will group either study variables or assays into ratios, defining a group for the nominator and a group for the denominator in the ratio calculation. These logical groups are then referenced from the quantification layers to indicate which group the values given are representing. As mzIdentML, mzQuantML allows the referencing of format elements in the W3C XML Schema Definition `@xsd:id` attribute style.
Unlike mzIdentML however, mzQuantML must accommodate documentation of the application of several different analytical methods in a specific order that matters for the resulting values, as is common in quantitative workflows. The `<Software>` elements in `<SoftwareList>` record software and version used, whereas `<DataProcessing>` in `<DataProcessingList>` connects the order in which the referenced `<Software>` is applied. Additionally, the input and output files can be detailed. The specific processing method used with the application of referenced software has to be given in flexible detail with `<cvParam>` elements. These CV terms can not only be used to detail the processing method, but for example the parameter settings, too.
Thus, the evidence trails for the MS workflow leading up to peptide or protein quantification as a whole can be described.

Integration of Identification Data

Any mzQuantML file reporting more than anonymous feature quantifications needs to integrate identification data. These are provided by separate analyses, themselves serialised in identification formats. The combination with mzIdentML files is recommended, especially for cases where information about the identifications used needs to go deeper into detail, though other identification result formats may work as well.
For `<PeptideConsensus>` elements, the connection is established with `<EvidenceRef>` elements that link both features and identifications coming from a separate identification format to a `<PeptideConsensus>` element. Specifically, the `<EvidenceRef>` element will list references of features and the assays providing evidence for the particular quantitative value and references to identifications within identification files matching the referenced features.
The `<Protein>` and `<ProteinGroup>` elements will retain the connection to the features implicitly through references to the `<PeptideConsensus>` or `<Protein>` elements respectively, following a chain of references leading up to the protein, connecting to the evidence providing the identification of the protein, for example, the `<ProteinDetectionHypothesis ↪ >` element in a mzIdentML file.
Since this might not be deemed a necessary level of detail for reporting in all uses of mzQuantML, the identification references are optional. For minimal stand-alone capabilities, the assumed sequence and modifications of a `<PeptideConsensus>` element can be given independent of any references to external files with identification information. This way, mzQuantML is still available in use cases where identification may not be available or poses an inconvenient threshold for adoption if incompatible identification formats are employed.

3.3.2 Results

Four major quantitative techniques are supported in the mzQuantML release (v1.0):

- MS1 label-free intensity
- MS1 label based (e.g., SILAC)
- MS2 tag-based (e.g., iTRAQ / TMT)
- Spectral counting

List.3.2 shows another `<AnalysisSummary>`, indicating its mzQuantML file is reporting on a SILAC experiment analysis, which is a MS1 label-based analysis. It reports on the raw feature quantitative values found and the respective peptide level, but not on any

protein level.

Listing 3.2: A modification location score CV term encoding example

```
<AnalysisSummary>
    <cvParam cvRef="PSI-MS" accession="MS:1002018" name="MS1 label
    ↪ -based analysis"/>
    <cvParam cvRef="PSI-MS" accession="MS:1001835" name="SILAC
        ↪ quantitation analysis"/>
    <cvParam cvRef="PSI-MS" accession="MS:1002001" name="MS1 label
        ↪ -based raw feature quantitation" value="true"/>
    <cvParam cvRef="PSI-MS" accession="MS:1002002" name="MS1 label
        ↪ -based peptide level quantitation" value="true"/>
    <cvParam cvRef="PSI-MS" accession="MS:1002003" name="MS1 label
        ↪ -based protein level quantitation" value="false"/>
    <cvParam cvRef="PSI-MS" accession="MS:1002004" name="MS1 label
        ↪ -based protein-group level quantitation" value="false"/>
</AnalysisSummary>
```

Support for Metabolomics

The format's data structures, while primarily designed to represent peptide and protein quantitative values, are similar to structures that would be required to represent small molecules for metabolomic studies. With minimal extension to the peptide-centric design, the schema can represent small molecule data. This is accomplished via the introduction of a `<SmallMolecule>` element, which is used to denote the individual molecules and can be used for reference in the QuantLayers just as the `<PeptideConsensus ↪ >` element. This represents a bridging connection of two bioinformatics communities already sharing a considerable amount of analysis methodology and tooling. The 'small molecules' addition marked a convergence of the two communities for mutual benefit, and was continued in the development of mzTab.
For mzQuantML with small molecules, identification data integration can be achieved by the use of similar references, however now to match entries in a spectral database given the database used for small molecule identification supports referenceable identifiers.

A Layered Design Helps Intermediate Results Exchange

As a result of the layered design, QuantLayers can represent quantitative data at each step of the analysis. To represent intermediate steps of the analysis, there can

be multiple <AssayQuantLayer>, <StudyVariableQuantLayer>, and <GlobalQuantLayer>, for <FeatureList> as well multiple <FeatureQuantLayers>, to report. Each layer is then to be referenced in the DataProcessingList as @InputObject_refs or @OutputObject_refs respectively to report the order of processing and the method of processing.
For a detailed evidence trace on the quantitative data that led up to the consensus of peptides studied and their included quantitative values, multiple elements of <FeatureList>, <SmallMoleculeList>, and <PeptideConsensusList> elements can be included. If several tools upstream of the analysis were providing files that eventually led to the consensus list, the values of these files can be kept in separate lists and the files connected as @InputObject_refs in the <DataProcessingList>. To give the file consumer a convenient option to pick the one list representing the final result, a mandatory boolean attribute is provided on the <PeptideConsensusList> called @finalResult.
The simple extraction of values from mzQuantML for further use in statistical software is another benefit of this design. A brief Python script (< 20 lines, an example can be found in Appendix C), can extract the abundance values stored in the <DataMatrix> of a target layer in an arbitrary mzQuantML file for direct use in statistical software given the layer is present. As such, mzQuantML is a format suited for archival and support of publications providing a solid base for reanalysis and reproduction. Each step of an analysis workflow can be captured at the level of the measured analyte (i.e., peptides or small molecules) and each experimental group of measurements. This, together with the high level of report detail possible for experimental design and analytical parameters, can provide a most comprehensive picture of a quantitative study.

Implementation in OpenMS

We implemented mzQuantML support for the FeatureFinderCentroided and SILACAnalyzer TOPP tools (from OpenMS v1.10). The FeatureFinderCentroided tool supports different quantitative methods by detecting quantitative features from the MS1 level and therefore acts as an intermediate-step tool towards a complete quantitative analysis as distinguished in the mzQuantML format and required for a valid file. The mzQuantML file support was hence dropped to avoid tool input creep in favour of the later described mzTab format support in the TOPP toolchain. The SILACAnalyzer supports the complete experiment analysis from the raw (unpicked) signal of multiplexed MS runs from metabolically labelled samples and subsequent export to mzQuantML. The ratios of identified peptides from differently labelled samples (i.e., <StudyVariable>) are reported in a <RatioQuantLayer> on <PeptideConsensusList> level and the raw intensities of the associated (paired) Features in a <FeatureQuantLayer> on <FeatureList> level. The

`SILACAnalyzer` tool was deprecated after a restructuring of the OpenMS core library dependencies for BSD-license permissiveness and a following reduction of the core TOPP toolset (from OpenMS v2.0).

3.3.3 Discussion

It is crucial for the concise communication and later appropriate re-use of findings from the analytical results of MS-based quantification experiments to capture the experimental design and reflect the analytical workflow metadata. The trend for ever larger studies and the fact that quantification is only useful in conjunction with identification data implies the contribution of many input files with different experimental groupings, originating from different individual MS experiments in the final results. To capture this data is an enormous task that mzQuantML is able to accommodate, however, full automation is often hindered by the confluence of files from different formats, many of which might not come equipped with the appropriate metadata, simply because file-producing software upstream of the analytical process may be oblivious to either the necessary requirements special to quantification or the overall experimental structure. OpenMS, too, lacks a sufficiently comprehensive experimental design representation, as this would involve necessary process introspection spanning multiple analytical workflow steps (i.e., tools), which cannot be provided in a flexible framework based on individual tools for each workflow step. Monolithic setups have the advantage there, but the greater disadvantage of creating a hard-to-maintain code base that can only cater to usually a single majority use case.
MzQuantML circumvents this downside partially with easy import/export of abundance matrices for use in statistical software. This also provides flexibility for multiple quantification use cases with different types of statistical techniques. Further flexibility is provided through the generalist way the study design is represented. The design feature of quantitative elements in layers allows tracking the analytical process and omits explicit modelling of an experimental design within the format. The format was also the first to specify a non-proteomics data representation for inter-domain use of LC-MS/MS data and analysis tools.
With the collective experience on the design of quantitative experiment representation and community feedback on the perceived requirements to improve reporting, we started the development of a more accessible and compact format that would improve communication and documentation of the broad analysis (including identification) and the end results, described in the next section.

3.4 The mzTab Format for MS Identifications and Quantifications

The PSI standard formats described in the previous sections are all designed in XML, which implies a minimum amount of human readability. XML is however best suited for machine readability, integration, and automation. In the rapidly developing field of MS proteomics, data analysis methods are being developed at a fast pace, sometimes not even with dedicated tools or focus on automation, and therefore often lack the maturity to integrate data in elaborate formats. Analysis development benefits from a concise, easy-to-access representation of data for quick exchange between parties involved. Experience shows, that in these stages of development, a likely choice will be a tabular format. Metadata and parameters can be kept in a less integrated fashion and data values encoded in a series of tables when strict automation is not yet required. To fill this gap, we developed mzTab[178] as extension to the PSI standard format family.

3.4.1 Methods

The primary goal of the format is an improved and structured sharing of experimental results, with an emphasis on sharing final results. This should help researchers to use and integrate available data more readily into their projects. Similarly, this should encourage the development of new tools for niche applications and small but specialised research communities by removing the necessity to parse big and complex XML files to integrate proteomics or metabolomics experiment results.
The methods section is split into two subsections of design rationale and structure (of mzTab) to better reflect the substance of developing, discussing, refining, and implementing file formats.

Design Rationale

As stated above, the primary goal is an uncomplicated sharing of (final) experimental results, so the format needs to be uncomplicated enough to be created manually and interpretable without specialised tools (or at least with only the most ubiquitous of tools). Therefore, the mode of tab-separated text files is ideal for the format, as it allows a tabular representation of data, and a text editor or better yet spreadsheet software can be found on most personal computers.
To also allow the combination of MS-based proteomics and metabolomics experimental results within a single file, the concept of multiple special purpose tables within separate sections of the format enables the capture of different types of data (sections)

and helps to solve issues with n:m relations within a type (e.g., spectrum to protein relations).
Defining separate tables within the format allows to further simplify the processing of data by combining identification and quantification in one file. To distinguish the different tables, a common first (section) column can be used to designate to which table a given row of data belongs. This also allows for more flexibility as the respective table header rows do not have to be completely predefined but can be adjusted to the given use case. As such, not only the different sections are distinguished via the section column, but also table header and rows.
The format should furthermore support two levels of detail: a final summary result as could be expected in the supplementary material of a publication, and an intermediate results level with more details, so that mzTab can be used by downstream (statistical) software in a more complex analysis workflow. The summary mode also aims to simplify access to identification/quantification data. Consequently, a minimum amount of information provided through mzTab needs to be defined, while optional data can be used to give the mzTab consumer a more detailed view of the data or to satisfy local workflow requirements. For further customizability, the inclusion of user-defined data should allow to extend and adapt the format to special needs and requirements.
The mzTab format also needs to represent a broad range of experiments and therefore also needs the experimental design reflected. Controlled vocabulary terms from existing ontologies can be used to provide semantic context to the information within and between the different levels and necessary metadata, assigning data to ms_runs (RawFile in mzQuantML), assays, samples, and study variables as described before in mzQuantML.
The issue of separating the data within the format into different tables (sections) will be explained in the following section.

Structure

Following the design goals and rationales, data within mzTab can be divided into five sections. The sections are represented as separate tables yet within the same file. Each table has a header and a row indicator in the first column to differentiate the tables. The mechanism is an adoption of a similar mechanism in the UniProt knowledgebase[179] '.DAT' file[vi]. Besides distinguishing the tables of each section, this also allows the introduction of commentary rows ("COM" row indicator), useful in method development and helps with later use of the file.

[vi] `https://web.expasy.org/docs/userman.html`, accessed 01.07.2024

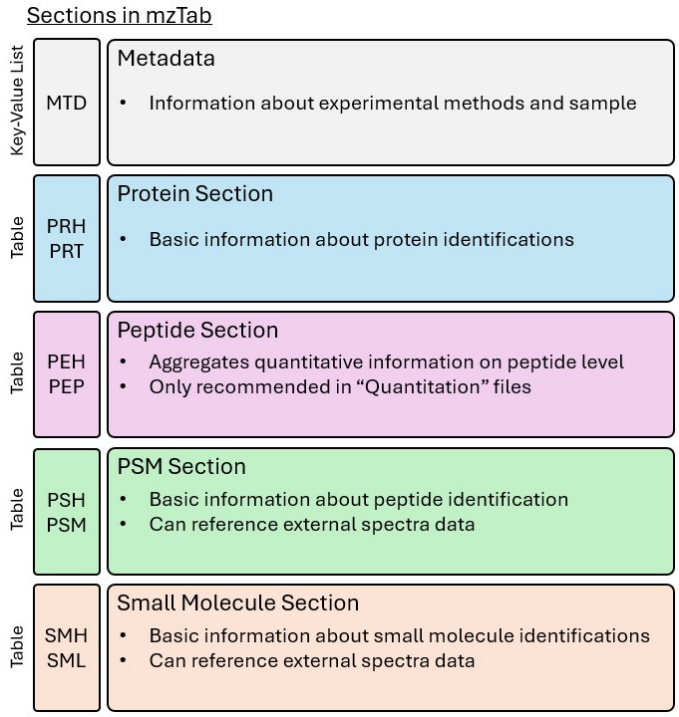


Figure 3.6: Overview of mzTab table sections with respective section indicator columns (left). Section indicators have a header (top) and row (bottom) indicator. The section's names and intended use (right) are listed in the sequence expected in a mzTab file.

Metadata

The metadata section differs from the subsequent sections as it is the only mandatory table and lacks a column header. It has only two columns beside the table indicator column: metadata definition type and value, effectively making each metadata row a key-value pair of metadata definitions. It therefore needs no explicit table column header as the other sections do and therefore has only the `MTD` section indicator (Fig. 3.6).
The metadata definition's key-value pairs provide essential and additional information on the dataset(s) reported on in a mzTab file. Essential information's metadata definitions must be present in any mzTab file:

- `"mzTab-version"`, `"mzTab-type"`, `"mzTab-mode"`, and `"description"`, to define the general function of the particular mzTab file,
- `"ms_run[1-n]-location"` to define the source peak files for which results are to be reported,
- `"protein_search_engine_score[1-n]"`, `"peptide_search_engine_score[1-n]"`,

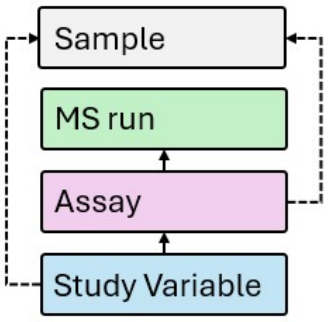
- `"psm_search_engine_score[1-n]"`, and `"smallmolecule_search_engine_score[1-n]"` for every search engine score reported in the file

- `"fixed_mod[1-n]"`, and `"variable_mod[1-n]"` to define the modifications included in the search

For definition keys that may occur multiple times, like `"ms_run"`, each individual definition is encoded with an appended counter in square brackets. Further sub-definitions, like `"ms_run[1-n]-location"`, are indicated by appending the type of specification to the respective definition key afterwards. The value of a definition is expected to be either a character string, number, CV term, or a comma-separated list of the same. This enumeration method to distinguish elements of the same type will be reused in later sections as well.
The mzTab standard specification details 52 other metadata definitions that can be included, resembling the common format elements of mzML, mzIdentML, or mzQuantML, detailing the instrument, software, search database used, amongst others. Most notable are a custom field and the definitions to describe the experimental design.
Another notable metadata definition is the "colunit". If specific unit uses are required for results reporting, they can be specified for each of the other four sections per column, unless the column has a dedicated unit specification/column, like the `"protein ↪ -quantification_unit"` column in the protein section table. Otherwise, the definition values must follow the encoding `"{column name}={Parameter defining the unit}"` where the parameter defining the unit must be an accession to a CV term defining the unit.

Experimental design representation

The model representation of the experimental design is adapted from the mzQuantML format and groups MS runs into assays, and the assays then grouped into study variables (or assays into groups of MS files, then into study variables in a multiplexed scenario) (Fig. 3.7). Each assay is defined in the metadata section and connected with one or more MS file(s) and, if applicable, a quantitative reagent definition. The MS files are defined by giving a location and format type. The study variables are defined by listing assay references and a description. The samples are referenced in the same way as the MS files from either assay or study variable and are themselves annotated with details in the metadata section. Furthermore, expected fixed or variable modifications to the sample's proteins are to be defined, as well as identification scores reported, and quantification units reported for each study variable or assay in a quantitative setting.



(a)

MTD	quantification_method	[MS, MS:1002038, unlabeled sample,]
MTD	ms_run[1]-location	file://path/f1.mzML
MTD	ms_run[2]-location	file://path/f2.mzML
MTD	ms_run[3]-location	file://path/f3.mzML
MTD	ms_run[4]-location	file://path/f4.mzML
MTD	ms_run[5]-location	file://path/f5.mzML
MTD	ms_run[6]-location	file://path/f6.mzML
MTD	assay[1]-ms_run_ref	ms_run[1]
MTD	assay[2]-ms_run_ref	ms_run[2]
MTD	assay[3]-ms_run_ref	ms_run[3]
MTD	assay[4]-ms_run_ref	ms_run[4]
MTD	assay[5]-ms_run_ref	ms_run[5]
MTD	assay[6]-ms_run_ref	ms_run[6]
MTD	study_variable[1]-assay_refs	assay[1], assay[2], assay[3]
MTD	study_variable[2]-assay_refs	assay[4], assay[5], assay[6]
MTD	study_variable[1]-description	heat shock response of control
MTD	study_variable[2]-description	heat shock response of treatment
MTD	study_variable[1]-sample_refs	sample[1]
MTD	study_variable[2]-sample_refs	sample[2]

(b)

Figure 3.7: The mzTab `MTD` section reflects the experimental design if used for results from multiple files. The example (right) shows the `MTD` section of a complete mzTab file reporting on a label-free experiment (l. 1) with six runs/assays, from two samples/study variables. The `assay[1..n]-quantification_reagent` entries, indicating no quantification reagent used, are omitted for readability. The schematic (left) indicates the connection of the respective mzTab keys forming an experimental design that is analogous to the previously seen mzQuantML experimental design representation. Each `study_variable` is described (l. 16-17), has a sample assigned (l. 18-19), and references the included `!`assays! (l. 14-15). The `assays` register the MS runs involved and can optionally reference their respective `sample` origins, should no `study_variable` be reported(not shown). `ms_run` details for the source files are required (l. 2-7).

CV parameter representation

Many table cell values can be represented with or added further semantic context by CV terms. The way CV terms are expressed in the XML-based PSI formats is via CV parameter elements that carry attributes for the CV term's origin, name, accession, and value. The design decision for mzTab was to encode this to fit into one table cell to avoid excessive bloat. The encoding is valid throughout the file's tables: `"[{label }, {accession}, {name}, {value}]"` with any field not available to be left empty. The label designates the term's origin via a unique label that is usually used as a prefix to the accessions (e.g., MS:1002601 for mzTab in the PSI-MS-CV). This encoding also allows to recreate the `<userParameter>` element known from XML-based PSI formats (i.e., local context terms without official CV term definition) as `"[,,{user parameter name}, {user parameter value}]"`.

CV terms may also be used for optional custom columns following the header format: `"opt_{object_id}_cv_{accession}_{name}"` as name for an optional column, where white spaces within the parameter's name have to be replaced by `"_"`. The `object_id` must be the identifier of the object the new column references, either assay or study variable, assigned in metadata. Values must then reflect the definition given for the term in the CV.

Multipurpose Design

The format has two types and two modes: identification or quantification, both in either summary or complete mode. The summary mode represents the final data for the experimental conditions analysed. For complete mode, full results per assay need to be included. The mode and type are two of the mandatory metadata definitions in the metadata section. Depending on mode and type, the list of required and optional columns for each section table can change.
The quantification type will of course require reporting quantitative values but may optionally report on identification type values, too. The complete column requirements can be found in the specification document.
For all types, modes, and section tables in general, missing values must be reported using "null". Multiple values within a table cell from separate measurements are to be listed using a vertical bar (`"|"`) delimiter, listings for one measurement like modifications for one sequence, are to be listed using a comma (`","`) delimiter.
The order of section tables (if present) must be as in Fig. 3.6, however, the order of columns is not specified, but recommended to appear in the same order as in the specification document. Additional columns may be added to the end of the table for all protein, peptide, PSM and small molecule sections as described in the CV parameter representation paragraph. These columns should represent information not included by default in the currently defined fields.

Reporting identifications

The report of proteomics identifications in mzTab can be done in the protein and PSM section tables, one protein/PSM per row. The peptide section is very similar to the PSM section and only used to report on identifications for quantitative results aggregated on the peptide level (see below), therefore not recommended in 'identification' files. The essential information mandatory to minimally report in any of the sections is the accession (or index) of the identified item as it appears in the search database used, the search engine and the search engine score. The first column to report is recommended to be the accession, for better comprehensibility. For PSM, the identified amino acid sequence is also mandatory. Related mandatory columns are database name, version, and species taxon. Protein groups are handled via the `ambiguity_members` column in the protein section, listing indistinguishable proteins by their accessions.

The identification process usually involves the assignment of a score to a peptide-spectrum pair or protein sequence, based on the spectrum evidence per LC-MS/MS

MTD	protein_ search_ engine_ score[1]	[MS,MS:1002252,Comet:xcorr,]
MTD	protein_ search_ engine_ score[2]	[MS,MS:1002052,MS-GF:SpecEValue,]
MTD	psm_ search_ engine_ score[1]	[MS,MS:1002252,Comet:xcorr,]
MTD	psm_ search_ engine_ score[2]	[MS,MS:1002052,MS-GF:SpecEValue,]
MTD	…	

PRH	accession	…	database	search_engine	best _search _engine _score[1]	best _search _engine _score[2]	…	protein _abundance _study _variable[1]	protein _abundance _study _variable[2]
PRT	P63017	…	UniProtKB	[...[1]]| [...[2]]	1.8991	5.89E-08	…	34.3	40.4351
PRT	…								

PSH	accession	sequence	PSM_ID	…	search_engine	search _engine _score[1]	search _engine _score[2]	modifications	retention _time	charge	…
PSM	P63017	…	1	…	[...[1]]| [...[2]]	2.2479	5.89E-08	null	1336.62	3	…
PSM	P63018	…	4103	…	[...[1]]| [...[2]]	2.0135	8.04E-08	null	1336.54	3	…
PSM	Q61699	…	2	…	[...[1]]| [...[2]]	1.8991	1.49E-06	9-UNIMOD:4	885.62	3	…
PSM	…										

Table 3.1: An abridged example of mzTab used to document the results of multiple search engine identifications on the same spectra. The respective search engine scores reported need to be specified in the `MTD` section (top), each for the `PRT` and `PSM` sections separately. The `search_engine` column informs on which identification score columns to expect and the type of score each column contains.

run. It is possible to report the results of multiple search engines in mzTab, and since the tables are designed to report one protein/PSM per row, a list ("|" separated) of successful search engines references (as listed in the metadata section) is expected in the `search_engine` column. The best scores (if multiple `ms_run` are considered) are then reported in n columns of `best_search_engine_score[1-n]`, with n according to the respective search engine's position in the `search_engine` list. In complete mode, it is also expected to itemise the scores per `ms_run` in `search_engine_score[1-n]_ms_run[1-n]` columns. For PSM of course, it only makes sense to report the `search_engine_score[1-n]` per search engine as each observation is intrinsically linked to one `ms_run`. PSM also require a unique identifier, so that if a PSM can be matched to multiple proteins, the same PSM can be represented on multiple rows with different protein accessions but the same PSM identifier.

The amino acid sequence required for the PSM section is not available for the protein section due to obvious readability issues. Modifications should be reported in all sections, with the exception of fixed modifications and labelling modifications, which only need to be present in the PSM section to improve readability. Consequently, modification reliability scores will be available at the PSM level only. Modifications are

reported in one column and follow an encoding that allows for the record of concise modification-type information via CV term use, and positional information and scoring (allowing for positional ambiguity). The encoding uses the previously described ("|") and "," delimiter for unambiguous separation of multiple pieces of information in one table cell, and is described in detail in the specification document.
A reliability classification can be used as an optional column in the protein, peptide, PSM, and small molecule section, to indicate identification reliability, depending on the used methods arbitrary reliability classification into three categories: high reliability ('1'), medium reliability ('2'), poor reliability ('3').
Depending on the section type and the report mode of the mzTab file (complete/-summary) a number of section-specific columns may be additionally required, like `spectra_ref` and `retention_time` in the PSM section. Complete requirement tables for all modes and sections to report identifications with mzTab can be found in the specification document.

Reporting Quantifications

Following the design goals to provide easier access to, in this case, quantification information, the mzTab quantification type must report the quantitative information on the level of study variables. This implies the aggregation of abundance values from several assays. The method of aggregation, however, is not part of the standard specification. More flexibly, it can be provided in a less stringent way via custom metadata and comments to support the file consumer. The protein, peptide, and small molecule sections report quantitative values in the columns associated with the `study_variable[1-n]` (e.g., `protein_abundance_study_variable[1-n]`), their units predefined in the metadata section. Quantitative information on study variable level is the most general result mzTab must provide in summary mode quantification-type files. The standard deviation and error for the given abundance values must also be provided. In complete type files, protein and peptide abundances can be given per assay in `protein_abundance_assay[1-n]` or `peptide_abundance_assay[1-n]` columns.
Most quantification methods involve a labelling agent to allow for multiplexed measurement. The quantification type mzTab therefore must provide metadata definitions about the quantification reagent (in form of a CV parameter) for each assay. For label-free analyses the PSI-MS CV term "label free sample" (MS:1002038) must be used in the metadata section.
For meaningful quantification, the complete PSM section identification column requirements are also required here. Similarly for the protein section, however, the number

of distinct and unique peptides and number of PSM per run are kept optional for conciseness. Instead, the peptide section needs to be present and provide essentially the same columns as the PSM section, although representing the present PSM evidence (in the PSM section) for a peptide through columns `retention_time_window` and `best_search_engine_score`, and `spectra_ref` reserved if quantification is done through MS2-label based approaches.

Reporting Small Molecules Analyses

For reporting small molecules in mzTab, the small molecules section table can be used. Within, small molecules are referenced through identifiers, as they would appear in a compound database like ChEBI or PubChem. For more detail, they also need to be assigned a chemical formula, SMILES and/or InChi identifier, and a free text description. Similar to the reporting of PSM, a database, version, species, spectrum reference, identification engine, and score also need to be reported for small molecules. Since the release of mzTab (v1.0), there have been ongoing efforts to better integrate small molecule data in mzTab, which resulted in a dedicated specification document, mzTab-M. The 'M' stands for metabolomics data, and the details of the format specialisation are beyond the scope of this work.

3.4.2 Results

The mzTab specification was developed over multiple HUPO-PSI meetings and regular teleconferences of the mzTab task-group members of PSI. With different types of columns, both mandatory, optional, as well as custom, section tables for multiple use cases and different application scenarios can be created. These can be used to describe both identification and quantification results. Many examples can be found in the mzTab repository[vii]

Usability

The simplest and likely most available method to use mzTab is with a text editor, built-in by default in any respectable operating system. More convenient and only slightly less prevalent is the use with a spreadsheet editor, making the formatting aspect of mzTab use intuitive. The use of mzTab is, however, not limited to spreadsheet or text editors, but can also be used with scripting languages popular with the analysis of experimental data like Python and R.

[vii] mzTab examples: `https://github.com/HUPO-PSI/mzTab/`, v1.0

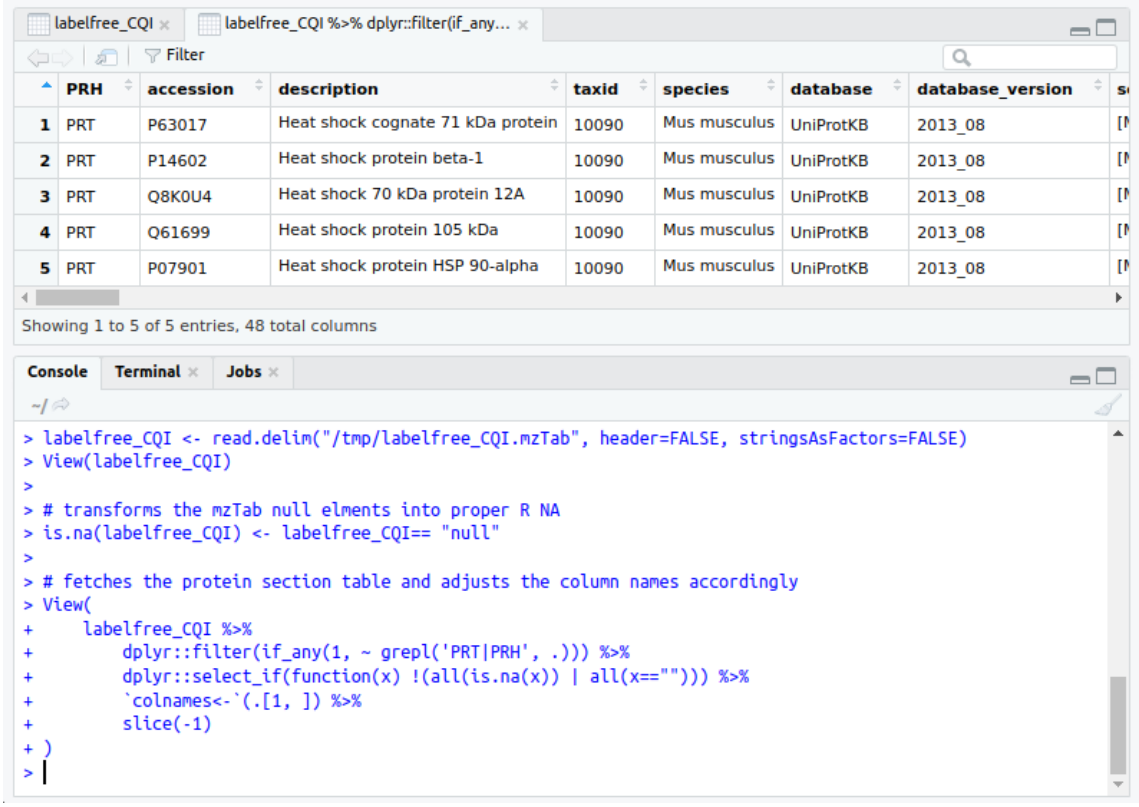


Figure 3.8: Little amounts of code (bottom) is needed to extract any of the mzTab section tables into an R data.frame, even without dependencies to specialised parsing software, but with reasonably standard packages in a basic data analysis setup – here "dplyr". Example loaded is a mzTab file released with the standard's documentation, shown here for visualisation within Rstudio, directly displaying the resulting data.frame (top).

The human accessibility of the format can perhaps be argued best through the fact that proteins/peptides/PSM that have been identified in only one, several, or all MS runs considered, can be easily spotted in Tab. 3.2). Within OpenMS, mzTab has been adopted through the implementation of the `MzTabExporter` TOPP tool. Through the exporter, mzTab can be leveraged in any OpenMS-based workflow producing OpenMS-internal quantification and identification files (featureXML, consensusXML, idXML), or mzIdentML files, as accepted input to the `MzTabExporter` to produce a shareable mzTab file. With this, it also provides an easier-to-use alternative to the previously described mzQuantML.

PSH	accession	sequence	PSM_ID	spectra_ref
PSM	Q61699	ALLRLHQECEKLK	3	ms_run[1]:scan=845
PSM	Q61699	ALLRLHQECEKLK	12	ms_run[2]:scan=1079
PSM	Q61699	ALLRLHQECEKLK	21	ms_run[3]:scan=611
PSM	Q61699	ALLRLHQECEKLK	29	ms_run[4]:scan=2280
PSM	Q8K0U4	AVVNGYSASDTVGAGFAQAK	2	ms_run[1]:scan=1300
PSM	Q8K0U4	AVVNGYSASDTVGAGFAQAK	20	ms_run[3]:scan=1066
PSM	Q8K0U4	AVVNGYSASDTVGAGFAQAK	28	ms_run[4]:scan=2735

Table 3.2: Through sorting the PSM table by the "sequence" or "accession" column, it becomes directly evident that peptide ALLRLHQECEKLK is detected in all ms_runs, however, peptide AVVNGYSASDTVGAGFAQAK is missing from ms_run[2]. This is an abridged example of the labelfree_CQI.mzTab file from the set of examples released with the mzTab specification.

Impact

With easily accessible methods to create, read, and share final experimental results, mzTab is well-suited for everyday use, especially for researchers outside the field of proteomics who lack specialized software to parse the existing PSI's XML-based standard file formats. This should furthermore encourage the development of small innovative research tools that can dispense with the requirement of parsing huge XML files, which might be outside the scope of many bioinformaticians.

Reporting complex analyses from similarly complex experimental methodology is hard to condense into a concise report that does not suffer from contextual omissions (i.e., useful only for researchers intimately familiar with context and implications). The use of CV terms for custom mzTab column definitions or within tabular cells of mzTab makes reporting more concise by falling back on existing definitions of key concepts, e.g., for different score types represented by an accession of a definition in a CV. On the other hand, custom columns in mzTab make automation harder to implement with mzTab as a general hand-over format.

Nevertheless, the low threshold of use and well-documented standardisation of mzTab and supporting tools can also be used to jump-start new analytical workflows. Although almost no workflow development has to start from the ground up, instead, reuses tools for generally applicable steps, it is still a considerable effort to develop and integrate tools necessary for a novel methodology. By adopting mzTab as handover format instead of a custom tab-delimited file format or even implementing XML-based standards, not only can valuable development time be saved but also result in a widely accepted format ready for sharing with the community[180–184].

The reception of mzTab can also be gauged by the number of public submissions made to PRIDE-Archive –the largest data repository for proteomics MS data– including mzTab files. As of 2023, there have been 2530 datasets with mzTab submitted to PRIDE, which is ∼10% of the total of submissions since the release of mzTab.

3.4.3 Discussion

mzTab is designed to give a high-level overview of analysis results, as it would usually be found in publications' supplementary materials. The latter would usually be in free-format of some of the various spreadsheet variations, or worse, in PDF, rendering programmatic access virtually impossible, and making re-use more difficult through conversion and reformatting. The mzTab format offers remedy with its simple yet structured design.
The format also features a more informative set of metadata than would usually be achieved with Excel or editors for simple csv/tsv, yet it can be constructed by hand without much more effort as needed to prepare the supplementary material to a publication, and with the same software a researcher would use otherwise. It is thus possible to replace those hard-to-reuse tables with mzTab, rendering the published results both human and machine-readable, offering a more consistent analysis result representation.
In general, however, mzTab does not replace the efforts necessary to re-use published data in metastudies as the reported quantitative values of mzTab should represent the final result of the performed data analysis. "The exact meaning of the values will thus depend on the used analysis pipeline and quantitation method and is not expected to be comparable across multiple mzTab files."[viii]
Within the family of PSI standard formats, mzTab is the first to embrace non-proteomics use cases for reporting identification and quantification of molecules by MS. The overlap of methodology is considerable, which in turn shapes the analysis result reports in a similar way. As such, it makes good sense to unify efforts for a human-accessible reporting format, as the fields see increasing cross-over in published studies due to their complementary nature and a common report format can only help efforts to combine approaches. However, mzTab metadata for small molecule identification results are currently restricted to score, FDR, sample processing, and fragmentation method, but with the derived efforts of mzTab-m improving applicability to non-proteomics fields can be expected.

[viii] mzTab specification document: `https://www.psidev.info/mztab`, v1.0

With easy options to read and write, the mzTab format is a readily available method for reporting data analysis results from mass spectrometry-based proteomics and metabolomics in general. However, since there are many specialist use cases, further canonicalisation and extended automation are only possible with considerable effort of specialisation (e.g., because of the presence of custom columns).

Chapter 4

Reporting Quality Control in Mass Spectrometry

This chapter includes partially identical or adapted content with permission from:

---

qcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments

Walzer M, Pernas LE, Nasso S, Bittremieux W, Nahnsen S, Kelchtermans P, Pichler P, van den Toorn HWP, Staes An, Vandenbussche J, Mazanek M, Taus T, Scheltema RA, Kelstrup CD, Gatto L, van Breukelen B, Aiche S, Valkenborg D, Laukens K, Lilley KS, Olsen JV, Heck ARJ, Mechtler K, Aebersold R, Gevaert K, Vizcaíno JA, Hermjakob H, Kohlbacher O, Martens L

(2014). Molecular & Cellular Proteomics, 13(8), 1905-1913; `https://doi.org/10.1074/mcp.M113.035907`

4.1 Introduction

Motivation

MS instrumentation is complex and sensitive to a host of external (e.g., temperature) and internal parameters (e.g., ESI efficiency). The system's input is usually coupled to an LC system to reduce sample complexity at any given measurement time point and this in turn to an autosampler. Minor differences in LC operation may change the elution order of peptides[185] or alter which peptides are selected for fragmentation[186,187] and MS2 generation. Experiment sample preparation involves multiple steps where contamination could be introduced or the sample be processed not as expected by the protocol followed (source, age, type of reactive agents)[188]. The complexity introduces many possible sources of variability, influencing overall repeatability and reproducibility. Quality control (QC) is increasingly recognized as a crucial aspect of mass spectrometry-

based proteomics, to record and report the variability and its potential sources. To maintain measurement reliability, detect failed or outlier measurements, maximise instrument up-time, and optimise maintenance schedules, a close watch on the quality of the measurements is key. Likewise, the data analysis has many dependent parameters that influence the outcome to some degree, and results can retrospectively inform about instrument or sample preparation performance. Therefore, a wide variety of quality parameters for different kinds of scenarios need to be watched and acted upon. The quality aspects of MS experiments and data analysis are also of great interest in the dissemination of research results, as they put the given results and claims into perspective to the measurement environment that gave rise to those results.
Naturally, a data exchange format that can accommodate this wide variety of quality parameters, versatile enough to support different kinds of experimental setups and use cases, would be of great benefit. With the previously described standard formats integrated into the OpenMS framework, versatile automated analysis workflows can be constructed to go alongside the experimental workflows of a wide variety of use cases involving MS. Since quality-related data may come from all steps of the workflow, such a format should also make it easy to add or extract additional information and work as a data hand-over format as well as a report format.

Background

A first call for coordinated efforts towards QC in MS was published with the Amsterdam Principles[189]. Since then, different proposals for the concrete application of QC and specific metrics in different use cases have been proposed[190–195]. Dedicated software packages, such as MSQC[195], QuaMeter[196], SIMPATIQCO[197], RawMeat by Vast Scientific (no longer supported, an open source software inspired by RawMeat is RawBeans[198]), PTXQC[199] and QCloud[200] have been implemented. Most serve distinct (but often overlapping) use cases, such as the inspection of single runs on a technical level, providing additional QC information on a data analysis, or augmenting core facility workflows. The QC metric definitions employed reflect the diversity of applications for QC in MS, all of which a unifying data format should optimally be able to support. As shown with the PSI-MS CV[201] use in mzML and in the PSI formats of the previous chapter, using CV-controlled data or parameter elements greatly improves the flexibility of a format, especially with future use cases and newly developed concepts (like novel QC metrics). Also shown in the previous chapter, using XML for creating a flexible hierarchical file schema makes data access easier, or in the case of a data-handover format, easier to append elements for new metric values within a data analysis workflow.

XML-based formats also provide a direct technical solution to rendering reports from files through extensible stylesheet language transformations (XSLT)[202], enabling the development of dedicated XSLT instructions for automatic report generation.
We therefore developed the qcML format, an XML-based standard that follows the design principles of the related mzML, mzIdentML, mzQuantML, and TraML standards from the HUPO-PSI.

4.2 Methods

The qcML format comprises an XML-based data standard and associated controlled vocabularies (CV) for storing various types of performance metrics, along with applicable metadata about the experiments. It is closely related to the HUPO-PSI data standards mzML, mzIdentML, and mzQuantML.
The methods section is split into two subsections of design rationale and structure (of qcML) to better reflect the substance of developing, discussing, refining, and implementing file formats.

Design Rationale

The qcML format was designed to address the two issues of storage and communication of quality control data and integrate into existing analysis workflows. In its implementation, it needs to reflect the qualitative aspects of either a single LC-MS/MS experiment or of a set of experiments in flexible grouping arrangement. This is to aid instrument 'health' monitoring and provide depth of support to a study's claims through introspection of the underlying data. The recorded quality parameters, in the following collectively called metrics, can be present in various forms, a single numerical value, a contextual tag or flag (e.g., a certain limit exceeded), multiple values either as list or in tabular form, or even an image (e.g., an image capture or summarising plot). New and custom metrics have to be accommodated without change in the format design.
The use cases to be supported are:

- Provide single experiment measurement (run) and analysis quality reports.
- Report on multiple groups of measurements (sets, e.g., as support material to a conducted study).
- Aide the longitudinal quality supervision and archival to support the high throughput operations of core facilities.

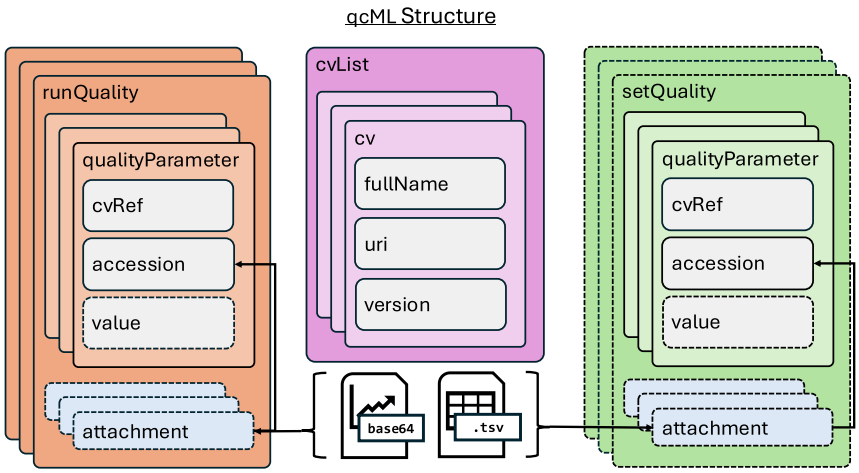


Figure 4.1: Overview of the top-level elements of qcML to represent both single-run (<runQuality>) and multiple-run (<setQuality>) quality metrics (<qualityParameter ↪ >), each with optional attachments in the form of tabular or image data (base64). The originating CV of terms used to define the quality metrics are referenced in a <cvList>.

- Support of and seamless integration into automated analysis workflows (plug-in style).
- The format should also allow for quick access in human- and machine-readable form, and optimally work both as a report for supporting background information in data sharing and ready for archival to preserve a record of the measurement circumstances, providing best practice of scientific data collection.

Structure

To accommodate the above-outlined use cases, especially to follow the analysis workflow with a record of quality assessments for each of the successive steps, the structure of a run- or set-quality should be a plain list, making merging and stacking a simple operation. To be able to report on single or multiple LC-MS/MS experiments, the format should have the option to contain metrics for individual runs and metrics for a group of runs. The nature of the data for quality assessment is in both cases the same, and a common underlying format structure can be reused to simplify implementation (Fig. 4.1).

To further ease complexity and allow for more code re-use, the basic structure for a metric representation can be taken from the <cvParam> concept of the previously introduced standard formats, here as a list of one or more <qualityParameter> elements for each LC-MS/MS representation or grouping thereof. This allows for the representation of single-value metrics or lists of values as required. The meaning of the value(s) is

defined via a CV term reference, first to a specific CV via `@cvRef` and a specific term therein with `@accession`. The CVs referenced are listed in a global `<cvList>` element, detailing each CV with a `@fullName`, `@uri`, and `@version`. To allow for more detailed data representation, optional attachments to the base metric must be possible. The `<attachment>` element allows for the representation of more extensive data in a separate list of elements. These can be tabular data (as previously seen in mzQuantML) represented with a `<table>` element inside the `<attachment>`, or base64 encoded (image) data with a `<binary>` element. The attachment is referencing a `<qualityParameter>` element and is itself defined via a CV term to provide meaning to the data.
The qcML format combines the above structure details into an XML-file format schema, that describes the quality of one or more runs, and optionally sets of runs described, and a listing of the controlled vocabularies that contain the metric or parameter definitions used throughout the file.

4.3 Results

The structure of the design is kept simple to ease implementation efforts and thus allow easy data access and data handover between data analysis tools and scripts.
By using a defined file structure, which can be validated using XML schema, as all previously described PSI XML standards, QC data can be represented in a compact and portable way. With the use of CV term-defined elements, the format is ready for future use cases and novel metrics. Through the simple design, aggregation of quality control metrics across experiments becomes a straightforward task through the ability to easily merge files or extract specific values.
We implemented QC tools in OpenMS to aid the collection and processing of QC data along a data analysis workflow (Tab. 4.1). Starting with a single MS run in mzML format, `QCCalculator` can produce an initial qcML file with basic quality parameters such as the number of spectra at given MS level. The breadth of basic quality parameters can be widened by additional input of identification data, as produced by the potentially earliest steps of a data analysis workflow. A data analysis workflow usually performs multiple sub-tasks, accruing additional data contributing to QC or warranting QC investigation for themselves. This data can be added over the course of the workflow (or post-hoc) with the `QCImporter`, importing calculated QC metric accession-value pairs to the qcML file, and the `QCEmbedder`, attaching additional tabular data or image data to the respective qcML. In the reverse, QCExporter can export metric values into tabular text format, and `QCExtractor` extract attachments, e.g., for downstream input to specialised (QC) scripts. The `QCMerger` tool allows merging two qcML files into one,

OpenMS tool name	Description
QCCalculator	Calculates basic quality parameters and metadata from MS experiments and subsequent analysis data, like experiment name, number of PSM, spectra S/N ratios, etc.
QCEmbedder	Attaches a table or image (PNG) to a given qc parameter entry
QCExtractor	Extracts a table attachment of a given qc parameter entry into tabular format
QCExporter	Exports the metric values of selected runs/sets into tabular format, optionally mapping the metric accessions to custom column header names
QCImporter	Imports multiple metric values into a qcML file, metrics need accessions in the column header, one run/set per row, optionally mapping custom table column names to accessions
QCMerger	Merges two qcML files together
QCShrinker	Remove selected attachments from a qcML file that are not needed anymore, e.g., for a final report

Table 4.1: QC workflow tools implemented in OpenMS

completing the tools needed to make qcML a QC data handover and collection format throughout a data analysis workflow. Any data in a qcML file not necessary at the end of a data analysis can be trimmed with the `QCShrinker`.

Though not its primary function, the qcML format can also easily serve as quality control report. A qcML document can contain both quality metric values derived and aggregated from MS experiments, as well as to give further detailed information on these metrics via attachments. These can also be plots visualising certain quality aspects of an experiment created from data collected during or at completion of a data analysis workflow with the help of `QCExporter` and `QCExtractor` and visualisation scripts (e.g., with R's ggplot). Through embedded XSLT instructions, the qcML document itself can be directly displayed in a browser (Fig. 4.2).

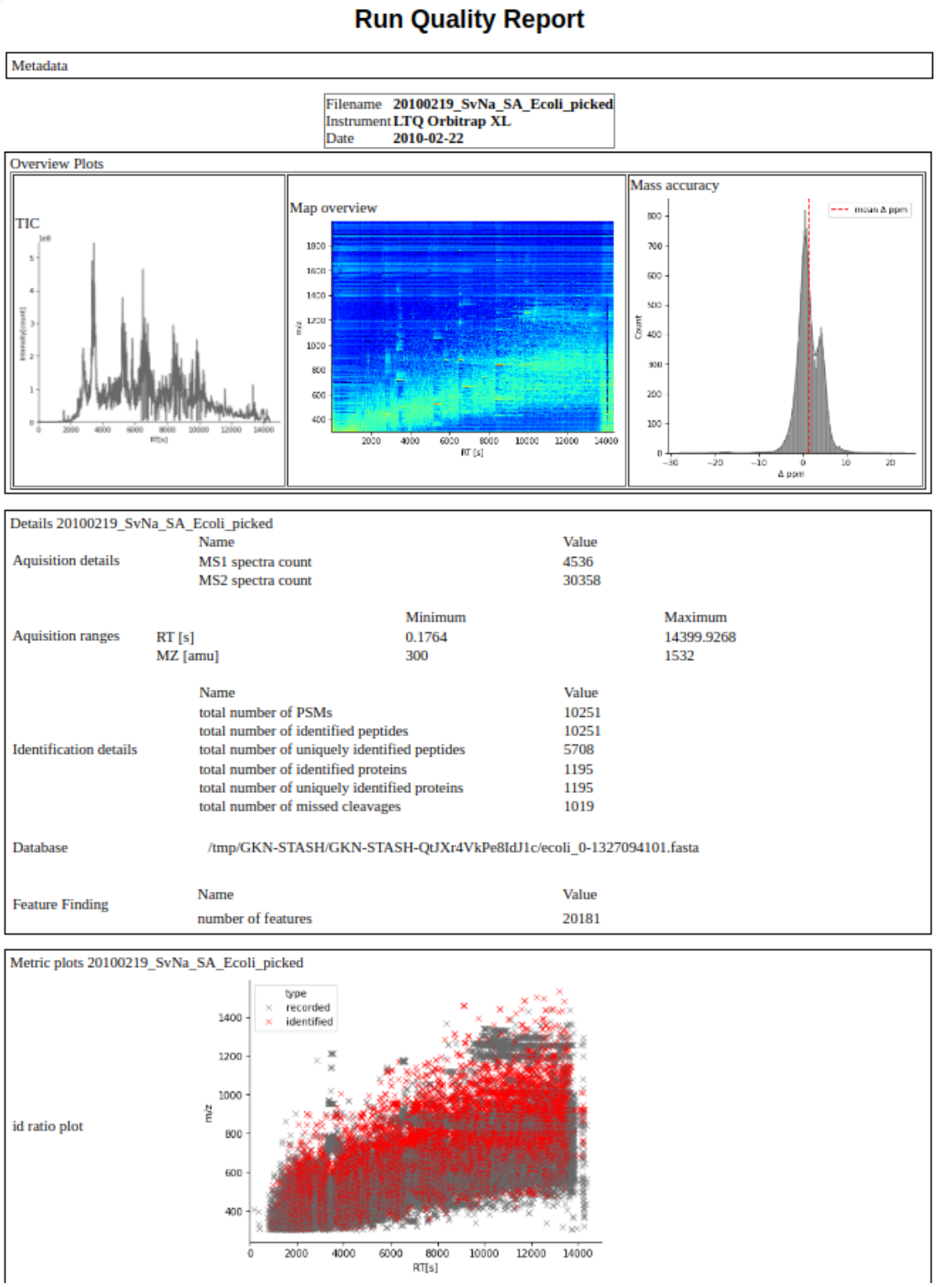


Figure 4.2: QC report rendered from the default XSLT included in the qcML file.

A report from a qcML file can be of high detail, given that both data visualisations and tabular data can be attached, to give e.g., a table detailing the confidence intervals, p-values, and test statistics of the dependent variables created by a statistics script for sets of MS runs in a study or identifications statistics visualisation like the mass error distribution of MS2 identifications for a single MS run (Fig. 4.2).
Many modern browsers have XML executing functions such as XSLT disabled for local files by default to mitigate security risks. Nevertheless, the qcML can be transformed into HTML or a similarly formatted PDF for display with little effort. XSLT transformation tools are widely available and can transform into PDF or HTML with a simple command (Lis. 4.1).

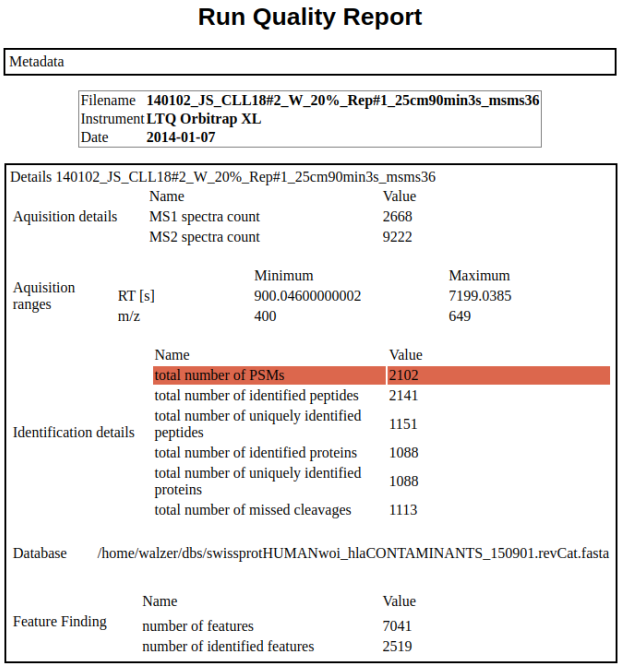
Listing 4.1: An example commandline call to transform a qcML file into HTML

```
/bin/bash$  xsltproc -v 141211_NT_HCC26SIMSYNPEPSinJY#4_msms5.qcML
    ↪  > 141211_NT_HCC26SIMSYNPEPSinJY#4_msms5.qcML.html
```

To support the core facility use case, where many LC-MS/MS experiments are measured on a day-to-day basis, hence process automation is necessary, integrated flagging thresholds can be easily accommodated with a customised XSLT stylesheet to indicate whether the metric has exceeded a specified limit.

Listing 4.2: An XLS stylesheet template to transform a qcML into a report with custom flagging for selected metrics, here highlighting in red if 'QC:0000029 ! total number of PSMs' is below 50% of 'QC:0000007 ! MS2 spectra count'.

```
<xsl:template match="ns:qualityParameter[(@accession = 'QC
    ↪ :0000029')]">
     <xsl:variable name="PSM" select="//ns:qualityParameter[
        ↪ @accession = 'QC:0000007']/@value"/>
     <xsl:choose>
          <xsl:when test="@value < $PSM div 2"> <!-- limit is <50%
    ↪ of MS2 recorded -->
               <td bgcolor="#DC674D">
               <xsl:value-of select="@name"/>
               </td>
               <td bgcolor="#DC674D">
               <xsl:value-of select="@value"/>
               </td>
          </xsl:when>
          <xsl:otherwise>
          ...
          </xsl:otherwise>
     </xsl:choose>
</xsl:template>
```



Run Quality Report

Metadata

Filename	**140102_JS_CLL18#2_W_20%_Rep#1_25cm90min3s_msms36**
Instrument	**LTQ Orbitrap XL**
Date	**2014-01-07**

Details 140102_JS_CLL18#2_W_20%_Rep#1_25cm90min3s_msms36

	Name	Value
Aquisition details	MS1 spectra count	2668
	MS2 spectra count	9222

		Minimum	Maximum
Aquisition ranges	RT [s]	900.04600000002	7199.0385
	m/z	400	649

	Name	Value
Identification details	total number of PSMs	2102
	total number of identified peptides	2141
	total number of uniquely identified peptides	1151
	total number of identified proteins	1088
	total number of uniquely identified proteins	1088
	total number of missed cleavages	1113

Database /home/walzer/dbs/swissprotHUMANwoi_hlaCONTAMINANTS_150901.revCat.fasta

	Name	Value
Feature Finding	number of features	7041
	number of identified features	2519

Figure 4.3: A red flag is raised for runs with fewer than 50% MS2 identified in this custom XSLT stylesheet (Lis. 4.2), overview plots in the qcML report transformation omitted for clarity.

An equivalent relational structure complementing the XML-based file format allows qcML data to be stored in a relational database[203] such as MySQL/MariaDB, PostgreSQL, or Oracle, to support archival purposes or longitudinal analysis, also a common requirement in core facilities. By limiting the structure to simple XML types and only minimal referencing between the `<qualityParameter>` and the `<attachment>`, uncomplicated database serialisation of the XML elements unlocks the option of long-term storage in laboratory information management system (LIMS). The schema is future-proofed by decoupling the metric representation from the structure itself. Both `<qualityParameter>` and `<attachment>` types are CV controlled, their value's meaning is CV-defined and the structure is a generic combination of XML built-in types with a CV term accession tagged for definition.

The first reason for a quality format is the same as with any of the other standardised formats. The field of proteomics needs to have standardised formats to fulfil the scientific premise of keeping good records. Storing and communicating this new type of information is currently not standardised, limiting the dissemination of quality control data along with experimental data. Even for just hundreds of files, all in different formats, the amount of work necessary to leverage the data contained for

new insights through reuse or even just review or reconsideration under the light of new insights, becomes insurmountable, let alone with thousands of files in thousands of studies, as was common and has ever since grown in extent due to the prevalent high throughput methodology in life science. This allows large facilities, employing LIMS to store quality information on their measurements, with little additional effort to make large volumes of data (think hundreds of thousands of runs) queryable efficiently regarding different qualitative aspects of the data.
Another reason is that the standard quality format qcML can be used as a report, ready for inspection of either a single file or groups of files as common in most studies published nowadays. This is important for concise communication of results between researchers, be it intra- or inter-domain, to elucidate the context surrounding the data used to conclude given claims. Especially for interdomain exchange, it is important to provide as much background detail as possible and leave as few details and assumptions surrounding the data unexplained as possible unclear or uncommented. As is common within many domains of research, details might be implicit to keep publications brief and focused on a narrow readership. In a similar manner, the qcML format can be used as archival information, summarising what data a particular archive may contain.
A third reason is flexibility. What to report largely depends on the use case: archival, analysis report, and data-sharing. However, all of these use cases can be accommodated thanks to the flexible way quality metrics are recorded. Reporting can be done on both the individual run and set-of-runs level which increases the possible depth of reporting. The metrics themselves also allow for a rich depth of data, primarily reporting a single numerical, string value, or boolean flag, but with the possibility to attach more values such as tabular data or even base64-encoded plots. This allows for a rich report, especially in combination with the XSLT feature of qcML, enabling qcML self-contained formatting instructions for display in HTML (i.e., view-compatibility with a browser), or PDF print. To stay updated with the current and future practices and methods evolving, the metrics are stored in a generic pattern of name-accession-value data, and defined via a CV, making updates and access to new metrics straightforward, compatible with automated or high-throughput application programming interfaces (APIs).
As such, qcML unifies the handling of quality control data by analysis or reporting tools, while maintaining a high level of transparency of the data to the user. As a dedicated format for quality control, qcML is the first of its kind.

Chapter 5

A Framework for Immunopeptidomics

This chapter includes partially identical or adapted content with permission from:

FRED 2: an immunoinformatics framework for Python

Schubert B., Walzer M., Brachvogel HP., Szolek A., Mohr C., Kohlbacher O.

(2016). Bioinformatics, 32(13), 2044–2046; `https://doi.org/10.1093/bioinformatics/btw113`

ImmunoNodes – graphical development of complex immunoinformatics workflows

Schubert B., de la Garza L., Mohr C., Walzer M., Kohlbacher O.

(2017). BMC Bioinformatics, 18(242); `https://doi.org/10.1186/s12859-017-1667-z`

5.1 Introduction

Motivation

Data analysis software frameworks, such as OpenMS, introduced in the Background chapter, can facilitate the development of new multiomics data analysis methods and workflows. Their effect is twofold: 1. through integration of (data sources and) software tools, they make a more flexible combination of methods possible; 2. they accommodate the rapid development of new data analysis tools by providing a stable structure of reusable concepts, such as data input parsing and output formatting to standard format as described in the Data Integration for Automated Workflows chapter. Computational approaches have also become common for immunology, aiding in research, by facilitating the process of epitope detection and helping with vaccine design. Through their application, the design of analysis pipelines that can handle large amounts of data is possible. Progress in the prediction of T-cell epitopes using machine learning methods particularly contributed in the publication of an abundance

of methods for different or partially overlapping use cases.
Many of these methods are freely available, many as web-services, or as standalone implementations of academic software. The issue with this type of software-service is the continued availability and maintenance, which can be addressed by integration into common (open source) frameworks. The FRED[204] was the first project to integrate different computational immunology methods of significant use-case coverage under one project. Its integration approach with a uniform interface also made comparison of different methods better possible, important to assess method applicability in the design of a given analysis workflow use case. FRED also laid the foundation for large-scale data handling capability in custom designed workflows.

Background

Defining for the functionality of a framework like FRED or OpenMS are the tools and method implementations included. For immunoinformatics, the key categories for which tools and methods have been published are epitope prediction, antigen processing, HLA typing, and vaccine design.
A prerequisite to generate an immune response from epitope-based vaccines are peptides that have a strong enough affinity to HLA for effective presentation[205]. Algorithms that predict the affinity of peptides to HLA are therefore at the forefront of immunoinformatics research tools to guide rational vaccine design. Binding affinity corresponds to the peptides' chemical properties to form weak molecular interactions with the binding site of the HLA molecule. HLA-class I and II are highly polymorphic and show specific peptide preferences[206,207]. There are several approaches to leverage the specificity for HLA-binding prediction, based on large amounts of experimental data, e.g., from SYFPEITHI[208] or IEDB[209]. The earliest approaches to the problem used position specific scoring matrices (PSSM) reflecting a motif, and assigned a binding score to candidate peptides through match against the motif (e.g., BIMAS[210], SYFPEITHI[208]). These linear approaches, however, fail to incorporate the spatial context in which the affinity properties of a given amino acid at one position may be influenced by amino acids at other positions in the peptide. More elaborate machine-learning methods use support vector machines (e.g., SVMHC[211]) or artificial neural networks (e.g., NetMHC[212]). Support vector machines classify peptides as binder or non-binder through the translation of the amino-acid sequence into a higher-dimensional space which is divided by a hyperplane defined through training data. Artificial neural networks represent a motif in terms of a weight-matrix over the connections of the neural network, peptide sequences encoded into a network input vector, and the weights assigned during training.

HLA ligand binding involves other, albeit less selective steps that impact the motif of naturally processed HLA peptides. Key steps are proteasomal cleavage and TAP transport. Proteasomal cleavage is the first step in the antigen processing pathway, for which specificity predictions can be made. Intracellular proteins are digested by the multicatalytic proteasome, generating specific cleavage patterns. Due to the scarcity of training data, most methods rely on in-vitro generated specificity measurements (e.g., PCM[213]), methods trained on in-vivo data do, however, exist (ProteaSMM)[214]. Another sequence specific step is the TAP transport, during which TAP binds intracellular peptides and delivers them into the endoplasmatic reticulum, where they can bind nascent HLA molecules. TAP binding are sequence specific and has been found to correlate with transport rates, but as with proteasomal cleavage, in-vivo data is scarce. For example, the initial SVMTAP[213] method, based on support vector regression, was trained on a set of ∼400 peptides with measured binding affinity to TAP.

HLA typing is essential for the development of vaccines, correct application of prediction methods, and especially important for the development of personalised therapies. With the introduction of molecular typing and NGS technology widely available in most clinical and research settings, algorithmic typing methods are now among the most important. Here, several methods have been published and considered in this work. OptiType[62] is a HLA genotyping algorithm based on integer linear programming, simultaneously considering all major and minor HLA Class I alleles. It can produce accurate four-digit HLA genotyping predictions from RNA, exome, or whole-genome sequencing data. Polysolver[215] uses a Bayesian probabilistic model to reassign reads from whole-exome sequencing that failed to map to the reference genome to HLA-class I references. The Seq2HLA[64] algorithm produces HLA-class I and II typing and expression evaluation from RNA-Seq data. It produces two-digit resolution typing from 'HLA groups' based on inter- and intra-allele sequence variability with a greedy algorithm based on read count maximization. ATHLATES[63], too, can produce HLA-class I and II typing from exome sequencing data. The approach used in the software package is the determination of the minimal (Hamming) distance of HLA allele pairs for each locus to the exons. Several publications give an deeper overview and benchmarking of the algorithms[216–219]

Another algorithmic application for immunology use cases can be found vaccine design. Since the HLA is highly polymorphic, the repertoires of potential HLA-binding peptides differs substantially. Economical and regulatory issues give strong incentives to identify the optimal set of peptides for a epitope-based vaccine and effective modes of assembly to improve vaccine response. With OptiTope, a selection of candidate vaccine peptides that maximizes the overall predicted immunogenicity through integer linear program-

ming can be found. The optimal assembly of such a candidate vaccine selection to form immunogenic structures can be investigated with a string-of-beads polypeptide approach[220] or epitope concatenation with optimal spacer sequences maximizing the cleavage probability[221].
Beside FRED, other resources making immunoinformatic tools collectively available, have been published.
EpiToolKit[222] is a web-based workbench for vaccine design. The tools offered cover epitope prediction methods, MHC genotyping, epitope assembly and selection, largely matching the methods described in this chapter. It is based on the open-source platform Galaxy[158], which allows the flexible combination of tools into a workflow. As a web-based platform, data needs to be uploaded to a server. The server also restricts tool-availability through its service lifetime.
The Immune Epitope Database and Analysis Resource[i] (IEDB)[223] primarily hosts curated immune epitope data related to T- and B-cell epitopes. As an analysis resource it is also home to a collection of published immunoinformatics tools. These include predictors for both MHC class I and II restricted T-cell epitopes, methods to predict linear B cell epitopes, epitope population coverage, epitope conservancy analysis, and epitope cluster analysis. IEDB tools are available as a web-service, with a RESTful interface to access tools and database.

5.2 Methods

We introduce FRED2, a modular Python framework for immunopeptidomics. FRED2 reflects the rework and extensive development of the Framework for Epitope Detection[204] (FRED), to align with the development of the source language and accommodate for better integration with other frameworks and workflow management software. A comprehensive basis of common data structures and framework functionality allows for the easy addition of new tools and extends code reusability. FRED2 is open-source software and released under a three-clause BSD license. The project design reflects the areas of immunoinformatics covered by FRED2 with a split into three sets of functionality modules.
The framework basis provides support for three fundamental capabilities necessary for immunoinformatics. Discovering and handling different HLA typings is important to accommodate for the immunological diversity present in any biological sample through HLA polygeny and polymorphy explained in the background chapter. For HLA typing,

[i] `http://www.iedb.org`

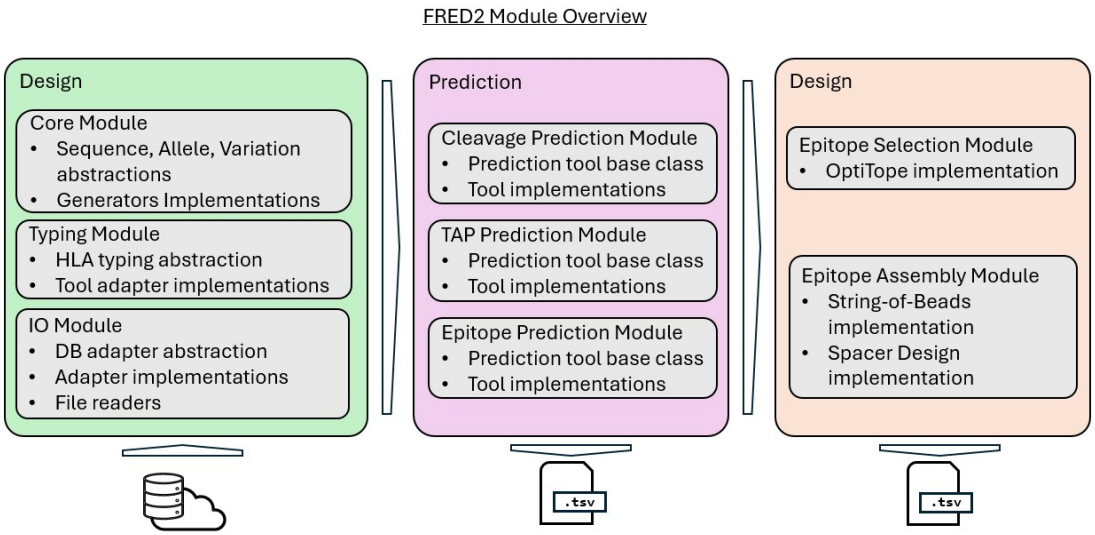


Figure 5.1: FRED 2 Module overview

FRED2 provides adapter methods for HLA typing approaches, such as OptiType[62], Polysolver[215], seq2HLA[64] and ATHLATES[63]. Handling biological sequences and variations at all major biological levels is the next core framework capability. Sequences can come from gene, transcript, or protein level, and need to be broken down to the immuno-peptide level for modelling immune system functionality. Sequence level transitions and integration of variations is managed by core-level functionality with sequence representation classes. Getting these sequences and variations in the first place is accomplished by interfacing with biological databases like BioMart[224], UniProt, RefSeq and Ensembl and reading common file formats such as VCF through dedicated adapters.

The prediction modules provides functionality for modelling core immune system properties. The prediction methods use established methods for the prediction of sequence behaviour in the immune system and are split into three packages of Epitope Prediction, TAP Prediction and Cleavage Prediction. Each package provides factory classes as entry points for the supported prediction methods to maintain a common codebase and familiar usage interface to improve development. These methods either reimplement published methods or provide command-line adapters to the tools implementing the published methods (see Tab. 5.1).

A third typical application of immunoinformatics is vaccine design. FRED2 provides modules for epitope selection and assembly. For population-based epitope selection, given a target population represented by their HLA alleles and virus proteins of interest, OptiTope[225], a highly flexible mathematical framework capable of expressing various

aspects of epitope-based vaccines, was reimplemented. OptiTope was integrated to select the most immunogenic epitopes that are constrained to cover at least a fraction of HLA alleles and antigens. All FRED2 prediction methods can interact with the reimplementation, overcoming previous limitations of the tool. To enable epitope assembly, FRED2 implements the traveling-salesperson (TSP) approach proposed by Toussaint et al.[220] and OptiVac[221] for string-of-beads design with optimal spacer sequences, which is similar to the approach taken by Antonets et al.[226]. An overview of the available tools and use case applications is given in Tab. 5.1.

Category	Method	Version	Use Case
Epitope Prediction	BIMAS[210]*	1	MHC-I binding
	SVMHC[211]*	1	
	ARB[227]*	1	
	SMM[228]*	1	
	SMMPMBEC[229]*	1	
	Comblib 2008[230]*	1	
	PickPocket[231]*	1.1	
	Epidemix[204]	1.1	
	NetMHC[232]*	4	
	NetMHCpan[233]*	3	
	HAMMER[234]*	1	
	TEPITOPEpan[235]*	1	MHC-II binding
	NetMHCII[236]*	2.2	
	NetMHCIIpan[237]*	3.1	
	SYFPEITHI[208]*	1	
	UniTope[238]*	1	T-cell epitope
	NetCTLpan[239]*	1.1	
	Callis propensity[240]*	1	Immunogenicity
Cleavage Prediction	ProteaSMM (C/S20)[214]*	1	Cleavage site
	PCM[213]*	1	
	Ginodi[241]	1	
	NetChop[242]*	3.1	
TAP Prediction	SVMTAP[213]*	1	TAP affinity
	SMMTAP[243]*	1	
	Additive matrix[244]*	1	
HLA Typing	OptiType[62]*	1	MHC-I typing
	Polysolver[215]	2.2	
	ATHLATES[63]	1	MHC-I/II typing
	Seq2HLA[64]*	2.2	
Epitope Assembly	TSP approach[220]*	1	String-of-beads design
	Spacer design[221]*	1	Spacer design
Epitope Selection	OptiTope[225]*	1	Vaccine design

Table 5.1: Supported immunoinformatics methods in FRED2, *: also available as part of ImmunoNodes (see 6.2.2)

With a view to the cascading effects when transitioning multiple sequence levels and integrating variants for multiple genes and heterozygous variations in estimating the immunogenicity of their peptide products on multiple alleles with multiple prediction methods, FRED2 sequence handling is built on generators. An exemplary use of

```
from Fred2.Core import *
from Fred2.IO import read_annovar_exonic
from Fred2.IO.MartsAdapter import MartsAdapter
from Fred2.IO.ADBAdapter import EIdentifierTypes
from Fred2.EpitopePrediction import EpitopePredictorFactory

vars = read_annovar_exonic("/tmp/annovar_excerpt.out")
mart = MartsAdapter(biomart="http://www.ensembl.org")
alleles = [Allele("A*02:01"), Allele("B*15:01")]
results = [EpitopePredictorFactory(e, version=v).predict(
                generate_peptides_from_proteins(
                    generate_proteins_from_transcripts(
                        generate_transcripts_from_variants(vars, mart,
                            ↪ EIdentifierTypes.REFSEQ))
                ,9),
                alleles=alleles)
            for e,v in [("Syfpeithi","1.0"), ("netmhc","4.0")]]
df = results[0].merge_results(results[1:])
df.head(n=8)
```

		A*02:01	B*15:01
Sequence	Method		
(A,A,A,Q,E,A,Q,A,D)	netmhc	0.016207	0.057122
	syfpeithi	9.000000	2.000000
(A,A,D,F,P,R,W,K,R)	netmhc	0.031507	0.015315
	syfpeithi	8.000000	1.000000
(A,A,F,L,A,G,L,L,S)	netmhc	0.120067	0.094033
	syfpeithi	8.000000	3.000000
(A,A,F,L,L,Q,H,V,Q)	netmhc	0.065083	0.135354
	syfpeithi	9.000000	3.000000

Figure 5.2: Generator and BioMart adapter use in FRED2. Top: a brief Python script for HLA-binding prediction from genetic variant products. Bottom: the script output, cut to the first four peptides, in its Python Pandas dataframe form.

the generators is given in Listing 5.2. Five variants, as created by ANNOVAR[245], nonsynonymous single nucleotide variants (SNV^{ns}) in *IL23R*, *ATG16*, and *NOD2*, and a frameshift deletion in *GJB2*, are integrated into their respective transcripts, fetched via the MartsAdapter from `ensembl.org` BioMart. FRED2 generators successively produce peptides of length 9 as input to the HLA-binding predictors SYFPEITHI and NetMHC. The combined results are kept in a multi-index Python Pandas dataframe.

5.3 Results

FRED2 is a framework for the development of immunoinformatic applications. For the implementation of new tools, FRED2 offers abstract base classes for various base methods (e.g., PSSM, SVM) to provide a common code base for common methodologies. The framework also provides representations for sequences and MHC-allele typings, necessary concepts in all immunoinformatics applications. To integrate existing tools to work seamlessly within the framework, FRED2 offers a abstract base class for external command-line tools, as used e.g., for NetMHC.
FRED2 has many immunoinformatics applications already integrated and covers three major areas of immunoinformatics: T-cell epitope prediction, HLA typing, epitope selection, and epitope assembly (Fig. 5.1, Tab. 5.1).

FRED 2 Immunoinformatics Use Cases

Epitope Prediction and Neoepitope Prediction

For a given set of HLA alleles and a given set of peptides, FRED2 can produce the HLA binding predictions for one or more chosen methods (Tab.5.1 - Epitope Prediction). Prediction scores are handled in a flexible tabular format. Peptide input can also come from protein sequences that FRED2 processes in a sliding window approach for a selected peptide size. With the input of genomic variants, FRED2 can generate all possible neoepitopes based on the annotated variants. Supported annotations come from ANNOVAR[245] and Variant Effect Predictor[246] using GRCh37 and GRCh38 human genome builds.

Cleavage Prediction

For a given set of protein sequences, FRED2 can produce the cleavage prediction under a chosen prediction model (ProteaSMM (C/S20)[214], PCM[213], NetChop[242], Tab. 5.1). The FRED2 interface allows for the specification of a target peptide length. The result are peptide sequences with their C-terminal cleavage score.

TAP Prediction

With the choice of three prediction models (SVMTAP[213], SMMTAP[243], Additive matrix[244], Tab. 5.1), FRED2 can produce TAP binding scores. Analogous to cleavage prediction, the FRED2 interface allows for the specification of a target peptide length. The result are peptide sequences with their predicted TAP score.

HLA Typing

HLA class I and II genotype can be inferred with typing methods integrated into FRED2. Depending on the method used (OptiType[62], Seq2HLA[64], Tab. 5.1), single-end whole exome, whole genome sequences, or RNA-Seq FASTQ files need to be provided. The most likely genotypes are reported in standard HLA nomenclature. The typing can be directly used in use cases that need HLA-allele type input.

Epitope Selection

From a set of input epitopes, the FRED2 implementation of OptiTope[225] can select a subset that maximizes the overall predicted immunogenicity. The integer linear programming-based approach also requires the assigned population frequencies for target HLA alleles, and optionally the number of epitopes to select and the percentage of HLA alleles and antigens that have to be covered by the selection. An additional constraint can be set for epitope conservation. The FRED2 implementation of OptiTope will calculate the conservation as the product of column-wise conservation from a multiple sequence alignment (MSA) of the epitopes. This constraint ensures that only epitopes that fulfil a user-defined conservation requirement will be considered. The set of input epitopes can be directly sourced from the results of the epitope prediction use case applications.

Epitope Assembly

Given a list of epitopes, the FRED2 epitope assembly implementations assemble a set of epitopes into an optimal vaccine construct to maximise the cleavage likelihood for full recovery of the individual epitopes after natural processing. Either a string-of-beads polypeptide can be constructed via a travelling salesman problem formulation (described by Toussaint et al.[220]) or a epitope concatenation with optimal spacer sequences maximizing the cleavage probability of the desired epitopes while simultaneously reducing the formation of neoepitopes (described in Schubert and Kohlbacher[221]). The set of input epitopes can be directly sourced from the results of the epitope prediction use case applications.

5.4 Discussion

We demonstrated the ready utility of FRED2 for immunoinformatics, with the support for multiple common-place use cases in the field. FRED2 is the (completely re-implemented) successor of FRED[204] with extensive tool support. We implemented

routines covering data pre-processing, HLA typing, epitope prediction, epitope selection, as well as epitope assembly.
Of particular use in the research of naturally processed HLA ligands can be the prediction of HLA-binding peptides for multiple alleles and multiple prediction methods. As seen in Lis. 5.2, predictions are ready for comparison after prediction and can be further processed together. For large studies or samples with particularly many variations, the generator based approach for neoepitope prediction can be memory conserving, as the sequences are generated on a need basis.
FRED2 provides a unified interface to many prediction tools. Many of the stand-alone HLA epitope prediction tools do not offer a unified interface and output format, which makes it difficult to use prediction methods interchangeably. One way to overcome these problems is web-based workbenches as offered by IEDB[247] or EpiToolKit[248]. But often data volume, speed, or legal restrictions (e.g., concerning data privacy) prevent the use of such applications. Additionally, there is usually a considerable lead time for the adoption of new or custom new methods. FRED2 offers custom application and workflow development with little overhead. These analysis tools in FRED2 have been stratified with a focus on the interoperability between the tools, including the prediction methods.
The complexity and development time of state-of-the-art immunoinformatics tasks is high. To maximize quality of the results and to decrease implementation time, it is common in many bioinformatics fields including immunoinformatics, that research analysis software makes use of already existing, thoroughly tested libraries. By building on top of popular modules such as BioPython[ii] and Pandas [iii], FRED2 allows rapid prototyping of complex and innovative immunoinformatics applications.
The open-source design of FRED2 was done with consideration towards flexibility to allow easy extension. With the benefits of a unified framework, in turn, the developer of a novel tool will not have to deal with data integration and can fall back on the unit tested components.
FRED and its successor FRED2 have proven the value of ensuring the availability of analysis tools, as, inevitable for most academic software, the developers change career, labs change focus, and websites go offline. For frameworks in general, the cumulative reduced maintenance efforts needed keep them compatible through iterations of current computing platforms, operating system, and programming language versions.
The structural parallels of frameworks like FRED2 and OpenMS, offer additional

[ii] `http://biopython.org`
[iii] `http://pandas.pydata.org`

benefits, which is the inherent compatibility within, and with flexible adaption, to interconnect with each other, which will be discussed in the next Chapter.

Chapter 6

Automated Workflows for Quality Control and Immunopeptidomics

This chapter includes partially identical or adapted content with permission from:

Workflows for automated downstream data analysis and visualization in large-scale computational MS

Aiche S., Sachsenberg T., Kenar E., Walzer M., Wiswedel B., Kristl T., Boyles M., Duschl A., Huber C.G., Berthold M.R., Reinert K. and Kohlbacher O.

(2015). Proteomics, 15, 1443-1447; https://doi.org/10.1002/pmic.201400391

ImmunoNodes – graphical development of complex immunoinformatics workflows

Schubert B., de la Garza L., Mohr C., Walzer M., Kohlbacher O.

(2017). BMC Bioinformatics, 18(242); https://doi.org/10.1186/s12859-017-1667-z

---

6.1 Introduction

Motivation

With the complexity of experimental methods in the life sciences, it is safe to assume that any complete data analysis will involve multiple steps, and thereby different tools. Common data standards for in- and output, and the frameworks' inherent compatibility, as described in the previous chapters, make this feasible.
The size of modern datasets and extent of studies makes it prohibitively time-consuming to manually start each analysis step's computation, wait for it to finish, then prepare the results to fit the input requirement for the next, then start the next, and so on. Here, automated successive execution of tools through batch execution, in the most

simple form via shell scripts[249], or, more evolved, through workflow management systems, can resolve some of the issues emerging. We will henceforth use 'workflow' as in the uninterrupted, consecutive and automated application of analysis tools to a particular set of data. Another expression which is often used in a similar context, 'pipeline', will be used as in drug development pipeline, including possibly multiple workflows, expert assessments, and (yet) non-automatable processes like literature research, model training, or follow-up experiments.
Further, most command line tools are not easily applied on large compute infrastructures or do not handle compute infrastructure failures in a way amenable to high-throughput use. Be it the recovery from a temporarily dysfunctional compute node or recovering from memory exhaustion during a particular analysis step for a particular part of data, discarding the successful computations and restarting from the beginning is clearly uneconomical. Failure strategies, and execution strategies in general, need to take the concrete compute infrastructure at hand into consideration, a feature that is largely out-of-scope for most (single-) tool developments but is addressed in overarching workflow management software[158,159,161]. The previous chapter illustrated how standard formats provide a stable interface between tools and harmonise the data handover. These are essential for robust workflows, i.e., swapping and updating tools without reimplementation of large parts of the workflow. With standard format integration also comes flexibility for the workflow design process, as chaining tools to achieve data analysis goals becomes straightforward, facilitating the direct in- and -output of software. It is therefore of interest to life-science research to have the tools used in the field fitted with standard format in- and output, and integrated into such workflow management software to leverage the advantages offered to analysis design and execution.

Background

The need for automation arose early in the history of computer-assisted data analysis. In 1979, the Bourne shell, a command-line interpreter with limited scripting capability was introduced to Version 7 Unix[250]. It was primarily aimed at improving working with the computer interactively, executing commands for the operating system, providing input and output redirection and, albeit limited, programmability through scripts. Later iterations of interactive command line interpreters, such as the Bourne-Again SHell (BASH), are to be found in virtually any Unix-like operating system, Linux being the most prevalent in the scientific communities' high-performance computing (HPC)

systems. Modern shell's scripting capabilities can be considered the most basic form of data analysis workflow system. Locally installed programs can be executed with local files, outputs redirected as input of successive program executions or to files, which are then used as input to the latter programs. These are however highly specific to the operating system and compute infrastructure for which they were developed, therefore lacking portability and reproducibility. The low-level functionality requires the scripting of recurring data manipulation operations to use them as workflows as defined above, which only further contributes to low portability, complex maintenance, and unimpressive user experience. Later in data analysis history, Platform Computing introduced Load Sharing Facility (LSF) (now IBM LSF) as workload management platform with shell script compatibility, extending the range of shell scripts to work efficiently with HPC systems on large datasets.
For the computational MS community, TOPP[251] introduced an intercompatible set of tools for proteomics data analysis, including signal processing, identification, and quantification. OpenMS[157] also provided a software framework with shared libraries of common functions and algorithms to speed up the development of new tools and ease the burden of maintenance. With TOPP, an automated workflow for proteomics data analysis could be scripted. Graphically assisted workflow editors like TOPP assistant (TOPPAS)[252] and Thermo Proteome Discoverer[253] were introduced to the community in the 2010s, which allow the easy construction of custom analysis workflows, using analysis software as building blocks, or nodes in a directed acyclic graph representing the data analysis flow. The 'building blocks' for TOPPAS are the tools of TOPP, where a notable tool, the `GenericWrapper`, allows other (external, command-line) tools to be used within its graphical user interface. Input or output incompatibilities had to be dealt with on the user side. Similar developments in the genomics community saw the introduction of Galaxy[158], which provides user-side graphical construction of a workflow of genomics tools in a web browser, and server-side data access and tool execution. KNIME[159] was introduced as a data analytics platform for visual design of workflows in a modular fashion, and offers a rich environment for data analysis, from statistics and visualisation to machine learning. The workflow design is modular, built through a succession of nodes for built-in data analysis tools, however no specific proteomics tools. The KNIME Community Extensions offer open-source KNIME nodes with bundled executables, ready for workflow design and data analysis. Extensions are built by different scientific communities, such as chemo- and bioinformatics, and enable building cross-domain workflows. A more generic approach to describe command-line tool based workflows for the scientific community was introduced with the Common Workflow Language[254] (CWL). As the name implies, workflows are described with

a declarative language and are independent of executors, which must be installed on the target compute infrastructure separately. There have also been efforts for an independent visual composer for CWL (`https://rabix.io/`), however not actively maintained anymore. The workflow orchestration system nextflow[161] has gained popularity within the computational proteomics community and foremost the genomics community, largely thanks to its data manipulation scripting capabilities and a built-in executor with a wide support of platforms to be executed on. In a similar fashion to CWL, nextflow workflows are implemented through a declarative language, and the input and output connections between tools have to be handled explicitly, therefore a robust knowledge of all tools involved in a workflow is required. Common to all workflow orchestration software is that in order to offer a specific command-line tool as a node, the tool's configuration options need to be known, among others the input/output types and available parameters. In nextflow, Galaxy, and CWL, these need to be described in the respective language or configuration environment. Similarly, but more framework neutral are CTD files[160], XML documents that contain all necessary information of a given tool. They are natively compatible with KNIME and TOPP, and through converters[i] also with CWL and Galaxy. They are a flexible add-on option for other frameworks such as FRED2 to integrate into larger workflow management systems. In the following, the focus will be exclusively on KNIME[159] based workflows, as the tools of interest for the following chapters have been first integrated into KNIME.

6.2 Methods

6.2.1 OpenMS in KNIME

The integration of OpenMS[169,251] into the workflow system KNIME[159] (v3 and later) extends data processing capabilities for proteomics data, allowing for sophisticated downstream analysis and visualisation. TOPP itself integrates external tools for extend pipeline capabilities (identification in particular) and extend popular tools with standard format compatibility as described in the Data Integration for Automated Workflows chapter. This is achieved with TOPP-adapters, for example, to established search engines such as MASCOT[255], X!Tandem[256], MS-GF+[173] or Comet[174].
OpenMS workflows can be constructed based on TOPP in the graphical user interface TOPPAS[252]. Designed for fully automated processing, TOPPAS does not provide functionality for downstream analysis and only limited visualization capabilities. KNIME

[i] `https://github.com/WorkflowConversion/CTDConverter`

offers a powerful and flexible workflow system combined with advanced data analytics, visualisation, and reporting capabilities. KNIME integrates nodes for machine learning, statistical data analysis, and interfaces to various scripting languages, for example, the statistical programming language R. KNIME's functionality can be easily extended with nodes provided via an online plugin repository (the so-called KNIME Community Extensions). As the execution of a KNIME workflow usually runs locally on a computer, it does not require extra IT security provisions beyond the usual steps on the operating system or file system level. For larger data, the execution can also be run on (often commercially) available cloud/cluster or server solutions, which meet today's security standards.

The integration of OpenMS into KNIME is based on the GenericWorkflowNodes project[ii] that generates KNIME nodes for any command line tool that provides a CTD description of the tool interface. We extended OpenMS to automatically generate CTD files for each TOPPtool. With the generated OpenMS KNIME nodes, a TOPPAS workflow can be recreated within KNIME. In contrast to TOPP, which' data input and output are file-based, KNIME's internal data flow is a table-based data exchange between nodes. To allow interaction between the file-based OpenMS nodes and regular KNIME nodes, we implemented a set of nodes to load the content of proteomics data files into KNIME tables. These nodes either use the OpenMS-specific `TextExporter` format or the recently published mzTab[257] format as input.

A KNIME/OpenMS workflow is composed of multiple nodes that are connected by ports. Ports represent single or multiple files that are passed from one tool to another. The number of incoming and outgoing ports depends on the individual tool and described by the tools' CTD, for example, a database search engine such as MS-GF+ or Comet will usually have two incoming ports, one for the file containing the spectra to be analysed and one for the protein sequence database to be searched. Nodes are drag-and-drop added to a virtual workbench and connected by drawing a line from the outgoing to the desired incoming port. For each generated connection between OpenMS nodes, the workflow engine will check if the file types are compatible, i.e., that only files of supported formats are given to a node. The parameters of nodes and their documentation are available via a configuration dialogue. On execution, each node checks, if the incoming data meets all requirements, for example, the correct type of input files.

Workflows are very rarely so simple that they contain only a linear sequence of nodes to process a single file. Therefore, additional nodes are provided to construct more complex workflows including loops and merge nodes. Loops allow applying the same

[ii] `https://github.com/genericworkflownodes`

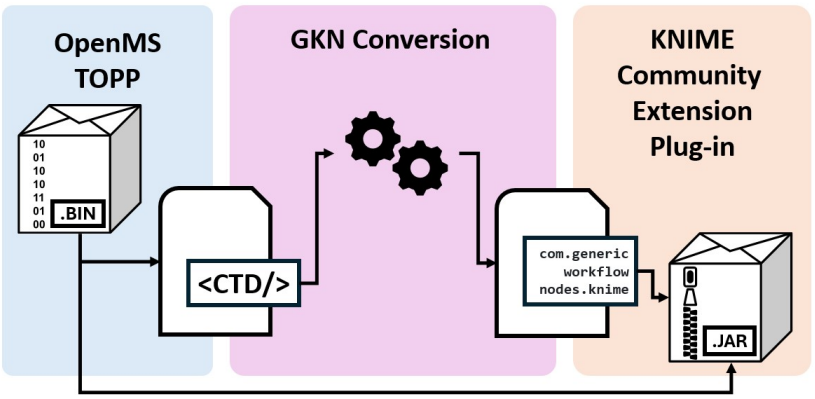


Figure 6.1: KNIME Integration of OpenMS' TOPP Tools

series of nodes to multiple input files one at a time. Merge nodes allow combining two files into a list that can be given to nodes that require more than one input file. To further structure a workflow, KNIME provides so-called Meta-nodes to group a collection of nodes. Grouping into Meta-nodes can be used to hide a complex series of nodes and instead provide a high-level view of the data flow.
KNIME can export complete, preconfigured workflows into self-contained ZIP files. Thus, workflows can easily be shared with collaborators, uploaded to web archives, or otherwise made accessible to the scientific community. Once configured, workflows can also be run from the command line for fully automated batch processing on a large number of files. With minor extensions, using KNIME's flow-variable concept, one can also configure KNIME such that the input files or other variable parameters of the workflow can be set from the command line when executed in batch mode. Consequently, workflows can be configured and tested in a desktop environment, using small subsets of the original dataset, and subsequently be deployed to large computational infrastructures.

6.2.2 FRED2 in KNIME: ImmunoNodes

Using the previously described GenericWorkflowNodes, CTD files for all compatible tools in FRED2 (Tab. 5.1) were generated as ImmunoNodes, the integration layer to KNIME. The ImmunoNode CTDs can then be used to automatically generate the KNIME plugin(Fig. 6.2). Based on FRED2, the ImmunoNodes also have dependencies to command line tools that are not bundled with FRED2. Several of these external tools are difficult to install or exclusively available on specific OS platforms. To address these

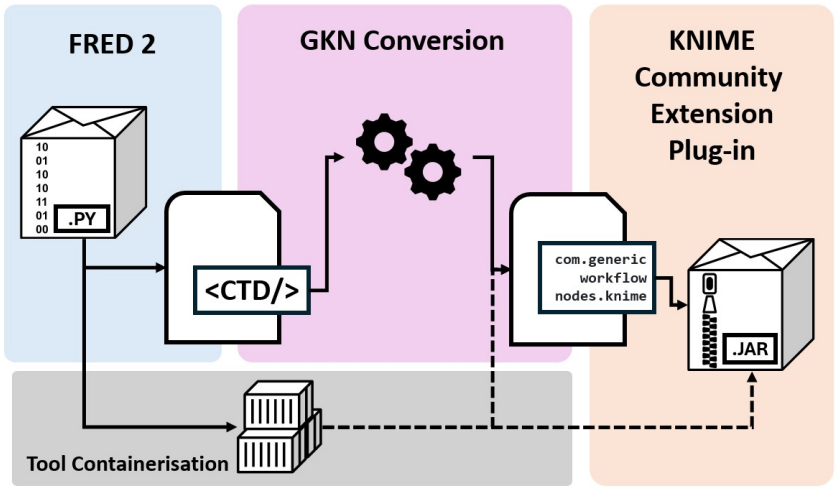


Figure 6.2: KNIME Integration of FRED2 Tools with the ImmunoNodes plug-in.

issues, we have extended generic KNIME node (GKN) to be able to execute command line tools provided within a Docker container. Docker as a software containerisation platform allows easy and platform-independent deployment of fully configured software suites, in this case, the ImmunoNodes tools to the KNIME plugin and subsequently the ImmunoNodes KNIME user. GKN scripts automatically generate the required Docker calls for the execution of an ImmunoNode in KNIME and handle the input and generated output redirection to-and-from the Docker container into KNIME. However, not all FRED2 tools are available as ImmunoNode tools due to either licensing or containerisation issues. ImmunoNodes offers twelve different nodes covering epitope, proteasomal cleavage, and TAP prediction, distance-to-self calculations of peptides, as well as HLA genotyping (Tab. 5.1). It also offers nodes for vaccine design including epitope selection and assembly.

6.2.3 Automated QC for HLA-ligandomics workflows

Built on existing analysis workflows for HLA-ligandomics MS data (Fig. 6.4), we used the OpenMS QC tools to collect qualitative data from the analysis in qcML and develop metrics to detect and address commonly occurring issues during acquisition and analysis. The base questions to be answered by the resulting metric values can be generalised as follows: Is the analytical system performing optimally? Can sources of analysis variability be identified? Both questions can be asked from the viewpoint of a particular run or over a span of instrument-operation time. And, are all runs of a particular set (e.g., a study variable) qualitatively the same or can we detect outliers? Combined, the metrics aim to provide effective quality control to MS acquisition and

analysis on the run level, instrument-performance level, and study level. We applied the quality measurements to 4 years of MS measurements of naturally processed HLA peptides from the Department of Immunology, University of Tübingen (n=12064). The runs were measured on a (LTQ) Orbitrap XL MS (Thermo Fisher) equipped with a nanoelectron spray ion source and coupled to an Ultimate 3000 RSLC Nano UHPLC System (Dionex). Gradient lengths varied between 90-215 minutes.

6.3 Results

The combination of different frameworks to create analysis tools that integrate data from different fields is an appealing option for integrated data analysis. The frameworks deal with the field-specific formats of (standardised) outputs, so the developers can focus on the analysis implementation. As discussed in the Data Integration for Automated Workflows chapter, OpenMS offers a framework to deal with data input and output, both custom and widely adopted standard formats. Additionally, many popular analysis methods and tools are integrated into TOPP, which makes it a tightly integrated platform, ideal for building robust data analysis workflows.

OpenMS TOPP Tool availability in KNIME

Integrating OpenMS through TOPP into KNIME enables the user to combine automated data processing of the raw MS data (signal processing, quantification, identification) with KNIME's data-mining and visualisation capabilities in a single workflow, for example, by directly integrating the initial raw data analysis with well-known R packages for proteomics data analysis (e.g., isobar[258]) or creating a vertically integrated workflow by utilising the available KNIME Community Extensions for cheminformatics (e.g., to visualise chemical structures of metabolites). The workflow can be constructed and configured in the KNIME GUI, which speeds up the design or restructuring for adaptation to new use cases. KNIME is easy to install on local computers, and simplifies the OpenMS installation to a plugin download. While the TOPP tools come with their own GUI, TOPPAS, it is limited in workflow design complexity and limited in the reach of tools. KNIME can also scale for deployment to larger or distributed compute resources. The initial data analysis results are compatible with KNIME's downstream data analytics tools, so results can be easily processed using the hundreds of KNIME nodes or custom R code implemented in the KNIME R nodes. Here, we present a workflow that takes advantage of the bi-directional data flow enabled by the OpenMS

KNIME integration and uses KNIME visualisation and KNIME R nodes to produce a rich QC report for a LC-MS/MS experiment (Fig. 6.3).

LC-MS/MS Quality Control Workflow

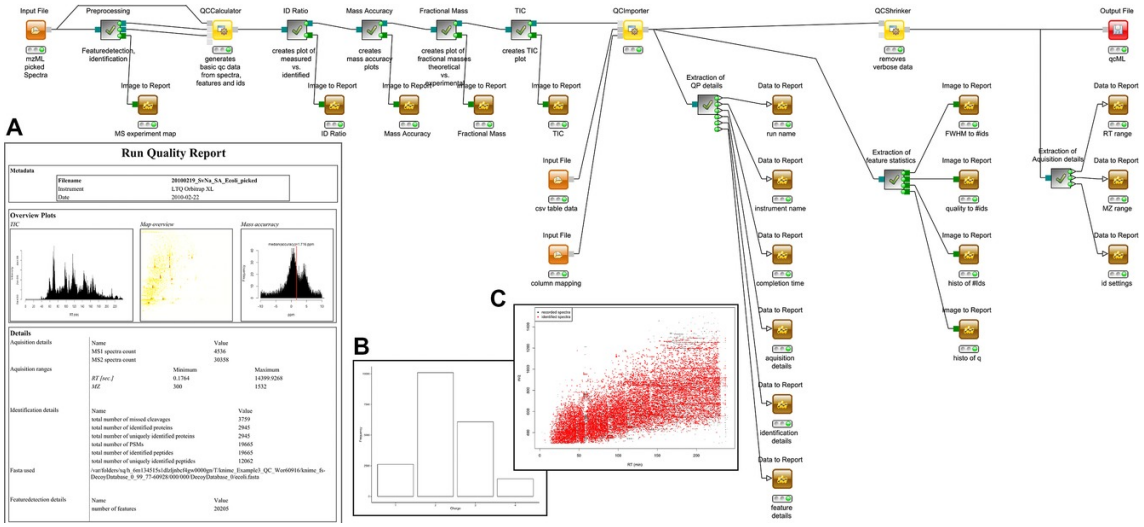


Figure 6.3: Proteomics workflow to compute and summarise a detailed quality report for an LC-MS/MS experiment. Insets (A-C) show parts of the generated quality report in qcML format, rendered in a web browser. (A) An overview of the analyzed file and details on the experiment, (B) charge distribution of found MS1 features, and (C) a plot comparing found MS1 features with peptides identified in the MS2 spectra.

With the increasing amount of data produced in LC-MS/MS experiments, quality control has become a crucial aspect of the day-to-day usage of MS. With the tools available through the OpenMS-KNIME integration, it is easy adapting existing analysis workflows to also incorporate quality control. For this, qcML[203], described in the Reporting Quality Control in Mass Spectrometry chapter, provides storage, transfer, and report rendering of quality metrics. Shown with the workflow in Fig. 6.3, reports can also be generated with a KNIME built-in reporting feature. The computation of the quality metrics, the generation of suitable plots, and integration in qcML can be achieved by adding few extra nodes to an existing workflow. A standard workflow for spectra identification at 5% FDR level and quantitative feature detection is collapsed into the 'preprocessing' metanode (see Fig. 6.4 for a full example of an identification workflow). Its output identifications and features, and also its original spectra input are input for the `QCCalculator` node, generating a basic qcML. The qcML file is handed along the consecutive metanodes that produce metric visualisations with R or Python and add them to the qcML. Optionally, the visualisations can also be added to the KNIME report. A `QCImporter` node exemplifies how additional QC data, not originating from

the workflow itself (e.g., protein content estimation or batch numbers of chemicals used during sample preparation), can be added to the qcML file. Several optional data extraction metanodes show how qcML data can be extracted and add more information that is natively presented in the qcML report rendering to the independent KNIME report. The 'Extraction of QP details' metanode reports metadata on the MS experiment like run name to KNIME. The 'Extraction of feature statistics' metanode reports the number of features and their charge states, etc. to KNIME. The 'Extraction of Acquisition details' metanode reports the RT and m/z acquisition ranges and identification settings to KNIME. The Fig. 6.3 insets A-C show the qcML report rendering in a browser as also seen in detail in the Reporting Quality Control in Mass Spectrometry chapter, and example metric visualisations. Metrics for MS experiments are presented and discussed in later sections of this chapter.

KNIME availability of FRED2 ImmunoNodes

Immunoinformatics has matured to the point where epitope prediction methods are widely used, from basic immunological research to translational applications[215,259]. High throughput methodologies based on liquid chromatography and MS have been successfully used to identify therapeutic targets for cancer immunotherapies[260–262]. However, these applications often require complex workflows combining multiple tools, multiple data sources and extensive pre- and post-processing.
With the integration of FRED2 into the KNIME Community Extension, and the OpenMS nodes, combined multiomics workflows can be constructed.

A Combined Multiomics Workflow: HLA Ligandomics Analysis

We showcase a KNIME-based application of computational MS and immunoinformatics, that leverages the Community Extension plug-in nodes of OpenMS and ImmunoNodes. The workflow represents a peptide identification analysis for ligandomics (Fig. 6.4). At the same time, this workflow exemplifies the synergistic effects of combining native KNIME nodes and different community extensions.

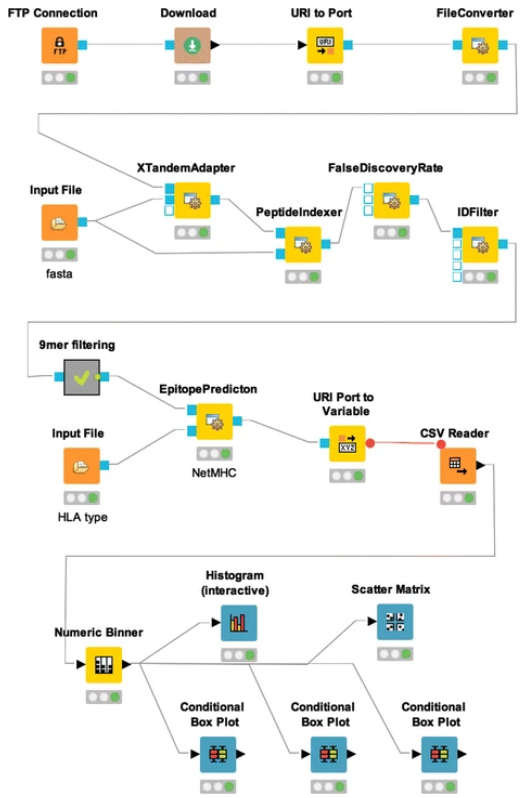


Figure 6.4: HLA ligandomics workflow combining native KNIME, OpenMS, and ImmunoNodes nodes. The workflow uses an online source for spectra data and performs mass spectra identification with the peptide search engine X!Tandem (`XTandemAdapter`). Identified 9-mer peptides are then annotated with their respective binding affinity as predicted by NetMHC using the EpitopePrediction node. Summary statistics and visualizations are generated with the use of native KNIME nodes.

The workflow extracts spectra data from PRIDE (file transfer protocol (FTP) Connection and Download node) and converts the data into standard format. Spectra identification is performed with the peptide search engine X!Tandem (`XTandemAdapter`), annotates the results with details of the given target/decoy database (`PeptideIndexer`), and calculates FDR (FalseDiscoveryRate). A filter node is applied to achieve 5% peptide-level FDR (`IDFilter`). The identified peptides are then selected for 9-mer length. This step comprises a number of native KNIME table manipulation and Python script nodes to accomplish the selection and is collapsed into a single KNIME Metanode for clarity. The peptides of appropriate length are then annotated with their respective binding affinity as predicted by NetMHC using the EpitopePrediction node. Finally, simple summary statistics and visualizations are generated with the use of native KNIME nodes. Exported KNIME workflows can easily be shared and reused by the whole

community[iii]. Reuse and customisation are simple, due to the parameter handling in KNIME, which greatly improves robustness, automatability, and workflow exchange.

Automated Workflow Quality Control Is Effective on Different Levels of Acquisition and Analysis of MS Experiments

Demonstrated in 6.3, fitting existing workflows with the OpenMS integrated QC tools (see Chapter 4), results in QC in parallel to the analysis with little overhead (Fig. 6.3). Here, we demonstrate how automated workflow QC is effective on different levels of the MS acquisition by showcasing selected metrics. Any HLA-peptide run measurement is the result of a complex sequence of processes, any of which can exhibit variable performance and hence introduce variability, or worse, compromise the successful measurement of the often limited amounts of available sample. This is generally true for most LC-MS/MS experiments with biological samples. The following metric descriptions show how to control for variability on different levels of acquisition and analysis in MS experiments.

Run Level

On the individual run level, QC metrics can be used to ensure the integrity of the run and to gather insights to identify issues with the analytical setup, including computational analysis. A significant amount of time is spent by instrument operators calibrating and optimising the instrument setup leading up to a measurable ion current[263] (i.e., the LC and the ESI) as it exhibits a parameter-rich environment, and is therefore also a major source for variability. As the ion current is the basic source of the run's spectra, a first measure for the QC on the individual run level is to monitor the total ion current (total ion current (TIC)) to check for interruptions and ensure there was a measurable ion current signal throughout the complete run time. We used the previously described QC tools from OpenMS with the HLA ligandomics analysis workflow to read the TIC as a basic QC metric and highlight interruptions in a TIC plot. In Fig. 6.5, the plotting script automatically highlights sections of RT where the signal dropped consecutively below $1e^3$ for more than 10 seconds. The 'gaps' in the measured ion current can occur inter alia because the electrospray of the ESI setup collapsed or not enough ionisable eluent exited the column. The TIC in the figure comes from run `150115_NT_HCC27_SIMSYNPEPSinJY_SIM#3_25cm90min3s_msms3`, a run that included spiked-in synthetic peptides measured in Targeted-SIM mode (denoted in the filename

[iii] `http://www.myexperiment.org/workflows/4947`

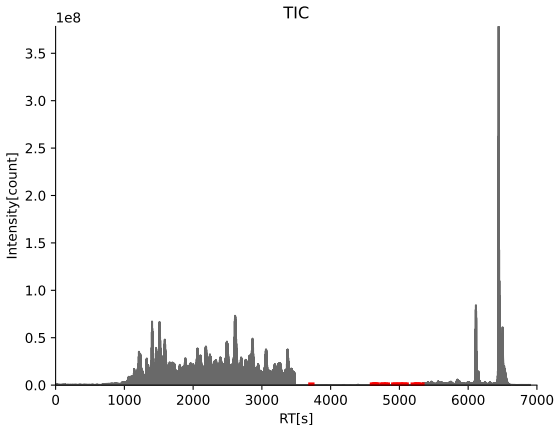


Figure 6.5: The total ion current of a MS run, reflecting the recorded ion signal over the LC gradient time. Signal gaps, marked in red, can occur from, e.g., a collapsed spray.

as SIM), eluting at known time points, after which an electrospray collapse did not cause sample measurement loss for the purpose of the run.
The signal threshold is not generalisable as the value range of signal intensity is influenced by the individual instrument and the measuring environment (e.g., different levels of noise for different instruments and different installations/setups/rooms).
A QC metric applied after the workflow's identification steps is the approximation of the observed mass error distribution. The mass error is measured as the difference of theoretical and experimental mass expressed in ppm for each identified spectrum (and its accepted PSM). Plotting the mass error along RT can yield further insight into the instrument's performance during the run, as seen in Fig. 6.6. The fitted LOWESS

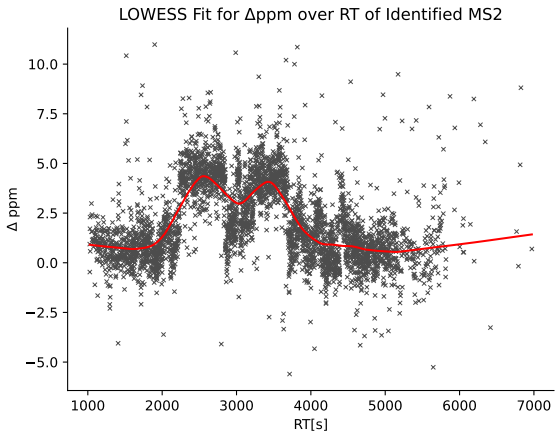


Figure 6.6: The mass errors from the identifications of a MS run plotted over the LC gradient time. Shown in red is a locally weighted scatterplot smoothing (LOWESS) fit of the data, highlighting two plateaus during which the mass error was at a consistently higher level, indicating a potential lockmass calibration issue.

function is optimally expected to take the value of the mean (i.e., $f(RT) = mean_{\Delta ppm}$). Deviations from a horizontal progression point to periods of RT where the instrument's acquisition is performing sub-optimally. The mass errors come from the identifications of the run `131223_DK_CLL56_postTherapy#1_2ndTry_W_16%_#3_msms4`. A temporary 'loss' of the external lock mass can result in a stretch of identifications for which' MS2 the recorded precursor mass was inaccurately recorded, resulting in a cluster of identifications with 'off-centre' mass error. Suboptimal lock mass correction, can stem from imperfect peaks caused by weak signals due to the method of recalibration[264] or the actual (temporary) loss of the known mass ambient ions registered for mass recalibration.
The identifications can be subjected to further scrutiny by relating the theoretical hydrophobicity of all the identified peptides to their spectras' respective order in the LC gradient. The hydrophobicity can be estimated by calculating the grand average of hydropathy (GRAVY) of all amino acids in the peptide, using a hydrophobicity scale[265].

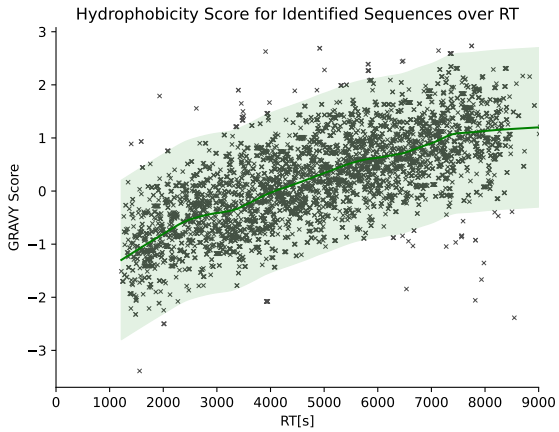


Figure 6.7: GRAVY hydrophobicity score calculated from identified sequences over the LC gradient time. The middle line in dark green shows the LOWESS fit of the data, and a ±1.5 'confidence' band around.

As seen in Fig. 6.7, the LC gradient of the run `150526_DK_JY_Standard15#1_postRSLCmaintenance_50cm195min3s_msms2` produces a clearly visible trend of hydrophilic peptides (negative values) eluting early to hydrophobic peptides (positive values) eluting later. Drawn around the LOWESS fit of the data is a ±1.5 band as rough guide for GRAVY score tolerance, since elution order of complex samples can vary, depending on factors other than estimated by the score (e.g., variability in the mobile phase pH, temperature, etc.). Though protein modifications can affect RT and are not considered in the calculation of the hydrophobicity score, late peptide identifications with a hydrophobicity score as low as some identifications at

the beginning of the measurement are unlikely to be the correct peptide assignments. However, only careful calibration with comprehensive standards of known synthetic peptides can produce a confidence band for a given LC implementation.

Instrument-Performance Level

An essential requirement for the successful analysis and reliability of results from (high-throughput) MS with automated workflows is the proper functioning of instrumentation[266–268]. The number of parameters involved in calibration, influencing the stability of performance make a stable performance challenging, especially when datasets involve runs collected over a period of weeks or months. Naturally, longitudinal QC monitoring of the results obtained from a particular instrument over its operation history can alert to decreased performance, give clues to the investigation of the parts of the instrument setup at fault, support inter-laboratory comparisons and the assessment of instrumental setups. Utility from monitoring general analysis results is however limited, as the measured samples are subject to variation from experiment to experiment, due to the sample nature, preparation protocols, data processing methodology, and instrument performance. The use of dedicated QC samples of known content[269] and consistent QC sample analysis workflow have been proposed[270] to minimise impacts from the former three sources of variability to control the latter instrument performance variability.
In the absence of dedicated QC sample protocols for HLA-peptide MS analysis, measurements of sample preparations of the JY cell line can be used as an approximation of known content. The JY cell line is an Epstein-Barr virus (EBV)-immortalised B cell lymphoblastoid line. Cell line samples have been suggested as complex QC sample for regular measurements[269] and the homozygous nature of HLA-A (A*02:01:01:01), HLA-B (B*07:02:01:01), and close relation of HLA-C alleles (C*07:02:01:01, C*07:02:01:03) in JY[271] reduce sample variability compared to heterozygous cell lines, making it an adequate complex QC sample for HLA-ligand MS analysis.
Here, we plotted four basic (observational count) QC metrics from a unified computational analysis of 52 JY samples measured over a ∼3-year time course (2013-2016; Fig. 6.8). The metrics observe the number of recorded MS2 spectra, the number of proteins from which peptides were detected, detected peptide features, and of those, the number with an assigned identification. As indicated by the markers in the plot, we noticed two sharp drops in performance, most notably in the per-run number of detected peptide features and number of acquired MS2 spectra. Dip a) occurred in a run directly before a separation column change, followed by a return to previous

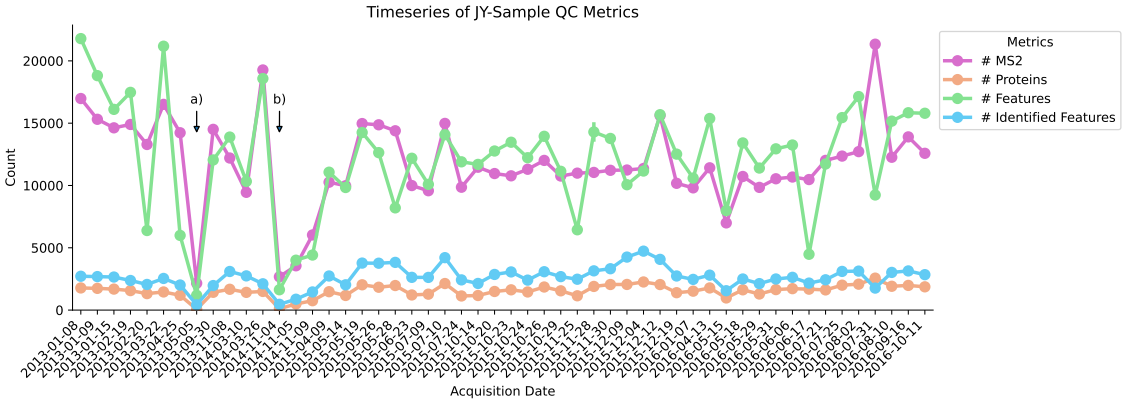
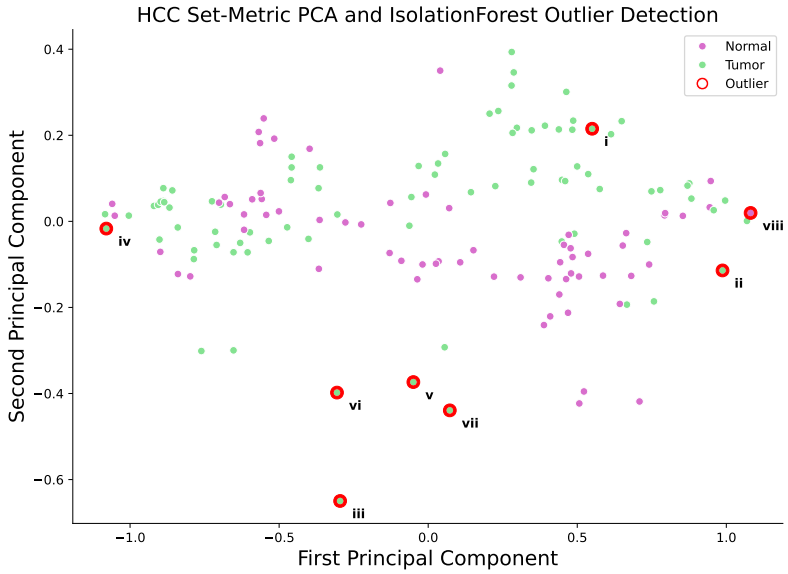


Figure 6.8: Timeseries of JY-sample runs as QC sample representations. Four different basic (observational count) metrics are plotted over instrument operation time course. Sharp drops in all metrics indicate possible instrument issues a)&b).

levels. A second dip b) occurred after the instrument's turbo-molecular pump exhibited performance issues, warranting replacement and additional replacement of the trap column (purpose: removal of anion or cation contaminants from eluents) to restore instrument performance to previous levels.

Study Level

On the study level, QC metrics can be used to check the integrity of the experimental design, assure comparability of the individual measurements within a study variable, and detect outliers, which may, if the study guidelines allow, be removed. Best practice may involve the removal of outliers after determination that a technical issue prevented a correct measurement. QC metrics can be used for the detection of outliers and to provide clues to where an issue might have compromised a measurement. In complex measurement systems such as LC-MS/MS, reliance on a single metric may reduce the potential of QC, especially to detect outliers and multivariate methods are preferred. The Isolation Forest algorithm[272] is an ensemble method, using multiple base models to derive a more robust predictive model. Here, the base models are Isolation Trees, performing recursive random partitioning of a multidimensional dataset (e.g., a set of metrics for each run under analysis), effectively isolating a datum by a number of successive partitions from the rest of the data, which can be represented by a tree structure. As such, Isolation Forests are a specialisation of Random Forests[273]. Outliers, by definition distinct and less frequent than normal observations, will be isolated with fewer partitions on average than 'normal' data. This distinguishes the Isolation Forest algorithm as an outlier detection method from other methods like clustering (e.g., KNN)



(a)

Outlier	Run Name	Probable Outlier Reason
i	131025_NT_HCC24_Tumor_W_20%_#1_50cm195min5s_msms6	Sample Issues
ii	140124_NT_HCC25_Tumor_W_20%_Rep#1_25cm90min3s_msms1	Shorter Gradient
iii	140221_nt_HCC27_Tumor_W_20%_Rep#1_25cm90min3s_msms1_SprayFailed	Spray Failed
iv	140704_NT_HCC30_Tumor_W_20%_Rep#1_25cm90min3s_msms25	Shorter Gradient
v	141203_NT_HCC26_Tumor_W_20%_Rep#4_SIM#3_msms3	SIM Mode
vi	150206_NT_HCC27_TUMOR_W_20%_Rep#2_SIM#1_50cm195min3s_msms10	SIM Mode
vii	150206_NT_HCC27_TUMOR_W_20%_Rep#3_SIM#2_50cm195min3s_msms11	SIM Mode
viii	150206_NT_HCC27_TUMOR_W_20%_Rep#4_SIM#3_50cm195min3s_msms12	SIM Mode

(b)

Figure 6.9: Multivariate isolation forest outlier detection applied to the set of HLA-ligand detection runs from hepatocellular carcinoma (HCC) tissue of the individualised immunotherapy study (Chapter 7). a) principal component analysis (PCA) 2D visualisation of the multivariate set metric (number of MS2 per run, number of PSM per run, number of peptide features detected per run, number of peptide features identified per run, number of source proteins identified; n=151), isolation forest outlier detection applied to the same set metric overlaid. b) Table of detected outlier runs and probable outlier reason. Run names are displayed as recorded in the acquisition system, the name encodes i.a. the acquisition date (e.g., '131025'), sample identifier (e.g., 'HCC24'), sample type (e.g., 'TUMOR') and measurement mode ('SIM#1_50cm195min' for SIM mode with a 50 cm LC column on a 195-minute gradient).

and classification approaches (e.g., support-vector machine (SVM)), which inherently detect or learn groups of observations as 'normal', and label the remainder as outliers. As such, the Isolation Forest algorithm is an unsupervised method, making it effective for unlabelled data, and well suited for application to new sets of data like a study dataset.

We performed the Isolation Forest outlier detection for demonstration on the dataset of all HLA-peptide detection runs from HCC samples included in the Personalised Immunopeptidomics study discussed in the Applications - Personalised Immunopeptidomics chapter. Included were both samples from normal and tumour tissue, with a varying number of runs per sample depending on sample availability, totalling n=151 runs. The dataset features, i.e., set-QC metrics, were chosen from the observational counts of the runs' MS2, detected peptide features, PSM, identified peptide features, and identified peptide source proteins.

For visualisation purposes, we performed a PCA dimensionality reduction of the metric dataset to highlight the detected outlier runs in 2D (Fig. 6.9 a)) and annotated the respective runs with probable outlier causation in a table (Fig. 6.9 b)). Outlier iii was found to have the electron spray failed during acquisition, and marked as such. As explored in Fig. 6.5, a failed spray can leave parts of the LC gradient without MS acquisition, hence making portions of the sample unavailable to peptide-feature detection and identification. Outliers ii, iv were measured with a shorter gradient, making them less comparable to the rest of the runs. For outliers $v, vi, vii, viii$ a semi-targeted acquisition mode was used. Here, parts of the LC gradient were focused on the acquisition of a specific mass range in a selected ion monitoring approach. The mass and RT ranges were derived from potentially mutated HLA-ligands as predicted with input from the sample's genetic sequencing and in-silico HLA-binding prediction (details see Chapter 7).

The sample from outlier i was found to be challenging and a second tissue lysate sample preparation was run on the same day (`131025_NT_HCC24_Tumor_LysatReRun_W_20%_#1 ↪ _50cm195min5s_msms12`). Some of the detected outliers are are not directly apparent as such from the PCA plot of first and second principal component, but appear different when considering their position in plots of the other principal component combinations (see Appendix). Also notable is that all detected outliers were from tumour tissue samples, underpinning that the detection of HLA-ligands from tumour tissue is complicated by sample varying factors such as volume and immunoprecipitation yield.

Add-on QC

Using the JY-sample measurements as a reference, though HLA peptide analysis measurements can vary widely depending on the sample source, the application of a QC add-on to the HLA peptide analysis workflow required to spend on average only 7.1% (∼20 s) of total CPU time on QC. The minimal overhead is due to the reuse of the intermediate results of the workflow for QC purposes and a lean implementation with efficient data integration.

6.4 Discussion

Structuring complex data analysis tasks into a collection of small, easily executable, simpler computations brings the benefit of adding a certain degree of reproducibility, an aspect desired in all scientific endeavours. From a software development perspective, this meshes well with the UNIX philosophy: write programs that do one thing and do it well. By definition, these tools cannot perform a complex analysis and therefore need software to orchestrate. Workflow orchestration software such as KNIME offers advantages to the data producer in need of developing custom analysis workflows to process their data. KNIME offers the visual construction of workflows through the chaining of nodes representing the workflow steps/data analysis tools necessary to complete a given analysis. Such workflows also represent a superior way to share and communicate complex analysis methods over monolithic software constructs. Workflows can be shared ready for re-use as we demonstrated with the combined multi-omics workflow for HLA peptide analysis (6.3). Their modularity allows for a better overview of the data analysis steps taken and better sustainability, should better alternatives for a given step become available or novel use cases/data necessitate slight changes in the workflow. Workflow orchestration also enables the automated, high-throughput analysis of large datasets. These, as discussed in the context of MS proteomics standard formats (Chapter 3), are becoming more common, as life-science disciplines are moving towards comprehensive data collections using systematic information capturing.
The frictionless combination of tools, next to operating system and other software requirements, depends heavily on the compatibility of inputs/outputs. The combination of data from different scientific disciplines is becoming more effective the better tools can interoperate. For this, as for any general use case involving the handover of data from one software to the next, it is highly advantageous to have standardised data formats to facilitate unambiguous data exchange between software, which has been discussed in detail in the previous chapters.

In this chapter, we demonstrated, how data analysis tools from these different disciplines can come together in shared workflows through the integration into a common workflow orchestration system, KNIME. KNIME has the advantage over web-based workbenches like Galaxy because of the better options to locally develop integration solutions and later run workflows flexibly on different (local) systems. This is in part due to the server-based setup of web-based solutions, which usually provide only limited or at least hard-to-access integration development opportunities. As a consequence, novel resources and methods for scientific disciplines with smaller communities, like immunoinformatics, are available first in local solutions such as KNIME. It's utility becomes apparent when comparing the visual capabilities to more specialised solutions like Galaxy and TOPPAS. Another benefactor for KNIME is a thriving extensions system to integrate more tools. For immunoinformatics, we showed with the implementation of FRED2, the benefits of a unified framework for novel tool development and the stratified integration into greater workflow orchestration with the ImmunoNodes. Being fully integrated into KNIME using GKN, the ImmunoNodes enable a wide audience to develop complex analysis workflows without the need to have mastered a programming language. Also, the complexity of installation and configuration of required third-party libraries has been lifted from the end user as a result of the provided Docker images. In a similar fashion, the provision of OpenMS and TOPP tools has been simplified for the data producer/analysis architect with the OpenMS community nodes. The same benefits as presented by the ImmunoNodes for the immunoinformatics community, are available for the proteomics community. Although the TOPP tools already had a visual workflow environment available with TOPPAS, the broader scope of compatible tools from different fields of the life sciences has the promise to significantly impact multi-omics study feasibility in the future. The easier sharing and reuse of workflows contribute to deeper insights into data and the development of more robust procedures in regards to reproducibility, due to the applicability to similar datasets from different data producers. To assure the comparability of those datasets and in general control the integrity of measurements, QC methods can be applied in the same workflow or dedicated workflows to conduct dedicated quality control of the involved instrumentation. Run-level monitoring is the first measure to control the success of a measurement and can be directly included in the data analysis workflow. The instrument setup of LC-MS/MS measurements usually involves many points of failure and QC in line with the data analysis makes particular issues easier to pinpoint. It is a rational choice to monitor two of the biggest contributors to variation, the LC and ESI, for each run. As demonstrated with the TIC plot (Fig. 6.5), using dedicated QC metrics to monitor usual suspects as sources of issues can be a simple but often time-saving action. We also rec-

ommend monitoring along the (often hours long) RT of a run to detect inconsistencies during the run as seen in Fig. 6.6, which can potentially be rectified post-acquisition with recalibration methods[274,275] to improve the data analysis.
Longitudinal QC monitoring has proven an effective method to proactively keep instrumentation in optimal working condition in proteomics[269,276,277] and related fields like metabolomics[278,279] to reduce sample waste due to in-measurement instrumentation failure or performance degradation. It shows promise to have the same effects for HLA-ligandomics, where low-concentration samples dominate. For such a specialised case, a dedicated QC sample measurement operating procedure would need to be designed. This would have to include 1) the design of a QC sample to mirror the type of peptides in the ranges of concentrations usually seen within a sample of HLA-ligandomics, *ii*) a strictly standardised sample preparation method, and *iii*) the establishment of a baseline for each combination of instrument setups and measurement modes. While the JY-sample measurements used in Fig. 6.8 can arguably serve as an appropriate base for such a QC sample, we still observed a good amount of variation in the measurements that can be in large parts attributed to the use of different instrument setups (e.g., different LC gradients) and less obviously to minor deviations from a common sample preparation method. With the outlined improvement in QC sample application, it is conceivable that a predictive maintenance regimen could be established improving the chances for predictive algorithms to detect arising issues and enable more timely maintenance interventions, improving overall consistency between measurements.
For consistency, especially in a study setting, (multivariate) set-metric QC can be of aid. To assure all of the measurements included in a study are comparable, variation should be attributable to either measurement sensitivity deviation, checked by comparison to QC sample runs over the study's data acquisition period, or variation due to sample concentration differences. For the latter, methods of outlier detection from multiple QC metric values can be used as demonstrated in Fig. 6.9. Variation from sample differences can arise either because of expected study variable differences (e.g., tumour vs. normal tissue) or unexpected differences hinting at sample preparation issues, either of which need to be explained should the study require the samples to be compared.
Isolation Forest is an unsupervised detection algorithm, i.e., it does not need to train on labelled data which is of advantage for the application to MS runs, where variation from many different sources can express in very different kinds of anomalous data, unlikely to be covered by even exhaustive training data set. Because Isolation Forest uses a random ensemble of isolation trees for anomaly estimation, the algorithm can be used online, training with subsampling from the input data set to be analysed. Another

benefit of the algorithm is the indifference to different feature scales, meaning any difference in value scaling of the combined QC metrics does not influence the detection. In many distance-based detection algorithms, e.g., KMeans, the feature scales may create an implicit weighting of feature importances, given they are different enough, and additional methods for normalisation need to be applied. The Isolation Forest algorithm depends on the characteristics of anomalies to be few and different. In general, outlier detection algorithms are less effective in detecting collective anomalies like runs affected by a defective part in the instrumentation. There, longitudinal monitoring with QC samples can help to detect systematic anomalies, given a sufficiently high QC sample frequency is applied.

Chapter 7

Applications - Personalised Immunopeptidomics

This chapter includes partially identical or adapted content with permission from:

Multi-omics discovery of exome-derived neoantigens in HCC

Löffler MW, Mohr C, Bichmann L, Freudenmann LK, Walzer M, Schroeder CM, Trautwein N, Hilke FJ, Zinser RS, Mühlenbruch L, Kowalewski DJ, Schuster H, Sturm M, Matthes J, Riess O, Czemmel S, Nahnsen S, Königsrainer I, Thiel K, Nadalin S, Beckert S, Bösmüller H, Fend F, Velic A, Maček B, Haen SP, Buonaguro L, Kohlbacher O, Stevanović S, Königsrainer A, HEPAVAC Consortium & Rammensee HG

(2019). Genome Medicine, 11(28). `https://doi.org/10.1186/s13073-019-0636-8`

7.1 Introduction

Motivation

Personalised Cancer Immunotherapy has great appeal because of its targeted approach and leveraging of the body's own defence mechanisms. Targeted, because it allows the selection of tumour specific antigens present on a patient tumours' cancer cells, introduced by cancer specific mutations or the change of expression. Though all cells accumulate mutations, including the usual and common hallmark mutations that are found in cancer, the mutations are random in a general sense, and present a tumour-specific pattern, not present in healthy tissue. This pattern should ideally trade down to the immunopeptides presented on the patients' cells, making cancer cells distinguishable by the host immune system with so-called neo-epitopes, and therefore present a target for personalised cancer vaccine therapy. A cancer vaccine should then elicit *de novo* T cell responses targeting the tumour cells, thereby leveraging the body's

own defence mechanisms against the cancer.
MS is essential in the development of personalised vaccines as it is able to identify and validate specific targets. One aspect of such an analysis, in particular in a clinical setting, is the quality control of the data being acquired. Another aspect is the automated processing for a timely and reproducbile analysis, affording the vaccine formulation and synthesis the necessary lead time. Based on the work of the previous chapters, we have thus developed a data analysis workflow able to analyse the multi-modal omics data generated for the analysis of personalised immunopeptidomes. As part of a larger study, we have applied the workflow to the clinical pipeline for exome-derived neoantigen discovery in HCC. We performed unprecedented in-depth multi-omics analyses encompassing whole exome and transcriptome sequencing, combined with proteome and HLA ligandome profiling in selected HCC patients aiming to obtain evidence for the natural presentation of exome-derived mutated HLA ligands.

Background

HCC is among the malignancies with the highest death toll on a global scale[280] and with very limited therapeutic options. Particularly in advanced stage, long-term survival is uncommon[281]. Although it has been shown that the microenvironment of the liver is tolerogenic and impairs immune responses[282], antigen-specific T cell responses do occur[283]. Since infiltration of HCCs with T cells[284] and spontaneous immune responses correlate with longer survival[285] but mostly prove weak and insufficient on their own, immunotherapies unleashing the immune system hold great promise.
Immune checkpoint (ICP) inhibitors demonstrating the potency and effectiveness of the immune system to fight malignancy[286] have set the stage for cancer immunotherapies. In contrast to established cytostatic treatments for cancer, this new class of drugs has enabled long-term survival in advanced and metastatic disease previously considered incurable[287]. However, although in some malignancies ICP inhibitors have proven highly effective, results for other cancers remain disappointing. One probable mode of action for ICP inhibitors is the induction and/or restoration of T cell effector functions against individual somatic tumour mutations presented by HLA molecules (i.e., mutated neoepitopes)[288]. Since these mutated HLA ligands were unacquainted to the immune system before carcinogenesis, they have been proposed as ideal tumour-specific targets[289,290].
In malignant melanoma (Mel), where ICP inhibitors were established first, mutational load was shown to strongly correlate with survival[291]. This has been corroborated in lung cancer[292] and colorectal carcinoma, where in the latter impressive survival

benefits remained strictly limited to mismatch repair-deficient carcinomas featuring very high numbers of mutations[293]. As elevated somatic mutation rates raise the odds for generating neoantigens, this supports the notion they may be critical for ICP inhibitor effectiveness[294]. Another line of evidence suggests that neoantigens recognised by T cells can generate impressive clinical effects, when identified and exploited for therapeutic purposes. This has been shown in remarkable case reports inter alia in advanced Mel[295] and metastatic cholangiocarcinoma[296].
With current affordable NGS and bioinformatics, an array of approaches predicting HLA-restricted neoantigens from virtually any tumour has emerged[297–299]. Indeed, at present most attempts are restricted to in silico analyses, lacking actual proof that the predicted neoantigens are relevant or even exist. So far, tangible evidence is scarce and mainly restricted to T cell recognition[300]. Therefore, one frequently missing link is proof of neoantigen presentation on native tumour tissue. Such an endeavours is very challenging and has been shown feasible in mouse models[301] and cell lines[302] but in human solid tumours hitherto merely in Mel at low numbers using MS, defining the current state-of-the-art[303,304].
Since both individual cancer traits and mutational load vary strongly between different tumour entities[305,306], these properties may ultimately restrict the foreseeable success and feasibility of neoantigen-targeted precision cancer medicine. In HCCs, only a small proportion of about 10% of patients showed mutations potentially accessible for drug therapy[307], whereas preliminary data for ICP inhibitors showed objective response rates in 15-20% of patients combined with a manageable safety profile[308], making neoantigens in principle an interesting case for precision cancer medicine.

7.2 Methods

HLA Ligandomics Data Analysis

MS data obtained from HLA immunoprecipitates was analysed using OpenMS(v2.3)[309] TOPP tools. Identification and post-scoring were performed using the TOPP adapters to Comet 2016.01 rev. 3[310] and Percolator 3.1.1[311] and filtered with a FDR threshold of 5% at PSM level. Database search was performed against a personalised version of the human reference proteome (UP000005640[i]), including the patient-specific mutanome without enzymatic restriction and methionine oxidation as the only variable modification.

[i] `https://www.uniprot.org/proteomes/UP000005640`, accessed: 29.02.2016

HLA Typing and Peptide Prediction

To define the sample-specific mutated peptide search space, peptides of 8-11 amino acid length were constructed by sliding a shifting window of the peptide length over the affected mutated positions. Resulting peptides were filtered against the human reference proteome (UP000005640[ii]) and the Ensembl proteome reference (release 84) to exclude peptides contained within wild-type proteins. Transcript information was retrieved via BioMart, based on the stable database version of GRCh37 [iii]. HLA typing at four-digit resolution using whole exome sequencing data was performed with OptiType[312] for HLA class I alleles as previously described[313] and confirmed in selected cases by molecular HLA typing (using clinically validated LUMINEX and sequence-based typing) during clinical routines. HLA-binding prediction was performed with SYFPEITHI[314], netMHC 4.0[315,316], and netMHCpan 3.0[317,318]. The workflow was implemented using FRED2[319]. Data management was performed through the qPortal instance at the Quantitative Biology Center (QBiC), Tübingen[320].

7.3 Materials and Data

Clinical Specimens

Clinical specimens from patients (n=16; median age: 74 years; min.-max. 55-85 years; 75% men) undergoing liver resection for HCCs, encompassing both non-malignant and malignant liver tissue as well as peripheral blood, were obtained directly after surgery and cryopreserved. HCC diagnosis and predominant tumour fraction within samples were histologically confirmed by an expert pathologist. All included patients were negative for chronic viral hepatitis (hepatitis B and C) and without systemic pretreatment for their malignancy.

Next-Generation Sequencing

DNA and RNA were extracted from fresh frozen tissue and PBMCs, respectively. After sample preparation and enrichment, paired-end whole exome sequencing and whole transcriptome sequencing were performed on an Illumina system at the Institute of Medical Genetics and Applied Genomics, Tübingen, Germany.

[ii] `https://www.uniprot.org/proteomes/UP000005640`, accessed: 29.02.2016
[iii] `http://feb2014.archive.ensembl.org`, release feb2014

Isolation of Naturally Presented HLA ligands from Tissues for HLA Ligandomics

HLA class I-peptide complexes were isolated from HCC and corresponding (non-malignant) liver tissue samples by immunoaffinity purification using the pan-HLA class I-specific monoclonal antibody W6/32[321] (produced in-house at the Department of Immunology, Tübingen, Germany) and eluted using 0.2% trifluoroacetic acid as described previously[322].

Analysis of HLA Ligands by LC-MS/MS

HLA class I ligand extracts were measured once or in multiple technical replicates, as described previously[322,323]. Samples were separated by HPLC and eluting peptides were analysed using CID in an online coupled Orbitrap MS. In addition to DDA, selected ion monitoring (SIM) and PRM targeted tandem MS (tMS2) was performed for selected samples to enhance the sensitivity and robustness of neoantigenic peptide identification.

Variant Calling

Reads were processed using the megSAP pipeline[iv] and the ngs-bits package[v] by the Department of Medical Genetics and Applied Genomics (Tübingen, Germany). Reads were mapped against the Genome Reference Consortium Human Build 37 (GRCh37) using BWA-mem[324]. Somatic variant calling was performed using Strelka and Strelka2[325,326] or with a proprietary software (CeGaT GmbH, Tübingen, Germany). Somatic mutations were annotated using SNPEff 4.1k[327].

Gene Expression Analysis

Gene expression values were calculated as fragments per kilobase of exon per million reads mapped (FPKM) of the corresponding transcripts and RNA tumour sequencing depth at the corresponding variant position. Mapping of RNA reads was done using TopHat 2 (v2.0.12)[328].

Protein In-Gel Digestion for Shotgun Protein Identification

Sample lysates were separated by SDS-PAGE. Coomassie-stained gel pieces were digested using trypsin. Peptides were desalted using C18 Stage tips. LC-MS/MS analyses were performed on an EasyLC nano-HPLC system (Proxeon Biosystems, Roskilde, Denmark) coupled to an LTQ Orbitrap Elite MS (ThermoFisher). Resulting data were

[iv] `https://github.com/imgag/megSAP`
[v] `https://github.com/imgag/ngs-bits`

processed with MaxQuant software suite v.1.5.2.8[329]. The human reference database was obtained from UniProt (containing 91,646 protein entries and 285 commonly occurring laboratory contaminants) and concatenated with the patient-specific mutanome. Endoprotease trypsin was fixed as enzyme with a maximum of two missed cleavages. Oxidation of methionines and N-terminal acetylation were specified as variable modifications, and carbamidomethylation of cysteines defined as a fixed modification. Initial maximum allowed mass tolerance was set to 6 ppm. Re-quantify was enabled. A FDR of 1% was applied at peptide and protein level. Proteome analysis was performed at the Proteome Center Tübingen (PCT), Germany.

Dataset Availability

The MS data generated and analysed during the study, including proteome and ligandome data, are publicly available through PRIDE (dataset identifier PXD013057). Somatic variant lists and expression data generated and analysed during the study are available from figshare[vi]. The whole exome and transcriptome sequencing raw data generated and analysed for the study are not publicly available, as participants did not provide respective informed consent for broad sharing of their data.

Alternative dataset for additional workflow validation: Mel

For additional workflow validation, we also processed a publicly available dataset[303] of somatic variants from whole exome sequencing (EGAS00001002050) and HLA-peptide MS data (PXD004894) of five Mel patients as a reference.

7.4 Results

7.4.1 A Multi-Omics Approach to Detect Mutated HLA Ligands in HCC

We performed analyses of malignant and non-malignant liver tissue, resected during surgery for HCCs, by a multi-omics approach encompassing analyses on exome (n=16), transcriptome (n=16), shotgun proteome (n=7), and HLA ligandome level (i.e., HLA-presented peptides; n=16). Multi-allelic HLA class I expression was confirmed in all patients of our HCC cohort. The overall aim of our research was to identify individual exome-derived somatic tumour mutations resulting in natural HLA ligands presented to T cells.

[vi] `https://figshare.com/s/6c09d3095a32402b4717`,`https://figshare.com/s/c02d184d8f55a813456a`

Detection of Somatic Variants in HCC

On average, we detected 151±40 somatic variants (Var) per HCC, including single nucleotide variants, small insertions/deletions, and frameshift variants; thereof, 44% (66±19) cause changes in the amino acid sequence of the encoded protein (i.e., non-synonymous variants; Var^{ns}), when referenced against DNA from blood. From these Var^{ns}, on average half were also detectable on transcript level (44±10%; Fig. 1a). Across all patients, we observed 1039 unique Var^{ns} in total, affecting 864 different genes and 45% of them (n=392) with additional evidence on RNA level (Var^{exp}). This translates to an average tumour mutational burden (TMB) (estimated as previously described[330]) of 1.89±0.49 per megabase observed in our HCC cohort.
Assessing mutational hotspots, we observed alterations (Var^{exp}) in β-*catenin* (*CTNNB1*; 50%) and in *neuroblastoma breakpoint family, member 1* (*NBPF1*; 38%), but also in genes encoding proteins typically expressed in the liver, such as *albumin* (*ALB*; 19%), *apolipoprotein b* (*APOB*; 13%), and γ-*glutamyltransferase* (*GGT1*; 19%). Var^{exp} frequently affected the HLA class II loci *HLA-DRB1* (6%), *HLA-DQA1* (13%), and *HLA-DRB5* (19%). However, due to the highly polymorphic nature of the HLA locus[331], variant detection in these regions is particularly error-prone and results should be cautiously interpreted as potential artifacts. For HLA class I loci, suitable computational pipelines for mutation detection are available[65], whereas for HLA class II to the best of our knowledge this is not the case. Overall, only 1.5% (6/392) of Var^{exp}-containing genes were shared among >2 patients and only one single mutation (in *NBPF1*; Chr. 1:16891365 G>T) reoccurred identically in three patients. Figure 7.1 shows the variation in numbers of Var, Var^{exp}, PubMed identifier (PMID), and PNE^{exp} in our HCC cohort. Considering established driver mutations included in the Cancer Gene Census[vii][332], we observed respective Var^{ns} in most of the HCCs (n=9; 1-3 Var^{ns} per patient), foremost the previously mentioned gene *CTNNB1* (n=8) but also the androgen receptor, *mediator complex subunit 12* (*MED12*), *nuclear receptor corepressor 1* (*NCOR1*), *neurogenic locus notch homolog protein 1* (*NOTCH1*) (all n=2), and *NOTCH2*/*PIK3CA* (n=1). Nevertheless, except from *CTNNB1*, Var^{ns} comprised in the Cancer Gene Census appeared rather infrequently among the examined HCCs.

[vii] `https://cancer.sanger.ac.uk/census`

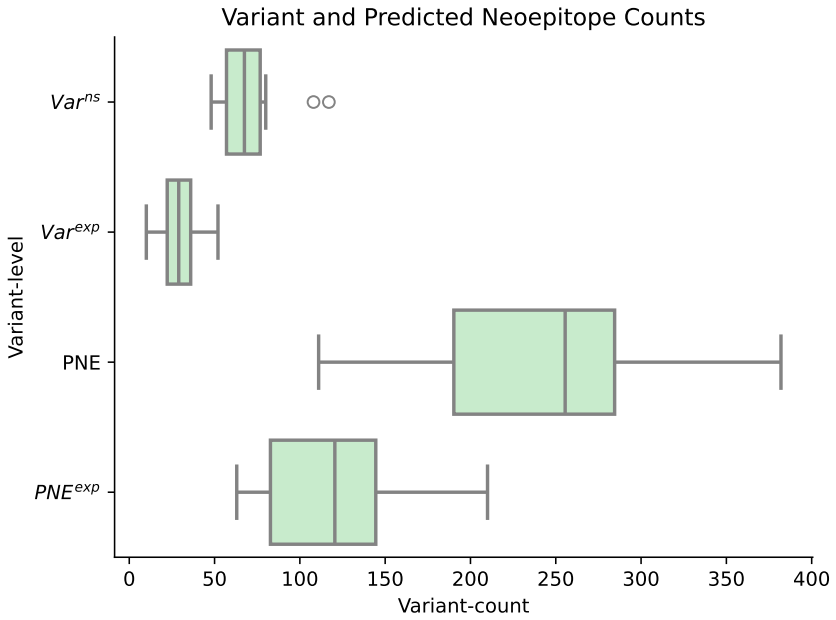


Figure 7.1: Variant-count variations between samples of the HCC cohort on different levels of detection and prediction.

7.4.2 Discovery of Mutation-Derived HLA Ligands on Different Omics Levels

Exome

As a first step, we sought to assess the number of mutation-derived PNE per patient. On average, 244±77 PNE per HCC patient were predicted to exceed the respective HLA class I alleles binding thresholds from 66±18 Var^{ns} (Fig. 7.2). The observed increase in PNE numbers compared to Var^{ns} is explained by the fact that Var^{ns} may give rise to multiple PNE due to the shifting window approach used with different peptide lengths (8-11 amino acids) as well as the HLA-binding prediction for up to six individual HLA alleles. Comparing the numbers of PNE to the numbers of protein-altering variants (Var^{ns}) resulted in a very weak correlation (Pearson's correlation coefficient r=0.38).

Transcriptome

When accounting for supplemental evidence for PNE on RNA level, numbers of predicted peptides (PNE^{exp}) decreased by half (49±8% of PNE), yielding an average of 118±40 PNE^{exp} per patient. The correlation between expressed protein-changing genomic variants (Var^{exp}) and PNE^{exp} was moderate (Pearson's correlation coefficient r=0.50).

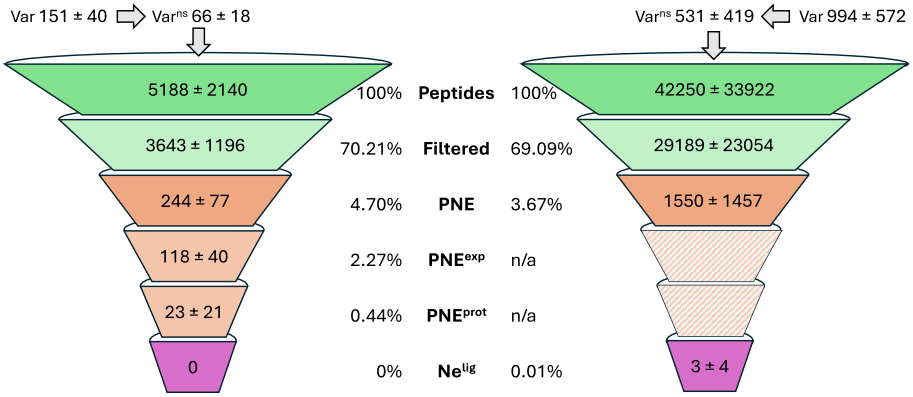


Figure 7.2: The number of potential neoepitopes within the HCC cohort (left) diminishes along the successive steps of analysis in the workflow. The analysis of the Mel dataset (right) start out with more variants in general, no expression or shotgun-proteomics is available, however neoepitopes could be detected.

Evidence for Mutated Proteins on Shotgun Proteome Level

To obtain the best available evidence for the presence of mutated proteins we first used shotgut MS to identify the proteins in HCC tissue samples (n=7). Evaluation of the identified proteins revealed support for the source proteins of a total of 159 PNE (17±14% of PNE^{exp}, on average for 23±21 PNE per patient). Only in one patient(HCC034), no protein source evidence for any PNE was found.
We found direct evidence for mutated proteins with identified peptides carrying single amino acid variations (SAV) corresponding to mutation sources for the PNE. We discovered one somatic mutation in *albumin* (*ALBK375E*) on proteome level represented by the tryptic peptide LAETYETTLEK in HCC025 (Fig. 7.3a), which was corroborated on both exome (Var^{ns}) and transcriptome (Var^{exp}) levels. Strikingly, we not only detected the tryptic wild-type peptide LAKTYETTLEK in the proteome of non-malignant liver tissue, but unexpectedly also the mutation-derived peptide LAETYETTLEK. This was corroborated by two additional replicate measurements from serum samples taken at different time points from the patient. Patient HCC025 showed tumour recurrence and active disease at both time points and the mutated peptide was detected in both samples, proving that the tumour synthesised a mutated *ALB* protein secreted into circulation. For HCC026, a Var^{exp} in the *ATP-dependent DNA helicase Q1* (*RECQL*; *H19R*) could be verified based on an additional tryptic cleavage site introduced through the arginine gained by mutation, which resulted in the proteotypic peptide AVEIQIQELTER. This peptide was not detected in the corresponding non-malignant liver tissue (Fig. 7.3b).

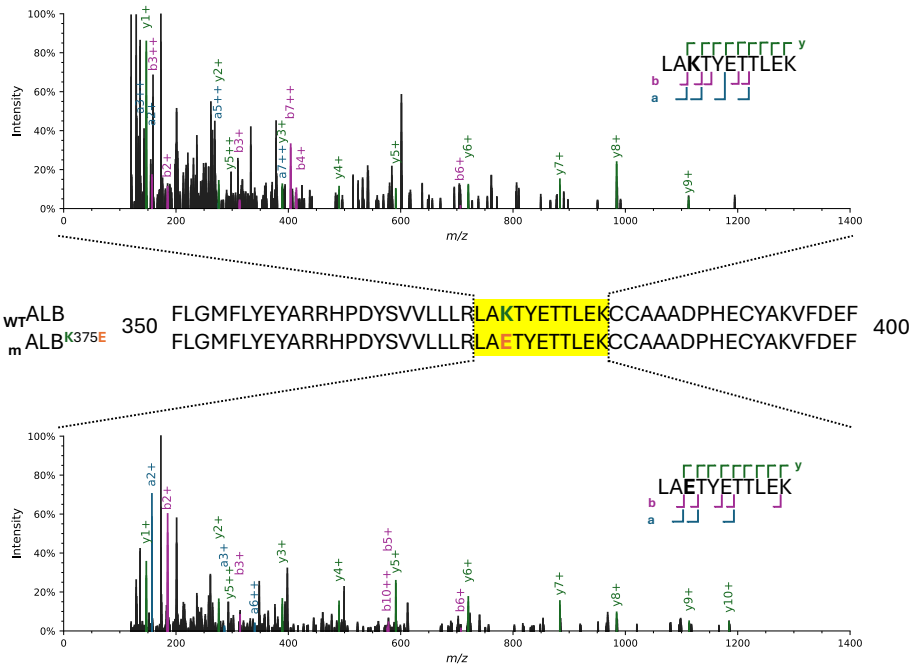


Figure 7.3: Evidence for mutated proteins in the shotgun proteome. Annotated spectra of *albumin* (*ALB*) showing sequences of wild-type (LAKTYETTLEK; top) and mutated (LAETYETTLEK; bottom) protein measured by LC-MS/MS.

HLA ligandome

To directly assess the presence of mutated HLA ligands we used LC-MS/MS measurements of isolates of naturally processed HLA peptides from HCCs and non-malignant liver tissues. These analyses yielded on average 1403±621 HLA class I-associated peptides from HCC and 1159±525 peptides from non-malignant liver tissue (FDR 5%, length 8-11 amino acids). On average, 51±11% of these peptides were shared between matching malignant and non-malignant liver tissue. When applying HLA-binding prediction on detected peptides, on average 1026±451 peptides per tumour (73±10%) and 867±450 peptides per non-malignant liver sample (72%±11%) exceeded the respective patients' HLA class I allotype binding thresholds. Of those, on average 58±12% peptides occurred in both matched malignant and non-malignant liver tissues. Importantly, we did not find any evidence for naturally presented mutated HLA ligands (NE^{lig}) in HCCs, independent of filtering criteria. However, in two HCC patients, we identified the wild-type sequence HLA ligand (WT^{lig}) to a corresponding PNE each.

Targeted mass spectrometry for discovery of mutated HLA ligands

As no NE^{lig} could be confirmed in HLA ligandomics data acquired in DDA mode, we used a SIM approach to avoid the sensitivity limitations inherent to DDA[333]. To corroborate the PNE^{prot} observed in *ALB* and *RECQL* by targeted MS approaches as well as other carefully selected PNE^{exp} in three chosen patients, we selected sets of PNE from three HCCs (HCC025-27) for SIM. Heavy isotope-labeled peptides were used as a reference and to improve the probability of detection. Nevertheless, we could not validate any of the candidates and comparisons of low confidence annotations with synthetic peptides did not yield evidence for peptide presentation.
Since peptides harbouring the mutations confirmed on proteome level (PNE^{prot}) seemed of particular interest (i.e., *ALB*K375E in HCC025 and *RECQL*H19R in HCC026), we additionally performed parallel reaction monitoring (PRM) measurements targeting the best ranking PNE as well as corresponding wild-type HLA ligands (WT^{lig}) covering the mutation site. Despite a high number of HLA class I peptides identified in DDA mode (HCC025 malignant: 5063; HCC025 non-malignant: 1497; HCC026 malignant: 3678; HCC026 non-malignant: 3197), PRM could not corroborate any of the PNE^{prot} as naturally presented HLA ligands in HCC.

Workflow Validation in Absence of Mutated HLA Ligands in HCC

To demonstrate the neoepitope identification workflow to its full extent, we additionally processed a publicly available dataset of somatic variants from five Mel patients as a reference[303]. A generally higher mutational burden of Mel in comparison to HCC was found from a survey of the cancer genome atlas (TCGA)[viii]: a mean number(±SD) of Var^{ns} of 90±100 for HCC ($n = 363$) and 461±761 for Mel ($n = 467$). This is reflected by the Var^{ns} from the Mel dataset and resulting PNE from our analysis: 531 Var^{ns} on average in the Mel dataset versus 66 in our HCC cohort; 1550 PNE on average in the Mel dataset versus 243 PNE in HCC (Fig. 7.2). The workflow analysis reconfirmed all of the NE^{lig} previously reported by Bassani-Sternberg et al.[303] and four additional NE^{lig} in two Mel samples that could be validated by matching spectra from synthetic peptides.

[viii] retrieved from Genomics Data Commons Data Portal `https://portal.gdc.cancer.gov/`, accessed 16.09.2018

7.5 Discussion

Neoepitopes, i.e., unique peptides derived from tumour-specific mutations presented as natural HLA ligands and recognised by T cells, have been suggested as highly attractive targets for cancer immunotherapy. There is mounting (albeit indirect) evidence to suggest that increased numbers of mutations may render malignancies more immunogenic through their neoantigenic repertoire (i.e., mutated HLA ligands) and therefore more amenable to immunotherapies[288]. Particularly for tumours that are characterised by a high TMB, a correlation with benefits of ICP inhibition has been shown[291–293,334].
One of the greatest challenges in understanding and ultimately harnessing this neoantigenic repertoire of cancers is the selection and validation of suitable targets from an array of PNE derived from computational algorithms[335]. From Fig. 7.2, it is plausible to assume that most PNE are irrelevant and would ultimately fail to make an impact on treatment outcomes of individual patients. On the other hand, the selection of a single suitable neoepitope may have unprecedented therapeutic consequences[296,336] and such a single neoepitope has already been shown to be a target of T cells induced by ICP inhibition[301]. Certainly, this notion is not limited to neoepitopes, but it also applies to tumour-associated antigens, which can possess a comparable immunogenicity[337]. Consequently, non-mutated tumour-specific or highly tumour-associated antigens should be considered prime choice for personalised immunotherapy, if they can be individually validated[338]. A knowledge-based approach for PNE prioritisation using previously measured wild-type HLA ligands (WT^{lig}) has been proposed[339]. As shown, when filtering detected HLA peptides for binding prediction, around 27-28% fell below the respective patients' HLA class I allotype binding thresholds. This can be either interpreted as a step to enrich for high probability HLA class I ligands, or as in silico accuracies that need to be considered in a prioritisation scenario.
Although many assumptions regarding mutated neoepitopes are theoretically and bio-mechanistically plausible[294], there remains a fundamental lack of knowledge concerning the precise immunological underpinnings behind tumour specificity[340] and therapeutic implications.
Moreover, biomarkers predicting response to ICP inhibitors with higher precision than TMB[334] are sought-after[341]. A respective biomarker might not only assess the odds for ICP therapy success but may simultaneously allow the development of tailored neoantigen-targeted immunotherapies.
In contrast to the vast array of data available relating to PNE[342,343], often derived from data of consortia like the international cancer genome consortium (ICGC) or TCGA, current physical evidence for exome-derived mutated HLA ligands (NE^{lig}) seems

anecdotal (reviewed by Freudenmann et al.[344]) and positive examples for finding this proverbial needle in the haystack are scarce. Lacking detection of mutated HLA ligands (NE^{lig}) does not equal their absence due to several reasons: inter alia (1) detection limits of the LC-MS/MS instrumentation, (2) lacking ionisability of respective peptides, (3) particularly strongly hydrophilic and hydrophobic peptides may be missed by the UHPLC method, (4) unknown temporal dynamics of the HLA ligandome[345]. Hence, to be able to benchmark our results obtained in HCC, we used the best evidence available to us. Our HCC dataset is characterised by close to 70 amino acid-changing mutations (Var^{ns}) on average translating to a TMB of about two per megabase, numbers corresponding very well with data from a comprehensive set of resectable HCCs[346]. These mutations encompass established hotspots, and a limited number of genes were found to be recurrently mutated[346], affecting the well-established CTNNB1 primarily but also NBPF1. The latter remained the only gene with a repeat identical mutation in our patient cohort, emphasising that in combination with an individual set of HLA class I allotypes, a neoepitope-targeted therapy needs to be strictly personalised[340]. Since in HCCs only about half of the initially 244 Var^{ns} could be corroborated by RNA level evidence (Var^{exp}), this bisected the computationally predicted neoepitope numbers to an average of 118 expressed PNE (PNE^{exp}). Further, the correlation of both PNE and PNE^{exp} numbers with mutation counts, showed only a weak correlation. This may imply that there is no direct interconnection between mutation frequency and respective HLA ligands but rather a probabilistic model applies[301], which is governed by the HLA ligandome with distinct rules of presentation[347]. Since we had shotgun proteomics data to the subset of our cohort available, we also assessed whether we could establish any additional physical evidence for the respective source proteins (PNE^{prot}) constituting the immediate proteomic context of NE^{lig}, which was the case in about one-fifth of PNE^{exp} and comprised about 10% of the initial PNE pool. Even less direct evidence of mutated proteins was found, with two out of seven patients where we could confirm a mutation in the proteome, once directly and in the other case through the introduction of an additional tryptic cleavage site by mutation. This is however in line with previously reported fractions of genomic alterations detectable on protein level by LC-MS/MS ($\sim$2%[348]). Nevertheless, since this neither implies the actual absence of synthesis of mutated proteins, nor the HLA presentation of a NE^{lig}, we assessed the eluted HLA ligands and searched for any PNE with actual evidence for HLA presentation by LC-MS/MS. Although the $\sim$1400 HLA-bound peptides detected on average in HCCs are generally comparable with the numbers previously published in solid cancers[323,349], they do fall short of the considerable depth reached in Mel, particularly in one single exceptional case, for which more than 20,000 HLA-bound

peptides were reported (Mel15[303]). We demonstrated with the application of our analysis workflow to the Mel data available, that results can be reproduced and improved. Improving identification sensitivity is essential for future approaches of the kind for Future. We used percolator to promote true target PSMs, and new developments show promise for further improvement. MSBooster[350] and Prosit[351] have provided approaches combining percolator with neural networks to increase sensitivity for this purpose.

However, in the direct cancer type comparison, it becomes particularly clear that Mel and HCC, despite both representing solid tumours, feature fundamental differences on a variety of biological levels. This notion is confirmed by an extensive analysis of 30 cancer types using comprehensive sequencing data from ICGC and TCGA[305], with striking differences concerning the PNE pool between HCCs and Mel or lung and colorectal cancer[343]. Those differences may imply disparities in antigenicity, determining the odds for immunotherapy success[294]. Indeed, we only found a single case with comparable Var^{ns} counts among Mel[303] similar to our relatively homogeneous HCC cohort, where a NE^{lig} could be verified.

In three other studies, all conducted on either Mel samples, Mel cell lines, or oesophageal adenocarcinoma (OAC), also a cancer with a high mutational burden, only very few neoantigens were detected (Mel: 2[352], cell lines: 3[304] & 2[353], OAC: 1[354]). The studies used bespoke workflows to identify neoantigen targets, that are hardly automatable. Noteworthy, Nicholas et al.[354] used pVAC-seq[355,356], a containerised collection of dedicated tools for identifying tumour neoantigens. However, all approaches lack the workflow orchestration integration our OpenMS and FRED2 combination affords.

Hence, chances for the presentation of exome-derived NE^{lig} in HCC may be commonly very low, possibly due to cancer immunoediting[357]. Furthermore, the odds for administering targeted therapies available to HCC patients in our cohort remained small as previously encountered[307]. For cancers with such limited target space, the scope may need to be widened to selections of tumour-associated antigens for immunotherapy to succeed, which also necessitates a better understanding of antigen presentation in such cases. There have been efforts towards the characterisation of the HLA-class I peptidome with the HLA-atlas[358] and the Human Immunopeptidome Project[359]. For confident TAA selection, immunopeptidome-wide association studies have to be applied on a large scale, for which OpenMS and FRED2 can provide the necessary automation.

Chapter 8

Conclusion

For personalised immunotherapy approaches or research with emergent experimental methodologies, a well-integrated analysis tool ecosystem which is also easy to integrate new tools into is key for flexible workflow design and adaptation to current needs without reimplementation of the foundational aspects. We showed how this can be achieved with the design and implementation of standards. The mzIdentML implementation in OpenMS opened up easy search engine and post-processing tool integration, broadening the ecosystem of analysis tools available for workflow design and subsequent high-throughput analysis. The design and implementation of the crosslinking use case illustrated this, as it acted as a keystone, bridging the development of a novel search engine for cross-linked peptides to existing visualisation solutions to form a better understanding of results. It also allows OpenPepXL to export the results into a standardised format accepted by major data repositories like PRIDE, usually a requirement for research publication in many established journals. As it is, there is a case to be made for the compounding effects of an early adoption of standards for emerging fields. The mzIdentML developments for crosslinking proved its utility to the field and resulted in increased interest in standard adoption, and further refinement in an upcoming version of the mzIdentML standard for a more detailed representation of crosslinking.

The development of mzQuantML has proven instrumental to consensus formation in quantitative proteomics research and a push towards common (reporting) goals. Its design greatly influenced the development of mzTab as a more accessible method to report commonly expected elements of analysis results. So, too, had the early exploration of embracing cross-field commonalities with other MS-based life sciences through the inclusion of small molecules, first mzQuantML, then in mzTab. This paid dividends in the development of a compatible mzTab-m standard for metabolomics, expansion of available tools, and opportunity for tighter collaboration between the fields. Reluctant adoption of mzQuantML revealed the need for flexibility in analysis workflow design as the highly dynamic field of quantitative methods is rich in diverse methodology and favours insular solutions for reporting. mzTab has found the lowest common

denominator with a spreadsheet-compatible method of reporting, its use however can still be complex due to the rich diversity in result elements important for particular methods. Future format developments might take advantage of recent developments to capture the relation between (proteomics) samples and the data generated in a standardised fashion[360] to reduce schema complexity by referencing the experimental design described with SDRF.

The inherent flexibility of qcML has proven optimal for the use cases of QC, which span acquisition and analysis. Potentially each step of a workflow is thus a QC monitoring opportunity, for which qcML's hierarchical structure with CV control provides simple add-on capabilities. In later and ongoing developments, the lessons from mzQuantML led to the development of a less demanding format basis using JSON instead of XML in the qcML successor format mzQC. Though virtually the same format structure was preserved, the new PSI standard for QC improved accessibility. The qcML format showcased the benefits of tight integration of QC processes along any MS-based data analysis workflow and its dual utility as hand-over and report format, continued by mzQC.

We illustrated in Chapter 6 the synergistic effects of standard file format implementation and workflow management integration. With standard formats for data hand-over, workflows can be designed without custom glue code, implicitly increasing the available toolset. This renders even complex workflows for advanced studies more transparent and transferable, therefore improving reproducibility. As exemplified in Chapters 6 and 7, workflows become complex when designing for comprehensive life sciences, demonstrated with the combination of a computational MS workflow and an immunopeptide workflow with applied QC. This also highlighted another benefit of workflow management integration: enabling high-throughput analysis for a laboratory's total data output of immunopeptide MS experiments. Particularly useful is automation for the longitudinal-QC use case, as the day-to-day ready use captures a consistent stream of instrument health data. We have shown the utility of longitudinal quasi-QC samples (JY), but we imagine a dedicated QC sample for immunopeptidomics run regularly to perform much better and for example, enable predictive maintenance. This, combined with the demonstrated QC metrics, can help boost instrument sensitivity, which has a dominating impact on immunopeptide MS measurements. The combination of workflows also underlines another set of synergistic effects through the development and integration of larger frameworks into workflow management systems as shown with FRED2. The development of inclusive frameworks has benefits of its own as discussed in Chapters 5 and 6. In short, they lower maintenance and make for easier integration of existing tools with inter-compatible base structures and a reduced development

burden for new tools and methods. FRED2, as the version in the name implies, reflects the consistent development of a software framework for computational immunology. This consistency has great value to the community and is seen to an even greater extent with OpenMS, with a track record spanning almost two decades and support from multiple institutional partners.

In future life-science studies, combining different disciplines will become important to consolidate new findings and develop novel approaches for disease diagnosis and therapy. The integration of OpenMS and the ImmunoNodes with KNIME offers great opportunities for combined experiment and analysis approaches through the community extensions. Finally, we concluded the integration of data and workflows with a demonstration of the application of omics-spanning data analysis in a pre-clinical study of personalised immunotherapy development for HCC. Unfortunately, we were not able to detect any neoepitopes on immunopeptide level, however, demonstrated the workflow functionality by reanalysis of a previously published melanoma dataset and detected the reported and additional neoepitopes. We believe the described approach holds promise, though apparently not all cancer types are equally suited. We showed that a higher mutational burden seems to coincide with successful neoepitope presentation and is therefore important for taking advantage of such personalised immunotherapy approaches. The ongoing research with similar approaches to other cancer types has provided further insight into potential applicability of the approach. Sustained and comprehensive studies including immunoinformatics and computational MS will be necessary for the discovery of novel methods to treat cancer and can now take advantage of existing solutions integrated into a wider ecosystem for analysis design.

We feel therefore confident that our workflows, or updated versions thereof, will have a place in future research in that direction. As the integration of immunopeptidomics and MS workflows is only one demonstration of the synergistic effects of data and tool integration for a wider community of life-science analysis development, we are equally confident, that other combinations will be able to take advantage of format standardisation and tool integration efforts. In general, we hope to have successfully demonstrated the benefits of data and analysis tool integration into software solutions for automated high-throughput analysis with a lasting impact on the quality and reusability of published research.

Bibliography

[1] Bruce Alberts and etc., editors. *Molecular biology of the cell*. CRC Press, Boca Raton, FL, 4 edition, March 2002. 5, 6

[2] Ron Milo, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer. BioNumbers–the database of key numbers in molecular and cell biology. *Nucleic Acids Res.*, 38(Database issue): D750–3, January 2010. 5

[3] Zimian Wang, Wei Shen, Donald P Kotler, Stanley Heshka, Lucian Wielopolski, John F Aloia, Miriam E Nelson, Richard N Pierson, Jr, and Steven B Heymsfield. Total body protein: a new cellular level mass and distribution prediction model. *Am. J. Clin. Nutr.*, 78(5):979–984, November 2003. 5

[4] Shintaro Iwasaki and Nicholas T Ingolia. PROTEIN TRANSLATION. seeing translation. *Science*, 352(6292):1391–1392, June 2016. 7

[5] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang,

M Coyne, C Dahlke, A Deslattes Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, February 2001. 7

[6] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[7] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, October 2004.

[8] Valerie A Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A Kitts, Terence D Murphy, Kim D Pruitt, Françoise Thibaud-Nissen, Derek Albracht, Robert S Fulton, Milinn Kremitzki, Vincent Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, William Chow, Joanna Collins, Glenn Harden, Timothy Hubbard, Sarah Pelan, Jared T Simpson, Glen Threadgold, James Torrance, Jonathan M Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Paul Peluso, Heng Li, Chen-Shan Chin, Adam M Phillippy, Richard Durbin, Richard K Wilson, Paul Flicek, Evan E Eichler, and Deanna M Church. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, 27(5):849–864, May 2017. 7

[9] F Sanger and A R Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94(3):441–448, May 1975. 7

[10] Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Mol. Cell*, 58(4):586–597, May 2015. 8

[11] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17(6):333–351, June 2016.

[12] Isidro Cortés-Ciriano, Doga C Gulhan, Jake June-Koo Lee, Giorgio E M Melloni, and Peter J Park. Computational analysis of cancer genome sequencing data. *Nat. Rev. Genet.*, 23(5):298–314, May 2022. 8

[13] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. 8

[14] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, March 2010. 8

[15] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, March 2009. 8

[16] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359, March 2012. 8

[17] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009. 8

[18] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4):R36, April 2013. 8

[19] Pascale Gerbault, Anke Liebert, Yuval Itan, Adam Powell, Mathias Currat, Joachim Burger, Dallas M Swallow, and Mark G Thomas. Evolution of lactase persistence: an example of human niche construction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 366(1566):863–877, March 2011. 8

[20] V M Ingram. Gene mutations in human hæmoglobin: The chemical difference between normal and sickle cell hæmoglobin. *Nature*, 180(4581):326–328, August 1957. 8

[21] Ilka Hoof, Debbie van Baarle, William H Hildebrand, and Can Keşmir. Proteome sampling by the HLA class I antigen processing pathway. *PLoS Comput. Biol.*, 8(5):e1002517, May 2012. 9

[22] Rebecca Mercier and Paul LaPointe. The role of cellular proteostasis in antitumor immunity. *J. Biol. Chem.*, 298(5):101930, May 2022.

[23] Priyanka S Rana, James J Ignatz-Hoover, and James J Driscoll. Targeting proteasomes and the MHC class I antigen presentation machinery to treat cancer, infections and age-related diseases. *Cancers (Basel)*, 15(23), November 2023. 9

[24] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22 (3):568–576, March 2012. 9

[25] Christopher T Saunders, Wendy S W Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, July 2012. 9

[26] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, 42(Database issue):D980–5, January 2014. 9

[27] S A Forbes, N Bindal, S Bamford, C Cole, C Y Kok, D Beare, M Jia, R Shepherd, K Leung, A Menzies, J W Teague, P J Campbell, M R Stratton, and P A Futreal. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, 39(Database): D945–D950, January 2011. 9

[28] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, August 2011. 9

[29] Lloyd M Smith, Neil L Kelleher, and Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat. Methods*, 10(3):186–187, March 2013. 9

[30] Zhongyan Li, Shangfu Li, Mengqi Luo, Jhih-Hua Jhong, Wenshuo Li, Lantian Yao, Yuxuan Pang, Zhuo Wang, Rulan Wang, Renfei Ma, Jinhan Yu, Yuqi Huang, Xiaoning Zhu, Qifan Cheng, Hexiang Feng, Jiahong Zhang, Chunxuan Wang, Justin Bo-Kai Hsu, Wen-Chi Chang, Feng-Xiang Wei, Hsien-Da Huang, and Tzong-Yi Lee. dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. *Nucleic Acids Res.*, 50(D1):D471–D479, January 2022. 10

[31] Katie Dunphy, Paul Dowling, Despina Bazou, and Peter O'Gorman. Current methods of post-translational modification analysis and their applications in blood cancers. *Cancers (Basel)*, 13 (8):1930, April 2021. 10

[32] A R Smith and B M Baker. SILv44 T cell receptor bound to HLA-A2 presenting gp100T2M peptide (IMDQVPFSV), November 2020. 10

[33] David Sehnal, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K Burley, Jaroslav Koča, and Alexander S Rose. Mol* viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.*, 49(W1):W431–W437, July 2021. 10

[34] Janko Nikolich-Zugich, Mark K Slifka, and Ilhem Messaoudi. The many important facets of t-cell repertoire diversity. *Nat. Rev. Immunol.*, 4(2):123–132, February 2004. 11

[35] Keith D Wilkinson. The discovery of ubiquitin-dependent proteolysis. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15280–15282, October 2005. 11

[36] Galen Andrew Collins and Alfred L Goldberg. The logic of the 26S proteasome. *Cell*, 169(5): 792–806, May 2017. 11

[37] Yosup Kim, Eun-Kyung Kim, Yoona Chey, Min-Jeong Song, and Ho Hee Jang. Targeted protein degradation: Principles and applications of the proteasome. *Cells*, 12(14), July 2023. 11

[38] Kirby Martinez-Fonts, Caroline Davis, Takuya Tomita, Suzanne Elsasser, Andrew R Nager, Yuan Shi, Daniel Finley, and Andreas Matouschek. The proteasome 19S cap and its ubiquitin receptors provide a versatile recognition platform for substrates. *Nat. Commun.*, 11(1):477, January 2020. 11

[39] Kenneth Murphy and Casey Weaver. *Janeway's Immunobiology*. Garland Science, London, England, 9 edition, March 2016. 11, 12

[40] K Maude Ashby and Kristin A Hogquist. A guide to thymic selection of T cells. *Nat. Rev. Immunol.*, 24(2):103–117, February 2024. 12

[41] D R Madden. The three-dimensional structure of peptide-MHC complexes. *Annu. Rev. Immunol.*, 13(1):587–622, 1995. 12

[42] P J Bjorkman, M A Saper, B Samraoui, W S Bennett, J L Strominger, and D C Wiley. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature*, 329(6139):506–512, 1987. 12

[43] A R Khan, B M Baker, P Ghosh, W E Biddison, and D C Wiley. CRYSTAL STRUCTURE OF HLA-A*0201/OCTAMERIC TAX PEPTIDE COMPLEX, February 2000. 13

[44] J Liu, Y Chen, L Lai, and E Ren. Crystal structure of HLA a*02:03 bound to HBV core 18-27, May 2011. 13

[45] R Stephens, R Horton, S Humphray, L Rowen, J Trowsdale, and S Beck. Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. *J. Mol. Biol.*, 291(4):789–799, August 1999. 13

[46] Nomenclature for factors of the HL-A system. *Bull. World Health Organ.*, 47(5):659–662, 1972. 13

[47] S G E Marsh, E D Albert, W F Bodmer, R E Bontrop, B Dupont, H A Erlich, M Fernández-Viña, D E Geraghty, R Holdsworth, C K Hurley, M Lau, K W Lee, B Mach, M Maiers, W R Mayr, C R Müller, P Parham, E W Petersdorf, T Sasazuki, J L Strominger, A Svejgaard, P I Terasaki, J M Tiercy, and J Trowsdale. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*, 75(4):291–455, April 2010. 13

[48] Dominic J Barker, Giuseppe Maccari, Xenia Georgiou, Michael A Cooper, Paul Flicek, James Robinson, and Steven G E Marsh. The IPD-IMGT/HLA database. *Nucleic Acids Res.*, 51(D1): D1053–D1060, January 2023. 13

[49] Faviel F Gonzalez-Galarza, Antony McCabe, Eduardo J Melo Dos Santos, James Jones, Louise Takeshita, Nestor D Ortega-Rivera, Glenda M Del Cid-Pavon, Kerry Ramsbottom, Gurpreet Ghattaoraya, Ana Alfirevic, Derek Middleton, and Andrew R Jones. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.*, 48(D1):D783–D788, January 2020. 14, 189

[50] Esteban Arrieta-Bolaños, Diana Iraíz Hernández-Zaragoza, and Rodrigo Barquera. An HLA map of the world: A comparison of HLA frequencies in 200 worldwide populations reveals diverse patterns for class I and class II. *Front. Genet.*, 14:866407, March 2023.

[51] S J Mack, B Tu, A Lazaro, R Yang, A K Lancaster, K Cao, J Ng, and C K Hurley. HLA-A, -b, -c, and -DRB1 allele and haplotype frequencies distinguish eastern european americans from the general european american population. *Tissue Antigens*, 73(1):17–32, January 2009. 14, 189

[52] Y Kawakami, P F Robbins, X Wang, J P Tupesis, M R Parkhurst, X Kang, K Sakaguchi, E Appella, and S A Rosenberg. Identification of new melanoma epitopes on melanosomal proteins recognized by tumor infiltrating T lymphocytes restricted by HLA-A1, -a2, and -A3 alleles. *J. Immunol.*, 161 (12):6985–6992, December 1998. 14

[53] M DiBrino, T Tsuchida, R V Turner, K C Parker, J E Coligan, and W E Biddison. HLA-A1 and HLA-A3 T cell epitopes derived from influenza virus proteins predicted from peptide binding motifs. *J. Immunol.*, 151(11):5930–5935, December 1993.

[54] John Sidney, Bjoern Peters, Nicole Frahm, Christian Brander, and Alessandro Sette. HLA class I supertypes: a revised and updated classification. *BMC Immunol.*, 9(1):1, January 2008. 14

[55] D R Madden, D N Garboczi, and D C Wiley. The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2. *Cell*, 75(4):693–708, November 1993. 14

[56] Anne Marie Asemissen, Ulrich Keilholz, Stefan Tenzer, Margret Müller, Steffen Walter, Stefan Stevanovic, Hansjörg Schild, Anne Letsch, Eckhard Thiel, Hans-Georg Rammensee, and Carmen Scheibenbogen. Identification of a highly immunogenic HLA-A*01-binding T cell epitope of WT1. *Clin. Cancer Res.*, 12(24):7476–7482, December 2006.

[57] Nicolas Parmentier, Vincent Stroobant, Didier Colau, Philippe de Diesbach, Sandra Morel, Jacques Chapiro, Peter van Endert, and Benoît J Van den Eynde. Production of an antigenic peptide by insulin-degrading enzyme. *Nat. Immunol.*, 11(5):449–454, May 2010. 14

[58] G D Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, January 2000. 14

[59] IUPAC-IUB Comm. on Biochem. Nomenclature. A one-letter notation for amino acid sequences. tentative rules. *Biochemistry*, 7(8):2703–2705, August 1968. 14

[60] Daniel M Tadros, Simon Eggenschwiler, Julien Racle, and David Gfeller. The MHC motif atlas: a database of MHC binding specificities and ligands. *Nucleic Acids Res.*, 51(D1):D428–D437, January 2023. 14

[61] P I Terasaki and J D Mcclelland. Microdroplet assay of human serum cytotoxins. *Nature*, 204 (4962):998–1000, December 1964. 15

[62] András Szolek, Benjamin Schubert, Christopher Mohr, Marc Sturm, Magdalena Feldhahn, and Oliver Kohlbacher. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*, 30(23):3310–3316, December 2014. 15, 89, 91, 93, 96

[63] Chang Liu, Xiao Yang, Brian Duffy, Thalachallour Mohanakumar, Robi D Mitra, Michael C Zody, and John D Pfeifer. ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res.*, 41(14):e142, August 2013. 89, 91, 93

[64] Sebastian Boegel, Martin Löwer, Michael Schäfer, Thomas Bukur, Jos de Graaf, Valesca Boisguérin, Ozlem Türeci, Mustafa Diken, John C Castle, and Ugur Sahin. HLA typing from RNA-Seq sequence reads. *Genome Med.*, 4(12):102, December 2012. 89, 91, 93, 96

[65] Sachet A Shukla, Michael S Rooney, Mohini Rajasagi, Grace Tiao, Philip M Dixon, Michael S Lawrence, Jonathan Stevens, William J Lane, Jamie L Dellagatta, Scott Steelman, Carrie Sougnez, Kristian Cibulskis, Adam Kiezun, Nir Hacohen, Vladimir Brusic, Catherine J Wu, and Gad Getz. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.*, 33(11):1152–1158, November 2015. 15, 127

[66] Carolyn K Hurley, Jane Kempenich, Kim Wadsworth, Jürgen Sauter, Jan A Hofmann, Daniel Schefzyk, Alexander H Schmidt, Pablo Galarza, Maria B R Cardozo, Malgorzata Dudkiewicz, Lucie Houdova, Pavel Jindra, Betina S Sorensen, Latha Jagannathan, Ankit Mathur, Tiina Linjama, Tigran Torosian, Rafi Freudenberger, Anastasios Manolis, John Mavrommatis, Nezih Cereb, Sigal Manor, Nira Shriki, Nicoletta Sacchi, Reem Ameen, Raewyn Fisher, Heather Dunckley, Irene Andersen, Ahmed Alaskar, Mohsen Alzahrani, Ali Hajeer, Dunia Jawdat, Grazia Nicoloso, Pawinee Kupatawintu, Louise Cho, Ashminder Kaur, Mats Bengtsson, and Jason Dehn. Common,

intermediate and well-documented HLA alleles in world populations: CIWD version 3.0.0. *HLA*, 95(6):516–531, June 2020. 15

[67] Paul Ehrlich. Über den jetzigen stand der karzinomforschung. *Nederlands Tijdschrift voor Geneeskunde*, 53:273–290, 1909. 15

[68] Cellular and humoral aspects of the hypersensitive states: A symposium at the new york academy of medicine. *J. Am. Med. Assoc.*, 170(7):883, June 1959. 15

[69] R T Prehn and J M Main. Immunity to methylcholanthrene-induced sarcomas. *J. Natl. Cancer Inst.*, 18(6):769–778, June 1957. 15

[70] Gavin P Dunn, Lloyd J Old, and Robert D Schreiber. The immunobiology of cancer immunosurveillance and immunoediting. *Immunity*, 21(2):137–148, August 2004. 15

[71] D Hanahan and R A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, January 2000. 15

[72] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144 (5):646–674, March 2011. 15, 16

[73] Karin E de Visser, Alexandra Eichten, and Lisa M Coussens. Paradoxical roles of the immune system during cancer development. *Nat. Rev. Cancer*, 6(1):24–37, January 2006. 15

[74] Joan Massagué. TGFbeta in cancer. *Cell*, 134(2):215–230, July 2008. 15

[75] Sergei I Grivennikov, Florian R Greten, and Michael Karin. Immunity, inflammation, and cancer. *Cell*, 140(6):883–899, March 2010. 16

[76] Eileen White, Cristina Karp, Anne M Strohecker, Yanxiang Guo, and Robin Mathew. Role of autophagy in suppression of inflammation and cancer. *Curr. Opin. Cell Biol.*, 22(2):212–217, April 2010. 16

[77] David G DeNardo, Pauline Andreu, and Lisa M Coussens. Interactions between lymphocytes and myeloid cells regulate pro- versus anti-tumor immunity. *Cancer Metastasis Rev.*, 29(2):309–316, June 2010. 16

[78] Mikala Egeblad, Elizabeth S Nakasone, and Zena Werb. Tumors as organs: complex tissues that interface with the entire organism. *Dev. Cell*, 18(6):884–901, June 2010.

[79] Subhra K Biswas and Alberto Mantovani. Macrophage plasticity and interaction with lymphocyte subsets: cancer as a paradigm. *Nat. Immunol.*, 11(10):889–896, October 2010. 16

[80] B Van den Eynde, B Lethé, A Van Pel, E De Plaen, and T Boon. The gene coding for a major tumor rejection antigen of tumor P815 is identical to the normal gene of syngeneic DBA/2 mice. *J. Exp. Med.*, 173(6):1373–1384, June 1991. 16

[81] P van der Bruggen, C Traversari, P Chomez, C Lurquin, E De Plaen, B Van den Eynde, A Knuth, and T Boon. A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science*, 254(5038):1643–1647, December 1991.

[82] S A Rosenberg. Progress in human tumour immunology and immunotherapy. *Nature*, 411(6835): 380–384, May 2001.

[83] Pierre G Coulie, Benoît J Van den Eynde, Pierre van der Bruggen, and Thierry Boon. Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nat. Rev. Cancer*, 14(2):135–146, February 2014. 16

[84] T Boon and P van der Bruggen. Human tumor antigens recognized by T lymphocytes. *J. Exp. Med.*, 183(3):725–729, March 1996. 16

[85] Hung T Khong and Steven A Rosenberg. Pre-existing immunity to tyrosinase-related protein (TRP)-2, a new TRP-2 isoform, and the NY-ESO-1 melanoma antigen in a patient with a dramatic response to immunotherapy. *J. Immunol.*, 168(2):951–956, January 2002.

[86] Deepa Kolaseri Krishnadas, Fanqi Bai, and Kenneth G Lucas. Cancer testis antigen and immunotherapy. *ImmunoTargets Ther.*, 2:11–19, April 2013.

[87] Poulam M Patel, Christian H Ottensmeier, Clive Mulatero, Paul Lorigan, Ruth Plummer, Hardev Pandha, Somaia Elsheikh, Efthymios Hadjimichael, Naty Villasanti, Sally E Adams, Michelle Cunnell, Rachael L Metheringham, Victoria A Brentville, Lee Machado, Ian Daniels, Mohamed Gijon, Drew Hannaman, and Lindy G Durrant. Targeting gp100 and TRP-2 with a DNA vaccine: Incorporating T cell epitopes with a human IgG1 antibody induces potent T cell responses that are associated with favourable clinical outcome in a phase I/II trial. *Oncoimmunology*, 7(6): e1433516, June 2018.

[88] A Marabelle, M G Fakih, J Lopez, M Shah, R Shapira-Frommer, K Nakagawa, H C Chung, H L Kindler, J A Lopez-Martin, W Miller, A Italiano, S Kao, S A Piha-Paul, J-P Delord, R R McWilliams, D Aurora-Garg, M Chen, F Jin, K Norwood, and Y-J Bang. Association of tumour mutational burden with outcomes in patients with select advanced solid tumours treated with pembrolizumab in KEYNOTE-158. *Ann. Oncol.*, 30(Supplement_5):v477–v478, October 2019. 16

[89] H G Rammensee, T Friede, and S Stevanoviíc. MHC ligands and peptide motifs: first listing. *Immunogenetics*, 41(4):178–228, 1995. 16

[90] S Stevanović and H Schild. Quantitative aspects of T cell activation–peptide generation and editing by MHC class I molecules. *Semin. Immunol.*, 11(6):375–384, December 1999.

[91] Felix Klug, Matthias Miller, Hans-Henning Schmidt, and Stefan Stevanovic. Characterization of MHC ligands for peptide based tumor vaccination. *Curr. Pharm. Des.*, 15(28):3221–3236, October 2009. 16

[92] John H Sampson, Amy B Heimberger, Gary E Archer, Kenneth D Aldape, Allan H Friedman, Henry S Friedman, Mark R Gilbert, James E Herndon, 2nd, Roger E McLendon, Duane A Mitchell, David A Reardon, Raymond Sawaya, Robert J Schmittling, Weiming Shi, James J Vredenburgh, and Darell D Bigner. Immunologic escape after prolonged progression-free survival with epidermal growth factor receptor variant III peptide vaccination in patients with newly diagnosed glioblastoma. *J. Clin. Oncol.*, 28(31):4722–4729, November 2010. 16

[93] Steffen Walter, Toni Weinschenk, Arnulf Stenzl, Romuald Zdrojowy, Anna Pluzanska, Cezary Szczylik, Michael Staehler, Wolfram Brugger, Pierre-Yves Dietrich, Regina Mendrzyk, Norbert Hilf, Oliver Schoor, Jens Fritsche, Andrea Mahr, Dominik Maurer, Verona Vass, Claudia Trautwein, Peter Lewandrowski, Christian Flohr, Heike Pohla, Janusz J Stanczak, Vincenzo Bronte, Susanna Mandruzzato, Tilo Biedermann, Graham Pawelec, Evelyna Derhovanessian, Hisakazu Yamagishi, Tsuneharu Miki, Fumiya Hongo, Natsuki Takaha, Kosei Hirakawa, Hiroaki Tanaka, Stefan Stevanovic, Jürgen Frisch, Andrea Mayer-Mokler, Alexandra Kirner, Hans-Georg Rammensee, Carsten Reinhardt, and Harpreet Singh-Jasuja. Multipeptide immune response to cancer vaccine IMA901 after single-dose cyclophosphamide associates with longer patient survival. *Nat. Med.*, 18(8):1254–1261, August 2012. 17

[94] Hans-Georg Rammensee and Harpreet Singh-Jasuja. HLA ligandome tumor antigen discovery for personalized vaccine approach. *Expert Rev. Vaccines*, 12(10):1211–1217, October 2013. 17

[95] C J Barnstable, W F Bodmer, G Brown, G Galfre, C Milstein, A F Williams, and A Ziegler. Production of monoclonal antibodies to group a erythrocytes, HLA and other human cell surface antigens-new tools for genetic analysis. *Cell*, 14(1):9–20, May 1978. 18

[96] Peter Parham and Frances M Brodsky. Partial purification and some properties of BB7.2 a cytotoxic monoclonal antibody with specificity for HLA-A2 and a variant of HLA-A28. *Hum. Immunol.*, 3 (4):277–299, December 1981. 18

[97] N Rebaï and B Malissen. Structural and genetic analyses of HLA class I molecules using monoclonal xenoantibodies. *Tissue Antigens*, 22(2):107–117, August 1983. 18

[98] Christian D Kelstrup, Rosa R Jersie-Christensen, Tanveer S Batth, Tabiwang N Arrey, Andreas Kuehn, Markus Kellmann, and Jesper V Olsen. Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field orbitrap mass spectrometer. *J. Proteome Res.*, 13(12): 6187–6195, December 2014. 18

[99] Ulises H Guzman, Ana Martinez-Val, Zilu Ye, Eugen Damoc, Tabiwang N Arrey, Anna Pashkova, Santosh Renuse, Eduard Denisov, Johannes Petzoldt, Amelia C Peterson, Florian Harking, Ole Østergaard, Rasmus Rydbirk, Susana Aznar, Hamish Stewart, Yue Xuan, Daniel Hermanson, Stevan Horning, Christian Hock, Alexander Makarov, Vlad Zabrouskov, and Jesper V Olsen. Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nat. Biotechnol.*, February 2024.

[100] Elizabeth S Hecht, Michaela Scigelova, Shannon Eliuk, and Alexander Makarov. Fundamentals and advances of orbitrap mass spectrometry, September 2019. 18, 22, 23, 24, 25

[101] Ole Schulz-Trieglaff, Rene Hussong, Clemens Gröpl, Andreas Hildebrandt, and Knut Reinert. A fast and accurate algorithm for the quantification of peptides from mass spectrometry data. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 473–487. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. 18

[102] UniProt Consortium. UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Res.*, 51(D1):D523–D531, January 2023. 18

[103] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, Toby Mathieson, Simone Lemeer, Karsten Schnatbaum, Ulf Reimer, Holger Wenschuh, Martin Mollenhauer, Julia Slotta-Huspenina, Joos-Hendrik Boese, Marcus Bantscheff, Anja Gerstmair, Franz Faerber, and Bernhard Kuster. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, May 2014. 18, 189

[104] Lihua Jiang, Meng Wang, Shin Lin, Ruiqi Jian, Xiao Li, Joanne Chan, Guanlan Dong, Huaying Fang, Aaron E Robinson, GTEx Consortium, and Michael P Snyder. A quantitative proteome map of the human body. *Cell*, 183(1):269–283.e19, October 2020. 18

[105] C E Shannon. Communication in the presence of noise. *Proc. IRE*, 37(1):10–21, January 1949. 18

[106] Jake A Melby, David S Roberts, Eli J Larson, Kyle A Brown, Elizabeth F Bayne, Song Jin, and Ying Ge. Novel strategies to address the challenges in top-down proteomics. *J. Am. Soc. Mass Spectrom.*, 32(6):1278–1294, June 2021. 19

[107] Julia Maria Burkhart, Cornelia Schumbrutzki, Stefanie Wortelkamp, Albert Sickmann, and René Peiman Zahedi. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *J. Proteomics*, 75(4):1454–1462, February 2012. 19

[108] M P Washburn, D Wolters, and J R Yates, 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.*, 19(3):242–247, March 2001. 19

[109] Hannes Hahne, Fiona Pachl, Benjamin Ruprecht, Stefan K Maier, Susan Klaeger, Dominic Helm, Guillaume Médard, Matthias Wilm, Simone Lemeer, and Bernhard Kuster. DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nat. Methods*, 10(10): 989–991, October 2013. 22

[110] Péter Judák, Janelle Grainger, Catrin Goebel, Peter Van Eenoo, and Koen Deventer. DMSO assisted electrospray ionization for the detection of small peptide hormones in urine by dilute-and-shoot-liquid-chromatography-high resolution mass spectrometry. *J. Am. Soc. Mass Spectrom.*, 28(8): 1657–1665, August 2017. 22

[111] A Makarov. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.*, 72(6):1156–1162, March 2000. 24

[112] Niels Hulstaert, Jim Shofstahl, Timo Sachsenberg, Mathias Walzer, Harald Barsnes, Lennart Martens, and Yasset Perez-Riverol. ThermoRawFileParser: Modular, scalable, and cross-platform RAW file conversion. *J. Proteome Res.*, 19(1):537–542, January 2020. 24

[113] Richard H Perry, R Graham Cooks, and Robert J Noll. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom. Rev.*, 27(6):661–699, November 2008. 24

[114] Jesper V Olsen, Lyris M F de Godoy, Guoqing Li, Boris Macek, Peter Mortensen, Reinhold Pesch, Alexander Makarov, Oliver Lange, Stevan Horning, and Matthias Mann. Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a c-trap. *Mol. Cell. Proteomics*, 4(12):2010–2021, December 2005. 24

[115] Jürgen Cox, Annette Michalski, and Matthias Mann. Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.*, 22(8):1373–1380, August 2011. 24

[116] Lindsay K Pino, Brian C Searle, Eric L Huang, William Stafford Noble, Andrew N Hoofnagle, and Michael J MacCoss. Calibration using a single-point external reference material harmonizes quantitative mass spectrometry proteomics data between platforms and laboratories. *Anal. Chem.*, 90(21):13112–13117, November 2018. 24

[117] Alexander Makarov, Eduard Denisov, Alexander Kholomeev, Wilko Balschun, Oliver Lange, Kerstin Strupat, and Stevan Horning. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.*, 78(7):2113–2120, April 2006. 25

[118] M W Senko, S C Beu, and F W McLaffertycor. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.*, 6(4):229–233, April 1995. 25

[119] Anastasia Kalli, Geoffrey T Smith, Michael J Sweredoski, and Sonja Hess. Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: focus on LTQ-Orbitrap mass analyzers. *J. Proteome Res.*, 12(7):3071–3086, July 2013. 26

[120] J Mitchell Wells and Scott A McLuckey. Collision-induced dissociation (CID) of peptides and proteins. In *Methods in Enzymology*, Methods in enzymology, pages 148–185. Elsevier, 2005. 26

[121] Jesper V Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy c-trap dissociation for peptide modification analysis. *Nat. Methods*, 4(9): 709–712, September 2007. 26

[122] John E P Syka, Joshua J Coon, Melanie J Schroeder, Jeffrey Shabanowitz, and Donald F Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.*, 101(26):9528–9533, June 2004. 26

[123] Hannes L Röst, George Rosenberger, Pedro Navarro, Ludovic Gillet, Saša M Miladinović, Olga T Schubert, Witold Wolski, Ben C Collins, Johan Malmström, Lars Malmström, and Ruedi Aebersold. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.*, 32(3):219–223, March 2014. 27

[124] Vadim Demichev, Christoph B Messner, Spyros I Vernardis, Kathryn S Lilley, and Markus Ralser. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods*, 17(1):41–44, January 2020. 27

[125] Fangfei Zhang, Weigang Ge, Lingling Huang, Dan Li, Lijuan Liu, Zhen Dong, Luang Xu, Xuan Ding, Cheng Zhang, Yingying Sun, Jun A, Jinlong Gao, and Tiannan Guo. A comparative analysis of data analysis tools for data-independent acquisition mass spectrometry. *Mol. Cell. Proteomics*, 22(9):100623, September 2023. 27

[126] Hanno Steen and Matthias Mann. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.*, 5(9):699–711, September 2004. 27

[127] Sem Tamara, Maurits A den Boer, and Albert J R Heck. High-resolution native mass spectrometry. *Chem. Rev.*, 122(8):7269–7326, April 2022.

[128] Florian Meier, Melvin A Park, and Matthias Mann. Trapped ion mobility spectrometry and parallel accumulation-serial fragmentation in proteomics. *Mol. Cell. Proteomics*, 20(100138):100138, August 2021.

[129] Edward M Marcotte. How do shotgun proteomics algorithms identify proteins? *Nat. Biotechnol.*, 25(7):755–757, July 2007.

[130] Hayley M Bennett, William Stephenson, Christopher M Rose, and Spyros Darmanis. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nat. Methods*, 20(3): 363–374, March 2023.

[131] Yi Yang and Liang Qiao. Data-independent acquisition proteomics methods for analyzing post-translational modifications. *Proteomics*, 23(7-8):e2200046, April 2023.

[132] Fangfei Zhang, Weigang Ge, Guan Ruan, Xue Cai, and Tiannan Guo. Data-independent acquisition mass spectrometry-based proteomics and software tools: A glimpse in 2020. *Proteomics*, 20 (17-18):e1900276, September 2020.

[133] Shannon Eliuk and Alexander Makarov. Evolution of orbitrap mass spectrometry instrumentation. *Annu. Rev. Anal. Chem. (Palo Alto Calif.)*, 8(1):61–80, 2015. 27

[134] Alessandro Tanca, Marcello Abbondio, Salvatore Pisanu, Daniela Pagnozzi, Sergio Uzzau, and Maria Filippa Addis. Critical comparison of sample preparation strategies for shotgun proteomic analysis of formalin-fixed, paraffin-embedded samples: insights from liver tissue. *Clin. Proteomics*, 11(1):28, July 2014. 27

[135] Frank Klont, Linda Bras, Justina C Wolters, Sara Ongay, Rainer Bischoff, Gyorgy B Halmos, and Péter Horvatovich. Assessment of sample preparation bias in mass spectrometry-based proteomics. *Anal. Chem.*, 90(8):5405–5413, April 2018. 27

[136] Matthias Berg, Axel Parbel, Harald Pettersen, David Fenyö, and Lennart Björkesten. Reproducibility of LC-MS-based protein identification. *J. Exp. Bot.*, 57(7):1509–1514, March 2006. 27

[137] Daniel Stalder, André Haeberli, and Manfred Heller. Evaluation of reproducibility of protein identification results after multidimensional human serum protein separation. *Proteomics*, 8(3): 414–424, February 2008.

[138] Simion Kreimer, Mikhail E Belov, William F Danielson, Lev I Levitsky, Mikhail V Gorshkov, Barry L Karger, and Alexander R Ivanov. Advanced precursor ion selection algorithms for increased depth of bottom-up proteomic profiling. *J. Proteome Res.*, 15(10):3563–3573, October 2016.

[139] Carolina Fernández-Costa, Salvador Martínez-Bartolomé, Daniel B McClatchy, Anthony J Saviola, Nam-Kyung Yu, and John R Yates, 3rd. Impact of the identification strategy on the reproducibility of the DDA and DIA results. *J. Proteome Res.*, 19(8):3153–3161, August 2020. 27

[140] P Roepstorff and J Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.*, 11(11):601, November 1984. 28

[141] R S Johnson, S A Martin, K Biemann, J T Stults, and J T Watson. Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Anal. Chem.*, 59(21):2621–2625, November 1987. 28

[142] Ioannis A Papayannopoulos. The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrom. Rev.*, 14(1):49–73, January 1995. 28

[143] Juri Rappsilber and Matthias Mann. What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.*, 27(2):74–78, February 2002. 28

[144] Steven Carr, Ruedi Aebersold, Michael Baldwin, Al Burlingame, Karl Clauser, Alexey Nesvizhskii, and Working Group on Publication Guidelines for Peptide and Protein Identification Data. The need for guidelines in publication of peptide and protein identification data: Working group on publication guidelines for peptide and protein identification data. *Mol. Cell. Proteomics*, 3(6): 531–533, June 2004. 28

[145] A Kenneth MacLeod, Padraic G Fallon, Sheila Sharp, Colin J Henderson, C Roland Wolf, and Jeffrey T-J Huang. An enhanced in vivo stable isotope labeling by amino acids in cell culture (SILAC) model for quantification of drug metabolism enzymes. *Mol. Cell. Proteomics*, 14(3): 750–760, March 2015. 29

[146] Sandra Orchard, Henning Hermjakob, and Rolf Apweiler. The proteomics standards initiative. *Proteomics*, 3(7):1374–1376, July 2003. 30, 41

[147] Eric W Deutsch, Juan Antonio Vizcaíno, Andrew R Jones, Pierre-Alain Binz, Henry Lam, Joshua Klein, Wout Bittremieux, Yasset Perez-Riverol, David L Tabb, Mathias Walzer, Sylvie Ricard-Blum, Henning Hermjakob, Steffen Neumann, Tytus D Mak, Shin Kawano, Luis Mendoza, Tim Van Den Bossche, Ralf Gabriels, Nuno Bandeira, Jeremy Carver, Benjamin Pullman, Zhi Sun, Nils Hoffmann, Jim Shofstahl, Yunping Zhu, Luana Licata, Federica Quaglia, Silvio C E Tosatto, and Sandra E Orchard. Proteomics standards initiative at twenty years: Current activities and future work. *J. Proteome Res.*, 22(2):287–301, February 2023. 30, 41

[148] Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, Andreas Römpp, Steffen Neumann, Angel D Pizarro, Luisa Montecchi-Palazzi, Natalie Tasman, Mike Coleman, Florian Reisinger, Puneet Souda, Henning Hermjakob, Pierre-Alain Binz, and Eric W Deutsch. mzML–a community standard for mass spectrometry data. *Mol. Cell. Proteomics*, 10(1):R110.000133, January 2011. 30, 41

[149] Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. Open mass spectrometry search algorithm. *J. Proteome Res.*, 3(5):958–964, September 2004. 32

[150] Sangtae Kim and Pavel A Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.*, 5(1):5277, October 2014.

[151] Robertson Craig and Ronald C Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, June 2004.

[152] Jimmy K Eng, Michael R Hoopmann, Tahmina A Jahan, Jarrett D Egertson, William S Noble, and Michael J MacCoss. A deeper look into comet–implementation and features. *J. Am. Soc. Mass Spectrom.*, 26(11):1865–1874, November 2015. 32

[153] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300, January 1995. 33

[154] John D Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *Ann. Stat.*, 31(6):2013–2035, December 2003. 33

[155] Joshua E Elias and Steven P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, 4(3):207–214, March 2007. 33

[156] Lukas Käll, John D Storey, Michael J MacCoss, and William Stafford Noble. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.*, 7 (1):29–34, January 2008. 34

[157] Marc Sturm, Andreas Bertsch, Clemens Gröpl, Andreas Hildebrandt, Rene Hussong, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, Alexandra Zerck, Knut Reinert, and Oliver Kohlbacher. OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics*, 9(1):163, March 2008. 35, 101

[158] Belinda Giardine, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, Webb Miller, W James Kent, and Anton Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, 15(10):1451–1455, October 2005. 37, 90, 100, 101

[159] Michael R Berthold, Nicolas Cebron, Fabian Dill, Thomas R Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. KNIME - the konstanz information miner. *SIGKDD Explor.*, 11(1):26–31, November 2009. 37, 100, 101, 102

[160] Luis de la Garza, Johannes Veit, Andras Szolek, Marc Röttig, Stephan Aiche, Sandra Gesing, Knut Reinert, and Oliver Kohlbacher. From the desktop to the grid: scalable bioinformatics via workflow conversion. *BMC Bioinformatics*, 17(1):127, March 2016. 37, 102

[161] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, 35 (4):316–319, April 2017. 37, 100, 102

[162] Yasset Perez-Riverol, Jingwen Bai, Chakradhar Bandla, David García-Seisdedos, Suresh Hewapathirana, Selvakumar Kamatchinathan, Deepti J Kundu, Ananth Prakash, Anika Frericks-Zipper, Martin Eisenacher, Mathias Walzer, Shengbo Wang, Alvis Brazma, and Juan Antonio Vizcaíno. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.*, 50(D1):D543–D552, January 2022. 40

[163] Florian Reisinger, Manuel Corpas, John Hancock, Henning Hermjakob, Ewan Birney, and Pascal Kahlem. ENFIN - an integrative structure for systems biology. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 132–143. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. 41

[164] Andrew R Jones, Michael Miller, Ruedi Aebersold, Rolf Apweiler, Catherine A Ball, Alvis Brazma, James Degreef, Nigel Hardy, Henning Hermjakob, Simon J Hubbard, Peter Hussey, Mark Igra, Helen Jenkins, Randall K Julian, Jr, Kent Laursen, Stephen G Oliver, Norman W Paton, Susanna-Assunta Sansone, Ugis Sarkans, Christian J Stoeckert, Jr, Chris F Taylor, Patricia L Whetzel, Joseph A White, Paul Spellman, and Angel Pizarro. The functional genomics experiment model (FuGE): an extensible framework for standards in functional genomics. *Nat. Biotechnol.*, 25(10): 1127–1133, October 2007. 41

[165] Andrew R Jones, Martin Eisenacher, Gerhard Mayer, Oliver Kohlbacher, Jennifer Siepen, Simon J Hubbard, Julian N Selley, Brian C Searle, James Shofstahl, Sean L Seymour, Randall Julian, Pierre-Alain Binz, Eric W Deutsch, Henning Hermjakob, Florian Reisinger, Johannes Griss, Juan Antonio Vizcaíno, Matthew Chambers, Angel Pizarro, and David Creasy. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics*, 11(7):M111.014381, July 2012. 41

[166] Mathias Walzer, Da Qi, Gerhard Mayer, Julian Uszkoreit, Martin Eisenacher, Timo Sachsenberg, Faviel F Gonzalez-Galarza, Jun Fan, Conrad Bessant, Eric W Deutsch, Florian Reisinger, Juan Antonio Vizcaíno, J Alberto Medina-Aunon, Juan Pablo Albar, Oliver Kohlbacher, and Andrew R Jones. The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Mol. Cell. Proteomics*, 12(8):2332–2340, August 2013. 41, 50, 185

[167] Andrew Keller, Jimmy Eng, Ning Zhang, Xiao-Jun Li, and Ruedi Aebersold. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.*, 1(1):2005.0017, August 2005. 41

[168] Eric W Deutsch, Luis Mendoza, David Shteynberg, Joseph Slagel, Zhi Sun, and Robert L Moritz. Trans-Proteomic pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin. Appl.*, 9(7-8):745–754, August 2015. 41

[169] Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, Xiao Liang, Sven

Nahnsen, Lars Nilse, Julianus Pfeuffer, George Rosenberger, Marc Rurik, Uwe Schmitt, Johannes Veit, Mathias Walzer, David Wojnar, Witold E Wolski, Oliver Schilling, Jyoti S Choudhary, Lars Malmström, Ruedi Aebersold, Knut Reinert, and Oliver Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods*, 13(9):741–748, September 2016. 41, 102

[170] Juan Antonio Vizcaíno, Gerhard Mayer, Simon Perkins, Harald Barsnes, Marc Vaudel, Yasset Perez-Riverol, Tobias Ternent, Julian Uszkoreit, Martin Eisenacher, Lutz Fischer, Juri Rappsilber, Eugen Netz, Mathias Walzer, Oliver Kohlbacher, Alexander Leitner, Robert J Chalkley, Fawaz Ghali, Salvador Martínez-Bartolomé, Eric W Deutsch, and Andrew R Jones. The mzIdentML data standard version 1.2, supporting advances in proteome informatics. *Mol. Cell. Proteomics*, 16(7): 1275–1285, July 2017. 42, 185

[171] Michael Sperberg-McQueen, Tim Bray, Eve Maler, François Yergeau, and Jean Paoli. Extensible markup language (XML) 1.0 (fifth edition). W3C recommendation, W3C, November 2008. https://www.w3.org/TR/2008/REC-xml-20081126/. 44

[172] Eugen Netz, Tjeerd M H Dijkstra, Timo Sachsenberg, Lukas Zimmermann, Mathias Walzer, Thomas Monecke, Ralf Ficner, Olexandr Dybkov, Henning Urlaub, and Oliver Kohlbacher. OpenPepXL: An open-source tool for sensitive identification of cross-linked peptides in XL-MS. *Mol. Cell. Proteomics*, 19(12):2157–2168, December 2020. 49, 195

[173] Sangtae Kim and Pavel A Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.*, 5(1):5277, October 2014. 49, 102

[174] Jimmy K Eng, Tahmina A Jahan, and Michael R Hoopmann. Comet: an open-source MS/MS sequence database search tool. *Proteomics*, 13(1):22–24, January 2013. 49, 102

[175] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, 4(11):923–925, November 2007. 49

[176] Eric W Deutsch, Nuno Bandeira, Yasset Perez-Riverol, Vagisha Sharma, Jeremy J Carver, Luis Mendoza, Deepti J Kundu, Shengbo Wang, Chakradhar Bandla, Selvakumar Kamatchinathan, Suresh Hewapathirana, Benjamin S Pullman, Julie Wertz, Zhi Sun, Shin Kawano, Shujiro Okuda, Yu Watanabe, Brendan MacLean, Michael J MacCoss, Yunping Zhu, Yasushi Ishihama, and Juan Antonio Vizcaíno. The ProteomeXchange consortium at 10 years: 2023 update. *Nucleic Acids Res.*, 51(D1):D1539–D1548, January 2023. 50

[177] Salvador Martínez-Bartolomé, Pierre-Alain Binz, and Juan P Albar. The minimal information about a proteomics experiment (MIAPE) from the proteomics standards initiative. In *Methods in Molecular Biology*, Methods in molecular biology (Clifton, N.J.), pages 765–780. Humana Press, Totowa, NJ, 2014. 52

[178] Johannes Griss, Andrew R Jones, Timo Sachsenberg, Mathias Walzer, Laurent Gatto, Jürgen Hartler, Gerhard G Thallinger, Reza M Salek, Christoph Steinbeck, Nadin Neuhauser, Jürgen Cox, Steffen Neumann, Jun Fan, Florian Reisinger, Qing-Wei Xu, Noemi Del Toro, Yasset Pérez-Riverol,

Fawaz Ghali, Nuno Bandeira, Ioannis Xenarios, Oliver Kohlbacher, Juan Antonio Vizcaíno, and Henning Hermjakob. The mztab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell. Proteomics*, 13 (10):2765–2775, October 2014. 63

[179] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Darren A Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L Yeh. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.*, 32(Database issue):D115–9, January 2004. 64

[180] Michael Turewicz, Michael Kohl, Maike Ahrens, Gerhard Mayer, Julian Uszkoreit, Wael Naboulsi, Thilo Bracht, Dominik A Megger, Barbara Sitek, Katrin Marcus, and Martin Eisenacher. BioInfra.Prot: A comprehensive proteomics workflow including data standardization, protein inference, expression analysis and data publication. *J. Biotechnol.*, 261:116–125, November 2017. 73

[181] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V Melnik, Michael J Meehan, Wei-Ting Liu, Max Crüsemann, Paul D Boudreau, Eduardo Esquenazi, Mario Sandoval-Calderón, Roland D Kersten, Laura A Pace, Robert A Quinn, Katherine R Duncan, Cheng-Chih Hsu, Dimitrios J Floros, Ronnie G Gavilan, Karin Kleigrewe, Trent Northen, Rachel J Dutton, Delphine Parrot, Erin E Carlson, Bertrand Aigle, Charlotte F Michelsen, Lars Jelsbak, Christian Sohlenkamp, Pavel Pevzner, Anna Edlund, Jeffrey McLean, Jörn Piel, Brian T Murphy, Lena Gerwick, Chih-Chuang Liaw, Yu-Liang Yang, Hans-Ulrich Humpf, Maria Maansson, Robert A Keyzers, Amy C Sims, Andrew R Johnson, Ashley M Sidebottom, Brian E Sedio, Andreas Klitgaard, Charles B Larson, Cristopher A Boya P, Daniel Torres-Mendoza, David J Gonzalez, Denise B Silva, Lucas M Marques, Daniel P Demarque, Egle Pociute, Ellis C O'Neill, Enora Briand, Eric J N Helfrich, Eve A Granatosky, Evgenia Glukhov, Florian Ryffel, Hailey Houson, Hosein Mohimani, Jenan J Kharbush, Yi Zeng, Julia A Vorholt, Kenji L Kurita, Pep Charusanti, Kerry L McPhail, Kristian Fog Nielsen, Lisa Vuong, Maryam Elfeki, Matthew F Traxler, Niclas Engene, Nobuhiro Koyama, Oliver B Vining, Ralph Baric, Ricardo R Silva, Samantha J Mascuch, Sophie Tomasi, Stefan Jenkins, Venkat Macherla, Thomas Hoffman, Vinayak Agarwal, Philip G Williams, Jingqui Dai, Ram Neupane, Joshua Gurr, Andrés M C Rodríguez, Anne Lamsa, Chen Zhang, Kathleen Dorrestein, Brendan M Duggan, Jehad Almaliti, Pierre-Marie Allard, Prasad Phapale, Louis-Felix Nothias, Theodore Alexandrov, Marc Litaudon, Jean-Luc Wolfender, Jennifer E Kyle, Thomas O Metz, Tyler Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav, Rolf Müller, Katrina M Waters, Wenyuan Shi, Xueting Liu, Lixin Zhang, Rob Knight, Paul R Jensen, Bernhard O Palsson, Kit Pogliano, Roger G Linington, Marcelino Gutiérrez, Norberto P Lopes, William H Gerwick, Bradley S Moore, Pieter C Dorrestein, and Nuno Bandeira. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.*, 34(8):828–837, August 2016.

[182] Pavel Sinitcyn, Hamid Hamzeiy, Favio Salinas Soto, Daniel Itzhak, Frank McCarthy, Christoph Wichmann, Martin Steger, Uli Ohmayer, Ute Distler, Stephanie Kaspar-Schoenefeld, Nikita Prianichnikov, Şule Yılmaz, Jan Daniel Rudolph, Stefan Tenzer, Yasset Perez-Riverol, Nagarjuna

Nagaraj, Sean J Humphrey, and Jürgen Cox. MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat. Biotechnol.*, 39(12):1563–1573, December 2021.

[183] Alberto Santos, Ana R Colaço, Annelaura B Nielsen, Lili Niu, Maximilian Strauss, Philipp E Geyer, Fabian Coscia, Nicolai J Wewer Albrechtsen, Filip Mundt, Lars Juhl Jensen, and Matthias Mann. A knowledge graph to interpret clinical proteomics data. *Nat. Biotechnol.*, 40(5):692–702, May 2022.

[184] Elena Krismer, Isabell Bludau, Maximilian T Strauss, and Matthias Mann. AlphaPeptStats: an open-source python package for automated and scalable statistical analysis of mass spectrometry-based proteomics. *Bioinformatics*, 39(8), August 2023. 73

[185] Amol Prakash, Parag Mallick, Jeffrey Whiteaker, Heidi Zhang, Amanda Paulovich, Mark Flory, Hookeun Lee, Ruedi Aebersold, and Benno Schwikowski. Signal maps for mass spectrometry-based comparative proteomics. *Mol. Cell. Proteomics*, 5(3):423–432, March 2006. 77

[186] Hongbin Liu, Rovshan G Sadygov, and John R Yates, 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.*, 76(14):4193–4201, July 2004. 77

[187] Lyris M F de Godoy, Jesper V Olsen, Gustavo A de Souza, Guoqing Li, Peter Mortensen, and Matthias Mann. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.*, 7(6):R50, 2006. 77

[188] Scott J Walmsley, Paul A Rudnick, Yuxue Liang, Qian Dong, Stephen E Stein, and Alexey I Nesvizhskii. Comprehensive analysis of protein digestion using six trypsins reveals the origin of trypsin as a significant source of variability in proteomics. *J. Proteome Res.*, 12(12):5666–5680, December 2013. 77

[189] Henry Rodriguez, Mike Snyder, Mathias Uhlén, Phil Andrews, Ronald Beavis, Christoph Borchers, Robert J Chalkley, Sang Yun Cho, Katie Cottingham, Michael Dunn, Tomasz Dylag, Ron Edgar, Peter Hare, Albert J R Heck, Roland F Hirsch, Karen Kennedy, Patrik Kolar, Hans-Joachim Kraus, Parag Mallick, Alexey Nesvizhskii, Peipei Ping, Fredrik Pontén, Liming Yang, John R Yates, Stephen E Stein, Henning Hermjakob, Christopher R Kinsinger, and Rolf Apweiler. Recommendations from the 2008 international summit on proteomics data release and sharing policy: The amsterdam principles. *J. Proteome Res.*, 8(7):3689–3692, July 2009. 78

[190] Cristina Chiva, Teresa Mendes Maia, Christian Panse, Karel Stejskal, Thibaut Douché, Mariette Matondo, Damarys Loew, Dominic Helm, Mandy Rettel, Karl Mechtler, Francis Impens, Paolo Nanni, Anna Shevchenko, and Eduard Sabidó. Quality standards in proteomics research facilities: Common standards and quality procedures are essential for proteomics facilities and their users. *EMBO Rep.*, 22(6):e52626, June 2021. 78

[191] Chris F Taylor, Norman W Paton, Kathryn S Lilley, Pierre-Alain Binz, Randall K Julian, Jr, Andrew R Jones, Weimin Zhu, Rolf Apweiler, Ruedi Aebersold, Eric W Deutsch, Michael J Dunn, Albert J R Heck, Alexander Leitner, Marcus Macht, Matthias Mann, Lennart Martens, Thomas A Neubert, Scott D Patterson, Peipei Ping, Sean L Seymour, Puneet Souda, Akira Tsugita, Joel Vandekerckhove,

Thomas M Vondriska, Julian P Whitelegge, Marc R Wilkins, Ioannnis Xenarios, John R Yates, 3rd, and Henning Hermjakob. The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.*, 25(8):887–893, August 2007.

[192] Ernesto S Nakayasu, Marina Gritsenko, Paul D Piehowski, Yuqian Gao, Daniel J Orton, Athena A Schepmoes, Thomas L Fillmore, Brigitte I Frohnert, Marian Rewers, Jeffrey P Krischer, Charles Ansong, Astrid M Suchy-Dicey, Carmella Evans-Molina, Wei-Jun Qian, Bobbie-Jo M Webb-Robertson, and Thomas O Metz. Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation. *Nat. Protoc.*, 16(8):3737–3760, August 2021.

[193] David L Tabb. Quality assessment for clinical proteomics. *Clin. Biochem.*, 46(6):411–420, April 2013.

[194] Christopher R Kinsinger, James Apffel, Mark Baker, Xiaopeng Bian, Christoph H Borchers, Ralph Bradshaw, Mi-Youn Brusniak, Daniel W Chan, Eric W Deutsch, Bruno Domon, Jeff Gorman, Rudolf Grimm, William Hancock, Henning Hermjakob, David Horn, Christie Hunter, Patrik Kolar, Hans-Joachim Kraus, Hanno Langen, Rune Linding, Robert L Moritz, Gilbert S Omenn, Ron Orlando, Akhilesh Pandey, Peipei Ping, Amir Rahbar, Robert Rivers, Sean L Seymour, Richard J Simpson, Douglas Slotta, Richard D Smith, Stephen E Stein, David L Tabb, Danilo Tagle, John R Yates, 3rd, and Henry Rodriguez. Recommendations for mass spectrometry data quality metrics for open access data (corollary to the amsterdam principles). *Mol. Cell. Proteomics*, 10(12): O111.015446, December 2011.

[195] Paul A Rudnick, Karl R Clauser, Lisa E Kilpatrick, Dmitrii V Tchekhovskoi, Pedatsur Neta, Niksa Blonder, Dean D Billheimer, Ronald K Blackman, David M Bunk, Helene L Cardasis, Amy-Joan L Ham, Jacob D Jaffe, Christopher R Kinsinger, Mehdi Mesri, Thomas A Neubert, Birgit Schilling, David L Tabb, Tony J Tegeler, Lorenzo Vega-Montoto, Asokan Mulayath Variyath, Mu Wang, Pei Wang, Jeffrey R Whiteaker, Lisa J Zimmerman, Steven A Carr, Susan J Fisher, Bradford W Gibson, Amanda G Paulovich, Fred E Regnier, Henry Rodriguez, Cliff Spiegelman, Paul Tempst, Daniel C Liebler, and Stephen E Stein. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell. Proteomics*, 9(2):225–241, February 2010. 78

[196] Ze-Qiang Ma, Kenneth O Polzin, Surendra Dasari, Matthew C Chambers, Birgit Schilling, Bradford W Gibson, Bao Q Tran, Lorenzo Vega-Montoto, Daniel C Liebler, and David L Tabb. QuaMeter: multivendor performance metrics for LC-MS/MS proteomics instrumentation. *Anal. Chem.*, 84 (14):5845–5850, July 2012. 78

[197] Peter Pichler, Michael Mazanek, Frederico Dusberger, Lisa Weilnböck, Christian G Huber, Christoph Stingl, Theo M Luider, Werner L Straube, Thomas Köcher, and Karl Mechtler. SIMPATIQCO: a server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on orbitrap instruments. *J. Proteome Res.*, 11(11):5540–5547, November 2012. 78

[198] David Morgenstern, Rotem Barzilay, and Yishai Levin. RawBeans: A simple, vendor-independent, raw-data quality-control tool. *J. Proteome Res.*, 20(4):2098–2104, April 2021. 78

[199] Chris Bielow, Guido Mastrobuoni, and Stefan Kempa. Proteomics quality control: Quality control software for MaxQuant results. *J. Proteome Res.*, 15(3):777–787, March 2016. 78

[200] Cristina Chiva, Roger Olivella, Eva Borràs, Guadalupe Espadas, Olga Pastor, Amanda Solé, and Eduard Sabidó. QCloud: A cloud-based quality control system for mass spectrometry-based proteomics laboratories. *PLoS One*, 13(1):e0189209, January 2018. 78

[201] Gerhard Mayer, Luisa Montecchi-Palazzi, David Ovelleiro, Andrew R Jones, Pierre-Alain Binz, Eric W Deutsch, Matthew Chambers, Marius Kallhardt, Fredrik Levander, James Shofstahl, Sandra Orchard, Juan Antonio Vizcaíno, Henning Hermjakob, Christian Stephan, Helmut E Meyer, Martin Eisenacher, and HUPO-PSI Group. The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. *Database (Oxford)*, 2013(0):bat009, March 2013. 78

[202] Michael Kay. XSL transformations (XSLT) version 2.0. W3C recommendation, W3C, January 2007. https://www.w3.org/TR/2007/REC-xslt20-20070123/. 79

[203] Mathias Walzer, Lucia Espona Pernas, Sara Nasso, Wout Bittremieux, Sven Nahnsen, Pieter Kelchtermans, Peter Pichler, Henk W P van den Toorn, An Staes, Jonathan Vandenbussche, Michael Mazanek, Thomas Taus, Richard A Scheltema, Christian D Kelstrup, Laurent Gatto, Bas van Breukelen, Stephan Aiche, Dirk Valkenborg, Kris Laukens, Kathryn S Lilley, Jesper V Olsen, Albert J R Heck, Karl Mechtler, Ruedi Aebersold, Kris Gevaert, Juan Antonio Vizcaíno, Henning Hermjakob, Oliver Kohlbacher, and Lennart Martens. qcML: an exchange format for quality control metrics from mass spectrometry experiments. *Mol. Cell. Proteomics*, 13(8):1905–1913, August 2014. 85, 107, 186

[204] Magdalena Feldhahn, Pierre Dönnes, Philipp Thiel, and Oliver Kohlbacher. FRED–a framework for t-cell epitope detection. *Bioinformatics*, 25(20):2758–2759, October 2009. 88, 90, 93, 96

[205] J W Yewdell and J R Bennink. Immunodominance in major histocompatibility complex class i-restricted T lymphocyte responses. *Annu. Rev. Immunol.*, 17(1):51–88, 1999. 88

[206] K Falk, O Rötzschke, S Stevanović, G Jung, and H G Rammensee. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature*, 351(6324):290–296, May 1991. 88

[207] J B Rothbard and W R Taylor. A sequence pattern common to T cell epitopes. *EMBO J.*, 7(1): 93–100, January 1988. 88

[208] H Rammensee, J Bachmann, N P Emmerich, O A Bachor, and S Stevanović. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219, November 1999. 88, 93

[209] Randi Vita, James A Overton, Jason A Greenbaum, Julia Ponomarenko, Jason D Clark, Jason R Cantrell, Daniel K Wheeler, Joseph L Gabbard, Deborah Hix, Alessandro Sette, and Bjoern Peters. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.*, 43(Database issue):D405–12, January 2015. 88

[210] K C Parker, M A Bednarek, and J E Coligan. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.*, 152(1): 163–175, January 1994. 88, 93

[211] Pierre Dönnes and Arne Elofsson. Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*, 3(1):25, September 2002. 88, 93

[212] S Buus, S L Lauemøller, P Worning, C Kesmir, T Frimurer, S Corbet, A Fomsgaard, J Hilden, A Holm, and S Brunak. Sensitive quantitative predictions of peptide-MHC binding by a 'query by committee' artificial neural network approach. *Tissue Antigens*, 62(5):378–384, November 2003. 88

[213] Pierre Dönnes and Oliver Kohlbacher. Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci.*, 14(8):2132–2140, August 2005. 89, 93, 95

[214] S Tenzer, B Peters, S Bulik, O Schoor, C Lemmel, M M Schatz, P-M Kloetzel, H-G Rammensee, H Schild, and H-G Holzhütter. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage,TAP transport and MHC class I binding. *Cell. Mol. Life Sci.*, 62(9): 1025–1037, May 2005. 89, 93, 95

[215] Sachet A Shukla, Michael S Rooney, Mohini Rajasagi, Grace Tiao, Philip M Dixon, Michael S Lawrence, Jonathan Stevens, William J Lane, Jamie L Dellagatta, Scott Steelman, Carrie Sougnez, Kristian Cibulskis, Adam Kiezun, Nir Hacohen, Vladimir Brusic, Catherine J Wu, and Gad Getz. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.*, 33(11):1152–1158, November 2015. 89, 91, 93, 108

[216] Denis C Bauer, Armella Zadoorian, Laurence O W Wilson, Natalie P Thorne, and Melbourne Genomics Health Alliance. Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. *Brief. Bioinform.*, page bbw097, November 2016. 89

[217] Maria Luisa Matey-Hernandez, Danish Pan Genome Consortium, Søren Brunak, and Jose M G Izarzugaza. Benchmarking the HLA typing performance of polysolver and optitype in 50 danish parental trios. *BMC Bioinformatics*, 19(1), December 2018.

[218] Xiangyong Li, Chi Zhou, Ke Chen, Bingding Huang, Qi Liu, and Hao Ye. Benchmarking HLA genotyping and clarifying HLA impact on survival in tumor immunotherapy. *Mol. Oncol.*, 15(7): 1764–1782, July 2021.

[219] Kazuma Kiyotani, Tu H Mai, and Yusuke Nakamura. Comparison of exome-based HLA class I genotyping tools: identification of platform-specific genotyping errors. *J. Hum. Genet.*, 62(3): 397–405, March 2017. 89

[220] Nora C Toussaint, Yaakov Maman, Oliver Kohlbacher, and Yoram Louzoun. Universal peptide vaccines - optimal peptide vaccine design based on viral sequence conservation. *Vaccine*, 29(47): 8745–8753, November 2011. 90, 92, 93, 96

[221] Benjamin Schubert and Oliver Kohlbacher. Designing string-of-beads vaccines with optimal spacers. *Genome Med.*, 8(1):9, January 2016. 90, 92, 93, 96

[222] Benjamin Schubert, Hans-Philipp Brachvogel, Christopher Jürges, and Oliver Kohlbacher. EpiToolKit–a web-based workbench for vaccine design. *Bioinformatics*, 31(13):2211–2213, July 2015. 90

[223] Yohan Kim, Alessandro Sette, and Bjoern Peters. Applications for t-cell epitope queries and tools in the immune epitope database and analysis resource. *J. Immunol. Methods*, 374(1-2):62–69, November 2011. 90

[224] Damian Smedley, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson, and Arek Kasprzyk. BioMart–biological queries made easy. *BMC Genomics*, 10(1):22, January 2009. 91

[225] Nora C Toussaint and Oliver Kohlbacher. OptiTope–a web server for the selection of an optimal set of peptides for epitope-based vaccines. *Nucleic Acids Res.*, 37(Web Server issue):W617–22, July 2009. 91, 93, 96

[226] Denis V Antonets and Sergei I Bazhan. PolyCTLDesigner: a computational tool for constructing polyepitope t-cell antigens. *BMC Res. Notes*, 6(1):407, October 2013. 92

[227] Huynh-Hoa Bui, John Sidney, Bjoern Peters, Muthuraman Sathiamurthy, Asabe Sinichi, Kelly-Anne Purton, Bianca R Mothé, Francis V Chisari, David I Watkins, and Alessandro Sette. Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*, 57(5):304–314, June 2005. 93

[228] Bjoern Peters and Alessandro Sette. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, 6(1): 132, May 2005. 93

[229] Yohan Kim, John Sidney, Clemencia Pinilla, Alessandro Sette, and Bjoern Peters. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a bayesian prior. *BMC Bioinformatics*, 10(1):394, November 2009. 93

[230] John Sidney, Erika Assarsson, Carrie Moore, Sandy Ngo, Clemencia Pinilla, Alessandro Sette, and Bjoern Peters. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Res.*, 4 (1):2, January 2008. 93

[231] Hao Zhang, Ole Lund, and Morten Nielsen. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics*, 25(10):1293–1299, May 2009. 93

[232] Massimo Andreatta and Morten Nielsen. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, 32(4):511–517, February 2016. 93

[233] Morten Nielsen and Massimo Andreatta. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.*, 8(1):33, March 2016. 93

[234] T Sturniolo, E Bono, J Ding, L Raddrizzani, O Tuereci, U Sahin, M Braxenthaler, F Gallazzi, M P Protti, F Sinigaglia, and J Hammer. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.*, 17(6): 555–561, June 1999. 93

[235] Lianming Zhang, Yiqing Chen, Hau-San Wong, Shuigeng Zhou, Hiroshi Mamitsuka, and Shanfeng Zhu. TEPITOPEpan: Extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules. *PLoS One*, 7(2):e30483, February 2012. 93

[236] Morten Nielsen, Claus Lundegaard, and Ole Lund. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*, 8(1):238, July 2007. 93

[237] Edita Karosiene, Michael Rasmussen, Thomas Blicher, Ole Lund, Søren Buus, and Morten Nielsen. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*, 65(10):711–724, October 2013. 93

[238] Nora C Toussaint, Magdalena Feldhahn, Matthias Ziehm, Stefan Stevanović, and Oliver Kohlbacher. T-cell epitope prediction based on self-tolerance. In *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, New York, NY, USA, August 2011. ACM. 93

[239] Thomas Stranzl, Mette Voldby Larsen, Claus Lundegaard, and Morten Nielsen. NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics*, 62(6):357–368, June 2010. 93

[240] Jorg J A Calis, Matt Maybeno, Jason A Greenbaum, Daniela Weiskopf, Aruna D De Silva, Alessandro Sette, Can Keşmir, and Bjoern Peters. Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.*, 9(10):e1003266, October 2013. 93

[241] Ido Ginodi, Tal Vider-Shalit, Lea Tsaban, and Yoram Louzoun. Precise score for the prediction of peptides cleaved by the proteasome. *Bioinformatics*, 24(4):477–483, February 2008. 93

[242] Morten Nielsen, Claus Lundegaard, Ole Lund, and Can Keşmir. The role of the proteasome in generating cytotoxic t-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, 57(1-2):33–41, April 2005. 93, 95

[243] Björn Peters, Sascha Bulik, Robert Tampe, Peter M Van Endert, and Hermann-Georg Holzhütter. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol.*, 171(4):1741–1749, August 2003. 93, 95

[244] Irini Doytchinova, Shelley Hemsley, and Darren R Flower. Transporter associated with antigen processing preselection of peptides binding to the MHC: a bioinformatic evaluation. *J. Immunol.*, 173(11):6813–6819, December 2004. 93, 95

[245] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, 38(16):e164, September 2010. 94, 95

[246] William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. Deriving the consequences of genomic variants with the ensembl API and SNP effect predictor. *Bioinformatics*, 26(16):2069–2070, August 2010. 95

[247] Randi Vita, James A Overton, Jason A Greenbaum, Julia Ponomarenko, Jason D Clark, Jason R Cantrell, Daniel K Wheeler, Joseph L Gabbard, Deborah Hix, Alessandro Sette, and Bjoern Peters. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.*, 43(Database issue):D405–12, January 2015. 97

[248] Benjamin Schubert, Hans-Philipp Brachvogel, Christopher Jürges, and Oliver Kohlbacher. EpiToolKit–a web-based workbench for vaccine design. *Bioinformatics*, 31(13):2211–2213, July 2015. 97

[249] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS Comput. Biol.*, 9(10):e1003285, October 2013. 100

[250] M. D. McIlroy. A research unix reader: Annotated excerpts from the programmer's manual, 1971-1986, June 1987. `https://archive.org/details/research-unix-reader`, accessed 01.07.2024. 100

[251] Oliver Kohlbacher, Knut Reinert, Clemens Gröpl, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, and Marc Sturm. TOPP–the OpenMS proteomics pipeline. *Bioinformatics*, 23(2):e191–7, January 2007. 101, 102

[252] Johannes Junker, Chris Bielow, Andreas Bertsch, Marc Sturm, Knut Reinert, and Oliver Kohlbacher. TOPPAS: a graphical workflow editor for the analysis of high-throughput proteomics data. *J. Proteome Res.*, 11(7):3914–3920, July 2012. 101, 102

[253] Benjamin C Orsburn. Proteome Discoverer-A community enhanced data processing suite for protein informatics. *Proteomes*, 9(1):15, March 2021. 101

[254] Michael R Crusoe, Sanne Abeln, Alexandru Iosup, Peter Amstutz, John Chilton, Nebojša Tijanić, Hervé Ménager, Stian Soiland-Reyes, Bogdan Gavrilović, Carole Goble, and The CWL Community. Methods included. *Commun. ACM*, 65(6):54–63, June 2022. 101

[255] D N Perkins, D J Pappin, D M Creasy, and J S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, December 1999. 102

[256] Robertson Craig and Ronald C Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, June 2004. 102

[257] Johannes Griss, Andrew R Jones, Timo Sachsenberg, Mathias Walzer, Laurent Gatto, Jürgen Hartler, Gerhard G Thallinger, Reza M Salek, Christoph Steinbeck, Nadin Neuhauser, Jürgen Cox,

Steffen Neumann, Jun Fan, Florian Reisinger, Qing-Wei Xu, Noemi Del Toro, Yasset Pérez-Riverol, Fawaz Ghali, Nuno Bandeira, Ioannis Xenarios, Oliver Kohlbacher, Juan Antonio Vizcaíno, and Henning Hermjakob. The mztab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell. Proteomics*, 13 (10):2765–2775, October 2014. 103, 186

[258] Florian P Breitwieser, André Müller, Loïc Dayon, Thomas Köcher, Alexandre Hainard, Peter Pichler, Ursula Schmidt-Erfurth, Giulio Superti-Furga, Jean-Charles Sanchez, Karl Mechtler, Keiryn L Bennett, and Jacques Colinge. General statistical modeling of data from protein relative expression isobaric tags. *J. Proteome Res.*, 10(6):2758–2766, June 2011. 106

[259] V Boisguérin, J C Castle, M Loewer, J Diekmann, F Mueller, C M Britten, S Kreiter, Ö Türeci, and U Sahin. Translation of genomics-guided RNA-based personalised cancer vaccines: towards the bedside. *Br. J. Cancer*, 111(8):1469–1475, October 2014. 108

[260] Daniel J Kowalewski, Stefan Stevanovic, Hans-Georg Rammensee, and Juliane S Stickel. Antileukemia t-cell responses in CLL - we don't need no aberration. *Oncoimmunology*, 4(7):e1011527, July 2015. 108

[261] Janet Kerstin Peper, Hans-Christian Bösmüller, Heiko Schuster, Brigitte Gückel, Helen Hörzer, Kevin Roehle, Richard Schäfer, Philipp Wagner, Hans-Georg Rammensee, Stefan Stevanović, Falko Fend, and Annette Staebler. HLA ligandomics identifies histone deacetylase 1 as target for ovarian cancer immunotherapy. *Oncoimmunology*, 5(5):e1065369, May 2016.

[262] Daniel J Kowalewski, Heiko Schuster, Linus Backert, Claudia Berlin, Stefan Kahn, Lothar Kanz, Helmut R Salih, Hans-Georg Rammensee, Stefan Stevanovic, and Juliane Sarah Stickel. HLA ligandome analysis identifies the underlying specificities of spontaneous antileukemia immune responses in chronic lymphocytic leukemia (CLL). *Proc. Natl. Acad. Sci. U. S. A.*, 112(2):E166–75, January 2015. 108

[263] Ryan M Taylor, Jamison Dance, Russ J Taylor, and John T Prince. Metriculator: quality assessment for mass spectrometry-based proteomics. *Bioinformatics*, 29(22):2948–2949, November 2013. 110

[264] Jesper V Olsen, Lyris M F de Godoy, Guoqing Li, Boris Macek, Peter Mortensen, Reinhold Pesch, Alexander Makarov, Oliver Lange, Stevan Horning, and Matthias Mann. Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a c-trap. *Mol. Cell. Proteomics*, 4(12):2010–2021, December 2005. 112

[265] J Kyte and R F Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1):105–132, May 1982. 112

[266] Lennart Martens, Juan Antonio Vizcaíno, and Roz Banks. Quality control in proteomics. *Proteomics*, 11(6):1015–1016, March 2011. 113

[267] Christopher R Kinsinger, James Apffel, Mark Baker, Xiaopeng Bian, Christoph H Borchers, Ralph Bradshaw, Mi-Youn Brusniak, Daniel W Chan, Eric W Deutsch, Bruno Domon, Jeff Gorman,

Rudolf Grimm, William Hancock, Henning Hermjakob, David Horn, Christie Hunter, Patrik Kolar, Hans-Joachim Kraus, Hanno Langen, Rune Linding, Robert L Moritz, Gilbert S Omenn, Ron Orlando, Akhilesh Pandey, Peipei Ping, Amir Rahbar, Robert Rivers, Sean L Seymour, Richard J Simpson, Douglas Slotta, Richard D Smith, Stephen E Stein, David L Tabb, Danilo Tagle, John R Yates, and Henry Rodriguez. Recommendations for mass spectrometry data quality metrics for open access data (corollary to the amsterdam principles). *J. Proteome Res.*, 11(2):1412–1419, February 2012.

[268] Thomas Köcher, Peter Pichler, Remco Swart, and Karl Mechtler. Quality control in LC-MS/MS. *Proteomics*, 11(6):1026–1030, March 2011. 113

[269] Peter Pichler, Michael Mazanek, Frederico Dusberger, Lisa Weilnböck, Christian G Huber, Christoph Stingl, Theo M Luider, Werner L Straube, Thomas Köcher, and Karl Mechtler. SIMPATIQCO: a server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on orbitrap instruments. *J. Proteome Res.*, 11(11):5540–5547, November 2012. 113, 119

[270] Cristina Chiva, Roger Olivella, Eva Borràs, Guadalupe Espadas, Olga Pastor, Amanda Solé, and Eduard Sabidó. QCloud: A cloud-based quality control system for mass spectrometry-based proteomics laboratories. *PLoS One*, 13(1):e0189209, January 2018. 113

[271] Sebastian Boegel, Martin Löwer, Thomas Bukur, Ugur Sahin, and John C Castle. A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. *Oncoimmunology*, 3(8):e954893, August 2014. 113

[272] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, December 2008. 114

[273] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001. 114

[274] Jürgen Cox, Annette Michalski, and Matthias Mann. Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.*, 22(8):1373–1380, August 2011. 119

[275] Purva Kulkarni, Filip Kaftan, Philipp Kynast, Aleš Svatoš, and Sebastian Böcker. Correcting mass shifts: A lock mass-free recalibration procedure for mass spectrometry imaging data. *Anal. Bioanal. Chem.*, 407(25):7603–7613, October 2015. 119

[276] Bryan A Stanfill, Ernesto S Nakayasu, Lisa M Bramer, Allison M Thompson, Charles K Ansong, Therese R Clauss, Marina A Gritsenko, Matthew E Monroe, Ronald J Moore, Daniel J Orton, Paul D Piehowski, Athena A Schepmoes, Richard D Smith, Bobbie-Jo M Webb-Robertson, Thomas O Metz, and TEDDY Study Group. Quality control analysis in real-time (QC-ART): A tool for real-time quality control assessment of mass spectrometry-based proteomics data. *Mol. Cell. Proteomics*, 17 (9):1824–1836, September 2018. 119

[277] Ernesto S Nakayasu, Marina Gritsenko, Paul D Piehowski, Yuqian Gao, Daniel J Orton, Athena A Schepmoes, Thomas L Fillmore, Brigitte I Frohnert, Marian Rewers, Jeffrey P Krischer, Charles Ansong, Astrid M Suchy-Dicey, Carmella Evans-Molina, Wei-Jun Qian, Bobbie-Jo M Webb-Robertson,

and Thomas O Metz. Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation. *Nat. Protoc.*, 16(8):3737–3760, August 2021. 119

[278] Jennifer A Kirwan, Helen Gika, Richard D Beger, Dan Bearden, Warwick B Dunn, Royston Goodacre, Georgios Theodoridis, Michael Witting, Li-Rong Yu, Ian D Wilson, and metabolomics Quality Assurance and Quality Control Consortium (mQACC). Quality assurance and quality control reporting in untargeted metabolic phenotyping: mQACC recommendations for analytical quality management. *Metabolomics*, 18(9):70, August 2022. 119

[279] David Broadhurst, Royston Goodacre, Stacey N Reinke, Julia Kuligowski, Ian D Wilson, Matthew R Lewis, and Warwick B Dunn. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics*, 14(6), June 2018. 119

[280] Lindsey A Torre, Freddie Bray, Rebecca L Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. Global cancer statistics, 2012. *CA Cancer J. Clin.*, 65(2):87–108, March 2015. 122

[281] European Association For The Study Of The Liver and European Organisation For Research And Treatment Of Cancer. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. *J. Hepatol.*, 56(4):908–943, April 2012. 122

[282] Tim Chan, Robert H Wiltrout, and Jonathan M Weiss. Immunotherapeutic modulation of the suppressive liver and tumor microenvironments. *Int. Immunopharmacol.*, 11(7):879–889, July 2011. 122

[283] Lisa H Butterfield, Antoni Ribas, Douglas M Potter, and James S Economou. Spontaneous and vaccine induced AFP-specific T cell phenotypes in subjects with AFP-positive hepatocellular cancer. *Cancer Immunol. Immunother.*, 56(12):1931–1943, December 2007. 122

[284] Wei Yao, Jun-Chuang He, Yan Yang, Jian-Ming Wang, Ya-Wei Qian, Tao Yang, and Lei Ji. The prognostic value of tumor-infiltrating lymphocytes in hepatocellular carcinoma: A systematic review and meta-analysis. *Sci. Rep.*, 7(1):7525, August 2017. 122

[285] Esther Unitt, Aileen Marshall, William Gelson, Simon M Rushbrook, Susan Davies, Sarah L Vowler, Lesley S Morris, Nicholas Coleman, and Graeme J M Alexander. Tumour lymphocytic infiltrate and recurrence of hepatocellular carcinoma following liver transplantation. *J. Hepatol.*, 45(2): 246–253, August 2006. 122

[286] James Larkin, Vanna Chiarion-Sileni, Rene Gonzalez, Jean Jacques Grob, C Lance Cowey, Christopher D Lao, Dirk Schadendorf, Reinhard Dummer, Michael Smylie, Piotr Rutkowski, Pier F Ferrucci, Andrew Hill, John Wagstaff, Matteo S Carlino, John B Haanen, Michele Maio, Ivan Marquez-Rodas, Grant A McArthur, Paolo A Ascierto, Georgina V Long, Margaret K Callahan, Michael A Postow, Kenneth Grossmann, Mario Sznol, Brigitte Dreno, Lars Bastholt, Arvin Yang, Linda M Rollin, Christine Horak, F Stephen Hodi, and Jedd D Wolchok. Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *N. Engl. J. Med.*, 373(1):23–34, July 2015. 122

[287] Dirk Schadendorf, F Stephen Hodi, Caroline Robert, Jeffrey S Weber, Kim Margolin, Omid Hamid, Debra Patt, Tai-Tsang Chen, David M Berman, and Jedd D Wolchok. Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma. *J. Clin. Oncol.*, 33(17):1889–1894, June 2015. 122

[288] Ton N Schumacher and Robert D Schreiber. Neoantigens in cancer immunotherapy. *Science*, 348 (6230):69–74, April 2015. 122, 132

[289] Hans-Georg Rammensee and Harpreet Singh-Jasuja. HLA ligandome tumor antigen discovery for personalized vaccine approach. *Expert Rev. Vaccines*, 12(10):1211–1217, October 2013. 122

[290] Robert H Vonderheide and Katherine L Nathanson. Immunotherapy at large: the road to personalized cancer vaccines. *Nat. Med.*, 19(9):1098–1100, September 2013. 122

[291] Alexandra Snyder, Vladimir Makarov, Taha Merghoub, Jianda Yuan, Jesse M Zaretsky, Alexis Desrichard, Logan A Walsh, Michael A Postow, Phillip Wong, Teresa S Ho, Travis J Hollmann, Cameron Bruggeman, Kasthuri Kannan, Yanyun Li, Ceyhan Elipenahli, Cailian Liu, Christopher T Harbison, Lisu Wang, Antoni Ribas, Jedd D Wolchok, and Timothy A Chan. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.*, 371(23):2189–2199, December 2014. 122, 132

[292] Naiyer A Rizvi, Matthew D Hellmann, Alexandra Snyder, Pia Kvistborg, Vladimir Makarov, Jonathan J Havel, William Lee, Jianda Yuan, Phillip Wong, Teresa S Ho, Martin L Miller, Natasha Rekhtman, Andre L Moreira, Fawzia Ibrahim, Cameron Bruggeman, Billel Gasmi, Roberta Zappasodi, Yuka Maeda, Chris Sander, Edward B Garon, Taha Merghoub, Jedd D Wolchok, Ton N Schumacher, and Timothy A Chan. Cancer immunology. mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*, 348(6230):124–128, April 2015. 122

[293] Dung T Le, Jennifer N Uram, Hao Wang, Bjarne R Bartlett, Holly Kemberling, Aleksandra D Eyring, Andrew D Skora, Brandon S Luber, Nilofer S Azad, Dan Laheru, Barbara Biedrzycki, Ross C Donehower, Atif Zaheer, George A Fisher, Todd S Crocenzi, James J Lee, Steven M Duffy, Richard M Goldberg, Albert de la Chapelle, Minori Koshiji, Feriyl Bhaijee, Thomas Huebner, Ralph H Hruban, Laura D Wood, Nathan Cuka, Drew M Pardoll, Nickolas Papadopoulos, Kenneth W Kinzler, Shibin Zhou, Toby C Cornish, Janis M Taube, Robert A Anders, James R Eshleman, Bert Vogelstein, and Luis A Diaz, Jr. PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.*, 372 (26):2509–2520, June 2015. 123, 132

[294] Matthew M Gubin and Robert D Schreiber. Cancer. the odds of immunotherapy success. *Science*, 350(6257):158–159, October 2015. 123, 132, 134

[295] Nienke van Rooij, Marit M van Buuren, Daisy Philips, Arno Velds, Mireille Toebes, Bianca Heemskerk, Laura J A van Dijk, Sam Behjati, Henk Hilkmann, Dris el Atmioui, Marja Nieuwland, Michael R Stratton, Ron M Kerkhoven, Can Keşmir, John B Haanen, Pia Kvistborg, and Ton N Schumacher. Tumor exome analysis reveals neoantigen-specific t-cell reactivity in an ipilimumab-responsive melanoma. *J. Clin. Oncol.*, 31(32):e439–e442, November 2013. 123

[296] Eric Tran, Simon Turcotte, Alena Gros, Paul F Robbins, Yong-Chen Lu, Mark E Dudley, John R Wunderlich, Robert P Somerville, Katherine Hogan, Christian S Hinrichs, Maria R Parkhurst, James C Yang, and Steven A Rosenberg. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science*, 344(6184):641–645, May 2014. 123, 132

[297] Edward F Fritsch, Mohini Rajasagi, Patrick A Ott, Vladimir Brusic, Nir Hacohen, and Catherine J Wu. HLA-binding properties of tumor neoepitopes in humans. *Cancer Immunol. Res.*, 2(6): 522–529, June 2014. 123

[298] Matthew M Gubin, Maxim N Artyomov, Elaine R Mardis, and Robert D Schreiber. Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J. Clin. Invest.*, 125 (9):3413–3421, September 2015.

[299] Mohini Rajasagi, Sachet A Shukla, Edward F Fritsch, Derin B Keskin, David DeLuca, Ellese Carmona, Wandi Zhang, Carrie Sougnez, Kristian Cibulskis, John Sidney, Kristen Stevenson, Jerome Ritz, Donna Neuberg, Vladimir Brusic, Stacey Gabriel, Eric S Lander, Gad Getz, Nir Hacohen, and Catherine J Wu. Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood*, 124(3):453–462, July 2014. 123

[300] Eric Tran, Mojgan Ahmadzadeh, Yong-Chen Lu, Alena Gros, Simon Turcotte, Paul F Robbins, Jared J Gartner, Zhili Zheng, Yong F Li, Satyajit Ray, John R Wunderlich, Robert P Somerville, and Steven A Rosenberg. Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science*, 350(6266):1387–1390, December 2015. 123

[301] Matthew M Gubin, Xiuli Zhang, Heiko Schuster, Etienne Caron, Jeffrey P Ward, Takuro Noguchi, Yulia Ivanova, Jasreet Hundal, Cora D Arthur, Willem-Jan Krebber, Gwenn E Mulder, Mireille Toebes, Matthew D Vesely, Samuel S K Lam, Alan J Korman, James P Allison, Gordon J Freeman, Arlene H Sharpe, Erika L Pearce, Ton N Schumacher, Ruedi Aebersold, Hans-Georg Rammensee, Cornelis J M Melief, Elaine R Mardis, William E Gillanders, Maxim N Artyomov, and Robert D Schreiber. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature*, 515(7528):577–581, November 2014. 123, 132, 133

[302] Mahesh Yadav, Suchit Jhunjhunwala, Qui T Phung, Patrick Lupardus, Joshua Tanguay, Stephanie Bumbaca, Christian Franci, Tommy K Cheung, Jens Fritsche, Toni Weinschenk, Zora Modrusan, Ira Mellman, Jennie R Lill, and Lélia Delamarre. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature*, 515(7528):572–576, November 2014. 123

[303] Michal Bassani-Sternberg, Eva Bräunlein, Richard Klar, Thomas Engleitner, Pavel Sinitcyn, Stefan Audehm, Melanie Straub, Julia Weber, Julia Slotta-Huspenina, Katja Specht, Marc E Martignoni, Angelika Werner, Rüdiger Hein, Dirk H Busch, Christian Peschel, Roland Rad, Jürgen Cox, Matthias Mann, and Angela M Krackhardt. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.*, 7(1):13404, November 2016. 123, 126, 131, 134

[304] Shelly Kalaora, Yochai Wolf, Tali Feferman, Eilon Barnea, Erez Greenstein, Dan Reshef, Itay Tirosh, Alexandre Reuben, Sushant Patkar, Ronen Levy, Juliane Quinkhardt, Tana Omokoko, Nouar Qutob,

Ofra Golani, Jianhua Zhang, Xizeng Mao, Xingzhi Song, Chantale Bernatchez, Cara Haymaker, Marie-Andrée Forget, Caitlin Creasy, Polina Greenberg, Brett W Carter, Zachary A Cooper, Steven A Rosenberg, Michal Lotem, Ugur Sahin, Guy Shakhar, Eytan Ruppin, Jennifer A Wargo, Nir Friedman, Arie Admon, and Yardena Samuels. Combined analysis of antigen presentation and t-cell recognition reveals restricted immune responses in melanoma. *Cancer Discov.*, 8(11): 1366–1375, November 2018. 123, 134

[305] Ludmil B Alexandrov, Australian Pancreatic Cancer Genome Initiative, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P Butler, Carlos Caldas, Helen R Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Natalie Jäger, David T W Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C Munshi, Hiromi Nakamura, Paul A Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V Pearson, Xose S Puente, Keiran Raine, Manasa Ramakrishna, Andrea L Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N Schumacher, Paul N Span, Jon W Teague, Yasushi Totoki, Andrew N J Tutt, Rafael Valdés-Mas, Marit M van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R Yates, Jessica Zucman-Rossi, P Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M Pfister, Peter J Campbell, Michael R Stratton, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, and ICGC PedBrain. Signatures of mutational processes in human cancer. *Nature*, 500(7463): 415–421, August 2013. 123, 134

[306] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, January 2014. 123

[307] Zhengyan Kan, Hancheng Zheng, Xiao Liu, Shuyu Li, Thomas D Barber, Zhuolin Gong, Huan Gao, Ke Hao, Melinda D Willard, Jiangchun Xu, Robert Hauptschein, Paul A Rejto, Julio Fernandez, Guan Wang, Qinghui Zhang, Bo Wang, Ronghua Chen, Jian Wang, Nikki P Lee, Wei Zhou, Zhao Lin, Zhiyu Peng, Kang Yi, Shengpei Chen, Lin Li, Xiaomei Fan, Jie Yang, Rui Ye, Jia Ju, Kai Wang, Heather Estrella, Shibing Deng, Ping Wei, Ming Qiu, Isabella H Wulur, Jiangang Liu, Mariam E Ehsani, Chunsheng Zhang, Andrey Loboda, Wing Kin Sung, Amit Aggarwal, Ronnie T Poon, Sheung Tat Fan, Jun Wang, James Hardwick, Christoph Reinhard, Hongyue Dai, Yingrui Li, John M Luk, and Mao Mao. Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res.*, 23(9):1422–1433, September 2013. 123, 134

[308] Anthony B El-Khoueiry, Bruno Sangro, Thomas Yau, Todd S Crocenzi, Masatoshi Kudo, Chiun Hsu, Tae-You Kim, Su-Pin Choo, Jörg Trojan, Theodore H Welling, 3rd, Tim Meyer, Yoon-Koo Kang, Winnie Yeo, Akhil Chopra, Jeffrey Anderson, Christine dela Cruz, Lixin Lang, Jaclyn Neely, Hao Tang, Homa B Dastani, and Ignacio Melero. Nivolumab in patients with advanced hepatocellular carcinoma (CheckMate 040): an open-label, non-comparative, phase 1/2 dose escalation and expansion trial. *Lancet*, 389(10088):2492–2502, June 2017. 123

[309] Hannes L Röst, Timo Sachsenberg, Stephan Aiche, Chris Bielow, Hendrik Weisser, Fabian Aicheler, Sandro Andreotti, Hans-Christian Ehrlich, Petra Gutenbrunner, Erhan Kenar, Xiao Liang, Sven Nahnsen, Lars Nilse, Julianus Pfeuffer, George Rosenberger, Marc Rurik, Uwe Schmitt, Johannes Veit, Mathias Walzer, David Wojnar, Witold E Wolski, Oliver Schilling, Jyoti S Choudhary, Lars Malmström, Ruedi Aebersold, Knut Reinert, and Oliver Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods*, 13(9):741–748, September 2016. 123

[310] Jimmy K Eng, Tahmina A Jahan, and Michael R Hoopmann. Comet: an open-source MS/MS sequence database search tool. *Proteomics*, 13(1):22–24, January 2013. 123

[311] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, 4(11):923–925, November 2007. 123

[312] András Szolek, Benjamin Schubert, Christopher Mohr, Marc Sturm, Magdalena Feldhahn, and Oliver Kohlbacher. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*, 30(23):3310–3316, December 2014. 124

[313] Markus W Löffler, P Anoop Chandran, Karoline Laske, Christopher Schroeder, Irina Bonzheim, Mathias Walzer, Franz J Hilke, Nico Trautwein, Daniel J Kowalewski, Heiko Schuster, Marc Günder, Viviana A Carcamo Yañez, Christopher Mohr, Marc Sturm, Huu-Phuc Nguyen, Olaf Riess, Peter Bauer, Sven Nahnsen, Silvio Nadalin, Derek Zieker, Jörg Glatzle, Karolin Thiel, Nicole Schneiderhan-Marra, Stephan Clasen, Hans Bösmüller, Falko Fend, Oliver Kohlbacher, Cécile Gouttefangeas, Stefan Stevanović, Alfred Königsrainer, and Hans-Georg Rammensee. Personalized peptide vaccine-induced immune response associated with long-term survival of a metastatic cholangiocarcinoma patient. *J. Hepatol.*, 65(4):849–855, October 2016. 124

[314] H Rammensee, J Bachmann, N P Emmerich, O A Bachor, and S Stevanović. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219, November 1999. 124

[315] Massimo Andreatta and Morten Nielsen. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, 32(4):511–517, February 2016. 124

[316] Morten Nielsen, Claus Lundegaard, Peder Worning, Sanne Lise Lauemøller, Kasper Lamberth, Søren Buus, Søren Brunak, and Ole Lund. Reliable prediction of t-cell epitopes using neural networks with novel sequence representations. *Protein Sci.*, 12(5):1007–1017, May 2003. 124

[317] Ilka Hoof, Bjoern Peters, John Sidney, Lasse Eggers Pedersen, Alessandro Sette, Ole Lund, Søren Buus, and Morten Nielsen. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*, 61(1):1–13, January 2009. 124

[318] Morten Nielsen and Massimo Andreatta. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.*, 8(1):33, March 2016. 124

[319] Benjamin Schubert, Mathias Walzer, Hans-Philipp Brachvogel, András Szolek, Christopher Mohr, and Oliver Kohlbacher. FRED 2: an immunoinformatics framework for python. *Bioinformatics*, 32(13):2044–2046, July 2016. 124, 187

[320] Christopher Mohr, Andreas Friedrich, David Wojnar, Erhan Kenar, Aydin Can Polatkan, Marius Cosmin Codrea, Stefan Czemmel, Oliver Kohlbacher, and Sven Nahnsen. qportal: A platform for data-driven biomedical research. *PLoS One*, 13(1):e0191603, January 2018. 124

[321] C J Barnstable, W F Bodmer, G Brown, G Galfre, C Milstein, A F Williams, and A Ziegler. Production of monoclonal antibodies to group a erythrocytes, HLA and other human cell surface antigens-new tools for genetic analysis. *Cell*, 14(1):9–20, May 1978. 125

[322] Daniel J Kowalewski and Stefan Stevanović. Biochemical large-scale identification of MHC class I ligands. In *Antigen Processing*, Methods in molecular biology (Clifton, N.J.), pages 145–157. Humana Press, Totowa, NJ, 2013. 125

[323] Markus W Löffler, Daniel J Kowalewski, Linus Backert, Jörg Bernhardt, Patrick Adam, Heiko Schuster, Florian Dengler, Daniel Backes, Hans-Georg Kopp, Stefan Beckert, Silvia Wagner, Ingmar Königsrainer, Oliver Kohlbacher, Lothar Kanz, Alfred Königsrainer, Hans-Georg Rammensee, Stefan Stevanović, and Sebastian P Haen. Mapping the HLA ligandome of colorectal cancer reveals an imprint of malignant cell transformation. *Cancer Res.*, 78(16):4627–4641, August 2018. 125, 133

[324] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009. 125

[325] Sangtae Kim, Konrad Scheffler, Aaron L Halpern, Mitchell A Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, Yeonbin Kim, Doruk Beyter, Peter Krusche, and Christopher T Saunders. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, 15(8): 591–594, August 2018. 125

[326] Christopher T Saunders, Wendy S W Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, July 2012. 125

[327] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, 6(2):80–92, April 2012. 125

[328] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4):R36, April 2013. 125

[329] Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, 26 (12):1367–1372, December 2008. 126

[330] Zachary R Chalmers, Caitlin F Connelly, David Fabrizio, Laurie Gay, Siraj M Ali, Riley Ennis, Alexa Schrock, Brittany Campbell, Adam Shlien, Juliann Chmielecki, Franklin Huang, Yuting He, James Sun, Uri Tabori, Mark Kennedy, Daniel S Lieber, Steven Roels, Jared White, Geoffrey A Otto, Jeffrey S Ross, Levi Garraway, Vincent A Miller, Phillip J Stephens, and Garrett M Frampton. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.*, 9(1), December 2017. 127

[331] Mark J P Chaisson, Richard K Wilson, and Evan E Eichler. Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.*, 16(11):627–640, November 2015. 127

[332] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nat. Rev. Cancer*, 4(3):177–183, March 2004. 127

[333] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928): 198–207, March 2003. 131

[334] Aaron M Goodman, Shumei Kato, Lyudmila Bazhenova, Sandip P Patel, Garrett M Frampton, Vincent Miller, Philip J Stephens, Gregory A Daniels, and Razelle Kurzrock. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol. Cancer Ther.*, 16(11):2598–2608, November 2017. 132

[335] The problem with neoantigen prediction. *Nat. Biotechnol.*, 35(2):97–97, February 2017. 132

[336] Vinod P Balachandran, Australian Pancreatic Cancer Genome Initiative, Marta Łuksza, Julia N Zhao, Vladimir Makarov, John Alec Moral, Romain Remark, Brian Herbst, Gokce Askan, Umesh Bhanot, Yasin Senbabaoglu, Daniel K Wells, Charles Ian Ormsby Cary, Olivera Grbovic-Huezo, Marc Attiyeh, Benjamin Medina, Jennifer Zhang, Jennifer Loo, Joseph Saglimbeni, Mohsen Abu-Akeel, Roberta Zappasodi, Nadeem Riaz, Martin Smoragiewicz, Z Larkin Kelley, Olca Basturk, Mithat Gönen, Arnold J Levine, Peter J Allen, Douglas T Fearon, Miriam Merad, Sacha Gnjatic, Christine A Iacobuzio-Donahue, Jedd D Wolchok, Ronald P DeMatteo, Timothy A Chan, Benjamin D Greenbaum, Taha Merghoub, and Steven D Leach. Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature*, 551(7681):512–516, November 2017. 132

[337] Sanja Stevanović, Anna Pasetto, Sarah R Helman, Jared J Gartner, Todd D Prickett, Bryan Howie, Harlan S Robins, Paul F Robbins, Christopher A Klebanoff, Steven A Rosenberg, and Christian S Hinrichs. Landscape of immunogenic tumor antigens in successful immunotherapy of virally induced epithelial cancer. *Science*, 356(6334):200–205, April 2017. 132

[338] Norbert Hilf, Sabrina Kuttruff-Coqui, Katrin Frenzel, Valesca Bukur, Stefan Stevanović, Cécile Gouttefangeas, Michael Platten, Ghazaleh Tabatabai, Valerie Dutoit, Sjoerd H van der Burg, Per Thor Straten, Francisco Martínez-Ricarte, Berta Ponsati, Hideho Okada, Ulrik Lassen, Arie Admon, Christian H Ottensmeier, Alexander Ulges, Sebastian Kreiter, Andreas von Deimling, Marco Skardelly, Denis Migliorini, Judith R Kroep, Manja Idorn, Jordi Rodon, Jordi Piró, Hans S Poulsen, Bracha Shraibman, Katy McCann, Regina Mendrzyk, Martin Löwer, Monika Stieglbauer,

Cedrik M Britten, David Capper, Marij J P Welters, Juan Sahuquillo, Katharina Kiesel, Evelyna Derhovanessian, Elisa Rusch, Lukas Bunse, Colette Song, Sandra Heesch, Claudia Wagner, Alexandra Kemmer-Brück, Jörg Ludwig, John C Castle, Oliver Schoor, Arbel D Tadmor, Edward Green, Jens Fritsche, Miriam Meyer, Nina Pawlowski, Sonja Dorner, Franziska Hoffgaard, Bernhard Rössler, Dominik Maurer, Toni Weinschenk, Carsten Reinhardt, Christoph Huber, Hans-Georg Rammensee, Harpreet Singh-Jasuja, Ugur Sahin, Pierre-Yves Dietrich, and Wolfgang Wick. Actively personalized vaccination trial for newly diagnosed glioblastoma. *Nature*, 565(7738):240–245, January 2019. 132

[339] Markus W Löffler, Christopher Mohr, Leon Bichmann, Lena Katharina Freudenmann, Mathias Walzer, Christopher M Schroeder, Nico Trautwein, Franz J Hilke, Raphael S Zinser, Lena Mühlenbruch, Daniel J Kowalewski, Heiko Schuster, Marc Sturm, Jakob Matthes, Olaf Riess, Stefan Czemmel, Sven Nahnsen, Ingmar Königsrainer, Karolin Thiel, Silvio Nadalin, Stefan Beckert, Hans Bösmüller, Falko Fend, Ana Velic, Boris Maček, Sebastian P Haen, Luigi Buonaguro, Oliver Kohlbacher, Stefan Stevanović, Alfred Königsrainer, HEPAVAC Consortium, and Hans-Georg Rammensee. Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma. *Genome Med.*, 11(1):28, April 2019. 132, 188

[340] Cory A Brennick, Mariam M George, William L Corwin, Pramod K Srivastava, and Hakimeh Ebrahimi-Nik. Neoepitopes as cancer immunotherapy targets: key challenges and opportunities. *Immunotherapy*, 9(4):361–371, March 2017. 132, 133

[341] Alessandra Cesano and Sarah Warren. Bringing the next generation of immuno-oncology biomarkers to the clinic. *Biomedicines*, 6(1):14, February 2018. 132

[342] Ti-Cheng Chang, Robert A Carter, Yongjin Li, Yuxin Li, Hong Wang, Michael N Edmonson, Xiang Chen, Paula Arnold, Terrence L Geiger, Gang Wu, Junmin Peng, Michael Dyer, James R Downing, Douglas R Green, Paul G Thomas, and Jinghui Zhang. The neoepitope landscape in pediatric cancers. *Genome Med.*, 9(1), December 2017. 132

[343] Gabriel N Teku and Mauno Vihinen. Pan-cancer analysis of neoepitopes. *Sci. Rep.*, 8(1):12735, August 2018. 132, 134

[344] Lena Katharina Freudenmann, Ana Marcu, and Stefan Stevanović. Mapping the tumour human leukocyte antigen (HLA) ligandome by mass spectrometry. *Immunology*, 154(3):331–345, July 2018. 133

[345] David Gfeller and Michal Bassani-Sternberg. Predicting antigen presentation—what could we learn from a million peptides? *Front. Immunol.*, 9, July 2018. 133

[346] Sung-Min Ahn, Se Jin Jang, Ju Hyun Shim, Deokhoon Kim, Seung-Mo Hong, Chang Ohk Sung, Daehyun Baek, Farhan Haq, Adnan Ahmad Ansari, Sun Young Lee, Sung-Min Chun, Seongmin Choi, Hyun-Jeung Choi, Jongkyu Kim, Sukjun Kim, Shin Hwang, Young-Joo Lee, Jong-Eun Lee, Wang-Rim Jung, Hye Yoon Jang, Eunho Yang, Wing-Kin Sung, Nikki P Lee, Mao Mao, Charles Lee, Jessica Zucman-Rossi, Eunsil Yu, Han Chu Lee, and Gu Kong. Genomic portrait of resectable hepatocellular carcinomas: implications of RB1 and FGF19 aberrations for patient stratification. *Hepatology*, 60(6):1972–1982, December 2014. 133

[347] Andreas O Weinzierl, Claudia Lemmel, Oliver Schoor, Margret Müller, Tobias Krüger, Dorothee Wernet, Jörg Hennenlotter, Arnulf Stenzl, Karin Klingel, Hans-Georg Rammensee, and Stefan Stevanovic. Distorted relation between mRNA copy number and corresponding major histocompatibility complex ligand density on the cell surface. *Mol. Cell. Proteomics*, 6(1):102–113, January 2007. 133

[348] Dongxue Wang, Basak Eraslan, Thomas Wieland, Björn Hallström, Thomas Hopf, Daniel Paul Zolg, Jana Zecha, Anna Asplund, Li-Hua Li, Chen Meng, Martin Frejno, Tobias Schmidt, Karsten Schnatbaum, Mathias Wilhelm, Frederik Ponten, Mathias Uhlen, Julien Gagneur, Hannes Hahne, and Bernhard Kuster. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.*, 15(2):e8503, February 2019. 133

[349] Heiko Schuster, Janet K Peper, Hans-Christian Bösmüller, Kevin Röhle, Linus Backert, Tatjana Bilich, Britta Ney, Markus W Löffler, Daniel J Kowalewski, Nico Trautwein, Armin Rabsteyn, Tobias Engler, Sabine Braun, Sebastian P Haen, Juliane S Walz, Barbara Schmid-Horch, Sara Y Brucker, Diethelm Wallwiener, Oliver Kohlbacher, Falko Fend, Hans-Georg Rammensee, Stefan Stevanović, Annette Staebler, and Philipp Wagner. The immunopeptidomic landscape of ovarian carcinomas. *Proc. Natl. Acad. Sci. U. S. A.*, 114(46):E9942–E9951, November 2017. 133

[350] Kevin L Yang, Fengchao Yu, Guo Ci Teo, Kai Li, Vadim Demichev, Markus Ralser, and Alexey I Nesvizhskii. MSBooster: improving peptide identification rates using deep learning-based features. *Nat. Commun.*, 14(1):4539, July 2023. 134

[351] Mathias Wilhelm, Daniel P Zolg, Michael Graber, Siegfried Gessulat, Tobias Schmidt, Karsten Schnatbaum, Celina Schwencke-Westphal, Philipp Seifert, Niklas de Andrade Krätzig, Johannes Zerweck, Tobias Knaute, Eva Bräunlein, Patroklos Samaras, Ludwig Lautenbacher, Susan Klaeger, Holger Wenschuh, Roland Rad, Bernard Delanghe, Andreas Huhmer, Steven A Carr, Karl R Clauser, Angela M Krackhardt, Ulf Reimer, and Bernhard Kuster. Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat. Commun.*, 12(1):3346, June 2021. 134

[352] Shelly Kalaora, Yochai Wolf, Tali Feferman, Eilon Barnea, Erez Greenstein, Dan Reshef, Itay Tirosh, Alexandre Reuben, Sushant Patkar, Ronen Levy, Juliane Quinkhardt, Tana Omokoko, Nouar Qutob, Ofra Golani, Jianhua Zhang, Xizeng Mao, Xingzhi Song, Chantale Bernatchez, Cara Haymaker, Marie-Andrée Forget, Caitlin Creasy, Polina Greenberg, Brett W Carter, Zachary A Cooper, Steven A Rosenberg, Michal Lotem, Ugur Sahin, Guy Shakhar, Eytan Ruppin, Jennifer A Wargo, Nir Friedman, Arie Admon, and Yardena Samuels. Combined analysis of antigen presentation and t-cell recognition reveals restricted immune responses in melanoma. *Cancer Discov.*, 8(11): 1366–1375, November 2018. 134

[353] Shelly Kalaora, Eilon Barnea, Efrat Merhavi-Shoham, Nouar Qutob, Jamie K Teer, Nilly Shimony, Jacob Schachter, Steven A Rosenberg, Michal J Besser, Arie Admon, and Yardena Samuels. Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget*, 7(5):5110–5117, February 2016. 134

[354] Ben Nicholas, Alistair Bailey, Katy J McCann, Oliver Wood, Robert C Walker, Robert Parker, Nicola Ternette, Tim Elliott, Tim J Underwood, Peter Johnson, and Paul Skipp. Identification of neoantigens in oesophageal adenocarcinoma. *Immunology*, 168(3):420–431, March 2023. 134

[355] Jasreet Hundal, Beatriz M Carreno, Allegra A Petti, Gerald P Linette, Obi L Griffith, Elaine R Mardis, and Malachi Griffith. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.*, 8(1):11, January 2016. 134

[356] Jasreet Hundal, Susanna Kiwala, Joshua McMichael, Christopher A Miller, Huiming Xia, Alexander T Wollam, Connor J Liu, Sidi Zhao, Yang-Yang Feng, Aaron P Graubert, Amber Z Wollam, Jonas Neichin, Megan Neveau, Jason Walker, William E Gillanders, Elaine R Mardis, Obi L Griffith, and Malachi Griffith. PVACtools: A computational toolkit to identify and visualize cancer neoantigens. *Cancer Immunol. Res.*, 8(3):409–420, March 2020. 134

[357] Hirokazu Matsushita, Matthew D Vesely, Daniel C Koboldt, Charles G Rickert, Ravindra Uppaluri, Vincent J Magrini, Cora D Arthur, J Michael White, Yee-Shiuan Chen, Lauren K Shea, Jasreet Hundal, Michael C Wendl, Ryan Demeter, Todd Wylie, James P Allison, Mark J Smyth, Lloyd J Old, Elaine R Mardis, and Robert D Schreiber. Cancer exome analysis reveals a t-cell-dependent mechanism of cancer immunoediting. *Nature*, 482(7385):400–404, February 2012. 134

[358] Ana Marcu, Leon Bichmann, Leon Kuchenbecker, Daniel Johannes Kowalewski, Lena Katharina Freudenmann, Linus Backert, Lena Mühlenbruch, András Szolek, Maren Lübke, Philipp Wagner, Tobias Engler, Sabine Matovina, Jian Wang, Mathias Hauri-Hohl, Roland Martin, Konstantina Kapolou, Juliane Sarah Walz, Julia Velz, Holger Moch, Luca Regli, Manuela Silginer, Michael Weller, Markus W Löffler, Florian Erhard, Andreas Schlosser, Oliver Kohlbacher, Stefan Stevanović, Hans-Georg Rammensee, and Marian Christoph Neidert. HLA ligand atlas: a benign reference of HLA-presented peptides to improve t-cell-based cancer immunotherapy. *J. Immunother. Cancer*, 9 (4):e002071, April 2021. 134

[359] Juan Antonio Vizcaíno, Peter Kubiniok, Kevin A Kovalchik, Qing Ma, Jérôme D Duquette, Ian Mongrain, Eric W Deutsch, Bjoern Peters, Alessandro Sette, Isabelle Sirois, and Etienne Caron. The human immunopeptidome project: A roadmap to predict and treat immune diseases. *Mol. Cell. Proteomics*, 19(1):31–49, January 2020. 134

[360] Chengxin Dai, Anja Füllgrabe, Julianus Pfeuffer, Elizaveta M Solovyeva, Jingwen Deng, Pablo Moreno, Selvakumar Kamatchinathan, Deepti Jaiswal Kundu, Nancy George, Silvie Fexova, Björn Grüning, Melanie Christine Föll, Johannes Griss, Marc Vaudel, Enrique Audain, Marie Locard-Paulet, Michael Turewicz, Martin Eisenacher, Julian Uszkoreit, Tim Van Den Bossche, Veit Schwämmle, Henry Webel, Stefan Schulze, David Bouyssié, Savita Jayaram, Vinay Kumar Duggineni, Patroklos Samaras, Mathias Wilhelm, Meena Choi, Mingxun Wang, Oliver Kohlbacher, Alvis Brazma, Irene Papatheodorou, Nuno Bandeira, Eric W Deutsch, Juan Antonio Vizcaíno, Mingze Bai, Timo Sachsenberg, Lev I Levitsky, and Yasset Perez-Riverol. A proteomics sample metadata representation for multiomics integration and big data analysis. *Nat. Commun.*, 12(1): 5854, October 2021. 136

[361] Stephan Aiche, Timo Sachsenberg, Erhan Kenar, Mathias Walzer, Bernd Wiswedel, Theresa Kristl, Matthew Boyles, Albert Duschl, Christian G Huber, Michael R Berthold, Knut Reinert, and Oliver Kohlbacher. Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry. *Proteomics*, 15(8):1443–1447, April 2015. 187

[362] Benjamin Schubert, Luis de la Garza, Christopher Mohr, Mathias Walzer, and Oliver Kohlbacher. ImmunoNodes - graphical development of complex immunoinformatics workflows. *BMC Bioinformatics*, 18(1):242, May 2017. doi: 10.1186/s12859-017-1667-z. 187

[363] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, Ingmarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, and Fredrik Pontén. Proteomics. tissue-based map of the human proteome. *Science*, 347(6220):1260419, January 2015. 189

[364] Gilbert S Omenn, Lydie Lane, Christopher M Overall, Young-Ki Paik, Ileana M Cristea, Fernando J Corrales, Cecilia Lindskog, Susan Weintraub, Michael H A Roehrl, Siqi Liu, Nuno Bandeira, Sudhir Srivastava, Yu-Ju Chen, Ruedi Aebersold, Robert L Moritz, and Eric W Deutsch. Progress identifying and analyzing the human proteome: 2021 metrics from the HUPO human proteome project. *J. Proteome Res.*, 20(12):5227–5240, December 2021.

[365] Pavel Sinitcyn, Alicia L Richards, Robert J Weatheritt, Dain R Brademan, Harald Marx, Evgenia Shishkova, Jesse G Meyer, Alexander S Hebert, Michael S Westphall, Benjamin J Blencowe, Jürgen Cox, and Joshua J Coon. Global detection of human variants and isoforms by deep proteome sequencing. *Nat. Biotechnol.*, 41(12):1776–1786, December 2023. 189

[366] Rhiannon Morris, Katrina A Black, and Elliott J Stollar. Uncovering protein function: from classification to complexes. *Essays Biochem.*, 66(3):255–285, August 2022. 190

[367] Junichi Ono, Yoshihiro Matsumura, Toshifumi Mori, and Shinji Saito. Conformational dynamics in proteins: Entangled slow fluctuations and nonequilibrium reaction events. *J. Phys. Chem. B*, 128(1):20–32, January 2024. 190

[368] Athit Kao, Chi-Li Chiu, Danielle Vellucci, Yingying Yang, Vishal R Patel, Shenheng Guan, Arlo Randall, Pierre Baldi, Scott D Rychnovsky, and Lan Huang. Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol. Cell. Proteomics*, 10(1):M110.002212, January 2011. 190

[369] Christoph Hage, Claudio Iacobucci, Anne Rehkamp, Christian Arlt, and Andrea Sinz. The first zero-length mass spectrometry-cleavable cross-linker for protein structure analysis. *Angew. Chem. Weinheim Bergstr. Ger.*, 129(46):14743–14747, November 2017. 190

[370] Sandro Holzer, Gianluca Degliesposti, Mairi L Kilkenny, Sarah L Maslen, Dijana Matak-Vinkovíc, Mark Skehel, and Luca Pellegrini. Crystal structure of the n-terminal domain of human timeless and its interaction with tipin. *Nucleic Acids Res.*, 45(9):5555–5563, May 2017. 191

[371] Konstantin E Komolov, Yang Du, Nguyen Minh Duc, Robin M Betz, João P G L M Rodrigues, Ryan D Leib, Dhabaleswar Patra, Georgios Skiniotis, Christopher M Adams, Ron O Dror, Ka Young Chung, Brian K Kobilka, and Jeffrey L Benovic. Structural and functional analysis of a β2-adrenergic receptor complex with GRK5. *Cell*, 169(3):407–421.e16, April 2017.

[372] Javier Fernandez-Martinez, Seung Joong Kim, Yi Shi, Paula Upla, Riccardo Pellarin, Michael Gagnon, Ilan E Chemmama, Junjie Wang, Ilona Nudelman, Wenzhu Zhang, Rosemary Williams, William J Rice, David L Stokes, Daniel Zenklusen, Brian T Chait, Andrej Sali, and Michael P Rout. Structure and function of the nuclear pore complex cytoplasmic mRNA export platform. *Cell*, 167(5):1215–1228.e25, November 2016.

[373] C Plaschka, L Larivière, L Wenzeck, M Seizl, M Hemann, D Tegunov, E V Petrotchenko, C H Borchers, W Baumeister, F Herzog, E Villa, and P Cramer. Architecture of the RNA polymerase II-Mediator core initiation complex. *Nature*, 518(7539):376–380, February 2015.

[374] Joost Snijder, Jan M Schuller, Anika Wiegard, Philip Lössl, Nicolas Schmelling, Ilka M Axmann, Jürgen M Plitzko, Friedrich Förster, and Albert J R Heck. Structures of the cyanobacterial circadian oscillator frozen in a fully assembled state. *Science*, 355(6330):1181–1184, March 2017. 191

[375] Meena Choi, Ching-Yun Chang, Timothy Clough, Daniel Broudy, Trevor Killeen, Brendan MacLean, and Olga Vitek. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, 30(17):2524–2526, September 2014. 194

Abbreviations

API Application Programming Interface: a set of specifications that a software can follow to access the services and resources provided by another software that implements the API. 86

BASH Bourne-Again Shell: a popular command-line interface for the operating system. 100

CID Collision-Induced Dissociation: a mass spectrometry technology to fragment primarily peptide ions. 26, 28, 125

CTD Common Tool Descriptors. 37, 102–104

CV Controlled Vocabulary. 46–49, 54, 56–58, 60, 66–68, 70, 73, 78–81, 85, 86, 136, 194

DDA Data-Dependent Acquisition. 27, 53, 125, 131

DIA Data-Independent Acquisition. 27

DNA Deoxyribonucleic Acid: a polynucleotide polymer carrying genetic information encoded in its seuqence of nucleotides. 6, 7, 9, 14, 15, 124, 129

ESI Electrospray Ionisation: a soft ionisation technique used for the mass spectrometry of peptides. 21, 25, 77, 110, 118

ETD Electron-Transfer Dissociation: a mass spectrometry technology to fragment primarily peptide ions. 26, 28

FDR False Discovery Rate: a method of conceptualising the rate of type I errors in null hypothesis testing. 33–35, 40, 74, 107, 109, 123, 126, 130

FFT Fast Fourier Transform : an algorithm that converts a signal from its original domain (often time or space or mass) to a representation in the frequency domain and vice versa. 22, 23

FRED Framework for Epitope Detection: a modular python framework for immunoinformatics development. 3, 36, 37, 88, 90–97, 102, 104, 105, 108, 118, 124, 134, 136, 137

FTP File Transfer Protocol: a standard communication protocol used for the transfer of computer files from a server to a client. 109

FWER Family-Wise Error Rate: represents the probability of making one or more false discoveries, or type I errors when performing multiple hypotheses tests. 33

FWHM Full Width At Half Maximum: a concept to describe resolution in mass spectrometry or the properties of a peak-map feature created from a eluting analyte in LC-MS. 23, 24

GKN Generic Knime Node. 105, 118

GUI Graphical User Interface: a form of user interface for visual interaction. 35, 106

HCC Hepatocellular Carcinoma: a type of primary liver cancer. 115, 116, 122–131, 133, 134, 137, 197, 198

HCD Higher-Energy Collisional Dissociation: a CID technique. 26

HLA Human Leukocyte Antigens: membrane bound proteins with peptide-presenting function as part of the immune system. 2, 9–19, 25–27, 31, 36, 88–97, 105, 106, 109, 110, 113, 115–117, 119, 122–128, 130–133, 189, 190, 197, 198

HPC High-Performance Computing: a computer cluster to solve advanced computation problems. 100, 101

HUPO Human Proteome Organization: an international scientific organization that promotes proteomics research and applications. 30, 41, 47, 51, 71, 79

I/O Input/Output: the input and output of a software program. 35

ICGC International Cancer Genome Consortium : a global collaboration to identify the genetic faults in 50 types of cancer. 132, 134

ICP Immune checkpoint. 122, 123, 132

INDEL Insertion-Deletion: term for an insertion or deletion of nucleotides in a genomic sequence. 8, 9

KNIME Konstanz Information Miner: a workflow management software. 36, 37, 101–110, 117, 118, 137

LC Liquid Chromatography: a technique for the separation of complex samples. 18, 19, 21–23, 25, 77, 110–113, 115, 116, 118, 119

LC-MS Liquid Chromatography Mass Spectrometry: LC coupled to a MS. 20, 52, 55–57

LC-MS/MS Liquid Chromatography Coupled Tandem Mass Spectrometry. 27, 49, 62, 68, 79, 80, 84, 107, 110, 114, 118, 125, 130, 133

LIMS Laboratory Information Management System. 85, 86

LOWESS Locally Weighted Scatterplot Smoothing: a local polynomial regression for scatterplot smoothing. 111, 112

LSF Load Sharing Facility. 101

m/z mass-to-charge ratio. 23, 24, 26, 27, 54, 195

Mel Malignant Melanoma. 122, 123, 126, 129, 131, 133, 134

MHC Major Histocompatibility Complex: a highly polymorphic region on chromosome 6 with genes particularly involved in immune functions. 6, 11, 12, 16

MS Mass Spectrometry. 2–5, 8, 18–22, 24, 25, 28–30, 32, 34, 35, 40–43, 49–53, 58, 61–63, 66, 72, 74, 77, 78, 82, 101, 105–108, 110, 111, 113, 116, 117, 119, 122, 123, 125, 126, 129, 136, 137, 195

MS1 Survey Scan: MS spectrum from scans of the whole (preset) mass range.. 25–27, 31

MS2 Tandem Mass Spectrum. 26, 29, 31, 77

NGS Next-Generation Sequencing. 1, 7, 15, 19, 89, 123

OpenMS Open Mass Spectrometry: a software framework for computational mass spectrometry. 3, 34–37, 41, 42, 47–50, 61, 62, 72, 78, 81, 87, 88, 97, 101–110, 118, 123, 134, 135, 137, 195

PCA Principal Component Analysis: a linear dimensionality reduction technique. 115, 116

PCR Polymerase Chain Reaction. 8, 15

PNE Predicted Neoepitopes: mutation-derived predicted neoepitopes. 127–134

PPI Protein–Protein Interaction. 9, 190

PRIDE Proteomics Identifications Database: the worlds largest repository for proteomics mass spectrometry data. 40, 74, 109, 126, 135

PSI Proteomics Standards Initiative: a working group of the Human Proteome Organization to define data standards for proteomics to facilitate data comparison, exchange and verification. 30, 41, 44, 47, 49–51, 63, 67, 71, 73, 74, 78, 79, 81, 136

PSM Peptide-Spectrum Match: a peptide sequence matched to a tandem mass spectrum, indicating a possibly identified peptide. 32–34, 43, 45, 46, 68–73, 82, 84, 111, 115, 116, 123, 134

PTM Post-Translational Modification: a covalent modification of proteins following protein biosynthesis, usually enzymatic. 9, 11

QC Quality Control. 77, 78, 81, 82, 105, 107, 110, 111, 113, 114, 116–120, 136

RNA Ribonucleic Acid. 6–8, 89, 96, 124, 125, 127, 128, 133

RT Retention Time. 19, 25, 29, 30, 108, 110–112, 116, 119

S/N Signal-to-Noise Ratio. 25

SILAC Stable Isotope Labeling By Amino Acids In Cell Culture. 29

SNP Single Nucleotide Polymorphism. 9

SNV Single Nucleotide Variation. 9, 94

SVM Support-Vector Machine. 116

TAA Tumour Associated Antigen. 16, 134

TAP Transporter Associated With Antigen Processing. 11, 105

TCGA The Cancer Genome Atlas: a catalogue of genomic alterations in cancer. 131, 132, 134

TCR T-Cell Receptor. 10–13

TIC Total Ion Current. 110, 118

TMB Tumour Mutational Burden: the number of somatic mutations per megabase of interrogated genomic sequence. 127, 132, 133

TOPP The Openms Proteomics Pipeline. 35, 36, 41, 48, 61, 62, 72, 101–104, 106, 118, 123

TOPPAS TOPP Assistant: a GUI to create, edit, and run TOPP workflows. 101–103, 106, 118

VCF Variant Call Format: a text format for storing gene sequence variations. 9, 91

XML Extensible Markup Language: a markup language and file format for storing, transmitting, and reconstructing data. 41, 44, 45, 47, 58, 63, 67, 73, 78, 79, 81, 84, 85, 102, 136

XSLT Extensible Stylesheet Language Transformations: a language designed for transforming XML documents. 79, 82, 84–86

Appendix A: Permissions and Contributions

The mzIdentML Data Standard Version 1.2, Supporting Advances in Proteome Informatics [170]

JAV, EWD, and ARJ designed research; JAV, GM, SRP, HB, MV, YP, TT, JU, ME, LF, JR, EN, **MW**, OK, AL, RJC, FG, SM, EWD, and ARJ performed research; JAV, GM, SRP, HB, MV, YP, TT, JU, ME, LF, JR, EN, **MW**, OK, AL, FG, SM, EWD, and ARJ contributed new reagents or analytic tools; JAV, EWD, and ARJ wrote the paper.

The mzQuantML Data Standard for Mass Spectrometry–based Quantitative Studies in Proteomics [166]

ARJ, OK, JAV, and **MW** designed research. **MW**, DQ, GM, JU, ME, TS, FFGG, JF, CB, EWD, FR, JAV, JAMA, JPA, OK, ARJ performed research. **MW**, DQ, and ARJ wrote the paper.

The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience.[257]

Permission to reuse text, figures, and charts was granted by the American Society for Biochemistry and Molecular Biology. Creative Commons Attribution (CC BY 4.0).
"This research was originally published in Molecular & Cellular Proteomics, Volume 13, Issue 10, 2765 - 2775; J. Griss, A.R. Jones, T. Sachsenberg, M. Walzer, L. Gatto, J. Hartler, G.G. Thallinger, R.M. Salek, C. Steinbeck, N. Neuhauser, J. Cox, S. Neumann, J. Fan, F. Reisinger, Q.W. Xu, N. del Toro, Y. Pérez-Riverol, F. Ghali, N. Bandeira, I. Xenarios, O. Kohlbacher, J.A. Vizcaíno, H. Hermjakob; The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. © the American Society for Biochemistry and Molecular Biology."
ARJ, OK, JAV, and HH designed research. JG, ARJ, TS, **MW**, LG, JH, GT, RMS, CS, NN, JC, SN, JF, FR, QX, NDT, YP, FG, NB, IX, and JAV performed research. JG, JAV, and HH wrote the paper.

qcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments[203]

Permission to reuse text, figures, and charts was granted by the American Society for Biochemistry and Molecular Biology. Creative Commons Attribution (CC BY 4.0).
"This research was originally published in Molecular & Cellular Proteomics, Volume 13, Issue 8, 1905 - 1913; M. Walzer, L.E. Pernas, S. Nasso, W. Bittremieux, S. Nahnsen, P. Kelchtermans, P. Pichler, H.W.P. van den Toor, A. Staes, J. Vandenbussche, M. Mazanek, T. Taus, R.A. Scheltema, C.D. Kelstrup, L. Gatto, B. van Breukelen, S. Aiche, D. Valkenborg, K. Laukens, K.S. Lilley, J.V. Olsen, A.J.R. Heck, K. Mechtler, R. Aebersold, K. Gevaert, J.A. Vizcaíno, H. Hermjakob, O. Kohlbacher, L. Martens; qcML: An Exchange Format for Quality Control Metrics from Mass Spectrometry Experiments. © the American Society for Biochemistry and Molecular Biology."
MW, LEP, SNas, SNah, PK, JAV, OK, and LM designed research; **MW**, LEP, SNah, PK, PP, HWv, AnS, JV, TTa, CDK, LG, BvB, KL, JAV, and LM performed research; **MW**, LEP, SNas, WB, SNah, PK, PP, AnS, JV, MM, TT, RAS, LG, SA, DV, JAV, and LM analyzed data; **MW**, LEP, SNas, WB, PK, PP, HWv, KL, KSL, JVO, AJH, KM, RA, KG, JAV, HH, OK, and LM wrote the paper.

FRED 2: an immunoinformatics framework for Python[319]

BS, **MW**, HPB, AS, CM, OK designed research; BS, **MW**, HPB, AS, CM performed research; BS and OK wrote the paper.

Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry[361]

CGH, MRB, KR, OK designed research; SA, TS, EK, **MW**, BW, TK, MB, AD performed research; SA and OK wrote the paper.

ImmunoNodes – graphical development of complex immunoinformatics workflows.[362]

OK designed research; BS, LG developed and implemented the method; CM implemented the distance-to-self nodes. **MW** contributed the ligandomics workflow. BS, LG, and OK wrote the paper.

Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma[339]

"This research was originally published in Genome Med 11, 28; M.W. Löffler, C. Mohr, L. Bichmann, L.K. Freudenmann, M. Walzer, C.M. Schroeder, N. Trautwein, F.J. Hilke, R.S. Zinser, L. Mühlenbruch, D.J. Kowalewski, H. Schuster, M. Sturm, J. Matthes, O. Riess, S. Czemmel, S. Nahnsen, I. Königsrainer, K. Thiel, S. Nadalin, S. Beckert, H. Bösmüller, F. Fend, A. Velic, B. Maček, S.P. Haen, L. Buonaguro, O. Kohlbacher, S. Stevanović, A. Königsrainer, HEPAVAC Consortium & H.G. Rammensee ; Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma. © the Authors"

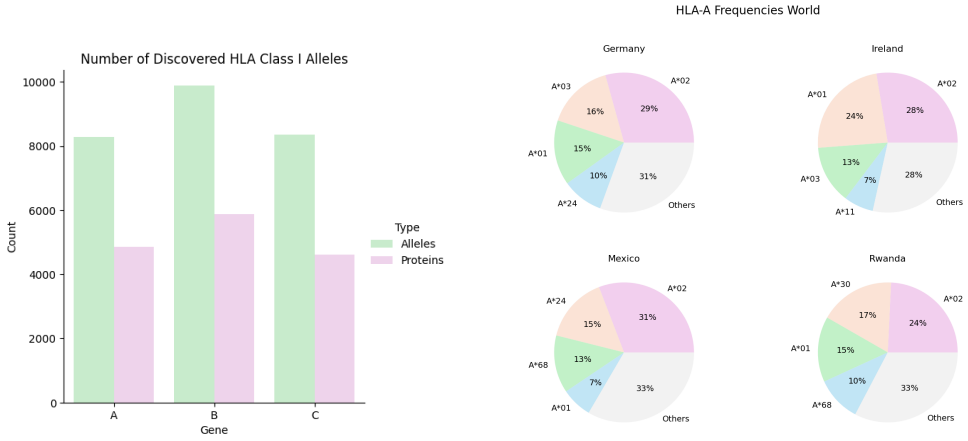
MWL, OR, BM, LBu, OK, SS, HEPAVAC and HGR are responsible for the study concept and design. MWL, LKF, NT, FJH, RSZ, LMue, DJK, HS, HB, and AV are responsible for the acquisition of data. MWL, CM, LBi, LKF, **MW**, CMS, FJH, RSZ, LMue, DJK, HS, MS, JM, SC, SNah, IK, KT, SNad, SB, FF, AV, BM, SPH, and HGR are responsible for the analysis and/or interpretation of data. MWL, CM, LBi, and LKF are responsible for the drafting of the manuscript. **MW**, CMS, NT, FJH, RSZ, LMue, DJK, HS, MS, JM, OR, SC, SNah, IK, KT, SNad, SB, HB, FF, AV, BM, SPH, LBu, OK, SS, AK, and HGR are responsible for the critical revision of the manuscript for important intellectual content. MWL, SNah, OK, AK, HGR, and HEPAVAC obtained funding. MWL, **MW**, CMS, NT, FJH, RSZ, LMue, DJK, HS, MS, JM, SC, KT, SNad, SB, HB, AV, SPH, LBu, and HEPAVAC are responsible for the administrative, technical, or material support. SNah, OR, FF, BM, OK, SS, AK, and HGR are responsible for the study supervision.

AD: Albert Duschl, AJH: Albert J.R. Heck, AK: Alfred Königsrainer, AL: Alexander Leitner, ARJ: Andrew Robert Jones, AS: András Szolek, AV: Ana Velic, AnS: An Staes, BM: Boris Maček, BS: Benjamin Schubert, BW: Bernd Wiswedel, BvB: Bas van Breukelen, CB: Conrad Bessant, CDK: Christian D. Kelstrup, CGH: Christian G. Huber, CM: Christopher Mohr, CMS: Christopher M. Schroeder, CS: Christoph Steinbeck, DJK: Daniel J. Kowalewski, DQ: Da Qi, DV: Dirk Valkenborg, EK: Erhan Kenar, EN: Eugen Netz, EWD: Eric W. Deutsch, FF: Falko Fend, FFGG: Faviel F. Gonzalez-Galarza, FG: Fawaz Ghali, FJH: Franz J. Hilke, FR: Florian Reisinger, GM: Gerhard Mayer, GT: Gerhard Thallinger, HB: Hans Bösmüller, HGR: Hans-Georg Rammensee, HH: Henning Hermjakob, HPB: Hans-Philipp Brachvogel, HS: Heiko Schuster, HWv: Henk W.P. van den Toorn, IK: Ingmar Königsrainer, IX: Ioannis Xenarios, JAMA: J. Alberto Medina-Aunon, JAV: Juan Antonio Vizcaíno, JC: Jürgen Cox, JF: Jun Fan, JG: Johannes Griss, JH: Jürgen Hartler, JM: Jakob Matthes, JP: Julianus Pfeuffer, JPA: Juan Pablo Albar, JR: Juri Rappsilber, JU: Julian Uszkoreit, JV: Jonathan Vandenbussche, JVO:J esper V. Olsen, KG: Kris Gevaert, KL: Kris Laukens, KM: Karl Mechtler, KR: Knut Reinert, KSL: Kathryn S. Lilley, KT: Karolin Thiel, LBi: Leon Bichmann, LBu: Luigi Buonaguro, LDG: Luis de la Garza, LEP: Lucia Espona Pernas, LF: Lutz Fischer, LG: Laurent Gatto, LKF: Lena Katharina Freudenmann, LM: Lennart Martens, LMue: Lena Mühlenbruch, MB: Matthew Boyles, ME: Martin Eisenacher, MM: Michael Mazanek, MRB: Michael R. Berthold, MS: Marc Sturm, MV: Marc Vaudel, **MW: Mathias Walzer**, MWL: Markus W. Löffler, NB: Nuno Bandeira, NDT: Noemi del Toro, NJ: Nico Jehmlich, NN: Nadin Neuhauser, NT: Nico Trautwein, OK: Oliver Kohlbacher, OR: Olaf Riess, PK: Pieter Kelchtermans, PP: Peter Pichler, QX: Qing-Wei Xu, RA: Ruedi Aebersold, RAS: Richard A. Scheltema, RJC: Robert J. Chalkley, RMS: Reza M. Salek, RS: Reza Salek, RSZ: Raphael S. Zinser, SA: Stephan Aiche, SB: Stefan Beckert, SC: Stefan Czemmel, SM: Salvador Martínez-Bartolomé, SN: Steffen Neumann, SNad: Silvio Nadalin, SNah: Sven Nahnsen, SNas:S ara Nasso, SPH: Sebastian P. Haen, SRP: Simon Perkins, SS: Stefan Stevanović, TK: Theresa Kristl, TS: Timo Sachsenberg, TT: Tobias Ternent, TTa: Thomas Taus, WB:Wout Bittremieux, YP: Yasset Pérez-Riverol

Appendix B: Background

HLA Polymorphy and Worldwide Population Diversity

Certain Alleles dominate in certain populations[49–51]. This is demonstrated for HLA-A in Fig. B.1 b). The four most frequent HLA-A alleles in four populations[i] (Germany, Ireland, Mexico City, Rwanda) are compared in pie charts, where HLA-A*02 is the only frequent commonality.



(a) **(b)**

Figure B.1: Examples for HLA polymorphy and worldwide population diversity. a) shows the number of different HLA class I alleles (left bars) and distinct HLA class I proteins (right bars) for HLA-alleles A, B, and C discovered in the worldwide population. b) shows the percentual share of the four most frequent HLA-A alleles in four populations.

Fig. B.1 a) shows the number of different HLA class I alleles (left bars) and distinct HLA class I proteins (right bars) for HLA-alleles A, B, and C discovered in the worldwide population[ii]. Polymorphy, Mendelian inheritance, and codominant expression result in a huge diversity of HLA types found in the worldwide population. However, certain Alleles dominate in certain populations[49–51].

Crosslinking

Mass spectrometry-based protein identification has been focused on the amino acid sequence and their detection in samples of tissue. We now have a basic if however still incomplete understanding of protein occurrence in different tissues[103,363–365]. Yet, knowledge of occurrence is not equal to understanding the function of proteins. Protein function is in large parts conducted

[i] Data taken from `https://www.allelefrequencies.net`, Germany (n=11407), Ireland South Pop 2 (n=17624), Mexico Mexico City Center (n=152), Rwanda (n=280), accessed 01.07.2024

[ii] `https://www.ebi.ac.uk/ipd/imgt/hla/about/statistics/`, accessed 01.07.2024

by interacting with other proteins. Natural proteins have tertiary and quaternary structure and lose their function when denatured. This implies their structure as part of their function, with parts of their amino acid sequence internal, others surface exposed, some with restricted access by crevices formed from other sections of their sequence as seen with the HLA molecules. Other protein structure investigation techniques such as electron microscopy (EM), NMR and crystallography have sample preparation requirements and necessary purification steps that keep them from observing PPI in situ. This is a complicating factor for function investigation, as proteins are known to undergo conformation changes in response to their environment and to activate or deactivate their function[366,367]. Protein crosslinking provides methods to investigate PPI in an unbiased fashion in complex protein mixtures. This is accomplished by the use of bifunctional cross-linking molecules that link two amino acids in close proximity with a covalent bond. Cross-linkers are molecules of reactive end groups divided by a spacer. The reactive groups targeting specific amino acids simplify data analysis. The spacer defines the spatial resolution as longer spacers have longer reach, so have more combinations available to connect residues. The limit is however only an upper-bound distance restraint, increasing data analysis complexity. After proteolytic digestion of the protein-protein complexes, four distinct peptide products are formed:

1. Single peptides not captured by the cross-linking reagent, therefore yielding no structural information

2. Monolinks, single peptides captured by only one side of the cross-linking reagent. They do however give evidence that the identified section of the protein was surface-exposed

3. Loop links, single peptides captured with both sides of the cross-linking agent. They, too, give evidence of surface-exposed protein sections.

4. Two peptides captured by the cross-linking reagent, providing distance information about protein tertiary (peptides from the same protein) or quaternary (peptides from different proteins) structure.

Cross-linked peptides are large and branched, compromising the fragmentation efficiency and resulting in complicated fragmentation spectra. The cross-linker also increases the number of potential sequence matches to be considered through the combinations of different peptides plus cross-linker matching the precursor mass. The quadratic inflation of the search space also makes false positive control more challenging. MS-cleavable cross-linkers like DSSO[368] and CDI[369] allow simpler mass spectrometry-based identification, reducing the problem to regular peptide-plus-modification to spectrum matching. The trade-off is the loss of direct linkage partner information, requiring post-identification analysis. It is therefore essential to provide and document more crosslinking details than with regular mass spectrometry-based identifications

for successful analysis result presentation, sharing, and integration. With a sufficient amount of cross-links identified, it is possible to detect and define protein interfaces. The results can even be leveraged to produce final protein complex models[370–374]. A combination with crystal structure and cryo-EM data promises to provide a better picture of spatial protein relations and a better understanding of protein functions.

Appendix C: Formats

Programmatic Accession of mzQuantML

The simple extraction of values from mzQuantML for further use in statistical software is another benefit of the mzQuantML design.

Listing C.1: A brief Python script can extract a target layer <DataMatrix> values with the help of XPATH and the mzQuantML schema.

```
from lxml.etree import parse
import pandas as pd
# load a mzQuantML file
q = parse('CPTAC-Progenesis-small-example.mzq')
ns={'n':'http://psidev.info/psi/pi/mzQuantML/1.0.0'}
# fetch the AssayQuantLayer from the PeptideConsensusList
# containing peptide normalised abundance (MS:1001891)
xml_aql = next(iter(q.xpath("//n:MzQuantML/n:PeptideConsensusList/
    ↪ n:AssayQuantLayer/n:DataType/n:cvParam[@accession='MS
    ↪ :1001891']/../..",namespaces=ns)), None)
if xml_aql is not None:
    # extract the normalised peptide abundance values into a proto
        ↪ -table
    peptide_quant_table = []
    for row in xml_aql.xpath('.//n:DataMatrix/n:Row',namespaces=ns
        ↪ ):
        peptide_quant_table.append([row.attrib['object_ref']] +
            ↪ row.text.split('␣'))
    # extract the table header
    column_header = next(iter(xml_aql.xpath('.//n:ColumnIndex',
        ↪ namespaces=ns)), None)
    if column_header is not None:
        peptide_quant_table_header = [['peptide'] + column_header.
            ↪ text.split('␣')]
    # assemble DataFrame and write tsv
    if column_header is not None and peptide_quant_table is not
        ↪ None:
        df = pd.DataFrame(peptide_quant_table,columns=
            ↪ peptide_quant_table_header)
        df.to_csv('normalised_peptide_quant.tsv', sep='\t', index=
            ↪ False)
else:
    print("QuantLayer␣not␣found")
```

The script can extract the abundance values stored in the <DataMatrix> of a target layer in an arbitrary mzQuantML file for direct use in statistical software given the layer is present. In the example script, the target layer is the <AssayQuantLayer> of <PeptideConsensusList>, i.e., the peptide abundances of confidently identified peptides in all assays included in the study. The correct layer is selected through the combination of the XPATH of the layer and its datatype accession (here: normalised peptide abundance, MS:1001891). The script will extract the <DataMatrix> values of the fitting layer into a pandas 'DataFrame' and subsequent '.tsv' file, ready for direct use in downstream analysis software such as MSstats[375]. Each assay is represented in a separate column, and each row is indexed with the respective peptide represented. The peptide is represented in the script through the @object_ref pointing each to a <PeptideConsensus> element. The @object_ref, by convention rather than norm, is usually encoded with the peptide sequence, charge, and running index (e.g., id="pep_AETDDGADVIR_2_10786"). Should a file be produced without informative @object_ref, the <PeptideSequence> and <EvidenceRef> can be extracted in similar brevity directly from the same <PeptideConsensusList> to be put as row index (not shown).

Updating mzIdentML to Report Cross-linking Experiment Spectra Identifications

Another design update to mzIdentML was the addition of a new mechanism to encode the identification of cross-linked peptides. With the present cross-linking techniques, one spectrum may identify more than one peptide, which are either the cross-linked peptides or fragments of the peptides and crosslinker. In case these values are reported, the protocol section needs to include the CV term "cross-linking search" (MS:1002494) as a use-case label.

The structure of the previous mzIdentML standard definitions restricted the reference of one spectrum identification to one peptide. This was solved with the formation of peptide pairs, annotating the respective <Peptide> elements with the cross-link modification and a pair-specific identifier. This is achieved with a CV term in each peptide's <Modification> element, either "cross-link donor" (MS:1002509) or "cross-link acceptor" (MS:1002510), the value of which for both must be a document-unique identifier for the peptide pair. The cross-link introduced mass difference will be represented by the 'donor' peptide's <Modification> element (@monoisotopicMassDelta attribute), whereas the 'acceptor' peptide's <Modification> element will have no nominal mass difference. The donor/acceptor nomenclature is arbitrary, the donor is by convention however the longer peptide.

A <SpectrumIdentificationItem> then references each peptide via the respective <PeptideEvidence> elements, through which they will be implicitly paired, but also need to be linked explicitly with a CV term "cross-link spectrum identification item" (MS:1002511) that will link the items by its shared @value attribute. Just as with the peptides' "cross-link donor" and "cross-link acceptor", the values need to be file-unique identifiers. The identification scores,

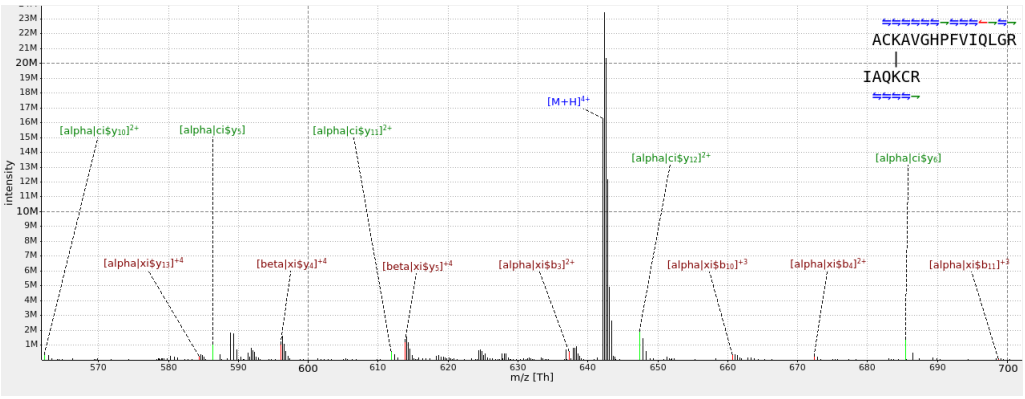


Figure C.1: Visualisation of cross-linked spectrum with annotated matched peaks and peptide sequence coverage loaded from a mzIdentML (v1.2) file. Visualisation with `TOPPView`, identified with OpenPepXL. The integration of cross-linking mzIdentML in OpenMS allows visualising identified cross-linked spectra with the OpenMS spectrum visualisation tool `TOPPView`. Data shown from MS run `C_Lee_141014_CRM_dialysis_NCE20_1` (mzML and mzIdentML), spectrum index 8821, 560-700 m/z of PXD014359. Labels show identified fragment peaks of the peptide chains. Upper-right corner shows the sequence coverage indicator. A one-sided arrow means the fragment starting at the marked residue and containing the rest of the peptide or peptide pair in the direction of the arrow was matched. A double arrow means fragments starting at the marked residue and containing the rest of the peptide or peptide pair in both directions were matched.

rank, and charge reported within the pair must be the same. The `@experimentalMassToCharge` and `@calculatedMassToCharge` attribute values of the `<SpectrumIdentificationItem>` elements must represent the mass values of both peptides and the cross-linker.

The implementation of the cross-linking use case, available in OpenMS (v2.5), can be utilised to write the results from OpenPepXL[172] in mzIdentML or to read cross-linking identification results for visualisation in `TOPPView` (Fig. C.1).

Appendix D: Automated Workflows for Quality Control and Immunopeptidomics

Outlier Detection from QC Metrics

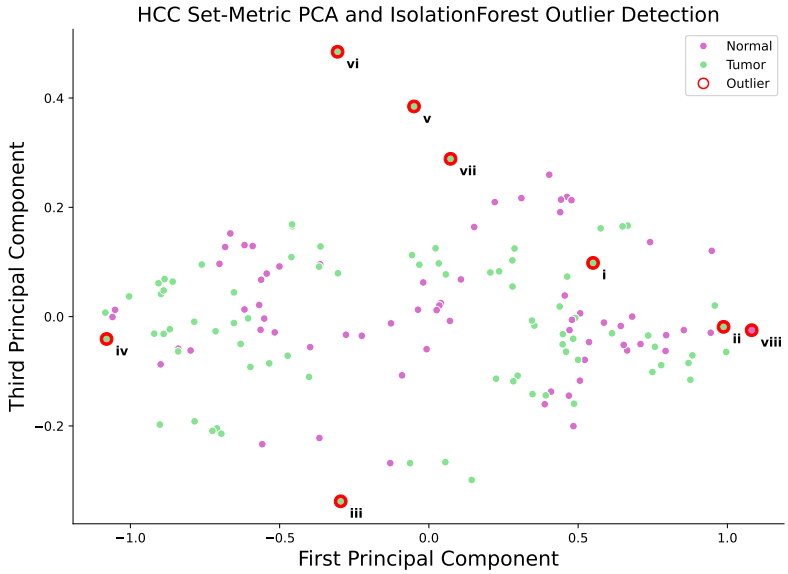


Figure D.1: Multivariate isolation forest outlier detection applied to the set of HLA-ligand detection runs from HCC tissue of the individualised immunotherapy study (Chapter 7). Visualised is the 2D combinations of principal component 1 and 3 to complement the visualisation in the main text.

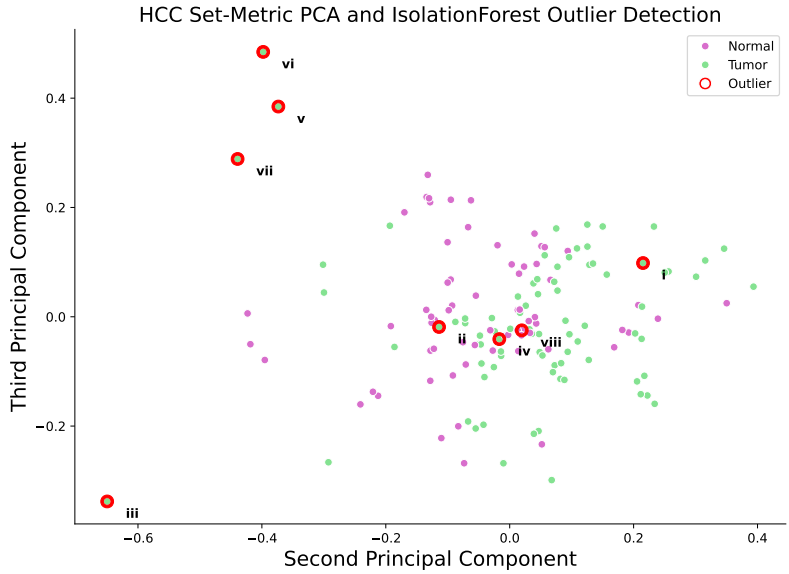


Figure D.2: Multivariate isolation forest outlier detection applied to the set of HLA-ligand detection runs from HCC tissue of the individualised immunotherapy study (Chapter 7). Visualised is the 2D combinations of principal component 2 and 3 to complement the visualisation in the main text.