Aus dem Adolf-Butenandt-Institut
Lehrstuhl Molekularbiologie
Inistitut der Ludwig-Maximilians-Universität München
Direktor: Prof. Dr. Peter B. Becker

Data Analysis for Genomics, Transcriptomics and Proteomics

Dissertation zum Erwerb des Doktorgrades der
Naturwissenschaften (Dr. rer. nat.) an der
Medizinischen Fakultät der
Ludwig-Maximilians-Universität München

vorgelegt von

Bo Sun

aus Henan, China

2023

Mit Genehmigung der Medizinischen Fakultät

der Universität München

Betreuer:	Prof. Dr. Axel Imhof
Zweitgutachte:	Prof. Dr. Kristian Unger
Dekan:	Prof. Dr. med. Thomas Gudermann

Tag der mündlichen Prüfung: 09. November 2023

Eidesstattliche Versicherung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Thema "Data Analysis for Genomics, Transcriptomics and Proteomics"
selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

Ort, Datum: Munich, 04.01.2023 Unterschrift: Bo Sun

Table of content

Abbreviations

MS: mass spectrometry

DDA: data dependent acquisition

DIA: data-independent acquisition

TOF-MS: time-of-flight mass spectrometry

SRM: selected reaction monitoring

PRM: parallel reaction monitoring

MC: missed cleavages

ROC: receiver operating characteristic

AUC: area under the curve

PSM: peptide spectrum matches

LSTM: long short-term memory

TCN: temporal convolutional network

CNN: convoulutional neural network

FDR: false discovery rate

PPV: positive predictive value

MCC: Matthews correlation coefficient

Publications

I. Thomas Andreas Gerland, **Bo Sun**, Pawel Smialowski, Andrea Lukacs, Andreas Walter Thomae, and Axel Imhof. 2017. "The Drosophila Speciation Factor HMR Localizes to Genomic Insulator Sites." *PLOS ONE* 12 (2): e0171798. doi:10.1371/journal.pone.0171798.

II. Bo Sun, Pawel Smialowski, Tobias Straub, and Axel Imhof. 2021. "Investigation and Highly Accurate Prediction of Missed Tryptic Cleavages by Deep Learning." *Journal of Proteome Research*. doi:10.1021/acs.jproteome.1c00346.

III. Bo Sun, Pawel Smialowski, Wasim Aftab, Andreas Schmidt, Ignasi Forne, Tobias Straub, and Axel Imhof. 2022. "Improving SWATH-MS analysis by Deep Learning." *Proteomics*, 2022, doi: 10.1002/pmic.202200179.

1. Contributions

This dissertation is presented in a cumulative way, which reflects the major achievements of my doctoral research. The data and results are shown in three publications, which are collaborative work with other colleagues and published in the journal of *PLOS ONE* (2017), *Journal of Proteome Research* (2021), and *Proteomics* (2022).

1.1 Contribution to paper I

Thomas Andreas Gerland, **Bo Sun**, Pawel Smialowski, Andrea Lukacs, Andreas Walter Thomae, and Axel Imhof. 2017. "The Drosophila Speciation Factor HMR Localizes to Genomic Insulator Sites." *PLOS ONE* 12 (2): e0171798. doi:10.1371/journal.pone.0171798.

This work was a collaboration work with Thomas Andreas Gerland. After we acquired raw data from ChIP, RT-PCR, and sequencing, I performed data analysis with Python, R, bash, and other related bioinformatic packages and resources on reads alignment, peak calling, motif search, peak annotation, identification of differential expressed genes and statistical analysis. Most of the results were visualized in different figures with R. For the preparation of the manuscript, I took part in the writing of the methods part and partial results, including revisions.

1.2 Contribution to paper II

Bo Sun, Pawel Smialowski, Tobias Straub, and Axel Imhof. 2021. "Investigation and Highly Accurate Prediction of Missed Tryptic Cleavages by Deep Learning." *Journal of Proteome Research*. doi:10.1021/acs.jproteome.1c00346.

This work was supported by and collaborated with Prof. Imhof and other co-authors. Prof. Imhof and I proposed the conceptualization and made investigations of other related work. Then I collected data, designed the workflow, and performed the formal analysis with bioinformatic tools and resources, especially TensorFlow and python for deep-learning, and R for statistical analysis. Moreover, I took part in the visualization of all results and preparation of the manuscript including writing and revisions.

1.3 Contribution to paper III

Bo Sun, Pawel Smialowski, Wasim Aftab, Andreas Schmidt, Ignasi Forne, Tobias Straub, and Axel Imhof. 2022. “Improving SWATH-MS analysis by Deep Learning.” *Proteomics*, 2022, doi: 10.1002/pmic.202200179.

This work was supported by and collaborated with Prof. Imhof and other co-authors. Prof. Imhof and I proposed the conceptualization. Then I made investigations of other related work including the collection of publicly accessible data and preliminary testing of different algorithms. I designed the procedure and workflow for data analysis including benchmarking, and applications. For the implementation of algorithms, TensorFlow and python were utilized for deep-learning, and R for statistical analysis. I also took part in the generation of all figures and preparation of the manuscript including writing and revisions.

2. Introduction

In recent years, modern biology has undergone unprecedented improvement on different levels, including genomics, transcriptomics, and proteomics. The accomplishment of the Human Genome Project (Venter et al., 2001; Lander et al., 2001) indicates a great accomplishment in genomics. Genomics has important applications in the diagnosis of diseases (Petersen et al., 2017), the development of medicine (Lu et al., 2014), synthetic biology (Baker et al., 2011), and so on. The information from the genome is then transcribed to RNA, the analysis of which is now known as transcriptomics, which studies the level of gene expression (Schena et al., 1995; Cheung et al., 1999). However, it is proteins that play key roles in the building of a cell and carry out different biochemical reactions, including metabolism, gene regulation, catalysis, molecular signaling, and physical interactions (Yates et al., 2009; Chen et al., 2020). The study of the identification, quantification, and localization of protein components of cells is known as proteomics (Aebersold & Mann, 2003; Yarmush et al., 2002). As an important method of the postgenomic era, proteomics has a substantial impact on the diagnosis of diseases, prognosis, and development of drugs (Aslam et al., 2017). Nowadays, both experimental and informatic methods have been developed to improve the analysis of proteomic data (Mallick et al., 2010; Choi et al., 2020). For the experimental aspect, mass spectrometers (MS) combined with liquid chromatography (LC) is the most popular and major technology that is critical to the fast development of proteomics (Aebersold & Mann, 2003). Through the last few decades, MS-based proteomics has achieved great success but still faces big challenges. Diverse technologies and strategies have been developed for different mass spectrometers, such as bottom-up and top-down strategies, which have been widely applied in proteomics research. The study of various levels of omics not only provides comprehensive insights into the composition of organisms, but also provides different aspects to the analysis, elucidate potential causative changes that lead to diseases, precision diagnosis of diseases, and finally the development of new drugs (Hasin et al., 2017).

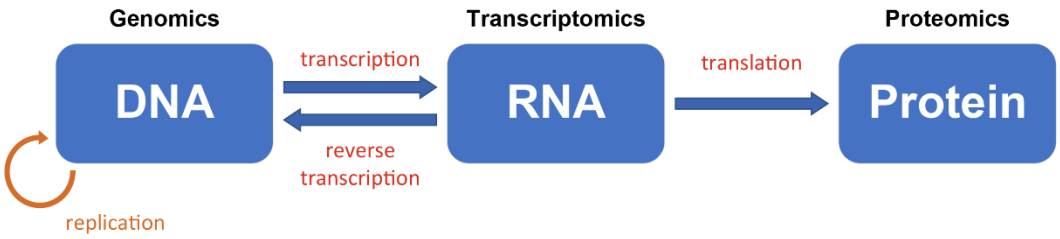


Figure 2.1 The illustration of the different levels of omics studies. The information of DNA flows to protein corresponding from genomics to proteomics.

Here, we discuss the data analysis for genomics, transcriptomics, and proteomics. Even though advanced instruments have provided powerful methods to study omics, however, complex data have been generated from such equipment which needs to be analyzed to reveal the biological meaning behind them (A.L.McGuire et al., 2020; Cristoni & Bernardi, 2004; Patel et al., 2021). Nowadays, both traditional statistics and machine learning methods have been applied to analyze omic data including qualitative and quantitative data. The identification of peptides and proteins in the organelle, cell, or tissue lysate is the focus of qualitative proteomic data, while quantitative proteomic data often includes the comparison of two or more biological states (Kumar et al., 2009). Computational algorithms and software for both qualitative and quantitative data have been developed. The algorithms that have been developed for omics can be used in the following applications: data preprocessing, statistical analysis, enrichment analysis, and so on (Chen et al., 2020). Besides, machine learning or deep-learning has more applications in the analysis of omic data, for example, the applications of machine learning to human genomics (Alharbi WS et al., 2022), phenotype prediction from transcriptomics data (Smith AM et al., 2020), and deciphering proteome profiling by deep-learning (Wang et al., 2020).

In this part, the major concepts of omics and related realms are introduced, then the computational and statistical methods for omics are discussed. Finally, the strategies used for the analysis of genome sequencing, gene expression, and mass spectrometry are introduced which I performed for my doctoral research.

2.1 Genomics

One century ago, the term “genome” was created to refer to the complete set of genes and chromosomes in an organism. However, “Genomics” is an inter-discipline in biology to study the mapping, sequencing, evolution, and editing of genomes. Genomics can be classified as “structural genomics” and “functional genomics” for different research aspects. Structural genomics focuses on the three-dimensional construction of proteins encoded by a certain genome (Hieter P et al., 1997). While functional genomics focuses on gene function and regulation, such as the dynamic expression of gene products in space, time, and disease (Przybyla L et al., 2022).

The accomplishment of the Human Genome Project (HGP) has brought fruitful contributions to the study of genomics. One major goal of the HGP was to create genetic and physical high-resolution maps for each human chromosome (Collins FS et al., 2003). The relative position of genes and DNA markers along the chromosome can be illustrated by genetic and physical maps. Recombination frequencies are used to figure out how far apart two points are on a genetic map, while the number of nucleotide pairs between loci is used to make a physical map. Genetic maps are an indispensable resource for the creation of physical maps. Both genetic and physical maps are important to elucidate the organization of a genome.

However, the decipherment of genomic DNA sequences is just the beginning of the exploration of the biological mechanisms behind the arrangement of nucleotides along chromosomes in a living organism. The generated vast amounts of sequence data trigger the development of bioinformatics for the elucidation of the expression and functions of all the genes in a genome. Nowadays, different computational methods have been applied to genomic analysis, including organization, analysis, understanding, visualization, and storage of genomic data (Diniz WJS et al., 2017). To achieve this goal, some public databases have been built to store the rapidly increasing genome data, such as European Molecular Biology Laboratory (EMBL), DNA Database of Japan (DDBJ), GenBank at the National Center for Biotechnology Information (NCBI); and also some functional databases, such as Reactome and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa M et al., 2017; Diniz WJS et al., 2017). By integrating these genomic data, the comparison among sequences can be done through alignment to elucidate the evolutionary relationship between genes, individuals, organisms, and others (Junqueira et al., 2014).

2.2 Transcriptomics

The total set of ribonucleic acid (RNA) molecules present at a particular developmental stage or physiological condition in a cell, tissue, or organism, is known as the transcriptome (Wolf JBW, 2013; Milward EA et al., 2016). All kinds of transcripts, including mRNAs, noncoding RNAs, and small RNAs, are covered by transcriptomics. Transcriptomics focuses on the study of structures, related genes, locations, functions, transcription, expression levels, trafficking, and degradation (Milward EA et al., 2016). With the study of transcriptomics, gene expression in an organism can be measured in certain conditions or tissues, which gives insights into the regulation of genes and more details of an organism's biology. Besides, the functions of previously unannotated genes can also be inferred through transcriptomic analysis (Lowe R et al., 2017). And nowadays transcriptomics also plays a key role in contemporary cancer medicine (Supplitt S et al., 2021).

The study of transcriptomics advances with the development of high-throughput technologies. RNA sequencing (RNA-seq) has been a ubiquitous technique for transcriptomic analysis, including the discovery of novel transcripts, analysis of differential gene expression (DGE), detection of allele-specific expression, and characterization of alternative splicing variants. Compared to other next-generation approaches, RNA-seq has higher resolution and coverage in characterizing the dynamic nature of the transcriptome (Kukurba KR et al., 2015). Besides, RNA-seq data are generated from functional genomic elements directly, most of them are protein-coding genes. A typical RNA-Seq experiment involves the isolation of RNA, then conversion to complementary DNA (cDNA), the preparation of the sequencing library, and its sequencing on an NGS platform. Nowadays, a parallel sequencing by-synthesis method is used by most high-throughput sequencing platforms to sequence tens of millions of sequence clusters (Lowe R et al., 2017).

With the accumulation of RNA-seq data, bioinformatic approaches are necessary to be developed to elucidate the biological meaning of such complex data. The general procedure for RNA-seq data analysis includes sequenced reads stored in FASTQ-format files generated from an NGS platform, alignment of these reads to reference genome, and gene expression quantification (Geraci F et al., 2020). However, several challenges to informatic analysis in RNA-seq need to be further addressed, such as the storage, retrieving, and processing of huge data, the errors in base-calling, and image analysis (Wang Z et al., 2009).

2.3 Proteomics

Proteins are polymers of amino acids with functional groups of amino (-NH3+) and carboxylate (-CO2-) and a specific side chain for distinct amino acids (Nelson et al., 2005). The structure of proteins can be defined on four distinct levels: the primary structure of a protein is specified by the sequence of its amino acids, which then directs the secondary structure by the proper folding of the polypeptide chain including the alpha helix, in which corkscrew shape is folded in a region of the polypeptides; however, the other common type of secondary structure, beta-strands are formed in a linear structure of polypeptides by bonding together. The three-dimensional structure of proteins is formed by the chemical interactions of turns and coils, resulting in the final protein (Chandrasekhar et al., 2014). Proteins play key roles in almost all biological processes, such as catalysis (Agarwal et al., 2006), molecular signaling (Yates et al., 2009), immune function, and gene regulation (Chen et al., 2020). Many illnesses are caused by aberrant protein function regulation, which is an important objective of biomedical research in the development of possible novel medications for disease therapy (LaBear, 2002). Moreover, the combination of the information of genome and proteome has been applied to develop new strategies for the designing of drugs for associated diseases (Chandrasekhar et al., 2014).

The term “proteome” is to describe the total number of proteins in a cell, as well as their localizations, physical interactions, and post-translational modifications (PTMs) at any given moment (Dupree et al., 2020; Aslam et al., 2017). The identification, quantification, and localization of protein components in cells are the primary research aspects of proteomics, which extends from protein expression profiling and signaling circuit analysis to the creation of protein biomarkers (Mallick et al., 2010; Yates et al., 2009).

Specifically, the scope of proteomics covers the following aspects: protein expression profiling, structural and functional proteomics, and so on (Graves et al., 2002). The structure and function are two major aspects of proteins which are also the key studies of proteomic research. For proteins, functions are determined by their structures. The purpose of structural proteomics is to find all of the proteins inside a protein complex or a particular cellular organelle, locate them, and describe their protein-protein interactions.

However, the advance of proteomics cannot be improved without the development of associated techniques (Aslam et al., 2017), such as conventional techniques like chromatography-based techniques, enzyme-linked immunosorbent assays (ELISA), and western

blotting. Some advanced techniques, including mass spectrometry (MS), and protein microarray.

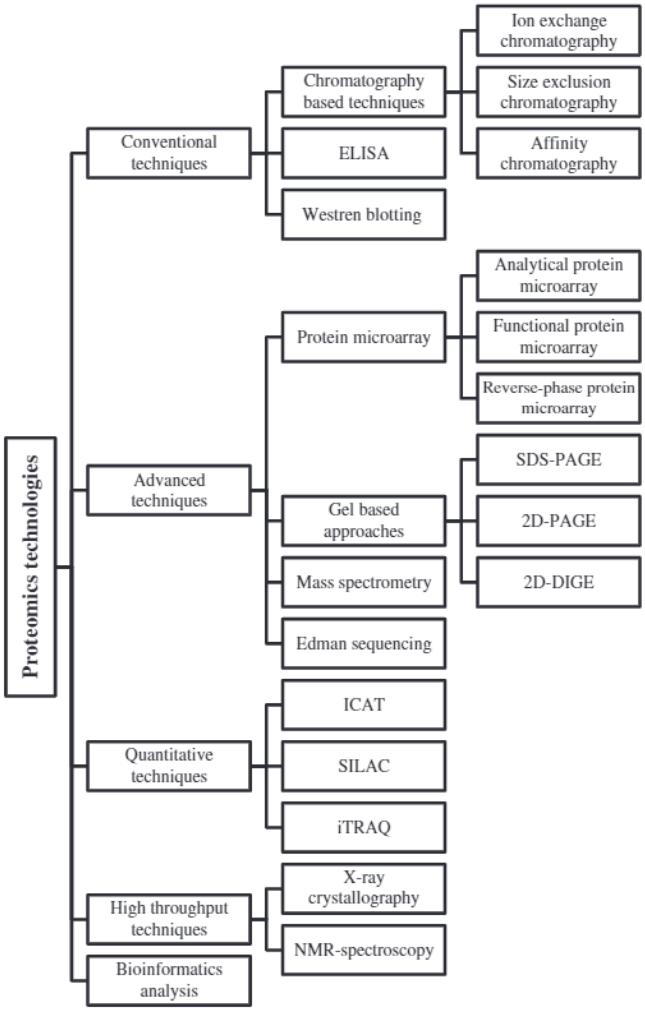


Figure 2.2 An overview of proteomics techniques (Aslam et al., 2017).

Nowadays, MS is becoming more essential in proteomics research since it complements other methods and can identify proteins in extremely small amounts, such as 1-10ng (Keshishian H et al., 2007).

There are two typical strategies for MS-based proteomics: bottom-up and top-down (Kar et al., 2017). For the bottom-up approach, peptides are generated from the digestion of proteins, which are then analysed in a mass spectrometer (Gillet et al., 2016). However, for top-down proteomics, intact protein ions or large protein fragments generated by electrospray ionization (ESI) are subjected to gas-phase fragmentation for mass spectrometry analysis (Toby et al., 2016; Donnelly et al., 2019; Chen et al., 2008).

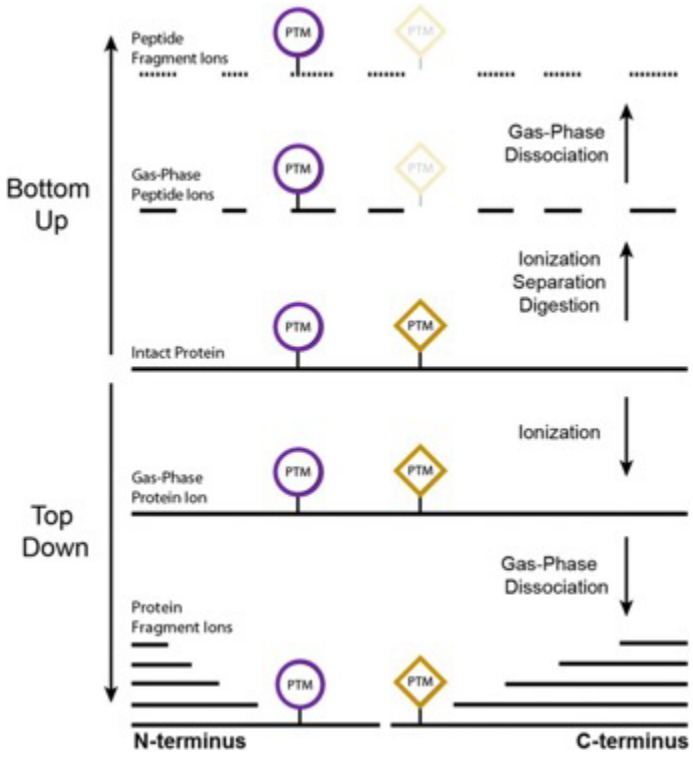


Figure 2.3 The comparison of bottom-up and top-down approaches in mass spectrometry. In the bottom-up technique, intact proteins are digested into peptides, which are then detected and fragmented using a mass spectrometer. While in the top-down approach, the intact protein is ionized directly which improves the coverage of protein sequence and the detection of PTMs (Catherman A. D. et al., 2014).

2.4 Next generation sequencing

Next generation sequencing (NGS) plays a key role in both genomics and transcriptomics. Compared to traditional Sanger sequencing (Sanger F et al., 1977), NGS is much cheaper, and faster, along with higher throughput in sequencing DNA. Millions of fragments of DNA in a single sample can be sequenced together due to the massively parallel sequencing technology of NGS, which leads to an entire genome can be sequenced in less than one day. With such high-throughput capability, NGS can identify disease-related genes and regulatory elements by sequencing the human genome. Besides, the complexity of genome can also be figured out through the performance of NGS (Grada A et al., 2013).

With the development of NGS, more studies including Chromatin immunoprecipitation followed by sequencing (ChIP–seq), genome-wide association (GWA) studies, and RNA sequencing (RNA-seq) become much easier than before (Hawkins RD et al., 2010). ChIP-seq is a method for profiling DNA-binding proteins, histone modifications, or

nucleosomes throughout the entire genome. Due to the remarkable advances of NGS, ChIP-seq achieves fewer artifacts and higher resolution with a larger dynamic range than previous methods. Transcriptional regulation can be elucidated through the genome-wide mapping of epigenetic marks and protein-DNA interactions. The gene regulatory network for different biological processes can be elucidated by precise mapping of binding sites for transcription factors (TFs), key transcriptional machinery, and other DNA-binding proteins (Farnham PJ et al., 2009; Park PJ et al., 2009). To reveal the above mechanisms, ChIP is the main technique for the detection of protein-DNA binding *in vivo* (Solomon MJ et al., 1988). Compared to the array method, the interested DNA fragments are directly sequenced rather than hybridization. The procedure of a ChIP-seq experiment is shown in Figure 2.4.

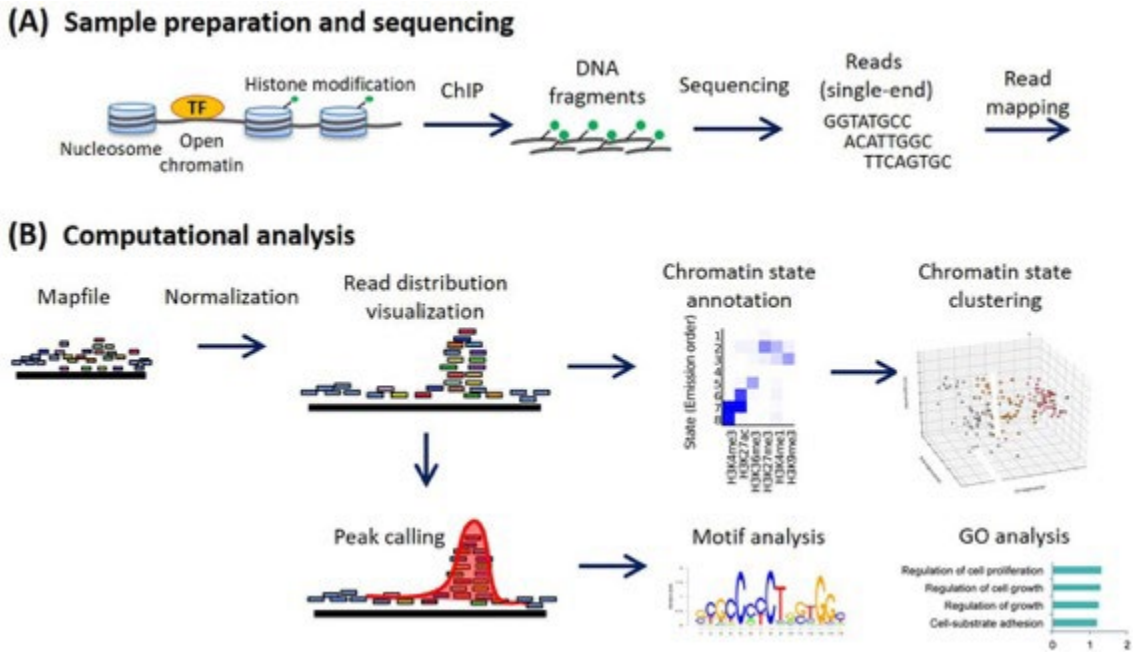


Figure 2.4 The procedures to perform a ChIP-seq experiment and computational analysis. (A) Preparation and sequencing of samples. (B) Data analysis for a typical ChIP-seq assay. (Nakato R et al., 2009).

However, there are still some challenges in both experimental and computational ChIP-seq analysis. For the experimental aspect, the artifacts can not be eliminated, especially at the end of each read. Besides, the enrichment of GC content in fragment selection, and the difficulties in loading the exact amount of sample for the generation of high-quality data. In addition, the current cost and availability of ChIP-seq still have potential space

to ameliorate. For the ChIP-seq data analysis, the amounts of different types of data such as image data, sequence tags, and alignment data increase dramatically, for which there are still problems in the storage and extraction of such high-throughput data. In addition, most non-unique tags in genome alignment are not handled properly. Besides, more reliable, and advanced software needs to be developed (Park PJ et al., 2009).

For transcriptomic analysis, RNA-seq has become a ubiquitous technique in gene expression studies. The quantification of gene expression includes the discovery of new transcripts, and the characterization of alternative splicing variants or novel cell types. Furthermore, the application of RNA-seq in clinical diagnosis has become true. Compared to traditional hybridization-based methods, RNA-seq can break the limitation of annotated genomic sequences to detect novel transcripts. Besides, RNA-seq is capable of locating precise transcription boundaries and revealing sequence variations such as SNPs (Cloonan N et al., 2008). In addition, RNA-seq has very few background signals compared to DNA microarrays, and unlimited upper levels for quantification. Thus, RNA-seq is a high-throughput and quantitative approach to studying gene expression levels by scanning the transcriptome. A typical RNA-seq experiment is shown in Figure 2.5.

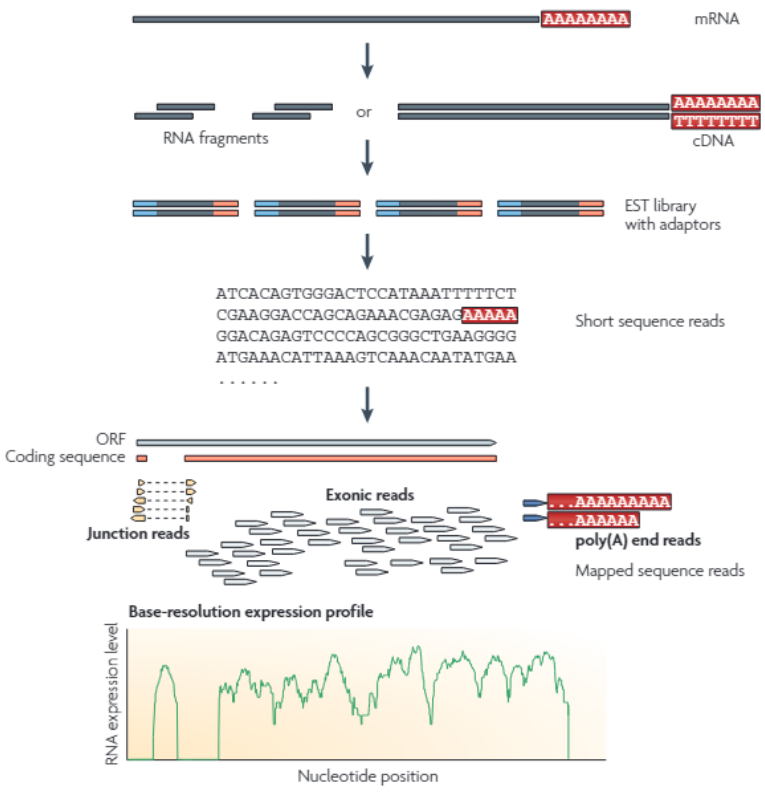


Figure 2.5 The procedure to perform a typical RNA-seq experiment. Firstly, a cDNA library is prepared by fragmentation of either RNA or DNA for long RNAs. Then each cDNA fragment is added with sequencing adaptors (blue) and a short sequence is generated. Three types of sequence reads, which are junction reads, exonic reads, and poly(A) end reads, are aligned to the reference genome or transcriptome. The bottom plot shows an expression profile for genes with the usage of the above three types of reads (Wang Z et al., 2009).

However, like other NGS techniques, RNA-seq also faces challenges on both sides of experiments and data analysis. For the experimental aspect, the complexity of the cDNA library building in profiling transcripts, biases generated from RNA fragmentation, and lack of strand information for cDNA analysis. Meanwhile, the challenges also come from informatics for RNA-seq analysis. First, like other NGS approaches, the continuously increasing amounts of data require efficient storage, retrieving, and processing. Second, the errors come from base-calling, image analysis, and low-quality reads. Third, efficient and simple computational algorithms need to be developed to identify novel splicing events between two distant sequences or genes. Last, the issue comes from the increased cost of greater sequence coverage (Wang Z et al., 2009; Geraci F et al., 2020).

2.5 Mass spectrometry

Nowadays, MS is one of the most important techniques for proteomics because of its ability to handle the complexities of proteomic research. Compared to other traditional techniques, MS can achieve in-depth proteomic information. Besides, the development of soft ionization techniques, such as matrix-assisted laser desorption/ionization (MALDI), electrospray ionization (ESI), and liquid chromatography (LC), is crucial to the successful performance of mass spectrometric analysis (Han et al., 2008; Aebersold et al., 2003).

The primary structural information of proteins, the amino acid sequences, can be obtained through MS analysis, which can then be used for the identification of proteins by searching databases. Besides, the type and location of protein modifications can also be determined by MS analysis (Graves et al., 2002). The information on proteins acquired by MS can be achieved through three stages: 1) sample preparation, 2) sample ionization, and 3) mass analysis. The technique of liquid chromatography, such as HPLC, can meet the

requirements for the purification and separation of peptides. Before the samples are analysed by the mass spectrometer, the molecules must be charged and dry, which can be accomplished by ESI and MALDI mentioned above. Generally, the integrated liquid-chromatography ESI-MS systems (LC-MS) are applied to more complex samples in contrast to MALDI-MS. Then, MS will measure the gas-phase ions produced by MALDI-MS or ESI-MS, which are nebulized into tiny, highly charged droplets in an electrospray ion source. After evaporation, the gas-phased multiply protonated peptides are subjected to the mass analyzer of the mass spectrometer, which measures their mass-to-charge ratio (m/z). Based on the mass spectra generated by the computer connected to the mass spectrometer, information about peptides and proteins can be acquired by matching against protein sequence databases (Aebersold et al., 2003).

For the applications of MS experiments, mass analysers are indispensable to mass spectrometers for their ability to store and separate ions based on m/z. There are different types of mass analysers, such as time-of-flight (TOF) and quadrupoles (Q), ion trap (IT), Orbitrap, and ion cyclotron resonance (ICR), with different unique properties, such as mass range, resolution, sensitivity, dynamic range, and analysis speed. These analyzers may be used alone or in combination to maximize the benefits of each (Aebersold et al., 2003; Yates et al., 2009).

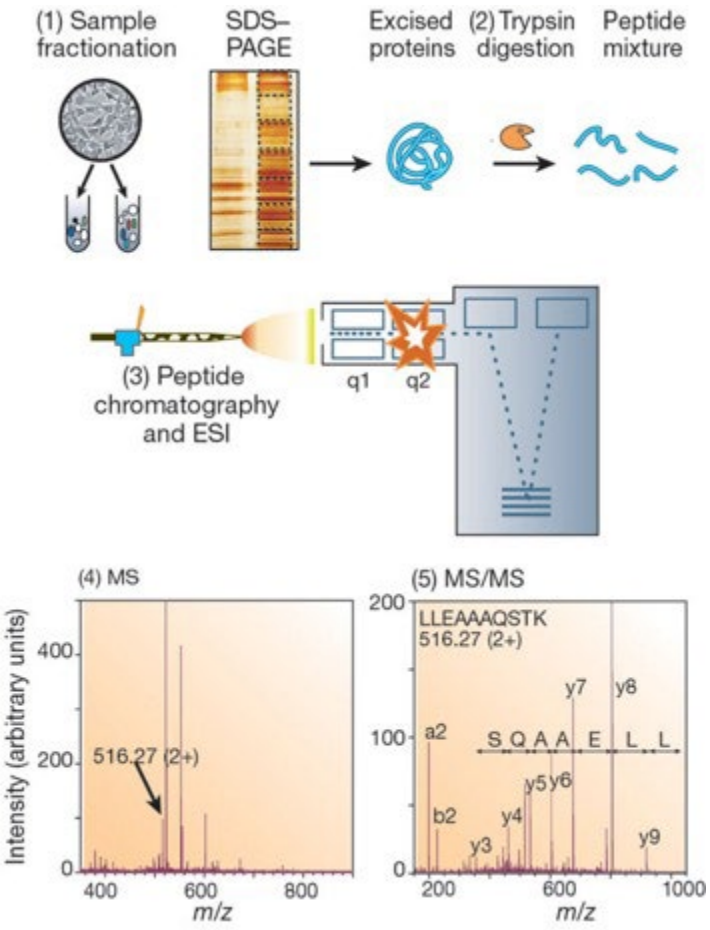


Figure 2.6 The workflow of MS-based proteomics experiment. Five stages for the MS-based proteomics experiments are illustrated. In stage 1, the protein samples are isolated from cells or tissues by biochemical fractionation or affinity selections. Then the proteins are digested by protease such as trypsin to peptides in SDS-PAGE in stage 2. In stage 3, the peptides are separated by LC and eluted into an electrospray ion source which then enter the mass spectrometer. Then the mass spectra for given peptides are taken at a specific time point in stage 4. In the last stage, the fragmentation of these peptides and a series of tandem mass spectrometric experiments are performed (Aebersold et al., 2003).

For the applications of MS experiments, several strategies are being widely used: data-dependent acquisition (DDA), data-independent acquisition (DIA), multiple reaction monitoring (MRM), and parallel reaction monitoring (PRM).

In the following part, the instruments and tools for mass spectrometry and the strategies of DDA and DIA will be introduced in detail.

2.5.1 The instruments and tools for mass spectrometry

A mass spectrometer consists of an ion source that converts biological molecules into gas-phase ions, a mass analyser that measures the mass-to-charge ratio (m/z) of the ionized analytes, and a detector that records the number of ions at each m/z value (Aebersold et al., 2003; Han et al., 2008).

Matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) are two soft ionization techniques that transfer analytes into the gas phase without extensive degradation, which enables proteins and peptides to be analysed by MS. The MALDI method involves transferring a laser-heated matrix to an acidified analyte, which causes the analyte's [M+H]+ ions to enter the gas phase (Yates et al., 2009). To achieve an acceptable signal-to-noise ratio for detection, the ionization of MALDI requires hundreds of laser shots to prepare enough energy (Liao et al., 1995). Because the produced ions are generally singly charged, MALDI is suited for top-down analysis of high-molecular-weight proteins. However, there are some drawbacks to the MALDI technique: the low shot-to-shot reproducibility and reliance on the sample preparation methods which leads to the improvements of this approach, such as the matrix-free MALDI techniques SALDI (Chen et al., 1998), DIOS (Shen et al., 2001), and atmospheric pressure MALDI (AP-MALDI) (Laiko et al., 2000).

Another important ionization technique is ESI. Compared to MALDI, ESI is driven by high voltage (2–6 kV) to produce ions from solution. The formation and desolvation of analyte-solvent droplets follow the formation and desolvation of an electrically charged spray in the physicochemical process of ESI (Yates et al., 2009). Unlike MALDI, the ions from ESI are multiple-charged species and sensitive to analyte concentration and flow rate. So some improvements have been proposed for ESI, such as micro and nano-ESI (Griffin et al., 1991; Emmett et al., 1994). The two ionization methods are usually chosen for different mass analyzers (Aebersold et al., 2003; Yates et al., 2009).

In a mass spectrometer, a mass analyser is indispensable for its ability to take and separate ions based on the mass-to-charge ratio (m/z). The mass analyzers can be divided into two categories: the trapping mass spectrometers, such as IT, Orbitrap, and FT-ICR; and the scanning and ion-beam mass spectrometers, such as TOF and Q. MALDI is usually chosen for TOF analyzers to measure the mass of intact peptides through pulsed analysis, whereas ESI has mostly been used in conjunction with ion-beam and trapping equipment. Several types of instrument configurations are widely used in proteomic research: ion

traps, such as LIT or LTQ (Hager et al., 2003); triple quadrupoles (TQ), LTQ-Orbitrap (Hu et al., 2005; Makarov et al., 2006), LTQ-FTICR (Syka et al., 2004; Breuker et al., 2008), Q-TOF (Morris et al., 1996; Shevchenko et al., 1997) and IT-TOF (Collings et al., 2001; Campbell et al., 1998). In this part, Orbitrap and Q-TOF will be introduced in detail.

The Orbitrap machine is widely used for proteomics for its high resolution (up to 150,000), high-mass accuracy (2–5 ppm), and good dynamic range greater than 10^3 (Hu et al., 2005). Within an orbitrap instrument, a static electric field is created, in which ions orbit and oscillate in the axial direction around a central electrode (Figure 2.7). A fast Fourier transform (FFT) algorithm (Senko et al., 1996) is then used to convert the overlapping frequencies into mass-to-charge spectra. Because of its high mass accuracy, Orbitraps can perform alternate data acquisition and data analysis approaches to achieve greater coverage and accuracy.

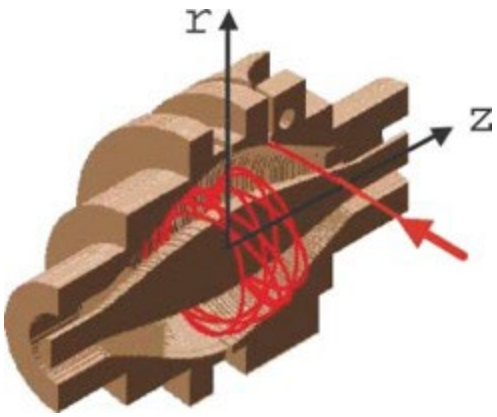


Figure 2.7 Cross section view of the Orbitrap mass analyzer. The injection point and pathways of ions in the mass analyzer are indicated in red arrows and lines respectively. The two perpendicular directions of the mass analyzer are shown in the z and r-axis (Hu et al.,2005).

Another advanced tandem mass spectrometer is a quadrupole time of flight instrument (Q-TOF), which provides high peak capacity, resolving power (e.g., RP ~ 10000), mass measurement accuracy (e.g., MMA ~ 10 ppm), spectral acquisition rates, and dynamic ranges (>3 orders of magnitude). As mentioned above, Q-TOF machine is usually coupled with ESI to perform better analysis than other modes.

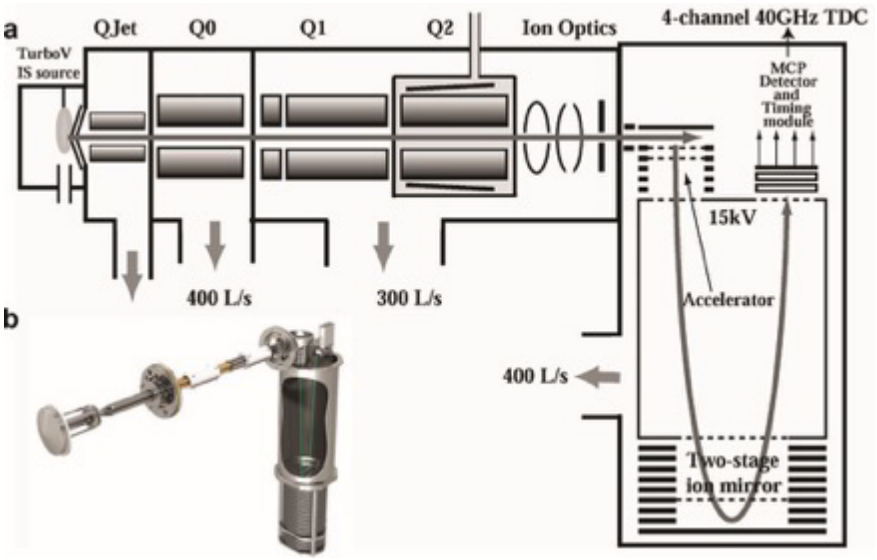


Figure 2.8 Schematic of the TripleTOF instrument. (a) Overall diagram of the TripleTOF. (b) An image of the TripleTOF machine (Andrews et al., 2011).

The TripleTOF instrument is one of the hybrid Q-TOFs (Andrews et al., 2011). With its features of high mass accuracy, resolution, speed, and sensitivity, TripleTOF is suitable for DDA analysis. Moreover, comprehensive and specific peptide quantification by DIA, such as SWATH (sequential window acquisition of all theoretical spectra) and MS (mass spectrometry), can also be operated on TripleTOF (Gillet et al., 2012). Next, the two strategies of DDA and DIA will be introduced.

2.5.2 DDA & DIA

With LC-MS/MS, DDA has been extensively utilized for the identification and quantification of protein groups in a range of biological samples (Bateman et al., 2014; Mann et al., 2001) for its breadth of detection, flexibility, and simplified settings and measurements (Hu et al., 2016). Single precursor ions are selected by mass spectrometer based on their abundance, which are isolated in MS1 scan and then fragmented in sequential MS2 scans in DDA mode. For each of the MS2 scans, a database search algorithm is applied to the analysis. Through the performance of DDA, thousands of proteins can be identified. However, the DDA approach has significant limitations: the irreproducibility and imprecision are produced by simply measuring the most abundant peptides, which means that low-abundance peptides might be overlooked. Moreover, accurate quantification is hindered by only one or two times of measurements for each peptide (Venable et al., 2004).

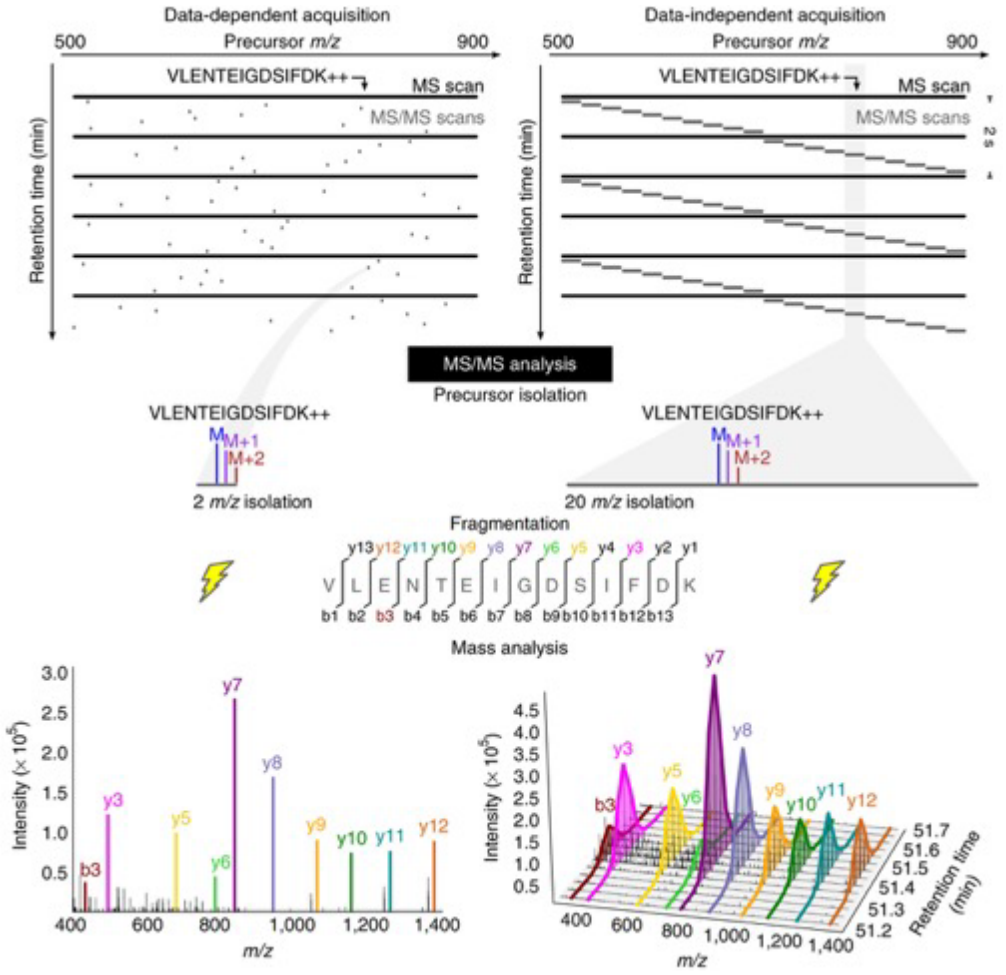


Figure 2.9 The illustrations for strategies of DDA and DIA. DDA records MS/MS spectra from individually isolated peptide precursors. DIA uses wide isolation windows to acquire fragment spectra from multiple precursors. These mixed spectra are then deconvoluted by dedicated software packages (Egertson et al., 2015).

DIA is a method for detecting all peptides within vast, pre-specified mass ranges by isolating, fragmenting, and analyzing all precursor ions using a high-resolution mass spectrometer (Hu et al., 2016). The benefits of PRM (high sensitivity and reproducibility) and DDA have been blended in DIA (broad protein coverage). Besides, DIA has higher sensitivity, reproducibility, and selectivity compared to DDA mode (Figure 2.10).

Figure 2.10 Performance profiles of DDA and DIA. Five metrics including sensitivity, reproducibility, selectivity, multiplexing, and ease of assay development are shown in radar graphs. For each metric, 4 indicates the best performance while 0 indicates the worst performance (Li et al., 2021).

Precursor ions are sampled and separated into consecutive small mass-to-charge (m/z) windows (5-25 Da) in Q1, which are then fragmented in Q2 (Figure 2.8a). The product ions within a certain m/z window are monitored by a high-resolution accurate-mass (HRAM) mass analyser in an unbiased and systematic manner (Shi et al., 2016; Huang et al., 2015). Then, highly complex MS2 spectra are generated from the co-fragmentation of peptides that belong to the same precursor. Because the connection between the precursor and its fragments is lacking, a spectra library based on DDA investigations is required to understand such complicated MS2 spectra (Gillet et al., 2012; Ludwig et al., 2018). The quantification analysis by DIA is generally comparable to those targeted methods because of its high reproducibility and mass accuracy. Compared to DDA, the mass spectra generated in DIA have an additional dimension, the retention time (RT), which makes the information of fragment ions can be extracted over time to promote the quantitative analysis of peptides and proteins. In general, DIA mass spectra quantitative findings are the total of the area under the curve of each fragment ion, which is preferable to DDA in terms of resolving quantitative information.

However, compared to some targeted methods, such as MRM and PRM, DIA has lower selectivity and sensitivity for its highly complex MS spectra. Moreover, the sensitivity, selectivity, and proteome coverage of a DIA assay can be affected by several factors, for example, the instrumentation, mass-to-charge (m/z) windows width, and spectral library. So the optimal settings for different parameters are necessary for the precise identification and quantification results.

2.6 Bioinformatics for NGS and MS

The development of omics is not only promoted by the improvement of instrumentation and experimental technologies but also by bioinformatics. A huge amount of data has been generated from different levels of omic research (Kumar et al., 2009; Magi A et al., 2010; Cristoni S et al., 2014). The main goal of bioinformatics on omics is to organize and interpret the biological meaning of data generated from experiments such as NGS and MS. For the NGS, we focus on the data analysis for ChIP-seq and RNA-seq.

Qualitative and quantitative proteomics data are still two major aspects of proteomic bioinformatics. Traditionally, only the most abundant proteins in gel electrophoresis can be analysed. However, with the development of mass spectrometry, the data at the proteome level can be analysed by computational methods. Bioinformatics for proteomics is rapidly evolving, and fields as diverse as mathematics, statistics, and computer science have been used to handle the challenges posed by such complex data. (Chen et al., 2020).

2.6.1 Traditional statistical methods for NGS and MS

For the data generated from the NGS technique, multiple computational methods have been developed for ChIP-seq and RNA-seq analysis. For the ChIP-seq analysis, the main goal is to map the interactions between proteins and DNA by the isolation of genomic fragments which interact with antibodies or DNA-binding proteins such as TFs (Park PJ et al., 2009). The reads generated from isolated genomic fragments mapped to the reference genome are used to identify enriched regions for functional factors (Nakato R et al., 2021). After the mapping of reads to the reference sequence, the discrimination of genomic regions enriched with reads from 'background' noise is necessary. To filter out noise coming from the background, a negative control can be used to generate a noise pattern used to compare with real data. In this way, the enriched genomic regions in the positive sample can be detected for further consideration, while the tags from the control experiment can be used as a background model. However, in experiments which are lack control samples, stochastic methods can be applied to estimate the background read levels. The hypothesis for such cases is that each genomic region to be extracted and sequenced has the same probability in a total random experiment. If we define the total number of tags as t, and the size of the genome as g, then t/g stands for the probability of one tag mapped in a given position. Thus, the probability of the expected number of tags within

a genomic region can be calculated, for example, by Poisson or negative binomial distributions. Then, through the sliding windows across the whole genome, the significance of the tag enrichment can be calculated. Several 'peak-calling' programs have been developed, such as MACS (Cokus SJ et al., 2008), QuEST (Valouev A et al., 2008), FindPeaks (Fejes AP et al., 2008), HOMER (Heinz S et al., 2010), ChIPseek (Chen T-W et al., 2014), in which several tools are prepared for ChIP-seq analysis. Moreover, the built-in background model can be used to estimate the significance of tag enrichment (Horner DS et al., 2010).

For the RNA-seq analysis, one-end or paired-end sequencing is applied to generate sequence reads from total or poly-A enriched RNAs. Generally, protein-coding mRNAs are detected based on the poly-A enrich fraction. However, the non-polyadenylated ncRNAs can be missed in this way. Then the total RNA can be randomly amplified to have a broad overview of the transcriptome. The read lengths range from 30bp to over 400bp, which are generated from different NGS platforms, such as ABI SOLiD, Illumina, and Roche 454 FLX (Horner DS et al., 2010). Different read lengths are selected for various applications, the short reads with high throughput are chosen for transcript quantification by tag profiling. While longer reads are more suitable for the determination of exon coordinates and relative quantification for expressed isoforms of full length. For the detection of novel splicing sites and variants, there are several mapping tools such as TopHat (Trapnell C et al., 2009), QPALMA (Bona FD et al., 2008) to split align reads against the reference genome. However, there are also some other tools developed such as RefSeq (Pruitt KD et al., 2007), ASPicDB (Castrignanò T et al., 2008) for the correct detection of exon boundary. Then the quality of mapped RNA-seq data needs to be assessed in depth, which can be done with Picard (http://broadinstitute.github.io/picard/), SAMTools (Li et al., 2009), Qualimap2 (Okonechnikov et al. 2016), RNASeQC (DeLuca et al. 2012).

For the quantification of gene expression levels, the concept of Reads Per Kilobase of exon model per Million mapped measures (RPKM) was introduced, which is calculated by the following Equation (1):

$$RPKM = \ 10^9 \times \frac{C}{N \cdot L} \qquad (1)$$

Where *C* is the number of reads mapped to the exons of genes, *N* is the total number of mappable reads for the experiment and *L* is the whole length of the exons (Mortazavi A et al., 2008).

While the analysis of differential gene expression, several programs have been developed: DESeq2 (Love et al., 2014), EdgeR (Robinson et al., 2010), CuffDiff2 (Trapnell et al., 2013), which are based on the counts of reads to infer genes or transcripts to calculate the significance of differentially expressed genes. Then the differential expressed genes can be performed for enrichment analysis by tools such as EnrichR (Kuleshov et al., 2016) or DAVID (Huang et al., 2018).

The first important goal of MS-based proteomics is the identification of peptides and proteins, in which case the determination of the sequence of peptides is crucial. Here, data preprocessing, statistical analysis, and enrichment analysis will be discussed for MS-based proteomics. Two approaches, including searching against the fragmentation spectra databases (Geer et al., 2004; Craig et al., 2004) and de novo peptide sequencing (Frank et al., 2005; Shevchenko et al., 1997), will be introduced.

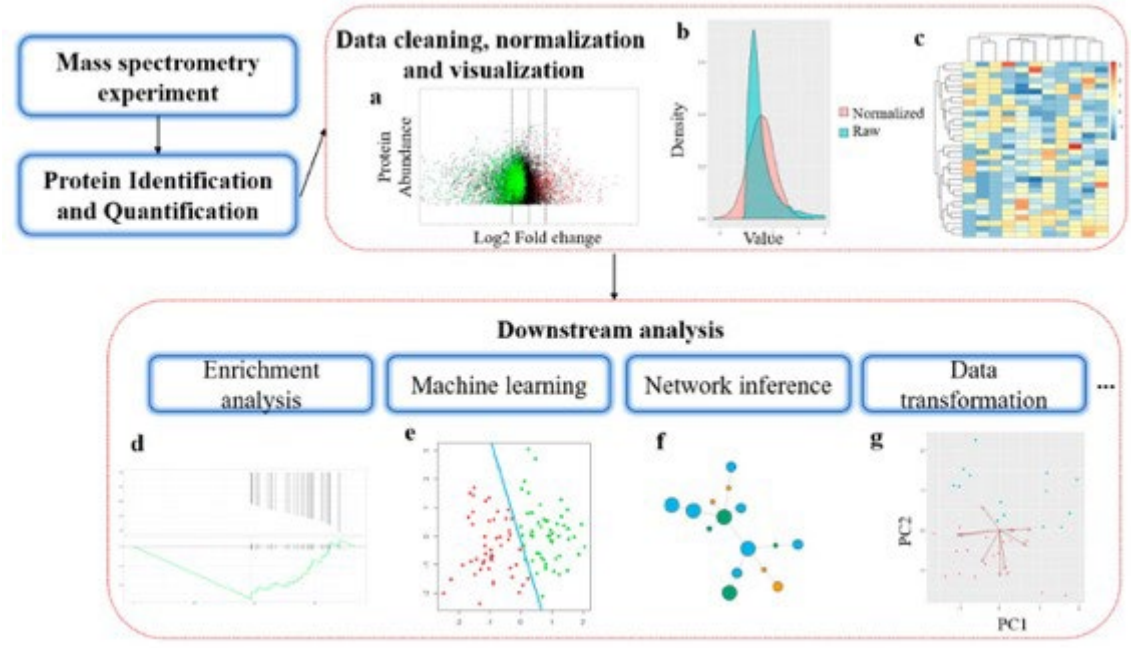


Figure 2.11 General workflow of bioinformatic analysis in mass spectrometry-based proteomics. (a) MA-plot for the differential abundance analysis of proteins. The X-axis indicates the log2 transformed fold change and Y-axis indicates the mean protein abundance from replicates. (b) Normalization of protein abundance data. (c) Heatmap for protein abundance with clustering. (d) Enrichment analysis for protein sets. The X-axis indicates the ranked positions in the protein list, Y-axis in the

above plot indicates the ranked list metric, which in the bottom plot indicates the running enrichment score. (e) Clustering on sample datasets based on machine learning. (f) Illustration of an interaction network inferred from proteomics data. (g) Dimensionality reduction of proteomics expression profile (Chen et al., 2020).

In the database searching approach, the fragmentation spectra with the highest peptide spectrum match (PSM) score are chosen as candidates for the query peptide. So the scoring function of PSMs is crucial to the database searching approach. Several tools have been developed to calculate the PSM score in database searching, for example, SEQUEST's scoring system is based on a normalized cross-correlation between the m/z predicted from sequences and the fragment ions found in mass spectrometers (Eng et al., 2008). Another popular software MASCOT (Perkins et al., 1999) applies probability-based scoring to determine the peptide sequence. Generally, a second round of searching against a decoy database is applied by some software, such as MASCOT (Perkins et al., 1999) and MaxQuant (Tyanova et al., 2016), to reduce FDRs after database searching (Elias et al., 2007).

In contrast to the database searching, Graphical Probabilistic Model (GPM) and Hidden Markov Model (HMM) are preferable choices for the de novo peptide sequencing, such as PepNovo (Frank et al., 2005) and NovoHMM (Fischer et al., 2005). Furthermore, to improve speed, several programs have merged de novo peptide sequencing with a database search strategy, such as InsPecT (Tanner et al., 2005) and DirecTag (Tabb et al., 2008).

After the identification of peptides, protein inference will be performed to reconstruct the peptide sequences into original proteins. Several models have been used during this step: probabilistic models (Nesvizhskii et al., 2003), Hierarchical Statistical Model (Shen et al., 2008), Bayesian inference Model (Li et al., 2009), and so on. However, for the quantitative analysis of proteins, two methods have been widely used: labeled methods and label-free methods. Various bioinformatic methods have been developed for both MS1 and MS2-based labelings, such as MaxQuant (Tyanova et al., 2016), PVIEW (Khan et al., 2009), iTracker (Shadforth et al., 2005), and IsobariQ (Arntzen et al., 2011).

Normalization is frequently required to handle MS-based proteomics data to eliminate any non-biological related variances and make downstream analysis more trustworthy. Several types of normalization methods have been developed based on different statistical hypotheses, such as logarithm transformation on the intensity values, linear regression-based normalization which are applied in RlrMA and LinRegMA (Valikangas et al.,

2016), local regression normalization (Berger et al., 2004), variance stabilization normalization (VSN) (Huber et al., 2002), quantile, median, and EigenMS (Bern et al., 2006). Furthermore, heatmaps and hierarchical clustering are also popular methods for visualizing and preprocessing proteomic input data. Another issue that lies in MS data is the missing values caused by stochasticity in sampling during experiments (Wei et al., 2018). Various methods have been proposed to address this issue, such as singular value decomposition (SVD) imputation (Bergamo et al., 2008), and empirical distribution sampling (Berg et al., 2019).

After the preprocessing of the raw data, more statistical techniques are needed for further analysis. One important work for proteomics analysis is differential expression profiling. T-test and ANOVA (analysis of variance) are two frequently used statistical procedures for determining significant changes by calculating p-values based on certain statistical hypotheses in this kind of research. However, a relatively large variance can be introduced because of the limited multiplexity in proteomics data. To address this issue, moderated t-statistics from the empirical Bayes procedure for Linear Models for Microarray Data (LIMMA) were proposed by Kammers et al. (Kammers et al., 2015). Then the FDR threshold is indispensable for the multiple performances of statistical tests. The Benjamini-Hochberg procedure (Iterson et al., 2010) and FDR estimation from permutation (Xie et al., 2005) are widely used for FDR-controlling.

Enrichment analysis is usually performed to find the overrepresented proteins in the predefined gene set of interest. By performing the enrichment analysis, the systemic hypotheses can be tested on proteomics data instead of the transcriptome. Some publicly available online databases, such as DAVID (Dennis et al., 2003) and STRING (Szklarczyk et al., 2017), include the ability to do enrichment analysis on gene sets based on prior information. PhosphoSitePlus and Signor both give enrichment analysis on modification position/type based on data gleaned via literature mining. Moreover, the enrichment analysis usually needs consistent identifiers which are converted from different databases, in which the conversion tasks can be carried out by some web services, such as PICR (Cote et al., 2007) and CRONOS (Waegele et al., 2009). The Gene Ontology (GO) annotation (Gene Ontology Consortium, 2004) is another notable use of enrichment analysis, which employs Fisher's exact test and the hypergeometric test to clarify the biological process in which chosen genes or proteins are engaged. In addition to the above applications, enrichment analysis is also usually performed on regulatory pathway networks and diseases, which is available on several databases, such as PANTHER (Mi et al., 2009),

KEGG (Kanehisa et al., 2017), and Reactome (Croft et al., 2011) for different pathways analysis. Besides, similar to gene set enrichment analysis (GSEA), protein set enrichment analysis (PSEA) is a popular enrichment approach that calculates the enrichment score based on the significant changes of proteins in abundance, which is available on the software PSEA-Quant (Lavallee-Adam et al., 2014).

2.6.2 Machine learning methods for proteomics

Nowadays, more and more machine learning methods have been put into the application to address biological questions from basic nucleotide and protein sequence analysis to systems biology. Machine learning, compared to classical statistics, builds predictive models based on useful features from large datasets, allowing intricate statistical principles to be learned and applied to new datasets for prediction. Based on the applications for different tasks, machine learning can be divided into two categories: one is supervised learning and the other is unsupervised learning. The input and output datasets are both labeled in supervised learning, but they are not in unsupervised learning. Moreover, based on the data types whether they are continuous or discrete, the tasks can also be divided into classification or regression. Several classical machine learning algorithms have been developed and applied in proteomics, such as Support Vector Machines (SVM), Bayesian classifiers, Random Forest, and Deep Neural Networks. These algorithms have been widely used in proteomic research. For example, the k-nearest neighbor (k-NN) algorithm has been used to predict the protein subcellular location based on its sequence (Huang et al., 2004). The combination of SVM and Bayesian classifier was used to predict the surface residues of proteins that participate in protein-protein interactions. (Yan et al., 2004). In addition, machine learning techniques can be used to reduce the dimensionality of high-dimensional proteomics data, which is another key application, such as Linear Discriminant Analysis (LDA), principal component analysis (PCA), and t-Distributed Stochastic Neighbor Embedding (t-SNE) are popular methods which are chosen for such propose (Chen et al., 2020).

In recent years, deep neural networks or deep-learning algorithms have been used to enhance feature selection, peptide identification, and protein inference for proteomics research. (Meyer et al., 2021). Deep neural networks have two major types: recurrent neural network (RNN) (Rumelhart et al., 1986) and convolutional neural network (CNN)

(Fukushima, 1980) based on fundamental tasks such as image and natural language processing. Different deep neural networks have different frameworks, which are characterized by different settings including the number of neurons, layers, and connections between layers (Wen B et al.,2020).

For the CNN, convolutional and pooling layers are fundamental to such architecture, which are usually followed by fully connected layers to process the final output generated from convolutional layers. In CNN, the backpropagation algorithm is used to train the convolution kernel. An important function of CNN for processing information is to extract high-level features by sliding filters on images or sequences in convolution operations. Then the patterns captured by convolutional layers are identified by pooling layers. The outputs from each neuron in CNN are controlled by the activation function. The widely used functions for activation layers include tanh, sigmoid, softmax, ReLU, and leaky ReLU. Through pooling layers, the pixel or sequence information is vectorized and concatenated, which then flows into dense layers. At the end of CNN, a loss layer is generally connected to adjust the performance of the model. CNN has been applied in medical image and sequence data analysis (Tang B et al., 2019; Wen B et al.,2020). The framework for a typical CNN is shown in Figure 2.12.

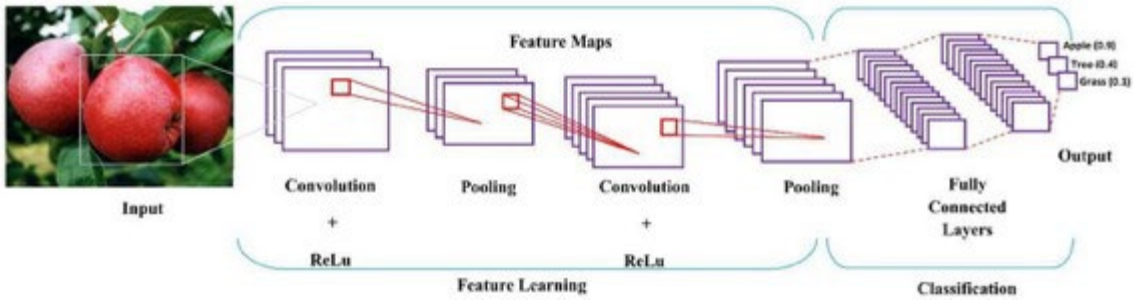


Figure 2.12 The framework for a typical CNN (Tang B et al., 2019).

For the RNN, different architectures and strategies were proposed, which include gated recurrent units (GRU) (Chung et al., 2014) and long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997). Like CNN, the process of information in RNN is also trained with the backpropagation algorithm. In RNN, most of the previous information can be utilized for the current status process, which is illustrated in Figure 2.13 and Equation (2)

$$H_n = \sigma_1 (W_{1,n}^T H_{n-1} + W_{2,n}^T X_n + b_{1,n}) \quad (2)$$

where H_n stands for the hidden layer neuron, $W_{1,n}$ and $W_{2,n}$ are weight matrix, $b_{1,n}$ represent a bias matrix, and $\sigma(\cdot)$ stands for an activation function.

While the total loss L_{total} from each hidden layer is shown as Equation (3)

$$L_{total} = \sum_{n=1}^{N} L_n = \sum_{n=1}^{N} L(\hat{Y}, Y) \quad (3)$$

By performing the backpropagation process, the parameters for each neuron are updated after each iteration. For the sequence inputs, RNN processes one element each time by using cyclic and recurrent units.

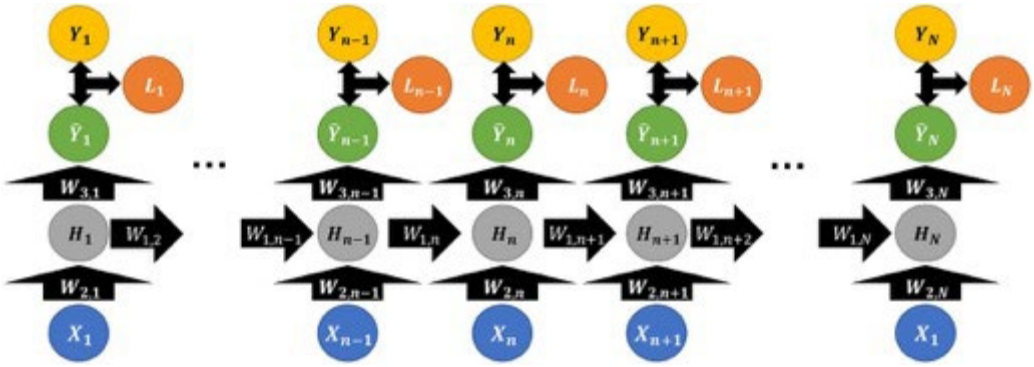


Figure 2.13 Schematic illustrations for RNN. Items X, Y, and W have the same meaning as Equation (2); while L_i stands for the loss function for a given step (Tang B et al., 2019).

However, the traditional RNN can not solve the long-time dependence problems very well, thus GRU and LSTM were proposed to address this issue. While GRU is a simplified and efficient version of LSTM. The architecture and information flow of LSTM is shown in Figure 2.14. In Figure 2.14, the information flow from past to new features for the input gate is shown as the yellow track. The green track denotes both an input gate and hidden layer neurons. While the blue track denotes the output gate which is influenced by the yellow track.

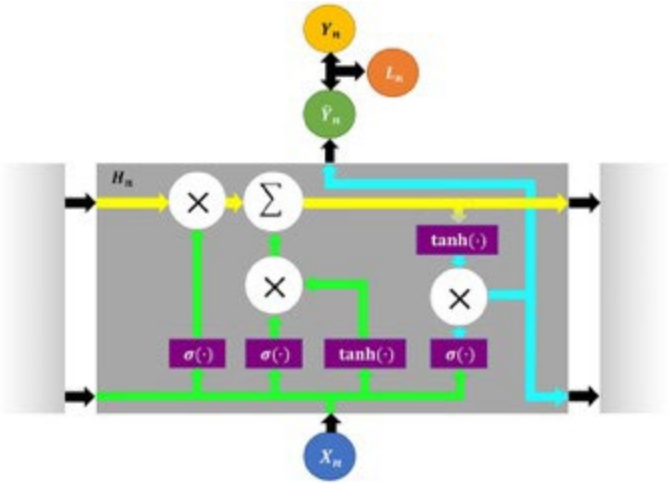


Figure 2.14 Schematic illustrations for LSTM. Items X, Y, and W have the same meaning as above (Tang B et al., 2019).

For the applications of deep-learning approaches to proteomics, the predictions based on the peptide sequences have been studied intensively, such as the retention time prediction (Ma et al., 2018; Yang et al., 2020) and fragment ion intensities (Zhou et al., 2017; Gessulat et al., 2019). The identifications of peptides and proteins are greatly improved by deep learning, such as the acquirement of features for LC-MS (Kantz et al., 2019). Moreover, de novo sequencing is also facilitated by the application of deep learning, such as the DeepNovo used for de novo sequencing by combining CNN and RNN (Tran et al., 2017). In my doctoral research, all the above deep-learning techniques were used for MS-based proteomics.

3. Summary

Genomics, transcriptomics, and proteomics are fundamental blocks that shaped modern biology. High throughput and large-scale techniques, such as next-generation sequencing (NGS) and mass spectrometry (MS), have been widely used in the life sciences. Due to the complexity of these data, the analysis needs to be done by sophisticated bioinformatic methods. During my doctorate research, I developed new computational methods and applied new strategies to advance the research in genomics, transcriptomics, and proteomics.

NGS has brought tremendous and numerous changes to genomic research by providing higher sensitivity, sequencing depth, and throughput compared with traditional sequencing methods, such as Sanger sequencing, qPCR, and microarrays. Benefiting from the advantages of NGS technology, RNA-seq has been widely used for the qualitative and quantitative analysis of genome wide changes in gene expression. Chromatin immunoprecipitation sequencing (ChIP-seq) as another popular application of NGS provides an efficient way to analyze the interaction between proteins and DNA. During my doctoral studies, I used these techniques to uncover the mechanisms behind the hybrid incompatibility between *Drosophila melanogaster* and *D.simulans.*

The loss of HMR in *D.melanogaster* leads to mitotic defects, increased transcription of transposable elements, and deregulated heterochromatic genes. Through the genome-wide analysis of HMR's localization by ChIP-seq, I found that genomic insulator sites bound by HMR can be grouped into two clusters. One set is composed of gypsy insulators, whereas the other is bordered by HP1a-bound areas of active genes. In *Hmr* mutant flies, the transcription of genes belonging to the latter group is severely disrupted in larval tissue and ovaries. These findings showed a novel connection between HMR and insulator proteins, indicating a possible role for genome organization in species development.

Beyond the study of particular genes, and RNA transcripts, I also dedicated my work towards improving proteomic research by accurately predicting fragmentation patterns of peptides in tandem mass spectrometry (MS) with deep-learning.

MS is an important and powerful technology for proteomic research. In recent years, with the development of both theoretical and industrial technology and methods, the research scope of proteome has improved at an unprecedented speed. SWATH-MS is a mass spectrometric technique that combines the advantages of targeted data analysis and combines

it with the speed of time-of-flight (ToF) mass spectrometers to improve peptide quantitation and identification in a data-independent acquisition (DIA) mode. SWATH-MS can analyze proteomes on a much larger scale than traditional methods such as data-dependent acquisition (DDA), parallel reaction monitoring (PRM), or selected reaction monitoring (SRM) due to its increased reproducibility and accuracy. Moreover, SWATH-MS shows a significant increase in the detection rates of peptides and proteins along with higher accurate quantifications.

However, mass spectra data generated by SWATH-MS showed a higher complexity compared to the traditional DDA mass spectrometry method. Therefore, more accurate data analysis strategies were required to address this complexity. At the beginning of my doctorate, SWATH-MS relied entirely on fragment libraries generated by DDA experiments, which greatly limited the number of detectible and identifiable peptides. Hence, the extension of the search space is crucial to improve both identification and quantitation on a proteome-wide scale, especially for SWATH-MS analysis.

With the development of new computational approaches to complex problems, more and more biological questions were addressed successfully. In this work, we applied such advanced methods to build a prediction framework that is composed of several tools: dpMS for mass spectra prediction, dpRT for retention time prediction, and dpMC for missed tryptic cleavages prediction, along with other new strategies to improve the effective search space for SWATH-MS in high quality. With the *in-silico* library, we can identify proteins and peptides that exceed the experimental library limitation. We demonstrated the reproducibility and efficiency of dpSWATH across different organisms from *D. melanogaster* and *H. sapiens* on a Q-TOF instrument. With different experimental conditions, dpSWATH can build highly reliable theoretical libraries for SWATH-MS analysis. Consequently, the new searching space has improved both sensitivity and specificity for SWATH-MS analysis at a higher level.

Within this thesis I summarize three publications I (co)authored: one of which is on the analysis of next generation sequencing, and the other two are on the work of predictions for mass spectrometry, which are listed above.

4. Zusammenfassung

Genomik, Transkriptomik und Proteomik sind grundlegende Bausteine, die die moderne Biologie geprägt haben. Hochdurchsatz- und groß angelegte Techniken wie die Hoch-Durchsatz Sequenzierung (NGS) und die Massenspektrometrie (MS) werden in den Biowissenschaften in großem Umfang eingesetzt. Aufgrund der Komplexität dieser Daten muss die Analyse mit ausgefeilten bioinformatischen Methoden durchgeführt werden. Während meiner Doktorarbeit habe ich neue Methoden entwickelt und neue Strategien angewandt, um die Forschung in den Bereichen Genomik, Transkriptomik und Proteomik voranzutreiben.

NGS hat die Genomforschung in vielerlei Hinsicht verändert, da es im Vergleich zu herkömmlichen Sequenzierungsmethoden wie Sanger-Sequenzierung, qPCR und Microarrays eine höhere Empfindlichkeit, Sequenzierungstiefe und einen höheren Durchsatz bietet. RNA-seq profitiert von den Vorteilen der NGS-Technologie und wurde in großem Umfang für die qualitative und quantitative Analyse genomweiter Veränderungen der Genexpression eingesetzt. Die Chromatin-Immunpräzipitations-Sequenzierung (ChIP-seq), eine weitere Anwendung von NGS, bietet eine effiziente Möglichkeit zur Analyse der Interaktion zwischen Proteinen und DNA. Während meines Promotionsstudiums habe ich diese Techniken eingesetzt, um die Mechanismen hinter der Hybridinkompatibilität zwischen Drosophila melanogaster und Drosophila simulans aufzudecken.

Der Verlust von HMR in D. melanogaster führt zu mitotischen Defekten, erhöhter Transkription von transposablen Elementen und deregulierten heterochromatischen Genen. Durch die genomweite Analyse der HMR-Lokalisierung mittels ChIP-seq habe ich herausgefunden, dass genomische Isolatorstellen, die von HMR gebunden werden, in zwei Gruppen unterteilt werden können. Die eine Gruppe besteht aus Gypsy-Insulatoren, während die andere von HP1a-gebundenen Bereichen aktiver Gene begrenzt wird. Bei Hmr-mutierten Fliegen ist die Transkription von Genen, die zur letzteren Gruppe gehören, im Larvengewebe und in den Eierstöcken stark gestört. Diese Ergebnisse zeigen eine neuartige Verbindung zwischen HMR und Isolatorproteinen, was auf eine mögliche Rolle der Genomorganisation bei der Entwicklung von Arten hinweist.

Neben der Untersuchung bestimmter Gene und RNA-Transkripte widmete ich meine Arbeit auch der Verbesserung der Proteomforschung durch die genaue Vorhersage von

Fragmentierungsmustern von Peptiden in der Tandem-Massenspektrometrie (MS) mit Hilfe von Deep-learning.

Die MS ist eine wichtige und leistungsfähige Technologie in der Proteomforschung. In den letzten Jahren hat sich der Umfang der Proteomforschung durch die Entwicklung sowohl theoretischer als auch experimenteller Technologien und Methoden dramatisch verbessert. SWATH-MS ist eine massenspektrometrische Methode, die die Vorteile der gezielten Untersuchung von individuellen Analyten mit der Geschwindigkeit von Flugzeit-Massenspektrometern kombiniert, um die Quantifizierung und Identifizierung von Peptiden in einer datenunabhängigen Messung (DIA) zu verbessern. SWATH-MS kann Proteome in einem viel größeren Umfang analysieren als herkömmliche Methoden wie die datenabhängige Messung (DDA), die parallele Messung von Fragmentübergängen (PRM) oder die Messung ausgewählter Fragmente (SRM), da es eine höhere Reproduzierbarkeit und Genauigkeit bietet. Darüber hinaus zeigt SWATH-MS eine signifikante Steigerung der Detektionsraten von Peptiden und Proteinen zusammen mit einer höheren Quantifizierungsgenauigkeit.

Die mit SWATH-MS erzeugten Massenspektren sind jedoch komplexer als bei der herkömmlichen DDA-Massenspektrometrie. Daher sind genauere Datenanalysestrategien erforderlich, um diese Komplexität zu bewältigen. Zu Beginn meiner Promotion stützte sich SWATH-MS ausschließlich auf Fragmentbibliotheken, die aus DDA-Experimenten stammten, was die Zahl der nachweisbaren und identifizierbaren Peptide stark einschränkte. Durch die von mir entwickelte Methode konnte ich den Suchraum deutlich erweitern, um sowohl die Identifizierung als auch die Quantifizierung auf proteomweiter Ebene zu verbessern.

Mit der Entwicklung neuer computergestützter Ansätze für komplexe Probleme konnten immer mehr biologische Fragen erfolgreich beantwortet werden. Die von mir entwickelte bioinformatische Methode besteht aus mehreren Komponenten: dpMS für die Vorhersage von Fragmentspektren, dpRT für die Vorhersage von Retentionszeiten und dpMC für die Vorhersage tryptischer Spaltungen, um den effektiven Suchraums für SWATH-MS zu erweitern. Mit der so (in-silico) generierten Bibliothek von Fragmentspektren konnte ich deutlich mehr Proteine und Peptide identifizieren. Ich konnte die Reproduzierbarkeit und Effizienz von dpSWATH durch Messung von Proteomen aus verschiedenen Organismen auf einem Q-TOF-Instrument nachgeweisen. Unter verschiedenen Versuchsbedingungen kann dpSWATH sehr zuverlässige theoretische Bibliotheken für die SWATH-MS-

Analyse erstellen und damit die Sensitivität als auch die Spezifität der SWATH-MS-Analyse verbessern.

In dieser Arbeit fasse ich drei Publikationen zusammen, die ich (mit-)verfasst habe: eine davon befasst sich mit der Analyse von Next Generation Sequencing, die beiden anderen mit der Arbeit an Vorhersagen für die Massenspektrometrie, die oben aufgeführt sind.

5. The Drosophila speciation factor HMR localizes to genomic insulator sites (Paper I)

The lethal interaction of the proteins encoded by the *Hmr* and *Lhr* genes can cause hybrid incompatibility between *Drosophila melanogaster* and *D.simulans*. HMR plays a key role in the mitotic process. In this study, we analyzed the function of HMR by genome-wide localization and chromatin immunoprecipitation. The result implicates genome organization playing a potential role in the formation of species by analyzing the connection between HMR and insulator proteins.

Thomas Andreas Gerland, **Bo Sun**, Pawel Smialowski, Andrea Lukacs, Andreas Walter Thomae, and Axel Imhof. 2017. "The Drosophila Speciation Factor HMR Localizes to Genomic Insulator Sites." *PLOS ONE* 12 (2): e0171798. doi:10.1371/journal.pone.0171798.

PLOS ONE

RESEARCH ARTICLE

The *Drosophila* speciation factor HMR localizes to genomic insulator sites

Thomas Andreas Gerland[1,2], Bo Sun[1], Pawel Smialowski[1,3], Andrea Lukacs[1], Andreas Walter Thomae[1,4], Axel Imhof[1,2]*

1 Biomedical Center, Histone Modifications Group, Department of Molecular Biology, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany, **2** Center for Integrated Protein Science Munich (CIPSM), Ludwig-Maximilians-Universität München, Munich, Germany, **3** Biomedical Center, Core Facility Computational Biology, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany, **4** Biomedical Center, Core Facility Bioimaging, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany

* imhof@lmu.de

OPEN ACCESS

Citation: Gerland TA, Sun B, Smialowski P, Lukacs A, Thomae AW, Imhof A (2017) The *Drosophila* speciation factor HMR localizes to genomic insulator sites. PLoS ONE 12(2): e0171798. doi:10.1371/journal.pone.0171798

Editor: Barbara Jennings, Oxford Brookes University, UNITED KINGDOM

Received: December 9, 2016

Accepted: January 26, 2017

Published: February 16, 2017

Data Availability Statement: ChIP-Seq data from this study are publicly available at NCBI GEO with the accession number (GSE86106).

Funding: The work was funded by a grant from the Deutsche Forschungsgemeinschaft (DFG) to Andrea Lukacs (QBM) and Axel Imhof (IM/9-1). B. S. was funded by the Chinese Scholarship Council. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Hybrid incompatibility between *Drosophila melanogaster* and *D. simulans* is caused by a lethal interaction of the proteins encoded by the *Hmr* and *Lhr* genes. In *D. melanogaster* the loss of HMR results in mitotic defects, an increase in transcription of transposable elements and a deregulation of heterochromatic genes. To better understand the molecular mechanisms that mediate HMR's function, we measured genome-wide localization of HMR in *D. melanogaster* tissue culture cells by chromatin immunoprecipitation. Interestingly, we find HMR localizing to genomic insulator sites that can be classified into two groups. One group belongs to *gypsy* insulators and another one borders HP1a bound regions at active genes. The transcription of the latter group genes is strongly affected in larvae and ovaries of *Hmr* mutant flies. Our data suggest a novel link between HMR and insulator proteins, a finding that implicates a potential role for genome organization in the formation of species.

Introduction

Biodiversity is the result of the emergence and the extinction of species. New species form by pre- and post-zygotic isolation mediated by genetic incompatibility [1]. One of the best characterized examples of hybrid incompatibility is the gene pair *Hybrid male rescue* (*Hmr*) and *Lethal hybrid rescue* (*Lhr*). *Hmr* and *Lhr* cause hybrid incompatibility between the closely related fly species *Drosophila melanogaster* and *D. simulans*. *Hmr* diverged in both *Drosophila* sibling species under positive selection [2]. HMR and LHR from both species interact physically and localize predominantly to centromeric regions [3]. A reduction of HMR expression results in a misregulation of transposable elements, satellite DNAs and heterochromatic genes [3–5]. The major difference between HMR and LHR in *D. melanogaster* and *D. simulans* is their substantial difference in protein amounts [3,6], which has been proposed to result in a lethal gain of function in male hybrids [3]. High levels of HMR and

Abbreviations: BEAF-32, Boundary element associated factor 32; ChIP, Chromatin immunoprecipitation; CID, Centromer identifier; CP190, Centrosomal Protein 190; CRISPR, clustered, regularly interspaced, short palindromic repeats; CTCF, CCCTC binding factor; DGRC, *Drosophila* genomics resource center; gRNA, guide RNA; GST, Glutathione S-transferase; *gtwin*, *gypsy-twin*; HMR, Hybrid male rescue; HP1a, Heterochromatin Protein 1a; LHR, Lethal hybrid rescue; Mod(mdg4), Modifier of mdg4; RNAi, RNA interference; Su(Hw), Suppressor of Hairy wing; TSS, Transcription start site.

LHR in hybrids and overexpression of these proteins in pure species lead to an increased number of binding sites of the complex [3]. Such spreading phenomena based on protein amount have been observed for several chromatin-associated complexes such as the dosage compensation complex [7,8], the polycomb complex [9] or components of pericentromeric heterochromatin [10,11]. In most cases, the precise mechanisms for targeting and spreading are not fully understood. Interestingly, several of the components involved in these processes show signs of adaptive evolution and differ substantially even in very closely related organisms [12–14]. This observation has spurred a model of a dynamic genome that drives the adaptive evolution of chromatin-associated factors [15].

Eukaryotic genomes of closely related species differ mostly in the amount and sequence of repetitive DNA [16–18]. This DNA is often derived from transposable elements, which are highly mutagenic and are therefore under tight transcriptional control by the cellular machinery. During evolution transposons or transposon-derived sequences occasionally adopted structural or novel *cis*-regulatory functions, thereby contributing to the evolution of new, species-specific, phenotypic traits [19–21]. Genomic insulators are a particular class of such novel, fast evolving, *cis*-regulatory elements that show signs of transposon ancestry [22,23]. A strong expansion of these elements is observed in arthropods, which also experienced a successive gain in the number of insulator binding proteins during evolution [24]. In fact, the *Drosophila* genome harbours a large variety of insulator proteins such as CTCF, BEAF-32, Su(Hw), Mod(mdg4) and CP190, which all affect nuclear architecture [25]. Different *Drosophila* species underwent multiple genomic rearrangements and transposon invasions [26,27], which presumably resulted in an adaptive response of regulatory DNA binding factors to maintain spatial and temporal gene expression. For example, binding sites for the insulator proteins BEAF-32 and CTCF show a high degree of variability when compared among very closely related species [26,27]. The gain of new insulator sites is associated with chromosome rearrangements, new born genes and species-specific transcription regulation [19,23]. Similar to insulator proteins, which tend to cluster in specific nuclear regions [28], the speciation factor HMR clusters at centromeres or pericentromeric regions in diploid cells [3,6] but is also detected at distinct euchromatic regions along the chromosome arms in polytene chromosomes [3]. A unifying feature for many of these sites is their close proximity to binding sits of the Heterochromatin Protein 1 (HP1a), a HMR interactor and a well-characterized heterochromatic mark.

Various studies describe HMR's localization to heterochromatin, but the molecular details on HMR's binding sites and its recruitment to these sites are not well understood. To get new insights into HMR's association to chromatin, we measured HMR's genome-wide localization by chromatin immunoprecipitation (ChIP) in the *D. melanogaster* embryonic S2 cell line. We demonstrate an extensive colocalization of HMR with a subset of insulator sites across the genome. HMR's binding to genomic *gypsy* insulators, which constitute the major group of its binding sites, is dependent on the residing insulator protein complex. In a second group, HMR borders heterochromatin together with the insulator protein BEAF-32. In agreement with previous low-resolution techniques in cell lines and fly tissue [3], these binding sites are enriched at pericentromeric regions, the cytological region 31 on the 2nd chromosome and the entire 4th chromosome. At most of these sites, HMR associates to the promoters of actively transcribed genes. Interestingly, these genes code for transcripts that have been reported to be downregulated in *Hmr* mutant larvae and ovaries. Altogether, our data provide evidence for a functional link between HMR and insulator proteins, which potentially results in hybrid incompatibilities due to the adaptive evolution of these genome-organizing complexes.

Materials and methods

Cell culture and RNAi

D. melanogaster S2-DRSC cells were obtained from the DGRC and grown at 26˚C in Schneider's *Drosophila* medium (Invitrogen) supplemented with 10% fetal calf serum and antibiotics (100 units/mL penicillin and 100 μg/mL streptomycin).

For RNAi experiments cells were incubated in serum-free medium containing 10 mg/mL dsRNA. After 1 hr of incubation, the serum-containing medium was supplied. Samples were taken after 7 days. The dsRNA was prepared using the MEGAScript T7 Transcription Kit (Thermo Fisher Scientific) following the manufacturers instructions with primers listed in S1 Table.

Chromatin immunoprecipitation, Real-Time PCR and sequencing

For chromatin immunoprecipitation (ChIP) cells were crosslinked with 1% formaldehyde for 5 min at room temperature. Upon cell lysis, protease inhibitors and proteasome inhibitor MG-132 (Enzo Life Sciences) were applied. The chromatin was isolated and sheared with adaptive focused acoustics (Covaris) to an average size of 200 base pair (bp). For each ChIP reaction, chromatin isolated from 1–2 x 10^6 cells was incubated with following antibodies precoupled to Protein A/G Sepharose: rat anti-HMR 2C10 (RRID: AB2569849) [3] with rabbit IgG anti-rat IgG (RRID: AB2339804), mouse anti-HP1a C1A9 (RRID: AB528276) [29], rabbit anti-H3 (RRID: AB302613), rabbit anti-H3K9me3 (RRID:AB2532132) and mouse anti-FLAG (RRID: AB262044). Real-Time PCR was performed with Fast SYBR Green master mix (Applied Biosystems) using a LightCycler 480 II (Roche). For deep sequencing, all libraries were prepared using MicroPlex (Diagenode) or NEBNext (NEB) Library Preparation kit and single-end, 50 bp sequenced with the Illumina HiSeq2000. An overview of all ChIP-Seq samples used and the number of uniquely aligned sequence reads is provided as S2 Table. A list of HMR peaks used for further analyses is provided as S3 Table. All sequencing data are publicly available as described below.

Data analysis

The raw reads were aligned to the *D. melanogaster* genome assembly (dm3) using Bowtie 2.2.6 with unique mapping criteria and exclusion of chromosome Uextra [30]. The raw read quality was accessed using FASTQC 11.5 [31] and read filtering was performed using FastX 0.0.13 [32]. Sequencing tracks were visualized using IGB [33] and IGV [34] genome viewers. Peak calling, motif search and peak annotation were performed using HOMER 4.8 with peak size of 200 bp [35] and ChIPseeks implementation of HOMER [36]. For downstream analysis, peaks identified in two out of three biological replicates were taken. Downstream analysis steps were performed using Python and R and parts of data preprocessing was done using ChipPeakAnno [37]. For repeat analysis, reads from ChIP-Seq experiments were mapped to RepBase version 19.10 [38] using Bowtie [30]. Only unique reads were kept for analysis. For each repetitive element, the log2 fold change was calculated. Following genome-wide binding data sets derived from S2 cells (unless stated otherwise) were used: CP190, Su(Hw), CTCF and mod(mdg4) from GEO GSE41354 [39], BEAF-32 from GEO GSE32815 [40]. RNA expression data for untreated S2 cells was taken from GEO GSE46020. For *D. melanogaster* larvae and ovaries, RNA-Seq data were taken from NCBI BioProject PRJNA236022 [4] and analyses were performed with cuffdiff 2 [41]. An extended description of the bioinformatics tools and methods used is provided in S1 Methods.

Western blot analysis

Samples were boiled in loading buffer, separated on SDS-PAGE gels (Serva), processed for western blot using standard protocols and detected using rat anti-HMR 2C10 (1:20) (RRID: AB2569849), rabbit anti-CP190 (RRID: AB2615894) [42], rabbit anti-H3K9me3 (1:2000) (RRID:AB2532132) and mouse anti-Tubulin (1:800) (RRID: AB2241150) antibodies. Secondary antibodies included sheep anti-mouse (1:5000) (RRID: AB772210), goat anti-rat (1:5000) (RRID: AB772207), donkey anti-rabbit (1:5000) (RRID: AB772206) coupled to horseradish peroxidase.

Data access

ChIP-Seq data from this study are publicly available at NCBI GEO (GSE86106).

Results

Genome-wide binding map of HMR in *D. melanogaster*

Immunohistological studies revealed a binding of HMR to centromeric or pericentromeric regions in diploid cells and to several euchromatic and telomeric regions in polytene chromosomes [3,6]. However, detailed information on HMR's binding to chromatin was so far lacking. To better understand the molecular mechanisms that govern HMR's binding within the genome, we mapped the genomic binding sites of HMR in cultured *D. melanogaster* S2 cells. We used a highly specific monoclonal antibody against HMR [3] to purify associated chromatin followed by next generation sequencing (ChIP-Seq) and derived a set of 794 HMR binding sites, which were present in at least two out of three biological replicates (Fig 1A). A composite plot of all HMR binding sites found in the genome revealed a sharp peak of HMR binding with a width of approximately 200 nucleotides, which is reminiscent of sequence specific transcription factors (Fig 1B). To validate the identified HMR binding sites, we applied multiple strategies. First, we measured enrichment of the HMR binding sites in ChIP experiment using an anti-FLAG antibody, an epitope that is not expressed in wild type cells (Fig 1B and S1A Fig). Second, we performed RNAi knock-down experiments to reduce HMR protein level and compared the enrichment of HMR between HMR RNAi treated cells and Control (Ctrl) RNAi treated cells. Although we observe an overall reduction of HMR binding at most HMR peaks (S1B Fig), we rarely see a complete loss of binding despite the high efficiency of the HMR knock-down. This apparent discrepancy suggests that chromatin-bound HMR is rather resistant towards a RNAi-mediated removal. The existence of such RNAi-resistant binding sites in ChIP experiments has been observed before and was attributed to high-affinity binding sites [43] or an incomplete removal of the chromatin-bound factors. Third, we used the CRISPR/*cas9* system to edit the HMR locus in S2 cells such that the cell line exclusively expresses an HMR allele, which carries a double FLAG-tag at the C-terminus. ChIP-qPCR using HMR and FLAG antibody in wild type and HMR-$Flag_2$ expressing cells showed specific and reproducible enrichment of HMR at selected HMR binding sites (S1C Fig).

We find HMR binding sites on all chromosomes and distributed along the whole chromosome arms with a marked increase in peak density at pericentromeric regions and at the 4th chromosome where we also observe a higher density of binding sites for HP1a, a known interaction partner of HMR [3,4] (Fig 1C and S1D Fig). Unfortunately, the centromeric regions are not present in the current *Drosophila* genome assembly, preventing read mapping and analysis in this area of the genome. However, the increased number of peaks at pericentromeric regions (S1D Fig) is consistent with the strong centromeric HMR signal we previously observed when staining S2 cells with an anti-HMR antibody [3]. Besides the pericentromeric region, we also

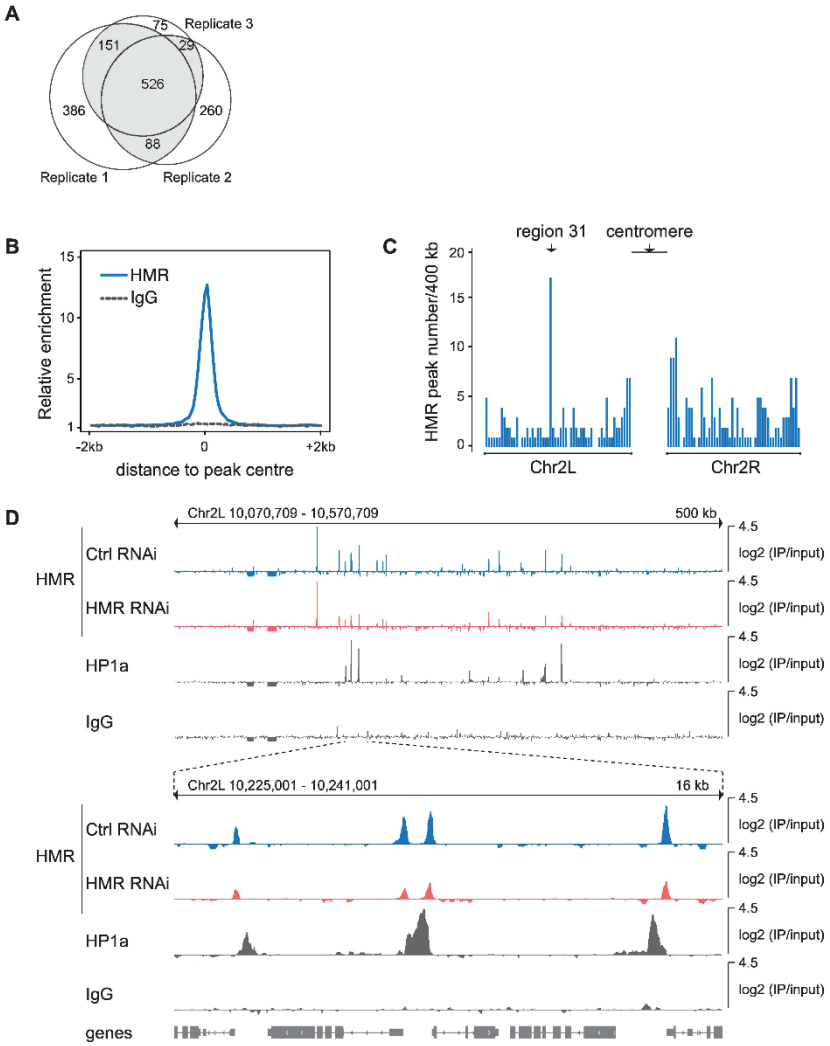


Fig 1. Identification of HMR binding sites in *D. melanogaster* S2 cells. (A) Venn diagram of HMR peaks showing the number of peaks identified in three independent biological replicates. Peaks identified in at least two out of three replicates were used for further analysis and are highlighted in *grey*. **(B)** Composite analysis of HMR and control IgG (anti-FLAG) ChIP signals at genomic HMR peak positions. **(C)** Histogram of HMR peak density across the left arm (2L) and right arm (2R) of the 2nd chromosome. The cytological region 31 and centromere-proximal regions are indicated **(D)** Genome browser view of HMR, HP1a and control IgG (anti-FLAG) ChIP signals at region 31. HMR ChIP signals obtained upon knock-down using control RNAi and HMR RNAi are shown with the same amplitude.

doi:10.1371/journal.pone.0171798.g001

observe a strong clustering of HMR peaks at the cytological region 31 on the left arm of the second chromosome where we also see HMR binding in polytene chromosomes [3]. Interestingly, HMR binding sites within these regions do not completely overlap with HP1a bound regions but rather localize at their edges (Fig 1D).

HMR binding sites largely overlap with genomic insulator sites

We next asked whether HMR binding sites are enriched for specific DNA sequence motifs. A motif analysis of HMR-bound regions revealed three DNA sequence motifs that were significantly enriched and present in up to 26% of all HMR peaks (Fig 2A). These motifs are highly related to the recognition motifs of the insulator DNA binding proteins Su(Hw) and BEAF-32 (S2A Fig), both containing a zinc-finger DNA-binding domain [44,45], suggesting that HMR binds to insulator regions. Indeed, we observe a substantial overlap of our HMR binding profiles with the published ones of known insulator proteins such as CP190, Mod(mdg4), Su (Hw), CTCF and BEAF32 [39,43] (Fig 2B and S2B Fig). Insulator binding sites can be subclassified depending on their composition of known insulator proteins [43]. One of the best characterized family of insulators are derived from the *gypsy* retrotransposon and are strongly bound by Su(Hw), Mod(mdg4) and CP190 [46–48]. Consistent with the strong enrichment of Su(Hw)-recognition motifs in the binding sites of HMR, we find about half of all HMR sites belonging to this *gypsy*-like family of insulators (Fig 2B). However, only 7% of all Su(Hw) binding sites and 11% of *gypsy*-like elements classified as bound by Su(Hw), Mod(mdg4) and CP190 are also bound by HMR.

Given the extensive colocalization of HMR with non-repetitive *gypsy*-like insulators (Fig 2C) and the effect of a *Hmr* mutations on the expression of retrotransposons [3,4], we wondered whether HMR is also enriched at repetitive DNA. We therefore mapped sequences obtained from our ChIP-Seq experiments using anti-HMR and anti-HP1a antibodies as well as published binding profiles for Su(Hw), Mod(mdg4)2.2 and CP190 [39] against the RepBase repeat database [38]. In agreement with previous studies we observe a strong enrichment of HP1a at the centromeric heterochromatin-associated Dodeca satellite (DMSAT6) [49] and the transposable elements Rt1a and Rt1b (DMRT1A, DMRT1B) [50] (S2C Fig). In contrast to HP1a, the only repetitive elements that show a substantial enrichment for HMR are the retrotransposons *gypsy*, and *gtwin* (Fig 3A). At these elements HMR binds together with Su(Hw), Mod(mdg4) and CP190 to the 5' insulator region (Fig 3B and S2D Fig).

A key element for the formation of insulator complexes at *gypsy*-like elements is the presence of the CP190 adaptor protein. A reduction of CP190 levels has been shown to strongly affect binding of insulator proteins to these elements but not to others [43]. To test whether CP190 also impacts the binding of HMR to *gypsy*-like binding sites, we performed RNAi knock-down experiments to reduce CP190 protein level (Fig 4A) and measured HMR binding. Strikingly, we observe a substantial reduction of HMR binding only for the *gypsy*-like group of binding sites (Fig 4B and 4C), suggesting that HMR's binding to the *gypsy*-like insulator class is indeed dependent on CP190. A HMR RNAi knock-down in contrast affects HMR binding equally in both classes (Fig 4D and S3 Fig). As insulator sites are known to contain less nucleosomes [51], nucleosome occupancy can serve as a proxy for insulator complex integrity at these sites [43]. We therefore performed a Histone H3 ChIP upon CP190 RNAi knock-down to monitor changes in insulator complex integrity [52]. Consistent with the importance of CP190 for maintaining the *gypsy* insulator, nucleosome occupancy only increases in the *gypsy*-like HMR binding sites (Fig 4C). Taken together, these results demonstrate an extensive colocalization of HMR with genomic insulator proteins, which play an important role in mediating its binding to chromatin.

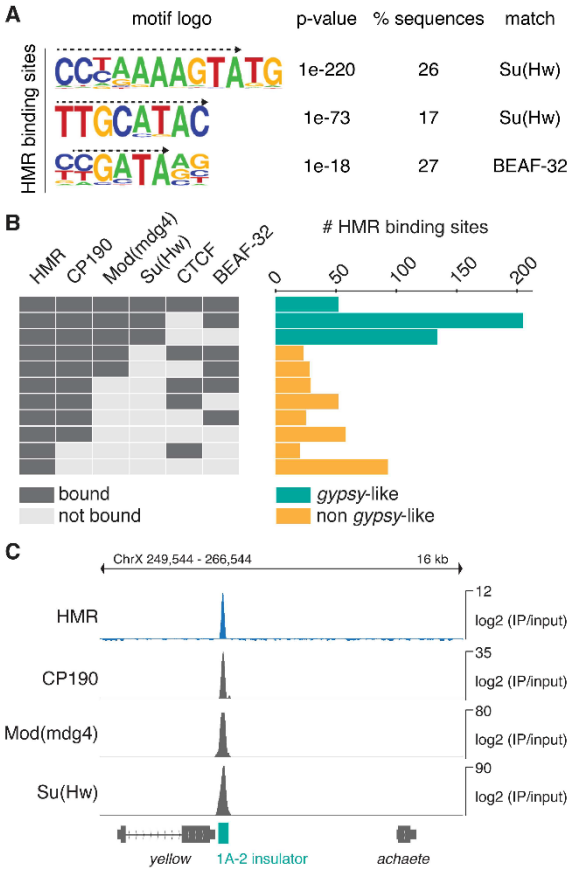


Fig 2. HMR localizes to genomic insulators and the *gypsy* transposon. (A) Sequence motifs identified within HMR peak regions. The corresponding motif logo, p-value of enrichment, percentage of regions with this motif and putative binding factors are indicated. Dashed arrows mark the sequence that matches the published binding sites of Su(Hw) and BEAF-32 (see also S2A Fig) **(B)** Peak overlap of HMR with peaks of the insulator proteins CP190, Mod(mdg4), Su(Hw), CTCF [39] and BEAF-32 [40]. The number of HMR peaks is indicated depending on their colocalization with known boundary factors. Groups with less than 11 members are not displayed. Su(Hw)-containing *gypsy*-like groups are depicted in *green*, non *gypsy*-like groups in *orange*. Combinations that contain less than ten HMR peaks are not shown. **(C)** Genome browser view of the Su(Hw) binding region 1A-2. ChIP signals of HMR and known *gypsy* binding factors are shown. The 1A-2 insulator is highlighted in *green*.

doi:10.1371/journal.pone.0171798.g002

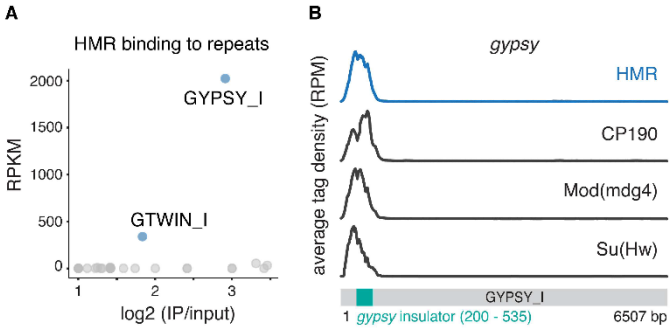


Fig 3. HMR localisation to repetitive elements. (A) HMR ChIP tag enrichment at repetitive DNA elements. To identify enriched sequences the enrichment (log2-fold) over input is plotted against the RPKM of an individual repeat sequence from RepBase. Repeats with less than 2-fold enrichment are not displayed. **(B)** ChIP tag density of HMR and the *gypsy*-insulator proteins CP190, Mod(mdg4), Su(Hw) [39] across the repetitive *gypsy* retrotransposon sequence. The *gypsy* insulator sequence at the 5' end is highlighted in *green*.

doi:10.1371/journal.pone.0171798.g003

HMR borders HP1a domains at active promoters

Although a large portion of HMR binding sites is associated with *gyspy and gypsy*-like insulators, there is a considerable number of HMR-bound peaks that do not localize with Su(Hw), Mod(mdg4) and CP190 (Fig 4B). We noticed that many of these non *gypsy*-like sites are in close proximity to HP1a bound regions (Fig 1D). Indeed, when we sorted all HMR peaks according to the presence of HP1a in their proximity, we observed an almost complete lack of *gypsy* insulator binding proteins at these sites (Fig 5A). Consistent with the lack of Su(Hw) binding to this class of HMR peaks, a motif search revealed no enrichment of the Su(Hw) recognition site among those peaks but rather an enrichment for BEAF-32 binding sites (Fig 5B). To better understand a possible role of HMR at these sites, which we termed class 1 binding sites, we analyzed them with regards to their annotation. Interestingly, almost all HP1a-associated HMR binding sites (90%) are in close proximity to transcriptional start sites (TSS), whereas the other HMR binding sites show a somewhat broader distribution among various functional elements (Fig 5C). Strikingly, HMR binds very closely to the TSS at the boundary between HP1a containing domains and the gene body (S4A Fig). The genes in proximity of these HMR binding sites are classified as transcriptionally active suggesting that HMR might prevent the repressive influence of HP1a on neighbouring genes (Fig 5D). To investigate whether HMR loss has an impact on HP1a or H3K9me3 domains at these genomic regions, we performed HP1a ChIP and H3K9me3 ChIP upon HMR knockdown. However, we could not confirm extensive spreading of the heterochromatin marks HP1a or H3K9me3 after HMR loss (S4C and S4D Fig). Nevertheless, the genes associated with this class of HMR binding sites are transcriptionally down-regulated in *Hmr* mutant larvae and ovaries (Fig 5E, S4E Fig and [5]). This seems to be particularly important within regions that are rich in heterochromatin such as the 4th chromosome or the pericentromeric regions where we find the class of HP1a-associated binding sites highly enriched (Fig 5F). In summary, we can classify HMR's genomic binding sites into two groups: One being associated with *gypsy* insulators, and another one associated with active promoters in pericentromeric heterochromatin where HMR borders

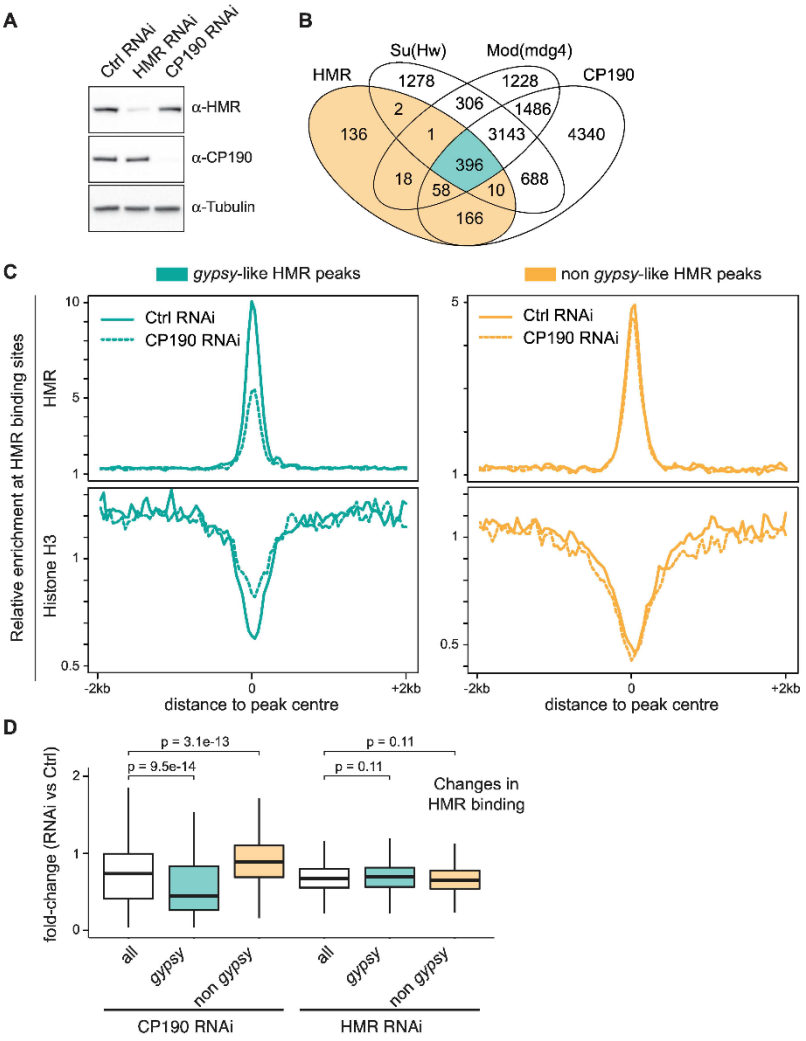


Fig 4. HMR genomic localization to *gypsy*-like insulator sites is dependent on CP190. (A) Western Blot of cell lysates after treatment with specific and control dsRNA shows an efficient knock-down of HMR and CP190. **(B)** Venn diagram of the overlap between HMR, CP190, Mod(mdg4) and Su(Hw) peaks [39] classifying HMR peaks as *gypsy*-like (highlighted in *green*) and non *gypsy*-like (highlighted in *orange*). **(C)** Composite analysis of HMR ChIP signals and Histone H3 ChIP signals at genomic HMR peak positions according to the groups defined in **(B)**. **(D)** Quantification of the fold-change of HMR ChIP enrichment upon CP190 RNAi and HMR RNAi. Box plots represent the fold-change of normalized HMR ChIP tag number aligned to 200 bp wide HMR peak regions. Peak regions with less than 50 aligned tags were excluded from the analysis. Significance of difference was estimated with p-values calculated with Wilcoxon rank sum test [72].

doi:10.1371/journal.pone.0171798.g004

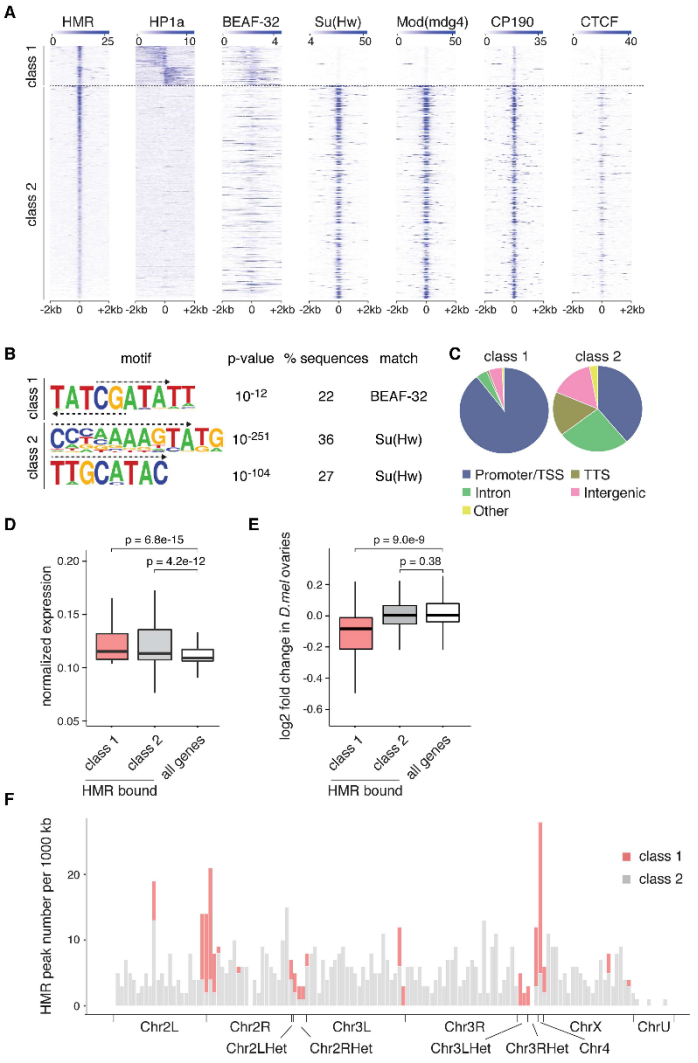


Fig 5. HMR borders HP1a together with BEAF-32 at the TSS of actively transcribed genes and enhances their transcription. (A) Heatmaps of HMR, HP1a, BEAF-32 [40], Su(Hw), Mod(mdg4), CP190 and CTCF [39] ChIP signals. All signals are centered around the HMR binding sites, clustered according to adjacent HP1a signals and sorted by HMR intensity. **(B)** Sequence motifs identified within HMR peak regions from class 1 and class 2 based on HOMER motif analysis. The corresponding motif logo, p-value of enrichment, percentage of regions with this motif, and putative binding factors are indicated. Dashed arrows

mark the sequence that matches the published binding sites of Su(Hw) and BEAF-32 (see also S2A Fig). **(C)** Distribution of class 1 and class 2 HMR peaks among various genomic landmarks. **(D)** Box plot showing the normalized RNA expression of all genes and HMR-bound genes (promoter/TSS annotated) in class 1 and in class 2. S2 cells RNA expression levels were used according to [73]. Significance of difference was estimated with p-values calculated with Wilcoxon signed rank test [72]. **(E)** Box plot showing the log2 fold change of protein coding gene transcripts of all analyzed genes and HMR-bound genes (promoter/TSS annotated) in class 1 and in class 2 comparing *Hmr* mutant against wild type flies. The RNA-Seq data comes from experiments done in *D. melanogaster* ovaries [4]. Significance of difference was estimated with p-values calculated with Wilcoxon rank sum test [72]. For both box plots the box represents the interval that contains the central 50% of the data with the line indicating the median. The length of the whiskers is 1.5 times the interquartile distance (IQD). **(F)** Histogram showing HMR peak density across the annotated *D. melanogaster* genome. Class 1 HMR binding sites are enriched at region 31, centromere-proximal regions and the 4th chromosome.

doi:10.1371/journal.pone.0171798.g005

HP1a-containing chromatin regions together with BEAF-32 and potentially promotes gene transcription.

Discussion

HMR localizes to centromeric and pericentromeric regions in *D. melanogaster* cell lines as well as in mitotically dividing embryonic cells where it has been suggested to act as a repressor of transposable elements [3–5]. Mutations in *Hmr* lead to overexpression of satellite DNA and transposable elements in ovaries and larvae [4]. Such a derepression is also observed in hybrid flies [53], where HMR and LHR levels are higher than the ones in pure species and result in a widespread distribution of the HMR/LHR complex [3]. To better understand the targeting principles that mediate HMR binding within the *D. melanogaster* genome, we wondered whether we could identify HMR binding sites by applying ChIP-Seq in the *D. melanogaster* S2 cell line. Combining this approach with RNAi mediated knockdown experiments we uncover a strong colocalization of HMR with *gypsy* insulator binding sites and demonstrate that HMR binding to these sites depends on the presence of the residing insulator protein complex. Notably, HMR associates only with a subset of all Su(Hw) binding sites, but almost all those sites can be classified as *gypsy*-like sites bound by CP190 and mod(mdg4) in addition to Su(Hw).

Besides dispersed binding of HMR at genomic *gypsy* insulator sites along the chromosome arms, we observe dense clusters of HMR binding sites around the centromere and on the 4th chromosome where it potentially serves to separate HP1a binding domains from highly active genes. This dense clustering of binding sites around the centromere correlates well with the strong colocalization of HMR signals with the centromeric H3 variant CID in immunolocalization experiments [3]. Due to its biochemical interaction and partial colocalization with the heterochromatin protein HP1a in *Drosophila* embryos, HMR has been suggested to be a *bona-fide* heterochromatin component [3,4,6,54]. However, in contrast to HP1a, we detect very distinct HMR binding sites within the genome. When we find HMR close to an HP1a binding domain, it rather borders it than covering the whole domain. The sharp HMR binding signals and the fact that almost all euchromatic HMR binding sites contain putative insulator elements, suggest a role of HMR in separating chromatin domains. A distinct boundary that separates constitutive heterochromatin from the core centromere has also been postulated by Olszak and colleagues who suggest that transition zones between heterochromatin and euchromatin are hotspots for sites of CID misincorporation [55]. Unfortunately, centromeres are notoriously difficult to study by next generation sequencing due to their highly repetitive nature [56,57]. In addition, the microscopic resolution is not sufficiently high to allow a distinction between a binding to the core centromere chromatin and the chromatin immediately adjacent to it. Therefore, we cannot rule out the possibility that HMR binds large domains at the central region of the *Drosophila* centromere. However, the fact that the purification of

chromatin containing the centromeric H3 variant CID did not identify HMR [58], suggests that it may very well also form a boundary between pericentromeric heterochromatin and the core centromere. To which extent and by which mechanism HMR fulfils a functional role at these genomic sites remains to be elucidated.

The genomic sites, where we find HMR bound next to an HP1a domain, are highly enriched for recognition sites of the insulator protein BEAF-32. Interestingly, a depletion of BEAF-32 in S2 cells results in an increased rate of mitotic defects [45], which is very reminiscent of the phenotype detected when HMR is depleted [3]. Similarly to flies carrying a mutation in the *Hmr* gene, flies in which BEAF-32 is only contributed maternally have defects in female fertility [59,60]. BEAF-32's role in maintaining associated promoter regions in an environment that facilitates high transcription levels [61] has been suggested to be functionally relevant for this phenotype [45]. Strikingly, we find most HMR/BEAF-32 binding sites located between HP1a containing heterochromatin and the transcription start site of a highly active gene. HP1a chromatin might fulfil a repressive function at these genomic regions and HMR might block this repressive impact on the neighbouring gene body. However, we do not see extensive spreading of HP1a or H3K9me3 upon HMR knockdown suggesting that the repressive effect is not directly mediated by HP1a binding or the HMR knock down not efficient enough. As there is evidence that HP1a can also promote gene transcription [62], HMR may also function as a co-activator for HP1a. Currently, we therefore consider HMR binding next to HP1a containing chromatin as a unifying feature of transcriptionally affected genes but can only speculate about potential mechanism by which HMR exerts its function.

Although HMR depletion has a substantial effect on the transcription of multiple transposons, we find HMR only enriched at the 5' insulator region of the *gypsy* or *gtwin* retrotransposons and to similar sites within the genome that are presumably derived from these elements. These sites are occupied by insulator proteins Su(Hw), CP190 and Mod(mdg4) and often display enhancer blocking activity in transgenic assays [43,63–65]. Artificial targeting of HMR to DNA placed between an enhancer and a promoter of a reporter gene can block the transcription activity [3], suggesting that HMR may indeed play a role in setting up endogenous boundary elements. Similar to what is known for Su(Hw), HMR binding to this class of binding sites is dependent on the presence of the structural protein CP190, which has a key function in the stabilization of insulator protein complexes [22]. However, as we do not observe a strong physical interaction between CP190 and HMR, the loss of HMR binding upon a reduction of CP190 levels may also be the result of increased nucleosome occupancy. Such increase in Histone H3 binding cannot be observed upon HMR removal suggesting that HMR acts downstream of CP190. Interestingly, CP190 loss impairs HMR binding to *gypsy*-like insulator sites but has weak effect on HMR binding to sites containing BEAF-32 recognition motifs. Notably, in contrast to BEAF-32, CP190 is not required for oogenesis [66], suggesting that the lack of HMR binding to the class 1 sites may be responsible for the female sterility phenotype observed in *Hmr* mutant flies.

How can we integrate our findings with the lethal phenotype of increased HMR/LHR levels in male hybrids? It is tempting to speculate that multiple additional binding sites that are observed in hybrids and on polytene chromosomes of fly strains over-expressing HMR [3] constitute boundary regions. An increased binding to such boundaries, which have been shown to cluster and form aggregates *in vivo* [48,67,68], may trigger a massive change in nuclear architecture. In turn, this could indirectly activate multiple transposable elements similar to what is observed when centromere clustering is disturbed [69]. Such a disturbed nuclear architecture may then trigger the activation of a cell cycle checkpoint which has been previously suggested to be a major cause of hybrid lethality [70,71].

Altogether, our data provide a novel link between HMR and *cis*-regulatory elements bound by insulator proteins. We speculate that divergent evolution of such genomic elements and their corresponding binding factors in sibling species is triggering hybrid incompatibilities.

Supporting information

S1 Fig. Control experiments of HMR ChIP-Seq studies. (A) Venn diagram showing the lack of overlap between HMR peaks (peaks identified in at least two out of three independent biological replicates, highlighted in *grey*) and control IgG (anti-FLAG) ChIP peaks (pool of peaks from two independent biological replicates). **(B)** Changes in HMR ChIP enrichment upon HMR RNAi versus a control RNAi (GST) in two biological replicates. Each data point represents a mapped HMR peak. The scatter plot on the left displays fold changes of normalized HMR ChIP tag number mapped to a 200 bp HMR peak region in two biological replicates. Peak regions with less than 50 aligned tags were excluded from the analysis. The histogram on the right shows the frequency of peaks displaying a reduction of HMR binding upon knockdown. Shown are average values of replicate 1 and replicate 2. **(C)** ChIP-qPCR showing specific HMR enrichment at HMR binding sites. HMR ChIP is enriched for HMR binding sites in both wild type and HMR-$Flag_2$ expressing cells. FLAG ChIP is enriched for HMR binding sites only in HMR-$Flag_2$ expressing cells but not in wild type cells lacking the $Flag_2$ epitope. Data are represented as mean ± SD of three technical replicates. **(D)** Genome browser view of HMR ChIP, HP1a ChIP and control IgG ChIP signal at a large centromere-proximal region at the right arm of the 2nd chromosome.
(TIF)

S2 Fig. Overlap of HMR binding sites with known insulator regions and repetitive DNA. (A) Sequence motifs identified within Su(Hw) [39] and BEAF-32 [40] peak regions. The corresponding motif logo, p-value of enrichment and percentage of regions with this motif are indicated. Dashed arrows mark the sequence that matches the published binding sites of Su(Hw) and BEAF-32. (B) Genome browser view of ChIP signals showing combinatoric binding pattern for HMR and the insulator proteins CP190, Mod(mdg4), Su(Hw), CTCF [39] and BEAF-32 [40]. (C) Su(Hw) and HP1a ChIP tag enrichment at repetitive DNA elements. Each point in the scatter plot represents the enrichment (log2 fold) over input and the RPKM of an individual repeat from Repbase. Repeats with less than 2-fold enrichment are not displayed. (D) ChIP tag density of HMR and the *gypsy*-insulator proteins CP190, Mod(mdg4), Su(Hw) [39] across the *gypsy-1twin* repeat.
(TIF)

S3 Fig. Selective effect of CP190 RNAi on HMR binding to *gypsy*-like elements. Composite analysis of HMR ChIP signal and Histone H3 ChIP signal at genomic HMR peak positions according to the groups defined in Fig 4B. The ChIP signals were obtained upon control RNAi and HMR RNAi. The HMR ChIP signals are similarly affected in both groups, whereas Histone H3 ChIP signals are retained.
(TIF)

S4 Fig. Additional information for HP1a-associated HMR binding sites. (A) Composite analysis of HMR, HP1a and BEAF-32 ChIP signals at class 1 genomic sites relative to the transcriptional start site (TSS) and the gene body. Shown are normalised and scaled read density plots **(B)** Peak overlap of HMR with peaks of the insulator proteins CP190, Mod(mdg4), Su (Hw), CTCF [39] and BEAF-32 [40] for class 1 and for class 2 HMR binding sites. **(C)** Composite analysis of HP1a and H3K9me3 ChIP signals at class 1 HMR binding sites after HMR knockdown. Class 1 is defined in Fig 5A but oriented according to HP1a ChIP signal. **(D)**

Western Blot analysis on cell lysates to assay protein levels after HMR knockdown. Tubulin protein detection served as control. **(E)** Same as described in Fig 5E, but the RNA-Seq data comes from experiments done in *D. melanogaster* male larvae [4].
(TIF)

S1 Table. List of primers used for CRISPR/*cas9* genome editing, RNAi experiments and ChIP Real-Time PCR. List of primers used in this study. Primers used in ChIP Real-Time PCR were designed with help of Primer3.
(DOCX)

S2 Table. ChIP-Seq sample overview and number of uniquely aligned sequence reads. ChIP-Seq sample overview and number of uniquely aligned sequence reads. The percentage of uniquely mapped reads in ChIP-Seq experiments can largely vary and depends on the nature of the ChIPed protein. Proteins that bind repetitive regions (such as HMR or HP1a) give substantially lower percentages of uniquely mapped reads.
(DOCX)

S3 Table. HMR peaks used for downstream analysis. HMR peak list derived from HOMER peak calling on three biological replicates (Fig 1A). First three columns provide information on the peak position within the genome (chromosome, peak start and end using dm3), followed by peak annotation obtained from ChIPseeks implementation of HOMER (Fig 5C) and classification according to adjacent HP1a signals (Fig 5A).
(XLSX)

S1 Methods. Supporting information on methods. *Hmr* gene editing using CRISPR/*cas9*, extended ChIP Real-Time PCR methods, extended ChIP-seq data analysis methods.
(DOCX)

Acknowledgments

We would like to thank all members of the Imhof group for the critical discussion and experimental support. The work was funded by a grant from the Deutsche Forschungsgemeinschaft (DFG) to Andrea Lukacs (QBM) and Axel Imhof (IM/9-1). B.S was funded by the Chinese Scholarship Council. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank S. Krebs and H. Blum (LAFUGA at Gene Center, LMU, Munich) for outstanding sequencing service and K. Förstemann and his group for help and assistance in applying the CRISPR/*cas9* system.

Author Contributions

Conceptualization: AI AW'I TAG.

Data curation: TAG BS.

Formal analysis: BS PS TAG.

Funding acquisition: AI.

Investigation: TAG.

Methodology: AI TAG.

Project administration: AI TAG.

Resources: AL TAG.

Software: BS.

Supervision: AI TAG AWT.

Validation: TAG.

Visualization: AI TAG.

Writing – original draft: AI TAG.

Writing – review & editing: AI TAG AL.

References

1. Maheshwari S, Barbash DA. The Genetics of Hybrid Incompatibilities. Annu Rev Genet. 2011; 45: 331–355. doi: 10.1146/annurev-genet-110410-132514 PMID: 21910629
2. Barbash DA, Awadalla P, Tarone AM. Functional divergence caused by ancient positive selection of a Drosophila hybrid incompatibility locus. PLoS Biol. 2004; 2: e142. doi: 10.1371/journal.pbio.0020142 PMID: 15208709
3. Thomae AW, Schade GOM, Padeken J, Borath M, Vetter I, Kremmer E, et al. A Pair of Centromeric Proteins Mediates Reproductive Isolation in Drosophila Species. Dev Cell. 2013; 27: 412–424. doi: 10.1016/j.devcel.2013.10.001 PMID: 24239514
4. Satyaki PRV, Cuykendall TN, Wei KH-C, Brideau NJ, Kwak H, Aruna S, et al. The hmr and lhr hybrid incompatibility genes suppress a broad range of heterochromatic repeats. Malik HS, editor. PLoS Genet. Public Library of Science; 2014; 10: e1004240.
5. Wei KHC, Clark AG, Barbash DA. Limited Gene Misregulation Is Exacerbated by Allele-Specific Upregulation in Lethal Hybrids between Drosophila melanogaster and Drosophila simulans. Mol Biol Evol. 2014; 31: 1767–1778. doi: 10.1093/molbev/msu127 PMID: 24723419
6. Maheshwari S, Barbash DA. Cis-by-Trans Regulatory Divergence Causes the Asymmetric Lethal Effects of an Ancestral Hybrid Incompatibility Gene. Begun DJ, editor. PLoS Genet. 2012; 8: e1002597. doi: 10.1371/journal.pgen.1002597 PMID: 22457639
7. Straub T, Grimaud C, Gilfillan GD, Mitterweger A, Becker PB. The chromosomal high-affinity binding sites for the Drosophila dosage compensation complex. PLoS Genet. 2008; 4: e1000302. doi: 10.1371/journal.pgen.1000302 PMID: 19079572
8. Gorchakov AA, Alekseyenko AA, Kharchenko P, Park PJ, Kuroda MI. Long-range spreading of dosage compensation in Drosophila captures transcribed autosomal genes inserted on X. Genes Dev. 2009; 23: 2266–2271. doi: 10.1101/gad.1840409 PMID: 19797766
9. Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, Bonnet J, et al. Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. Cell. Elsevier; 2011; 144: 214–226.
10. Tartof KD, Hobbs C, Jones M. A structural basis for variegating position effects. Cell. 1984; 37: 869–878. PMID: 6086148
11. Talbert PB, Henikoff S. Spreading of silent chromatin: inaction at a distance. Nat Rev Genet. 2006; 7: 793–803. doi: 10.1038/nrg1920 PMID: 16983375
12. Rodriguez MA, Vermaak D, Bayes JJ, Malik HS. Species-specific positive selection of the male-specific lethal complex that participates in dosage compensation in Drosophila. Proc Natl Acad Sci USA. 2007; 104: 15412–15417. doi: 10.1073/pnas.0707445104 PMID: 17878295
13. Ross BD, Rosin L, Thomae AW, Hiatt MA, Vermaak D, la Cruz de AFA, et al. Stepwise Evolution of Essential Centromere Function in a Drosophila Neogene. Science. 2013; 340: 1211–1214. doi: 10.1126/science.1234393 PMID: 23744945
14. Bayes JJ, Malik HS. Altered Heterochromatin Binding by a Hybrid Sterility Protein in Drosophila Sibling Species. Science. 2009; 326: 1538–1541. doi: 10.1126/science.1181756 PMID: 19933102
15. Sawamura K. Chromatin Evolution and Molecular Drive in Speciation. International Journal of Evolutionary Biology. 2012; 2012: 1–9.
16. Lerat E, Burlet N, Biémont C, Vieira C. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. Gene. Elsevier B.V; 2011; 473: 100–109.
17. Vieira C, Biémont C. Transposable element dynamics in two sibling species: Drosophila melanogaster and Drosophila simulans. Genetica. 2004; 120: 115–123. PMID: 15088652

18. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007; 450: 203–218. doi: 10.1038/nature06341 PMID: 17994087
19. Ni X, Zhang YE, Nègre N, Chen S, Long M, White KP. Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome. PLoS Biol. 2012; 10: e1001420. doi: 10.1371/journal.pbio.1001420 PMID: 23139640
20. Indjeian VB, Kingman GA, Jones FC, Guenther CA, Grimwood J, Schmutz J, et al. Evolving New Skeletal Traits by cis-Regulatory Changes in Bone Morphogenetic Proteins. Cell. 2016; 164: 45–56. doi: 10.1016/j.cell.2015.12.007 PMID: 26774823
21. Dai H, Chen Y, Chen S, Mao Q, Kennedy D, Landback P, et al. The evolution of courtship behaviors through the origination of a new gene in Drosophila. National Acad Sciences; 2008; 105: 7478–7483.
22. Bushey AM, Dorman ER, Corces VG. Chromatin insulators: regulatory mechanisms and epigenetic inheritance. Mol Cell. 2008; 32: 1–9. doi: 10.1016/j.molcel.2008.08.017 PMID: 18851828
23. Yang J, Ramos E, Corces VG. The BEAF-32 insulator coordinates genome organization and function during the evolution of Drosophila species. Genome Res. 2012; 22: 2199–2207. doi: 10.1101/gr.142125.112 PMID: 22895281
24. Heger P, George R, Wiehe T. Successive gain of insulator proteins in arthropod evolution. Evolution. 2013; 67: 2945–2956. doi: 10.1111/evo.12155 PMID: 24094345
25. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell. Elsevier Inc; 2012; 148: 458–472.
26. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, et al. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature. 2007; 450: 219–232. doi: 10.1038/nature06340 PMID: 17994088
27. Bosco G, Campbell P, Leiva-Neto JT, Markow TA. Analysis of Drosophila species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics. 2007; 177: 1277–1290. doi: 10.1534/genetics.107.075069 PMID: 18039867
28. Labrador M, Corces VG. Setting the boundaries of chromatin domains and nuclear organization. Cell. 2002; 111: 151–154. PMID: 12408858
29. James TC, Elgin SC. Identification of a nonhistone chromosomal protein associated with heterochromatin in Drosophila melanogaster and its gene. Mol Cell Biol. 1986; 6: 3862–3872. PMID: 3099166
30. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10: R25. doi: 10.1186/gb-2009-10-3-r25 PMID: 19261174
31. Andrews S. FATSQC, a quality control tool for high throughput sequence data. In: http://www.bioinformatics.babraham.ac.ukprojectsfastqc. 2010.
32. Hannon GJ. FASTX Toolkits FASTA/Q short reads preprocessing kit [Internet].
33. Nicol JW, Helt GA, Blanchard SG, Raja A, Loraine AE. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics. 2009; 25: 2730–2731. doi: 10.1093/bioinformatics/btp472 PMID: 19654113
34. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinformatics. 2013; 14: 178–192. doi: 10.1093/bib/bbs017 PMID: 22517427
35. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38: 576–589. doi: 10.1016/j.molcel.2010.05.004 PMID: 20513432
36. Chen T-W, Li H-P, Lee C-C, Gan R-C, Huang P-J, Wu TH, et al. ChIPseek, a web-based analysis tool for ChIP data. BMC Genomics. BioMed Central; 2014; 15: 539.
37. Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics. 2010; 11: 237. doi: 10.1186/1471-2105-11-237 PMID: 20459804
38. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015; 6: 11. doi: 10.1186/s13100-015-0041-9 PMID: 26045719
39. Ong C-T, Van Bortle K, Ramos E, Corces VG. Poly(ADP-ribosyl)ation Regulates Insulator Function and Intrachromosomal Interactions in Drosophila. Cell. Elsevier Inc; 2013; 155: 148–159.
40. Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, et al. Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res. 2011; 21: 147–163. doi: 10.1101/gr.110098.110 PMID: 21177972

41. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013; 31: 46–53. doi: 10.1038/nbt.2450 PMID: 23222703

42. Frasch M, Glover DM, Saumweber H. Nuclear antigens follow different pathways into daughter nuclei during mitosis in early Drosophila embryos. J Cell Sci. 1986; 82: 155–172. PMID: 3098744

43. Schwartz YB, Linder-Basso D, Kharchenko PV, Tolstorukov MY, Kim M, Li H-B, et al. Nature and function of insulator protein binding sites in the Drosophila genome. Genome Res. Cold Spring Harbor Lab; 2012; 22: 2188–2198.

44. Spana C, Harrison DA, Corces VG. The Drosophila melanogaster suppressor of Hairy-wing protein binds to specific sequences of the gypsy retrotransposon. Genes Dev. 1988; 2: 1414–1423. PMID: 2850261

45. Emberly E, Blattes R, Schuettengruber B, Hennion M, Jiang N, Hart CM, et al. BEAF Regulates Cell-Cycle Genes through the Controlled Deposition of H3K9 Methylation Marks into Its Conserved Dual-Core Binding Sites. Misteli T, editor. PLoS Biol. 2008; 6: e327–15.

46. Ghosh D, Gerasimova TI, Corces VG. Interactions between the Su(Hw) and Mod(mdg4) proteins required for gypsy insulator function. EMBO J. 2001; 20: 2518–2527. doi: 10.1093/emboj/20.10.2518 PMID: 11350941

47. Gause M, Morcillo P, Dorsett D. Insulation of enhancer-promoter communication by a gypsy transposon insert in the Drosophila cut gene: cooperation between suppressor of hairy-wing and modifier of mdg4 proteins. Mol Cell Biol. 2001; 21: 4807–4817. doi: 10.1128/MCB.21.14.4807-4817.2001 PMID: 11416154

48. Pai C-Y, Lei EP, Ghosh D, Corces VG. The centrosomal protein CP190 is a component of the gypsy chromatin insulator. Mol Cell. Elsevier; 2004; 16: 737–748.

49. Abad JP, Carmena M, Baars S, Saunders RD, Glover DM, Ludeña P, et al. Dodeca satellite: a conserved G+C-rich satellite from the centromeric heterochromatin of Drosophila melanogaster. Proc Natl Acad Sci USA. 1992; 89: 4663–4667. PMID: 1584802

50. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, et al. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. 2002; 3: RESEARCH0084. doi: 10.1186/gb-2002-3-12-research0084 PMID: 12537573

51. Bartkuhn M, Straub T, Herold M, Herrmann M, Rathke C, Saumweber H, et al. Active promoters and insulators are marked by the centrosomal protein 190. EMBO J. EMBO Press; 2009; 28: 877–888.

52. Bohla D, Herold M, Panzer I, Buxa MK, Ali T, Demmers J, et al. A functional insulator screen identifies NURF and dREAM components to be required for enhancer-blocking. Dean A, editor. PLoS ONE. 2014; 9: e107765. doi: 10.1371/journal.pone.0107765 PMID: 25247414

53. Kelleher ES, Edelman NB, Barbash DA. Drosophila Interspecific Hybrids Phenocopy piRNA-Pathway Mutants. Noor MAF, editor. PLoS Biol. 2012; 10: e1001428. doi: 10.1371/journal.pbio.1001428 PMID: 23189033

54. Brideau NJ, Barbash DA. Functional conservation of the Drosophila hybrid incompatibility gene Lhr. BMC Evol Biol. BioMed Central Ltd; 2011; 11: 57.

55. Olszak AM, van Essen D, Pereira AJ, Diehl S, Manke T, Maiato H, et al. Heterochromatin boundaries are hotspots for de novo kinetochore formation. Nat Cell Biol. 2011; 13: 799–808. doi: 10.1038/ncb2272 PMID: 21685892

56. Sun X, Wahlstrom J, Karpen G. Molecular structure of a functional Drosophila centromere. Cell. 1997; 91: 1007–1019. PMID: 9428523

57. Sun X, Le HD, Wahlstrom JM, Karpen GH. Sequence Analysis of a Functional Drosophila Centromere. Genome Res. 2003; 13: 182–194. doi: 10.1101/gr.681703 PMID: 12566396

58. Barth TK, Schade GOM, Schmidt A, Vetter I, Wirth M, Heun P, et al. Identification of novel Drosophila centromere-associated proteins. Proteomics. 2014; 14: 2167–2178. doi: 10.1002/pmic.201400052 PMID: 24841622

59. Aruna S, Flores HA, Barbash DA. Reduced fertility of Drosophila melanogaster hybrid male rescue (Hmr) mutant females is partially complemented by Hmr orthologs from sibling species. Genetics. 2009; 181: 1437–1450. doi: 10.1534/genetics.108.100057 PMID: 19153254

60. Roy S, Gilbert MK, Hart CM. Characterization of BEAF mutations isolated by homologous recombination in Drosophila. Genetics. Genetics; 2007; 176: 801–813. doi: 10.1534/genetics.106.068056 PMID: 17435231

61. Jiang N, Emberly E, Cuvier O, Hart CM. Genome-wide mapping of boundary element-associated factor (BEAF) binding sites in Drosophila melanogaster links BEAF to transcription. Mol Cell Biol. American Society for Microbiology; 2009; 29: 3556–3568.

62. Yasuhara JC, Wakimoto BT. Oxymoron no more: the expanding world of heterochromatic genes. Trends Genet. Elsevier; 2006; 22: 330–338.

63. Parnell TJ, Viering MM, Skjesol A, Helou C, Kuhn EJ, Geyer PK. An endogenous suppressor of hairy-wing insulator separates regulatory domains in Drosophila. Proc Natl Acad Sci USA. 2003; 100: 13436–13441. doi: 10.1073/pnas.2333111100 PMID: 14597701

64. Golovnin A, Biryukova I, Birukova I, Romanova O, Silicheva M, Parshikov A, et al. An endogenous Su (Hw) insulator separates the yellow gene from the Achaete-scute gene complex in Drosophila. Development. 2003; 130: 3249–3258. PMID: 12783795

65. Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, et al. A cis-regulatory map of the Drosophila genome. Nature. 2011; 471: 527–531. doi: 10.1038/nature09990 PMID: 21430782

66. Baxley RM, Soshnev AA, Koryakov DE, Zhimulev IF, Geyer PK. The role of the Suppressor of Hairy-wing insulator protein in Drosophila oogenesis. Dev Biol. 2011; 356: 398–410. doi: 10.1016/j.ydbio.2011.05.666 PMID: 21651900

67. Gerasimova TI, Byrd K, Corces VG. A chromatin insulator determines the nuclear localization of DNA. Mol Cell. 2000; 6: 1025–1035. PMID: 11106742

68. Gerasimova TI, Corces VG. Polycomb and trithorax group proteins mediate the function of a chromatin insulator. Cell. 1998; 92: 511–521. PMID: 9491892

69. Padeken J, Mendiburo MJ, Chlamydas S, Schwarz H-J, Kremmer E, Heun P. The Nucleoplasmin Homolog NLP Mediates Centromere Clustering and Anchoring to the Nucleolus. Mol Cell. Elsevier Inc; 2013;: 1–14.

70. Bolkan BJ, Booker R, Goldberg ML, Barbash DA. Developmental and Cell Cycle Progression Defects in Drosophila Hybrid Males. Genetics. 2007; 177: 2233–2241. doi: 10.1534/genetics.107.079939 PMID: 17947412

71. Phadnis N, Baker EP, Cooper JC, Frizzell KA, Hsieh E, la Cruz de AFA, et al. An essential cell cycle regulation gene causes hybrid inviability in Drosophila. Science. 2015; 350: 1552–1555. doi: 10.1126/science.aac7504 PMID: 26680200

72. Wilcoxon F. Individual comparisons by ranking methods. Biometrics bulletin. 1945; 1s: 80–83.

73. Rus F, Flatt T, Tong M, Aggarwal K, Okuda K, Kleino A, et al. Ecdysone triggered PGRP-LC expression controls Drosophila innate immunity. EMBO J. 2013; 32: 1626–1638. doi: 10.1038/emboj.2013.100 PMID: 23652443

Supporting information

Figure S1

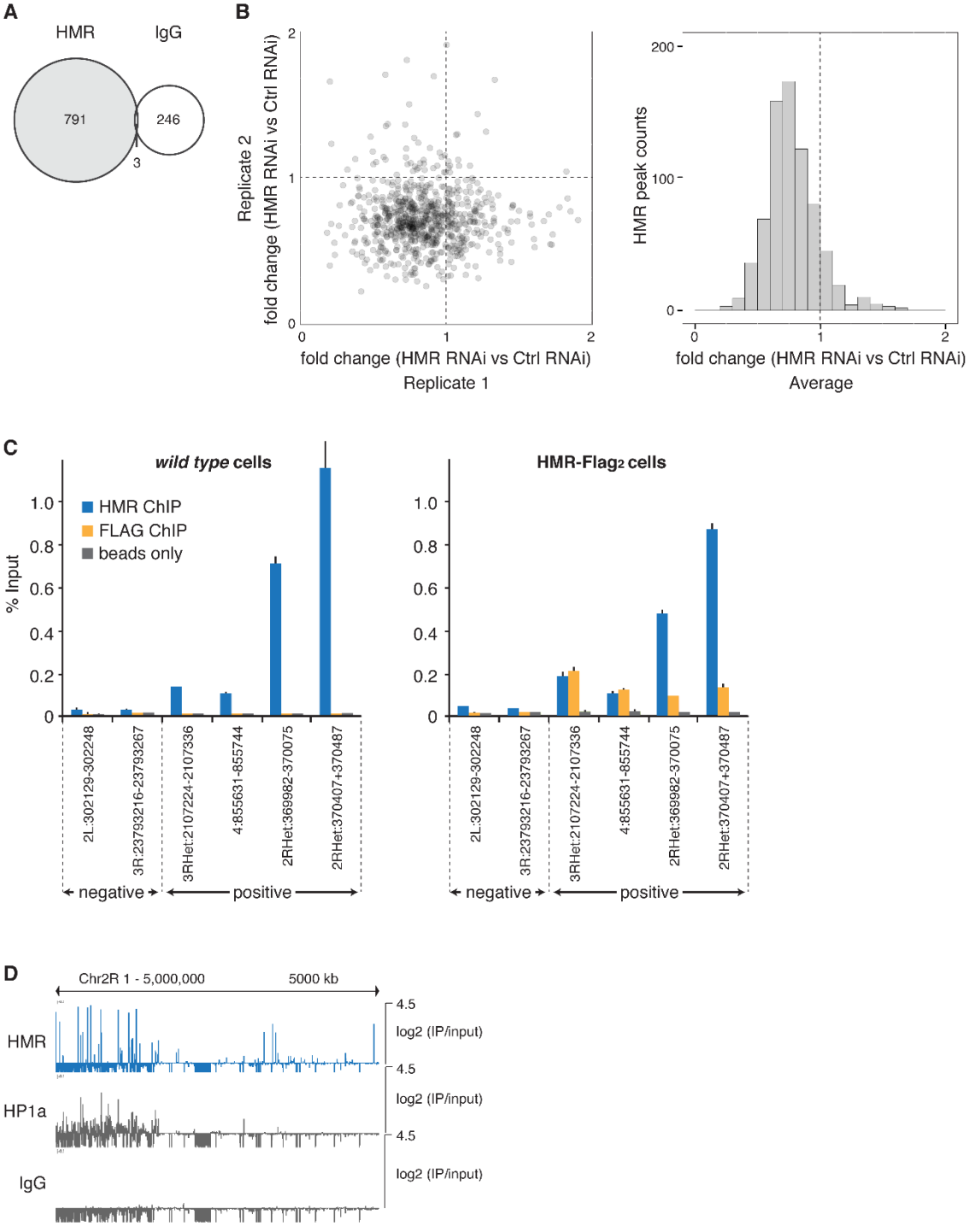


Figure S2

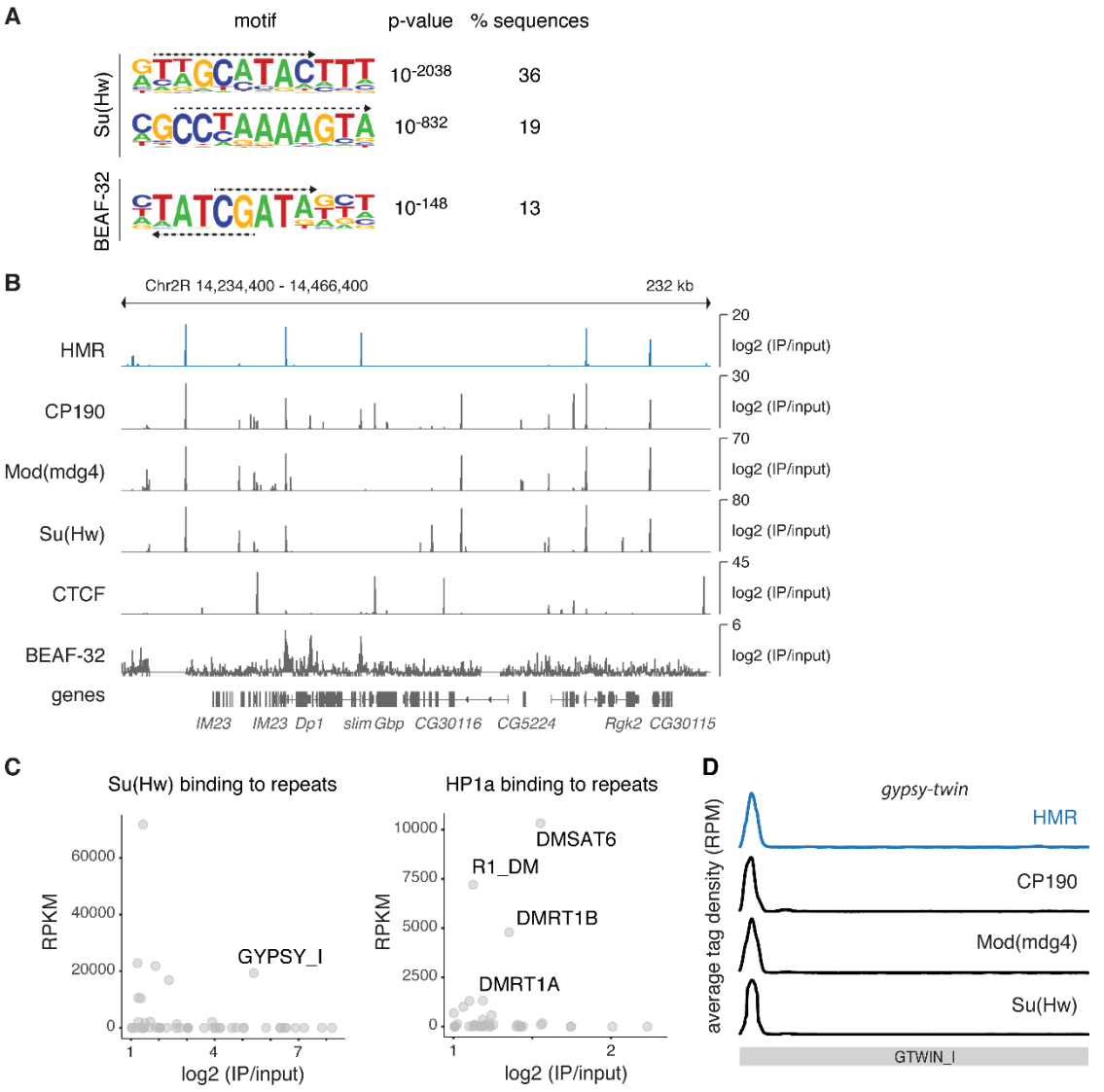

A
motif
p-value
% sequences
Su(Hw)
10-2038
36
10-832
19
BEAF-32
10-148
13
B
Chr2R 14,234,400 - 14,466,400
232 kb
HMR
CP190
Mod(mdg4)
Su(Hw)
CTCF
BEAF-32
genes
log2 (IP/input)
IM23
IM23
Dp1
slim
Gbp
CG30116
CG5224
Rgk2
CG30115
C
Su(Hw) binding to repeats
HP1a binding to repeats
RPKM
GYPSY_I
DMSAT6
R1_DM
DMRT1B
DMRT1A
log2 (IP/input)
D
gypsy-twin
average tag density (RPM)
HMR
CP190
Mod(mdg4)
Su(Hw)
GTWIN_I

Figure S3

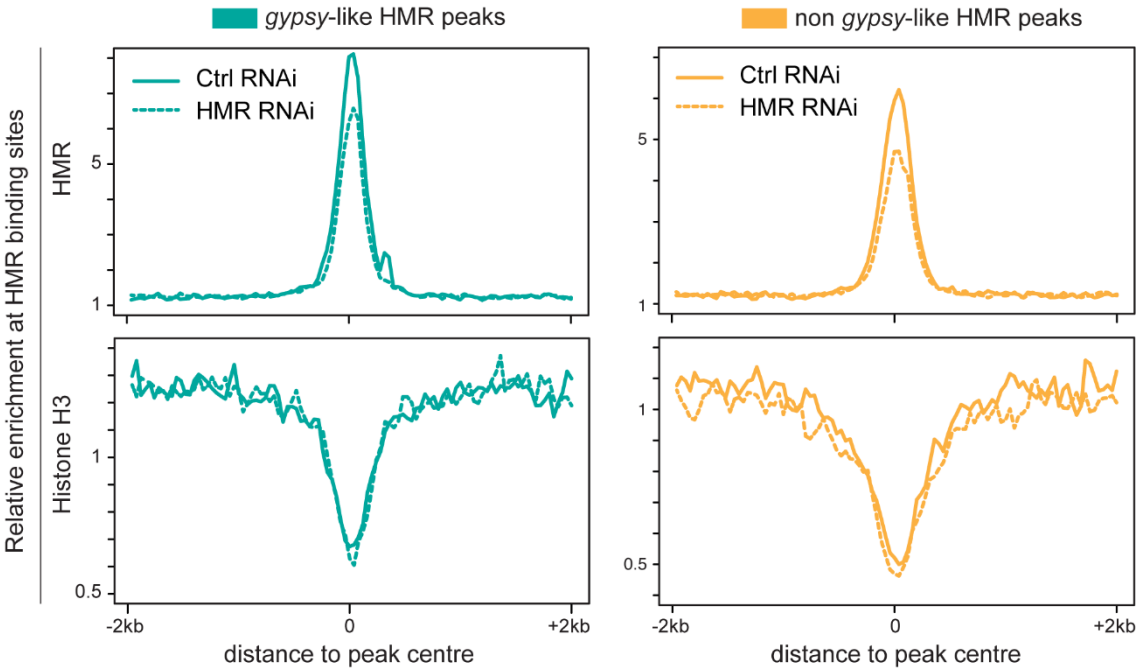
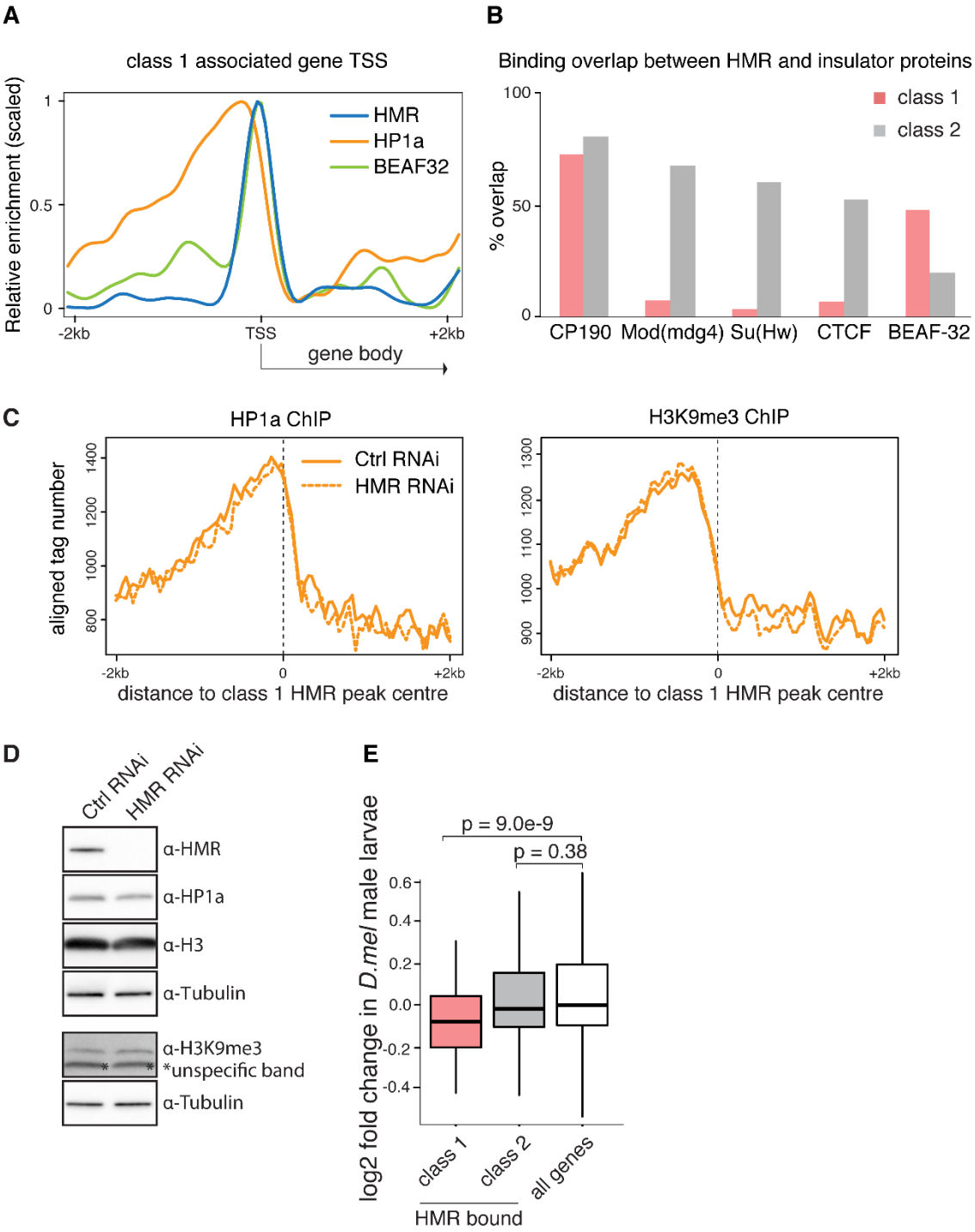


Figure S4


A
class 1 associated gene TSS
Relative enrichment (scaled)
1
0.5
0
HMR
HP1a
BEAF32
-2kb
TSS
+2kb
gene body
B
Binding overlap between HMR and insulator proteins
% overlap
100
50
0
class 1
class 2
CP190
Mod(mdg4)
Su(Hw)
CTCF
BEAF-32
C
HP1a ChIP
aligned tag number
1400
1200
1000
800
Ctrl RNAi
HMR RNAi
-2kb
0
+2kb
distance to class 1 HMR peak centre
H3K9me3 ChIP
1300
1200
1100
1000
900
-2kb
0
+2kb
distance to class 1 HMR peak centre
D
Ctrl RNAi
HMR RNAi
α-HMR
α-HP1a
α-H3
α-Tubulin
α-H3K9me3
*unspecific band
α-Tubulin
E
p = 9.0e-9
p = 0.38
log2 fold change in D.mel male larvae
0.6
0.4
0.2
0.0
-0.2
-0.4
class 1
class 2
all genes
HMR bound

Name	Sequence (5'-3')	Publication
CRISPR/cas9 system based HMR epitope tagging		
Lig4 RNAi f	TAATACGACTCACTATAGGGCCCAATGATCC AAAGTGTTTTTGCA	[1]
Lig4 RNAi r	TAATACGACTCACTATAGGGAAGTAGGATGC CTTCGCGA	[1]
oligo scaffold	GTTTTAGAGCTAGAAATAGCAAGTTAAAATA AGGCTAGTCCGTTATCAACTTGAAAAAGTGG CACCGAGTCGGTGC	[1]
oligo CRISPR target with T7 prom.	TAATACGACTCACTATAGCCACCGCCTTAGC TCTCGAAACTTTGTTTTAGAGCTA	this study
primer antis. scaffold	GCACCGACTCGGTGCCACT	[1]
U6-gRNA sense	GCTCACCTGTGATTGCTCCTAC	[1]
U6-gRNA antisense	GCTTATTCTCAAAAAAGCACCGACTCGGTGC CACT	[1]
HMRtar sense	TGGGCCTACGCCGTCGGTAACTTGTCCACGG CCAGTCAGGATACACTGCTCGGCAAGATGAC GCAGCTGTTCTCTAAATACGCCAAGGTCAAT CCGCCACCGCCTGGATCTTCCGGATGGCTCG AG	this study
HMRtar antisense	ACGGCGAAAGTTCTTACAGAGAATATGTATG ACTAAACTACGTGTGCCAAAAGTTTCGAGAG GAAGTTCCTATTCTCTAGAAAGTATAGGAAC TTCCATATG	this study
Hmr CDS sense	TATAAGCAGGTGAAGCCGAAC	this study
Hmr downstream antisense	TGCCCTCATCGCTATCATTCTG	this study
RNAi knockdown experiments		
CP190 RNAi f	TAATACGACTCACTATAGGGCCTGGCTGTGC CTGAGA	[2]
CP190 RNAi r	TAATACGACTCACTATAGGGCTGGTAGACTT ATGTCCGAAA	[2]
GST RNAi f	TTAATACGACTCACTATAGGGAGAAGTTTGA ATTGGGTTTGGAGTTTCC	[3]
GST RNAi r	TTAATACGACTCACTATAGGGAGATCGCCAC CACCAAACGTGG	[3]
HMR RNAi f	TTAATACGACTCACTATAGGGAGAGATGTGG AGGTCATAGAGAATCCGCCAATG	[3]
HMR RNAi r	TTAATACGACTCACTATAGGGAGAACCTTGT TGTGCAGGGAGTCCTCCGTC	[3]
ChIP Real-Time PCR		
2L:302129-302248 for	CACAGCAACGAAGCTCTCTG	this study
2L:302129-302248 rev	AGCATAGTGACCCGCATCTC	this study
3R:23793216-23793267 for	GAGCAAGAACAGCAGCTACTTTGT	this study
3R:23793216-23793267 rev	CACCTTGACGTTGTTGGGAAT	this study

3RHet:2107224-2107336 for	AACCCTATCCAAATTTCGAACC	this study
3RHet:2107224-2107336 rev	AGCCAAGATGAAGTCGATGC	this study
4:855631-855744 for	TAAACTCAGCCCTGCATTCC	this study
4:855631-855744 rev	GTGTTAAACCAATCCGAGACATC	this study
2RHet:369982-370075 for	CATTTGACTTCTTCGACACGAC	this study
2RHet:369982-370075 rev	GACACTGATTTACACAAAGCACAAC	this study
2RHet:370407-370487 for	TGCATACCCTACAAATAGTTTTGC	this study
2RHet:370407-370487 rev	TTGATCGGCTAAGTGAAGTGG	this study

S1 Table. List of primers used in this study. Primers used in ChIP Real-Time PCR were designed with help of Primer3 [4].

References

1. Böttcher R, Hollmann M, Merk K, Nitschko V, Obermaier C, Philippou-Massier J, et al. Efficient chromosomal gene modification with CRISPR/cas9 and PCR-based homologous recombination donors in cultured Drosophila cells. Nucleic Acids Res. Oxford University Press; 2014;42: e89–e89. doi:10.1093/nar/gku289
2. Van Bortle K, Ramos E, Takenaka N, Yang J, Wahi JE, Corces VG. Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. Genome Res. 2012;22: 2176–2187. doi:10.1101/gr.136788.111
3. Thomae AW, Schade GOM, Padeken J, Borath M, Vetter I, Kremmer E, et al. A Pair of Centromeric Proteins Mediates Reproductive Isolation in Drosophila Species. Dev Cell. 2013;27: 412–424. doi:10.1016/j.devcel.2013.10.001
4. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. Oxford University Press; 2007;23: 1289–1291. doi:10.1093/bioinformatics/btm091

Treatment	Sample	Number of reads	Number of unique reads
untreated	HMR_1	20989662	7321851
untreated	HP1a_1	21130231	9615590
untreated	IgG_1	15461190	8941371
untreated	Input_1	17888908	10427301
untreated	HMR_2	22695222	11790768
untreated	HP1a_2	21119605	3793451
untreated	IgG_2	29198102	17068323
untreated	Input_2	23224706	14681351
untreated	HMR_3	21035775	11727510
untreated	Input_3	23103165	14734701
untreated	HP1a_4	27968435	13525876
untreated	Input_4	34527460	21798351
Ctrl RNAi	HMR_CtrlRNAi_1	24593353	8637377
Ctrl RNAi	HP1a_CtrlRNAi_1	7603697	765717
Ctrl RNAi	H3_CtrlRNAi_1	21646028	13497422
Ctrl RNAi	H3K9me3_CtrlRNAi_1	20709056	9610889
Ctrl RNAi	Input_CtrlRNAi_1	23510417	14443029
HMR RNAi	HMR_HMRRNAi_1	22969101	11868648
HMR RNAi	HP1a_HMRRNAi_1	18370927	7419588
HMR RNAi	H3_HMRRNAi_1	16887197	10562127
HMR RNAi	H3K9me3_HMRRNAi_1	19310617	8553421
HMR RNAi	Input_HMRRNAi_1	22252385	14445163
CP190 RNAi	HMR_CP190RNAi_1	23814936	10825131
CP190 RNAi	H3_CP190RNAi_1	20909517	13085885
CP190 RNAi	Input_CP190RNAi_1	24311469	15526600
Ctrl RNAi	HMR_CtrlRNAi_2	24910232	14947403
Ctrl RNAi	Input_CtrlRNAi_2	19192206	11772305
HMR RNAi	HMR_HMRRNAi_2	25880670	15528028
HMR RNAi	Input_HMRRNAi_2	26225419	16101359

S2 Table. ChIP-Seq sample overview and number of uniquely aligned sequence reads. The percentage of uniquely mapped reads in ChIP-Seq experiments can largely vary and depends on the nature of the ChIPed protein. Proteins that bind repetitive regions (such as HMR or HP1) give substantially lower percentages of uniquely mapped reads [1,2].

References

1. Jung YL, Luquette LJ, Ho JWK, Ferrari F, Tolstorukov M, Minoda A, et al. Impact of sequencing depth in ChIP-seq experiments. Nucleic Acids Res. 2014;42: e74–e74. doi:10.1093/nar/gku178

2. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. Lewitter F, editor. PLoS Comput Biol. 2013;9: e1003326–8. doi:10.1371/journal.pcbi.1003326

S3 Table. HMR peaks used for downstream analysis.

HMR peak list derived from HOMER peak calling on three biological replicates (Fig 1A). First three columns provide information on the peak position within the genome (chromosome, peak start and end using dm3), followed by peak annotation obtained from ChIPseeks implementation of HOMER (Fig 5C) and classification according to adjacent HP1a signals (Fig 5A).

https://doi.org/10.1371/journal.pone.0171798.s007

(XLSX)

Supplemental methods

Hmr* gene editing using CRISPR/*cas9

Endogenous tagging of *Hmr* in S2 cells was performed with support of Prof. Klaus Förstemann and colleagues and was performed exactly as described in [1]. We used *D. melanogaster* S2-DRSC cells in combination with U6-driven guide-RNA construct generated by overlap extension PCR. *Hmr*-specific reagents are listed in S1 Table.

Extended ChIP Real-Time PCR methods

Specific primers (S1 Table) were designed with help of Primer3 [2]. After purification, input DNA was diluted 500-fold, immunoprecipitated DNA was diluted 10-fold before Real-Time PCR. Real-Time PCR was performed in 10 µL reaction volume with 5 µL 2x Fast SYBR Green master mix (Applied Biosystems), 1 µL 3 mM Primer forward, 1 µL 3 mM Primer reverse, 2 µL DNA template and 1 µL H_2O on a LightCycler 480 II (Roche). The PCR program was 20 seconds at 95°C; 45 cycles of 95°C for 3 seconds and 60°C for 30 seconds.

The sample's Ct values (number of cycles required for the fluorescent signal to cross the threshold) reported by the LightCycler 480 II (Roche) software were used to calculate the percentage of immunoprecipitated DNA with respect to the input DNA. The percentage (Input %) value is

$$\% \text{ input} = e^{\Delta Ct} \times df \times vf \times 100$$

with ΔCt = Ct (Input) - Ct (ChIP), e = primer pair efficiency (close to 2 or equals 2) calculated with LightCycler 480 II (Roche) software on a serial dilution of template DNA, df = Dilution factor taking dilution of DNA template into account (df = 10/500, see above) and vf = Volume factor taking starting volume of ChIP and Input into account [vf = starting volume(Input)/starting volume(ChIP)].

Extended ChIP-seq data analysis methods

The raw reads were aligned to the *D. melanogaster* genome assembly (UCSC dm3) using Bowtie (version 2.2.6) [3] and excluding chromosome Uextra [3]. Only uniquely mapped reads are kept using samtools (version 1.2) [4]. The raw read quality was accessed using FASTQC (version 11.5) [5] and reads filtering was performed using FastX (version 0.0.13) [6]. Sequencing tracks of both fold enrichment and log (of base 2) transformation with parameter settings *–m FE* and *–m logLR –p 0.00001* were generated using MACS (version 2.1.1) [7], which were then visualized using IGB [8] and IGV [9] genome viewers. Peak calling was performed using HOMER 4.8 with parameter settings *-style factor -size 200 -fragLength 200 -inputFragLength 200* [10]. Motif search and peak annotation were performed using ChIPseeks implementation of HOMER [11].

For downstream analysis, peaks identified in two out of three biological replicates were taken. Downstream analysis steps were performed using Python and R and parts of data preprocessing was done using ChipPeakAnno [12]. For the clustering of HMR peaks according to adjacent HP1a ChIP signals, three clusters were generated with K-means algorithm [13].

For repeat analysis, reads from ChIP-Seq experiments were mapped to RepBase version 19.10 [14] using bowtie [3]. Only unique reads were kept for analysis. For each repetitive element log (of base 2) fold change was calculated. For the read density tracks, deepTools (version 2.3.3.5) [15] with parameter sets *--ratio ratio --pseudocount=1* was utilized to normalize against the control.

Following genome-wide binding data sets derived from S2 cells (unless stated otherwise) were used: CP190, Su(Hw), CTCF and mod(mdg4) from GEO GSE41354 [16], BEAF-32 from GEO GSE32815 [17]. RNA expression data for untreated S2 cells was taken from GEO GSE46020. For *D. melanogaster* larvae and ovaries, RNA-Seq data were taken from NCBI BioProject PRJNA236022 [18] and analyses were performed with cuffdiff 2 [19].

References

1. Böttcher R, Hollmann M, Merk K, Nitschko V, Obermaier C, Philippou-Massier J, et al. Efficient chromosomal gene modification with CRISPR/cas9 and PCR-based homologous recombination donors in cultured Drosophila cells. Nucleic Acids Res. Oxford University Press; 2014;42: e89–e89. doi:10.1093/nar/gku289

2. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. Oxford University Press; 2007;23: 1289–1291. doi:10.1093/bioinformatics/btm091

3. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10: R25. doi:10.1186/gb-2009-10-3-r25

4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. Oxford University Press; 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

5. Andrews S. FATSQC, a quality control tool for high throughput sequence data. In: http://www.bioinformatics.babraham.ac.ukprojectsfastqc. 2010.

6. Hannon GJ. FASTX Toolkits FASTA/Q short reads preprocessing kit [Internet].

7. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. BioMed Central; 2008;9: R137. doi:10.1186/gb-2008-9-9-r137

8. Nicol JW, Helt GA, Blanchard SG, Raja A, Loraine AE. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics. 2009;25: 2730–2731. doi:10.1093/bioinformatics/btp472

9. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinformatics. 2013;14: 178–192. doi:10.1093/bib/bbs017

10. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38: 576–589.

doi:10.1016/j.molcel.2010.05.004

11. Chen T-W, Li H-P, Lee C-C, Gan R-C, Huang P-J, Wu TH, et al. ChIPseek, a web-based analysis tool for ChIP data. BMC Genomics. BioMed Central; 2014;15: 539. doi:10.1186/1471-2164-15-539

12. Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics. 2010;11: 237. doi:10.1186/1471-2105-11-237

13. MacQueen J. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1: Statistics. Some methods for classification and …; 1967.

14. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015;6: 11. doi:10.1186/s13100-015-0041-9

15. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. Oxford University Press; 2016;44: W160–5. doi:10.1093/nar/gkw257

16. Ong C-T, Van Bortle K, Ramos E, Corces VG. Poly(ADP-ribosyl)ation Regulates Insulator Function and IntrachromosomalInteractions in Drosophila. Cell. Elsevier Inc; 2013;155: 148–159. doi:10.1016/j.cell.2013.08.052

17. Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, et al. Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res. 2011;21: 147–163. doi:10.1101/gr.110098.110

18. Satyaki PRV, Cuykendall TN, Wei KH-C, Brideau NJ, Kwak H, Aruna S, et al. The hmr and lhr hybrid incompatibility genes suppress a broad range of heterochromatic repeats. Malik HS, editor. PLoS Genet. Public Library of Science; 2014;10: e1004240. doi:10.1371/journal.pgen.1004240

19. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31: 46–53. doi:10.1038/nbt.2450

6. Investigation and highly accurate prediction of missed tryptic cleavages by deep-learning (Paper II)

Trypsin has been widely used in MS analysis for its exclusive cleavages at the C-terminus of lysine and arginine of peptide bonds. During the past few years, people had put a lot of effort into the highly accurate predictions of missed tryptic cleavages for the improvement of identifications and quantifications of proteins. In this work, we achieved high accuracy for the predictions of missed tryptic cleavages by deep-learning. With such a highly accurate prediction tool, we believe people can leverage its power to improve the performance of MS analysis.

Bo Sun, Pawel Smialowski, Tobias Straub, and Axel Imhof. 2021. "Investigation and Highly Accurate Prediction of Missed Tryptic Cleavages by Deep Learning." *Journal of Proteome Research*. doi:10.1021/acs.jproteome.1c00346.

Journal of proteome research

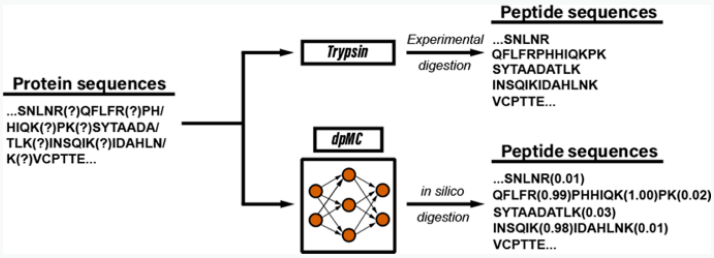
pubs.acs.org/jpr Article

Investigation and Highly Accurate Prediction of Missed Tryptic Cleavages by Deep Learning

Bo Sun, Pawel Smialowski, Tobias Straub, and Axel Imhof*

Cite This: J. Proteome Res. 2021, 20, 3749−3757 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Trypsin is one of the most important and widely used proteolytic enzymes in mass spectrometry (MS)-based proteomic research. It exclusively cleaves peptide bonds at the C-terminus of lysine and arginine. However, the cleavage is also affected by several factors, including specific surrounding amino acids, resulting in frequent incomplete proteolysis and subsequent issues in peptide identification and quantification. The accurate annotations on missed cleavages are crucial to database searching in MS analysis. Here, we present deep-learning predicting missed cleavages (dpMC), a novel algorithm for the prediction of missed trypsin cleavage sites. This algorithm provides a very high accuracy for predicting missed cleavages with area under the curves (AUCs) of cross-validation and holdout testing above 0.99, along with the mean F1 score and the Matthews correlation coefficient (MCC) of 0.9677 and 0.9349, respectively. We tested our algorithm on data sets from different species and different experimental conditions, and its performance outperforms other currently available prediction methods. In addition, the method also provides a better insight into the detailed rules of trypsin cleavages coupled with propensity and motif analysis. Moreover, our method can be integrated into database searching in the MS analysis to identify and quantify mass spectra effectively and efficiently.

KEYWORDS: *trypsin, missed cleavage, prediction, deep learning, mass spectrometry*

INTRODUCTION

Given its high specificity and stability, trypsin is the major protease used in shotgun proteomics that cleaves the C-terminal of arginine or lysine. The proteolytic products of trypsin are then analyzed by tandem mass spectrometry (MS). Then, the generated fragment spectra of selected peptide ions are matched to theoretical spectra for peptide identification. However, cleavages are frequently incomplete, and the missed cleavage rates of up to 40%[1] are regularly observed in large-scale proteomic studies. So far, the probability of cleavage for tryptic peptide prediction has been based on the Keil rules[2,3] that describe a blockage of digestion when arginine or lysine is followed by proline and a reduction of cleavage when acidic amino acids flank either side of the corresponding arginine and lysine. However, such fixed rules cannot fully explain all experimentally observed missed cleavages, leading to a flurry of approaches to better explain and predict missed cleavages.[1,4−6] Moreover, as trypsin cleavage is essentially a probabilistic event, no fixed rules could fully explain which peptide bonds will be cleaved and which ones will not be cleaved. The oversimplified cleavage rules applied in experimental data analysis can often lead to false or inaccurate identifications and quantifications based on the peptide spectrum matches (PSMs).[5,7] Accurate annotations of missed tryptic cleavages will remove unlikely sequences and lower the complexity of the database, which, in turn, will result in increased sensitivity and specificity, while decreasing the analysis time.[1,2,6]

The precise quantitation of proteins in shotgun proteomics highly depends on the number of proteotypic peptides detected. Errors introduced by wrongly assigned peptides carrying missed tryptic cleavages and unusual fragmentation patterns can therefore result in inaccurate quantitation. An improved prediction of the detectability of such peptides

Received: April 27, 2021
Published: June 17, 2021

https://doi.org/10.1021/acs.jproteome.1c00346
J. Proteome Res. 2021, 20, 3749−3757

Table 1. Statistics for the Datasets Used in This Work[a]

instrument	species (data source)	# of total proteins groups	# of total peptides	# peptides used for dpMC [training/testing/holdout]
TTOF 5600	*H. sapiens* (HeLa)	4298	35,624	8756 [7092/788/876]
timsTOF Pro	*H. sapiens* (HeLa)	6946	54,049	4616 [3739/415/462]
Q Exactive	*H. sapiens* (HeLa)	19,524	86,940	17,210 [13,940/1549/1721]
Q Exactive	*D. melanogaster*	9287	119,212	21,156 [17,136/1904/2116]
Q Exactive	*M. musculus*	17,037	59,210	4640 [3758/418/464]
Q Exactive	*S. cerevisiae*	4541	69,618	8896 [7205/801/890]

[a]Six independent data sets from different species and instruments are shown. The number of total peptides, the total protein groups in the search raw files, and the peptides used for training and testing are listed. The number in square brackets shows the number of peptides used for 10-fold cross-validation and holdout testing by dpMC. Equal number of cleavages and missed cleavages are contained in each of the six data sets.

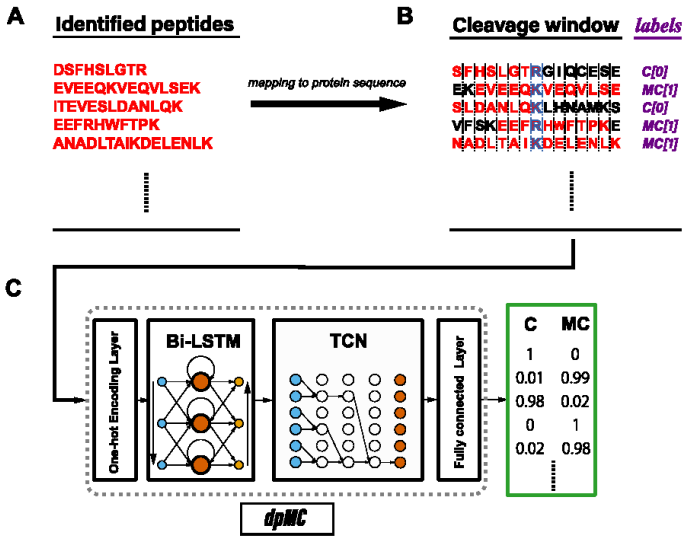


Figure 1. Workflow to build dpMC. (A) Illustrations for the peptides chosen from the search raw files that are generated from MaxQuant or Spectronaut. (B) Cleavage windows built for the corresponding peptides in (A), the peptide sequences in cleavage windows are marked in red, the adjacent sequences from the corresponding proteins are marked in black. The lysines or arginines in the middle position of the cleavage windows are marked with blue frames denoting the (missed) cleavage sites. For each peptide, the corresponding label is indicated as C (cleavage) or MC (missed cleavage) with 0 or 1 in the square brackets, respectively. (C) Illustration of the architecture of dpMC is shown in dotted gray block. Major parts of the framework are shown in four solid black blocks. The probabilistic output from dpMC is shown in solid green block. C: cleavage; MC: missed cleavage.

through an improved prediction of tryptic missed cleavages has been shown to substantially improve the quantitation of the corresponding proteins.[8−10]

In recent years, deep learning has been successfully applied for different prediction tasks including natural language processing (NLP),[11] picture recognition,[12] and weather forecasting,[13] and so forth. In this work, we leverage the powerful capability of deep learning for the prediction of sequence processing to make highly accurate predictions of missed cleavages by trypsin. Long short-term memory (LSTM) has been successfully applied in plenty of time-series work, including NLP, labeling for pictures, and so forth. Compared to the traditional recurrent neural network (RNN), the temporal convolutional network (TCN) performs better on different time series and long memory tasks including copy memory, adding problems, and so forth. Meanwhile, the TCN also shows advantages on flexible receptive field sizes, stable gradients, and speed. Instead of using gated cells, the TCN takes the advantages of one-dimensional convolutional neural network (1D-CNN), through the dilated connection of neurons in different CNN layers, enabling it to keep the distant information of long sequences.

Based on the framework we built using the algorithms mentioned above, we achieved the testing AUC of above 0.99 for different species and experimental conditions. Moreover, we analyzed sequence features determining the efficiency of the cleavages and missed cleavages by trypsin from statistical and deep-learning perspectives. Also, in order to make highly accurate predictions for given species, experiment and measurement setup, a fine-tuning strategy was proposed in this analysis. The source code of our method is available online at https://github.com/dpMC-sun/dpMC.

METHODS

Data Acquisition

Six data sets from the proteomes of four different species, *H. sapiens*, *D. melanogaster*, *M. musculus* and *S. cerevisiae*, were

https://doi.org/10.1021/acs.jproteome.1c00346
J. Proteome Res. 2021, 20, 3749−3757

used to train, validate, and test the performance of deep-learning predicting missed cleavages (dpMC), one of which was recorded on TripleTOF 5600 (PXD009273[14]), one on timsTOF Pro (PXD014777[15]), and four on Q Exactive Orbitrap (PXD007158,[16] PXD010627,[17] PXD013478,[18] and PXD018100[19]). The details of the number of peptides used by dpMC are shown in Table 1.

Searching of Raw Files and Preprocessing of Data

For the timsTOF Pro and Q Exactive Orbitrap data sets, database searching was performed using MaxQuant[20] and the *peptide.txt* files in the repository used for dpMC analysis. For the MaxQuant *peptides.txt* file, the searching parameters were set as the description in the *parameters.txt* file using trypsin or trypsin/P as enzyme cleavage rules. Only nonreverse and nonpotential contaminant peptides with posterior error probability (PEP) less than 0.01 were kept. For all the peptide entries in the search results, only arginine (R) or lysine (K) in the C-terminal of the peptides and a maximum of two missed cleavages in the peptides were kept.

For the data sets of HeLa from TripleTOF 5600, the 15 pSALIC DDA files of HeLa cell line data sets were searched by Pulsar in Spectronaut (14.2.200619, Biognosys AG, Schlieren, Switzerland) to generate library files used for dpMC in which trypsin/P was set as the enzyme/cleavage rules with missed cleavages less than 2, and the length of peptides was set in the range from 7 to 60. Variable modifications including acetyl (protein N-terminal) and oxidation (M), in addition to the fixed modification carbamidomethyl (C) were searched.

To achieve high confident cleavage information, peptides that were identified with and without tryptic cleavages were discarded. A cleavage window of at most 15 amino acid length, with 7 amino acids N-terminally and C-terminally of the putative cleavage sites was used. As the identified peptides had a minimum length of 7 amino acids, only the peptides containing 7 amino acids N-terminally of the putative cleavage site were kept, while all peptides that have at least one amino acid C-terminally of the putative cleavage sites were kept (Figure 1A,B, Table 1). The cleavage windows that contain "X" or "U" are not used for training and prediction.

Classification Algorithms

The TCN connected to bidirectional LSTM (Bi-LSTM) was adopted to build dpMC. For dpMC, first, we encoded the sequence of cleavage windows built from the processed identified peptides from a search engine by one-hot encoding with 15 time steps and 21 input dimensions that are encoded by all 20 amino acids and one blank positional codes.

Then, the input was smoothed by 256 filters with the kernel size of 9, through the skipped connection of 7 dilations of 1, 2, 4, 8, 16, 32, and 64 positions, respectively, to build the TCN block with which two dense layers of 512 units and the connection between those two layers were formed. For both training and testing data sets, missed cleavage sequences were marked as positive class—"1", and the cleavage sequence was set as negative class—"0". The probabilistic output of the last dense layer was between 0 and 1, with both indications for cleavages and missed cleavages. We chose *softmax*[21] as the activation function of the last dense layer and *categorical crossentropy*[22] as the loss function with *Adam*[23] as the optimizer (Figure 1C and Figure S1). The outputs of dpMC are the probabilities of both classes with arginine or lysine at position P1 within the cleavage window. A value of 0.5 was set as the threshold for labeling the predictions on given cleavage windows. Twenty epochs and 64 batch size were set as the default values for training dpMC. dpMC was developed in Python 3.6.5 (Anaconda3 5.2.0 64-bit) using keras (Version 2.3.1) with tensorflow-gpu (Version 1.13.1) backend. dpMC is open-source and freely available on Github.

Validation and Testing of Prediction Performance

To test the performance of the dpMC algorithm, specific data sets from different sources and different instruments were used. Each data set containing an equal number of fully cleaved peptides and peptides resulting from one or two missed cleavages was split into two parts: 9/10 of which was used for cross-validation and 1/10 of which was used for holdout testing. In all 20 epochs of dpMC training, a training−validation strategy was applied using a 10-fold cross-validation for each of the data sets. For each cycle, 5% of the training data sets were used for validation to calculate the validation loss; afterward, the trained model with the minimum validation loss was chosen for further testing (Figure S2).

In this work, sensitivity, specificity, and precision are also known as the positive predictive value (PPV), F1-score, and Matthews correlation coefficient (MCC), and the area under the receiver operating characteristic curve (AUC) was adopted to evaluate the performance of dpMC on tryptic cleavages and missed cleavages that are calculated as follows:

$$sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$specificity = \frac{TN}{FP + TN} \tag{2}$$

$$precision \text{ or } PPV = \frac{TP}{TP + FP} \tag{3}$$

$$F1\ score = \frac{2TP}{2TP + FN + FP} \tag{4}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(FP + TN)(TN + FN)}} \tag{5}$$

$$false\ positive\ rate = \frac{FP}{FP + TN} \tag{6}$$

where TP, FP, TN, and FN are true-positive, false-positive, true-negative, and false-negative rates, respectively. The AUC is calculated by integrating the receiver operating characteristic curve (ROC), which is formed by the relation between the sensitivity (true-positive rate) and the FP rate (eq 6) of the classifier.

The AUC equals the probability that a classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. An AUC of 0.5 corresponds to a random, and an AUC of 1.0 corresponds to a perfect predictor.

Cleavages labeled as 0 and missed cleavages labeled as 1 are used for this analysis for both ground true and predicted values. Both the optimized[24] and standard cutoffs of 0.5 are used for labeling the predicted values for further analysis, which are labeled as 1 above the cutoffs and as 0 below the cutoffs. For the 10-fold cross-validation testing, the mean values of all measurements stated above were calculated along with the holdout testing for further testing and comparison with other prediction algorithms.

https://doi.org/10.1021/acs.jproteome.1c00346
J. Proteome Res. 2021, 20, 3749−3757

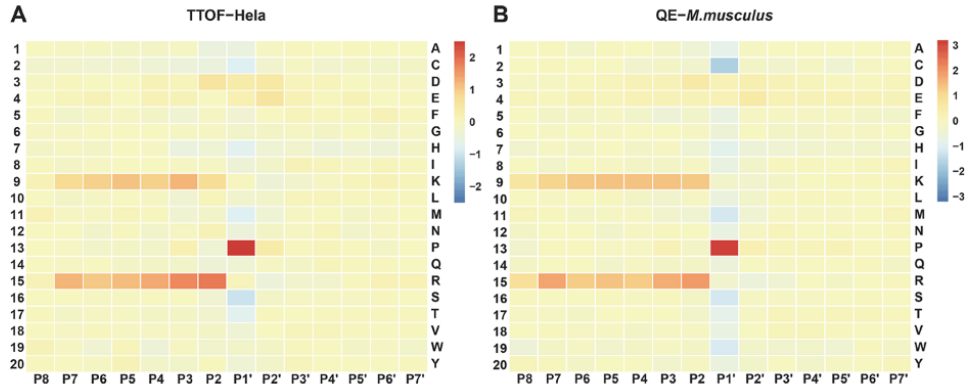


Figure 2. Heatmap of the log ratios of missed cleavages to cleavages for each amino acid at each position around cleaved or missed cleaved sites. (A) Heatmap based on the data sets of HeLa from TripleTOF 5600. (B) Heatmap based on the data sets of *M. musculus* from Q Exactive.

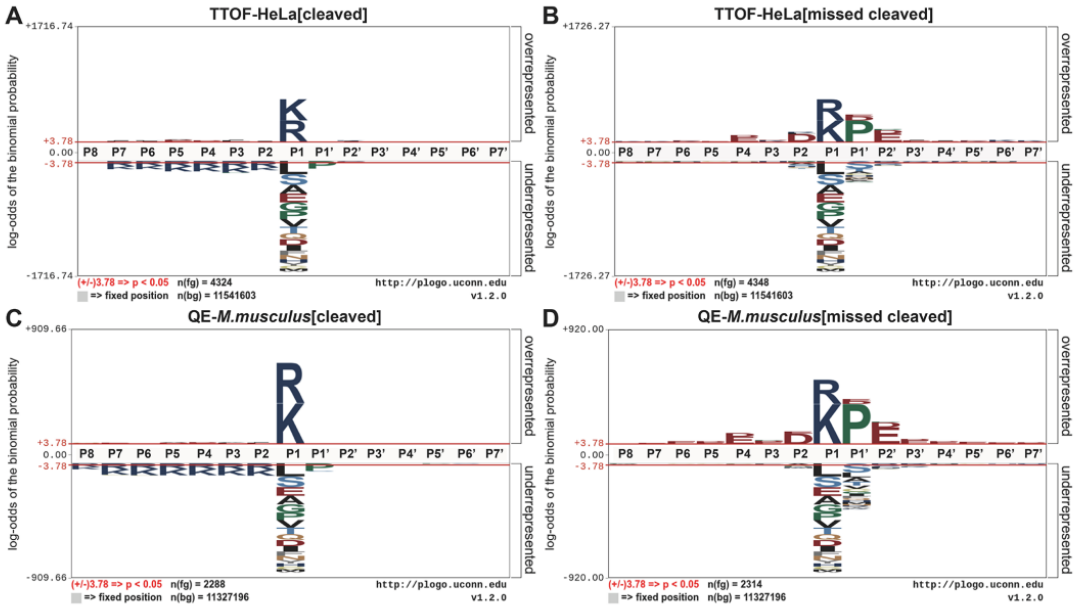


Figure 3. Probabilistic motifs for the cleavage windows of cleaved and missed cleaved sites. The probabilistic motifs of cleavage windows of (A) cleaved sites and (B) missed cleaved sites based on the data sets of HeLa from TripleTOF 5600, and the probabilistic motifs of cleavage windows of (C) cleaved sites and (D) missed cleaved sites based on the data sets of *M. musculus* from Q Exactive.

RESULTS

Frequencies of Amino Acids in Cleavage Windows

In this work, we adopted the general model of the enzymatic cleavage of subsite nomenclature by Schechter and Berger,[25,26] which annotates the amino acids around the putative cleavage sites [P8-P(7−2)-P1-P1′-P(2′-6′)-P7′], where P1 denotes the position of the potential cleavage.

In 1992, Keil summarized the digestion rules of trypsin based on the frequencies of amino acids in P2 and P1′ around arginine and lysine.[3] Here, we considered all amino acids within a window of 15 amino acids around the putative cleavage site. To determine the potential influence of a given amino acid on a tryptic cleavage, we calculated the log ratios of occurrences for a given residue at a particular position in noncleaved versus cleaved windows to reveal the propensities of amino acids around P1.

Our analysis demonstrated the previously described Keil rules, suggesting that proline located at position P1′ strongly interferes with trypsin cleavage. Interestingly, we found that this effect also extended to position P2′ and further amino acids, although to a lower extent (Figure 2 and Figure S3). Consistent with previous findings,[3,27,28] the acidic amino acids aspartate and glutamate at P2, P1′, and P2′ reduce trypsin's cleavage efficiency. Moreover, additional lysine and arginine

https://doi.org/10.1021/acs.jproteome.1c00346
J. Proteome Res. 2021, 20, 3749−3757

Table 2. K-Mer Analysis of the Most Significant Motifs[a]

Triple TOF:HeLa								Q Exactive *M. musculus*							
Cleaved				Missed Cleaved				Cleaved				Missed Cleaved			
Enriched		Depleted		Enriched		Depleted		Enriched		Depleted		Enriched		Depleted	
position	#kmer	position	#kmer	position	#kmer	position	#kmer	position	#kmer	position	#kmer	position	#kmer	position	#kmer
P8		P8	R	P8	M	P8	K	P8	P	P8	K	P8		P8	P
P7	L	P7	K	P7		P7	P	P7		P7		P7	A	P7	K
P6	L	P6	K	P6		P6	P	P6		P6		P6		P6	K
P5	E	P5		P5,P4	LE	P5	K	P5	P	P5		P5,P4	LE	P5	K
P4	P	P4		P4	E	P4	K	P4	E	P4		P4	E	P4	K
P3,P2,P1	LEK	P3		P3,P2,P1	ADK	P3	K	P3,P2,P1	LLK	P3		P3,P2,P1	EDK	P3	K
P2,P1	AR	P2	K	P2,P1	DK	P2	P	P2,P1	LK	P2		P2,P1	DK	P2	P
P1	R	P1		P1	K	P1		P1	K	P1		P1,P1′	KD	P1	
P1′	M	P1′	R	P1′	P	P1′	K	P1′		P1′		P1′	P	P1′	K
P2′		P2′	E	P2′	E	P2′	K	P2′		P2′		P2′	E	P2′	K
P3′		P3′		P3′	A	P3′	K	P3′		P3′		P3′		P3′	K
P4′		P4′		P4′	I	P4′	K	P4′		P4′		P4′		P4′	K
P5′		P5′		P5′	A	P5′	K	P5′		P5′		P5′	L	P5′	K
P6′		P6′		P6′		P6′	P	P6′		P6′		P6′		P6′	K
P7′		P7′		P7′		P7′	P	P7′		P7′		P7′		P7′	P

[a]The most significant motifs of HeLa from TripleTOF 5600 and *M. musculus* from Q Exactive analyzed by kpLogo. The most significant enriched or depleted motifs around cleaved or missed cleaved sites are shown in red.

Table 3. Performance of dpMC Evaluated on Holdout Testing Datasets with a Cutoff of 0.5[a]

instrument	species (data source)	fine-tuning	sensitivity	specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H. sapiens* (HeLa)	no	0.9569	0.9678	0.9679	0.9624	0.9248	0.9959
TTOF 5600	*H. sapiens* (HeLa)	yes	0.9569	0.9793	0.9791	0.9679	0.9365	0.9959
timsTOF Pro	*H. sapiens* (HeLa)	no	0.9502	0.9668	0.9633	0.9567	0.9172	0.9957
timsTOF Pro	*H. sapiens* (HeLa)	yes	0.9593	0.9793	0.9770	0.9680	0.9387	0.9968
Q Exactive	*H. sapiens* (HeLa)	no	0.9826	0.9558	0.9570	0.9696	0.9387	0.9959
Q Exactive	*H. sapiens* (HeLa)	yes	0.9756	0.9721	0.9722	0.9739	0.9477	0.9969
Q Exactive	*D. melanogaster*	no	0.9622	0.9593	0.9595	0.9609	0.9216	0.9938
Q Exactive	*D. melanogaster*	yes	0.9660	0.9707	0.9706	0.9683	0.9367	0.9943
Q Exactive	*M. musculus*	no	0.9839	0.9630	0.9683	0.9760	0.9470	0.9976
Q Exactive	*M. musculus*	yes	0.9879	0.9815	0.9839	0.9859	0.9694	0.9994
Q Exactive	*S. cerevisiae*	no	0.9675	0.9533	0.9572	0.9623	0.9209	0.9940
Q Exactive	*S. cerevisiae*	yes	0.9567	0.9626	0.9651	0.9609	0.9193	0.9957

[a]Performances of dpMC on six independent data sets from different species and instruments are shown. PPV, positive predicted value; MCC, Matthews correlation coefficient; and AUC, area under the curve.

residues are also enriched in peptides with a missed cleaved site (Figure 2 and Figure S3).

In addition to these previous findings, we found that the amino acids serine, methionine, histidine, threonine, and cysteine apparently facilitate trypsin cleavage as they are underrepresented at P1′ in peptides containing a missed cleavage site. Our analyses show that these amino acids also affect the cleavage efficiency when residing at a more distal position to various extents (Figure 2 and Figure S3).

Motifs in Cleavage Windows

To identify distinct amino acid patterns in the cleavage window, probabilities and k-mer of amino acids were analyzed with pLogo[29] and kpLogo.[30] We performed the motif analysis for single amino acid using a probabilistic approach based on binomial testing by pLogo. As expected, lysine and arginine are highly enriched at position P1 but virtually absent at other positions in fully cleaved peptides (Figure 3A,C, and Figures S4A, S4C, S4E, and S4G). In fully cleaved peptides, we do not observe a significant enrichment of amino acid patterns around P1. However, in peptides that contain at least one missed cleavage site, proline is frequently observed at P1′, and aspartate and glutamate are distributed in a broader range around the missed cleavage site. Interestingly, although proline at position P1′ is the major contributor for missed cleavages in all experiments, the relative additional contribution of aspartate and glutamate at position P1′ varies. In contrast, at P2′, the major contribution to cleavage resistance is made by aspartate and glutamate with only a minor, yet significant contribution of proline. (Figure 3B, S4B, and S4H). The pLogo analysis also shows that there is a less-pronounced effect of specific amino acids N-terminally of the tryptic cleavage (P1−P8). However, P2 is clearly enriched for aspartate in peptides carrying a missed cleavage. This is in contrast with P2′, where glutamate has a bigger contribution. All of these detailed novel findings that are based on the binomial probability analysis extended our understanding of the impact of proline on missed cleavages.

To identify potential motifs that are significantly enriched around the cleaved or missed cleaved sites, we also performed k-mer analysis using kpLogo.[30] This different algorithm revealed leucine as the most likely amino acid present at P3 of the cleaved sites in all data sets, regardless of the species and instruments used (Table 2 and Tables S6 and S7). Similar to the analysis of pLogo,[29] we also found lysine, arginine, proline, aspartate, and glutamate depleted in fully cleaved peptides.

When we applied k-mer analysis to detect motifs around the missed cleaved sites, we identified more complex motifs compared to the ones around the cleaved sites (Table 2 and Tables S6 and S7). This is similar to what we observed when performing the pLogo analysis. However, kpLogo also led to the identification of different motifs such as "DK", starting at P2. The k-mer algorithm also suggests a stronger enrichment of glutamate at P4 and P2′ in all data sets. Furthermore, it hints at a stronger interference of the combination of proline at P1′ and glutamate at P2′ with the tryptic cleavage than the pLogo analysis. Moreover, we observed the depletion of lysine, arginine, and proline across the entire window in peptides containing missed cleavage sites, which is in contrast with the notion that tryptic cleavage is blocked by dibasic sites.[4] Our analysis suggests that trypsin can still fully cleave a polypeptide

https://doi.org/10.1021/acs.jproteome.1c00346
J. Proteome Res. 2021, 20, 3749−3757

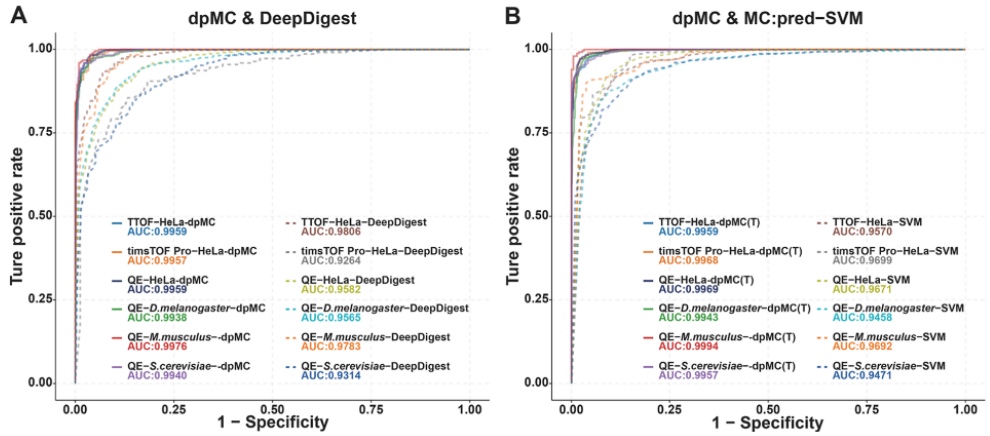


Figure 4. ROC curves of dpMC with DeepDigest and SVM. Each ROC curve is based on the performance on corresponding holdout-testing data sets. dpMC with fine-tuning is denoted with "T" in the bracket, while no fine-tuning is denoted without "T" in the bracket. (A) ROC curves of dpMC without fine-tuning are denoted as "dpMC" and DeepDigest as "DeepDigest". The performance of dpMC without fine-tuning is plotted in solid lines while dotted lines for DeepDigest. (B) ROC curves of dpMC with fine-tuning are denoted as "dpMC(T)" and MC:pred-SVM as "SVM". The performance of dpMC with fine-tuning is plotted in solid lines while dotted lines for SVM from MC:pred.

carrying long stretches of lysine and arginine, which has been demonstrated before.[2] Despite the subtle differences, the probabilistic approach as well as the k-mer analysis showed that the amino acids surrounding the putative cleavage sites substantially contribute to the cleavage efficiency in a more complex manner than previously anticipated and should therefore be considered when predicting tryptic peptides.

These findings prompted us to develop new deep-learning strategies to better address this issue. Our newly developed algorithm, dpMC, can learn even subtle patterns from specific data sets, thereby improving the accuracy of cleavage prediction (see the Methods section).

Performance of dpMC on Testing Datasets

We tested the performance of dpMC on the holdout data set and by cross-validation testing (Figure S2). For the cross-validation testing, the mean measurements were calculated for each data set. Under the cutoff of 0.5, the testing on the corresponding holdout data sets for all six data sets resulted in the AUCs ranging from 0.9938 to 0.9976 for the *D. melanogaster* data set and the *M. musculus* data set from Q Exactive (Table 3). While upon the cross-validation testing, the mean AUCs of the ROC range from 0.9904 to 0.9949 for the *S .cerevisiae* data set and the HeLa data set from Q Exactive (Table S2). For further fine-tuning of the algorithm, some of the best-trained models were used. For the fine-tuning of HeLa from TripleTOF 5600, *D. melanogaster*, *M. musculus*, and *S. cerevisiae* from Q Exactive, the best-trained model with a validation loss of 0.06731 from HeLa from Q Exactive was used. For the fine-tuning of HeLa from timsTOF Pro and Q Exactive, the best-trained model with a validation loss of 0.02913 from *M. musculus* from Q Exactive was used. For the testing on the holdout data set, the AUCs of improved models range from 0.9943 to 0.9994 for the *D. melanogaster* data set and the *M. musculus* data set from Q Exactive (Table 3). The mean AUCs of improved models range from 0.9912 to 0.9967 for the *S .cerevisiae* data set and the *M. musculus* data set from Q Exactive on the cross-validation testing data sets (Table S2).

In addition to the performance measured with AUCs, we checked the PPV, F1-score, and MCC, along with sensitivity and specificity (Table 3 and S2). For the overall measurements, the high F1-score ranging from 0.9609 to 0.9859 and MCC ranging from 0.9193 to 0.9694 with fine-tuning on the holdout-testing data sets suggest that dpMC achieved high performance on the whole level of prediction on cleavages and missed cleavages, which is also observed on the cross-validation testing data sets. On the other side, a high PPV achieved by dpMC coupled with high sensitivity suggests the accurate classification of missed cleavages. Meanwhile, the specificity also maintained at a high level indicates the high accuracy of classification for cleaved sites. In addition to the standard cutoff of 0.5, we also checked all the measurements with an optimized cutoff,[4,24] which was selected based on the testing performance. However, the differences of all measurements based on two cutoffs mentioned above are very small, which range from 0 to 0.0317 on the holdout-testing data sets (Table 3 and S1). Therefore, in the application of dpMC, we use the static classification threshold of 0.5, which was chosen for the predicted labels.

Benchmarking of dpMC

We compared the performance of dpMC with known tools for cleavage prediction such as DeepDigest,[6] SVM,[1] and information theory[4] approaches within MC:pred and PeptideCutter (https://www.expasy.org/resources/peptidecutter) from ExPASy[31] (Figure 4, Figure S5, and Supplementary Note). All the abovementioned tools, including dpMC, were tested on the same corresponding holdout data set of the six data sets (Table 1). For all measurements, the standard cutoff of 0.5 and optimized cutoffs were used for benchmarking (Tables S1−S5). It turned out that dpMC is much more robust with regard to the cutoff used than the other four algorithms. Hence, for the other programs, single optimized cutoffs are suggested,[4] which hampers their general usability. Moreover, even when using optimized cutoffs, dpMC still clearly outperformed the other prediction algorithms.

https://doi.org/10.1021/acs.jproteome.1c00346
J. Proteome Res. 2021, 20, 3749−3757

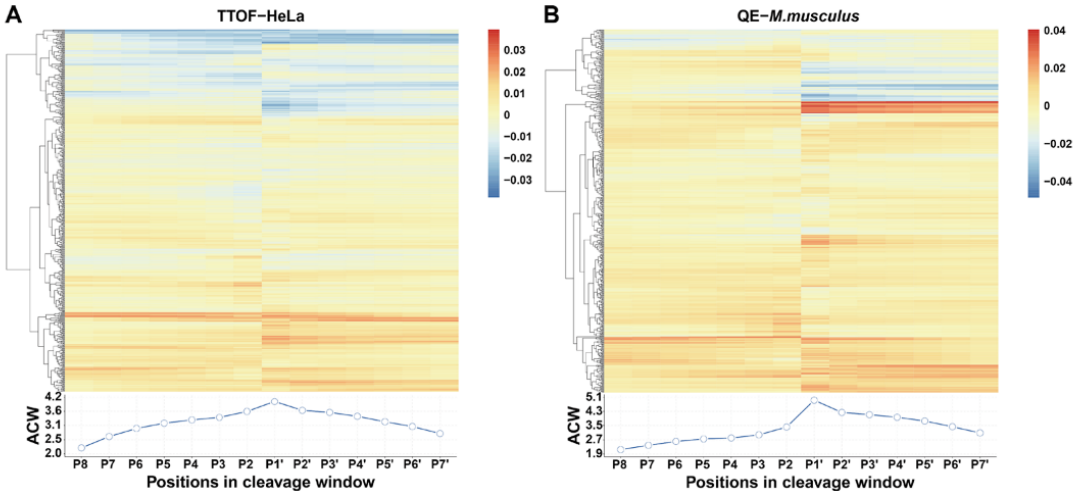


Figure 5. Heatmap and line plot of feature maps extracted from the Bi-LSTM layer. The heatmaps of activation values from 512 neurons of Bi-LSTM in the model trained on the data sets of (A) HeLa from TripleTOF 5600 and (B) *M. musculus* from Q Exactive are shown in the upper part; the line plots of ACW of the feature map are shown below the heatmaps. The positions of amino acids in the cleavage window for both the heatmap and line plot are plotted corresponding to each other.

In addition to the algorithms mentioned above, there are other approaches adopted to predict missed cleavages. For example, cleavage prediction using decision trees (CP-DT[5]) uses decision tree ensembles to predict the missed tryptic cleavages; however, because of the expiration of CP-DT's online resource, we could not compare it with dpMC, while according to the developers, CP-DT showed a maximal AUC of 0.90.

Features Driving Classification

Several deep-learning algorithms were adopted to build dpMC. In order to understand the process and mechanisms behind the decisions made on the prediction for missed cleavages, we extracted the feature map based on the activation outputs from the Bi-LSTM layer. The activation function "tanh" was used in this work, which gives both positive and negative outputs for the prediction used for updates of weights from one neuron. For the recurrent activation function, we chose "sigmoid", which is a classical method to deal with the information that flows into neurons. The feature map of the Bi-LSTM layer shows what features were detected and used for prediction. The weights assigned to a specific amino acid within the cleavage window reveal the extent and confidence for each amino acid position that Bi-LSTM made to predict missed cleavages. In this work, the absolute cumulative weights (ACWs) of all 512 filters from the Bi-LSTM layer for 14 amino acids except P1 were calculated (Figure 5 and Figure S6). For all six data sets, the apex point of ACW and the pattern of weights at P1′ are more distinct compared with other positions, indicating that dpMC extracts most of the features from P1′, and the lowest at either P8 or P7′, which fits the previous conclusions summarized by other studies[3,28] that the impact of flanking amino acids on the cleavages is strong and becomes weaker as the distance increases to the cleavage sites. In addition, the trends of the ACW curves are different for different data sets and experimental conditions. For the data sets from Q Exactive, the impact of the N-terminal amino acids is much less than that of the C-terminal amino acids of cleavage sites (Figure 5B, and Figures S6B, S6C, and S6D), while for the data sets from TripleTOF and timsTOF (Figure 5A and Figure S6A), the impact of the N-terminal amino acids is more similar to that of the C-terminal amino acids. This suggests an impact of different instruments on the detection of missed tryptic cleavages. Combined with the previous frequencies of amino acids for each position and motif analysis, the feature maps illustrate the impact of weights of these flanking amino acids on missed tryptic cleavages.

DISCUSSION

Mass spectrometry (MS)-based proteomic research relies on the matching between the experimental and in silico fragmentation patterns of peptide ions for their identification. The generation of theoretical mass spectra in the searching database depends on the correct prediction of trypsin cleavage to improve the identification and quantification rates of both shotgun data dependent acquisition (DDA) and data independent acquisition (DIA) proteomic analysis. However, existing tools still cannot meet the requirements for accurate references in database searching. One reason is the deficiency in their design and capabilities of such prediction algorithms and the other is that even though all peptides are digested by trypsin, the discrepancies in the efficiency of trypsin digestion and their detection in different experimental setups lead to an apparently different missed cleavage rate. The generalization by fixed rules or trained models based on only a few general data sets is usually not optimal. Therefore, a new accurate strategy needs to be proposed and applied.

In this work, we propose a novel method, dpMC, using a deep-learning framework to predict the probabilities of missed cleavages by trypsin for given peptide sequences with a high accuracy. dpMC outperforms other existing tools in different measurements of classification on the prediction of trypsin cleavages and missed cleavages based on the publicly available

https://doi.org/10.1021/acs.jproteome.1c00346
J. Proteome Res. 2021, 20, 3749−3757

trained models or algorithms. Using dpMC, one can train the model on custom data sets that come from specific experiments with the fine-tuning strategy. Moreover, we also studied the mechanisms behind the cleavages and missed cleavages by trypsin by extracting the feature maps used by dpMC, which allows us to better understand the importance of features for learning and decision processes in the prediction by dpMC. Through the analysis of the frequencies and motifs of amino acids around the cleaved or missed cleaved sites, some conclusions are consistent with previous findings,[2,3] and more novel and clear investigations have been made to improve the understanding of underlying patterns.

Database searching of mass spectra generated by the trypsin cleavage of protein mixtures is a major aspect of the identification and quantification of proteins in MS-based proteomics. Missed cleavages result in an incomplete coverage of the identified proteins and hamper protein grouping, which is frequently used for protein quantitation. Therefore, an accurate prediction of missed cleavages is essential for precise protein quantitation. In addition, a more accurate prediction of missed cleavages substantially decreases the search space and improves the rates of identification and quantification.[1,2]

In recent years, researchers have also put a lot of effort on the prediction of surrogate parameters for detectability[10,32,33] to identify the most detectable peptides in a theoretical proteome. These peptides are then used to generate spectral libraries in silico, which can be used for protein identification and quantitation in complex mixtures. The improved prediction method for missed cleavages provided by dpMC will greatly facilitate the identification of suited peptides that can be used for spectral libraries.

All the functions are available in an open-source program that researchers can easily use to perform training and digestion for custom data sets from different species and experiments. For each site prediction, dpMC provides both classification and its probabilities. For the future application of dpMC, users can integrate it into database searching algorithms or use it directly to precalculate classifications for specific experiments and data sets.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00346.

Performance comparison of dpMC with other existing tools for predictions of missed tryptic cleavages (PDF).

Kmer analysis by kpLogo on the sequence of cleavage windows of *S. cerevisiae* from Q Exactive (xlsx)

Kmer analysis by kpLogo on the sequence of missed cleavage windows of *S. cerevisiae* from Q Exactive (xlsx)

■ AUTHOR INFORMATION

Corresponding Author

Axel Imhof − *Biomedical Center, Protein Analysis Unit, Faculty of Medicine, Ludwig-Maximilians-Universität München, 82152 Planegg-Martinsried, Germany;* orcid.org/0000-0003-2993-8249; Phone: 0049 89 218075420; Email: Imhof@lmu.de

Authors

Bo Sun − *Biomedical Center, Protein Analysis Unit, Faculty of Medicine, Ludwig-Maximilians-Universität München, 82152 Planegg-Martinsried, Germany*

Pawel Smialowski − *Institute of Stem Cell Research, Helmholtz Center Munich, German Research Center for Environmental Health, 85764 Munich, Germany; Biomedical Center, Computational Biology Unit, Faculty of Medicine, Ludwig-Maximilians-Universität München, 82152 Planegg-Martinsried, Germany*

Tobias Straub − *Biomedical Center, Computational Biology Unit, Faculty of Medicine, Ludwig-Maximilians-Universität München, 82152 Planegg-Martinsried, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.1c00346

Funding

Bo Sun was funded by the Chinese Scholarship Council (201506230154). Work in the Imhof lab was funded by the grants of the Deutsche Forschungsgemeinschaft (CRC1064 and 1309).

Notes

The authors declare no competing financial interest.

This paper was intended for the Software Tools and Resources 2021 Special Issue, published as the April 2, 2021 issue of *J. Proteome Res.* (Vol. 20, No. 4).

■ ACKNOWLEDGMENTS

We would like to thank Ignasi Forne, Wasim Aftab, and Andreas Schmidt for suggestions and discussions.

■ REFERENCES

(1) Siepen, J. A.; Keevil, E. J.; Knight, D.; Hubbard, S. J. Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *J. Proteome Res.* **2007**, *6*, 399−408.

(2) Yen, C. Y.; Russell, S.; Mendoza, A. M.; Meyer-Arendt, K.; Sun, S.; Cios, K. J.; Ahn, N. G.; Resing, K. A. Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and SCX elution rules from data mining of MS/MS spectra. *Anal. Chem.* **2006**, *78*, 1071−1084.

(3) Keil, B., *Specificity of proteolysis*; Springer-Verlag: Berlin-Heidelberg-NewYork 1992, 335.

(4) Lawless, C.; Hubbard, S. J. Prediction of missed proteolytic cleavages for the selection of surrogate peptides for quantitative proteomics. *OMICS* **2012**, *16*, 449−456.

(5) Fannes, T.; Vandermarliere, E.; Schietgat, L.; Degroeve, S.; Martens, L.; Ramon, J. Predicting tryptic cleavage from proteomics data using decision tree ensembles. *J. Proteome Res.* **2013**, *12*, 2253−2259.

(6) Yang, J.; Gao, Z.; Ren, X.; Sheng, J.; Xu, P.; Chang, C.; Fu, Y. DeepDigest: Prediction of Protein Proteolytic Digestion with Deep Learning. *Anal. Chem.* **2021**, *93*, 6094−6103.

(7) Jesse, G. M. *In Silico* Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage. *ISRN Comput. Bio.* **2014**, *2014*, 1.

(8) Zohora, F. T.; Rahman, M. Z.; Tran, N. H.; Xin, L.; Shan, B.; Li, M. DeepIso: A Deep Learning Model for Peptide Feature Detection from LC-MS map. *Sci. Rep.* **2019**, *9*, 17168.

(9) Cheng, H.; Rao, B.; Liu, L.; Cui, L.; Xiao, G.; Su, R.; Wei, L. PepFormer: End-to-End Transformer-Based Siamese Network to Predict and Enhance Peptide Detectability Based on Sequence Only. *Anal. Chem.* **2021**, *93*, 6481−6490.

(10) Gao, Z.; Chang, C.; Yang, J.; Zhu, Y.; Fu, Y. AP3: An Advanced Proteotypic Peptide Predictor for Targeted Proteomics by Incorporating Peptide Digestibility. *Anal. Chem.* **2019**, *91*, 8705−8711.
(11) Cambria, T. Y. D. H. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computat. Intellig. Magaz.* **2018**, *13*, 55−75.
(12) CarlosAffonso, A. L. D.; AntunesVieira, F. H.; de Leon Ferreirade Carvalho, A. C. P. Deep learning for biological image classification. *Exp. Syst. Applic.* **2017**, *85*, 114−122.
(13) Yan, J.; Mu, L.; Wang, L.; Ranjan, R.; Zomaya, A. Y. Temporal Convolutional Networks for the Advance Prediction of ENSO. *Sci. Rep.* **2020**, *10*, 8055.
(14) Liu, Y.; Mi, Y.; Mueller, T.; Kreibich, S.; Williams, E. G.; Van Drogen, A.; Borel, C.; Frank, M.; Germain, P. L.; Bludau, I.; Mehnert, M.; Seifert, M.; Emmenlauer, M.; Sorg, I.; Bezrukov, F.; Bena, F. S.; Zhou, H.; Dehio, C.; Testa, G.; Saez-Rodriguez, J.; Antonarakis, S. E.; Hardt, W. D.; Aebersold, R. Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat. Biotechnol.* **2019**, *37*, 314−322.
(15) Prianichnikov, N.; Koch, H.; Koch, S.; Lubeck, M.; Heilig, R.; Brehmer, S.; Fischer, R.; Cox, J. MaxQuant Software for Ion Mobility Enhanced Shotgun Proteomics. *Mol. Cell. Proteomics* **2020**, *19*, 1058−1069.
(16) Hammerschmidt, P.; Ostkotte, D.; Nolte, H.; Gerl, M. J.; Jais, A.; Brunner, H. L.; Sprenger, H. G.; Awazawa, M.; Nicholls, H. T.; Turpin-Nolan, S. M.; Langer, T.; Kruger, M.; Brugger, B.; Bruning, J. C. CerS6-Derived Sphingolipids Interact with Mff and Promote Mitochondrial Fragmentation in Obesity. *Cell* **2019**, *177*, 1536−1552.e23.
(17) Gartner, S. M. K.; Hundertmark, T.; Nolte, H.; Theofel, I.; Eren-Ghiani, Z.; Tetzner, C.; Duchow, T. B.; Rathke, C.; Kruger, M.; Renkawitz-Pohl, R. Stage-specific testes proteomics of Drosophila melanogaster identifies essential proteins for male fertility. *Eur. J. Cell Biol.* **2019**, *98*, 103−115.
(18) Just, P. A.; Charawi, S.; Denis, R. G. P.; Savall, M.; Traore, M.; Foretz, M.; Bastu, S.; Magassa, S.; Senni, N.; Sohier, P.; Wursmer, M.; Vasseur-Cognet, M.; Schmitt, A.; Le Gall, M.; Leduc, M.; Guillonneau, F.; De Bandt, J. P.; Mayeux, P.; Romagnolo, B.; Luquet, S.; Bossard, P.; Perret, C. Lkb1 suppresses amino acid-driven gluconeogenesis in the liver. *Nat. Commun.* **2020**, *11*, 6127.
(19) Velazquez, D.; Albacar, M.; Zhang, C.; Calafi, C.; Lopez-Malo, M.; Torres-Torronteras, J.; Marti, R.; Kovalchuk, S. I.; Pinson, B.; Jensen, O. N.; Daignan-Fornier, B.; Casamayor, A.; Arino, J. Yeast Ppz1 protein phosphatase toxicity involves the alteration of multiple cellular targets. *Sci. Rep.* **2020**, *10*, 15613.
(20) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.
(21) Goodfellow, I. B., Yoshua; Courville, A. *Softmax Units for Multinoulli Output Distributions*; MIT Press, 2016; 180−184.
(22) Ian Goodfellow, Y. B.; Courville, A., *Deep Learning*; MIT Press, 2016; 73−75.
(23) Diederik, P. K.; Adam, J. B. A Method for Stochastic Optimization. *arXiv* **2014**, *2014*, 1−15.
(24) Hao, Z. D. A. Y.; *Report ROC: An Easy Way to Report ROC Analysis*; CRAN, 2020, 1−5.
(25) Berger, A.; Schechter, I. On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun.* **1967**, *27*, 157.
(26) Schechter, I.; Berger, A. On the active site of proteases. III. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochem. Biophys. Res. Commun.* **1968**, *32*, 898.
(27) Needleman, S. B. *Protein Sequence Determination: A Sourcebook of Methods and Techniques*; Springer, 2013; 348.
(28) Heissel, S.; Frederiksen, S. J.; Bunkenborg, J.; Hojrup, P. Enhanced trypsin on a budget: Stabilization, purification and high-temperature application of inexpensive commercial trypsin for proteomics applications. *PLoS One* **2019**, *14*, No. e0218374.
(29) O'Shea, J. P.; Chou, M. F.; Quader, S. A.; Ryan, J. K.; Church, G. M.; Schwartz, D. pLogo: a probabilistic approach to visualizing sequence motifs. *Nat. Methods* **2013**, *10*, 1211−1212.
(30) Wu, X.; Bartel, D. P. kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. *Nucleic Acids Res.* **2017**, *45*, W534−W538.
(31) Artimo, P.; Jonnalagedda, M.; Arnold, K.; Baratin, D.; Csardi, G.; de Castro, E.; Duvaud, S.; Flegel, V.; Fortier, A.; Gasteiger, E.; Grosdidier, A.; Hernandez, C.; Ioannidis, V.; Kuznetsov, D.; Liechti, R.; Moretti, S.; Mostaguir, K.; Redaschi, N.; Rossier, G.; Xenarios, I.; Stockinger, H. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* **2012**, *40*, W597−W603.
(32) Yang, Y.; Liu, X.; Shen, C.; Lin, Y.; Yang, P.; Qiao, L. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat. Commun.* **2020**, *11*, 146.
(33) Eyers, C. E.; Lawless, C.; Wedge, D. C.; Lau, K. W.; Gaskell, S. J.; Hubbard, S. J. CONSeQuence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. *Mol. Cell. Proteomics* **2011**, *10*, No. M110.003384.

NOTE ADDED AFTER ASAP PUBLICATION

This paper was published on the Web on June 17, 2021. The tables were reformatted for better clarity, and text had been missing from the first paragraph of the "Features Driving Classification" was inserted. The corrected version was reposted on June 22, 2021.

https://doi.org/10.1021/acs.jproteome.1c00346
J. Proteome Res. 2021, 20, 3749−3757

Supplementary information

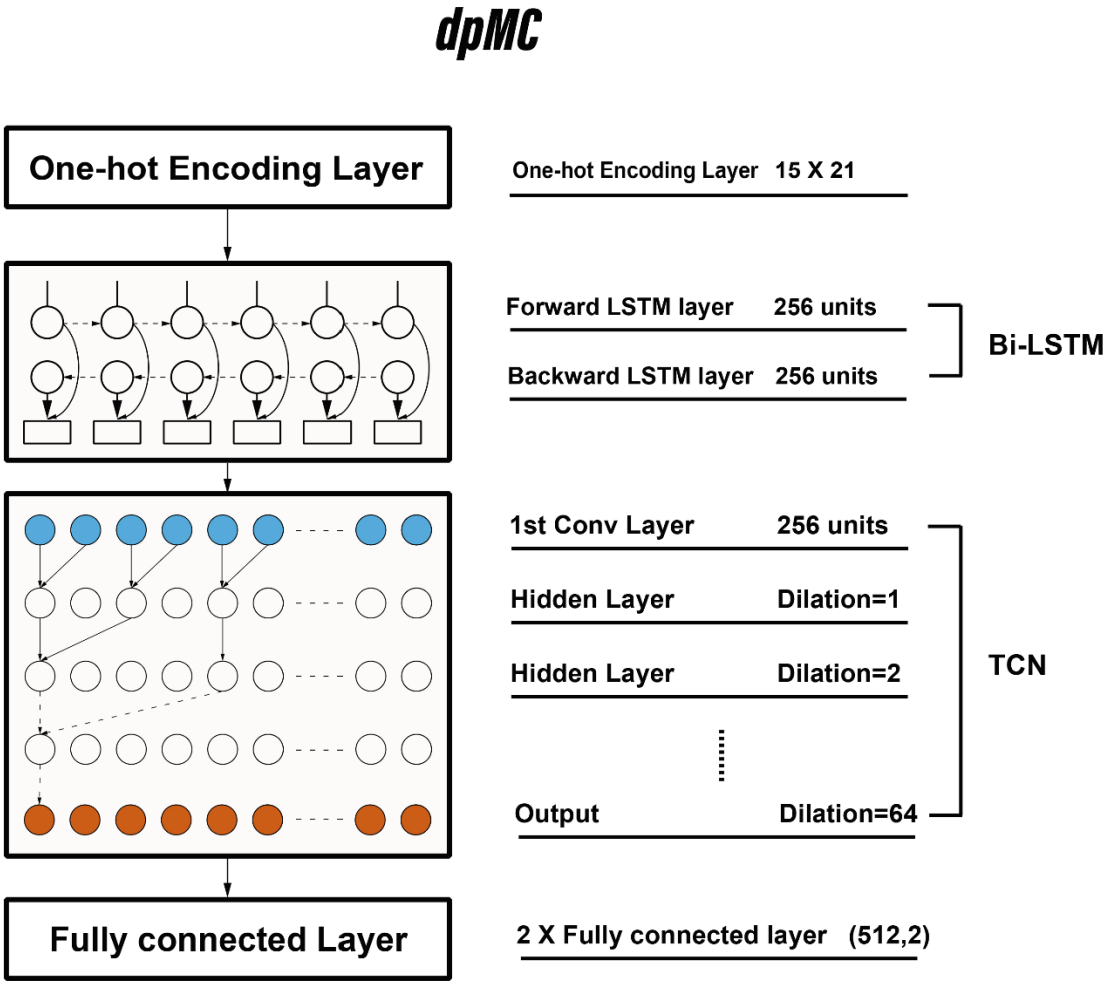


Figure S1. The details of the architecture of dpMC. The numbers of dimensions of one-hot encoding layer and numbers of neurons in other layers are noted on the right side of the illustration of the structure of dpMC.

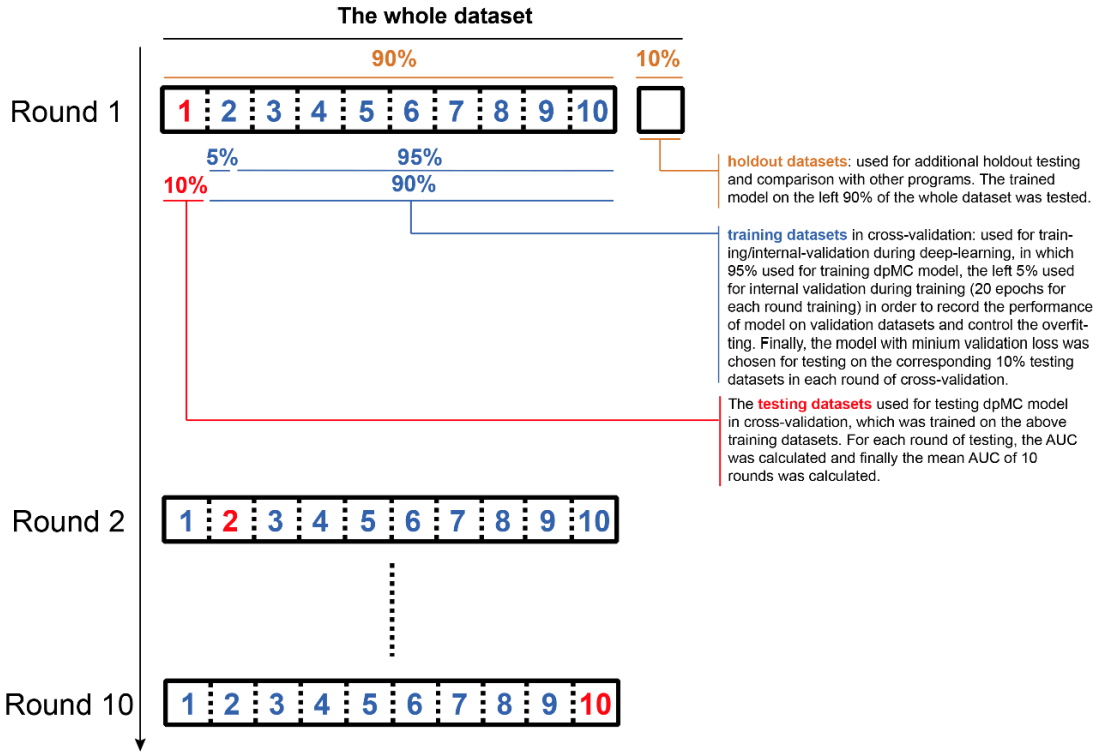


Figure S2. The workflow on the conduction of training, testing and validation procedure. Both 10-fold and holdout testing are shown, also including the internal training and validation during deep-learning by dpMC.

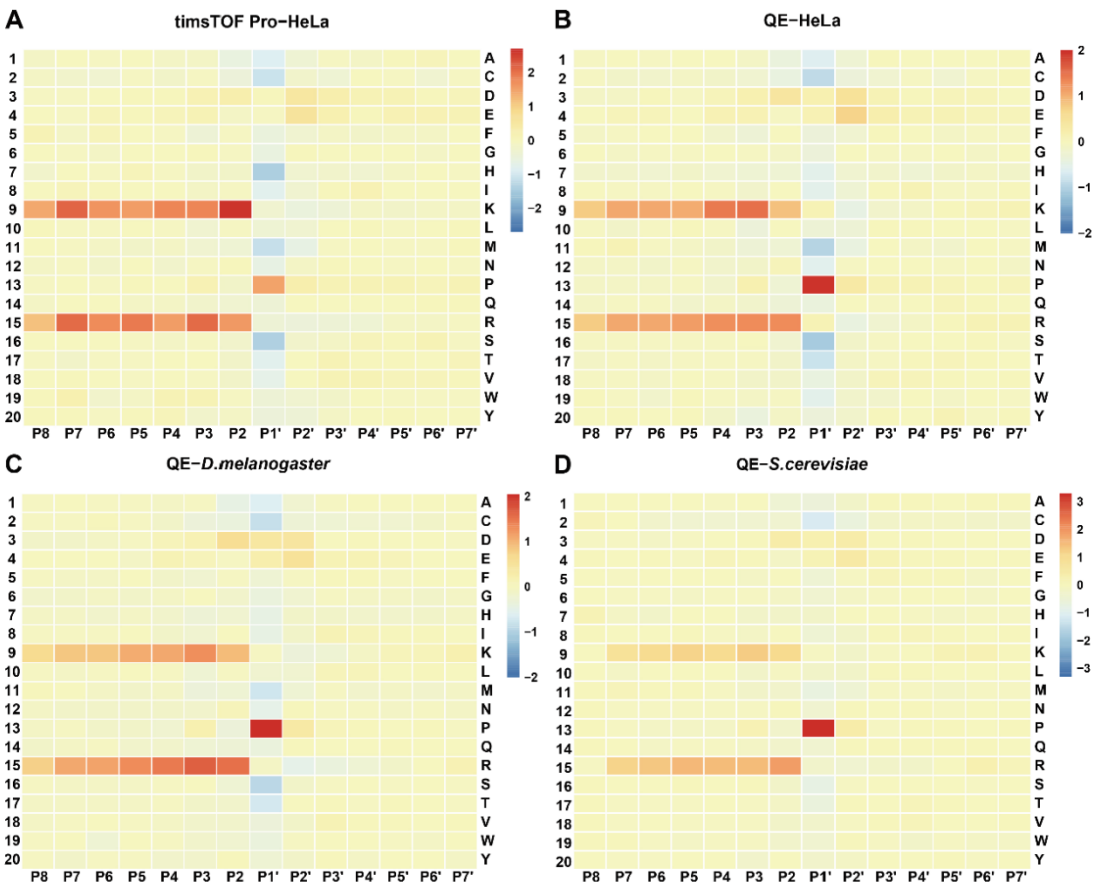


Figure S3. The heatmap of log ratios of amino acids of missed cleavages to cleavages at each position (A) The heatmap based on the datasets of HeLa from timsTOF Pro. (B) The heatmap based on the datasets of HeLa from Q Exactive. (C) The heatmap based on the datasets of *D.melanogaster* from Q Exactive. (D) The heatmap based on the datasets of *S.cerevisiae* from Q Exactive.

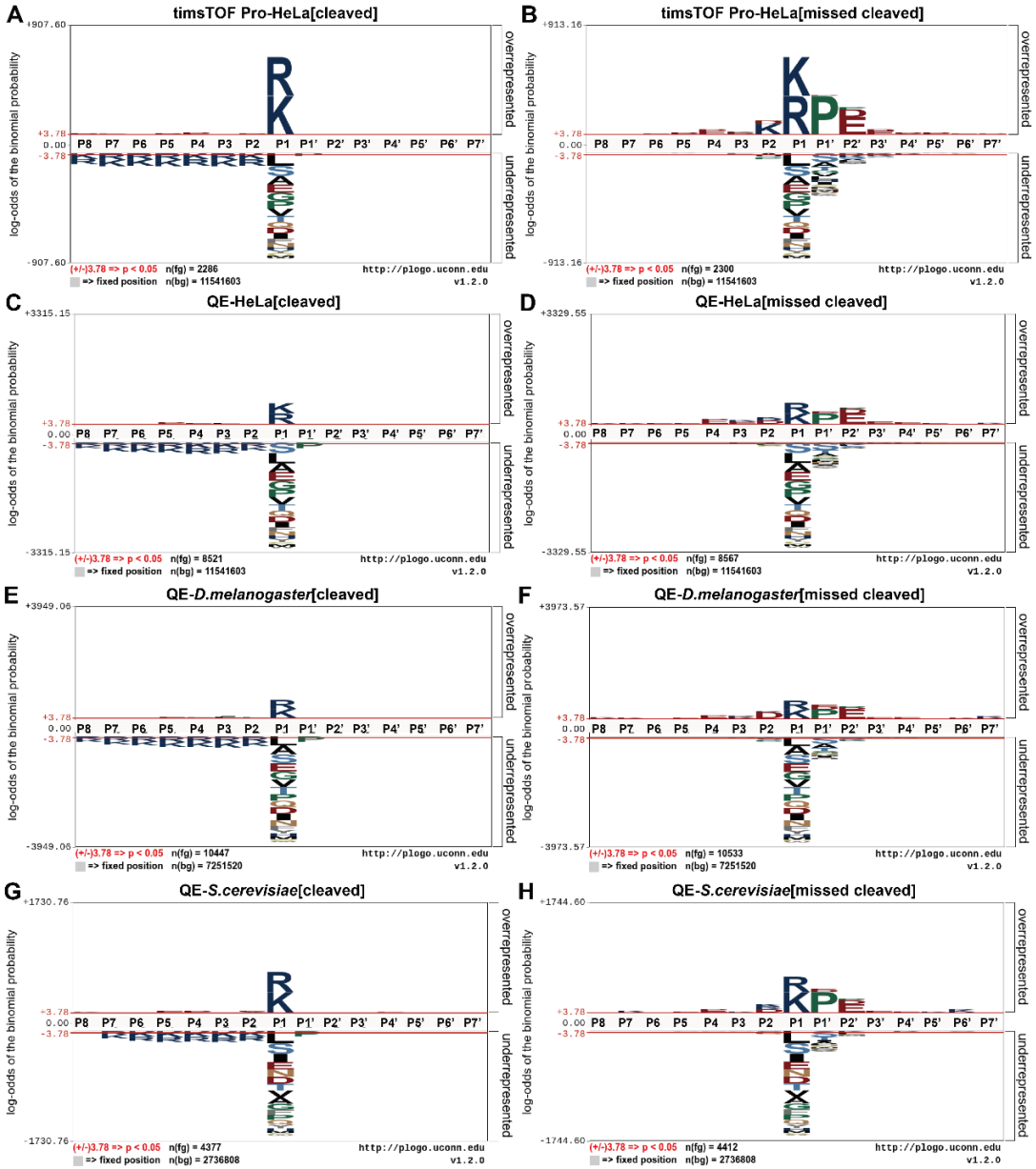


Figure S4. The probabilistic motifs for cleavage windows of cleaved and missed cleaved sites. The probabilistic motifs of cleavage windows of (A) cleaved sites and (B) missed cleaved sites based on the datasets of HeLa from timsTOF Pro. The probabilistic motifs of cleavage windows of (C) cleaved sites and (D) missed cleaved sites based on the datasets of HeLa from Q Exactive. The probabilistic motifs of cleavage windows of (E) cleaved sites and (F) missed cleaved sites based on the datasets of *D.melanogaster*. The probabilistic motifs of cleavage windows of (G) cleaved sites and (H) missed cleaved sites based on the datasets of *S.cerevisiae*.

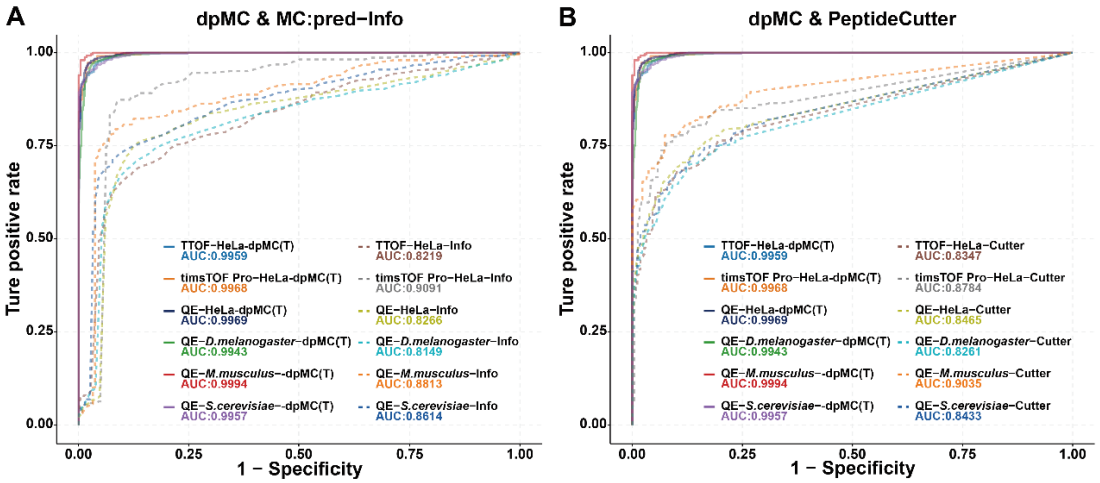


Figure S5. The ROC curves of dpMC, information theory and PeptideCutter. Each ROC curve is based on the performance on corresponding holdout testing datasets. dpMC with fine-tuning is denoted with 'T' in the bracket. (A) The ROC curves of dpMC with fine-tuning are denoted as 'dpMC(T)' and information theory from MC:pred as 'Info'. The performance of dpMC with fine-tuning is plotted in solid lines while dotted lines for information theory. (B) The ROC curves of dpMC with fine-tuning are denoted as 'dpMC(T)' and PeptideCutter as 'Cutter'. The performance of dpMC with fine-tuning is plotted in solid lines while dotted lines for PeptideCutter.

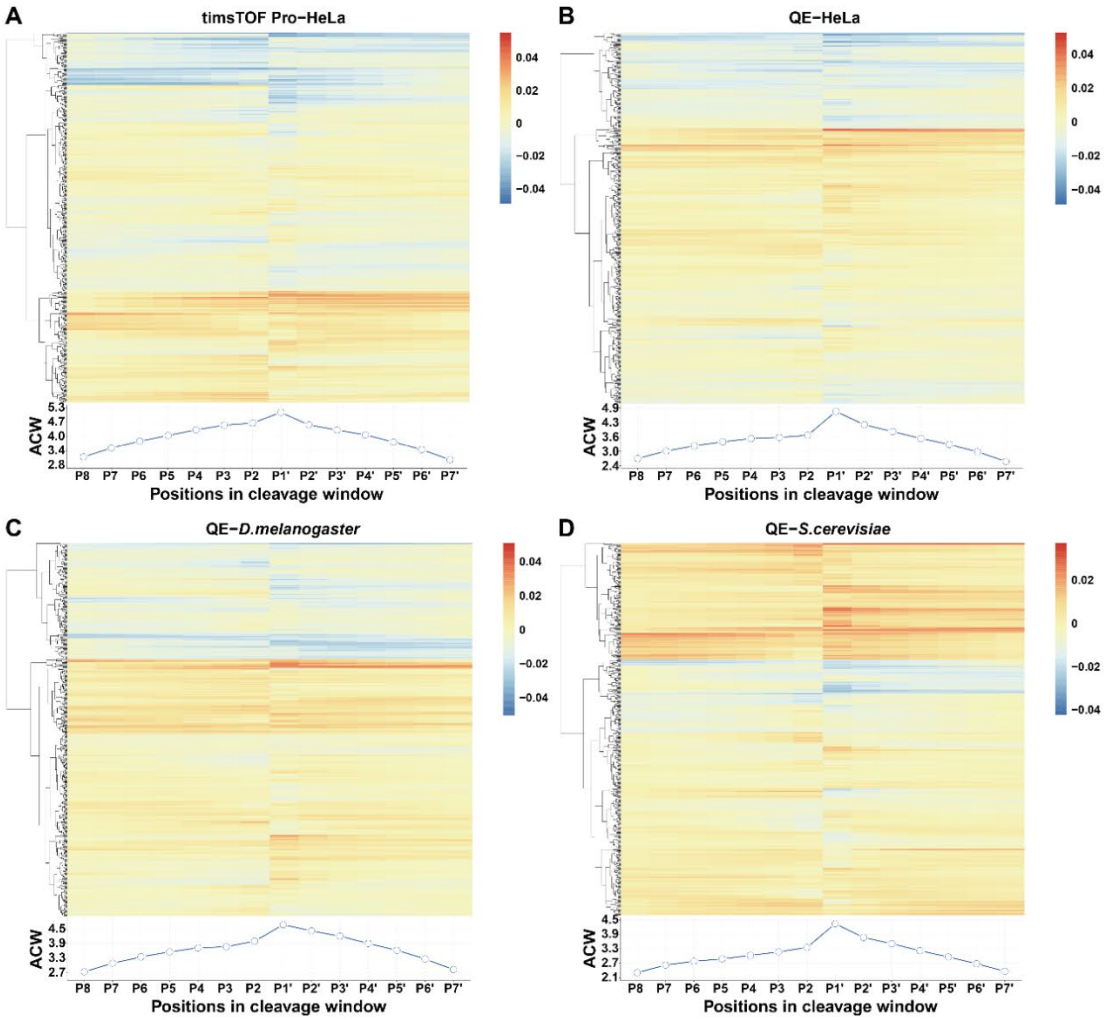


Figure S6. The heatmap and line plot of feature maps extracted from Bi-LSTM layer. The heatmaps of activation values from 512 neurons of Bi-LSTM in model trained on datasets of (A) HeLa from timsTOF Pro, (B) HeLa from Q Exactive, (C) *D.melanogaster* from timsTOF Pro and (D) *S.cerevisiae* from Q Exactive are shown in upper part, the line plots of Absolute Culumative Weights (ACW) of the feature maps are shown below the heatmaps. The positions of amino acids in the cleavage window for both heatmap and line plot are plotted in corresponding to each other.

Table S1. Performance of dpMC evaluated on holdout testing datasets with optimzed cutoffs.

Instrument	Species (data source)	Fine-tuning	Sensitivity	Specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H.sapiens* (HeLa)	No	0.9819	0.9540	0.9558	0.9687	0.9362	0.9959
TTOF 5600	*H.sapiens* (HeLa)	Yes	0.9637	0.9747	0.9748	0.9692	0.9385	0.9959
timsTOF Pro	*H.sapiens* (HeLa)	No	0.9819	0.9544	0.9518	0.9666	0.9366	0.9957
timsTOF Pro	*H.sapiens* (HeLa)	Yes	0.9729	0.9793	0.9773	0.9751	0.9521	0.9968
Q Exactive	*H.sapiens* (HeLa)	No	0.9768	0.9663	0.9667	0.9717	0.9431	0.9959
Q Exactive	*H.sapiens* (HeLa)	Yes	0.9779	0.9721	0.9723	0.9751	0.9500	0.9969
Q Exactive	*D.melanogaster*	No	0.9736	0.9527	0.9537	0.9636	0.9265	0.9938
Q Exactive	*D.melanogaster*	Yes	0.9717	0.9688	0.9689	0.9703	0.9405	0.9943
Q Exactive	*M.musculus*	No	0.9597	0.9907	0.9917	0.9754	0.9509	0.9976
Q Exactive	*M.musculus*	Yes	0.9798	0.9954	0.9959	0.9878	0.9753	0.9994
Q Exactive	*S.cerevisiae*	No	0.9589	0.9766	0.9779	0.9683	0.9357	0.9940
Q Exactive	*S.cerevisiae*	Yes	0.9481	0.9790	0.9799	0.9637	0.9275	0.9957

Performance of dpMC on 6 independent datasets from different species and instruments are shown. PPV, Positive Predicted Value; MCC, Matthews Correlation Coefficient; AUC, Area Under Curve.

Table S2. Performance of dpMC evaluated by 10 folds cross-validation with cutoff as 0.5.

Instrument	Species (data source)	Fine-tuning	Sensitivity	Specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H.sapiens* (HeLa)	No	0.9569	0.9500	0.9500	0.9531	0.9074	0.9920
TTOF 5600	*H.sapiens* (HeLa)	Yes	0.9567	0.9566	0.9567	0.9566	0.9135	0.9921
timsTOF Pro	*H.sapiens* (HeLa)	No	0.9536	0.9632	0.9634	0.9583	0.9171	0.9917
timsTOF Pro	*H.sapiens* (HeLa)	Yes	0.9654	0.9620	0.9629	0.9640	0.9275	0.9930
Q Exactive	*H.sapiens* (HeLa)	No	0.9697	0.9637	0.9638	0.9667	0.9334	0.9949
Q Exactive	*H.sapiens* (HeLa)	Yes	0.9661	0.9688	0.9689	0.9675	0.9351	0.9951
Q Exactive	*D.melanogaster*	No	0.9562	0.9650	0.9646	0.9603	0.9213	0.9929
Q Exactive	*D.melanogaster*	Yes	0.9569	0.9648	0.9646	0.9607	0.9219	0.9931
Q Exactive	*M.musculus*	No	0.9732	0.9518	0.9526	0.9627	0.9254	0.9941
Q Exactive	*M.musculus*	Yes	0.9643	0.9779	0.9770	0.9704	0.9425	0.9967
Q Exactive	*S.cerevisiae*	No	0.9441	0.9604	0.9598	0.9517	0.9050	0.9904
Q Exactive	*S.cerevisiae*	Yes	0.9501	0.9574	0.9569	0.9534	0.9076	0.9912

The mean values of each measurement index based on 10-fold cross-validation are shown for each dataset. PPV, Positive Predicted Value; MCC, Matthews Correlation Coefficient; AUC, Area Under Curve.

Table S3. Performance of dpMC evaluated by 10-fold cross-validation with optimized cutoff.

Instrument	Species (data source)	Fine-tuning	Sensitivity	Specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H.sapiens* (HeLa)	No	0.9550	0.9621	0.9617	0.9582	0.9173	0.9920
TTOF 5600	*H.sapiens* (HeLa)	Yes	0.9658	0.9559	0.9563	0.9610	0.9220	0.9921
timsTOF Pro	*H.sapiens* (HeLa)	No	0.9712	0.9589	0.9605	0.9657	0.9304	0.9917
timsTOF Pro	*H.sapiens* (HeLa)	Yes	0.9672	0.9686	0.9697	0.9683	0.9362	0.9930
Q Exactive	*H.sapiens* (HeLa)	No	0.9751	0.9626	0.9633	0.9691	0.9379	0.9949
Q Exactive	*H.sapiens* (HeLa)	Yes	0.9701	0.9697	0.9698	0.9699	0.9398	0.9951
Q Exactive	*D.melanogaster*	No	0.9632	0.9629	0.9630	0.9631	0.9262	0.9929
Q Exactive	*D.melanogaster*	Yes	0.9625	0.9664	0.9664	0.9643	0.9291	0.9931
Q Exactive	*M.musculus*	No	0.9654	0.9737	0.9726	0.9688	0.9394	0.9941
Q Exactive	*M.musculus*	Yes	0.9764	0.9792	0.9787	0.9775	0.9557	0.9967
Q Exactive	*S.cerevisiae*	No	0.9423	0.9697	0.9686	0.9552	0.9125	0.9904
Q Exactive	*S.cerevisiae*	Yes	0.9448	0.9713	0.9707	0.9574	0.9167	0.9912

The mean values of each measurement index based on 10 folds cross-validation are shown for each dataset.

Table S4. Performance of three existing programs evaluated on holdout testing datasets with cutoff as 0.5.

MC:pred-DeepDigest

Instrument	Species (data source)	Sensitivity	Specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H.sapiens* (HeLa)	0.9796	0.8253	0.8504	0.9104	0.8146	0.9806
timsTOF Pro	*H.sapiens* (HeLa)	0.9457	0.6556	0.7158	0.8148	0.6283	0.9264
Q Exactive	*H.sapiens* (HeLa)	0.9710	0.7395	0.7887	0.8704	0.7303	0.9582
Q Exactive	*D.melanogaster*	0.9585	0.7833	0.8159	0.8815	0.7534	0.9565
Q Exactive	*M.musculus*	1.0000	0.7083	0.7974	0.8873	0.7405	0.9783
Q Exactive	*S.cerevisiae*	0.9372	0.7079	0.7760	0.8490	0.6628	0.9314

MC:pred-SVM

Instrument	Species (data source)	Sensitivity	Specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H.sapiens* (HeLa)	0.9728	0.7241	0.7814	0.8667	0.7195	0.9570
timsTOF Pro	*H.sapiens* (HeLa)	1.0000	0.6556	0.7270	0.8419	0.6983	0.9699
Q Exactive	*H.sapiens* (HeLa)	0.9942	0.6907	0.7629	0.8633	0.7188	0.9671
Q Exactive	*D.melanogaster*	0.9556	0.7360	0.7839	0.8613	0.7090	0.9458
Q Exactive	*M.musculus*	0.9919	0.6852	0.7834	0.8754	0.7114	0.9692
Q Exactive	*S.cerevisiae*	0.9675	0.6822	0.7667	0.8555	0.6780	0.9471

MC:pred-Info

Instrument	Species (data source)	Sensitivity	Specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H.sapiens* (HeLa)	0.5261	0.9471	0.9098	0.6667	0.5217	0.8219
timsTOF Pro	*H.sapiens* (HeLa)	0.7873	0.9295	0.9110	0.8447	0.7241	0.9091
Q Exactive	*H.sapiens* (HeLa)	0.5947	0.9279	0.8920	0.7136	0.5542	0.8266
Q Exactive	*D.melanogaster*	0.5449	0.9442	0.9072	0.6808	0.5334	0.8149
Q Exactive	*M.musculus*	0.7177	0.9583	0.9519	0.8184	0.6965	0.8813
Q Exactive	*S.cerevisiae*	0.5758	0.9626	0.9433	0.7151	0.5838	0.8614

PeptideCutter

Instrument	Species (data source)	Sensitivity	Specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H.sapiens* (HeLa)	0.2925	1.0000	1.0000	0.4526	0.4139	0.8347
timsTOF Pro	*H.sapiens* (HeLa)	0.4434	0.9876	0.9703	0.6087	0.5137	0.8784
Q Exactive	*H.sapiens* (HeLa)	0.3368	0.9965	0.9898	0.5026	0.4435	0.8465
Q Exactive	*D.melanogaster*	0.3107	0.9972	0.9910	0.4730	0.4233	0.8261
Q Exactive	*M.musculus*	0.4758	1.0000	1.0000	0.6448	0.5587	0.9034
Q Exactive	*S.cerevisiae*	0.3290	1.0000	1.0000	0.4951	0.4437	0.8433

Table S5. Performance of three existing programs evaluated on holdout testing datasets with optimized cutoff.

MC:pred-DeepDigest

Instrument	Species (data source)	Sensitivity	Specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H.sapiens* (HeLa)	0.9274	0.9310	0.9317	0.9295	0.8585	0.9806
timsTOF Pro	*H.sapiens* (HeLa)	0.8552	0.8672	0.8552	0.8552	0.7225	0.9264
Q Exactive	*H.sapiens* (HeLa)	0.9431	0.8349	0.8512	0.8948	0.7826	0.9582
Q Exactive	*D.melanogaster*	0.8933	0.8903	0.8908	0.8920	0.7836	0.9565
Q Exactive	*M.musculus*	0.9435	0.9120	0.9249	0.9341	0.8560	0.9783
Q Exactive	*S.cerevisiae*	0.8463	0.8551	0.8631	0.8546	0.7015	0.9314

MC:pred-SVM

Instrument	Species (data source)	Sensitivity	Specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H.sapiens* (HeLa)	0.9070	0.8966	0.8989	0.9029	0.8036	0.9570
timsTOF Pro	*H.sapiens* (HeLa)	0.9819	0.8506	0.8577	0.9156	0.8398	0.9699
Q Exactive	*H.sapiens* (HeLa)	0.9152	0.9128	0.9131	0.9142	0.8280	0.9671
Q Exactive	*D.melanogaster*	0.8754	0.9044	0.9018	0.8884	0.7801	0.9458
Q Exactive	*M.musculus*	0.8992	0.9630	0.9654	0.9311	0.8639	0.9692
Q Exactive	*S.cerevisiae*	0.9113	0.8435	0.8627	0.8863	0.7565	0.9471

MC:pred-Info

Instrument	Species (data source)	Sensitivity	Specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H.sapiens* (HeLa)	0.6667	0.8897	0.8596	0.7510	0.5707	0.8219
timsTOF Pro	*H.sapiens* (HeLa)	0.8688	0.9129	0.9014	0.8848	0.7824	0.9091
Q Exactive	*H.sapiens* (HeLa)	0.7573	0.8663	0.8501	0.8010	0.6273	0.8266
Q Exactive	*D.melanogaster*	0.6874	0.8903	0.8626	0.7651	0.5900	0.8149
Q Exactive	*M.musculus*	0.7944	0.9167	0.9163	0.8510	0.7164	0.8813
Q Exactive	*S.cerevisiae*	0.7143	0.9229	0.9091	0.8000	0.6515	0.8614

PeptideCutter

Instrument	Species (data source)	Sensitivity	Specificity	PPV	F1-score	MCC	AUC
TTOF 5600	*H.sapiens* (HeLa)	0.6825	0.8989	0.8725	0.7659	0.5955	0.8347
timsTOF Pro	*H.sapiens* (HeLa)	0.7738	0.9004	0.8769	0.8221	0.6796	0.8784
Q Exactive	*H.sapiens* (HeLa)	0.7398	0.8581	0.8393	0.7864	0.6022	0.8465
Q Exactive	*D.melanogaster*	0.6893	0.8742	0.8459	0.7596	0.5734	0.8261
Q Exactive	*M.musculus*	0.7782	0.9259	0.9234	0.8446	0.7120	0.9034
Q Exactive	*S.cerevisiae*	0.7208	0.8668	0.8538	0.7817	0.5940	0.8433

Table S6. The significant motifs around cleaved sites identified by kpLogo. The significance of motifs refers to the *corrected.p,* more significant if the value is higher. The positive statistics indicate the enriched motifs while the negative indicates the depleted ones.

Table S7. The significant motifs around missed cleaved sites identified by kpLogo.

Supplementary Note. Performance comparison of dpMC with other existing tools for predictions of missed tryptic cleavages.

All the performance comparisons of dpMC with other existing tools were done using the same corresponding holdout dataset of the six datasets (Table 1) in this study. For the performance on missed tryptic cleavages predictions by dpMC was compared with DeepDigest[1] (access date 2021-04 http://fugroup.amss.ac.cn/software/DeepDigest/DeepDigest.html). The testing model was provided along with the code by DeepDigest and trypsin associated files were utilized (i.e. Trypsin.h5 and Trypsin.json). For the code running, parameters in the command line tool were specified as default but the number of missed cleavages was set as 8 (--missedcleavages=8) to include all possible missed cleavages in the testing datasets.

Also, we compared dpMC with both SVM[2] and information theory[3] approaches provided within MC:pred (access date 2021-04 http://king.smith.man.ac.uk/mcpred/). For the comparison with SVM or information theory, the predictor was selected as "SVM" or "Info theory" correspondingly. Besides, we also compared dpMC with PeptideCutter (access date 2021-04 https://www.expasy.org/resources/peptidecutter) from ExPASy[4], and used the function of the "sophisticated model" on trypsin by selecting corresponding options.

For all the above existing tools, only the predictions for each missed cleavage at P1 in cleavage windows were used for further comparison with dpMC. No training or refinement functions for specific datasets or for certain experimental conditions were provided by the developers. Based on our studies we think that for an optimal performance experiment-specific training models are required, which is provided by dpMC.

1. Yang, J.; Gao, Z.; Ren, X.; Sheng, J.; Xu, P.; Chang, C.; Fu, Y., DeepDigest: Prediction of Protein Proteolytic Digestion with Deep Learning. *Anal Chem* **2021,** *93* (15), 6094–6103.
2. Siepen, J. A.; Keevil, E. J.; Knight, D.; Hubbard, S. J., Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *J Proteome Res* **2007,** *6* (1), 399-408.
3. Lawless, C.; Hubbard, S. J., Prediction of missed proteolytic cleavages for the selection of surrogate peptides for quantitative proteomics. *OMICS* **2012,** *16* (9), 449-56.
4. Artimo, P.; Jonnalagedda, M.; Arnold, K.; Baratin, D.; Csardi, G.; de Castro, E.; Duvaud, S.; Flegel, V.; Fortier, A.; Gasteiger, E.; Grosdidier, A.; Hernandez, C.; Ioannidis, V.; Kuznetsov, D.; Liechti, R.; Moretti, S.; Mostaguir, K.; Redaschi, N.; Rossier, G.; Xenarios, I.; Stockinger, H., ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res* **2012,** *40* (Web Server issue), W597-603.

7. Improving SWATH-MS analysis by Deep-learning（Paper III)

Sequential windowed acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH-MS) is a popular approach for MS analysis, which is a DIA method that can be applied at an unprecedented speed. However, such analysis needs high quality and extensive searching space to cover all the theoretical peptide candidates. Generally, the search libraries are created by data-dependent acquisition (DDA) experiments. In order to improve the search space of SWATH-MS analysis, we developed the tool for building a high-quality theoretical library for SWATH-MS analysis.

Bo Sun, Pawel Smialowski, Wasim Aftab, Andreas Schmidt, Ignasi Forne, Tobias Straub, and Axel Imhof. 2022. “Improving SWATH-MS analysis by Deep Learning.” *Proteomics*, 2022, doi: 10.1002/pmic.202200179.

Received: 5 May 2022 | Revised: 22 November 2022 | Accepted: 21 December 2022

DOI: 10.1002/pmic.202200179

Proteomics
Proteomics and Systems Biology

RESEARCH ARTICLE

Improving SWATH-MS analysis by deep-learning

Bo Sun[1] | Pawel Smialowski[2,3] | Wasim Aftab[1] | Andreas Schmidt[1] | Ignasi Forne[1] | Tobias Straub[3] | Axel Imhof[1]

[1] Faculty of Medicine, Biomedical Center, Protein Analysis Unit, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany

[2] Institute of Stem Cell Research, Helmholtz Center Munich, German Research Center for Environmental Health, Germany

[3] Faculty of Medicine, Biomedical Center, Computational Biology Unit, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany

Correspondence
Biomedical Center, Protein Analysis Unit, Faculty of Medicine, Ludwig-Maximilians-Universität München, Großhaderner Strasse 9, 82152 Planegg-Martinsried, Germany.
Email: imhof@lmu.de

Funding information
Chinese Scholarship Council, Grant/Award Number: 201506230154; Deutsche Forschungsgemeinschaft, Grant/Award Numbers: 2133249687, 219249687, 325871075; German Federal Ministry of Education and Research, Grant/Award Number: BMBF FKZ161L0214F

Abstract

Data-independent acquisition (DIA) of tandem mass spectrometry spectra has emerged as a promising technology to improve coverage and quantification of proteins in complex mixtures. The success of DIA experiments is dependent on the quality of spectral libraries used for data base searching. Frequently, these libraries need to be generated by labor and time intensive data dependent acquisition (DDA) experiments. Recently, several algorithms have been published that allow the generation of theoretical libraries by an efficient prediction of retention time and intensity of the fragment ions. Sequential windowed acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH-MS) is a DIA method that can be applied at an unprecedented speed, but the fragmentation spectra suffer from a lower quality than data acquired on Orbitrap instruments. To reliably generate theoretical libraries that can be used in SWATH experiments, we developed deep-learning for SWATH analysis (dpSWATH), to improve the sensitivity and specificity of data generated by Q-TOF mass spectrometers. The theoretical library built by dpSWATH allowed us to increase the identification rate of proteins compared to traditional or library-free methods. Based on our analysis we conclude that dpSWATH is a superior prediction framework for SWATH-MS measurements than other algorithms based on Orbitrap data.

KEYWORDS
proteomics, deep learning, spectral Library, data independent acquisition

1 | INTRODUCTION

The analysis of the proteomic composition of biological samples promises to provide a rich source of information, which could greatly improve our molecular understanding of a wide range of biological processes. It has the potential to revolutionize molecular diagnostics and treatment of disease. Despite a substantial improvement of the instruments (mostly mass spectrometers) used to perform proteomic measurements, the field still suffers from a substantial undersampling of peptides in shot gun proteomics studies (also called data dependent acquisition or DDA) and therefore a very low coverage of all possible peptides. To overcome this problem data independent acquisition (DIA) strategies have been developed that result in the fragmentation of all possible ions, which should (at least in theory) substantially improve peptide coverage. To achieve this task, extremely fast tandem mass spectrometers (such a quadrupol time of flight or Q-TOF instruments) need to be used, which results in a decrease of fragment spectrum

Abbreviations: BiLSTM, Bidirectional long-short term memory; CNN, Convolutional neural network; DDA, data-dependent acquisition; DIA, data-independent acquisition; dpMC, deep learning for missed cleavage; dpMS, deep learning for MS fragment ion prediction; dpRT, deep learning for retention time prediction; dpSWATH, deep learning for SWATH analysis; FDR, false discovery rate; LC-MS, Liquid chromatography coupled mass spectrometry; PCC, Pearson correlation coefficient; PSM, Peptide spectral matches; Q-TOF, Quadrupol Time of Flight; RNN, Recurrent neural network; RPKM, reads per kilobase per million mapped reads; SWATH-MS, sequential window acquisition of all theoretical mass spectra.

quality. One of these methods is the so called sequential window acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH-MS) using a quadrupole-TOF instrument [1]. In SWATH-MS mode, typically a precursor ion (MS1) spectrum is recorded, followed by a series of fragment ion (MS2) spectra recordings with wide precursor isolation windows (for example 25 m/z). A comprehensive data set is recorded through repeated cycling of consecutive precursor isolation windows over a defined mass range, which includes continuous information on all detectable fragment and precursor ions [1]. SWATH-MS has been implemented in many aspects of research, which including quantitative proteomics [2], clinical biomarker research [3], histone post-translational modification (PTM) analysis [4] and the analysis of protein-protein interactomes [5].

In addition to a better peptide coverage SWATH-MS also has advantages in reproducibility [6] and speed of analysis [7] and allows a retrospective targeting [1], which is not possible when using targeted workflows.

A disadvantage of all DIA methods is the requirement of specific fragment ion libraries for identification. Currently most of these libraries are experimentally generated using DDA measurements of a highly fractionated sample pool measured prior to SWATH-MS acquisition on the same instrument [8]. A lot of efforts have been put into building the assay library to improve the coverage and quality of proteomic research [9]. In 2016, J. Wu et al. have compared the SWATH mass spectrometry performance using local seed libraries integrated with external assay libraries and local assay libraries alone [10] and showed that the first one had a better performance with regard to peptide identification and quantification. In addition, software tools like SpectraST [11] have been developed to improve the building of consensus mass spectrum libraries [12].

Nowadays, deep-learning methods have empowered proteomic research. Especially the predictions based on the information inferred from peptide sequence have gained a lot of attention, such as the prediction of retention time [13] and fragment ion intensities [14, 15]. In addition to the prediction of peptide properties, deep-learning is also used for the identification of peptides and proteins. For example, the detection of LC-MS features is performed by deep-learning models [16]. Moreover, the deep-learning approach can also be used for de novo sequencing, such as the work that has been done by DeepNovo [17].

More recently, tools have been developed that allow an extension of the used libraries by applying both experimental [12, 18] and theoretical approaches [14, 15, 19]. However, most of the theoretical approaches used mass spectra recorded in an orbitrap instrument, which are of higher quality than the ones measured in a Q-TOF mass spectrometer. To improve the SWATH-MS analysis, we developed a novel framework and strategy to build high-quality in silico libraries by deep-learning.

2 | MATERIALS AND METHODS

2.1 | Datasets

2.1.1 | Datasets used for training and testing of dpSWATH

We used datasets generated by TripleTOF 5600 and 6600 (ABSciex, Concord, Ontario, Canada) from *Homo sapiens* and *Drosophila melanogaster*, respectively. We used the DDA datasets from the Pan-Human project (PXD000953) [12] as a pre-training datasets for the TripleTOF 5600 measurements. All peptide spectra matches (PSMs) of the Pan-Human project were extracted from file *PHL.pep.xml* and split into training and testing datasets. For the training dataset, we selected 2,000,029 unmodified PSMs containing 94,878 unique unmodified peptides. To test the model, we used 499,999 unmodified PSMs with 23,436 unique peptides. No DIA datasets were used for testing on Pan-Human project.

For further training and testing we used a DDA dataset of Hela extracts (PXD009273) [20]. To retrain the retention time and mass spectral models to build the in silico library, 12 DIA datasets were used for DIA searching followed by identification and quantification of the proteins.

For TripleTOF 6600, an aliquot corresponding to 500 μg of proteins *of a Drosophila* embryo extract [21] was precipitated with TCA. The protein pellet was dissolved in 6 M urea for subsequent protein cleavage by LysC and trypsin, disulfide reduction and alkylation with DTT and iodoacetamide, respectively. The obtained polypeptide mixture was desalted over C18 stage tips before further high pH-reversed phase separation. Individual fractions were injected onto an Exigent 425 nanoLC system, operated in micro-flow mode at 5 μl/min and separated on a 300 μm x 15 cm column directly coupled to the TripleTOF 6600 mass spectrometer (both ABSciex). For peptide separation a 50 min gradient from 2% to 35% acetonitrile in water was employed followed by 5 min washing at 80% acetonitrile. Peptides eluting from the column were detected in information-dependent detection mode acquiring a survey scan from 350 to 1500 m/z. Maximally 20 precursors with charge state 2+ or higher and a signal intensity of min. 160 counts were selected for MS/MS analysis to obtain high quality data for peptide identification. To further increase the number of detected peptides and proteins, DDA experiments of 72 fractions of a Drosophila embryo extract fractionated by size exclusion chromatography (Superose 6 10/300 GE Healthcare, Chicago, IL). All the PSM information was extracted using ProteinPilot (ABSciex, Concord, Ontario, Canada) or SpectroMine (Biognosys AG, Schlieren, Switzerland) and deposited on the Pride database (PXD038407). To evaluate the performance of dpSWATH, the precursors of 72 fractionated DDA runs including peptide sequences and precursor charges were extracted from

experimental library based on the searching results of Pulsar in Spectronaut (15.2.210819, Biognosys AG, Schlieren, Switzerland). For the 72 fractionated library, 3655 unique peptides were extracted to transfer-train the retention time and mass spectral models, the left 40,000 peptides with corresponding precursor charges were used to build the validation library with prediction by dpSWATH.

For all of the training, testing and validation datasets, the fragment ions were normalized which divided by the highest peak for each mass spectrum. The minimum and maximum length of peptides is 7 and 60 respectively and, the precursor charge ranges from 1 to 6 and a maximum charge of fragment ions of 2+.

2.1.2 | Datasets used for building the theoretical library

Fasta files of *D. melanogaster and H. sapiens* were downloaded from FlyBase (http://flybase.org/) and UniProt (https://www.uniprot.org/) respectively. Then the protein sequences were selected based on the entries recorded in the DDA libraries. For *D. melanogaster*, 5006 protein groups were extracted while 10524 and 4460 protein groups were extracted from the in Pan-Human library or the DDA experiment prepared from HeLa extracts respectively. The peptide sequences were prepared based on the cleavage standard rules of trypsin [22]. Up to two missed cleavages and cleavages followed by proline were predicted by dpMC. The length of the peptides is from 7 to 60 and the charges for peptides range from 2+ to 4+. Data of mRNA expression profiles for different stages of embryos of *D. melanogaster* (gene_rpkm_report_fb_ 2017_05.tsv) were downloaded from Flybase, while mRNA expression data for HeLa cell-lines were used from the ProteomeXchange repository (PXD009273) [20].

2.2 | Preprocessing of datasets for modeling (dpMScore)

All datasets were preprocessed using the newly developed dpMScore and used for both training and testing of the performance of dpSWATH (Figure 1). dpMScore uses hierarchical clustering to choose the most abundant and consistent fragmentation of each peptide. The dpMScore is calculated by the following formula:

$$\text{dpMScore} = -ln(Dist)\sum_{1}^{Nc}\prod_{i}^{Nc} pi \exp^{\prod_{1}^{Nc} pi\left(1-\prod_{1}^{Nc} pi\right)}$$

where Dist is the distance among different fragmentations for the same peptide, which ranges from 0.01 to 0.2 based on Pearson Correlation Coefficient (PCC); *Nc* is the number of clusters split at one certain bar; p_i is the proportion of i_{th} cluster calculated by the number of fragmentations in this cluster divided by the total number of fragmentations for this peptide, which ranges from 1 to *Nc*.

The dpMScore was only calculated for peptides that had more than three replicates whereas peptides with less than three replicates were kept in the training or testing datasets for dpSWATH without a score attached to it.

2.3 | Retention time prediction

As the prediction of retention time is crucial to SWATH-MS analysis, we developed dpRT as part of the dpSWATH program for a highly accurate retention time prediction and an increased sensitivity and identification of peptides and proteins (Figure 1, Figure S1). The framework of dpRT takes advantage of both convolutional neural network (CNN) and recurrent neural network (RNN) with self-attention mechanism (Figures 1 and S1). CNN performs very well on image and lingual work which benefits from its powerful feature extraction function. In dpRT, we use one-dimensional CNN as feature extractor to analyze the peptide sequence by setting the kernel size as 3. It is beneficial for next level RNN to use these features to predict the fragment ions' intensities. As for the RNN work, we chose two parallel bidirectional Long-Short Term Memory (BiLSTM) layers. The BiLSTM is very good at dealing with sequence or sentence cases, which has the advantage of processing information in both directions; for each predicted vector, BiLSTM makes the prediction combining the past and future states simultaneously. However, BiLSTM still shows lack of capability of dealing with long sequences, which could be complemented by the advantage of self-attention algorithm which is able to assign different weights to different features and has strong capability to deal with long sequences. Besides the BiLSTM layers, we also adopted self-attention layers to enhance the capability of model on dealing with the distant information along the sequences. Then two dense layers with 256 units and 1 unit respectively were connected to above RNN layers to generate the single predicted value.

2.4 | Fragment ion prediction

For the prediction of fragment ions, we developed dpMS. In dpMS, we also used one-dimensional CNN as feature extractor to analyze the peptide sequence by setting the kernel size as 2, in this way CNN could extract features from each two adjacent amino acids which have a strong and direct effect to the fragment ions that lies between them, which is beneficial for next level RNN to use these features to predict the fragment ions' intensities. As for the RNN work, we keep the similar architecture as dpRT but modify the units of RNN and self-attention layer with width as 49. For the fragment ions used to construct mass spectra, we take b ions and y ions that are generated by one time-distributed dense layer as the output layer of dpMS and the dimension of output is 59*4.

2.5 | DDA library generation

The search engine Pulsar in Spectronaut (15.2.210819, Biognosys AG, Schlieren, Switzerland) was used to build all the above the DDA

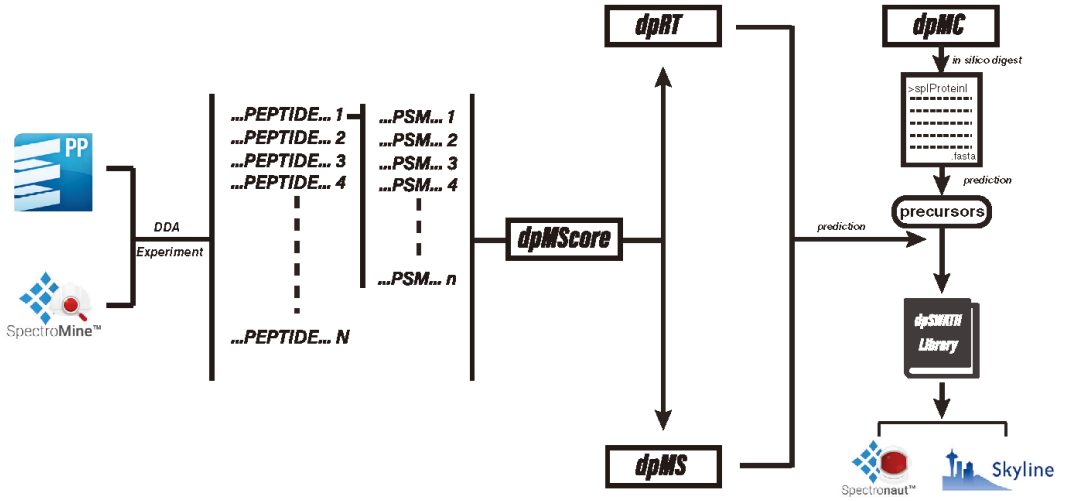


FIGURE 1 The workflow of dpSWATH and strategies applied in this study. Datasets from either ProteinPilot or SpectroMine can be analyzed, and the generated library can be used by Spectronaut or Skyline

libraries. The public Pan-Human library is pre-deposited in the Spectronaut software. To measure the performance of dpSWATH, we built the experimental library of unmodified peptides with length from 7 to 60 amino acids, precursor charges from 1+ to 6+ and, set Cysteine carbamidomethylation as fixed structural modification and no variable modification are selected. The maximum missed cleavage was set as 2.

2.6 | Construction of dpSWATH library

After the prediction of fragment ions' intensities and retention time, we assembled the two parts' results into .txt file which could be read by Spectronaut. The .txt file stores all available mass spectra to build the searching library. We put all of the 10 necessary information (Supplementary Note 1) of mass spectra including the predicted fragment ions and retention time which suggested by Spectronaut into the .txt file (Figure 1). Besides, we also prepared the script for building the library for Skyline.

3 | RESULTS

3.1 | Preprocessing of the datasets

Compared to an orbitrap mass spectrometer, the fragment spectra analyzed within a TripleTOF mass analyzer show a higher variability [11, 23]. The selection of the representative mass spectrum is therefore crucial for efficient identification and quantification of the corresponding peptide. Currently, most spectral libraries were built with the consensus mass spectra from PSMs using clustering algorithms such as SpectraST [11]. The selection of the consensus spectrum is often based on selecting the spectra with a minimal Q-value. However, for TripleTOF datasets, even PSMs with similar Q-values show big differences in the intensities of individual fragment ions as shown in Figure 2. For the training of dpSWATH, we therefore devised dpMScore to preprocess MS datasets and select the most abundant and consistent mass spectra for a given peptide with the same precursor charge. In dpMScore, we take the similarities among mass spectra into consideration and choose the cluster with the smallest distance and the largest number of spectra (see Section 2). In this way, the clusters of mass spectra were not only determined by the intensities, but also by the number of detected fragment ions and their proportions.

3.2 | Benchmarking of dpSWATH

Tandem MS spectra are strongly affected by many different experimental conditions ranging from sample preparation to instrumental set up to ambient environmental conditions such as temperature, humidity, or electrical interference (Gallien et al., 2013). We thereby designed dpSWATH in such a way that it can be trained and tested using data measured on multiple different instruments and under variable conditions and used transfer learning to construct reliable libraries.

To prepare a high-quality predicted library, the algorithm should therefore be able to efficiently extract associated features, which affect the mass spectrum pattern and retention time. To do this, we put the convolutional layer as the first layer to extract the features at a deep level automatically. To address the issue of identifying very long peptides (e.g., longer than 40 amino acids), we also used a self-attention layer to deal with longer sequence peptides (Figure S1).

First, we split the Pan-Human datasets from TripleTOF 5600 into training, validation, and testing datasets into a ratio of 8:1:1. By

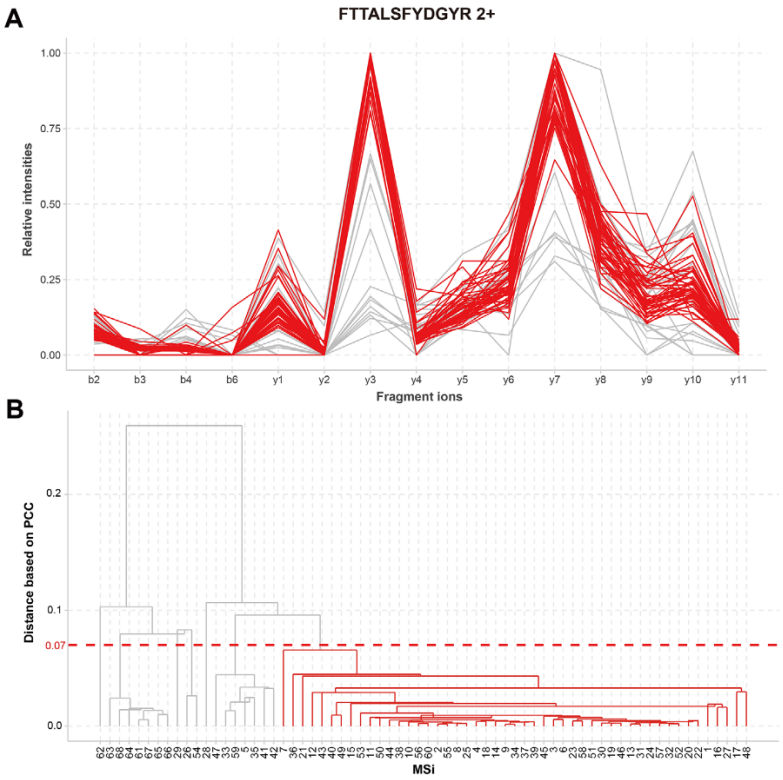


FIGURE 2 Different mass spectra patterns for the same precursor on the same condition. (A) The line plot of mass spectra of peptide "FTTALSFYDGYR" with precursor charge 2, the fragment ions' names are shown in x-axis with relative intensities in y-axis. All mass spectra pattern for this peptide are shown, the red patterns are the cluster chosen based on the dpMScore, the left gray pattern are filtered out by dpMScore for this peptide; (B) the clustering diagram for this peptide, the index of different mass spectra for this peptide are shown in x-axis with the distance among different mass spectra shown in y-axis. The chosen cluster are shown red corresponding to the red mass spectra in (A), which are chosen on the threshold at distance based on PCC 0.07

applying dpMScore as described in the methods, dpMS has achieved a median Pearson Correlation Coefficient (PCC) of 0.968 and median dot-product of 0.973 between the observed and predicted mass spectra (Figure 3A). For all validation and testing datasets, the peptides were not shown in the training datasets. We then applied transfer-learning on human datasets of TripleTOF 6600 with the trained model on TripleTOF 5600 to predict the fragmentation spectra of 57157 peptides from *D. melanogaster*. When doing this, we achieved a median Pearson Correlation Coefficient (PCC) of 0.980 and median dot-product 0.983 between observed and predicted mass spectra (Figure 3A,B). The similarities between observed and predicted mass spectra can directly affect the success of the identification and quantitation of proteins and peptides in the downstream analysis. Compared with DeepDIA on the same datasets, dpMS achieved much higher accuracy, which benefits the following analysis. Besides the higher accuracy given by dpMS, the capability of prediction for the longest sequence has been up to 60 and up to 6 of the highest precursor charges.

Then we applied the same strategy to the prediction of retention time. To eliminate the differences among different experiments and facilitate the prediction, we applied indexed retention time (iRT) throughout this study. The information of retention time can provide a reliable coordinate for mapping corresponding peptides [24, 25] and is usually combined with other analytical coordinates (m/z, intensity) for a reliable identification and quantification [25]. Therefore, we developed dpRT as part of the dpSWATH framework to facilitate the generation of building reliable in silico libraries (Figures 3C and S2).

Based on the high accurate prediction on mass spectra and retention time, we benchmarked the performance of dpSWATH by integrating the results from dpMS and dpRT on the validation datasets (see Section 2), which contains 40,000 peptides in the library. From the results, we got more peptides and proteins compared to the experimental 72 fractionated library and the library built by DeepDIA (Figure 3E,G). Compared to the experimental identified 28,940 peptides and 3301 protein groups, dpSWATH identified 31,012 peptides and 3545 protein groups, which are also more than the results from DeepDIA which

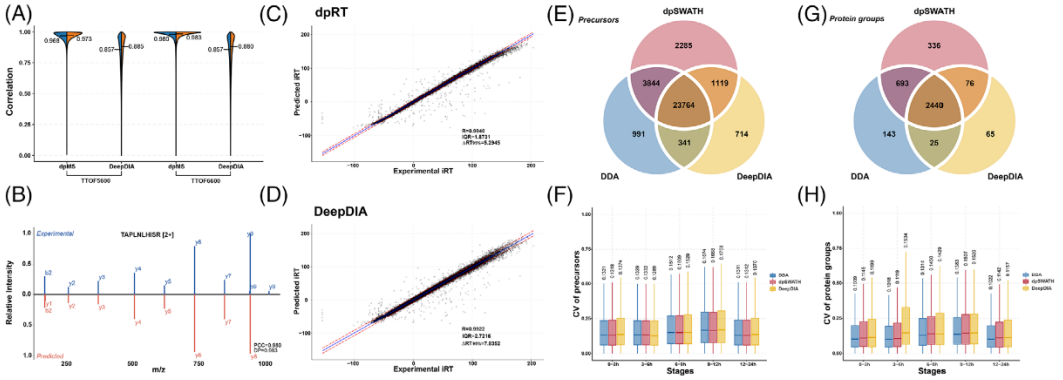


FIGURE 3 Benchmarking of dpSWATH. (A) The performance of dpMS and DeepDIA on datasets from TripleTOF 5600 and TripleTOF 6600, the blue histograms show the distribution of PCC, while the orange histograms show the distribution of dot-product, the median PCC and dot-product are shown. (B) The mirror plot for peptide 'TAPLNLHISR' with precursor charge of 2. The experimental mass spectra is shown in upper blue while the predicted shown in lower red. (C) The prediction of retention time by dpRT on datasets of *D. melanogaster*, the correlation of PCC, interquantile range (IQR) and distance of 95% datapoints are shown; (D) the prediction of retention time by DeepDIA on the same datasets as (C); (E) The overlapping of precursors among libraries of DDA, dpSWATH and DeepDIA. (F) The coefficient of variance (CV) of precursors for each stage of the embryo development in *D. melanogaster*. (G) The overlapping of protein groups among libraries of DDA, dpSWATH and DeepDIA. (H) The coefficient of variance (CV) of protein groups for each stage of the embryo development in *D. melanogaster*

identified 25,938 peptides and 2606 proteins. The libraries built by experimental (DDA) or theoretical approaches (dpSWATH, DeepDIA) are based on very different strategies. The DDA library was built on the identified PSMs of given precursors, which was based on the consensus mass spectra generation algorithm like SpectraST. For the library built by dpSWATH, the training process was based on the filtered PSMs, and then the mass spectra pattern and retention time were predicted by dpMS and dpRT, respectively. The library built by DeepDIA, only the PSMs with minimum Q-values were used for training which leads to relatively higher specificity but lower sensitivity.

To estimate the applicability of theoretical libraries, we also measured the coefficient of variance when quantifying protein groups from two technical replicates of five different developmental stages of *Drosophila* embryos (Figure 3F,H). In each case the CV is very similar between analyses made using the dpSWATH predicted library and derived from a DDA experiment (Figure S3A–S3F). Even when comparing the quantification of individual precursor ions both DDA and dpSWATH libraries performed equally well (Figure S3G). From this comparative analysis we conclude that dpSWATH not only identifies more peptides and protein groups, but also provides a robust and reproducible quantitation similar to the DDA approach but on this higher number of identified peptides.

3.3 | The interpretation of mass spectra on amino acids level

To understand the inner mechanism of our algorithm we investigated the amino acid contributions and therefore analyzed the impact of different amino acids on the prediction of the pattern of mass spectra.

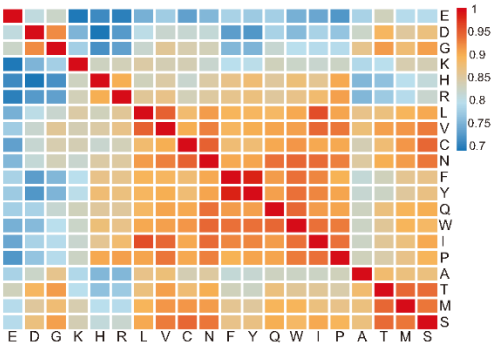


FIGURE 4 The heatmap of the correlation among amino acids based on their features

In the process of predicting mass spectra by dpMS, the properties of amino acids are encoded into each neuron, which is given different weights depending on the peptide sequence. The heatmap illustrates the weights of each amino acid assigned during predictions. We could see that some amino acids such as the aromatic amino acids F and Y cluster together due to their biochemical properties and structures (Figure 4).

3.4 | Missed cleavage prediction by dpMC

In proteomic analysis, trypsin is widely used to digest proteins into peptides. Despite being a robust and efficient protease

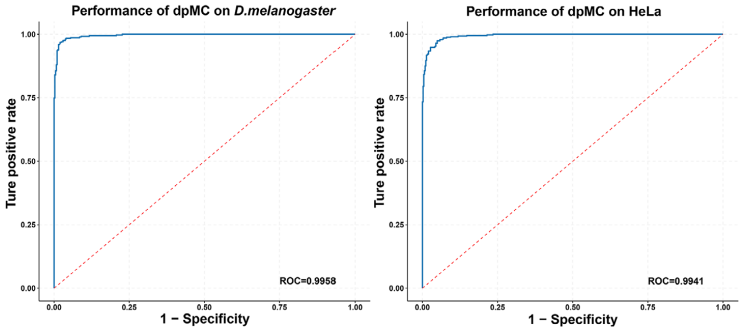


FIGURE 5 The performance of dpMC on datasets of *D. melanogaster* and HeLa

tryptic cleavage rarely reaches a100% efficiency. To predict the sites of inefficient cleavage most search engines use the Keil rules [22]. When it comes to building a large library based on entire proteomes, one problem is how to accurately predict missed cleavages. DeepDIA simply adopts the Keil rules to fully predict missed cleavages. To improve this prediction, we developed dpMC [26] (Figure 5). For the application of dpMC in dpSWATH, we also optimized parameters for the combinations of trypsin and LysC.

Besides, since the training of dpMC is based on the detected peptides in experiments, so the cleaved peptides also have the information of detectability, which is mentioned by AP3 [27]. Thus, the candidate peptides are most detectable for DIA analysis. In this way, we not only reduce the search time while maintaining a high specificity, but also improve the recovery rate and control FDR of theoretical libraries effectively.

3.5 | SWATH-MS analysis improvement using theoretical libraries generated by dpSWATH

Combined with the predictions of peptide fragment spectra by dpMS, retention time by dpRT and accurate missed cleavage sites by dpMC, we built an in silico library of all proteins detected in DDA experiments. To generate the theoretical library, we prepared the precursor candidates for each protein group detected in DDA library. This resulted in a library based on 5006 protein groups in *D. melanogaster*. We prepared the library using either the same peptide entries as observed in the DDA library, or the library predicted from the protein groups identified up to two missed cleaved sites by Keil rules or up to two missed cleaved sites predicted by dpMC. Then we searched data from the corresponding SWATH runs using these libraries with the same settings in Spectronaut. A comparison showed that we got the most identifications with the library built by dpSWATH with the predicted missed cleavages by dpMC. For the DDA based theoretical library, the protein groups' recovery rate of the library from 66.36% (3560/5006) to 95.65% (4788/5006) of the DDA library. We also performed the searching with directDIA 2.0 developed by Spectronaut, which showed only half of the identifications compared to the library built by dpSWATH (Figure 6).

Next, we built an in silico library referring to the peptide entries in Pan-Human library [12] and a DDA library of HeLa extracts [20] from TripleTOF 5600. We built the libraries with similar strategies except the library with up to two missed cleaved sites by Keil rules. Similar to our *Drosophila* data set, we got the most identifications of protein groups when we searched publicly available SWATH runs of HeLa extracts using a library built by dpSWATH with the predicted missed cleavages by dpMC (Figures 6C, S4C,E). Also in this case, the recovery rate of the Pan-Human library increased substantially from 32.40% (3410/10524) to 69.09% (7271/10524). The recovery rate increases when we use a library based on the entries from a DDA experiment performed on HeLa extracts, which is due to the same source. Even in this case the library built by dpSWATH performs better than the library built from experimental data only (86.32% (3850/4460) to 96.39% (4299/4460)) (Figure S4E). We also performed searches using directDIA 2.0 in Spectronaut, which resulted in far less identifications than by dpSWATH (Figure 6).

The prior DDA analysis to define the proteomic space used for the generation of a theoretical library was essential to keep a low FDR of the DIA search. In fact, when the library is generated from the entire proteome many DIA searches result in a low rate of peptide identifications and quantifications, which is often due to a high FDR. To limit the search space without the need of a prior extensive DDA measurement we built the in silico libraries based on transcriptomic data from the corresponding source. To do this, theoretical fragment spectra were generated from all protein candidates where the corresponding gene had an average number of reads per kilobase per million mapped reads (RPKM) [28] greater than or equal to 1. For the different developmental stages of *D. melanogaster*, this resulted in the inclusion of 17299 proteins. Compared with the DDA library, this strategy resulted in a much higher identification of protein groups (6156/3322), and peptides (64782/31538). The same effect is also observed when using the transcriptomic data from HeLa cells where we predicted the fragment spectra of peptides derived from 8758 proteins (Figure S4).

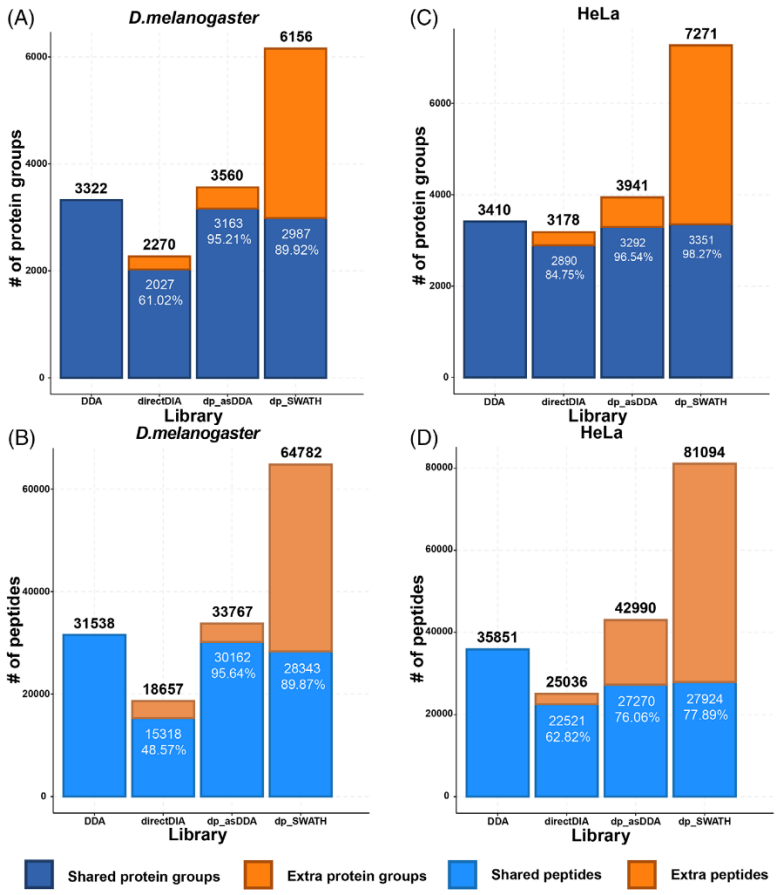


FIGURE 6 The identifications of protein groups and peptides from different libraries. (A) The number of identified protein groups on *D.melanogaster*. "DDA" indicates results from 72 fractionated DDA library; "directDIA" indicates the number of identified protein groups on *D. melanogaster* by directDIA 2.0; "dp_asDDA" indicates results from the in silico library on the same entries as DDA library; "dp_SWATH " indicates results from the in silico library on the digested FASTA sequence of the transcriptome based library by dpMC with up to two missed cleavages combined with no missed cleavages. (B) The number of identified peptides on *D. melanogaster*. (C) The number of identified protein groups on HeLa datasets from TripleTOF 5600 refer to the PanHuman library; "DDA" indicates results from experimental Pan-Human library; "directDIA" indicates the number of identified protein groups on HeLa datasets by directDIA 2.0; "dp_asDDA" indicates results from the in silico library on the same entries as experimental Pan-Human library; "dp_SWATH" indicates results from the in silico library on the digested FASTA sequence of the transcriptome based library by dpMC with up to two missed cleavages combined with no missed cleavages. (D) The number of identified peptides on HeLa datasets from TripleTOF 5600 compared to the PanHuman library. Identifications overlapped with the DDA-based libraries are denoted as "shared." Novel identifications by in silico libraries are denoted as "extra." The numbers and sensitivities of protein groups or peptides are shown

A detailed analysis of the correlation between the predicted mass spectra and the measured ones revealed the strong benefit of using dpMScore, which turned out to be crucial for building high quality libraries on Q-TOF datasets (Figure S5).

For the improved identifications, an estimate of the FDR control is crucial. We estimated the FDR by including predicted spectra from other species. Identifications from these species were counted as false positives. For these libraries from other species, we also digested the protein sequences with dpMC and predict the intensities and retention time by dpMS and dpRT respectively. We prepared the libraries of other species with the same number of proteins as the corresponding libraries built above for *D. melanogaster* and HeLa. We used a *S. cerevisiae* library of 5006 proteins which corresponds to *D. melanogaster* DDA library, the library of *C. elegans* and *D. discoideum* containing 10,524 proteins which corresponds to Pan-Human library, and the library of *S.cerevisiae* including 4460 proteins corresponding to the HeLa DDA library. For the transcriptome wide library, 17,299 proteins and 8758 proteins from *C. elegans* and *D. discoideum*

were prepared for entrapment library of *D.melanogaster* and HeLa, respectively.

We applied the entrapment strategy by pooling the entrapment libraries with their corresponding target libraries together to check the false positives, which revealed the false positives identified by the interferences of each other species. By calculation of the entrapments in the DIA searches based on DDA measurements or the transcriptome, we found the FDR was slightly higher when using larger libraries. For both DDA based libraries of *D. melanogaster* and HeLa, the FDR was around 1%, while it was around 2% for the transcriptome (Figure S6). Such a streamlining of the library is also intrinsically achieved by the use of an accurate prediction algorithm for missed cleavages such as dpMC. Based on the above FDR analysis, we showed the robustness of our method and strategy to buihighly accurate spectral libraries for SWATH-MS analysis.

In agreement with previous findings, the correlation (PCC) between the logarithmically transformed abundance of gene expression (RPKM) and protein intensities is rather moderate with a PCC value of 0.55 and 0.52 for *D. melanogaster* and HeLa respectively (Figure S7). Besides, for different scales of libraries built for *D. melanogaster*, the similarities between replicates for each stage of embryo development were also shown in Figure S8, in which the high correlations between replicates indicate the high quality of in silico libraries built by dpSWATH.

4 | DISCUSSION

The accurate theoretical prediction of peptide fragment spectra holds great promise for an improved quantification of entire proteomes using DIA methods such as SWATH-MS. Recently different models were developed to achieve a higher quality when predicting mass spectra. For example, Prosit [15] uses Collision Energy as an additional feature to train their model. However, for Q-TOF instruments the collision energy only marginally increases the accuracy of prediction [23], suggesting that many other subtle factors that could also affect the behavior of mass spectra. To consider such other, potentially unknown factors, we developed dpMScore to filter out the unreliable fragments spectra, which resulted in a more consistent and high-quality training and testing datasets for dpSWATH, in particular when using lower quality Q-TOF data.

The highly accurate prediction of mass spectra pattern and retention time makes SWATH-MS analysis methods more widely applicable. The reliable and effective workflow of dpSWATH, enables an fast generation and an efficient use of theoretical libraries. Based on the predicted library we built for *D. melanogaster* and *H. sapiens* (HeLa), we identified more proteins and peptides compared to an experimental library. This increase on the proteome coverage will favor a more comprehensive analysis of the biological system of interest.

During the development of the algorithm and its application to a wide range of data sets, we realized that the selection of consensus fragment mass spectra based on the dpMScore clustering algorithm is especially important for lower quality MS/MS spectra as the once recorded with non-trapping Q-TOF instruments. As these fragment spectra are substantially influenced by a various extrinsic factor such as the build of the instrument, humidity, temperature external electric fields et cetera, we suggest building the theoretical library based on the training datasets on the same platform and experimental conditions. Moreover, it turned out that the accuracy of peptide identification can be substantially improved by reducing the search space when building in silico libraries. In our proof-of concept studies we did this by applying a highly accurate prediction of missed tryptic cleavages using dpMC and a restriction to the proteins that are known to be expressed in the samples. The information about the proteins expressed in the studied sample(s) can be relatively easily gathered by RNA-Seq analysis or by a deep proteomic analysis of a pool of all samples. Based on our analysis, the transcriptome-based theoretical library showed the highest identification rate while maintaining FDR as the library based on a DDA measurement.

In summary, dpSWATH allows a robust and reliable prediction of fragment spectra that can be used in SWATH analyses therefore allowing a rapid and efficient quantification of a higher number of proteins and peptides compared to the classical DDA experiments or DIA experiments that rely on experimentally generated libraries.

ACKNOWLEDGEMENTS

B.S. was funded by the Chinese Scholarship Council (201506230154). Work in the lab of A.I. and T.S. were funded by grants of the Deutsche Forschungsgemeinschaft (CRC1064 (project number: 2133249687 (A.I.) and 219249687 (T.S.)) and 1309 (project number 325871075 (A.I.)) and the German Federal Ministry of Education and Research (BMBF FKZ161L0214F, ClinspectM).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The method and tools are open source on https://github.com/dpSWATH-sun/dpSWATH

ORCID

Axel Imhof https://orcid.org/0000-0003-2993-8249

REFERENCES

1. Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., & Aebersold, R. (2018). Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular Systems Biology*, *14*(8), e8126. https://doi.org/10.15252/msb.20178126
2. Krasny, L., Bland, P., Kogata, N., Wai, P., Howard, B. A., Natrajan, R. C., & Huang, P. H. (2017). SWATH mass spectrometry as a tool for quantitative profiling of the matrisome. *Scientific Reports*, *7*, 45913. https://doi.org/10.1038/srep45913
3. Liu, Y., Hüttenhain, R., Collins, B., & Aebersold, R. (2013). Mass spectrometric protein maps for biomarker discovery and clinical research. *Expert Review of Molecular Diagnostics*, *13*(8), 811–825. https://doi.org/10.1586/14737159.2013.845089
4. Sidoli, S., Lin, S., Xiong, L., Bhanu, N. V., Karch, K. R., Johansen, E., Hunter, C., Mollah, S., & Garcia, B. A. (2015). Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH) analysis for

characterization and quantification of histone post-translational modifications. *Molecular & Cellular Proteomics*, *14*(9), 2420–2428. https://doi.org/10.1074/mcp.O114.046102
5. Lambert, J.-P., Ivosev, G., Couzens, A. L., Larsen, B., Taipale, M., Lin, Z.-Y., Zhong, Q., Lindquist, S., Vidal, M., Aebersold, R., Pawson, T., Bonner, R., Tate, S., & Gingras, A.-C. (2013). Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nature Methods*, *10*(12), 1239–1245. https://doi.org/10.1038/nmeth.2702
6. Collins, B. C., Hunter, C. L., Liu, Y., Schilling, B., Rosenberger, G., Bader, S. L., Chan, D. W., Gibson, B. W., Gingras, A.-C., Held, J. M., Hirayama-Kurogi, M., Hou, G., Krisp, C., Larsen, B., Lin, L., Liu, S., Molloy, M. P., Moritz, R L., Ohtsuki, S., … Aebersold, R. (2017). Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nature Communications*, *8*(1), 291. https://doi.org/10.1038/s41467-017-00249-5
7. Messner, C B., Demichev, V., Bloomfield, N., Yu, J. S. L., White, M., Kreidl, M., Egger, A.-S., Freiwald, A., Ivosev, G., Wasim, F., Zelezniak, A., Jürgens, L., Suttorp, N., Sander, L. E., Kurth, F., Lilley, K S., Mülleder, M., Tate, S., & Ralser, M. (2021). Ultra-fast proteomics with Scanning SWATH. *Nature Biotechnology*, *39*(7), 846–854. https://doi.org/10.1038/s41587-021-00860-4 From NLM
8. Röst, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S. M., Schubert, O. T., Wolski, W., Collins, B. C., Malmström, J., Malmström, L., & Aebersold, R. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology*, *32*(3), 219–223. https://doi.org/10.1038/nbt.2841
9. Schubert, O. T., Gillet, L. C., Collins, B. C., Navarro, P., Rosenberger, G., Wolski, W. E., Lam, H., Amodei, D., Mallick, P., Maclean, B., & Aebersold, R. (2015). Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nature Protocols*, *10*(3), 426–441. https://doi.org/10.1038/nprot.2015.015
10. Wu, J. X., Song, X., Pascovici, D., Zaw, T., Care, N., Krisp, C., & Molloy, M. P. (2016). SWATH mass spectrometry performance using extended peptide MS/MS assay libraries. *Molecular & Cellular Proteomics*, *15*(7), 2501–2514. https://doi.org/10.1074/mcp.M115.055558
11. Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., Stein, S. E., & Aebersold, R. (2008). Building consensus spectral libraries for peptide identification in proteomics. *Nature Methods*, *5*(10), 873–875. https://doi.org/10.1038/nmeth.1254
12. Rosenberger, G., Koh, C. C., Guo, T., Röst, H. L., Kouvonen, P., Collins, B. C., Heusel, M., Liu, Y., Caron, E., Vichalkovski, A., Faini, M., Schubert, O T., Faridi, P., Ebhardt, H. A, Matondo, M., Lam, H., Bader, S L., Campbell, D S., Deutsch, E. W., … Aebersold, R. (2014). A repository of assays to quantify 10,000 human proteins by SWATH-MS, *Science Data*, *1*, 140031. https://doi.org/10.1038/sdata.2014.31
13. Yang, Y., Liu, X., Shen, C., Lin, Y., Yang, P., & Qiao, L. (2020). In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature Communications*, *11*(1), 146. https://doi.org/10.1038/s41467-019-13866-z
14. Zhou, X.-X., Zeng, W.-F., Chi, H., Luo, C., Liu, C., Zhan, J., He, S.-M, & Zhang, Z. (2017). pDeep: Predicting MS/MS spectra of peptides with deep learning. *Analytical Chemistry*, *89*(23), 12690–12697. https://doi.org/10.1021/acs.analchem.7b02566
15. Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., Reimer, U., Ehrlich, H.-C., Aiche, S., Kuster, B., & Wilhelm, M. (2019). Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, *16*(6), 509–518. https://doi.org/10.1038/s41592-019-0426-7
16. Kantz, E. D., Tiwari, S., Watrous, J. D., Cheng, S., & Jain, M. (2019). Deep neural networks for classification of LC-MS spectral peaks. *Analytical Chemistry*, *91*(19), 12407–12413. https://doi.org/10.1021/acs.analchem.9b02983 From NLM
17. Tran, N. H., Zhang, X., Xin, L., Shan, B., & Li, M. (2017). De novo peptide sequencing by deep learning. *PNAS*, *114*(31), 8247–8252. https://doi.org/10.1073/pnas.1705691114 From NLM
18. Wang, M., Wang, J., Carver, J., Pullman, B. S., Cha, S. W., & Bandeira, N. (2018). Assembling the community-scale discoverable human proteome. *Cell Systems*, *7*(4), 412–421.e5 e415. https://doi.org/10.1016/j.cels.2018.08.004
19. Guan, S., Moran, M. F., & Ma, B. (2019). Prediction of LC-MS/MS properties of peptides from sequence by deep learning. *Molecular & Cellular Proteomics*, *18*(10), 2099–2107. https://doi.org/10.1074/mcp.TIR119.001412
20. Liu, Y., Mi, Y., Mueller, T., Kreibich, S., Williams, E G., Van Drogen, A., Borel, C., Frank, M., Germain, P.-L., Bludau, I., Mehnert, M., Seifert, M., Emmenlauer, M., Sorg, I., Bezrukov, F., Bena, F. S., Zhou, H., Dehio, C., Testa, G.,… Aebersold, R. (2019). Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nature Biotechnology*, *37*(3), 314–322. https://doi.org/10.1038/s41587-019-0037-y
21. Völker-Albert, M. C., Pusch, M. C., Fedisch, A., Schilcher, P., Schmidt, A., & Imhof, A. (2016). A quantitative proteomic analysis of in vitro assembled chromatin. *Molecular & Cellular Proteomics*, *15*(3), 945–959. https://doi.org/10.1074/mcp.M115.053553
22. Keil, B. (1992). *Specificity of proteolysis* (p. 335). Springer-Verlag Berlin-Heidelberg.
23. Ammar, C., Berchtold, E., Csaba, G., Schmidt, A., Imhof, A., & Zimmer, R. (2019). Multi-reference spectral library yields almost complete coverage of heterogeneous LC-MS/MS data sets. *Journal of Proteome Research*, *18*(4), 1553–1566. https://doi.org/10.1021/acs.jproteome.8b00819
24. Searle, B. C., Swearingen, K. E., Barnes, C. A., Schmidt, T., Gessulat, S., Küster, B., & Wilhelm, M. (2020). Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nature Communications*, *11*(1), 1548. https://doi.org/10.1038/s41467-020-15346-1 From NLM
25. Van Puyvelde, B., Willems, S., Gabriels, R., Daled, S., De Clerck, L., Vande Casteele, S., Staes, A., Impens, F., Deforce, D., Martens, L., Degroeve, S., & Dhaenens, M. (2020). Removing the hidden data dependency of DIA with predicted spectral libraries. *Proteomics*, *20*(3-4), 1900306. https://doi.org/10.1002/pmic.201900306 From NLM
26. Sun, B., Smialowski, P., Straub, T., & Imhof, A. (2021). Investigation and highly accurate prediction of missed tryptic cleavages by deep learning. *Journal of Proteome Research*, *20*(7), 3749–3757. https://doi.org/10.1021/acs.jproteome.1c00346 From NLM
27. Gao, Z., Chang, C., Yang, J., Zhu, Y., & Fu, Y. (2019). AP3: An advanced proteotypic peptide predictor for targeted proteomics by incorporating peptide digestibility. *Analytical Chemistry*, *91*(13), 8705–8711. https://doi.org/10.1021/acs.analchem.9b02520
28. Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226 From NLM

SUPPORTING INFORMATION

Additional supporting information may be found online https://doi.org/10.1002/pmic.202200179 in the Supporting Information section at the end of the article.

How to cite this article: Sun, B., Smialowski, P., Aftab, W., Schmidt, A., Forne, I., Straub, T., & Imhof, A. (2023). Improving SWATH-MS analysis by deep-learning. *Proteomics*, e2200179. https://doi.org/10.1002/pmic.202200179

Supplementary information

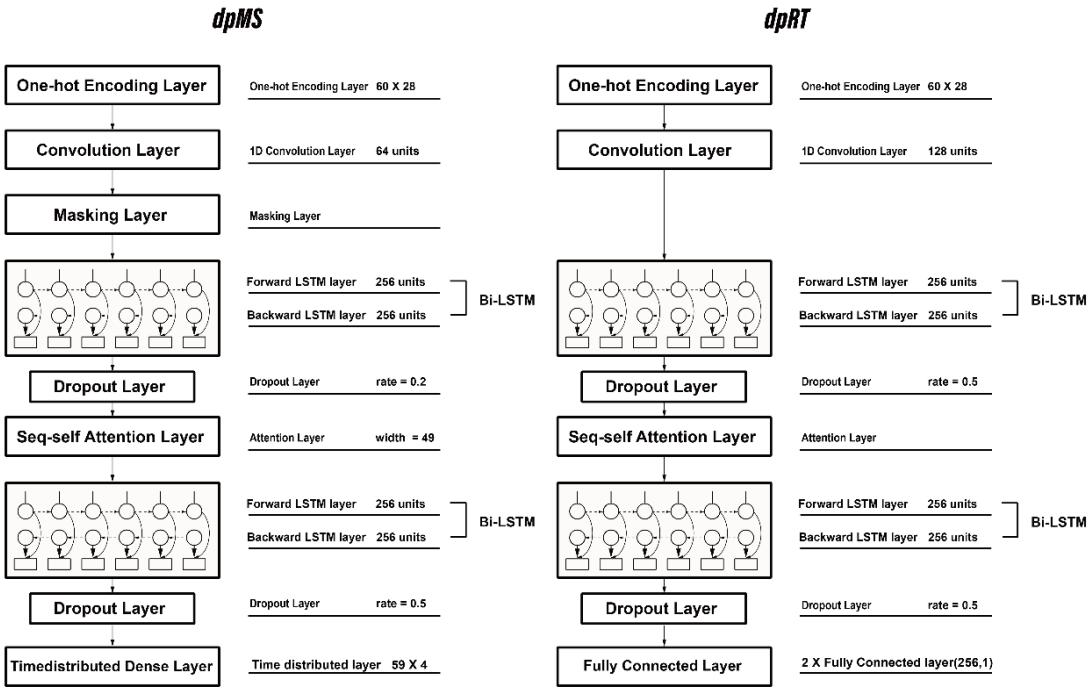


Figure S1. The architecture of dpMS and dpRT.

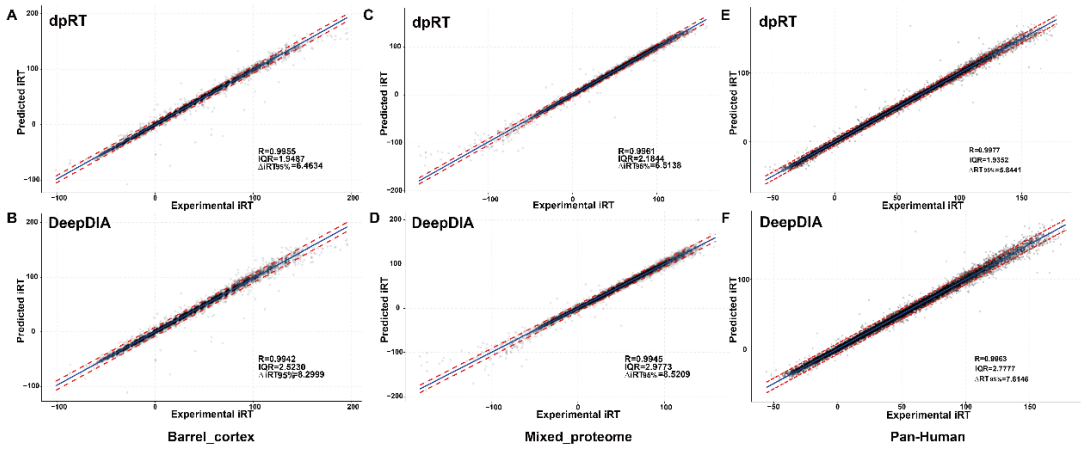


Figure S2. The performance of dpRT and DeepDIA on three different datasets. A, the performance of dpRT on barrel cortex datasets. B, the performance of DeepDIA on barrel cortex datasets. C, the performance of dpRT on Mixed proteome datasets. D, the performance of DeepDIA on Mixed proteome datasets. E, the performance of dpRT on Pan-Human datasets. F, the performance of DeepDIA on Pan-Human datasets.

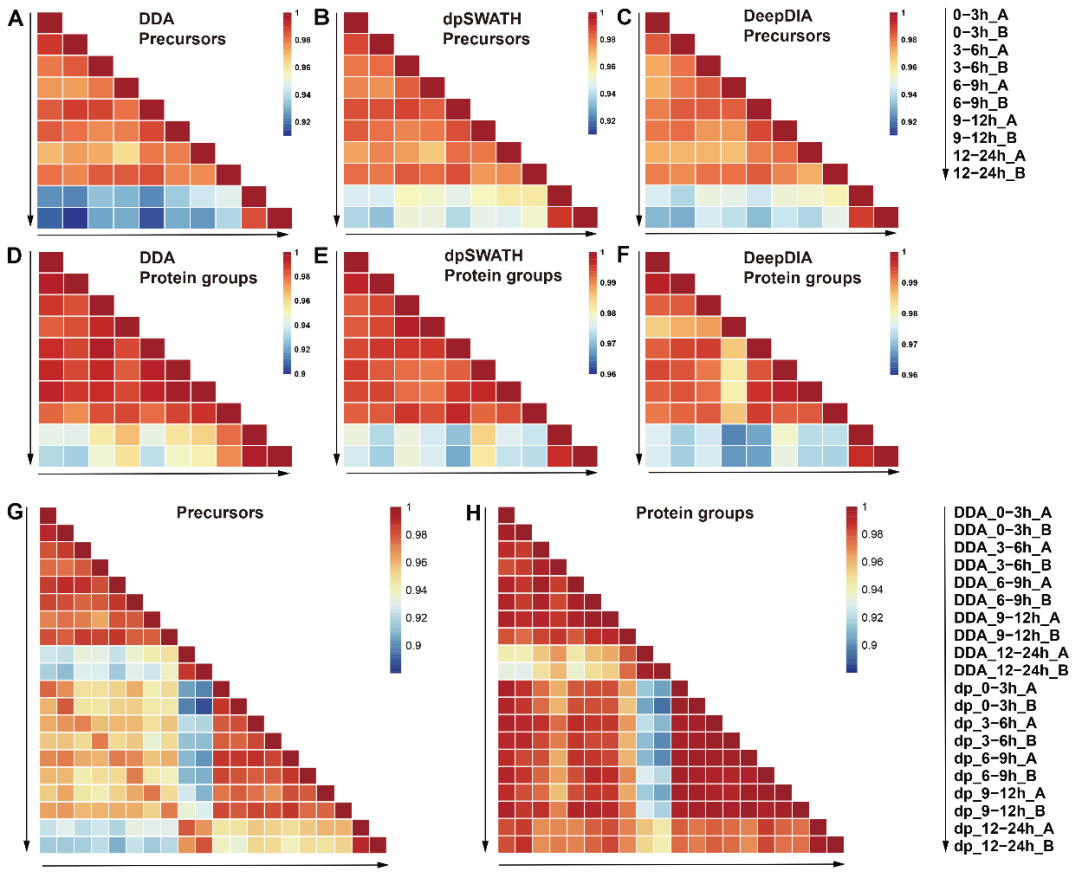


Figure S3. The correlation of quantifications among technical replicates for 5 stages. A, quantatification of precursors correlation heatmap based on 72 fractionated DDA library. B, quantatification of precursors correlation heatmap based on 72 fractionated dpSWATH library. C, quantatification of precursors correlation heatmap based on 72 fractionated DeepDIA library. D, quantatification of protein groups heatmap based on 72 fractionated DDA library. E, quantatification of protein groups heatmap based on 72 fractionated dpSWATH library. F, quantatification of protein groups heatmap based on 72 fractionated DeepDIA library. G, quantatification of precursors correlation heatmap based on among 72 fractionated DDA and dpSWATH library. H, quantatification of protein groups correlation heatmap based on among 72 fractionated DDA and dpSWATH library.

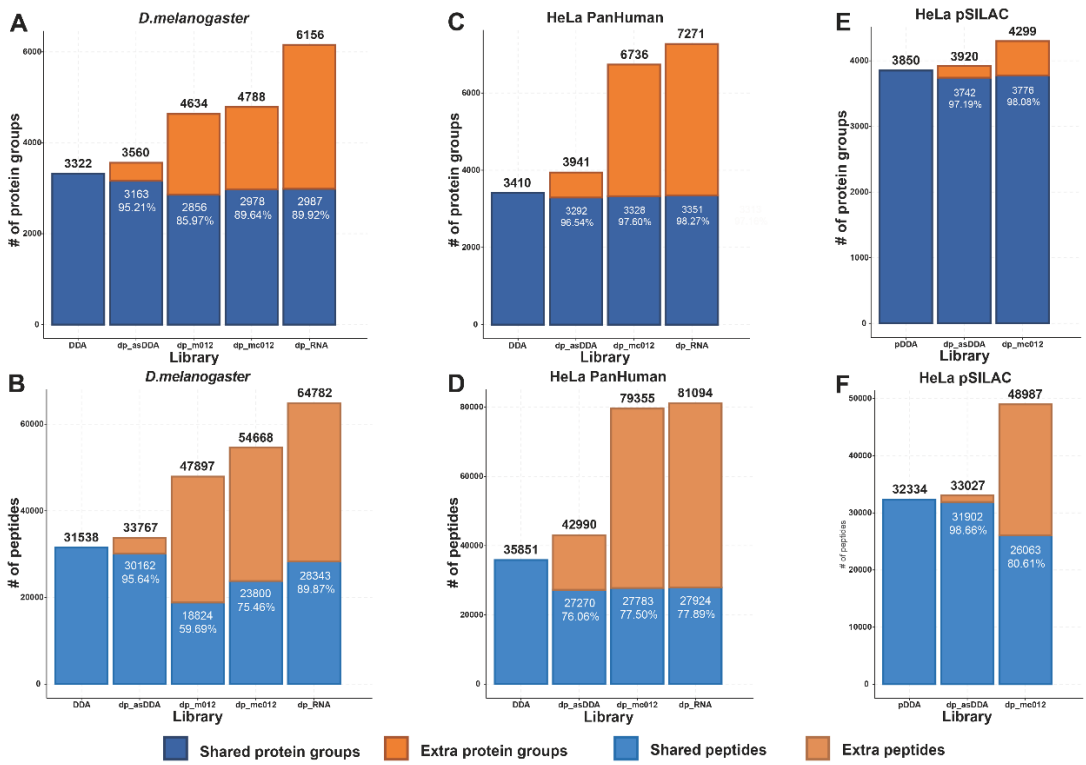


Figure S4. The identifications of protein groups and peptides from different libraries. A, the number of identified protein groups on *D.melanogaster*. "DDA" indicates results from 72 fractionated DDA library; "dp_asDDA" indicates results from the *in-silico* library on the same entries as DDA library; "dp_m012 " indicates results from the *in-silico* library on the digested fasta sequence of the same proteins as DDA library by Keil rules with up to 2 missed cleavages in Spectronaut; "dp_mc012" indicates results from the *in-silico* library on the digested fasta sequence of the same proteins as DDA library by dpMC with up to 2 missed cleavages combined with no missed cleavages; "dp_RNA" indicates results from the transcriptome based *in-silico* library. B, the number of identified peptides on *D.melanogaster*. C, the number of identified protein groups on HeLa datasets from TripleTOF 5600 compared to the PanHuman library. "DDA" indicates resutls from experimental Pan-Human library. D, the number of identified peptides on HeLa datasets from TripleTOF 5600 compared

to the PanHuman library. E, the number of identified protein groups on HeLa datasets from TripleTOF 5600 compared to the pSILAC library. "DDA" indicates resutls from experimental pSILAC DDA library. F, the number of identified peptides on HeLa datasets from TripleTOF 5600 compared to the pSILAC library. Identifications overlapped with the DDA-based libraries are denoted as "shared". Novel identifications by in-silico libraries are denoted as "extra". The numbers and sensitivities of protein groups or peptides are shown.

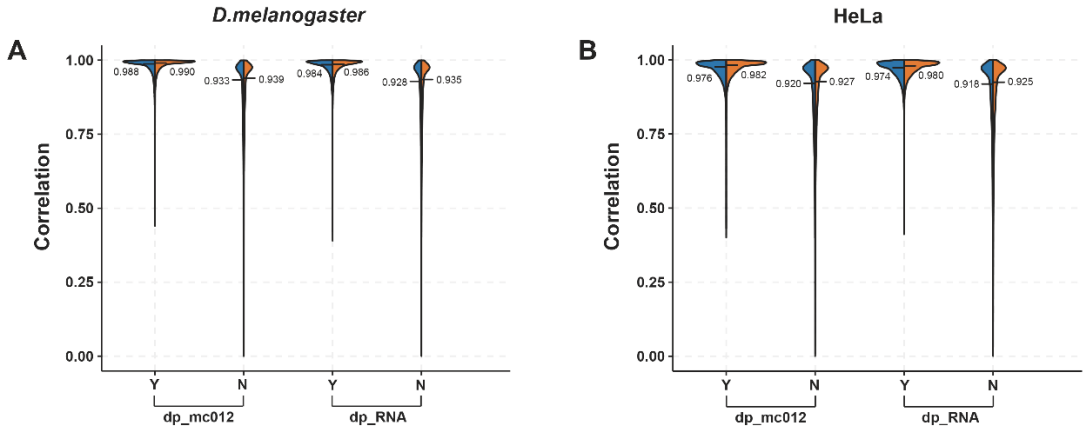


Figure S5. The impact of dpMScore on the performance of dpMS. A. "dp_mc012" indicates results from the *in-silico* library of *D.melanogaster* on the digested fasta sequence of the same proteins as DDA library by dpMC with up to 2 missed cleavages combined with no missed cleavages; "dp_RNA" indicates results from the transcriptome based *in-silico* library. "Y" denotes the results were processed with

dpMScore, "N" denotes the results were processed without dpMScore. The correlations were calculated between the predicted intensities and the final SWATH-MS intensities. B. The correlation between the predicted intensities and the final SWATH-MS intensities for HeLa datasets.

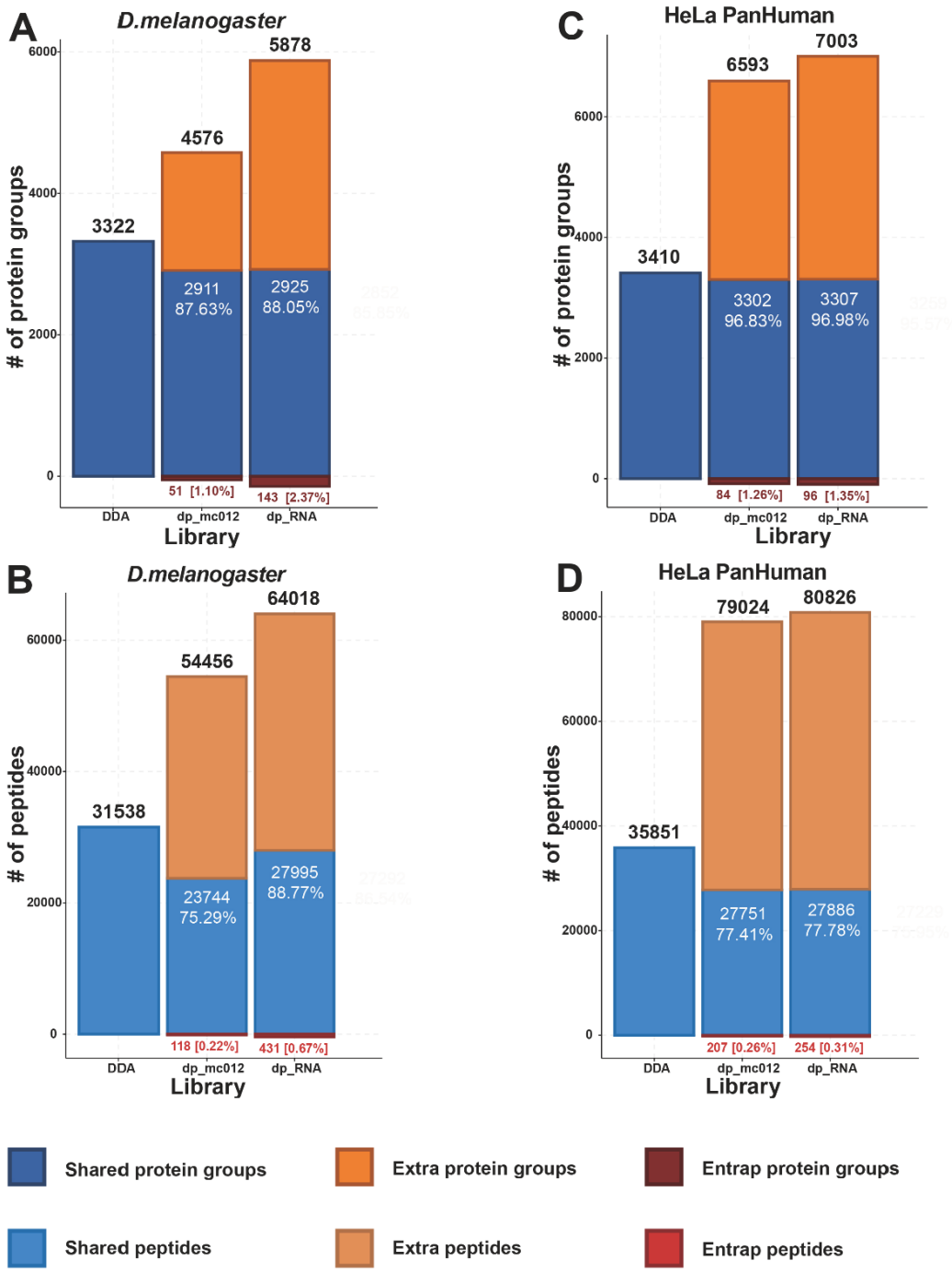


Figure S6. Estimation of FDR on the three datasets used in this study. A, the number of protein groups identified by entrapment library on *D.melanogaster*. "DDA" indicates the 72 fractionated DDA library. "dp_mc012" indicates results from the *in-silico* library on the digested fasta sequence of the same proteins as DDA library by dpMC with up to 2 missed cleavages combined with no missed cleavages but pooled

with entrapment library built with *S.cerevisiae*; "dp_RNA" indicates results from the *in-silico* transcriptome based library pooled with entrapment library built with *C.elegans and D.discoideum*; B, the number of peptides identified by entrapment library. C, the number of protein groups identified by entrapment library on HeLa compared to Pan-Human library. D, the number of peptides identified by entrapment library on HeLa compared to Pan-Human library. Identifications overlapped with the DDA-based libraries are denoted as "shared". Novel identifications by *in-silico* libraries are denoted as "extra". The numbers, sensitivities and entrapment percentages of protein groups or peptides are shown.

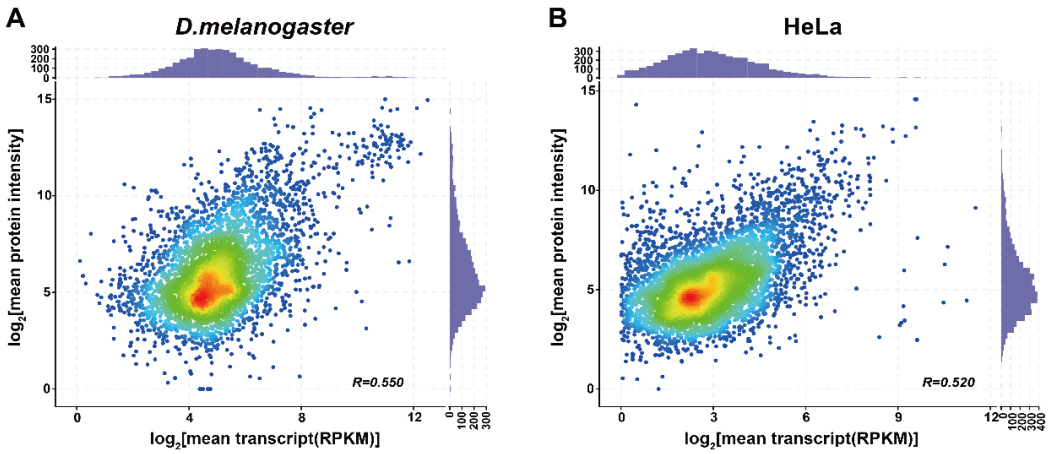


Figure S7. The correlation between gene expression data and protein intensities. A. the log-log scatterplot for *D.melanogaster*. B. the log-log scatterplot for HeLa. Both gene expression data and protein intensities are logarithm transformed. The counts of data points for X-axis and Y-axis are shown on the marginal sides of both scatter plots.

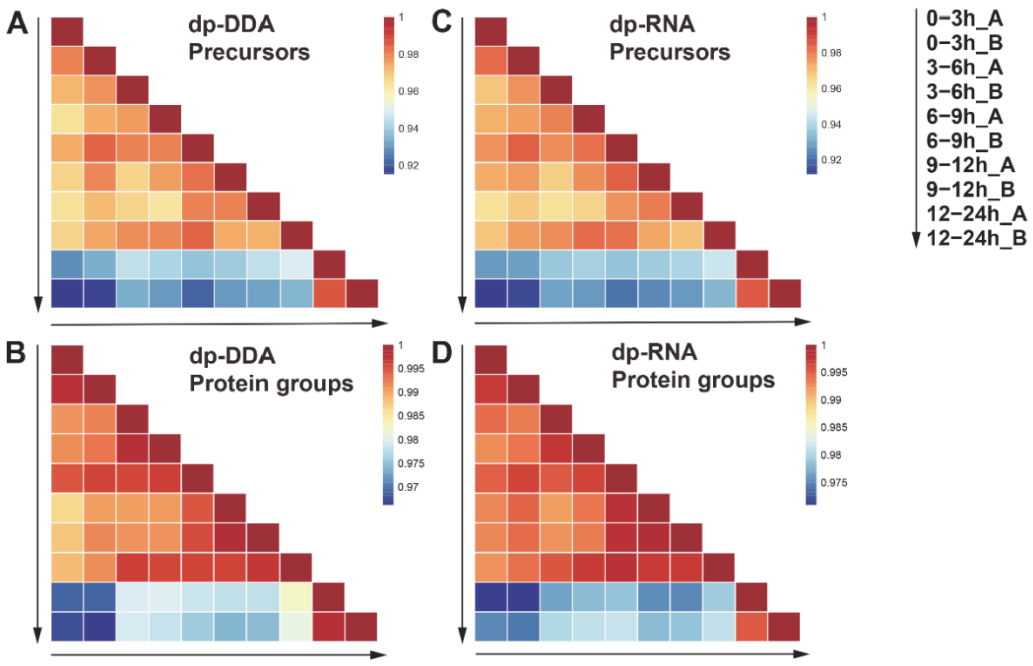


Figure S8. The correlation of quantifications among technical replicates for 5 stages of different scale of in-silico libraries. A, quantatification of precursors correlation heatmap of *in-silico* library generated based on the protein entries from 72 fractionated DDA library. B, quantatification of protein groups correlation heatmap of in-silico library generated based on the protein entries from 72 fractionated DDA library. C, quantatification of precursors correlation heatmap of HeLa transcriptome based dpSWATH library. D, quantatification of protein groups correlation heatmap of HeLa transcriptome based dpSWATH library.

Supplementary Note 1. The information of the generated library by dpSWATH for Spectronaut

The generated library by dpSWATH for Sepctronaut (15.2.210819, Biognosys AG, Schlieren, Switzerland) contains 10 necessary features in .txt format:

1. **StrippedSequence**. The stripped amino acid sequence of the peptide excluding any modifications.
2. **ModifiedSequence**. To specify the amino acid sequence including modifications in case that the peptide is modified. For now, dpSWATH only provides the fixed modificaiton of carbamidomethyl (C), and no variable modifications were specified.
3. **PrecursorCharge.** The peptide precursor ion charge.
4. **iRT.** The peptide retention time in the reverse phase chromatography converted into iRT space.
5. **PrecursorMz.** The *in silico* calculated m/z of the peptide precursor ion.
6. **FragmentMz.** The in silico calculated m/z of the peptide fragment ion.
7. **FragmentType.** The peptide fragment ion type. "y" ions and "b" ions are used in this work.
8. **FragmentNumber.** The peptide frament ion number.
9. **FragmentCharge.** The peptide fragment ion charge formatted as a number.
10. **RelativeFragmentIntensity.** The relative peptide fragment ion intensity expressed as a percentage of the most intense fragment ion.

Note: ***More detailed information about the above features used in the library built by dpSWATH can be refered in the User Manual of Spectronaut.***

Supplementary Note 2. The information of the generated library by dpSWATH for Skyline

The generated library by dpSWATH for Skyline[1, 2] contains 10 necessary features in .xml format:

For the tag of <**SPECTRUM**>, 11 features are included:

1. **charge.** The peptide precursor ion charge.
2. **elution.** The predicted retention time by dpRT.
3. **elutionpeakwidthfwhm.** As all the peaks were *in silico* generated, so we assign 0 for this element.
4. **msid.** We put 1 for this element.
5. **precursorelution.** The predicted retention time by dpRT.
6. **precursormass.** The *in silico* calculated m/z of the peptide precursor ion.
7. **precursorsignal.** We put 1 for this element.
8. **precursorsignalacquisition.** We put 1 for this element.
9. **sumofms2counts.** We leave it as blank for this element.
10. **xml:id.** The index for each peptide.
11. **yscale.** We put 1 for this element.

For the tag of <**MATCH**>, 13 elements are included:

1. **charge.** The peptide precursor ion charge.
2. **confidence.** The evalution of quality for each spectrum by Skyline. In this work, we put high confidence for each spectrum as 0.9999.
3. **confidence_prior.** We put -1 for this element.
4. **da_delta.** We put 0.005 for this element.
5. **eval.** We put very small value for this element.
6. **mod_prob.** We put 0 for this element.
7. **mz.** The *in silico* calculated m/z of the peptide precursor ion.

8. **pid.** We leave it as blank for this element.
9. **pm.** The indexes of the peaks recorded for given spectrum.
10. **score.** We put 10 for this element.
11. **searches.** The index for each peptide.
11. **seq.** The amino acid sequence of the peptide.
12. **type.** We put 0 for this element.
13. **xml:id.** The index for each peptide.

For the tag of < **MSMSPEAKS**>, 3 elements are included:

1. **attributes.** This is a fixed item for this element, which is " MOZ TO CHARGE1,CHARGE STATE,PEAK HEIGHT" for the following information of each peak.
2. **size.** The count of all the peaks.
3. **sp.** We put 0 for this element.

For the records of each peak in the body of <**CDATA**>, 3 features were recorded:

1. **MOZ TO CHARGE1.** The in silico calculated m/z of the peptide fragment ion.
2. **CHARGE STATE.** The peptide fragment ion charge formatted as a number.
3. **PEAK HEIGHT.** The relative peptide fragment ion intensity expressed as a percentage of the most intense fragment ion.

Note: ***More detailed information about the above features used in the library built by dpSWATH can be refered in the Tutorial of Skyline.***

1. MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010,** *26* (7), 966-8.
2. Egertson, J. D.; MacLean, B.; Johnson, R.; Xuan, Y.; MacCoss, M. J., Multiplexed peptide analysis using data-independent acquisition and Skyline. *Nat Protoc* **2015,** *10* (6), 887-903.

Discussion

Genomics, transcriptomics, and proteomics have been experiencing a fast-developing period for the benefit given by the improvement of NGS and MS. However, with the increase of data generated from different NGS and MS platforms, the complexity of such data is also posing challenges for analysis. ChIP-seq has become an indispensable technique for the analysis of the interactions between DNA and proteins. In the study of the mechanisms of HMR in *D.melanogaster*, the localization of HMR to genomic insulator sites was elucidated by ChIP-seq, which revealed the relationship between HMR and insulator proteins. Meanwhile, RNA-seq has been a ubiquitous method to analyze gene expression on the transcriptomic level. Through the analysis of the expression of HMR-bound genes by RNA-seq, we find the associated genomic insulator sites can be divided into two clusters, in which one set is bordered by HP1a-bound areas of active genes, whereas the other is composed of gypsy insulators.

The complex fragment ion spectra generated during a data-independent acquisition (DIA) method of proteomics experiments result in much larger and richer information to be analyzed compared to the more classical data-dependent acquisition methods (DDA). One of the challenges is posed by the construction of high-quality fragment libraries, which is crucial for both the identification and quantification of proteins. The spectral library is one important part of the analysis of spectra generated by DIA. Traditionally, the spectral library is prepared by DDA experiments. However, the coverage and recovery rate of such an experimental library cannot fully meet the requirements of DIA analysis for its limited detected precursors. In recent years, some algorithms have been developed for this purpose, however, the training of such models is based on Orbitrap mass spectrometers, which have more consistent and higher-quality mass spectra compared to the Q-TOF machine. On the other hand, the Q-TOF platform can provide people with a high-speed acquisition of tandem mass spectra. In this work, we developed a framework by deep learning, dpSWATH, to extend the search space for high-quality SWATH-MS analysis on the Q-TOF platform. The mass spectra generated from the Q-TOF platform have lower quality and reproducibility, which needs to filter noises and prepare more consistent mass spectra for training models on such a platform. For this reason, we designed an algorithm, dpMScore, to perform the clustering of mass spectra from a Q-TOF instrument.

The two major components of *in silico* library are the predicted intensities and retention time of given precursors, for which we developed dpMS and dpRT. We then prepared the

theoretically digested peptide candidates for the theoretical library for SWATH-MS analysis. Compared to the experimental library, much more peptide candidates for given protein groups can be integrated into the theoretical library for the Q-TOF platform. All the predictions by dpSWATH can lead to more identifications and accurate quantifications of protein groups and peptides compared to the experimental library. However, with the increase of precursors in the theoretical library, the false positives also increase. So the control of FDR for the precursor candidates for the theoretical library is crucial for the building of high-quality searching space for SWATH-MS analysis. To decrease the FDR issue for a such big searching library, we prepared the theoretical digested peptides by dpMC, which predicts the missed tryptic cleavages. By the application of dpMC, we can not only decrease the size of the theoretical library but also prepare the theoretical library with more detectable precursor candidates. Based on the above strategies, we construct a high-quality search space for SWATH-MS analysis.

Moreover, the *in silico* library was also built based on the transcriptome data, which achieved more protein groups with acceptable FDR compared to the experimental library. However, the whole proteome-wide analysis still needs to be explored. The FDR increased dramatically when the number of uncertain peptide entries in the search library. So more accurate predictions of detectable peptide candidates need to be performed for the building of a theoretical library. Furthermore, we will continue to fuel the model with more high-quality Q-TOF mass spectra data from more experiments or by improving the algorithm of filtering noises, to improve the whole performance of dpSWATH model. Thus, we believe the development of dpSWATH can improve SWATH-MS analysis on a deeper scale with sample-specific detected candidates on proteome-wide analysis.

Recently, some other advanced SWATH-MS methods have been developed, such as the Scanning SWATH (Messner CB et al., 2021), and Zeno SWATH MS (Wang Z et al., 2022). By application of such techniques, more proteins are identified with less volume of samples at ultra-fast speed compared to traditional DIA methods. However, the bioinformatic methods for such analysis still need to be developed and optimized, either with traditional statistical or machine learning approaches, for example, to build project-specific in silico libraries or optimized library-free approach for ultra-fast SWATH-MS analysis.

References

Lander ES, Linton LM, Birren B, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. https://doi.org/10.1038/35057062

Venter JC, Adams MD, Myers EW, et al (2001) The Sequence of the Human Genome. Science 291:1304–1351. https://doi.org/10.1126/science.1058040

Petersen B-S, Fredrich B, Hoeppner MP, et al (2017) Opportunities and challenges of whole-genome and -exome sequencing. Bmc Genet 18:14. https://doi.org/10.1186/s12863-017-0479-5

Lu Y-F, Goldstein DB, Angrist M, Cavalleri G (2014) Personalized Medicine and Human Genetic Diversity. Csh Perspect Med 4:a008581. https://doi.org/10.1101/cshperspect.a008581

Baker M (2011) The next step for the synthetic genome. Nature 473:403–408. https://doi.org/10.1038/473403a

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science 270:467–470. https://doi.org/10.1126/science.270.5235.467

Cheung VG, Morley M, Aguilar F, et al (1999) Making and reading microarrays. Nat Genet 21:15–19. https://doi.org/10.1038/4439

Yates JR, Ruse CI, Nakorchevsky A (2009) Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. Biomed Eng 11:49–79. https://doi.org/10.1146/annurev-bioeng-061008-124934

Chen C, Hou J, Tanner JJ, Cheng J (2020) Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. Int J Mol Sci 21:2873. https://doi.org/10.3390/ijms21082873

Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422:198–207. https://doi.org/10.1038/nature01511

Yarmush ML, Jayaraman A (2002) ADVANCES IN PROTEOMIC TECHNOLOGIES. Annu Rev Biomed Eng 4:349–373. https://doi.org/10.1146/annurev.bioeng.4.020702.153443

Aslam B, Basit M, Nisar MA, et al (2017) Proteomics: Technologies and Their Applications. J Chromatogr Sci 55:182–196. https://doi.org/10.1093/chromsci/bmw167

Keshishian H, Addona T, Burgess M, et al (2007) Quantitative, Multiplexed Assays for Low Abundance Proteins in Plasma by Targeted Mass Spectrometry and Stable Isotope Dilution*. Mol Cell Proteomics 6:2212–2229. https://doi.org/10.1074/mcp.m700354-mcp200

Mallick P, Kuster B (2010) Proteomics: a pragmatic perspective. Nat Biotechnol 28:695–709. https://doi.org/10.1038/nbt.1658

Choi M, Carver J, Chiva C, et al (2020) MassIVE.quant: a community resource of quantitative mass spectrometry–based proteomics datasets. Nat Methods 17:981–984. https://doi.org/10.1038/s41592-020-0955-0

Liebal UW, Phan ANT, Sudhakar M, et al (2020) Machine Learning Applications for Mass Spectrometry-Based Metabolomics. Metabolites 10:243. https://doi.org/10.3390/metabo10060243

Zhao Y-Y, Lin R-C (2014) UPLC–MSE application in disease biomarker discovery: The discoveries in proteomics to metabolomics. Chem-biol Interact 215:7–16. https://doi.org/10.1016/j.cbi.2014.02.014

Nicholson JK, Connelly J, Lindon JC, Holmes E (2002) Metabonomics: a platform for studying drug toxicity and gene function. Nat Rev Drug Discov 1:153–161. https://doi.org/10.1038/nrd728

Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. Genome Biol 18:83. https://doi.org/10.1186/s13059-017-1215-1

McGuire AL, Gabriel S, Tishkoff SA, et al (2020) The road ahead in genetics and genomics. Nat Rev Genet 21:581–596. https://doi.org/10.1038/s41576-020-0272-6

Cristoni S, Bernardi LR (2014) Bioinformatics in mass spectrometry data analysis for proteomics studies. Expert Rev Proteomic 1:469–483. https://doi.org/10.1586/14789450.1.4.469

Kumar C, Mann M (2009) Bioinformatics analysis of mass spectrometry-based proteomics data sets. Febs Lett 583:1703–1712. https://doi.org/10.1016/j.febslet.2009.03.035

Alharbi WS, Rashid M (2022) A review of deep learning applications in human genomics using next-generation sequencing data. Hum Genomics 16:26. https://doi.org/10.1186/s40246-022-00396-x

Smith AM, Walsh JR, Long J, et al (2020) Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. Bmc Bioinformatics 21:119. https://doi.org/10.1186/s12859-020-3427-8

Xu LL, Young A, Zhou A, Röst HL (2020) Machine Learning in Mass Spectrometric Analysis of DIA Data. Proteomics 20:1900352. https://doi.org/10.1002/pmic.201900352

Wang S, Li W, Hu L, et al (2020) NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. Nucleic Acids Res gkaa498-. https://doi.org/10.1093/nar/gkaa498

Raja K, Patrick M, Gao Y, et al (2017) A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries. Int J Genomics 2017:1–10. https://doi.org/10.1155/2017/6213474

Horgan RP, Kenny LC (2011) 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. Obstetrician Gynaecol 13:189–195. https://doi.org/10.1576/toag.13.3.189.27672

Vailati-Riboni M, Osorio JS, Trevisi E, et al (2017) Supplemental Smartamine M in higher-energy diets during the prepartal period improves hepatic biomarkers of health and oxidative status in Holstein cows. J Anim Sci Biotechno 8:17. https://doi.org/10.1186/s40104-017-0147-7

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63. https://doi.org/10.1038/nrg2484

Milward EA, Shahandeh A, Heidari M, et al (2016) Encyclopedia of Cell Biology. Horiz Integration 160–165. https://doi.org/10.1016/b978-0-12-394447-4.40029-5

Lowe R, Shirley N, Bleackley M, et al (2017) Transcriptomics technologies. Plos Comput Biol 13:e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Supplitt S, Karpinski P, Sasiadek M, Laczmanska I (2021) Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. Int J Mol Sci 22:1422. https://doi.org/10.3390/ijms22031422

Kukurba KR, Montgomery SB (2015) RNA Sequencing and Analysis. Cold Spring Harb Protoc 2015:pdb.top084970. https://doi.org/10.1101/pdb.top084970

Geraci F, Saha I, Bianchini M (2020) Editorial: RNA-Seq Analysis: Methods, Applications and Challenges. Frontiers Genetics 11:220. https://doi.org/10.3389/fgene.2020.00220

Anderson L, Seilhamer J (1997) A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18:533–537. https://doi.org/10.1002/elps.1150180333

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between Protein and mRNA Abundance in Yeast. Mol Cell Biol 19:1720–1730. https://doi.org/10.1128/mcb.19.3.1720

Hieter P, Boguski M (1997) Functional Genomics: It's All How You Read It. Science 278:601–602. https://doi.org/10.1126/science.278.5338.601

Przybyla L, Gilbert LA (2022) A new era in functional genomics screens. Nat Rev Genet 23:89–103. https://doi.org/10.1038/s41576-021-00409-w

Collins FS, Morgan M, Patrinos A (2003) The Human Genome Project: Lessons from Large-Scale Biology. Science 300:286–290. https://doi.org/10.1126/science.1084564

Diniz WJS, Canduri F (2017) REVIEW-ARTICLE Bioinformatics: an overview and its applications. Genet Mol Res 16:. https://doi.org/10.4238/gmr16019645

Nelson DL, Cox MM (2005). Principles of Biochemistry (4th ed.). New York: W. H. Freeman ISBN 0-7167-4339-6.

Chandrasekhar K, Dileep A, Lebonah DE, Kumari JP (2014) A Short Review on Proteomics and its Applications. Int Lett Nat Sci 17:77–84. https://doi.org/10.18052/www.scipress.com/ilns.17.77

Agarwal PK (2006) Enzymes: An integrated view of structure, dynamics and function. Microb Cell Fact 5:2. https://doi.org/10.1186/1475-2859-5-2

LaBaer J (2002) Genomics, proteomics, and the new paradigm in biomedical research. Genet Med 4:2S-9S. https://doi.org/10.1097/00125817-200211001-00002

Dupree EJ, Jayathirtha M, Yorkey H, et al (2020) A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. Proteomes 8:14. https://doi.org/10.3390/proteomes8030014

O'Farrell P (1975) High resolution two-dimensional electrophoresis of proteins. J Biol Chem 250:4007–4021. https://doi.org/10.1016/s0021-9258(19)41496-8

Klose J (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. Humangenetik 26:231–243. https://doi.org/10.1007/bf00281458

Kar UK, Simonian M, Whitelegge JP (2017) Integral membrane proteins: bottom-up, top-down and structural proteomics. Expert Rev Proteomic 14:715–723. https://doi.org/10.1080/14789450.2017.1359545

Gillet LC, Leitner A, Aebersold R (2015) Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. Annu Rev Anal Chem 9:1–24. https://doi.org/10.1146/annurev-anchem-071015-041535

Toby TK, Fornelli L, Kelleher NL (2016) Progress in Top-Down Proteomics and the Analysis of Proteoforms. Annu Rev Anal Chem 9:499–519. https://doi.org/10.1146/annurev-anchem-071015-041550

Chen CC, Greene PG, Crick A (1998) Does entrepreneurial self-efficacy distinguish entrepreneurs from managers? J Bus Venturing 13:295–316. https://doi.org/10.1016/s0883-9026(97)00029-3

Donnelly DP, Rawlins CM, DeHart CJ, et al (2019) Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. Nat Methods 16:587–594. https://doi.org/10.1038/s41592-019-0457-0

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc National Acad Sci 74:5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Grada A, Weinbrecht K (2013) Next-Generation Sequencing: Methodology and Application. J Invest Dermatol 133:1–4. https://doi.org/10.1038/jid.2013.248

Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genet 11:476–486. https://doi.org/10.1038/nrg2795

Farnham PJ (2009) Insights from genomic profiling of transcription factors. Nat Rev Genet 10:605–616. https://doi.org/10.1038/nrg2636

Park PJ (2009) ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680. https://doi.org/10.1038/nrg2641

Solomon MJ, Larsen PL, Varshavsky A (1988) Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. Cell 53:937–947. https://doi.org/10.1016/s0092-8674(88)90469-2

Nakato R, Sakata T (2021) Methods for ChIP-seq analysis: A practical workflow and advanced applications. Methods 187:44–53. https://doi.org/10.1016/j.ymeth.2020.03.005

Cloonan N, Forrest ARR, Kolle G, et al (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5:613–619. https://doi.org/10.1038/nmeth.1223

Han X, Aslanian A, Yates JR (2008) Mass spectrometry for proteomics. Curr Opin Chem Biol 12:483–490. https://doi.org/10.1016/j.cbpa.2008.07.024

Graves PR, Haystead TAJ (2002) Molecular Biologist's Guide to Proteomics. Microbiol Mol Biol R 66:39–63. https://doi.org/10.1128/mmbr.66.1.39-63.2002

Hu A, Noble WS, Wolf-Yadlin A (2016) Technical advances in proteomics: new developments in data-independent acquisition. F1000research 5:F1000 Faculty Rev-419. https://doi.org/10.12688/f1000research.7042.1

Liao P, Allison J (1995) Dissecting matrix-assisted laser desorption/ionization mass spectra. J Mass Spectrom 30:763–766. https://doi.org/10.1002/jms.1190300517

Shen Z, Thomas JJ, Averbuj C, et al (2001) Porous Silicon as a Versatile Platform for Laser Desorption/Ionization Mass Spectrometry. Anal Chem 73:612–619. https://doi.org/10.1021/ac000746f

Laiko VV, Baldwin MA, Burlingame AL (2000) Atmospheric Pressure Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. Anal Chem 72:652–657. https://doi.org/10.1021/ac990998k

Griffin PR, Coffman JA, Hood LE, Yates JR (1991) Structural analysis of proteins by capillary HPLC electrospray tandem mass spectrometry. Int J Mass Spectrom 111:131–149. https://doi.org/10.1016/0168-1176(91)85052-n

Emmett MR, Caprioli RM (1994) Micro-electrospray mass spectrometry: Ultra-high-sensitivity analysis of peptides and proteins. J Am Soc Mass Spectr 5:605–613. https://doi.org/10.1016/1044-0305(94)85001-1

Hager JW, Blanc JCYL (2003) High-performance liquid chromatography–tandem mass spectrometry with a new quadrupole/linear ion trap instrument. J Chromatogr A 1020:3–9. https://doi.org/10.1016/s0021-9673(03)00426-6

Hu Q, Noll RJ, Li H, et al (2005) The Orbitrap: a new mass spectrometer. J Mass Spectrom 40:430–443. https://doi.org/10.1002/jms.856

Makarov A, Denisov E, Lange O, Horning S (2006) Dynamic range of mass accuracy in LTQ orbitrap hybrid mass spectrometer. J Am Soc Mass Spectr 17:977–982. https://doi.org/10.1016/j.jasms.2006.03.006

Syka JEP, Marto JA, Bai DL, et al (2004) Novel Linear Quadrupole Ion Trap/FT Mass Spectrometer: Performance Characterization and Use in the Comparative Analysis of Histone H3 Post-translational Modifications. J Proteome Res 3:621–626. https://doi.org/10.1021/pr0499794

Breuker K, Jin M, Han X, et al (2008) Top-down identification and characterization of biomolecules by mass spectrometry. J Am Soc Mass Spectr 19:1045–1053. https://doi.org/10.1016/j.jasms.2008.05.013

Morris HR, Paxton T, Dell A, et al (1996) High Sensitivity Collisionally-activated Decomposition Tandem Mass Spectrometry on a Novel Quadrupole/Orthogonal-acceleration Time-of-flight Mass Spectrometer. Rapid Commun Mass Sp 10:889–896. https://doi.org/10.1002/1097-0231/rcm615

Shevchenko A, Chernushevich I, Ens W, et al (1997) Rapid ‘de novo’ peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. Rapid Commun Mass Sp 11:1015–1024. https://doi.org/10.1002/1097-0231/rcm958

Collings BA, Campbell JM, Mao D, Douglas DJ (2001) A combined linear ion trap time-of-flight system with improved performance and MSn capabilities. Rapid Commun Mass Sp 15:1777–1795. https://doi.org/10.1002/rcm.440

Campbell JM, Collings BA, Douglas DJ (1998) A new linear ion trap time-of-flight system with tandem mass spectrometry capabilities. Rapid Commun Mass Sp 12:1463–1474. https://doi.org/10.1002/(sici)1097-0231(19981030)12:20<1463::aid-rcm357>3.0.co;2-h

Senko MW, Canterbury JD, Guan S, Marshall AG (1996) A High-performance Modular Data System for Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. Rapid Commun Mass Sp 10:1839–1844. https://doi.org/10.1002/(sici)1097-0231(199611)10:14<1839::aid-rcm718>3.0.co;2-v

Andrews GL, Simons BL, Young JB, et al (2011) Performance Characteristics of a New Hybrid Quadrupole Time-of-Flight Tandem Mass Spectrometer (TripleTOF 5600). Anal Chem 83:5442–5446. https://doi.org/10.1021/ac200812d

Gillet LC, Navarro P, Tate S, et al (2012) Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis*. Mol Cell Proteomics 11:O111.016717. https://doi.org/10.1074/mcp.o111.016717

Bateman NW, Goulding SP, Shulman NJ, et al (2014) Maximizing Peptide Identification Events in Proteomic Workflows Using Data-Dependent Acquisition (DDA)*. Mol Cell Proteomics 13:329–338. https://doi.org/10.1074/mcp.m112.026500

Mann M, Hendrickson RC, Pandey A (2001) ANALYSIS OF PROTEINS AND PROTEOMES BY MASS SPECTROMETRY. Annu Rev Biochem 70:437–473. https://doi.org/10.1146/annurev.biochem.70.1.437

Venable JD, Dong M-Q, Wohlschlegel J, et al (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods 1:39–45. https://doi.org/10.1038/nmeth705

Egertson JD, MacLean B, Johnson R, et al (2015) Multiplexed peptide analysis using data-independent acquisition and Skyline. Nature protocols 10:887–903. https://doi.org/10.1038/nprot.2015.055

Shi T, Song E, Nie S, et al (2016) Advances in targeted proteomics and applications to biomedical research. Proteomics 16:2160–2182. https://doi.org/10.1002/pmic.201500449

Huang Q, Yang L, Luo J, et al (2015) SWATH enables precise label-free quantification on proteome scale. Proteomics 15:1215–1223. https://doi.org/10.1002/pmic.201400270

Ludwig C, Gillet L, Rosenberger G, et al (2018) Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol Syst Biol 14:e8126. https://doi.org/10.15252/msb.20178126

Magi A, Benelli M, Gozzini A, et al (2010) Bioinformatics for Next Generation Sequencing Data. Genes-basel 1:294–307. https://doi.org/10.3390/genes1020294

Cokus SJ, Feng S, Zhang X, et al (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452:215–219. https://doi.org/10.1038/nature06745

Valouev A, Johnson DS, Sundquist A, et al (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5:829–834. https://doi.org/10.1038/nmeth.1246

Fejes AP, Robertson G, Bilenky M, et al (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24:1729–1730. https://doi.org/10.1093/bioinformatics/btn305

Heinz S, Benner C, Spann N, et al (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 38:576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Chen T-W, Li H-P, Lee C-C, et al (2014) ChIPseek, a web-based analysis tool for ChIP data. Bmc Genomics 15:539. https://doi.org/10.1186/1471-2164-15-539

Horner DS, Pavesi G, Castrignanò T, et al (2010) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief Bioinform 11:181–197. https://doi.org/10.1093/bib/bbp046

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Bona FD, Ossowski S, Schneeberger K, Rätsch G (2008) Optimal spliced alignments of short sequence reads. Bioinformatics 24:i174–i180. https://doi.org/10.1093/bioinformatics/btn300

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65. https://doi.org/10.1093/nar/gkl842

Castrignanò T, D'Antonio M, Anselmo A, et al (2008) ASPicDB: A database resource for alternative splicing analysis. Bioinformatics 24:1300–1304. https://doi.org/10.1093/bioinformatics/btn113

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Okonechnikov K, Conesa A, García-Alcalde F (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 32:292–294. https://doi.org/10.1093/bioinformatics/btv566

DeLuca DS, Levin JZ, Sivachenko A, et al (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics 28:1530–1532. https://doi.org/10.1093/bioinformatics/bts196

Mortazavi A, Williams BA, McCue K, et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628. https://doi.org/10.1038/nmeth.1226

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140. https://doi.org/10.1093/bioinformatics/btp616

Trapnell C, Hendrickson DG, Sauvageau M, et al (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31:46–53. https://doi.org/10.1038/nbt.2450

Kuleshov MV, Jones MR, Rouillard AD, et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44:W90–W97. https://doi.org/10.1093/nar/gkw377

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57. https://doi.org/10.1038/nprot.2008.211

Geer LY, Markey SP, Kowalak JA, et al (2004) Open Mass Spectrometry Search Algorithm. J Proteome Res 3:958–964. https://doi.org/10.1021/pr0499491

Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20:1466–1467. https://doi.org/10.1093/bioinformatics/bth092

Frank A, Pevzner P (2005) PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. Analytical Chemistry 77:964–973. https://doi.org/10.1021/ac048788h

Eng JK, Fischer B, Grossmann J, MacCoss MJ (2008) A Fast SEQUEST Cross Correlation Algorithm. J Proteome Res 7:4598–4602. https://doi.org/10.1021/pr800420s

Perkins DN, Pappin DJC, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3567. https://doi.org/10.1002/1522-2683/aid-elps3551

Tyanova S, Temu T, Cox J (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc 11:2301–2319. https://doi.org/10.1038/nprot.2016.136

Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature Methods 4:nmeth1019. https://doi.org/10.1038/nmeth1019

Fischer B, Roth V, Roos F, et al (2005) NovoHMM: A Hidden Markov Model for de Novo Peptide Sequencing. Anal Chem 77:7265–7273. https://doi.org/10.1021/ac0508853

Tanner S, Shu H, Frank A, et al (2005) InsPecT: Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra. Anal Chem 77:4626–4639. https://doi.org/10.1021/ac050102d

Tabb DL, Ma Z-Q, Martin DB, et al (2008) DirecTag: Accurate Sequence Tags from Peptide MS/MS through Statistical Scoring. J Proteome Res 7:3838–3846. https://doi.org/10.1021/pr800154p

Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. Anal Chem 75:4646–4658. https://doi.org/10.1021/ac0341261

Shen C, Wang Z, Shankar G, et al (2008) A hierarchical statistical model to assess the confidence of peptides and proteins inferred from tandem mass spectrometry. Bioinformatics 24:202–208. https://doi.org/10.1093/bioinformatics/btm555

Li YF, Arnold RJ, Li Y, et al (2009) A Bayesian Approach to Protein Inference Problem in Shotgun Proteomics. J Comput Biol 16:1183–1193. https://doi.org/10.1089/cmb.2009.0018

Tyanova S, Albrechtsen R, Kronqvist P, et al (2016) Proteomic maps of breast cancer subtypes. Nat Commun 7:10259. https://doi.org/10.1038/ncomms10259

Khan Z, Bloom JS, Garcia BA, et al (2009) Protein quantification across hundreds of experimental conditions. Proc National Acad Sci 106:15544–15548. https://doi.org/10.1073/pnas.0904100106

Shadforth IP, Dunkley TP, Lilley KS, Bessant C (2005) i-Tracker: For quantitative proteomics using iTRAQ^{TM}. Bmc Genomics 6:145–145. https://doi.org/10.1186/1471-2164-6-145

Arntzen MØ, Koehler CJ, Barsnes H, et al (2011) IsobariQ: Software for Isobaric Quantitative Proteomics using IPTL, iTRAQ, and TMT. J Proteome Res 10:913–920. https://doi.org/10.1021/pr1009977

Välikangas T, Suomi T, Elo LL (2016) A systematic evaluation of normalization methods in quantitative label-free proteomics. Brief Bioinform bbw095. https://doi.org/10.1093/bib/bbw095

Berger JA, Hautaniemi S, Järvinen A-K, et al (2004) Optimized LOWESS normalization parameter selection for DNA microarray data. Bmc Bioinformatics 5:194. https://doi.org/10.1186/1471-2105-5-194

Huber W, Heydebreck A von, Sültmann H, et al (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18:S96–S104. https://doi.org/10.1093/bioinformatics/18.suppl_1.s96

Bern M, Goldberg D (2006) De Novo Analysis of Peptide Tandem Mass Spectra by Spectral Graph Partitioning. J Comput Biol 13:364–378. https://doi.org/10.1089/cmb.2006.13.364

Wei L, Xing P, Shi G, et al (2018) Fast prediction of protein methylation sites using a sequence-based feature selection technique. Ieee Acm Transactions Comput Biology Bioinform 16:1–1. https://doi.org/10.1109/tcbb.2017.2670558

Bergamo GC, Dias CT dos S, Krzanowski WJ (2008) Distribution-free multiple imputation in an interaction matrix through singular value decomposition. Sci Agric 65:422–427. https://doi.org/10.1590/s0103-90162008000400015

Berg P, McConnell EW, Hicks LM, et al (2019) Evaluation of linear models and missing value imputation for the analysis of peptide-centric proteomics. Bmc Bioinformatics 20:102. https://doi.org/10.1186/s12859-019-2619-6

Kammers K, Cole RN, Tiengwe C, Ruczinski I (2015) Detecting significant changes in protein abundance. Eupa Open Proteom 7:11–19. https://doi.org/10.1016/j.euprot.2015.02.002

Iterson M van, Boer JM, Menezes RX (2010) Filtering, FDR and power. Bmc Bioinformatics 11:450–450. https://doi.org/10.1186/1471-2105-11-450

Xie Y, Pan W, Khodursky AB (2005) A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. Bioinformatics 21:4280–4288. https://doi.org/10.1093/bioinformatics/bti685

Dennis G, Sherman BT, Hosack DA, et al (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4:R60. https://doi.org/10.1186/gb-2003-4-9-r60

Szklarczyk D, Morris JH, Cook H, et al (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res 45:D362–D368. https://doi.org/10.1093/nar/gkw937

Côté RG, Jones P, Martens L, et al (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. Bmc Bioinformatics 8:401. https://doi.org/10.1186/1471-2105-8-401

Waegele B, Dunger-Kaltenbach I, Fobo G, et al (2009) CRONOS: the cross-reference navigation server. Bioinformatics 25:141–143. https://doi.org/10.1093/bioinformatics/btn590

Consortium GO (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261. https://doi.org/10.1093/nar/gkh036

Mi H, Thomas P (2009) Protein Networks and Pathway Analysis. Methods Mol Biology 563:123–140. https://doi.org/10.1007/978-1-60761-175-2_7

Kanehisa M, Furumichi M, Tanabe M, et al (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353–D361. https://doi.org/10.1093/nar/gkw1092

Junqueira DM, Braun RL and Verli H (2014). Alinhamentos. In: Bioinformática da biologia à flexibilidade molecular (Verli H, ed.). SBBq, São Paulo, 38-61.

Wolf JBW (2013) Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Mol Ecol Resour 13:559–572. https://doi.org/10.1111/1755-0998.12109

Croft D, O'Kelly G, Wu G, et al (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39:D691–D697. https://doi.org/10.1093/nar/gkq1018

Lavallée-Adam M, Rauniyar N, McClatchy DB, Yates JR (2014) PSEA-Quant: A Protein Set Enrichment Analysis on Label-Free and Label-Based Protein Quantification Data. J Proteome Res 13:5496–5509. https://doi.org/10.1021/pr500473n

Huang Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20:21–28. https://doi.org/10.1093/bioinformatics/btg366

Yan C, Dobbs D, Honavar V (2004) A two-stage classifier for identification of protein–protein interface residues. Bioinformatics 20:i371–i378. https://doi.org/10.1093/bioinformatics/bth920

Meyer JG (2021) Deep learning neural network tools for proteomics. Cell Reports Methods 1:100003. https://doi.org/10.1016/j.crmeth.2021.100003

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536. https://doi.org/10.1038/323533a0

Fukushima K (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 36:193–202. https://doi.org/10.1007/bf00344251

Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Arxiv

Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Comput 9:1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Ma C, Ren Y, Yang J, et al (2018) Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. Analytical chemistry 90:10881–10888. https://doi.org/10.1021/acs.analchem.8b02386

Yang Y, Liu X, Shen C, et al (2020) In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. Nat Commun 11:146. https://doi.org/10.1038/s41467-019-13866-z

Zhou X-X, Zeng W-F, Chi H, et al (2017) pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. Analytical chemistry 89:12690–12697. https://doi.org/10.1021/acs.analchem.7b02566

Gessulat S, Schmidt T, Zolg D, et al (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nature Methods 16:509–518. https://doi.org/10.1038/s41592-019-0426-7

Kantz ED, Tiwari S, Watrous JD, et al (2019) Deep Neural Networks for Classification of LC-MS Spectral Peaks. Anal Chem 91:12407–12413. https://doi.org/10.1021/acs.analchem.9b02983

Serrano G, Guruceaga E, Segura V (2019) DeepMSPeptide: peptide detectability prediction using deep learning. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz708

Tran N, Zhang X, Xin L, et al (2017) De novo peptide sequencing by deep learning. Proceedings of the National Academy of Sciences 114:8247–8252. https://doi.org/10.1073/pnas.1705691114

Siepen JA, Keevil E-J, Knight D, Hubbard SJ (2007) Prediction of Missed Cleavage Sites in Tryptic Peptides Aids Protein Identification in Proteomics. J Proteome Res 6:399–408. https://doi.org/10.1021/pr060507u

Yen C-Y, Russell S, Mendoza AM, et al (2006) Improving Sensitivity in Shotgun Proteomics Using a Peptide-Centric Database with Reduced Complexity: Protease Cleavage and SCX Elution Rules from Data Mining of MS/MS Spectra. Anal Chem 78:1071–1084. https://doi.org/10.1021/ac051127f

Keil B (1992) Specificity of proteolysis. Springer-Verlag Berlin-Heidelberg-NewYork:335

Lawless C, Hubbard SJ (2012) Prediction of missed proteolytic cleavages for the selection of surrogate peptides for quantitative proteomics. OMICS 16 (9):449-456. doi:10.1089/omi.2011.0156

Fannes T, Vandermarliere E, Schietgat L, Degroeve S, Martens L, Ramon J (2013) Predicting tryptic cleavage from proteomics data using decision tree ensembles. J Proteome Res 12 (5):2253-2259. doi:10.1021/pr4001114

Yang J, Gao Z, Ren X, Sheng J, Xu P, Chang C, Fu Y (2021) DeepDigest: Prediction of Protein Proteolytic Digestion with Deep Learning. Anal Chem. doi:10.1021/acs.analchem.0c04704

Meyer JG (2014) In Silico Proteome Cleavage Reveals Iterative Digestion Strategy for High Sequence Coverage. ISRN Comput Biol 2014. doi:10.1155/2014/960902

Zohora FT, Rahman MZ, Tran NH, Xin L, Shan B, Li M (2019) DeepIso: A Deep Learning Model for Peptide Feature Detection from LC-MS map. Sci Rep 9 (1):17168. doi:10.1038/s41598-019-52954-4

Cheng H, Rao B, Liu L, Cui L, Xiao G, Su R, Wei L (2021) PepFormer: End-to-End Transformer-Based Siamese Network to Predict and Enhance Peptide Detectability Based on Sequence Only. Anal Chem. doi:10.1021/acs.analchem.1c00354

Gao Z, Chang C, Yang J, Zhu Y, Fu Y (2019) AP3: An Advanced Proteotypic Peptide Predictor for Targeted Proteomics by Incorporating Peptide Digestibility. Anal Chem 91 (13):8705-8711. doi:10.1021/acs.analchem.9b02520

Cambria TYDHSPE (2018) Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence Magazine 13 (3):55-75. doi:10.1109/MCI.2018.2840738

Carlos Affonso ALD, Fábio Henrique AntunesVieira,André Carlos Ponce de Leon Ferreirade Carvalho, (2017) Deep learning for biological image classification. Expert Systems with Applications 85:114-122. doi:10.1016/j.eswa.2017.05.039

Yan J, Mu L, Wang L, Ranjan R, Zomaya AY (2020) Temporal Convolutional Networks for the Advance Prediction of ENSO. Sci Rep 10 (1):8055. doi:10.1038/s41598-020-65070-5

A. SIB (1967) On the size of the active site in proteases. I. Papain. Biochemical and Biophysical Research Communication (27):157

A. SIB (1968) On the active site of proteases. 3. Mapping the active site of papain; specific peptide inhibitors of papain. Biochemical and Biophysical Research Communication (32):898

Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, Stockinger H (2012) ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40 (Web Server issue):W597-603. doi:10.1093/nar/gks400

Cambria TYDHSPE (2018) Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence Magazine 13 (3):55-75. doi:10.1109/MCI.2018.2840738

Carvalho, (2017) Deep learning for biological image classification. Expert Systems with Applications 85:114-122. doi:10.1016/j.eswa.2017.05.039

Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26 (12):1367-1372. doi:10.1038/nbt.1511

Diederik P. Kingma JB (2014) Adam: A Method for Stochastic Optimization. arXiv

Eyers CE, Lawless C, Wedge DC, Lau KW, Gaskell SJ, Hubbard SJ (2011) CONSeQuence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. Mol Cell Proteomics 10 (11):M110 003384. doi:10.1074/mcp.M110.003384

Gartner SMK, Hundertmark T, Nolte H, Theofel I, Eren-Ghiani Z, Tetzner C, Duchow TB, Rathke C, Kruger M, Renkawitz-Pohl R (2019) Stage-specific testes proteomics of Drosophila melanogaster identifies essential proteins for male fertility. Eur J Cell Biol 98 (2-4):103-115. doi:10.1016/j.ejcb.2019.01.001

Goodfellow IB, Yoshua; Courville, Aaron (2016) Softmax Units for Multinoulli Output Distributions. Deep Learning. MIT Press,

Hammerschmidt P, Ostkotte D, Nolte H, Gerl MJ, Jais A, Brunner HL, Sprenger HG, Awazawa M, Nicholls HT, Turpin-Nolan SM, Langer T, Kruger M, Brugger B, Bruning JC (2019) CerS6-Derived Sphingolipids Interact with Mff and Promote Mitochondrial Fragmentation in Obesity. Cell 177 (6):1536-1552 e1523. doi:10.1016/j.cell.2019.05.008

Hao ZDaY (2020) reportROC: An Easy Way to Report ROC Analysis.

Heissel S, Frederiksen SJ, Bunkenborg J, Hojrup P (2019) Enhanced trypsin on a budget: Stabilization, purification and high-temperature application of inexpensive commercial trypsin for proteomics applications. PLoS One 14 (6):e0218374. doi:10.1371/journal.pone.0218374

Ian Goodfellow YB, and Aaron Courville (2016) Deep Learning. MIT Press,

Just PA, Charawi S, Denis RGP, Savall M, Traore M, Foretz M, Bastu S, Magassa S, Senni N, Sohier P, Wursmer M, Vasseur-Cognet M, Schmitt A, Le Gall M, Leduc M, Guillonneau F, De Bandt JP, Mayeux P, Romagnolo B, Luquet S, Bossard P, Perret C (2020) Lkb1 suppresses amino acid-driven gluconeogenesis in the liver. Nat Commun 11 (1):6127. doi:10.1038/s41467-020-19490-6

Liu Y, Mi Y, Mueller T, Kreibich S, Williams EG, Van Drogen A, Borel C, Frank M, Germain PL, Bludau I, Mehnert M, Seifert M, Emmenlauer M, Sorg I, Bezrukov F, Bena FS, Zhou H, Dehio C, Testa G, Saez-Rodriguez J, Antonarakis SE, Hardt WD, Aebersold R (2019) Multi-omic measurements of heterogeneity in HeLa cells across laboratories. Nat Biotechnol 37 (3):314-322. doi:10.1038/s41587-019-0037-y

Needleman SB (2013) Protein Sequence Determination: A Sourcebook of Methods and Techniques.

O'Shea JP, Chou MF, Quader SA, Ryan JK, Church GM, Schwartz D (2013) pLogo: a probabilistic approach to visualizing sequence motifs. Nat Methods 10 (12):1211-1212. doi:10.1038/nmeth.2646

Prianichnikov N, Koch H, Koch S, Lubeck M, Heilig R, Brehmer S, Fischer R, Cox J (2020) MaxQuant Software for Ion Mobility Enhanced Shotgun Proteomics. Mol Cell Proteomics 19 (6):1058-1069. doi:10.1074/mcp.TIR119.001720

Velazquez D, Albacar M, Zhang C, Calafi C, Lopez-Malo M, Torres-Torronteras J, Marti R, Kovalchuk SI, Pinson B, Jensen ON, Daignan-Fornier B, Casamayor A, Arino J (2020) Yeast Ppz1 protein phosphatase toxicity involves the alteration of multiple cellular targets. Sci Rep 10 (1):15613. doi:10.1038/s41598-020-72391-y

Wu X, Bartel DP (2017) kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. Nucleic Acids Res 45 (W1):W534-W538. doi:10.1093/nar/gkx323

Yan J, Mu L, Wang L, Ranjan R, Zomaya AY (2020) Temporal Convolutional Networks for the Advance Prediction of ENSO. Sci Rep 10 (1):8055. doi:10.1038/s41598-020-65070-5

Yen CY, Russell S, Mendoza AM, Meyer-Arendt K, Sun S, Cios KJ, Ahn NG, Resing KA (2006) Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and SCX elution rules from data mining of MS/MS spectra. Anal Chem 78 (4):1071-1084. doi:10.1021/ac051127f

Ammar C, Berchtold E, Csaba G, Schmidt A, Imhof A, Zimmer R (2019) Multi-Reference Spectral Library Yields Almost Complete Coverage of Heterogeneous LC-MS/MS Data Sets. J Proteome Res 18 (4):1553-1566. doi:10.1021/acs.jproteome.8b00819

Collins BC, Gillet LC, Rosenberger G, Rost HL, Vichalkovski A, Gstaiger M, Aebersold R (2013) Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system. Nat Methods 10 (12):1246-1253. doi:10.1038/nmeth.2703

Collins BC, Hunter CL, Liu Y, Schilling B, Rosenberger G, Bader SL, Chan DW, Gibson BW, Gingras AC, Held JM, Hirayama-Kurogi M, Hou G, Krisp C, Larsen B, Lin L, Liu S, Molloy MP, Moritz RL, Ohtsuki S, Schlapbach R, Selevsek N, Thomas SN, Tzeng SC, Zhang H, Aebersold R

(2017) Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. Nat Commun 8 (1):291. doi:10.1038/s41467-017-00249-5

Gao Y, Wang X, Sang Z, Li Z, Liu F, Mao J, Yan D, Zhao Y, Wang H, Li P, Ying X, Zhang X, He K, Wang H (2017) Quantitative proteomics by SWATH-MS reveals sophisticated metabolic reprogramming in hepatocellular carcinoma tissues. Sci Rep 7:45913. doi:10.1038/srep45913

Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, Bonner R, Aebersold R (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics 11 (6):O111 016717. doi:10.1074/mcp.O111.016717

Guan S, Moran MF, Ma B (2019) Prediction of LC-MS/MS Properties of Peptides from Sequence by Deep Learning. Mol Cell Proteomics 18 (10):2099-2107. doi:10.1074/mcp.TIR119.001412

Krasny L, Bland P, Kogata N, Wai P, Howard BA, Natrajan RC, Huang PH (2018) SWATH mass spectrometry as a tool for quantitative profiling of the matrisome. J Proteomics 189:11-22. doi:10.1016/j.jprot.2018.02.026

Lam H, Deutsch EW, Eddes JS, Eng JK, Stein SE, Aebersold R (2008) Building consensus spectral libraries for peptide identification in proteomics. Nat Methods 5 (10):873-875. doi:10.1038/nmeth.1254

Lambert JP, Ivosev G, Couzens AL, Larsen B, Taipale M, Lin ZY, Zhong Q, Lindquist S, Vidal M, Aebersold R, Pawson T, Bonner R, Tate S, Gingras AC (2013) Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. Nat Methods 10 (12):1239-1245. doi:10.1038/nmeth.2702

Liu Y, Huttenhain R, Collins B, Aebersold R (2013a) Mass spectrometric protein maps for biomarker discovery and clinical research. Expert Rev Mol Diagn 13 (8):811-825. doi:10.1586/14737159.2013.845089

Liu Y, Huttenhain R, Surinova S, Gillet LC, Mouritsen J, Brunner R, Navarro P, Aebersold R (2013b) Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS. Proteomics 13 (8):1247-1256. doi:10.1002/pmic.201200417

Liu Y, Mi Y, Mueller T, Kreibich S, Williams EG, Van Drogen A, Borel C, Frank M, Germain PL, Bludau I, Mehnert M, Seifert M, Emmenlauer M, Sorg I, Bezrukov F, Bena FS, Zhou H, Dehio C, Testa G, Saez-Rodriguez J, Antonarakis SE, Hardt WD, Aebersold R (2019) Multi-omic measurements of heterogeneity in HeLa cells across laboratories. Nat Biotechnol 37 (3):314-322. doi:10.1038/s41587-019-0037-y

Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R (2018) Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol Syst Biol 14 (8):e8126. doi:10.15252/msb.20178126

Ortea I, Rodriguez-Ariza A, Chicano-Galvez E, Arenas Vacas MS, Jurado Gamez B (2016) Discovery of potential protein biomarkers of lung adenocarcinoma in bronchoalveolar lavage fluid by SWATH MS data-independent acquisition and targeted data extraction. J Proteomics 138:106-114. doi:10.1016/j.jprot.2016.02.010

Rosenberger G, Koh CC, Guo T, Rost HL, Kouvonen P, Collins BC, Heusel M, Liu Y, Caron E, Vichalkovski A, Faini M, Schubert OT, Faridi P, Ebhardt HA, Matondo M, Lam H, Bader SL, Campbell DS, Deutsch EW, Moritz RL, Tate S, Aebersold R (2014) A repository of assays to quantify 10,000 human proteins by SWATH-MS. Sci Data 1:140031. doi:10.1038/sdata.2014.31

Rosenberger G, Liu Y, Rost HL, Ludwig C, Buil A, Bensimon A, Soste M, Spector TD, Dermitzakis ET, Collins BC, Malmstrom L, Aebersold R (2017) Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. Nat Biotechnol 35 (8):781-788. doi:10.1038/nbt.3908

Rost HL, Rosenberger G, Navarro P, Gillet L, Miladinovic SM, Schubert OT, Wolski W, Collins BC, Malmstrom J, Malmstrom L, Aebersold R (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol 32 (3):219-223. doi:10.1038/nbt.2841

Schubert OT, Gillet LC, Collins BC, Navarro P, Rosenberger G, Wolski WE, Lam H, Amodei D, Mallick P, MacLean B, Aebersold R (2015) Building high-quality assay libraries for targeted analysis of SWATH MS data. Nat Protoc 10 (3):426-441. doi:10.1038/nprot.2015.015

Sidoli S, Lin S, Xiong L, Bhanu NV, Karch KR, Johansen E, Hunter C, Mollah S, Garcia BA (2015) Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH) Analysis for Characterization and Quantification of Histone Post-translational Modifications. Mol Cell Proteomics 14 (9):2420-2428. doi:10.1074/mcp.O114.046102

Sun B, Smialowski P, Straub T, Imhof A (2021) Investigation and Highly Accurate Prediction of Missed Tryptic Cleavages by Deep Learning. J Proteome Res 20 (7):3749-3757. doi:10.1021/acs.jproteome.1c00346

Tiwary S, Levy R, Gutenbrunner P, Salinas Soto F, Palaniappan KK, Deming L, Berndl M, Brant A, Cimermancic P, Cox J (2019) High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. Nat Methods 16 (6):519-525. doi:10.1038/s41592-019-0427-6

Volker-Albert MC, Pusch MC, Fedisch A, Schilcher P, Schmidt A, Imhof A (2016) A Quantitative Proteomic Analysis of In Vitro Assembled Chromatin. Mol Cell Proteomics 15 (3):945-959. doi:10.1074/mcp.M115.053553

Wang H, Shi T, Qian WJ, Liu T, Kagan J, Srivastava S, Smith RD, Rodland KD, Camp DG, 2nd (2016) The clinical impact of recent advances in LC-MS for cancer biomarker discovery and verification. Expert Rev Proteomics 13 (1):99-114. doi:10.1586/14789450.2016.1122529

Wang M, Wang J, Carver J, Pullman BS, Cha SW, Bandeira N (2018) Assembling the Community-Scale Discoverable Human Proteome. Cell Syst 7 (4):412-421 e415. doi:10.1016/j.cels.2018.08.004

Zhou XX, Zeng WF, Chi H, Luo C, Liu C, Zhan J, He SM, Zhang Z (2017) pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. Anal Chem 89 (23):12690-12697. doi:10.1021/acs.analchem.7b02566

Messner CB, Demichev V, Bloomfield N, et al (2021) Ultra-fast proteomics with Scanning SWATH. Nat Biotechnol 39:846–854. https://doi.org/10.1038/s41587-021-00860-4

Wang Z, Mülleder M, Batruch I, et al (2022) High-throughput proteomics of nanogram-scale samples with Zeno SWATH MS. Elife 11:e83947. https://doi.org/10.7554/elife.83947

Acknowledgements

I would like to thank my supervisor Prof. Dr. Axel Imhof, for offering me the opportunity to join his lab and for the excellent guidance and support throughout my whole doctoral study. During the past few years, I have been very grateful for your support and supervision of my study.

Special thanks go to Dr. Straub and Dr. Smialowski, who have always inspired me with smart ideas and suggestions when I faced difficult problems in bioinformatics, besides, they are very heartful for helping me to have a good study attitude.

I also want to thank my collaborators, Dr. Gerland, Dr. Lukacs, Dr. Kochanova, Dr. Aftab, Dr. Schmidt, Dr. Forne for their great help in the proteomic study and research, I have learned a lot of valuable knowledge about experiments and skills in processing biological datasets.

To all of the members in our lab and department, I would like to express my sincere thanks to those who are so helpful and kind to help me in the lab and department, you are the best people I have met in my life!

Last but not least, heartily thankful to my family, especially thanks to my mother, who cares for me so much during my study in Germany.

Thank you all so much!

Curriculum vitae

The CV is not accessible in the public version.