PhD degree in Molecular Medicine

(Curriculum in Bioinformatics)

European School of Molecular Medicine (SEMM)

University of Milan and University of Naples “Federico II”

Computational Methods to Study Known and Novel Protein Post-Translational Modifications by Mass Spectrometry

Enrico Massignani

European Institute of Oncology IRCCS (IEO), Milan

Matricola n. R12106

Supervisor: **Dr. Tiziana Bonaldi**, European Institute of Oncology IRCCS (IEO), Milan

Table of Contents

Figure Index

Table Index

Abbreviations

ACN	Acetonitrile
ADMA	Asymmetric di-methyl-Arginine
AML	Acute Myeloid Leukemia
CDDP	Cisplatin
CID	Collision Induced Dissociation
D3	D3-Acetyl
Da	Dalton
DDA	Data Dependent Acquisition
DDR	DNA damage response
DIA	Data indepenedent acquisition
dRT	Retention time difference
ECD	Electron capture dissociation
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FDR	False discovery rate
GAR	Glycine-Arginine rich
GO:BP	Gene ontology biological processes
GO:CC	Gene ontology cellular component
GO:MF	Gene ontology molecular function
H1F0K125me	Histone H1F0 Lysine 125 mono-methylation
H1F0S131me	Histone H3 Serine 131 mono-methylation
H3K27ac	Histone H3 Lysine 9 acetylation
H3K27me3	Histone H3 Lysine 9 tri-methylation
H3K36me2	Histone H3 Lysine 36 di-methylation
H3K4me3	Histone H3 Lysine 4 tri-methylation
H3K9ac	Histone H3 Lysine 9 acetylation
H3K9me3	Histone H3 Lysine 27 tri-methylation
H3S28me	Histone H3 Serine 28 mono-methylation
H3T32me	Histone H3 Threonine 32 mono-methylation
H4K16ac	Histone H4 Lysine 16 acetylation
H4K20me3	Histone H4 Lysine 20 tri-methylation
H4K8ac	Histone H4 Lysine 8 acetylation

H4R3me2a	Histone H4 Arginine 3 simmetric di-methylation
H4R3me2s	Histone H4 Arginine 3 asimmetric di-methylation
HAT	Histone acetyltransferase
HCD	High-energy Collision Dissociation
HCD	Higher-energy collision dissociation
HDAC	Histone deacetylase
hmSILAC	heavy methyl SILAC
hnRNP	heterogeneous ribonucleoprotein
HpH	High-pH
hPTM	Histone post-translational modification
iMethyl-SILAC	isomethionine methyl SILAC
IP	Immunoprecipitation
iTRAQ	Isobaric tag for relative and absolute quantitation
KD	Knock-down
KDM	Lysine demethylase
LC	Liquid Chromatography
LcPrb	Localization Probability
LDC	Large Drosha complex
LLPS	Liquid-liquid phase separation
LogRatio	Log2-transformed H/L ratio
m/z	Mass/charge ratio
MCC	Matthews correlation coefficient
ME	Mass Error
ML	Machine learning
MMA	Mono-methyl-Arginine
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
PEP	Posterior Error Probability
PIC	Phenylisocyanate
PKMT	Protein Lysine methyltransferase
PPI	protein:protein interactions
PRMT	Protein Arginine methyltransferase
Pro	Propionyl

PSM	Peptide-spectrum match
PTM	Post-translational modification
RBP	RNA-binding protein
RDM	Arginine demethylase
REAC	Reactome pathways
ROC curve	Receiving Operator Characteristics curve
RP	Reversed-phase
SAM	S-Adenosil-Methionine
SASP	Senescence-Associated Secretory Phenotype
SDMA	Symmetric di-methyl-Arginine
SF	Splicing factor
SILAC	Stable isotope labelling with amino acids in cell culture
TMT	Tandem mass tag
TNBC	Triple negative breast cancer
WB	Western blot
WT	Wild-type

1. ABSTRACT

A post-translational modifications (PTM) is the covalent modification of a protein after its synthesis, by either addition or removal of functional groups. PTMs significantly increase the complexity of the proteome by allowing each protein to exist in different forms, which can have different activity, stability or binding specificity. Because of their central role in protein regulation, PTMs are often deregulated in many diseases, especially cancer.

Since all PTMs introduce a change in the proteins mass, one of the leading techniques to study PTMs is Tandem Mass Spectrometry (MS/MS), which can measure the mass of molecules with high precision and resolution. Here, we apply computational methods to MS data, in order to study PTMs from two points of view.

In one project, we conducted a comprehensive analysis of Arginine (R) methylation at a global level. To achieve this, we significantly improved hmSEEKER, our in-house developed computational tool for the analysis of MS data from heavy methyl SILAC (hmSILAC) labelled samples, by implementing a machine learning model to identify methyl-peptides with high confidence. The hmSILAC-validated dataset was then combined with SILAC-based quantitative methyl-proteomics data from a set of SILAC experiments in which we profiled R methylation changes in response to different stimuli (e.g. Cisplatin treatment; inhibition of the major R methyltransferases; PRMT1 expression modulation) to generate the ProMetheus database (ProMetheusDB) of high-confidence methyl-sites. The in-depth analysis of R-methyl-sites inside ProMetheusDB reinforced the notion that protein R methylation modulates protein:RNA interactions but also provided new insights, such as the presence of several R-methyl-proteins involved in metabolism and immune response-related pathways or the fact that R methylation correlates differently with S/T-Y phosphorylation in response to different stimuli. Moreover, we employed computational methods to identify a number of protein:protein interactions that could be affect by this PTM and experimentally validated one of them. Finally, to fully exploit the potential of hmSILAC and hmSEEKER, we explored the application of our pipeline to the annotation of unconventional methyl-sites, which are largely uncharacterized.

Since MS has emerged as a powerful tool not only to characterize known PTMs but also to discover new ones, in a second project, we tried to expand the annotation of histone PTMs.

Histone lysine (K) acetylation and methylation are routinely used identify cancer subtypes and assess their severity; however, the unbiased nature of MS analysis has revealed the existence of several additional modifications that have yet to be systematically studied. Although theoretically possible, using MS to profile all PTMs that can occur on histones is impractical at the moment, due to limitations in the post-acquisition processing of the MS data. As a matter of fact, the most common computational tools for the identification of modified peptides are not suited to search more than 5-6 PTMs at a time, and this impairs the analysis of the combinatorial nature and cross-talk of histone PTMs.

To overcome this limitation, we took advantage of a novel peptide search engine named ionbot, which can perform an “open modification search” to identify an arbitrarily large number of modifications. We reasoned that such a tool would be suitable for the detection of hyper-modified histone peptides. Upon filtering of the results, we were able to annotate not only novel histone PTMs, such as short-chain acylations, but also several amino acid substitutions that could have a biological impact on the interactions between histone and other proteins (i.e. DAXX).

The analysis presented here permitted us to better understand the extent, dynamicity and biological role of protein R methylation and identified several modifications and mutation events on histone proteins. We believe further optimization and application of these methods will lead to the discovery of novel regulatory axes and the development of new cancer therapies.

2. INTRODUCTION

2.1. Protein post-translational modifications (PTMs)

Protein post-translational modifications (PTMs) consist in the covalent addition or removal of functional groups from specific amino acid residues within a protein sequence after said protein has been synthesized (Figure 1). PTMs fulfil a crucial role in countless biological processes because they allow cells to modify the physicochemical properties of a protein and therefore can rapidly regulate its function, its stability, or its binding specificity [1,2]. Some of the most common PTMs include proteolytic cleavage, phosphorylation, acetylation and ubiquitination. Because PTMs act as regulators of proteins stability and activity, their dysregulation is often associated with a variety of diseases, such as cancer, cardiovascular diseases and neurodegeneration [3-6].

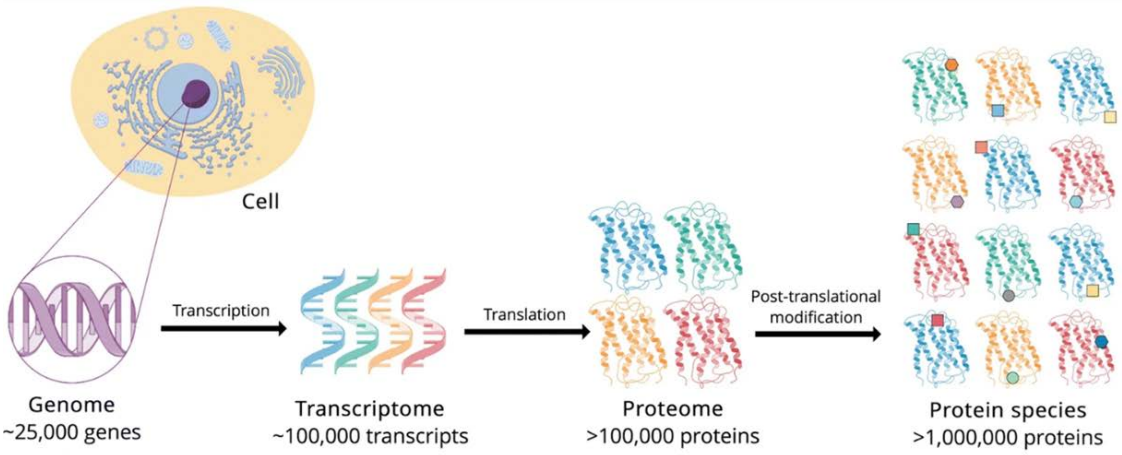


Figure 1. Increase in complexity of genetic information. *Alternative splicing can produce several transcripts and proteins starting from the same gene; proteins can then be further modified by post-translational modifications to regulate their function. Adapted from [7].*

2.1.1. Protein methylation

Protein methylation is a widespread PTM whereby one or more methyl (CH_3) groups are covalently attached to a residue. Lysine (K) methylation has been extensively studied in the context of histone modifications [8]. In the last decade, Arginine (R) methylation has also emerged as an important PTM not only on histones but also on non-histone proteins and there

is mounting evidence that R methylation is involved in almost every aspect of cell physiology, from RNA splicing and DNA repair to signal transduction and modulation of the immune response [9-11]. Methylation events have been observed also on Aspartate (D), Glutamate (E), Asparagine (N), Glutamine (Q), Histidine (H), Serine (S) and Threonine (T) [12]. For example, Q methylation of histone H2A is a nucleolus-specific PTM recognized by RNA polymerase I [13]; H methylation has been observed on actin and myosin and is involved in smooth muscle contraction [14]; finally, D residues that spontaneously convert into iso-aspartate during protein ageing are methylated by Protein L-isoaspartate O-methyltransferase (PCMT1) [15]. However, besides a recent study on H methylation [16], methylation of these residues has yet to be systematically analysed.

The biological donor of methyl groups in the enzymatic reaction of protein methylation is S-Adenosyl-Methionine (SAM), which is synthesized by the enzyme Methionine adenosyltransferase (MAT, most commonly known as SAM synthase) from Methionine (M), using ATP [17].

K can be mono-, di- or tri-methylated by protein lysine methyltransferases (PKMTs). R can be mono-methylated (MMA) or di-methylated, and the di-methylation can be symmetrical (SDMA) if the two methyl groups are bound to different nitrogen atoms of the guanidino group or asymmetrical (ADMA), if they are bound to the same atom (Figure 2). The different kinds of R methylation are deposed by different enzymes belonging to the protein arginine methyltransferase (PRMT) family: type I PRMTs (including PRMT1, PRMT2, PRMT3, PRMT4/CARM1, PRMT6 and PRMT8) catalyse the formation of MMA and ADMA; type II PRMTs (PRMT5 and PRMT9) catalyse the formation of MMA and SDMA; PRMT7, which is the only type III PRMT, catalyses MMA only [18]. PRMT1 is the most active enzyme among type I PRMTs and PRMTs in general, being overall responsible for ~85% of the methylation events in the human cell [19], while PRMT5 is the main type II PRMT [20].

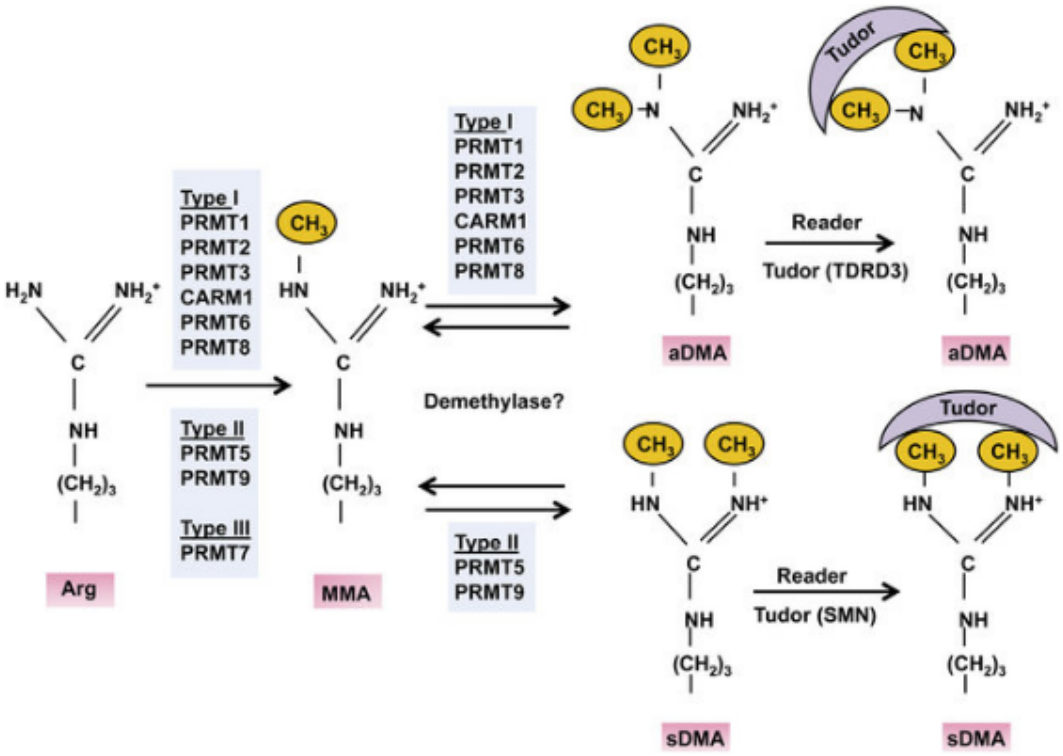


Figure 2. Schematic representation of unmodified, mono-methylated (MMA), symmetrically di-methylated (SDMA) and asymmetrically di-methylated (ADMA) arginine. *MMA formation can be catalysed by all PRMTs, while SDMA and ADMA formation are specific to Type I and Type II PRMTs respectively. ADMA and SDMA are then recognized by separate reader proteins owing to their different steric hindrance. Adapted from [21].*

An important feature of most PTMs is that they are reversible. Within the scope of protein methylation, removal of K methylation marks is carried out by lysine demethylases (KDMs), but there is an ongoing debate in the scientific community regarding the existence of arginine demethylases (RDMs). The KDM enzyme JMJD6 has been reported to catalyse R demethylation of both histone and non-histone substrates yet the biochemical evidence for its activity is still ambiguous [22]. Another possibility is that protein arginine deiminase 4 (PAD4) could counteract R methylation by removing an N atom from the guanidino group of arginine residues, producing citrulline as a result [23]. However, citrulline has different physicochemical properties than R and PAD4 can also catalyse the deimination of unmodified and glycated Arginines [24], therefore it is not clear whether PAD4 can be classified as a proper RDM or not.

Overall, methylation does not change the charge state of R or K, but alters their steric hindrance and reduces their ability to form hydrogen bonds with other molecules, thus affecting their interactions with other biomolecules, both positively and negatively [25,26]. Along this line, the molecular difference between ADMA and SDMA is crucial because the two isobaric modifications have different bulk, hence they are recognized and bound by distinct protein domains (Figure 2) and can lead to completely different functional outcomes as exemplified by asymmetrical and symmetrical di-methylation on R3 of histone H4 (H4R3me2a and H4R3me2s) which are associated to transcriptional activation and repression, respectively [27].

2.1.2. Protein arginine methyltransferases (PRMTs)

Protein R methylation and PRMTs have emerged as promising therapeutic targets since PRMTs are involved in a wide array of biological processes. PRMT1 was shown to be involved in DNA-Damage Response (DDR) by directing its activity towards chromatin proteins (such as RBMX, CHTOP and DDX17 [9]) and DNA damage proteins (such as MRE11 and 53BP1 [28]); in addition, PRMT1 asymmetrically di-methylates H4R3 upon replicative stress and induces cell expression of senescence-associated secretory phenotype genes [9]. PRMT4/CARM1 regulates nonsense-mediated decay and pre-mRNA splicing [29], and PRMT7 is involved in the response to stresses such as heat and proteasome inhibition [30]. In addition, PRMT1, PRMT4, PRMT5 and PRMT7 have all been reported to catalyse methylation on RNA-binding proteins (RBPs), weakening their interactions with RNA molecules, with implications in the regulation of mRNA splicing, miRNA maturation, translation and RNP granules assembly [9,10,31-35].

2.1.3. Protein methylation in disease

Overexpression of PRMTs is frequently associated with various types of cancer, including both solid tumours (i.e., brain, breast, lung, colon, bladder, head and neck cancer) and haematological malignancies (such as leukaemia) [36]. For instance, high levels of PRMT1 in breast cancer promote cell migration and metastasis formation, while overexpression of CARM1 blocks myeloid differentiation in Acute Myeloid Leukemia (AML) [37]. Aberrant levels of PRMTs have also been observed in other diseases such as neurodegeneration and cardiovascular diseases, where the accumulation of ADMA inhibits the activity of Nitric Oxide Synthase (NOS) and reduces the bioavailability of nitric oxide, a potent vasodilator

[38]. Moreover, there is growing evidence that PRMTs are also implicated in muscle development and regeneration and muscle cells metabolism [39].

Hence, there is a great interest in developing therapies that are based on the pharmacological modulation of protein R methylation levels. In fact, several PRMTs inhibitors are being developed, with some already undergoing clinical trials (trial identifiers NCT03573310, NCT02783300 and NCT03614728) [36,40].

2.1.4. Histone post-translational modifications (hPTMs)

Post-translational modifications of histone proteins (hPTMs) represent a crucial regulatory mechanism of chromatin structure and function, with implications in processes such as DNA transcription and repair, development, ageing and pathogenesis [41].

The state of chromatin is controlled by the balance between activating histone marks (such as H3K4me3, H3K9ac, H3K36me2 and H4K8ac) and repressive ones (such as H3K9me3, H3K27me3 and H4K20me3) [42]. However, histones can be modified by different PTMs and at different sites simultaneously, resulting in a cross-talk between the different marks. This cross-talk can occur at the level of the individual site, the histone tail, or the nucleosome [43-45]. An interesting combination of hPTMs is found within the 'bivalent domains' of embryonic stem cells, where the H3K4me3 active mark and the H3K27me3 repressive mark co-exist [46]. Bivalent domains enable embryonic stem cells to quickly activate or repress the expression of crucial developmental genes but are lost during differentiation.

Although researchers have mostly focused on the most abundant histone marks (namely K methylation and K acetylation; Figure 3), the list of potential hPTMs is constantly growing. In the last decade, several novel hPTMs, such as acylation, lipidation, biotinylation and serotonylation, have been discovered to have a biological impact on nucleosome stability and gene expression [47]. Moreover, some of these hPTMs (such as acylation and glycation) appear not to be enzymatic, thus challenging the assumption that every hPTMs has its own writers; however, the fact that their erasers have been identified suggests that they indeed have a functional role in the cell.

Analysis of all the PTMs that could occur on histones is challenging for several reasons. From a biochemical point of view, many of the newly discovered hPTMs have very low abundance,

hence they not only require highly sensitive methods to be identified but it is also difficult to distinguish the genuine hPTMs from the artefacts that might be introduced during sample preparation [48]. Most importantly, hPTMs are often present in combination with each other in what is known as the "Histone Code", thus it is often necessary to profile several hPTMs at once [49]. While the latter problem can theoretically be addressed by mass spectrometry, most computational tools for the analysis of mass spectrometry data are not designed to identify more than ~5 PTMs at the same time, thus limiting this approach to the study of the more common hPTMs (e.g. methylation, acetylation, phosphorylation). A possible solution to this issue is discussed in the following chapters.

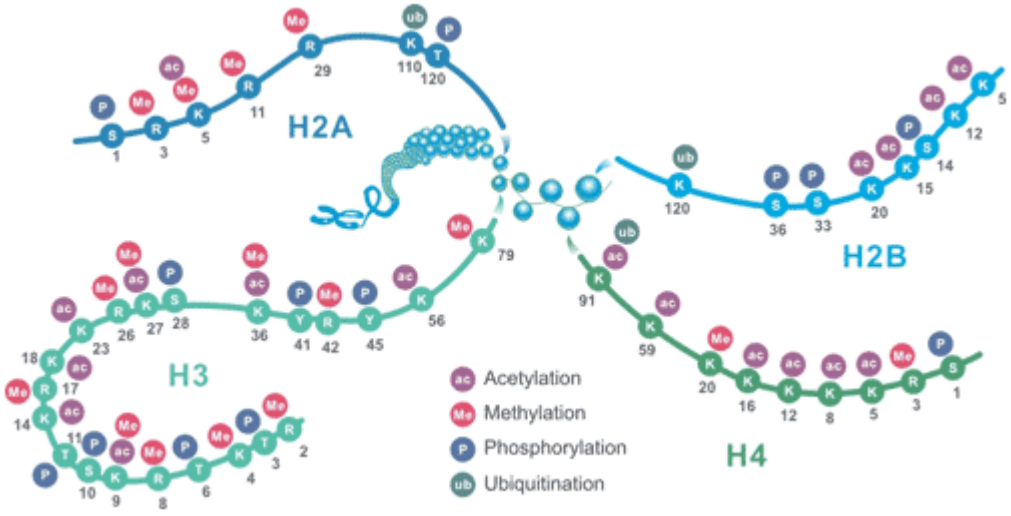


Figure 3. Schematic representation of common histone modification sites. *(Taken from www.cusabio.com.)*

2.1.5. Histone modifications in disease

Because hPTMs are key players of epigenetics, dysregulation or mutation of the enzymes that are responsible for deposing, removing or binding these modifications (termed "writers", "erasers" and "readers" respectively) is commonly observed in disease. For example, the global reduction of H4K16ac and H4K20me3 levels is considered a hallmark of many human cancers [50]; in addition, cancer cells show altered levels of H3K4me3 [51], H3K9me [52] and H3K27me3 [53]. Moreover, aberrant levels of histone acetylation, methylation and phosphorylation have been observed in neurodegenerative disorders, such as Alzheimer's, Parkinson's and Huntington's diseases [54].

Histone modifications are routinely used as biomarkers to evaluate the severity of cancer and predict its evolution [55]. In addition, small-molecule drugs that modulate the activities of hPTMs writers and erasers have attracted the interest of the scientific community [56]. One limitation of current approaches is that they focus on a few methylated and/or acetylated K residues on histones H3 and H4, whereas little is known about alterations of other hPTMs and their combinatorial patterns. For instance, cellular metabolism produces a variety of short-chain acyl-CoA molecules, such as crotonyl-CoA, propionyl-CoA and butyryl-CoA. These molecules are similar to acetyl-CoA and can be recognized by histone acetyltransferase (HATs), resulting in crotonylation, propionylation and butyrylation of K residues, which have been linked to the regulation of cell metabolism [57].

Therefore, the profiling of hPTMs in a comprehensive and unbiased way could allow us to uncover novel biomarkers and therapeutic targets.

2.2. Mass Spectrometry (MS)

Mass Spectrometry (MS) is an analytical technique that is widely adopted for the identification and quantitation of proteins in biological samples. MS permits the measurement of the exact molecular weights of molecules; this allows distinguishing proteins and peptides that carry PTMs from their unmodified counterparts since PTMs inevitably introduce a mass change. Moreover, high-resolution mass spectrometers can identify small modifications, such as mono-methylation (Δmass +12.04 Da), or distinguish between PTMs that have similar masses, such as acetylation and tri-methylation (Δmass +42.01 and +42.04 Da, respectively). In the last decades, several technological improvements led to the development of mass spectrometers characterized by progressively higher resolution, sensitivity and performance.

Nowadays, the vast majority of MS-based proteomics data is obtained through Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS). In its classical workflow, known as the “bottom-up” approach, proteins are digested with specific proteases into peptides before the MS analysis. The protein composition of the sample is then inferred from the peptides identified. The most widely used protease in MS-based proteomics is trypsin, which cleaves at the C-terminus of K and R residues unless they are followed by a proline (P). Tryptic peptides have an average length of ~12 amino acids and present a positively charged residue at the C-terminus, making them optimal for MS/MS analysis. However, it has been shown that the use

of multiple proteases in parallel can greatly improve the protein sequence coverage [58]. One promising alternative protease is LysArginase, which cleaves at the N-terminus of K and R, thus mirroring the cleavage of trypsin [59]; like tryptic peptides, peptides obtained from LysArginase digestion are 12-amino acids long on average but present a positive charge acid at the N-terminus. Some proteases are suited for specific applications: for instance, ArgC and LysC, which respectively cleave at the C-terminus of R or K residues only, are often used to digest histone proteins (see chapter 2.8).

Upon digestion, peptides are separated on a reversed-phase (RP) chromatographic column by Liquid Chromatography (LC), to separate them according to their hydrophobicity [60]. The RP-LC column is directly linked to the ion source of the mass spectrometer, which ionizes the peptides as they elute from the column. The mass spectrometer first measures the mass/charge (m/z) ratio of the whole peptides in what is called MS1 or MS scan. Afterwards, the most intense peptide ions are isolated, fragmented and subject to another round of acquisition, termed MS2 or MS/MS. The spectra obtained by recording the m/z ratios of the fragment ions are finally analysed with dedicated bioinformatics tools that identify the peptides based on their MS/MS patterns.

2.3. The Q Exactive Orbitrap mass spectrometer

Of all the mass spectrometers available, instruments based on Orbitrap mass analyzers have become the instruments of choice for the majority of “bottom-up” proteomics experiments, thanks to their high resolution and precision. All the MS data presented in this thesis were acquired on a Q Exactive mass spectrometer [61] (Figure 4).

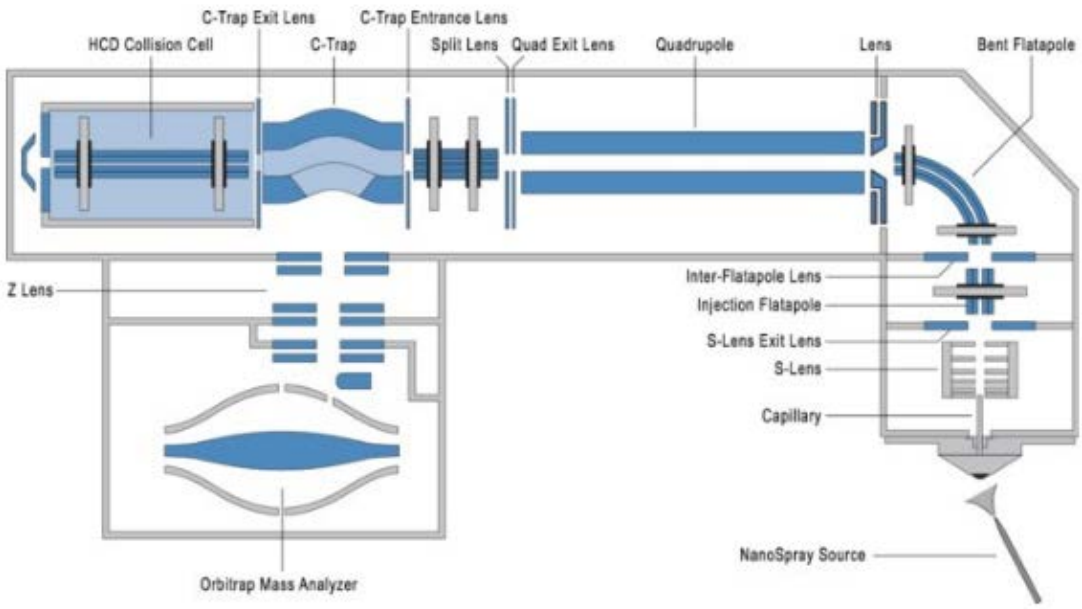


Figure 4. Schematic representation of Q Exactive mass spectrometer. *Taken from [61].*

2.3.1. Ion source: Electrospray ionization

Mass spectrometers measure the m/z ratio of charged gas-phase molecules, therefore peptides must be brought into the gas phase and ionized. In a Q Exactive, ions are generated via Electrospray ionization (ESI) [62]. ESI sources generate gas-phase ions from peptides in an aqueous solution and their charge is controlled by the pH of the solution (Figure 5). At acidic pH values, carboxyl groups and amine groups are protonated and give a positive charge to peptides, while at basic pH, de-protonation of the same groups confers a negative charge. Because fragmentation of the peptide ions is favoured by positive charges, ESI of peptides is usually done in the positive ion mode. ESI can produce multiply charged ions: in fact, tryptic peptides will always carry at least two positive charges (one from the amine at the N-terminus and one from the K/R residue at the C-terminus), but peptide ions with 3 or more charges are also possible if the peptide sequence contains multiple K/R/H residues. During the ESI process, the application of a high voltage (2–6 kV) between the end of the LC column and the entrance of the mass spectrometer forms an electrically charged spray that causes de-solvation of peptide droplets and the formation of ions. The heated capillary and the sheath gas flow, present at the mass spectrometer inlet, help this process. Because they produce ions from a solution, ESI sources are usually combined “on-line” with the LC instruments for continuous

analysis of the sample. Usually, an RP-LC column is used to separate the peptides at low pH: this reduces the sample complexity and also confers a positive charge to the peptides for the subsequent ionization. An important development of ESI includes nano-ESI sources, where the flow rates are lowered to nL/min to improve the sensitivity and increase the concentration of the analytes.

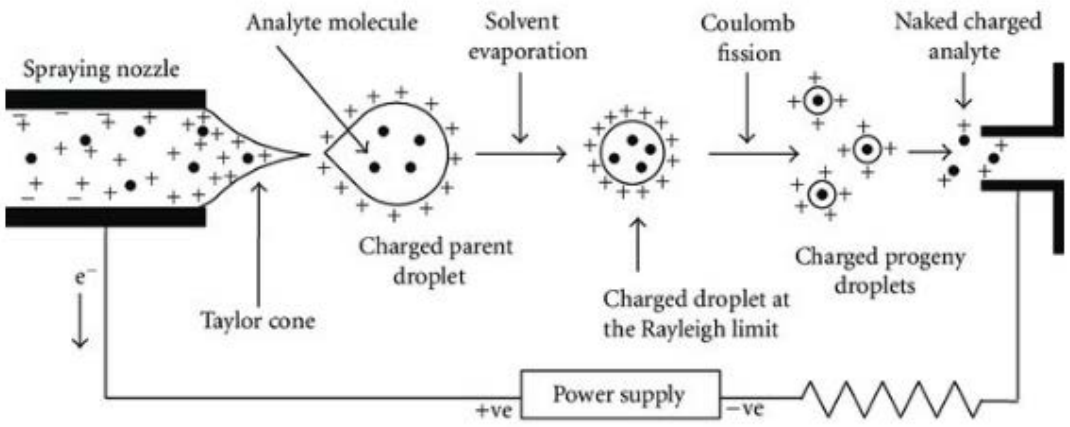


Figure 5. Schematic representation of the electrospray ionization (ESI) process. *Peptides elute from the HPLC column and are ionized by a high voltage applied between the capillary and the mass spectrometer. The charged solvent forms several small droplets and, once it evaporates, analyte ions are released into the gas phase. Taken from [62].*

2.3.2. Mass analyzer and fragmentation techniques

The mass analyser is the central core of the mass spectrometer and its main role consists in the storage and separation of ions based on their m/z. In a Q Exactive instrument, the mass analyzer is a Quadrupole, which applies a magnetic field that causes ions to oscillate as they move through the instrument [63]. Depending on the magnetic field, only ions with a certain m/z will maintain a stable trajectory and reach the C-trap, where they are then guided to the Orbitrap analyzer, while other ions will collide with the rods (Figure 6).

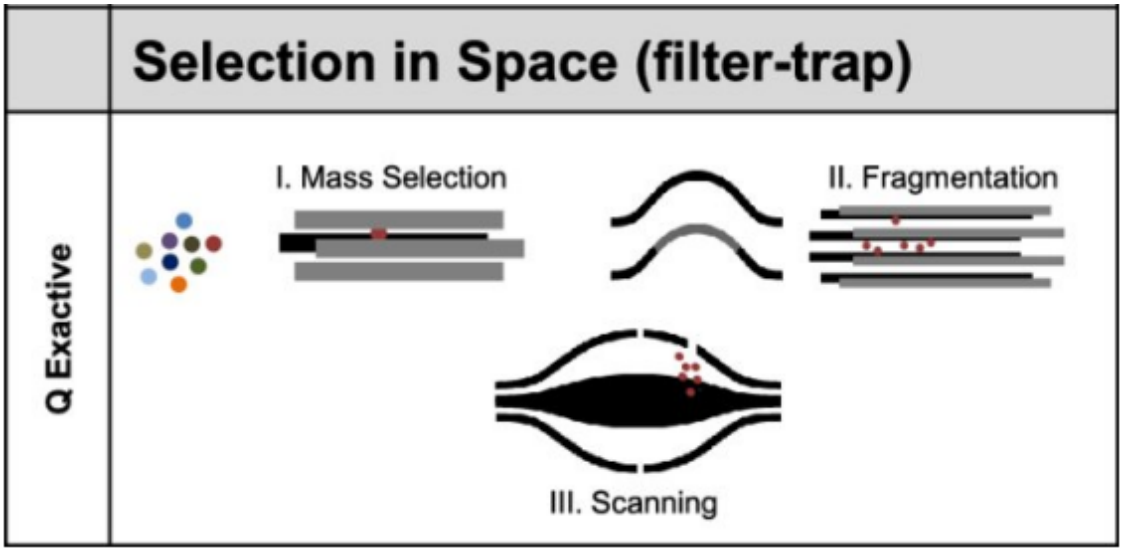


Figure 6. Schematic representation of the Q Exactive mass analyzer and and detector. *The Q Exactive instrument performs mass selection is "in space" as only ions with a specified m/z range have stable trajectories inside the quadrupole (I) and can be transferred to the storage or fragmentation (II) devices before Orbitrap analysis (III). Adapted from [61].*

In a classical proteomics workflow, the mass spectrometer works in Data-Dependent Acquisition (DDA) mode, whereby peptide ions in a given time window are fragmented starting from the most intense ones; as a result, DDA is biased towards the most intense ions (i.e. the most abundant peptides/proteins in the sample). Therefore, it is essential to reduce the sample complexity before the LC-MS/MS analysis, for example via an off-line chromatographic fractionation [9,64]. For the same reason, peptides bearing PTMs of interest have to be enriched from biological samples, since said PTMs are usually substoichiometric and would be masked by the much more abundant unmodified peptides. Ions isolated in this way are then fragmented to produce fragment ions, which allows the reconstruction of the peptides amino acid sequence and the identification and localization of PTMs, if any.

Different fragmentation techniques can be used, but the most common are the collision-induced dissociation (CID) and the higher energy collision dissociation (HCD) methods. In these types of fragmentation, protonated peptide ions undergo multiple collisions with rare gas atoms and beak at the peptide bonds (-CO-NH-), leading to the generation of fragment ions. CID and HCD are both suitable for the analysis of PTMs [65] but produce only limited information for peptides longer than 15 amino acids. In these cases, electron capture

dissociation (ECD) and electron transfer dissociation (ETD) can be used, which induce fragmentation of the peptide backbone based on gas-phase reactions exploiting either thermal electrons or the formation of radical ions [66]. Upon fragmentation, if the charge is retained by the C-terminal peptide fragment, a y-ion is generated; otherwise, a b-ion is generated. The "ladders" of b- and y-ions that result from the fragmentation can then be used to deduce the amino acid sequence [67]. When a peptide bond is broken, whether a b- or y-ion is produced depends on the composition of the fragments: fragments containing basic amino acids (i.e. amino acids that naturally acquire positive charges, such as K, R and H), are more likely to retain the charge and be recorded by the detector. As such, the MS/MS spectra of tryptic peptides, which end with a K or R, are usually dominated by y-ions, whereas LysArginase digestion favours the detection of b-ions because it produces peptides with a K/R residue at the N-terminus.

2.3.3. Detector

The Orbitrap mass analyzer [68] is characterized by a coaxial central spindle electrode surrounded by a barrel-like electrode. It produces an electrical field that causes ions to oscillate around a central electrode in ring shapes. Each ion has a specific oscillation frequency according to its m/z value, which is recorded by the detector and Fourier-transformed to generate high-resolution mass spectra (Figure 6).

2.4. Computational methods to identify peptides by MS

MS raw data files generated by the mass spectrometer are interpreted by mapping the MS/MS spectra back to the peptide that was subjected to fragmentation. Several computational methods have been developed to analyze MS/MS spectra, which can be classified into three categories: de novo sequencing; protein database searching; spectral library searching. The MS data shown in this thesis was analyzed by database searching.

2.4.1. The database searching strategy

Protein database searching is currently the most popular method to identify peptides and proteins from MS raw data. Database search computational tools perform an in silico digestion of a reference proteome and generate theoretical MS/MS spectra for the obtained peptides, which are subsequently compared to the experimental spectra acquired on the instrument

(Figure 7) [69]. Although each database search engine employs a different scoring function, most of them score a peptide-spectrum match (PSM) based on the number of fragment ions that are in common between the theoretical and experimental spectra, as well as the total intensity of the matched ions. This strategy was pioneered by SEQUEST [70] and Mascot [71] in the '90s and was later implemented by other search engines such as OMSSA, X!Tandem, Andromeda and MS-GF+ [72-75].

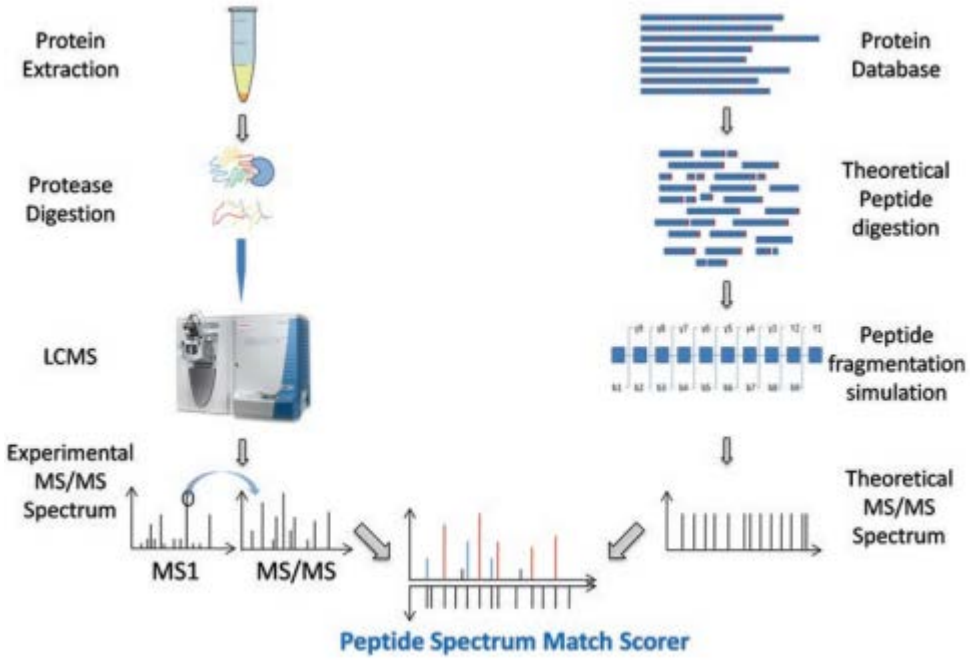


Figure 7. Schematic representation of the database search process. *Experimental MS/MS spectra obtained from the mass spectrometer are compared to theoretical MS/MS spectra generated in silico. Peptide-spectrum matches are scored based on the number and the intensity of the ions in common between theoretical and experimental spectra. Adapted from [76].*

Despite being routinely used, database search tools have two critical flaws: first, repeated identification of the same peptide is time-consuming; second, searching for large protein databases or databases that contain PTMs leads to an exponential increase of the candidate peptides that need to be compared with the query spectrum, increasing the chances of a peptide being matched at random.

One limitation of traditional database search engines is that, when generating the theoretical spectra library, they assign the same intensity value to all predicted peaks; this is a simplification, as some fragment ions are more likely to be generated than others, depending on both the amino acid composition of the peptides themselves and the fragmentation method used (see chapter 2.3.2). However, newer search engines are starting to incorporate tools to predict fragment ion intensities, thus producing theoretical spectra that are closer to the real ones [77].

2.4.2. Challenges in the identification of modified peptides

Since all PTMs introduce a mass difference (Δm/z) in peptides, MS has emerged as an advantageous technique to study them. The general workflow used for protein identification usually has to be adapted to address some issues that are intrinsic to PTM studies. First, PTMs are substoichiometric and therefore their analysis requires enrichment steps and high sensitivity of detection. Second, the covalent bond between the PTM and the amino acid side chain of the peptide can be labile (as for phosphorylation), thus it is difficult to maintain the peptide in its modified form during sample preparation and ionization. Third, some PTMs are highly dynamic and thus the enzymes responsible for removing the PTM of interest need to be inactivated. From a computational point of view, PTMs can be divided into three groups:

1. in vivo modifications (i.e. phosphorylation);
2. in vitro artefacts generated by sample handling (i.e. methionine oxidation or D/E methyl-esterification);
3. in vitro modifications introduced on purpose as part of the sample preparation protocol (i.e. Cysteine carbamidomethylation, Phenylisocyanate, mass tags).

On the one hand, the derivatization of peptides with chemical modification during sample preparation has very high efficiency, meaning that these PTMs can be treated as “fixed modifications”: in this case, the database search tool will simply replace all instances of a compatible residue in the database with the modified version of that residue. On the other hand, in vivo PTMs and artefacts must be treated as “variable modifications” because it is not known in advance which peptides will be modified, and on which specific residue, so the search engines must generate theoretical spectra not only for unmodified peptides but also for any possible modified form of those peptides. This is especially problematic in the case of peptides that have multiple potential modification sites and/or when many PTMs are searched

simultaneously because it leads to an exponential increase in the size of the database [69,78]. Thus, variable PTMs produce an increase in the database search space which results in a higher proportion of random matches. As a consequence, the software is forced to apply more stringent filtering of the PSMs to ensure the final FDR is below 1%, leading to an increased number of missed identifications (i.e. false negatives).

Localization of the PTMs on the correct residues of a given peptide is also challenging, as it requires high coverage of the peptide sequence, and several tools have been developed to evaluate the confidence of modification sites assignments [79].

2.5. PTMs profiling by quantitative proteomics

Accurate qualitative identification of peptides and proteins in a sample is indispensable for any proteomics research; however, quantitative data is also necessary to understand the biological mechanisms under study. Mass spectrometry is not a proper quantitative technique for the following reasons: first, the intensity of each peptide ion is proportional to its abundance but also depends on the physicochemical properties of the peptide (i.e. molecular composition, charge and hydrophobicity); second, the comparative analysis between two different LC-MS/MS runs is influenced by external variations like temperature and chromatography reproducibility; finally, in DDA methods, the choice of the precursor ions to be fragmented is stochastic and depends mostly on their abundance, therefore there is no guarantee the same peptides will be identified across all MS/MS runs, especially low abundant ones such as modified peptides.

Different strategies have been developed to quantify proteins by mass spectrometry, which can be divided into label-free strategies, whereby the abundance of proteins in different LC-MS/MS runs is compared, and isotope labelling strategies, where samples are labelled to introduce a mass difference which allows them to be distinguished at the MS or MS/MS level [80,81].

2.5.1. Stable Isotope Labelling with Amino acids in Cell culture (SILAC)

To produce the quantitative data presented here, we employed a metabolic labelling strategy named Stable Isotope Labelling with Amino acids in Cell culture (SILAC), in which isotope-labelled amino acids are added to the cell growth medium to be incorporated in the proteome

during protein biosynthesis (Figure 8) [82]. A major advantage of SILAC labelling is that corresponding samples can be mixed immediately after cell harvesting, minimizing errors due to handling of the samples and leading to intrinsically higher accuracy compared to other approaches (i.e. label-free, TMT, iTRAQ). Traditionally, the amino acids used to metabolically label proteins are K8 and R10 because they ensure that, upon trypsin digestion, all peptides produced by a protein contain at least one isotopically labelled residue (except the C-terminal one). Upon MS/MS analysis, each peptide is detected as a pair of peaks (heavy and light) separated by a specific Δm/z which depends on the peptide sequence and the labelling strategy.

Since its development, SILAC has been successfully and widely applied to proteomics studies, including the profiling of PTMs [83-85]. This approach can be modified to compare multiple conditions at the same time: for example, in triple SILAC, a third sample of cells is labelled with "medium" amino acids (usually K4 and R6), and the three samples are mixed in 1:1:1 proportion; this allows the distinction of three different proteomes during the MS analysis. SILAC labelling can also be used to directly label PTMs, as in the heavy methyl SILAC (hmSILAC) approach (see chapter 2.7.2). Another important development of SILAC is the Super-SILAC, which allow quantifying proteins from animal or human tissues. In super-SILAC, a heterogeneous mix of heavy-labelled cell lines is spiked in a set of samples to serve as an internal standard (see chapter 1.8.1).

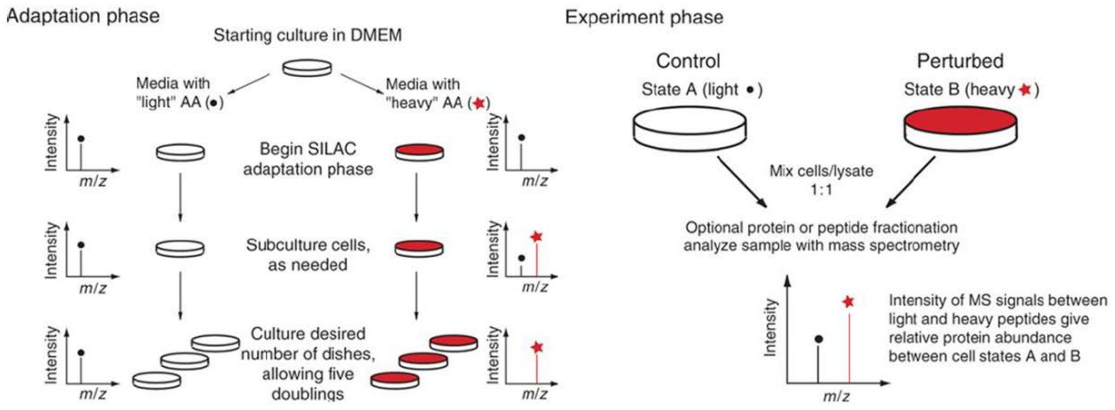


Figure 8. Schematic representation of a SILAC experiment. *SILAC allows up to three experimental conditions (Light, Medium and Heavy) to be compared. Compared to methods like TMT, SILAC has the advantage that sample cells can be mixed immediately after*

harvesting, so that all samples undergo exactly the same steps of sample preparation, thus achieving high accuracy. Adapted from [82].

2.6. MS-based analysis of PTMs with MaxQuant

Analysis of the MS raw data was performed with MaxQuant, a popular platform widely used in MS-based proteomics [86].

2.6.1. SILAC pair detection

MaxQuant efficiently supports SILAC experiments for quantitative proteomics. Potential SILAC pairs are required to have the same charge and a Pearson correlation value of their peak profiles greater than 0.6. In addition, they need to have a specific delta mass, which depends on the peptide sequence and labelling strategy. However, because SILAC pairs are detected before peptide identification, this delta mass is unknown. Thus, MaxQuant assumes a maximum of three labelled amino acids per peptide, and tests each possible combination of labels. This assumption only works when all or most of the labelled amino acids are also cleavage sites, such as when trypsin is used alongside labelled R and K. The SILAC ratio of a peak pair is then calculated as the slope of a straight line passing through the centroids of the two peaks.

2.6.2. Andromeda database search

Since its first publication, MaxQuant has regularly received updates, the most notable being the development of Andromeda [75], a dedicated probability-based protein database search engine. Like all database search engines, Andromeda first generates a list of all possible peptides generated from the protein sequences in a database based on the user-defined digestion rule and considering all possible combinations of indicated variable modifications. Then, each experimental MS/MS spectrum is compared to the theoretical spectra of all peptides that have a mass compatible with the precursor mass of the experimental spectrum. For each peptide, Andromeda calculates the probability that at least k of the n theoretical fragment ions are present in the experimental spectrum. Andromeda only considers the q most intense ions for every interval of 100 Th, so the probability of observing each theoretical ion individually is always $q/100$. This calculation is repeated with different q values, and only the

highest score is kept. The lower *q* is, the higher the Andromeda score is for given *n* and *k* values. Consequently, the score will be higher if the most intense peaks are matched since these peaks will be considered by Andromeda even at lower values of *q*.

2.6.3. False Discovery Rate (FDR) and Posterior Error Probability (PEP)

The output of Andromeda is a list of PSMs ranked according to the search score. The overall score distribution of PSMs in a proteomics experiment can be represented as a mixture of two distributions: one for correct identifications and one for incorrect ones. The method for discriminating the two distributions is the so-called Target Decoy Approach [87]. In this approach, the real protein database (Target database) is searched together with a database of random sequences (Decoy database); this decoy database is usually obtained by reverting or shuffling the target sequences, to ensure it has the same size and amino acid composition of the target database. The rationale of this strategy is that all identifications coming from the decoy database will be random matches and their score distribution will reflect the score distribution of random matches that originate from the target database. Therefore, if *N* peptides are identified with score >*S* in the decoy, then it can be assumed that the set of peptides identified in the target database with score >*S* also contains *N* false positives. Most proteomics pipelines, including MaxQuant, automatically filter the results of the database search so that the final PSM list has an FDR of 1% (Figure 9). Following the same approach, it is possible to estimate the number of false identifications at a specific score value: the number of decoy PSMs with score=*S* divided by the number of target PSMs with score=*S* is termed Posterior Error Probability (PEP) and represents the probability of a PSM with that specific score being a false positive.

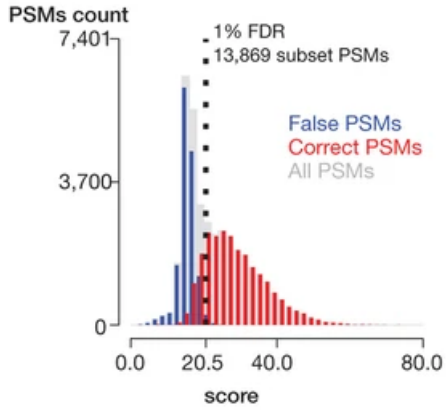


Figure 9. Histograms of correct and incorrect PSM scores. *Adapted from [88].*

2.7. Approaches for the study of global protein methylation

2.7.1. Non-MS-based methods

Several analytical techniques have been developed over the years to study the methyl-proteome [89]. A method that was widely used to detect methyl-proteins is radiolabelling with 14C- or 3H-labelled SAM, followed by gel electrophoresis separation of proteins and scintillation counting to measure the number of radioactive methyl groups transferred to each protein. However, radiolabelling was challenging because of the weak radioactive signal emitted by 14C and 3H. Also, in vivo radiolabelling did not permit the identification of the methylated protein or the localization of the methylated residue, thus requiring further analytical steps.

Antibodies specifically raised against methylations, both global (anti-pan-methyl antibodies) and specific for proteins or sites of interest (e.g. anti-Histone H4R3me2s), have successfully helped to characterise the activity of several histone methyltransferases and demethylases [22,90]. Despite being widely used, the sensitivity of this method is strongly dependent on the specific antibody and quantifying methylation changes across different conditions is difficult due to the antibody readout being non-linear. An advantage of this technique, however, is the possibility to raise antibodies that are site-specific and capable of distinguishing between different forms of methylation, such as SDMA and ADMA.

A third strategy uses SAM-analogues containing a terminal alkynyl group that is transferred by methyltransferases to methylated protein together with the methyl group; afterwards, a probe is covalently bound to the alkynyl group via “click” chemistry [91]. This method is suited for studying methylation induced by a specific methyltransferase but requires the ad hoc production of analogues that are compatible with the PRMT of interest or, conversely, the engineering of the PRMT itself [92].

2.7.2. MS-based methods

In recent years, MS-based approaches have emerged as a powerful analytical strategy for global investigation of protein PTMs, including methylation, thanks to the possibility of distinguishing modifications on different residues and profile modification changes in quantitative experiments. Despite these advantages, MS-based identification of protein

methylation is challenging and prone to high FDR, as shown by the works from the groups of Oreste Acuto and Marc Wilkins [11,93]. The limitations of protein methylation profiling by MS are the following:

1. As already discussed, methylation is sub-stoichiometric, thus requiring extra steps of methyl-peptides enrichment and off-line fractionation during the sample preparation.
2. Several pairs of amino acids have chemical structures that differ from each other for the presence of an additional methyl group, such as Glycine and Alanine, or Alanine and Valine As such, several amino acid substitutions result in mass differences that can be misinterpreted as methylations.
3. In vitro artefacts, such as methyl-esterification of D and E, are also isobaric to mono-methylation and can lead to false-positive identifications.
4. Protein methylation can exist in multiple forms and even restricting the analysis to the most common ones (mono-, di- and tri-methylated K and mono- and di-methylated R) still results in five variable modifications that need to be specified in the database search engine. Thus, false-positive identifications can arise from the search space explosion, as discussed in Chapter 1.4.2. This becomes a critical issue when the database search is expanded to methylated residues beyond K and R.

2.7.3. High-confidence identification of in vivo methyl-peptides with hmSILAC

To control the FDR associated with MS-based methyl-proteomics experiments, the group of Matthias Mann proposed the heavy methyl SILAC (hmSILAC) strategy [94]. hmSILAC follows the same workflow as standard SILAC, with the difference that the “heavy” growth medium contains [$^{13}CD_3$]-Methionine. This “heavy” methionine is metabolically converted by the cells into [$^{13}CD_3$]-SAM (Figure 10A). Cells incorporate heavy Methionine into the backbone of newly synthesized proteins and, in addition, the heavy methyl-groups are transferred by PKMTs and PRMTs from [$^{13}CD_3$]-SAM to their substrates. As a result, in hmSILAC, the mass difference between “heavy” and “light” peptides is encoded by the methyl groups themselves. This strategy is advantageous for two main reasons: First, heavy methyl groups can only be added through enzymatic reactions, therefore only peptides that are methylated in vivo and those containing an M residue will be detected as a doublet; second, the $\Delta m/z$ between heavy and light peptides depends on the peptide methylation state, permitting the distinction of mono-, di- and tri-methylated residues (Figure 10B).

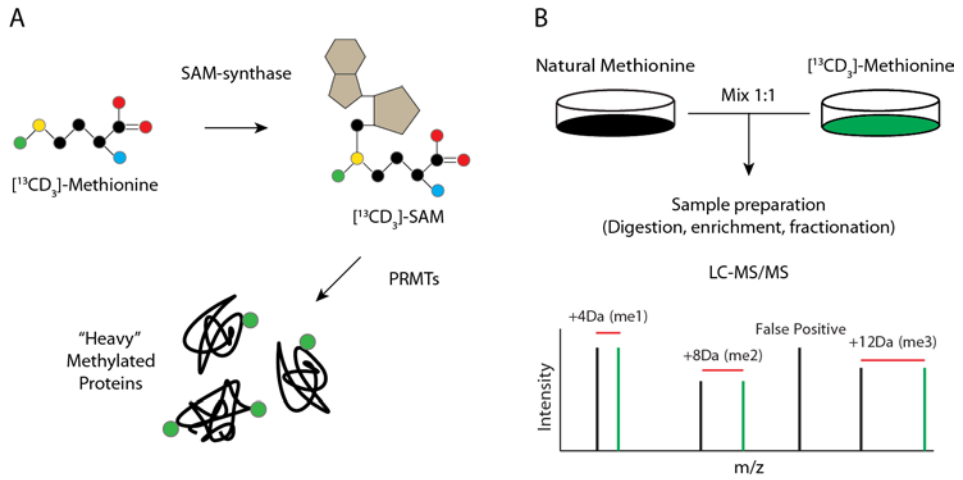


Figure 10. Schematic representation of the hmSILAC strategy. A) *Heavy Methionine is converted into heavy SAM and methyltransferases transfer heavy methyl groups to proteins.* ***B)*** *Upon mixing Light- and Heavy-labelled samples 1:1, methylated peptides will be detected by MS as pairs of co-eluting peaks with the same intensity and separated by a specific mass difference. Instead, false positives are detected as single peaks because they can only exist in their Light form.*

In 2015 a variant of hmSILAC name iso-methionine methyl-SILAC (iMethyl-SILAC) was proposed by the Acuto group [11]. In iMethyl-SILAC, the "light" cells are labelled with [$^{13}C_4$]-Methionine, which is isobaric to [$^{13}CD_3$]-Methionine but, due to a different distribution of the isotopes in its structure, does not produce heavy methyl-groups upon the methyltransferase enzymatic reaction. The advantage of iMethyl-SILAC is that any detected doublet can be unambiguously interpreted as a methylated peptide, whereas, in traditional hmSILAC, M-containing peptides also produce doublets, leading to more complex MS spectra.

At the time of writing, the hmSILAC and iMethyl-SILAC strategies are still rarely used, mostly because searching for hmSILAC peptide pairs is a computationally demanding task and there are few tools able to perform this kind of analysis. For instance, MaxQuant is not able to correctly calculate the $\Delta m/z$ of SILAC pairs if the label is encoded by a variable modification because the detection of peak pairs is performed before the identification step and the assumption of a maximum of three labelled amino acids per peptide does not hold. One algorithm for the analysis of hmSILAC data is MethylQuant, which, starting from a list of methylated peptides, searches for their hmSILAC counterparts in the raw MS data. This tool

was successfully applied in two works, to identify with high confidence methyl-sites in human T cells and yeast cells, respectively [11,95]. Within this context, one of the aims of my PhD was to develop a new tool that would allow us to robustly and automatically analyse hmSILAC data previously processed by the MaxQuant algorithm. By using MaxQuant as the starting point for the analysis of hmSILAC data, hmSILAC experiments could be integrated with quantitative data from SILAC experiments to produce high-quality functional data of methyl-proteome dynamics, retaining low FDRs thanks to hmSILAC validation.

2.8. MS-based profiling of histone PTMs

Bottom-up MS-based proteomics has widely and successfully been used for the identification of hPTMs. One important caveat compared to global PTMs analysis is that histones are rich in K and R residues, thus trypsin digestion generates short peptides that cannot be detected in standard LC-MS/MS. For this reason, bottom-up approaches usually employ the ArgC protease, which cleaves only at the C-terminus of arginine residues, is usually used for histone digestion in "bottom-up" experiments. When ArgC does not work (i.e. in the case of in-gel digestions), K residues on histones are modified with chemical products, such as acetic or propionic anhydride. Chemically derivatized K residues are not cleaved by trypsin, thus producing "ArgC-like" digestion (Figure 11) [96]. Furthermore, chemical derivatization, enhances the hydrophobicity of short histone peptides, increasing their retention time in standard RP chromatography. When peptides are derivatized with acetic anhydride, since K acetylation is an important hPTM, a deuterated anhydride is employed to introduce "heavy" acetyl groups (D3, Δmass +45.03 Da), which can be distinguished from in vivo acetylations, which are "light" (Δmass +42.01 Da). Another widely-used derivatization that increases hydrophobicity is the addition of phenyl-isocyanate (PIC) to the N-terminus of peptides, which is usually done in combination with propionylation (Pro-PIC protocol) [97].

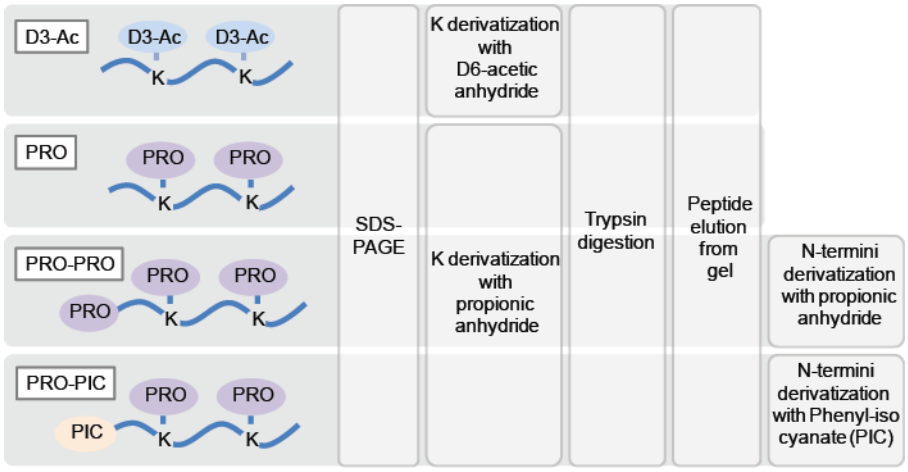


Figure 11. Schematic representation of the different histone digestion strategies. *Adapted from [98].*

2.8.1. Super-SILAC for hPTMs quantitation in clinical samples

The accurate quantitation of hPTMs is crucial not only in basic research but also in clinical applications. In recent years, MS has become the method of choice to identify and quantitate hPTMs, yet the traditional SILAC labelling strategy, which is widely used to profile global PTMs changes, excludes clinical samples as they cannot be metabolically labelled. However, these limitations can be overcome by employing the Super-SILAC strategy, whereby a mixture of cell lines is labelled with heavy amino acids. Heavy-labelled histones from the cell lines can therefore be "spiked in" the unlabelled samples to serve as an internal standard that allows all the samples to be compared with each other (Figure 12) [99].

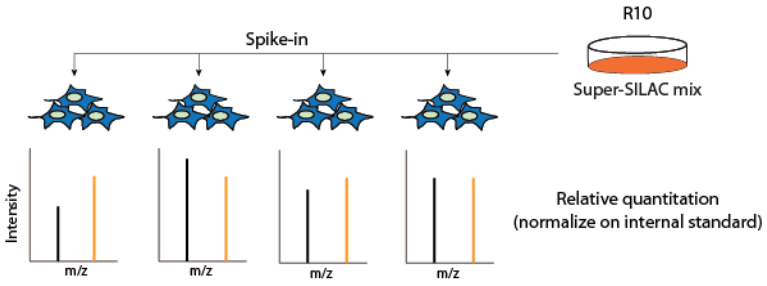


Figure 12. Schematic representation of the Super-SILAC strategy. *Unlabelled experimental samples are quantified relative to a heavy-labelled spike-in standard, allowing the comparison of multiple samples at a time, including samples that cannot be SILAC-labelled, such as clinical samples.*

2.8.2. Unbiased hPTMs analysis by Open Modification Search

A critical step during the analysis of histone MS raw data is the choice of variable modifications to specify during the database search. As pointed out in Chapter 1.4.2, common database search tools produce unreliable results when multiple modifications are searched. In the case of histones, this problem is exacerbated by the presence of several neighbouring residues that can be modified, making the localization of hPTMs challenging [100]. Furthermore, some combinations of PTMs have the same total mass, which leads to the presence of many isobaric peptides that can only be discriminated by few specific MS/MS fragment ions.

In the last two decades, several bioinformatics tools have been developed that allow researchers to consider arbitrarily large lists of variable PTMs without incurring into search space explosion, known as “open modification search” or “blind search” engines. The earliest implementation of an open modification search engine is MS-Alignment [101], which uses dynamic programming to align an experimental spectrum to the theoretical spectrum of an unmodified peptide (Figure 13). In MS-Alignment, all potential matches between experimental and theoretical peaks are represented in a matrix, whose top-left and bottom-right corners represents the N-terminus and C-terminus of the peptide, respectively, and a spectrum interpretation is represented as a path that connects these two termini; if the delta m/z between two peaks equals the mass of an amino acid in the peptide sequence, the peaks are connected with a “diagonal jump”, otherwise they are connected with an “oblique jump”, which introduces a delta m/z (i.e. a PTM) in the peptide sequence. The optimal alignment should introduce a total delta m/z equal to the delta m/z between the mass of the precursor peptide and that of the theoretical peptide, and connect the peaks with the highest total intensity. This strategy allows MS-Alignment to evaluate all modified forms of a peptide at the same time.

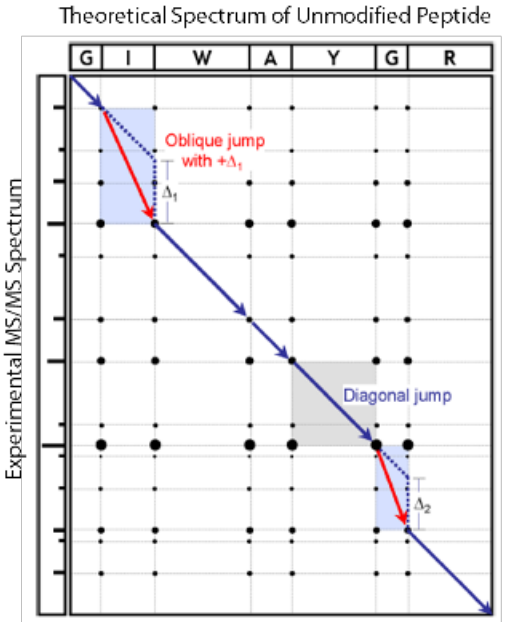


Figure 13. Schematic representation of the MS-Alignment algorithm. *Each dot in the matrix is a possible match between an experimental MS/MS peak and a peptide fragment. "Diagonal jumps" (blue) connect two peaks whose mass difference equals an unmodified amino acid, while "oblique jumps" (red) also introduce the mass shift of a PTM. Adapted from [102].*

Subsequent open search tools sped up the identification process by using sequence tags to restrict the list of candidate sequences and to evaluate and infer the PTMs [103-105]. MODa takes advantage of both sequence tags and spectral alignment [106]. Alternative approaches, such as ModifiComb [107], do not directly sequence peptides, but compare the precursor masses of identified and unidentified peptides and, based on the most frequently observed delta m/z values, produce a list of the putative PTMs present in the sample. Methods based on spectral networks align MS/MS spectra without any reference to a database: instead, a good alignment between two spectra that have different precursor masses would suggest that one is a modified form of the other, with the delta m/z of the precursor matching the m/z of the PTM [108]. Finally, newer open search tools take advantage of machine learning models to predict a peptide retention time and MS/MS fragmentation pattern. Having access to theoretical spectra that accurately represent the real fragmentation pattern peptides, tools such as ionbot [77] and MSFragger [109]are much more accurate than traditional database search engines and achieve high accuracy even when searching spectra against large databases, as in the case of open search.

3. AIM OF THE WORK

Protein PTMs are implicated in almost all cellular processes and represent an important field of study for both basic and clinical research. Therefore, in this thesis, I sought to develop and apply computational methods to address some of the challenges posed by PTMs, in particular, global Arginine methylation and histone PTMs.

In fact, it is now consolidated that Arginine methylation and protein Arginine methyltransferases (PRMTs) are implicated in important biological processes (such as DNA repair, transcription, RNA splicing and translation) and that their dysregulation can give rise to a variety of diseases, including cancer and neurodegeneration. Thus, there is growing interest in researching protein methylation, but its systematic analysis by MS suffers from a high false discovery rate due to the existence of amino acid substitutions and chemical artefacts that are isobaric to this PTM.

Histone PTMs are also known to play a central role in cancer onset by controlling the state of chromatin and the expression of genes. However, research in this field has traditionally focused on acetylation and mono-, di- and tri-methylation of few K residues on histones H3 and H4 for two main reasons: they are the most abundant histone PTMs and the most common computational tools cannot reliably identify combinations of several PTMs from MS data. In spite of numerous studies highlighting the presence of several low-abundance histone PTMs beyond acetylation and methylation, and although computational tools that could theoretically identify any number of PTMs in one run are available, these methods have not been systematically applied to histone MS data yet. Within this context, we were interested in exploring the application of a novel machine learning (ML)-based “open search” software, named ionbot, to complex histone MS data.

With this premise, my first goal was to carry out a re-analysis of all the hmSILAC and SILAC methyl-proteomics data available in our lab to generate a comprehensive database of methylation sites. To achieve this, I significantly improved the hmSEEKER computational tool by training an ML model to identify methyl-peptides with high confidence, without any bias that could be introduced by the manual setting of the parameters.

The updated version of hmSEEKER allowed us to, on the one hand, expand the current annotation of the R-methyl-proteome and, on the other hand, to explore the annotation of less-known methyl-marks such as Histidine methylation, Aspartate methylation, etc.

In parallel, I analysed histone MS data with ionbot in order to broaden the spectrum of PTMs that can be profiled. This unbiased analysis of histone modifications will provide better tools to classify cancer subtypes and predict the evolution of the disease, while possibly unravelling novel epigenetic mechanisms underlying tumour onset and development.

4. MATERIALS AND METHODS

4.1. Cell culture

4.1.1. hmSILAC labelling of samples

For hmSILAC, SK-OV-3, NB4, HeLa and U2OS cells were cultured in ‘‘Light’’ and ‘‘Heavy’’ hmSILAC RPMI media (PAA, custom) supplemented with L-Arginine (Sigma-Aldrich, A6969) L-Lysine (Sigma-Aldrich, L8662), plus either L-[$^{13}CD_3$]-Methionine (M4, heavy, Sigma-Aldrich, 299154) or L-[$^{12}CH_3$]-methionine (M0, light, Sigma-Aldrich, M5308), respectively. The concentration of L-Methionine was 30 mg/L. The hmSILAC media were then supplemented with 10% dialyzed FBS (GIBCO, Life Technologies 26400-044), 1% glutamine, 100 U/ml Penicillin and 100 mg/ml Streptomycin.

4.1.2. SILAC labelling of samples

For SILAC, SK-OV-3, NB4, HeLa and U2OS cells were grown in ‘‘Light’’, ‘‘Medium’’ and ‘‘Heavy’’ SILAC RPMI (Thermo Fisher Scientific, # 89984), supplemented with either L-Arginine, L-Lysine or their medium (L-$^{13}C_6$-Arginine, Sigma-Aldrich; L-D_4-Lysine, Sigma-Aldrich) or heavy (L-$^{13}C_6{}^{15}N_4$-Arginine, Sigma-Aldrich; L-$^{13}C_6{}^{15}N_2$-Lysine, Sigma-Aldrich) isotope-counterparts. Arginine and Lysine were added at a concentration of 84 mg/L and 146 mg/L, respectively. The SILAC media were then supplemented with 10% dialyzed FBS (GIBCO, Life Technologies 26400-044), 1% glutamine, 100 U/ml Penicillin and 100 mg/ml Streptomycin.

4.1.3. Super-SILAC labelling of histone samples

18 samples of breast cancer cell lines, 9 treated with histone deacetylase (HDAC) inhibitor Panobinostat and 9 untreated, were grown in a “light” medium; in parallel, MDA-MB-231, MDA-MB-468, MDA-MB-453 and MDA-MB-361 breast cancer cells lines were grown in SILAC-DMEM (Euroclone) supplemented with 2 mM L-Glutamine, 146 mg/l of L-Lysine (Sigma-Aldrich), 84 mg/l L-$^{13}C_6{}^{15}N_4$-Arginine (Sigma-Aldrich), 10% dialyzed serum (Life Technologies) and penicillin/streptomycin for at least 8 doublings to obtain complete labelling with R10. The Super-SILAC was then spiked into the light samples and histones were subsequently obtained through acidic extraction

4.2. Protein extraction and digestion

For the global analysis of protein methylation, cell pellets were lysed in urea lysis buffer (9 M urea, 20 mM Hepes (pH 8.0)) supplemented with 1x protease and phosphatase inhibitors cocktail (Roche), sonicated and cleared by ultracentrifugation (20,000g for 15 min at 15°C). For in-solution digestion, 200µg of proteins was reduced by adding 4.5 mM dithiothreitol (DTT) (Sigma-Aldrich) for 30 min at 55°C, alkylated with 5.5 mM iodoacetamide (10% (v/v) for 15 min at RT in the dark; Sigma-Aldrich), and digested overnight with sequencing-grade trypsin (1:100 (w/w); Promega) after a four-fold dilution in 25 mM ammonium bicarbonate solution. Protease digestion was terminated by the addition of 1% trifluoroacetic acid (TFA) to adjust pH < 3. Debris was removed by centrifugation for 15 min at 1780g at RT. Peptides were dried with a vacuum concentrator, re-suspended into 300 µL of 0.1% TFA and off-line High pH fractionated by Pierce™ High pH Reversed-Phase Peptide Fractionation Kit (Thermo Fisher Scientific). Eluted fractions were dried with a vacuum concentrator and resuspended in an aqueous 0.1% TFA solution prior to analysis by LC-MS/MS. After lyophilisation, each fraction was dissolved in 250 mL of 1x Immuno-Affinity Purification Buffer (IAP Buffer, #9993, Cell Signaling Technologies) and subjected to R-methyl-peptides enrichment using the anti-pan-methyl-R antibody-conjugated beads (PTMScan Asymmetric Di-Methyl-Arginine Motif [adme-R] Kit #13474; PTMScan Mono-Methyl-Arginine Motif [mme-RG] Kit #12235; PTMScan Symmetric Di-Methyl Arginine Motif [sdme-RG] Kit #13563; Cell Signaling Technologies) following the manufacturer's instruction. After peptides incubation with the antibody-conjugated beads for 2 hours at 4°C, the immunoprecipitates were washed twice in ice-cold IAP Buffer, followed by three washes in water; then, bound methyl peptides were eluted with 2x 50 mL 0.15% TFA. Peptide eluates were desalted on RP C18 StageTip microcolumns, as described previously [110] and subjected to a second round of trypsin digestion prior to LC-MS/MS analysis.

4.2.1. Extraction and protease digestion of hmSILAC-labelled histones

hmSILAC labelled HeLa S3 cells (L and H channels mixed in 1:1 ratio) were resuspended in lysis buffer (10% sucrose, 0.5mM EGTA, 60mM KCl, 15mM NaCl, 15mM HEPES, 0.5mM PMSF, 5 µg/ml Aprotonin, 5 µg/ml Leupeptin, 1mM DTT, 5 mM NaButyrate, 5 mM NaF, 30 µg/ml Spermine, 30 µg/ml Spermidine and 0.5% Triton X-100) and nuclei were separated from cytoplasm by centrifugation on sucrose cushions for 30 min at 3695g at 4°C. Histones

were then extracted through 0.4 M hydrochloric acid for 5 hours at 4°C and dialyzed overnight in CH_3COOH 100 mM Dialysed histones were then lyophilized and either kept at -80 until use or directly resuspended in milliQ water before sample processing prior to MS analysis. To maximize the protein sequence coverage, four different aliquots of 5 μg histones each were in-solution digested overnight using different proteases, such as ArgC, Trypsin, LysargiNase and LysC. Proteolytic peptides were then desalted and concentrated by micro-chromatography onto SCX and C18 Stage Tips micro-column, prior to LC-MS/MS analysis.

4.2.2. Preparation of Super-SILAC histone samples

For each run and sample, 5 μg of histones were mixed with an equal amount of Super-SILAC-labelled histones and separated on a 17% SDS-PAGE gel. For in-gel digestions, a band corresponding to the histone octamer (H3, H4, H2A, H2B) was excised, chemically acylated with D6-acetic anhydride (D3 protocol) or propionic anhydride (PRO-PIC protocols) and in-gel digested with trypsin, obtaining an "ArgC-like" digestion. The extraction of the digested peptides from the gel was performed with acetonitrile 50% and 100%, without formic acid, which would impair the subsequent derivatization steps. In addition, for the PRO-PIC protocol, the samples were concentrated after elution to a volume below 3 μl, diluted to 9 μl with water and derivatized with phenyl isocyanate (PIC). The derivatization with PIC was initially performed as described in [97]. The samples were buffered to pH 8.5 by adding 1 μl of 1 M triethylammonium bicarbonate buffer, 3 μl of a freshly prepared 1% v/v PIC solution in acetonitrile was added (17 mM final concentration), and the mixture was incubated for 60 min at 37 °C. Finally, the samples were acidified by addition 8 μl of 1% trifluoroacetic acid (TFA).

4.3. LC-MS/MS

For the global analysis of protein methylation, peptide mixtures were analysed by online nano-flow liquid chromatography-tandem mass spectrometry using an EASY-nLC 1000 (Thermo Fisher Scientific) connected to a Q Exactive instrument (Thermo Fisher Scientific) through a nano-electrospray ion source. The nano-LC system was operated in one column set-up with a 50cm analytical column (75 μm inner diameter) packed with C18 resin (easySpray PEPMAP RSLC C18 2M 50cm x 75 M, Fisher Scientific) configuration. Solvent A was 0.1% formic acid (FA) and solvent B was 0.1% FA in 80% ACN. Samples were injected in an aqueous 0.1% TFA solution at a flow rate of 500 nL/min. Peptides were separated with a gradient of 5–

40% solvent B over 90 min followed by a gradient of 40–60% for 10 min and 60–80% over 5 min at a flow rate of 250 nL/min in the EASY-nLC 1000 system. The Q-Exactive was operated in the data-dependent acquisition (DDA) mode to automatically switch between full-scan MS and MS/MS acquisition. Survey full-scan MS spectra (from m/z 300-1150) were analysed in the Orbitrap detector with resolution R=35,000 at m/z 400. The ten most intense peptide ions with charge states ≥2 were sequentially isolated to a target value of 3e6 and fragmented by Higher Energy Collision Dissociation (HCD) with a normalized collision energy setting of 25%. The maximum allowed ion accumulation times were 20 ms for full scans and 50 ms for MS/MS and the target value for MS/MS was set to 10^6. The dynamic exclusion time was set to 20s.

Histone-derived samples were analysed by LC-MS/MS on a 25 cm reverse phase C18 column (inner diameter 75 μm) connected to a Q Exactive orbitrap instrument. Solvent A was 0.1% formic acid (FA) and solvent B was 0.1% FA in 80% ACN. Samples were injected in an aqueous 0.1% TFA solution at a flow rate of 500 nL/min. After sample loading in the column, a gradient of 0–40% solvent B over 100 min, followed by a gradient of 40–60% solvent B in 5 min and 60–95% solvent B over 5 min at a flow rate of 250 nL/min, for peptide elution. The Q-Exactive was operated in the data-dependent mode (DDA) to automatically switch between full-scan MS and MS/MS acquisition. Survey full-scan MS spectra (from m/z 300-1350) were analysed in the Orbitrap detector with resolution R=70,000 at m/z 200. The ten most intense peptide ions with charge states ≥2 were sequentially isolated to a target value of 3e6 and fragmented by Higher Energy Collision Dissociation (HCD) with a resolution of R = 17,500.

4.4. MS data analysis

4.4.1. hmSILAC methyl-peptides identification with MaxQuant

MS raw data were analysed using the integrated MaxQuant software v1.6.2.10, using the Andromeda search engine [75,86]. In the global parameters sections, the estimated FDR of all peptide identifications was set to a maximum of 1%. The main search was performed with a mass tolerance of 4.5 ppm. A maximum of 3 missed cleavages was permitted, and the minimum peptide length was fixed at 6 amino acids. The June 2020 version of the Uniprot reference proteome (Proteome ID: UP000005640) was used for peptide identification. In the

group-specific parameters section, we indicated different parameters, according to the subset of raw data to be analysed.

Each hmSILAC MS data file was analysed twice to identify Heavy and Light methyl-peptides separately. In the “Light” analysis, mono-methylation of K/R (+14.02 Da), di-methylation of K/R (+28.03 Da), tri-methylation of K (+42.04 Da) and oxidation of M (+15.99 Da) were specified as variable modifications, while Carbamidomethylation of Cysteine (+57.02 Da) was indicated as fixed modification. In the “Heavy” analysis, heavy mono-methylation of K/R (+18.03 Da), heavy di-methylation of K/R (+36.07 Da), heavy tri-methylation of K (+54.11 Da) and oxidation of M were specified as variable modifications, while Carbamydomethylation of Cysteine and isotope-labelled Methionine (M4; +4.022 Da) were indicated as fixed modifications. Enzyme specificity was set to either Trypsin/P or LysargiNase. Upon re-analysis of the samples to identify non-canonical methyl-sites, the same parameters were indicated but mono-methylation (heavy or light) was allowed on K, R, D, E, H, Q, N, S and T.

For the analysis of non-canonical methylation sites on histones, MS raw data were also analysed twice. Mono-methylation (light or heavy) was allowed on K, R, D, E, H, Q, N, S and T and we included K acetylation (+42.01 Da) as a variable modification to address the fact that this modification is very abundant and likely coexists with methylation on histones. Enzyme specificity was set to Trypsin/P, ArgC, LysC or LysargiNase. Finally, to reduce search complexity, a database containing only human histone sequences was used.

4.4.2. Validation of methyl-peptides with hmSEEKER 2.0

Identification of light and heavy methyl-peptides doublets was carried out with hmSEEKER [111]. Methyl-peptides identified by MaxQuant were filtered to only retain peptides with Andromeda Score >25, Delta Score >12 and Localization Probability of the methylation sites >0.75. hmSILAC doublets were reconstructed with hmSEEKER. Initially, hmSILAC v1.0 called a methyl-peptide doublet when two peaks had the same charge, $|ME| < 2$ ppm, $|dRT| < 0.5$ min and $|LogRatio| < 1$. The Machine Learning model developed in this study within hmSEEKER v2.0 was trained using Python package Scikit-learn v0.23.1, as described in chapter 5.1.1.

4.4.3. SILAC methyl-peptides identification

MS raw data were analysed using the integrated MaxQuant software v1.3.0.5 or v1.5.2.8. Global parameters were the same the same as the hmSILAC analysis. In the group-specific parameters section, we indicated K8+R10 and/or K4+R6 as SILAC labels; N-terminal acetylation (+42.01 Da), M oxidation, mono-methyl-K/R and di-methyl-K/R as variable modifications; Carbamidomethylation of C as fixed modification. Enzyme specificity was set to Trypsin/P.

4.4.4. Quantitative analysis of SILAC methyl-proteomics data

The MaxQuant evidence.txt file was first filtered: potential contaminants and reverse sequences were removed; methyl-peptides were required to have an Andromeda score >25 and individual modifications were required to have a Localization Probability >0.75. For the methyl-peptides quantified more than once, the median SILAC ratio was calculated. Then, methyl-peptide SILAC ratios were normalized on the respective protein SILAC ratios calculated using unmodified peptides in the "input" experiments. To determine which methyl-peptides were significantly up- or down-regulated by each stimulus, the mean (μ) and standard deviation (σ) of the distribution of the unmodified peptide SILAC ratios (which are expected not to change) was calculated and a $\mu \pm 2\sigma$ cut-off was applied to the distributions of the modified peptides of the respective experiment.

4.4.5. Detection of SDMA- and ADMA-specific neutral losses

To search for symmetric and asymmetric arginine methylations within the hmSILAC peptide sequences, we configured MaxQuant by adding the neutral losses of light monomethylamine/dimethylamine [NH_2CH_3, 31.04 Da; $NH(CH_3)_2$, 45.06 Da] (Figure 14A) and heavy monomethylamine/dimethylamine [$NH_2(^{13}CD_3)$, 35.06 Da; $NH(^{13}CD_3)_2$, 53.10 Da] (Figure 14B) to the variable modifications di-methyl-K/R and di-methyl4-K/R, respectively.

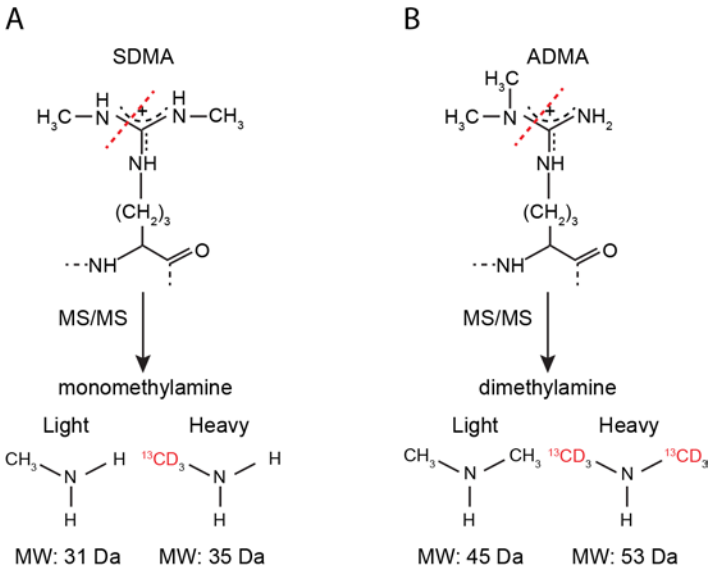


Figure 14. Schematic representation of how neutral losses arise from MS/MS fragmentation of di-methyl-R. A) *Fragmentation of the guanidino group of MMA and SDMA residues releases monomethylammine.* ***B)*** *Fragmentation of the guanidino group of ADMA residues releases dimethylammine. Adapted from [31].*

To search for symmetric and asymmetric arginine methylations within the SILAC dataset, we configured MaxQuant by adding the neutral losses of light monomethylamine/dimethylamine [NH_2CH_3, 31.04 Da; $NH(CH_3)_2$, 45.05 Da] and loss of heavy monomethylamine/ dimethylamine [$^{15}NH_2CH_3$, 32.04 Da; $^{15}NH(CH_3)_2$, 46.05 Da] to the variable modification di-methyl-R.

MaxQuant viewer was used for manual inspection of the MS/MS spectra of the identified dimethylated peptides to check the presence.

4.4.6. Unrestrictive analysis of histone MS data with ionbot

MS raw data were analyzed with ionbot [77], a novel search engine developed by the group of Lennart Martens at Ghent University. The “Unexpected modification search” and “Mutation search” options of IonBot were activated and peptides were searched against a database of human proteins. An in-house Perl script was then used to remove results decoy peptides and peptides with q-value >0.01 from ionbot results and map identified PTMs onto histone proteins. Identified amino acid substitutions were submitted to Mechismo to predict their functional implications.

4.5. Functional and structural analysis of the ProMetheusDB

Motif analysis was done using pLogo [112]. Sequence windows centred on regulated (or unchanging) R-methyl-sites were submitted as foreground; sequence windows centred on all identified R-methyl-sites (including non-quantified ones) were submitted as background; the resulting position weight matrices were then downloaded and visualized with the Logomaker Python package [113]. Functional enrichment analysis was performed using the "gprofiler2" R package [114]; Gene Ontology terms, Reactome [115] and KEGG pathways, and CORUM complexes were used as data sources; only the most significant, non-redundant terms are reported in the figures. Overlap of the annotated R-methyl-sites with InterPro domains [116] was performed with an in-house Python script; the InterPro database was filtered to include only regions classified as "Domain" or "Homologous superfamily". Mapping of modification sites on interaction surfaces was performed with the Mechismo web application [117], with a stringency threshold set to "medium". The database of currently annotated phosphorylation sites was downloaded on 05-28-19 from Phosphosite Plus [118]. Protein:protein functional interaction networks were generated within Cytoscape [119] using the Reactome plugin; protein:protein physical interactions were downloaded from the IMEX database [120]; network analysis was performed with Cytoscape and the Python package Pyntacle [121]. Fisher's exact tests were performed with the Scipy package in Python. Bar plots and box plots were generated with the Seaborn package in Python, network displays were generated in Cytoscape and the UpSets plot representation of methylation sites on proteins was generated with R.

4.6. Protein immunoprecipitation (IP) of NONO and Western Blot (WB) analysis of its R methylation state and co-IP of PSPC1

IP of NONO was performed starting from 1mg of HeLa whole-cell extract (WCE). Briefly, 30e6 HeLa cells were harvested, washed twice with cold PBS and re-suspended in 2 volumes of RIPA Buffer (10 mM Tris pH 8, 150 nM NaCl, 0.1 % SDS, 1 % Triton, 1mM EDTA, 0.1% Na-Deoxycholate, 1mM PMSF, 1mM DTT and 1x Protease and Phosphatase Inhibitors cocktail (Roche), supplemented with 10 kU of Benzonase (Merck Life Science)). The suspension was rotated on a wheel for 45 min at RT (vortex every 10 min), centrifuged at 12.000 g for 1h at 4°C and the supernatant was transferred into a new Eppendorf tube. Proteins were quantified by BCA colourimetric assay (Pierce BCA Protein assay kit) and 1 mg of WCE

was used for the IP, 8% of which was saved as Input (80 µg to be divided into 4 SDS-PAGE gels). The WCE used for the IP was rotated at 4°C overnight with 4µg of anti-NONO/p54 (sc-376865 Santa Cruz). G-protein-coupled magnetic beads (Dynabeads, Thermo Fisher Scientific) were saturated with a blocking solution (0.5% BSA) and rotated at 4°C overnight on a wheel. The following day, the beads were added to the lysate in 1: 100 proportion with the primary antibodies and incubated for 3 hours at 4°C on the wheel; the captured complexes were washed 4 times with the RIPA Buffer and then incubated 10 min at 95° with LSD Sample Buffer (2X) supplemented with 100 mM DTT to elute the immunoprecipitated proteins. Equal protein amounts were separated by SDS-PAGE electrophoresis (Nupage Novex 4-12% Bis-tris Gel 1.5 mm, Thermo Fisher) and transferred on Transfer membrane (Immobilon-P, Merck Millipore) by wet-transfer method. Membrane blocking was performed with 10% BSA/TBS 0.1% Tween-20 for 1h at RT and followed by overnight incubation with the selected primary antibodies and subsequent incubation with the HRP-conjugated secondary antibodies (Cell Signaling Technology) for 1h at RT. Proteins were detected by ECL (Bio-Rad). The following primary antibodies were used:

- anti-NONO (SC-376865, 1:500) was purchased from Santa Cruz;
- anti-ADMA (ASYM24 07-414, 1:1000) and anti-SDMA (SYM10 07-412, 1:2000) were purchased from Millipore;
- anti-MMA (D5A12; 1:1000) was purchased from Cell Signaling Technology
- anti-PSPC1 (A302-461, 1:5000) were purchased from Bethyl Laboratories;
- anti-GAPDH (Ab9484, 1:3000) was purchased from Abcam.

Quantification of the signal intensity for each band was performed by Fiji software [122] and signal intensity was normalized at 4 different levels:

- quantification of NONO in the Input was normalized on GAPDH (as loading control);
- quantification of NONO in its IP was normalized on the previous normalized input for each condition;
- R methylation (MMA, ADMA or SDMA) and PSPC1 signals were normalized on the amount of normalized NONO in the IP;
- signal intensity in MS023 condition was normalized on DMSO (untreated).

5. RESULTS

5.1. Annotation of hmSILAC-validated methyl-proteome

5.1.1. Implementation of Machine learning into hmSEEKER 2.0

Heavy methyl SEEKER (hmSEEKER) is a computational tool we first published in 2019 with the aim of providing the research community with a new method to identify methyl-peptides from hmSILAC MS data [111]. hmSEEKER distinguished true hmSILAC doublets from false positives based on their retention time difference (dRT), their log-transformed intensity ratio (LogRratio) and the difference between expected and observed mass shift (Mass Error, ME). In the initial version of hmSEEKER, the cut-offs for these parameters were provided by the users and if no cut-off values were specified, the tool would perform the analysis with its default cut-offs. To calculate optimal default cut-offs, we generated a dataset of M-containing peptides that did not bear any methylation, labelled “True positives”; in fact, the labelling with $^{13}CD_3$-M causes M-containing peptides to generate hmSILAC doublets that have the same properties as those generated by methyl-peptides (Figure 15A). To model “True Negatives”, we created a second dataset of unmodified peptides that did not contain M and randomly assigned methylations to them, to mimic an erroneous identification by MaxQuant (Figure 15B). The datasets were then analyzed and hmSEEKER parameters were manually adjusted to maximize the number of “True peptides” recovered while minimizing the number of “True negative” identifications [111].

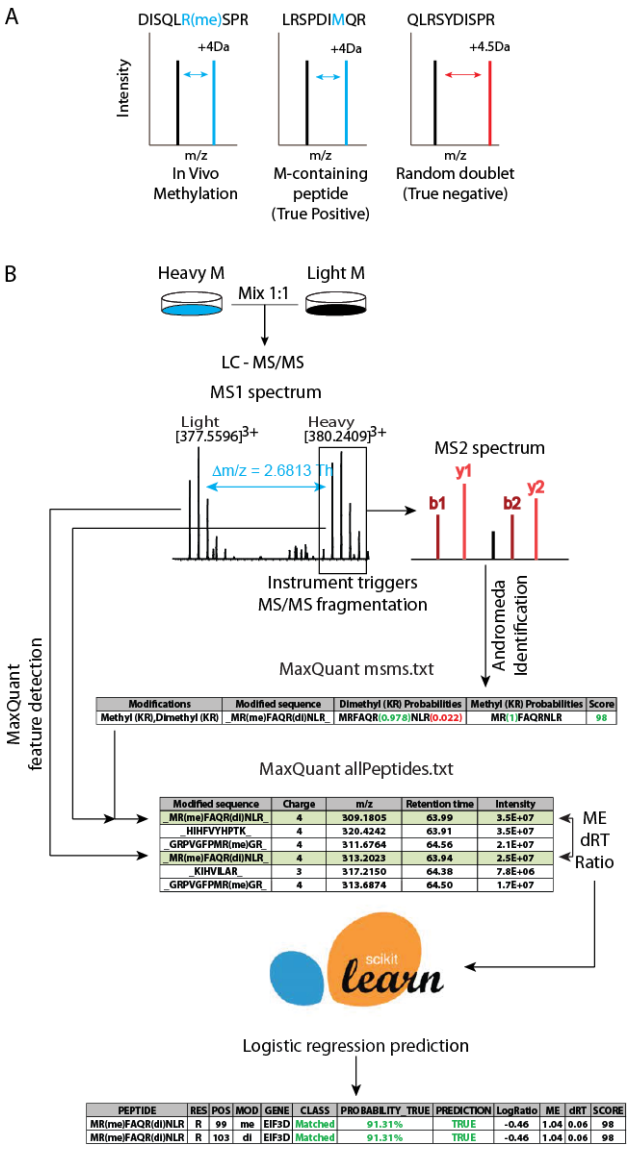


Figure 15. Schematic representation of hmSEEKER workflow. A) *Schematic representation of true positive and true negative doublets. Our rationale for optimizing hmSEEKER parameters, which was later also deployed for training the machine learning model, was that M-containing peptides produce hmSILAC doublets that are indistinguishable from those generated by methyl-peptides.* ***B)*** *Schematic representation of the hmSEEKER workflow upon metabolic labelling with stable-isotope encoded Methionine (M). Cells grown in "light" (M0) and "heavy" (M4) media are mixed 1:1, then proteins are extracted, digested and analyzed by LC-MS/MS. MaxQuant detects MS1 peaks in the MS raw data. Peaks with an associated MS2 spectrum are processed by the database search engine Andromeda to obtain peptide identification. The hmSEEKER software reads MaxQuant peptide identifications and, for each methyl-peptide that passes the quality filtering, finds its corresponding MS1 peak, then*

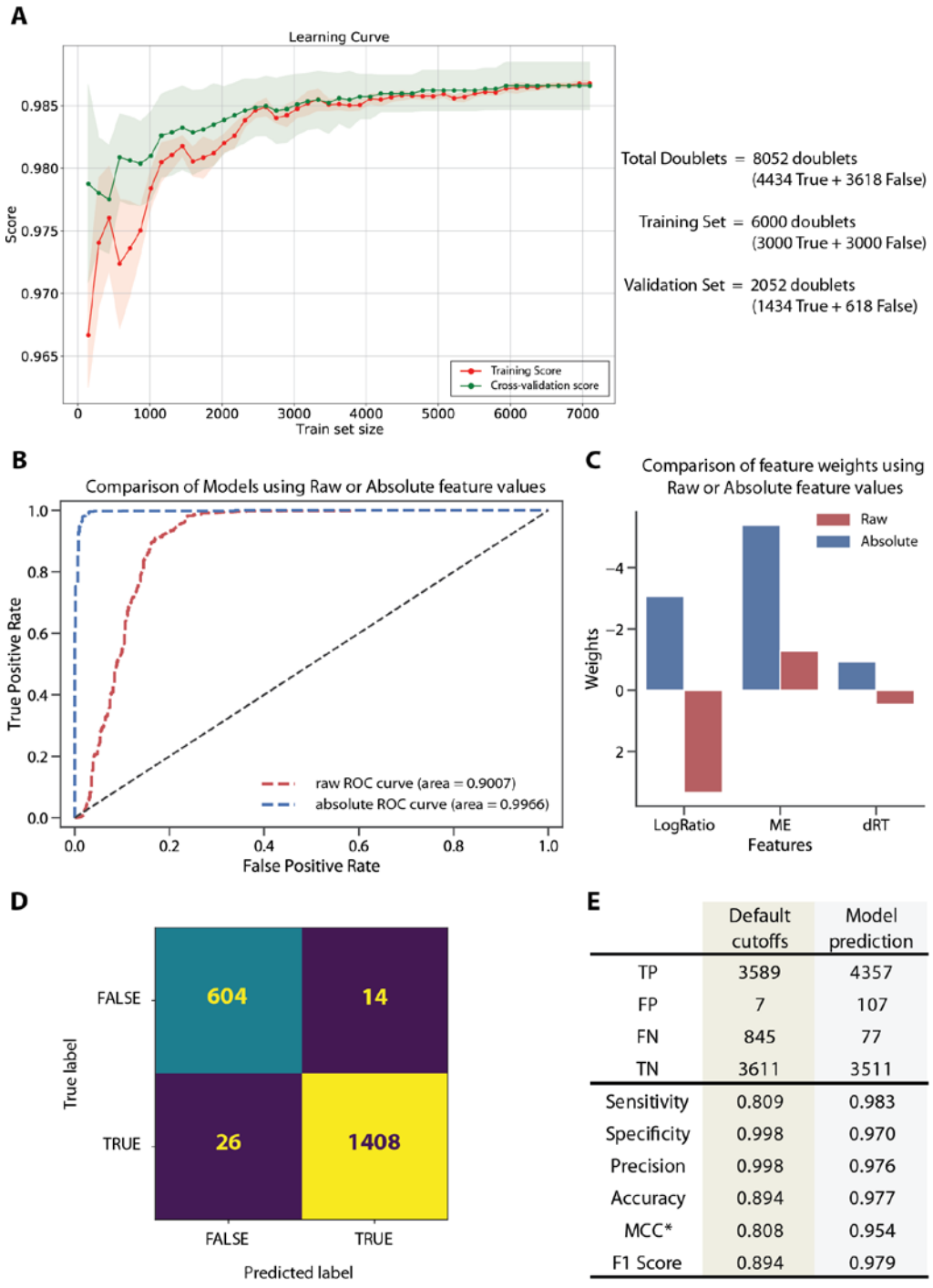
searches its heavy/light counterpart. A peak doublet is defined by the difference in the retention time, the intensity ratio and the deviation between expected and observed m/z delta; these three parameters are used to predict if the peak pair is a real hmSILAC doublet or a false positive.

However, the default parameters we defined were very conservative and while they resulted in a low FDR, they also showed a suboptimal sensitivity. To address this point and improve the performance of hmSEEKER, we trained a machine-learning (ML) logistic regression model to discriminate between putative true and putative false hmSILAC doublets without relying on user-defined criteria. This model allowed us to increase the number of hmSILAC-validated methyl-peptides in our dataset without compromising the FDR and provides more rigorous criteria for the identification of hmSILAC doublets compared to the empirical cut-offs. To obtain a dataset for training the model, we followed the same strategy employed previously in Massignani et al., i.e. using M-containing peptides as "mock methyl-peptides" and the rest of the unmodified peptides as "negative control". The "True positive" and "True negative" peptides were processed with hmSEEKER using the following cut-offs: |ME|< 100 ppm, |dRT| < 25 and |LogRatio| < 25. These cut-offs allowed us to retrieve as many putative doublets as possible, to have enough data for the training of the model. Doublets generated by a "mock methyl-peptide" were then filtered to only include the so-called "Matched" doublets (i.e. those where both heavy and light peptide are identified, which are the most confident) and labelled as "True" (n=4434); doublets generated by a "true negative" peptide were labelled "False" (n=3618). In total, the dataset used to train the predictor consisted of 8052 doublets.

The independent variables (features) used within the Logistic Regression model were the ME, dRT and LogRatio; the dependent variable was whether the doublet was a true hmSILAC doublet ("True") or a random peak pair ("False"). After plotting the model learning curve to estimate the optimal training set size (Figure 16A) the model was trained on 6000 doublets (3000 True + 3000 False) using stratified five-fold cross-validation. During the training process, to reduce the impact of potential outliers, we applied a quantile transformation to the features, which were then normalized based on their median and IQR values. Moreover, we observed that taking the absolute values of the features improved the performance of the model, allowing it to reach an area under the ROC curve >0.99 (Figure 16B). Inspection of the logistic regression coefficients revealed that ME was the most important feature in

distinguishing true and false doublets as it had the largest absolute weight (-5.393), followed by LogRatio (weight: -3.067), whereas dRT (weight: -0.927) appeared to be the least important (Figure 16C). Upon validation of the model on the remaining 2052 doublets, we found that only 40 (<2%) doublets were incorrectly labelled by the logistic regression (Figure 16D).

We then tested whether the inclusion of the ML model improved hmSEEKER v1.0 performance by comparing it to the default cut-offs defined in [111]: the entire set of 8052 doublets used for the training of the model was reanalysed twice, first with the default cut-offs (i.e. |ME| < 2 ppm, |dRT| < 0.5 min, |LogRatio| < 1) and then by using the model predictions. The logistic regression showed an increase in sensitivity and accuracy and a negligible reduction in specificity and precision; the metrics F1 score and Matthew's Correlation Coefficient (MCC) were also improved upon applying the model (Figure 16E). Overall, these results prove our initial suspicion that the initial cut-offs were very conservative.



	Default cutoffs	Model prediction
TP	3589	4357
FP	7	107
FN	845	77
TN	3611	3511
Sensitivity	0.809	0.983
Specificity	0.998	0.970
Precision	0.998	0.976
Accuracy	0.894	0.977
MCC*	0.808	0.954
F1 Score	0.894	0.979

Figure 16. Training and evaluation of the Machine Learning model within hmSEEKER v2.0. ***A)*** *The optimal training set size was determined by plotting the model learning curve, which reaches a plateau when the training set size is ~6000.* ***B)*** *Receiving Operator Characteristics (ROC) curves obtained by testing the models trained by using either the raw or absolute value of the features.* ***C)*** *Representation of the ML model weights with (blue) and without (red) taking the absolute values of the features.* ***D)*** *Confusion matrix generated by*

applying the model to the validation set. ***E)*** *Comparison of the performance of hmSEEKER with (New ML Model) and without (Old Cut-offs) the ML predictor (*MCC = Matthews correlation coefficient).*

5.1.2. Re-annotation of the human methyl-proteome with hmSEEKER v2.0

We first applied hmSEEKER v1.0 as described in [111] to our hmSILAC data. Briefly, the data consisted of samples from four different cell lines (HeLa, SK-OV-3, NB4 and U2OS) which were subjected to different biochemical pipelines before LC-MS/MS, such as immuno-enrichment of methyl-peptides with CST PTMScan kits or immuno-enrichment of proteins with antibodies against methyl-R, methyl-K or Large Drosha Complex (LDC) components (Figure 17A, **Table 1**). These enrichment methods were combined with separation techniques such as polyacrylamide gel electrophoresis, isoelectric focusing and HpH-RP liquid chromatography to reduce sample complexity and further boost the identification of methyl-peptides). Samples were then acquired on a Q Exactive orbitrap instrument and the resulting MS raw data were processed with MaxQuant to identify R- and K-methyl-sites.

The high-confidence methyl-proteome obtained from hmSEEKER v1.0 contained 2174 methyl-peptides, mapping to 490 different proteins, which carried a total of 1324 methyl-sites and 1735 methylation events (Figure 17B, left).

By re-analysing the hmSILAC dataset with hmSEEKER v2.0, we annotated 2688 methyl-peptides mapping on 703 proteins and 2191 methylation events distributed on 1750 methyl-sites. Overall, we were able to identify 23% more methyl-peptides, 32% more methyl-sites and 26% more methylation events (Figure 17B, right).

A

Summary of hmSILAC experiments

Cell line(s)	Enrichment Antibody specif city		Fraction(s)
HeLa			Nucleus, cytosol
HeLa	Protein IP	MMA, ADMA, SDMA	Nucleus, cytosol
HeLa	Protein IP	Kme1, Kme2, Kme3	Nucleus, cytosol
HeLa	Protein IP	LDC components	Whole cell extract
HeLa, NB4, SK-OV-3, U2OS	Peptide IP	MMA, ADMA, SDMA	Whole cell extract

B

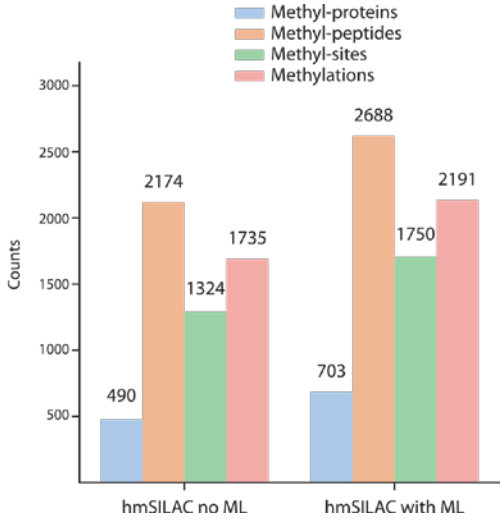


Figure 17. Annotation of the human methyl-proteome with hmSEEKER 2.0. A) *Summary of hmSILAC experiments analyzed.* ***B)*** *Composition of the high-confidence hmSILAC R-methyl-proteome before (left) and after (right) the implementation of the machine learning model.*

5.2. Analysis of Dynamic SILAC

To understand the role of methylation in cancer cells, the analysis of the hmSILAC data was paralleled by SILAC-based methyl-proteomics experiments, which were performed by my colleagues in the Bonaldi group **(Table 2).**

5.2.1. Relocalization of PRMT1 to chromatin upon genotoxic stress

To investigate the role of R methylation in the response to cisplatin (CDDP, a crosslinking agent that induces replicative stress) in ovarian cancer cells, we carried out three independent R-methyl-proteomics experiments, each in two replicates (forward and reverse). By combining SILAC labelling, immunoenrichment with anti-pan-MMA and ADMA antibodies, High-pH RP fractionation and MS-based analysis (Figure 18A), we reproducibly quantified 411 R-methyl-peptides, mapping to 157 proteins and bearing 413 methylation events distributed on 354 R-sites. The analysis of the six paired CDDP/untreated (CDDP/UT) experiments allowed defining statistically significant changes in R methylation induced by the drug, with 58 and 18 R-methyl-peptides being significantly down- and up-regulated, respectively (Figure 18B). Motif analysis performed on these regulated peptides identified an RG/RGG motif that is known to be preferentially targeted by PRMTs, especially PRMT1 (Figure 18C). Moreover,

we found that most of the down-regulated methyl-peptides belong to stress granule proteins, such as G3BP1, hnRNPA1 and KHDRBS1.

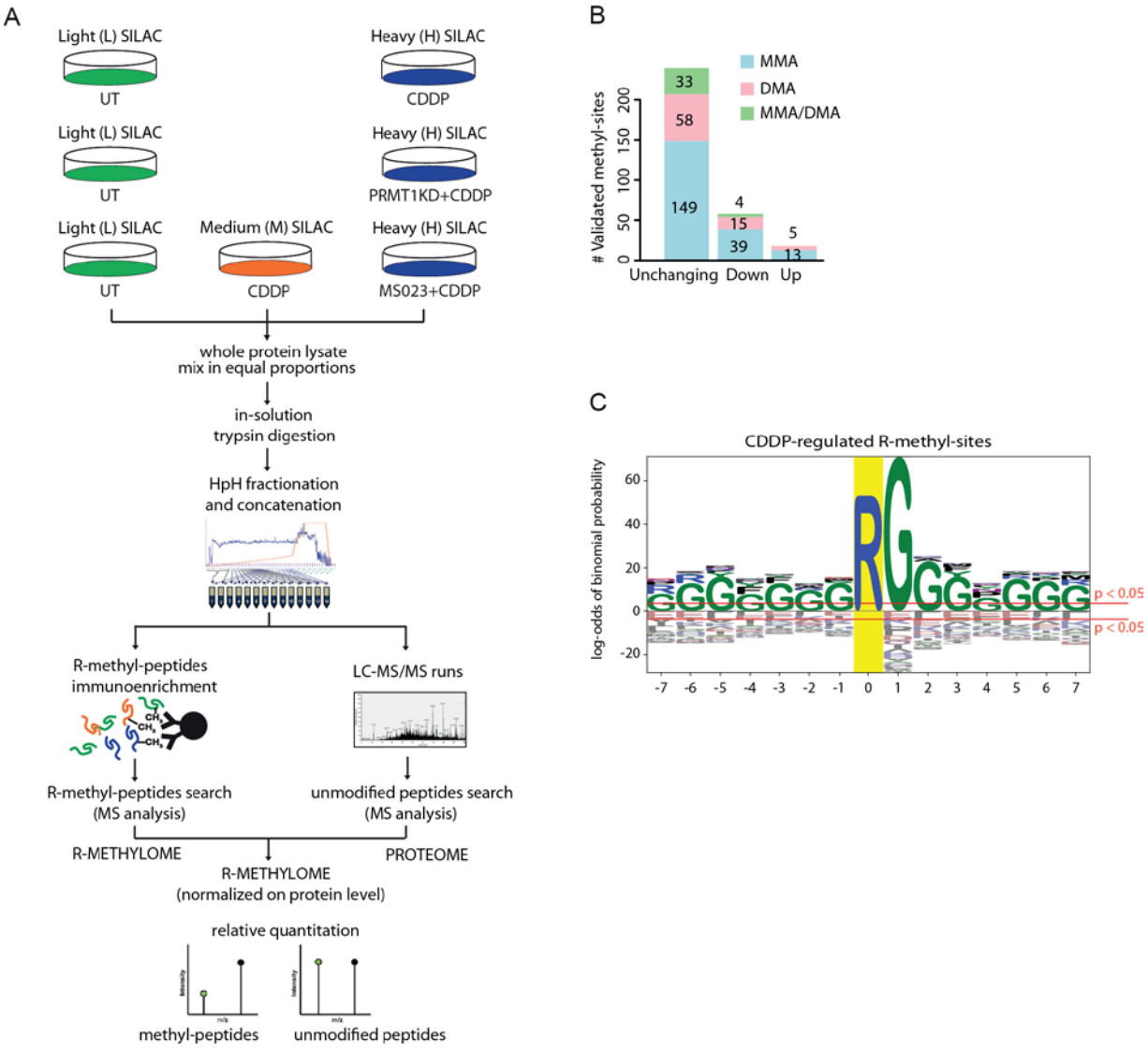


Figure 18. Methyl-proteome profiling of ovarian cancer cells upon CDDP treatment. A) *Schematic representation of the experimental setup of SILAC labelling coupled with affinity enrichment of methyl-peptides, followed by MS analysis with MaxQuant software for methyl-peptides identification and quantification.* ***B)*** *Bar graph showing the number of SILAC-quantified methyl-sites grouped based on the modification level and the response to CDDP. DMA, methyl-sites exclusively di-methylated; MMA, methyl-sites exclusively mono-methylated; MMA/DMA, methyl-sites found both mono- and di-methylated.* ***C)*** *Logo analysis of the methyl-sites regulated by CDDP, performed using as background random sequences from the human proteome. Adapted from [9].*

To address the PRMT1 dependency of the changes induced by CDDP, I compared the regulation of R-methyl-peptides elicited by CDDP in control and PRMT1-depleted cells. Unsupervised clustering of the peptides SILAC ratios in the different conditions (CDDP/UT, CDDP+KD/UT, CDDP+KD/CDDP) identified five methyl-peptides clusters corresponding to different regulation patterns (Figure 19): Cluster 1 includes peptides that are unresponsive in all conditions; Cluster 2 consists of methyl-peptides down-regulated by CDDP irrespective of PRMT1 KD; vice versa, cluster 3 consists of methyl-peptides down-regulated by PRMT1 KD irrespective of CDDP; Cluster 4 includes peptides that are down-regulated upon CDDP treatment but up-regulated by PRMT1 depletion, probably through substrates scavenging by other PRMTs; finally, Cluster 5 comprises peptides up-regulated by CDDP and reduced in PRMT1-depleted cells, a behaviour that identifies them as CDDP-responding specific targets of PRMT1. Our hypothesis that peptides in clusters 3 and 5 are bona fide targets of PRMT1 is corroborated by the fact that, while mono-methylated peptides were equally distributed in the five clusters, di-methylated peptides were mostly found in clusters 3 and 5; moreover, visual inspection of the MS/MS spectra of these peptides demonstrated that 70% of them carried the ADMA methylation mark.

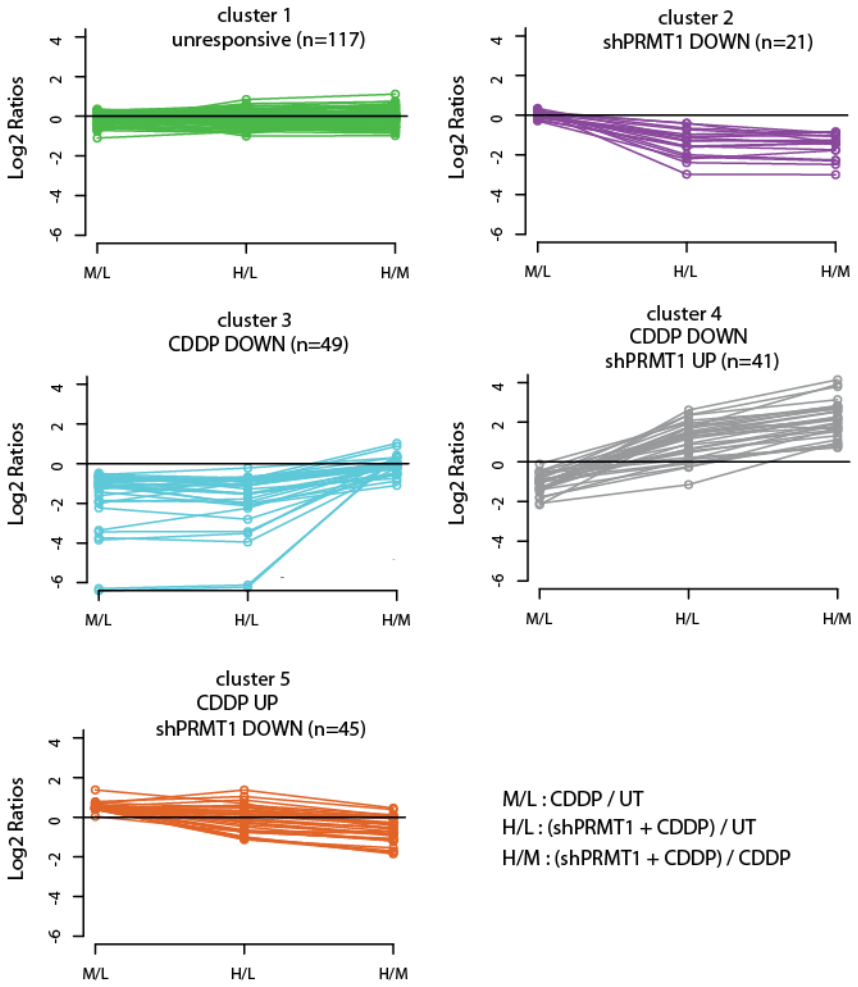


Figure 19. PRMT1 dependency of CDDP-induced methyl-proteome changes. *Unsupervised clustering analysis of the SILAC ratios of methyl peptides in response to CDDP and PRMT1 knock-down reveals five major patterns of changes. Only methyl-peptides reproducibly identified in the forward and reverse experiments are shown. Adapted from [9].*

Overall, this analysis showed that PRMT1-mediated protein methylation is involved in the cellular response observed upon CDDP treatment. Subsequent experiments revealed that the resistance of cancer cells to genotoxic stress caused by the CDDP treatment is indeed mediated by PRMT1 relocalization from the cytoplasm to chromatin. Briefly, phosphorylated PRMT1 is recruited to chromatin at sites of DNA damage, where it asymmetrically di-methylated histone H4 R3; the deposition of the H4R3me2a mark then triggers the expression of Senescence-Associated Secretory Phenotype (SASP) genes, which cause the arrest of the cell cycle and prevent apoptosis, eventually leading to cancer cells becoming resistant to chemotherapy. Hence, targeting PRMT1 by small molecule inhibitors may overcome SASP

genes activation and increase the effectiveness of CDDP-based cancer treatment. These results were published in [9].

5.2.2. Profiling of PRMT5 targets to uncover PRMT5 sequence specificity

As the most active type II PRMT, PRMT5 has emerged as an attractive drug target in the last few years, and intense research has been devoted to designing selective and potent small-molecule inhibitors targeting its methyltransferase activity. However, in-depth research on PRMT5 substrates to understand the consequences of its inhibition on the cellular methyl-proteome was lagging. To fill this gap, we profiled PRMT5 substrates in a SILAC setup, using pan-methyl antibodies to enrich peptides carrying MMA and SDMA from total extracts of HeLa cells, treated with either the selective PRMT5 inhibitor GSK591 or its inactive structural analogue SGC20969.

The experiment was carried out in two biological replicates, in forward and reverse experimental setups, in which SILAC labels were swapped among the two conditions. Upon MS acquisition and processing of the data through the MaxQuant algorithm, we robustly quantified in both biological replicates 686 R-methyl peptides, with 507 R-methyl-sites including 470 mono-methylations and 115 di-methylations, with 92 R that were identified as both mono- and di-methylated.

The analysis of normalized methyl-peptide SILAC ratios revealed that the inhibition of the major type II PRMT had an overall mild effect on the R-methyl-proteome with only 86 (12.9%) methyl-sites down-regulated and 56 (8.4%) up-regulated in both forward and reverse replicates (Figure 20A). The di-methylations were particularly enriched in the subset of peptides down-regulated by GSK591 (27.9%) compared to the up-regulated (21.4%) or the unchanging (20.7%) subsets, in line with the expected decrease of DMA following the inhibition of a type II PRMT. We also observed an increase of MMA on various sites paralleling a reduction in DMA at the same sites (Figure 20B). This suggests that the paradoxical increase of MMA after PRMT5 inhibition could actually result from the partial loss of SDMA at PRMT5 target sites.

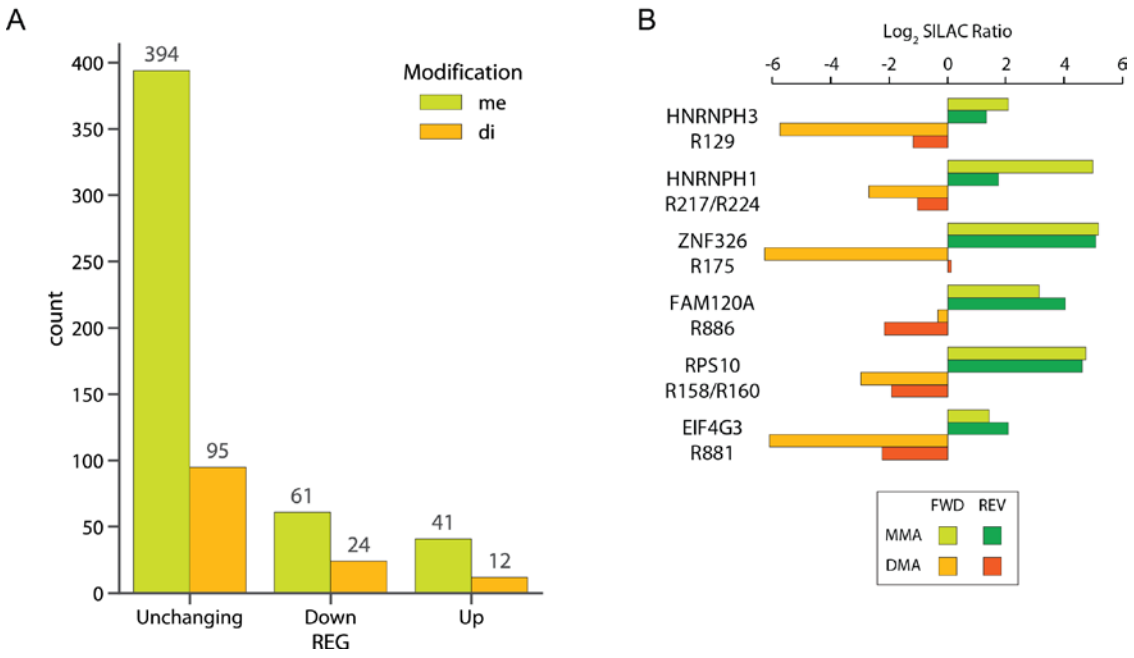


Figure 20. Profiling of methylation changes in HeLa cells treated with PRMT5 inhibitor. A) *Histogram indicating the number of R-sites carrying mono- and di-methylation within each group of sites differentially responding to GSK591.* ***B)*** *Bar graph displaying the opposite trend of response to GSK591 between mono- and di-methylation for the peptides carrying the indicated methylated R sites, as quantified in both forward and reverse experiments and expressed as log2-transformed SILAC ratios (PRMT5i/control). Adapted from [31].*

To establish whether the methyl-sites changing upon GSK591 treatment display a specific sequence motif, we analyzed them with pLogo using the whole R-methyl-sites dataset as background. The motif analysis revealed that the regulated methyl peptides show significant enrichment for Glycine (G) residues at positions −1 and +1 with respect to the modified R (Figure 19, top). To ensure the analysis was not biased by the antibodies used to enrich the peptides, we also analysed regulated sites identified in the Input samples before the immunoenrichment: we found that G was still enriched at position −1 (Figure 19, bottom). This step confirmed that the GR/GRG motif was genuinely associated with the activity of PRMT5.

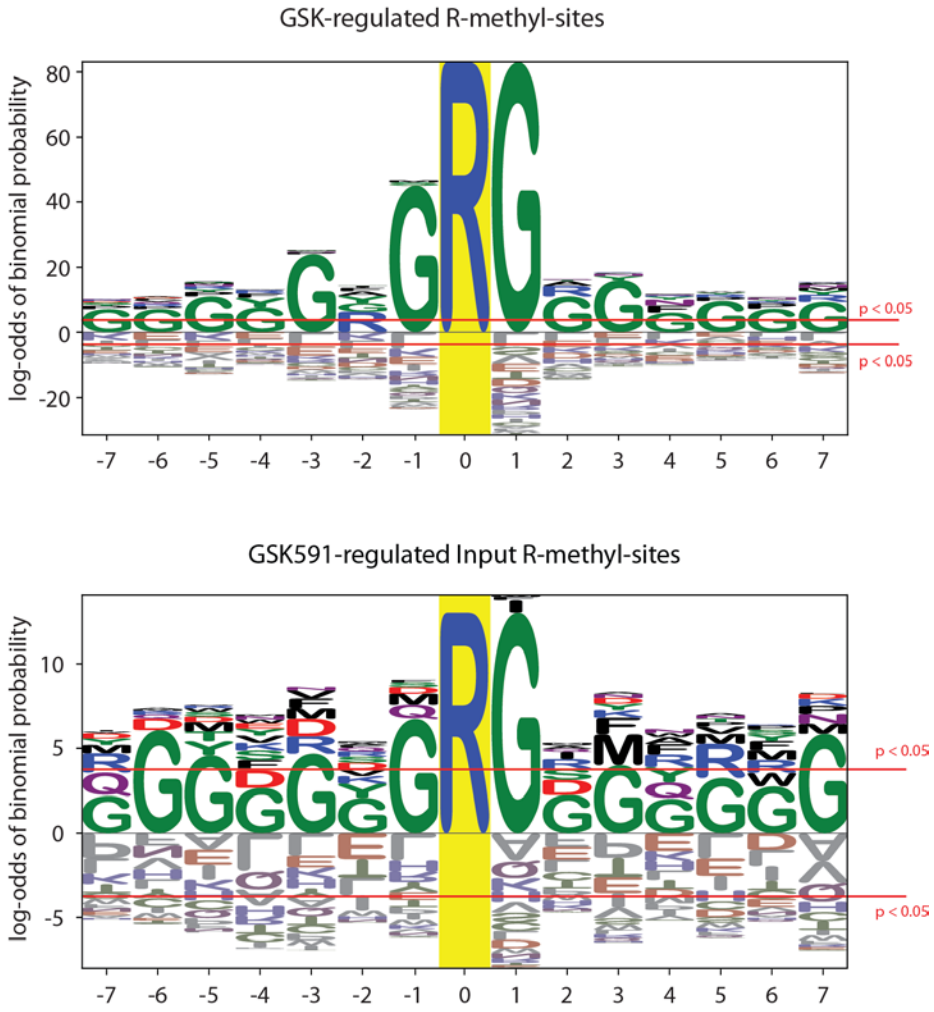


Figure 21. Motif analysis performed on GSK591-regulated methyl-sites. *The analysis shows the overrepresentation of glycine residues at positions −1 and + 1 with respect to the methylated R in the peptides regulated by GSK591 (i.e. PRMT5 targets).*

Several proteins involved in interactions with DNA and RNA, as well as other proteins, present repetitive and intrinsically disordered regions that are rich in G and R residues (termed glycine-arginine-rich, or GAR, domains). Within these domains, RG and GR motifs often overlap with each other, generating GRG/GRGG motifs that could represent shared targets of PRMT1 and PRMT5. The GR sequence belongs to the family of arginine- and glycine-rich motifs, whose asymmetrical and symmetrical arginine di-methylation affects protein-DNA and protein-protein interactions. Hence, it is very likely that these RGG domains carry both ADMA and SDMA in human cells, making the selective immunoenrichment of one di-methylation form over the other an indispensable step before LC-MS.

This observation prompted us to search for the neutral loss of monomethylamine and dimethylamine among the identified di-methylated peptides, as a way to distinguish the different di-methyl-R forms [123]. Notably, we identified the loss of monomethylamine in the MS2 spectra of all significantly down-regulated R-methyl peptides, confirming that they were true SDMA-bearing peptides (Figure 22). Conversely, the peptides displaying loss of dimethylamine, and not monomethylamine, were all found in the group of the ones unchanging upon PRMT5 inhibition. These results were published in [31].

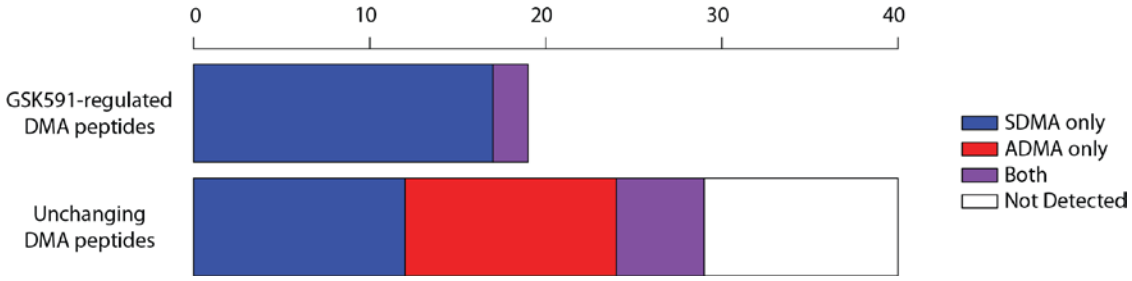


Figure 22. Neutral losses identified in the dimethyl-R-peptides MS/MS spectra. *The bar graph showing the number of R-methyl peptides displaying the neutral loss of monomethylamine (derived from SDMA) or dimethylamine (derived from ADMA) in the subsets of the GSK591-responding or not regulated. Adapted from [31].*

5.2.3. Targeting of RNA Splicing Catalysis through PRMTs inhibition

The observation that RNA splicing factors (SFs) are often mutated in cancer has led to clinical efforts to treat drug-resistant leukaemia with inhibitors of the spliceosome [124]. To uncover new means to therapeutically impact the process of splicing, our collaborators from the group of Ernesto Guccione sought to identify proteins with functional relationships to components of the core splicing machinery. In line with the notion that RNA-binding proteins (RBPs), including many SFs, are R-methylated, the specific PRMT5 inhibitor GSK591 and the type I PRMTs inhibitor MS023 emerged as promising candidates due to their capability to preferentially kill SF-mutant cells over their wild-type (WT) counterparts. Given these premises, we set out to explore the mechanistic basis for the link between inhibition of R methylation and its effects on SF-mutant leukaemia.

In brief, we performed two pairs of SILAC experiments (forward and reverse). In the forward experiments, cells cultured in the medium complemented with heavy-labelled Arginine and

Lysine were treated with GSK591 or MS023, while cells cultured with light amino acids were treated with vehicle; in the reverse experiments, the culture media were swapped. We quantified a total of 391 and 735 R-methyl-peptides in the GSK591 and MS023 experiments, respectively. Specifically, in the GSK591 treatment experiment, we found 299 peptides bearing MMA, 40 bearing DMA, and 52 bearing both modifications; in the MS023 experiment, we found 433 MMA peptides, 200 DMA peptides, and 102 peptides bearing both modifications. In total, these peptides carried 1188 R methylation events and were distributed on 219 different proteins. Analysis of the R-methyl-peptide SILAC ratios, normalized by the respective protein ratios, revealed that both MS023 and GSK591 caused a prominent down-regulation of methylation, with 135 (16%) and 49 (15%) methyl-peptides being significantly decreased, while only 97 (11%) and 4 (1%) were up-regulated, respectively (Figure 23).

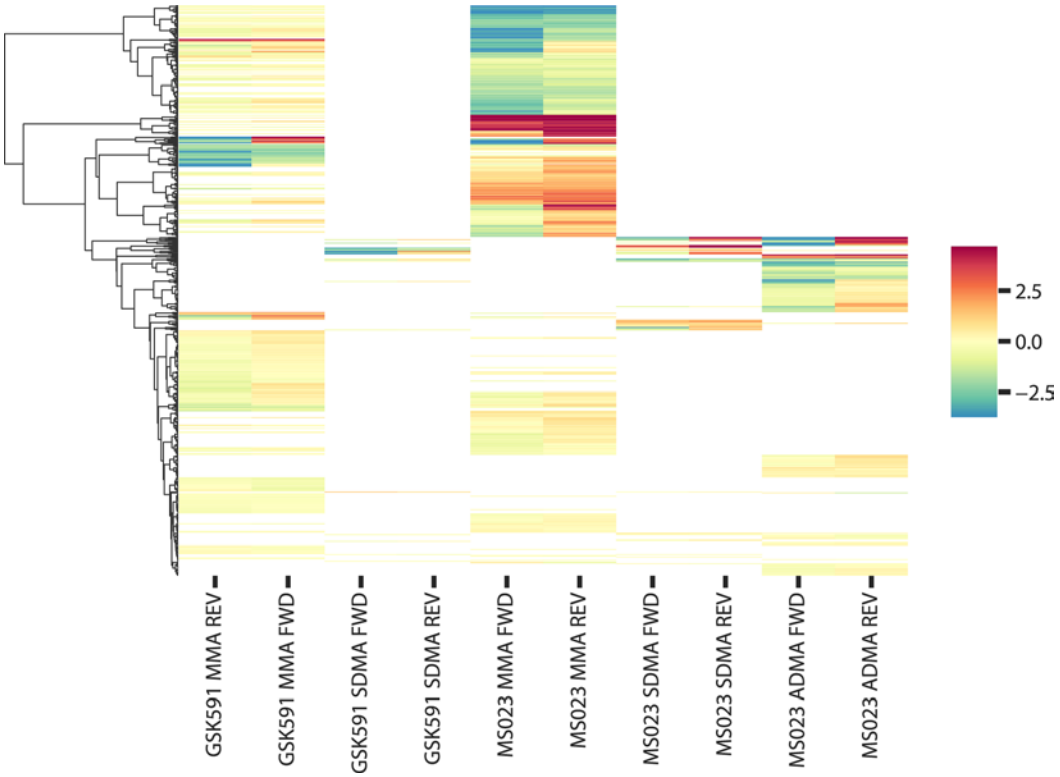


Figure 23. Heatmap of log2 SILAC ratios of each methyl-peptide identified and quantified from the SILAC experiment.

Motif analysis of the regulated peptides revealed that G was enriched around the methylation sites, as expected from sites that are targeted by PRMT1 and PRMT5 (Figure 24). This analysis also corroborated our previous finding that PRMT5 preferentially targets R-sites with

a G in position -1 (i.e. GR/GRG motifs), whereas PRMT1 and the other type I PRMTs target mostly RG/RGG motifs.

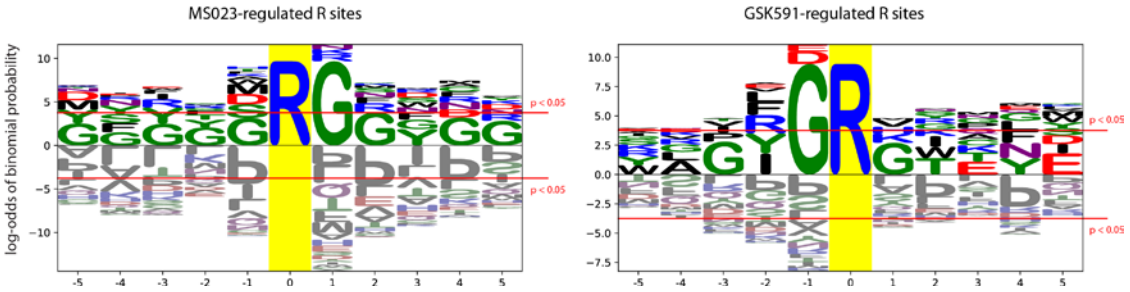


Figure 24. Motif analysis performed on methyl-sites regulated upon treatment with MS023 or GSK591. *The results indicate the consensus sequences significantly enriched in the methyl-peptides regulated by MS023 (left) or GSK591 (right).*

Building on these preliminary results, our collaborators from the Guccione group assessed the effect of combinatorial treatment of SF-mutant leukaemia cells with MS023 and GSK591. The treatment induced a unique pattern of splicing alterations that caused the mutant cells to undergo apoptosis. These results were published in [10].

5.2.4. Impact of PRMT1 modulation on microRNA biogenesis

Our group's preliminary data [125] revealed for the first time that many protein components of the Large Drosha Complex (LDC) were methylated. The LDC consists of RBPs that comprise the DEAD-box helicases DDX5 and DDX17 (also known as p68 and p72, respectively), several heterogeneous ribonucleoproteins (hnRNPs), the FET protein family EWSR, and other factors [126]. At the core of the LDC is the Microprocessor, a smaller complex formed by type III RNase Drosha, which processes primary miRNAs (pri-miRNAs) into precursor miRNAs (pre-miRNAs), and its interacting partner DGCR8, which binds the pri-miRNA. Notably, some LDC components such as ILF3, EWSR1, FUS and TAF15 were already known to be targets of PRMT1 [127,128], and R methylation is known to affect protein:protein and protein:RNA interactions. Based on this initial evidence, we performed a more in-depth analysis of R methylations within the LDC in order to understand how this PTM can regulate miRNA biogenesis.

We analysed the MS/MS spectra of the hmSILAC peptides detected in the LDC protein IP experiments and found that the majority of R di-methylations on LDC components are ADMA (Figure 25), pointing towards the involvement of the major type I PRMT, PRMT1, in the regulation of the complex activity and composition. This was later confirmed by observing a down-regulation of the miR-15a/16 and miR-17-92 miRNA clusters, whose deregulation was already associated with deregulation of some Microprocessor-associated proteins [129], in PRMT1 knock-down HeLa cells.

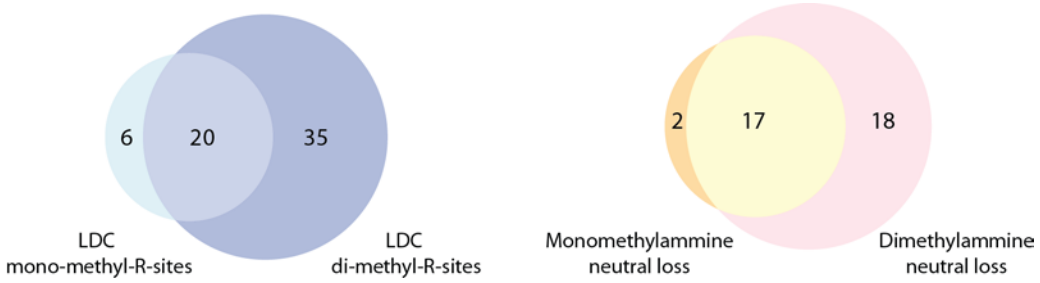


Figure 25. Composition of the LDC methyl-proteome. Left) *Overlap between mono- and di-methyl-R-sites identified on the components of the LDC.* ***Right)*** *Overlap between MS/MS spectra of di-methyl-R-peptides displaying a neutral loss of dimethylammine (associated to ADMA) or monomethylammine (associated to SDMA). Adapted from [34].*

To gain insights into the PRMT1-mediated regulation of the LDC activity we used SILAC to profile the composition and methylation state of the complex in PRMT1 knock-down (KD) or overexpressing (OE) HeLa cells (Figure 26A). After mixing, H and L cells were fractionated into nuclear and cytoplasmic extracts and a fraction of the nuclear extract was directly subjected to MS analysis for protein profiling of LDC. The rest was used as input to carry out three co-IPs in parallel, using DDX5, DGCR8 and FUS as baits. Each experiment was performed in three biological replicates. Although protein SILAC ratios indicated that the overall composition of the LDC was not altered upon PRMT1 modulation, 26 R-methyl-peptides were significantly regulated. These peptides could be divided into two clusters: cluster A, which included peptides hyper-methylated upon PRMT1 KD and hypo-methylated upon PRMT1 OE, and cluster B, which instead comprised peptides hyper-methylated by PRMT1 OE and hypo-methylated by PRMT1 KD (Figure 26B).

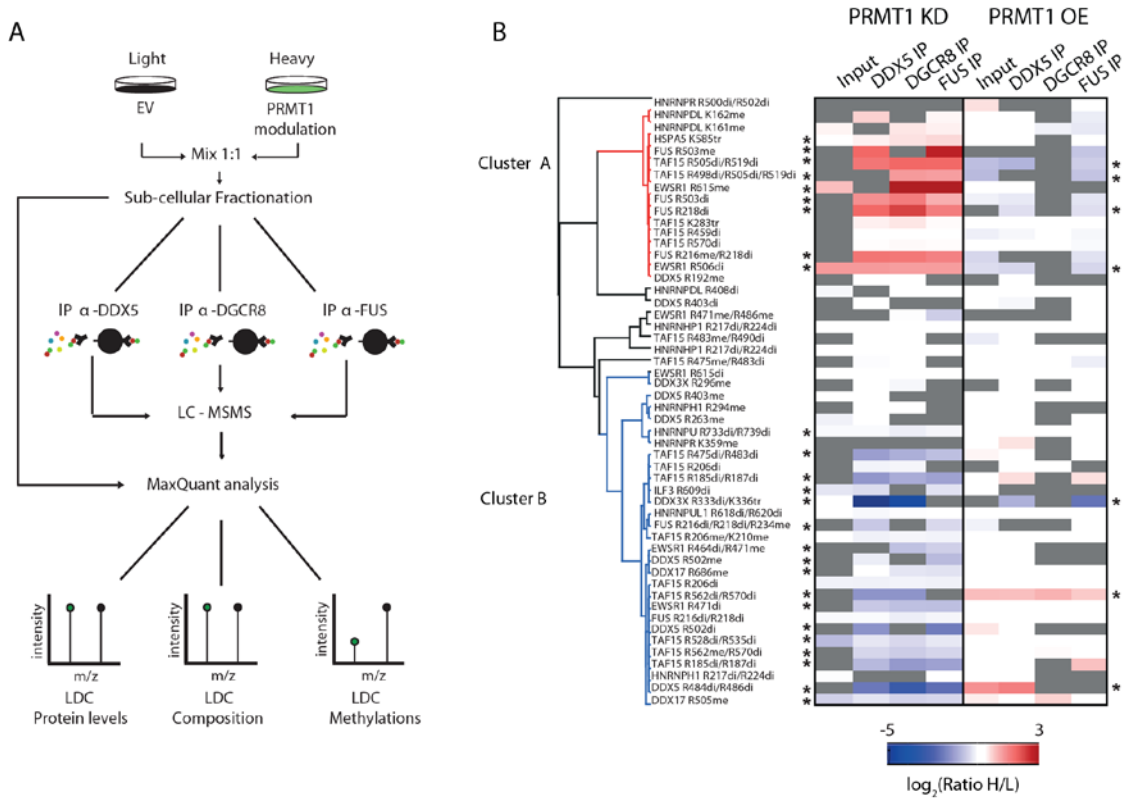


Figure 26. Methyl-proteomics profiling of LDC components upon PRMT1 modulation. A) *Schematic representation of the experiment.* ***B)*** *Heatmap representation of methyl-peptides SILAC rations in response to PRMT1 knock-down or overexpression. Adapted from [34].*

Motif analysis revealed that peptides in cluster B were enriched for the RG and RGG motifs, whereas cluster A peptides were only enriched for the RG motif (Figure 27). Therefore, we hypothesize that peptides in cluster B might be specific targets of PRMT1 (as suggested by their dynamic changes upon modulation of the methyltransferase), while peptides in cluster A are shared substrates of many PRMTs and are subject to scavenging by other PRMTs when PRMT1 is depleted. Overall, we identified and characterized by quantitative proteomics 29 R-methyl-sites on 8 LDC subunits (TAF15, FUS, EWSR1, ILF3, DDX3X, DDX3Y, DDX5 and DDX17) that are significantly regulated in dependence of PRMT1. Also, cluster B included peptides from DDX17, DDX3X, DDX5 and HNRNPH1 proteins, which we identify as novel targets of PRMT1. These results were published in [34].

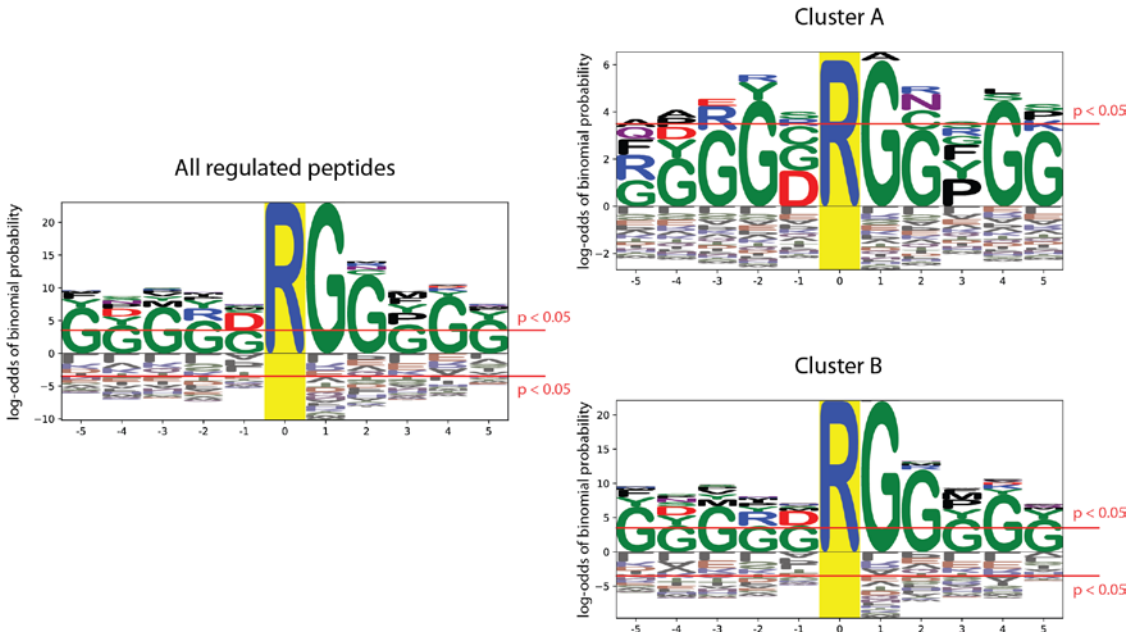


Figure 27. Motif analysis of methyl-peptides regulated upon PRMT1 modulation. *Peptides in cluster B show a stronger enrichment of G in position +2 compared to peptides in cluster A, suggesting that they are specific targets of PRMT1.*

5.3. Integration of SILAC and hmSILAC into ProMetheusDB

Upon completing the analysis of methyl-proteome changes in response to CDDP treatment, PRMT1 modulation, and PRMT5 and type I PRMTs inhibition, we integrated the "static" hmSILAC database with "dynamic" data from the SILAC experiments (summarized in Figure 28A). Notably, we included the final database, which was named ProMetheusDB, all the significantly regulated methyl-peptides quantified in the SILAC that were not already identified in the hmSILAC experiments. This was done under the assumption that methyl-peptides displaying statistically significant changes to a biological stimulus are likely to be enzymatic, whereas amino acid substitutions and chemical artefacts introduced by sample preparation should remain unaffected by perturbations. We thus obtained a final, larger dataset consisting of 2865 methyl-peptides, 723 methyl-proteins, 1814 methyl-sites and 2271 methylation events, which corresponds to our reference ProMetheusDB dataset of MS-identified high-quality K/R-methylations (Figure 28B, left).

A

Summary of SILAC experiments

Cell line(s)	Enrichment	Antibody specificity	Stimulus
NB4	Peptide IP	MMA, SDMA	GSK591 vs UT
NB4	Peptide IP	MMA, ADMA, SDMA	MS023 vs UT
HeLa	Peptide IP	MMA, SDMA	GSK591 vs UT
SK-OV-3	Peptide IP	MMA, ADMA	CDDP vs UT / PRMT1 KD + CDDP vs UT / PRMT1 KD + CDDP vs CDDP
SK-OV-3	Peptide IP	MMA, ADMA	CDDP vs UT / MS023 + CDDP vs UT / MS023 + CDDP vs CDDP
SK-OV-3	Peptide IP	MMA, ADMA, SDMA	CDDP vs UT / PRMT1 KD vs UT
HeLa	Protein IP	LDC components	PRMT1 KD vs UT / PRMT1 OE vs UT

B

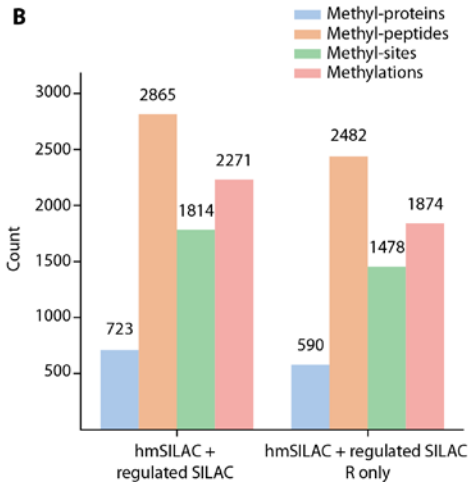


Figure 28. Integration of SILAC and hmSILAC datasets to generate ProMetheusDB. ***A)*** *Summary of SILAC experiments used to integrate our high-confidence methyl-proteome. Only peptides that were significantly regulated in one or more experiments were included, based on the assumption that dynamically regulated peptides are unlikely to be artefacts.* ***B)*** *Composition of the R-methyl-proteome after inclusion of the regulated SILAC methyl-peptides.*

To characterise the features of ProMetheusDB, we first checked its composition in terms of K and R methylation events. To avoid misassignment of R methylation events to K, and vice versa, K methylation was included in the MaxQuant database search; however, we expected the overall coverage of the R-methyl-proteome in the ProMetheusDB to be much larger than that of K-methyl-proteome, due to the employment of antibodies raised against methyl-Rs in most of our experiments. We identified 113 proteins bearing K-methyl-sites or combinations of K- and R-methyl-sites (Figure 29A), although peptides on which K and R methylation coexist were overall quite rare (n=88, Figure 29B), suggesting that K- and R-methyl-sites are not necessarily neighbouring but rather located in different protein regions. Subsequent analyses were focused on the R-methyl-proteome, which included 2482 R-methyl-peptides, 590 proteins, 1472 R-sites and 1874 R-modification (Figure 28B, right).

Afterwards, we asked how many peptides of the "R-methyl-peptides only" dataset were identified in the hmSILAC and SILAC experiments (Figure 29C) and found that most of the R-methyl-peptides were identified in the hmSILAC data (1945) or both (364); overall, the SILAC data contributes to 173 R-methyl-peptides (7% of the entire R-methyl-proteome). In addition, we compared the R-methyl-sites in ProMetheusDB to the hmSILAC-validated R-

methyl-sites in PhosphositePlus. We found that 752 (51%) of the sites in ProMetheusDB are already annotated, while the remaining 720 (49%) are novel (figure 29D). To uncover a possible cross-talk between the two degrees of R methylation, we compared the identified mono-methyl-sites and di-methyl-sites and observed that, of the total 1478 R sites, 783 are only mono-methylated, 290 are only di-methylated and 405 are identified in both forms (Figure 29E).

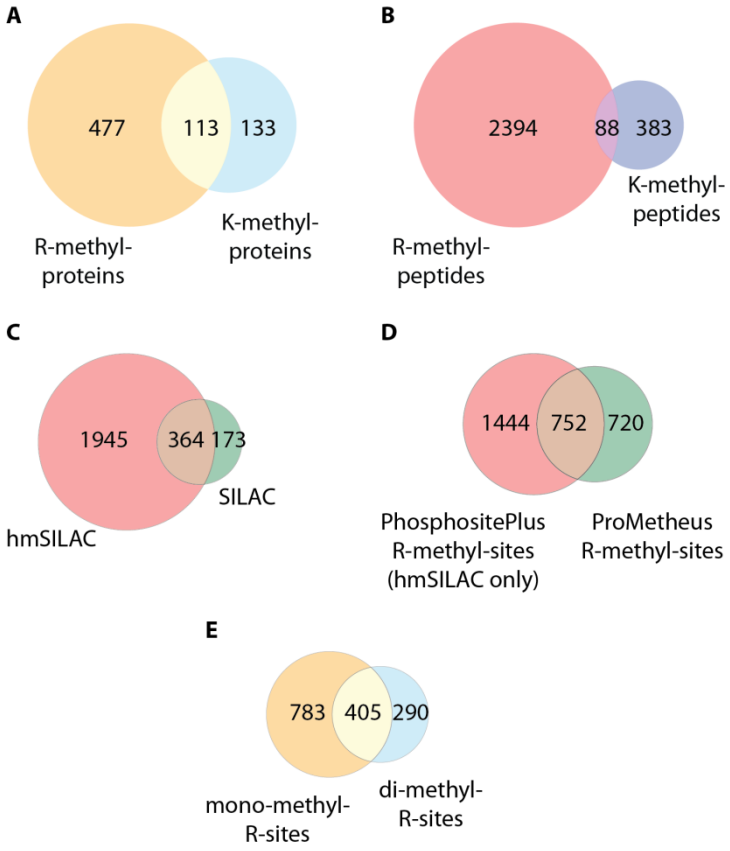


Figure 29. Analysis of the composition of ProMetheusDB. ***A-B)*** *Overlap between R-methylated and K-methylated peptides and proteins, respectively.* ***C)*** *Comparison between methyl-peptides identified with high confidence in the hmSILAC and SILAC experiments.* ***D)*** *Venn diagram comparing R-methyl-peptides in ProMetheusDB and annotated in Phosphosite Plus. E) Venn diagram comparing R-methyl-sites that have been identified as mono- or di-methylated.*

To verify if these counts of mono- and di-methylated R residues reflected a bias in the sample preparation protocol, we grouped our experiments based on whether the anti-pan-R-methyl antibodies from CST PTMScan kits were used for methyl-peptides enrichment before MS.

Indeed, we observed that mono- and di-methyl-R sites were equally represented when no affinity enrichment of R methylations was performed (Figure 30).

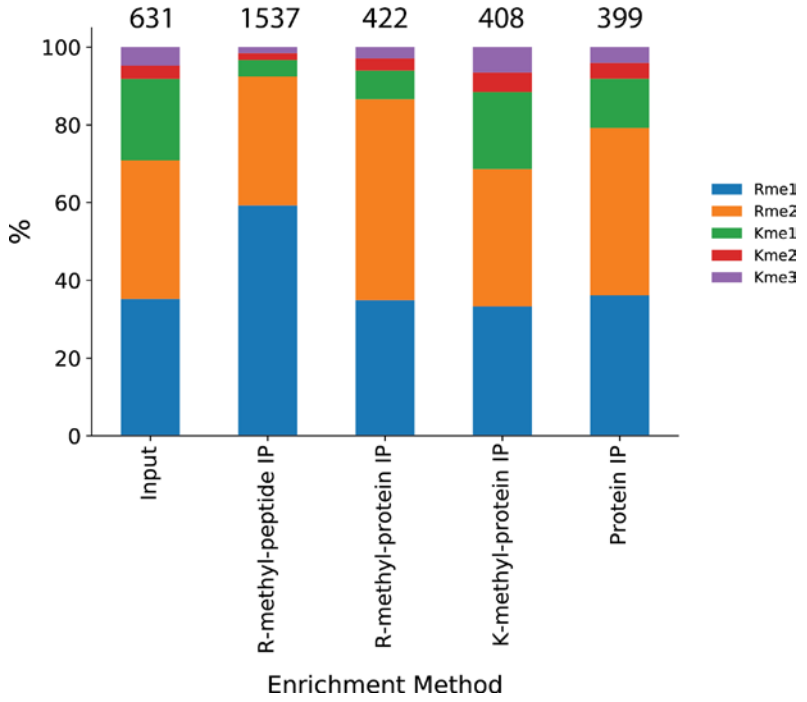


Figure 30. Bar chart displaying the fraction of methylation marks identified in each sub-category of hmSILAC experiments. *Numbers on top of each bar indicate the total number of modifications annotated.*

5.4. Functional analysis of the R-methyl-proteome

To explore the potential impact of protein R methylation on cellular processes, we used Cytoscape to generate a network of protein functional interactions, starting from the full list of R-methyl-proteins included in our final ProMetheusDB. Clustering analysis highlighted the presence of eight protein clusters enriched for distinct biological pathways (Figure 31): the largest cluster includes several components of the spliceosome and other proteins involved in mRNA processing, for a total of over 100 proteins (Figure 29), confirming the known role of R methylation in RNA metabolism that has been already described by both our group and others [10,34,130]. The remaining clusters included around 20-30 proteins each and were enriched for terms such as “DNA-binding transcription factor activity”, “Lipid metabolism”, “Cytoskeleton” and “Mitotic Anaphase and Metaphase”, “Fc Gamma receptor (FCGR)-

dependent phagocytosis", "Translation", "Antigen processing" and "Chromatin organization" (Figure 32).

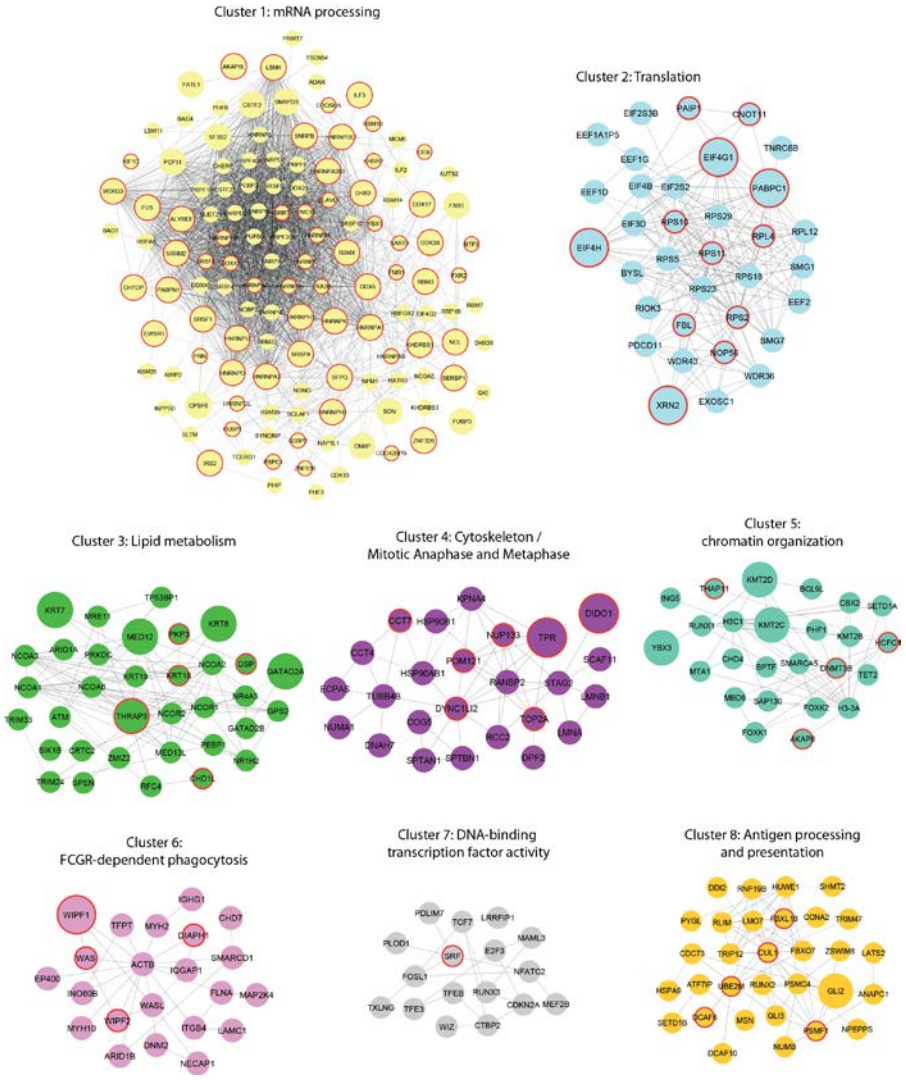


Figure 31. Protein clusters identified in the functional interaction network of R-methylated proteins. *Larger nodes represent proteins with 5 or more methylation sites, while red borders highlight proteins bearing regulated methylation sites.*

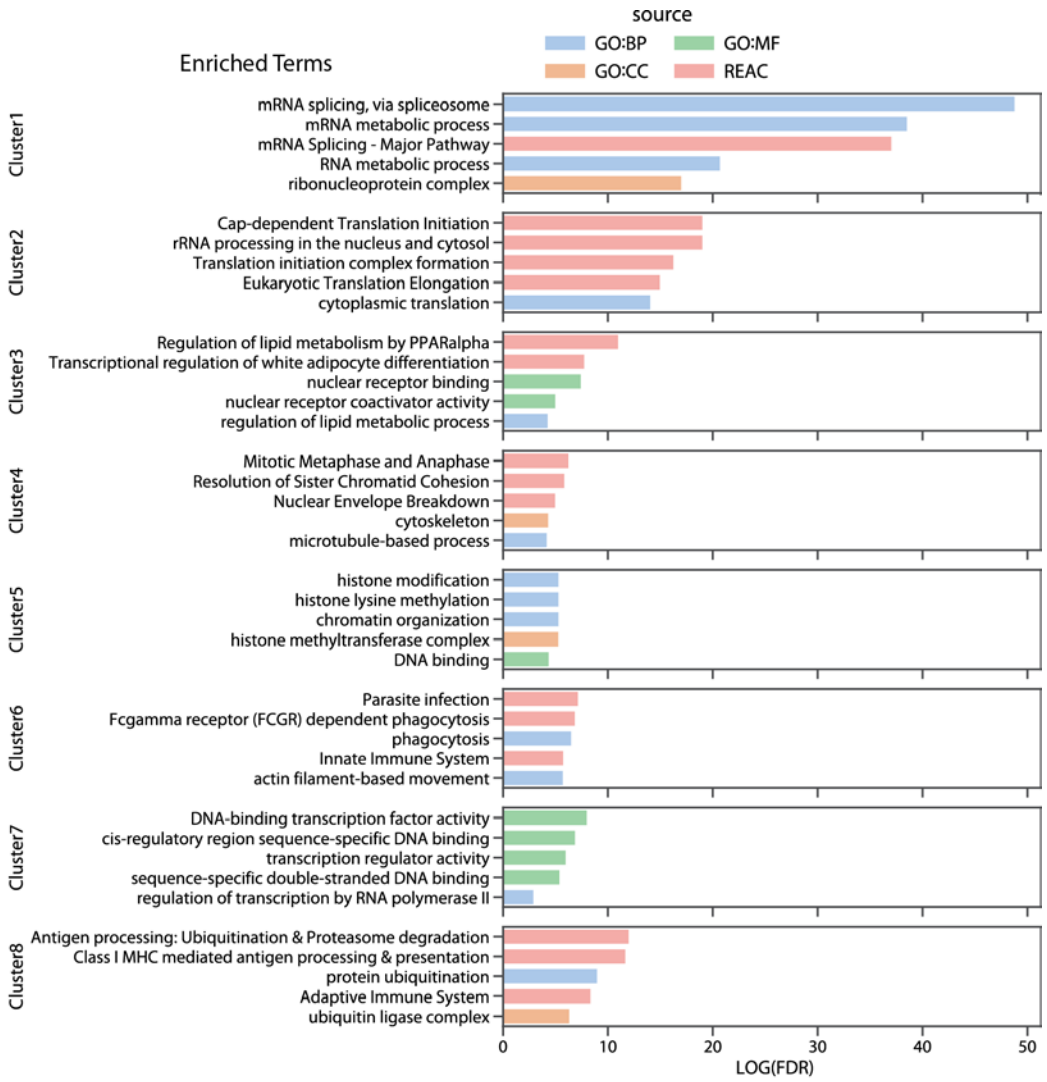


Figure 32. Functional enrichment of protein within each cluster. *The top 5 most enriched functional terms for each protein cluster are shown (GO:BP = Gene ontology biological processes; GO:CC = gene ontology cellular component; GO:MF = Gene ontology molecular function; REAC = Reactome pathways).*

Having previously observed a strong tendency of methylated proteins to be part of complexes [125], we asked whether this observation still held in the newly annotated, expanded methyl-proteome: we found that hyper-methylated proteins (bearing five or more R methyl-sites) presented significantly higher node degree and network centrality, not only compared to unmethylated proteins (detected in the Input samples) but also to hypo-methylated ones (presenting one or two R-methyl-sites) (Figure 33).

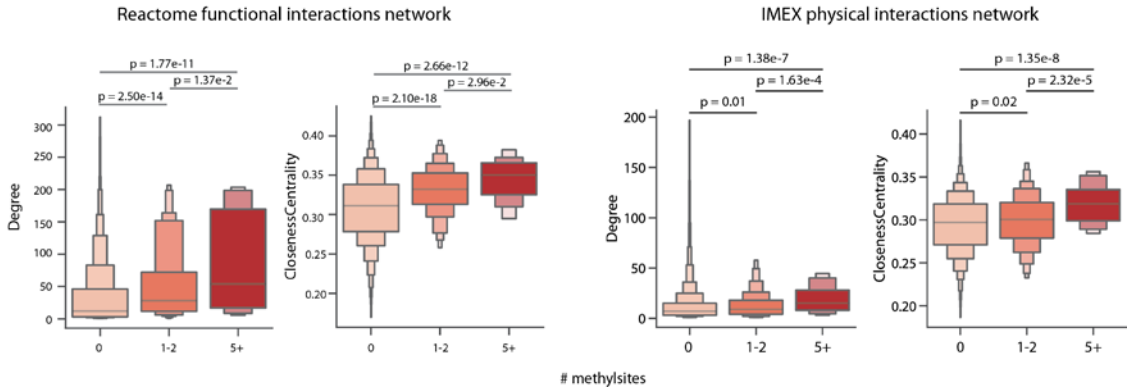


Figure 33. Network topology analysis to compare unmethylated, hypo-methylated and hyper-methylated proteins. *Boxen plot representation of node degree and closeness centrality for each category of proteins. The degree and centrality of proteins increase at increasing numbers of R-methyl-sites (p-values calculated with Kruskal-Wallis test). The analysis was initially performed on a network of functional interactions and the results were later corroborated by repeating the analysis on a network of physical interactions.*

As expected, several high-centrality, hyper-methylated proteins are known RBPs, such as RBMX, NONO, various ribosomal proteins, nucleolin and several HNRNPs [131]; yet, we find interesting the presence of R-methyl-proteins in pathways related to lipid metabolism, cytoskeleton organization and phagocytosis. These proteins include the stress granule nucleator G3BP1, which is also part of DNA/RNA-sensing pathways implicated in the regulation of innate immunity [132,133]; the cytoskeletal protein Actin (ACTB); the nuclear receptors NCOA2 and NCOA3, that are involved in metabolism, inflammation and adipocytes differentiation [134]; CUL1, a component of several E3 ubiquitin-protein ligase complexes [135]; SMARCA5, a helicase whose nucleosome-remodelling activity has a role in transcription, phosphorylation of H2AX, and maintenance of chromatin structures during DNA replication [136]. When taken together, these proteins and pathways suggest a role of R methylation in innate and adaptive immunity, in line with evidence linking PRMT1 to macrophage differentiation and apoptosis, inflammation and cytokine production [137-139]. Moreover, although all the cited proteins were already known to be methylated, we found several novel R-methyl-sites on RBMX (R383, R388), NONO (R142), SMARCA5 (R616), HSP90B1 (R51, R557) and ACTB (R206, R312).

As a further analysis, we integrated the network with the quantitative data on methylation dynamics obtained from the SILAC experiments, to identify which proteins present at least one R-methyl-site that is significantly regulated in at least one SILAC experiment (i.e. regulated in the same way in one pair of Forward and Reverse replicates) in response to different perturbations. Proteins featuring at least one significantly regulated R-site show higher node degree and centrality in the network (Figure 34).

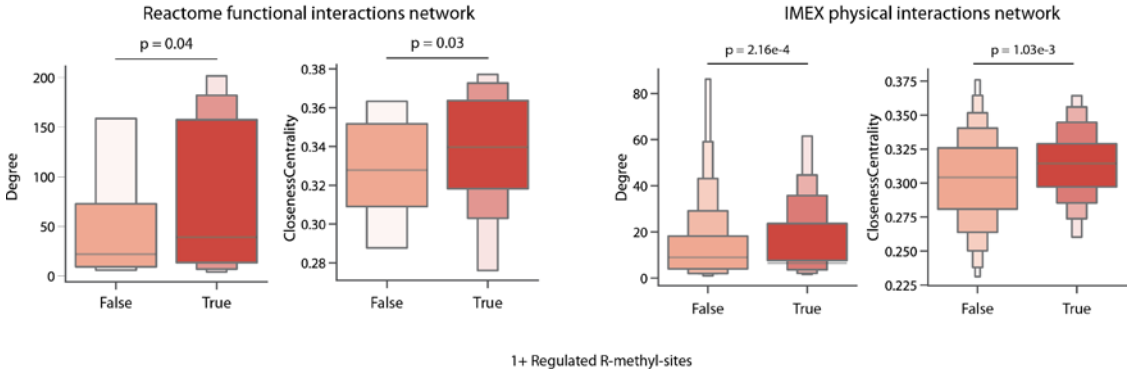


Figure 34. Network topology analysis to compare proteins bearing significantly regulated R-methyl-sites or not. *Boxen plot representation of node degree and closeness centrality for each category of proteins. Proteins bearing at least 1 regulated R-methyl-site have a significantly higher degree and network centrality than the rest (p-values calculated with Mann-Whitney test). The analysis was initially performed on a network of functional interactions and the results were later corroborated by repeating the analysis on a network of physical interactions.*

In addition, we observed that this group of proteins is also enriched for gene ontology terms like “RNA binding”, “RNA splicing” and “nucleic acid transport” (Figure 35, top), whereas proteins without regulated methylations were not enriched for any specific category (Figure 35, bottom). This result may suggest that there are two categories of R-methyl-sites: one includes sites that play a role in the regulation of protein-RNA interactions while the other comprises “structural” methylations, which are unperturbed by biological stimuli and occur on a wider range of proteins.

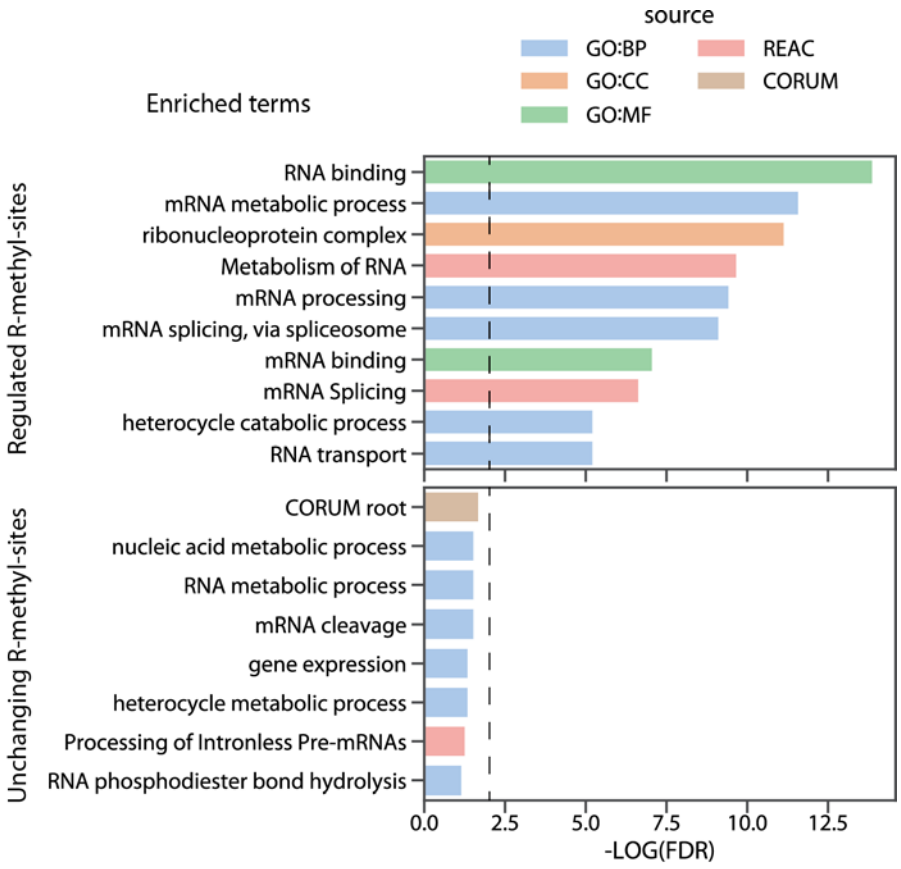


Figure 35. Functional enrichment of proteins being regulates R-methyl-sites (top) or not (bottom)

RBPs are often involved in the process of Liquid-Liquid Phase Separation (LLPS). To see if there was a connection between this process and R methylation, we intersected ProMetheusDB with PhaSepDB, a database of proteins that are involved in LLPS. We found that more than half of the proteins in our database were also annotated as part of one MLO inside PhaSepDB, reinforcing the concept that R methylation is crucial for MLOs assembly (Figure 36).

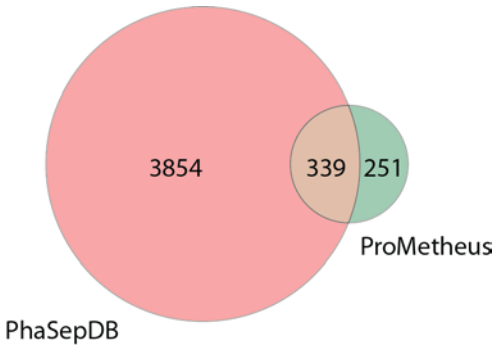


Figure 36. Intersection of PhaSepDB and ProMetheusDB. *57% of R-methyl-proteins are involved in the process of liquid-liquid phase separation (LLPS).*

Finally, having already observed during the analysis of the individual SILAC experiments that the major type I and II PRMTs (PRMT1 and PRMT5) preferentially target R-sites within glycine/arginine-rich (GAR) motifs, we repeated the motif analysis on the overall database. We found that regulated sites are surrounded by G residues, indicating that they are methylated by PRMT1 and PRMT5 (Figure 37, left). Instead, R-methyl-sites that emerged as unchanging in the different functional states did not produce a significant enrichment for any amino acid consensus motif (Figure 37, right): this result may suggest that unchanging methylations are deposed by methyltransferases other than the best-known ones, which may recognize different sequence motifs, yet to be characterised.

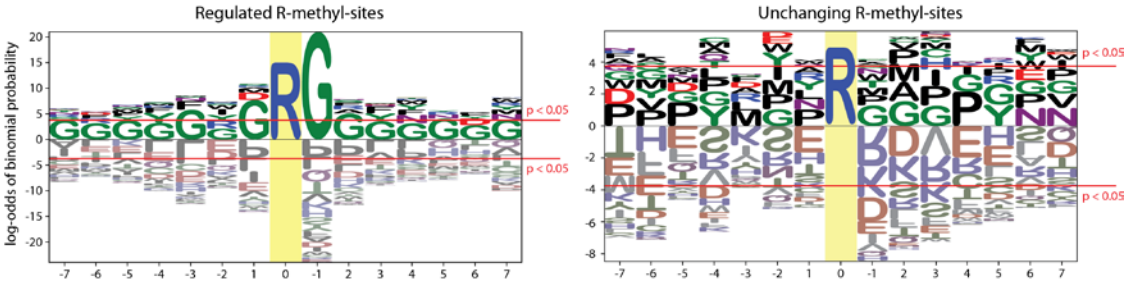


Figure 37. Motif analysis performed on significantly regulated (left) and unchanging (righty) R-sites. *Logos were generated using the full list of identified methyl-sites as background.*

5.5. Structural analysis of R-methyl-protein regions

In the last two decades, growing evidence indicated that intrinsically disordered regions of proteins play a role in different processes (such as protein-protein and protein-nucleic acid interactions and LLPS) and are often regulated by PTMs [140-142]. Having previously observed the preferential localisation of R-methyl-site in unstructured regions, we wanted to verify whether the functional link between R methylation and specific protein structures could be confirmed in our large-scale, high-quality R-methyl-proteome. Hence, we overlapped the R-methyl-sites in the ProMetheusDB with disordered regions and domains annotated in MobiDB [143] and confirmed that the vast majority (77%) of the R-methyl-sites are indeed located in disordered regions. When we carried out the same analysis on a subset of 1500 non-methylated R-sites, randomly extracted from the UniProt human proteome (version 2020_01), we observed a more even distribution between structured domains and disordered regions (Figure 38A). In fact, upon performing a Fisher's exact test on the counts of methylated or unmodified R located in structured or disordered regions, we found that the distribution was not random; this result corroborates previous evidence that R-methyl-sites tend to occur much more frequently within low complexity and disordered regions.

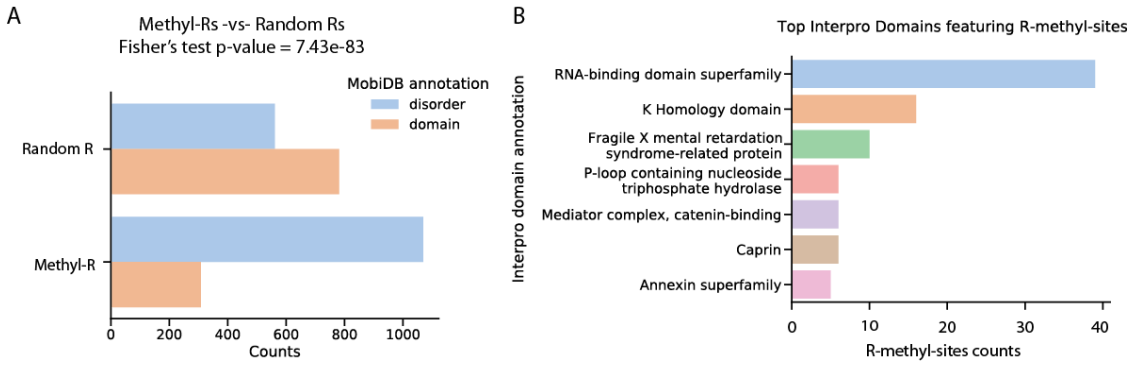


Figure 38. Structural analysis of R-methyl-sites. A) *Counts of R-methyl-sites that occur in regions that are annotated as domains or disordered in the MobiDB database. Counts of Random R sites are also included as a comparison and a p-value was calculated with Fisher exact test.* ***B)*** *Counts of R-methyl-sites occurring on specific protein domains, as annotated in the InterPro database.*

Despite their low number, we focused on the 308 R-methyl-sites located in structured regions and mapped them to InterPro domains [116] to investigate their possible association with

specific protein structures. We found that most R-methyl-sites localize within “RNA Binding Domain” or “K Homology domain”, or on the “Fragile X mental retardation syndrome-related” family of proteins (figure 38B). Besides the “RNA Binding Domain”, whose enrichment was expected, the K Homology domain is also mainly located in RBPs (specifically heterogeneous nuclear ribonucleoproteins, [144]) and the Fragile X mental retardation syndrome-related proteins also belong to the RBP family; hence, these results are consistent with the notion that RNA binding and processing are highly dependent on this PTM.

We then set out to investigate the physical interaction of R-methyl-proteins with other types of biomolecules. To do so, we analysed ProMetheusDB using Mechismo, a web application that maps alterations of amino acid residues (induced by mutations or PTMs) onto protein crystal structures to identify modifications sites that putatively engage in physical interactions [117,145]. We identified 26 R-methyl-sites that were involved in a total of 65 interactions with nucleic acids (n = 11), chemical compounds (n = 7), other copies of the same protein (n = 5) and other proteins (n = 42) (Figure 39). The presence of more interactions with proteins rather than with nucleic acids for these modifications located in structured domains is surprising, given that methyl-proteins are enriched for RBPs, but it can be explained by the fact that many RBPs bind RNA through low complexity regions for which crystal structures are not available. Still, these results support the hypothesis that R methylation can also modulate interactions with biomolecules other than RNAs, such as protein:protein and protein:chemical.

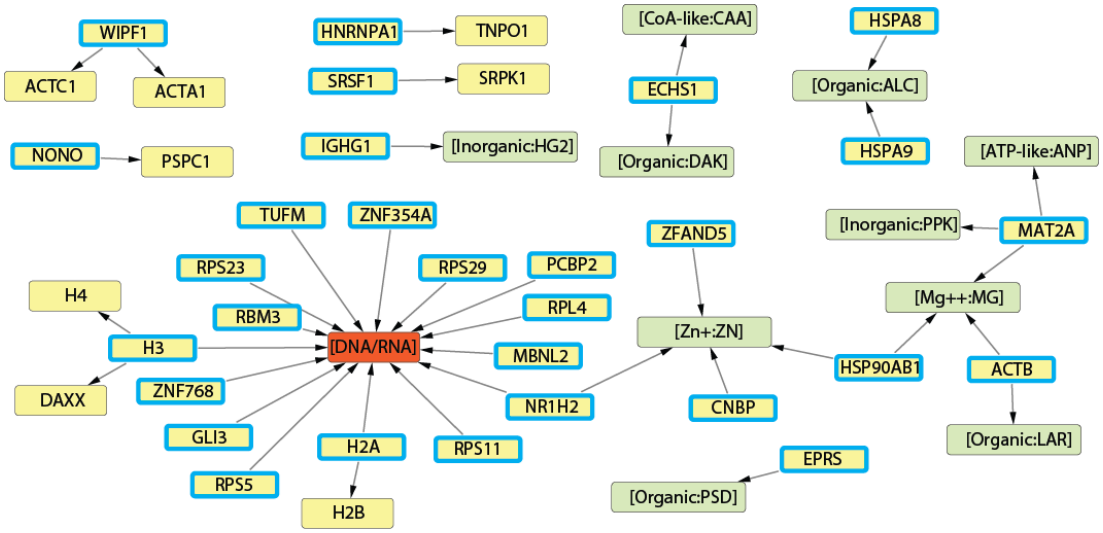


Figure 39. Network obtained from Mechismo. *Methyl-proteins in ProMetheusDB and their known or predicted interaction partners are shown. DNA/RNA and chemicals are highlighted in orange and green, respectively. R-methylated proteins are indicated by a blue border.*

We followed up the R-methyl-sites indicated by Mechismo analysis as putatively involved in protein-:rotein interactions: based on reports that methylation of the guanidino group of R can exert distinct effects on protein:protein interactions by reducing the number of hydrogen bond donors on this amino acid while also increasing its hydrophobicity [146], we hypothesized that methylation could, on the one hand, enhance interactions between R and hydrophobic residues and, on the other hand, inhibit the interactions between R and negatively charged amino acids. Based on this, we inspected the crystal structures of the methyl-proteins bound to their respective interactors that emerged from the Mechismo analysis, to formulate specific predictions on the impact of methylation on specific pairs of protein interactors.

An interesting case study highlighted by the structural analysis is represented by the protein pair NONO:PSPC1, where R256 of NONO is found at the interface with PSPC1 (Figure 40A). We, therefore, assumed that methylation of R256 would increase its hydrophobicity, thus stabilizing the NONO: PSPC1 interaction interface, which also entails apolar residues such as L222 of PSPC1. To experimentally verify this prediction, we set up the NONO IP in HeLa cells and profiled PSCP1 co-immunoprecipitation efficiency, both in basal conditions and upon treatment with the PRMT type I inhibitor MS023, which typically leads to ADMA reduction and MMA/SDMA increase, depending on the enzyme processivity on a specific site and/or the scavenging effects by other PRMTs, already reported [147,148]. When we profiled

both NONO R methylation state and interaction with PSPC1, we observed that MS023 induced the decrease of ADMA on NONO, with a parallel increase of MMA and - to a minor extent - of SDMA; in parallel, a mild increase of the amount of PSPC1 co-immunoprecipitated was also observed, which confirms that increased methylation levels favour NONO:PSPC1 interaction (Figure 40B).

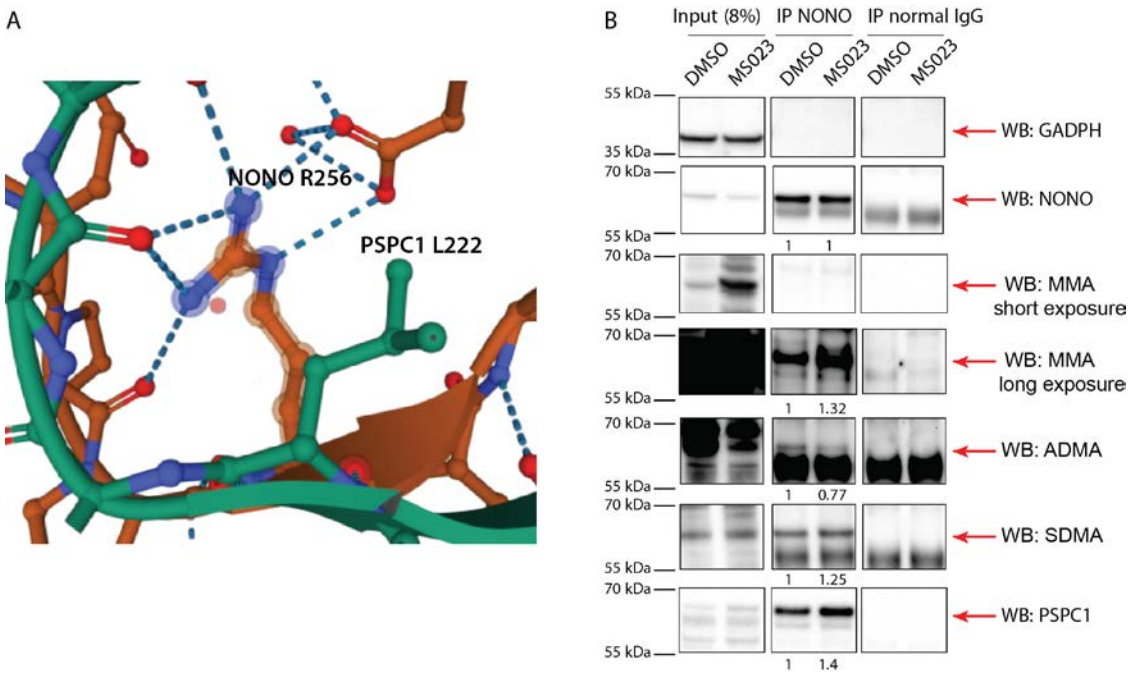


Figure 40. Assessment of NONO:PSPC1 interaction dependency on R methylation. A) *Crystal structure of NONO (orange) and PSPC1 (green) showing how R256 of NONO, which we identify as methylated, interacts with L222 of PSPC1, suggesting a potential role in the binding of the two proteins.* ***B)*** *IP of NONO and western blot (WB) analysis of NONO, its R methylation state and PSPC1 co-IP, in cells untreated or MS023-treated. Increased levels of MMA and SDMA on NONO and of co-immunoprecipitated PSPC1 were observed upon MS023 treatment, compared to DMSO. Quantification of the signal intensity for each band was performed by Fiji software and normalized as described in Material & Methods.*

5.6. Cross-talk between R methylation and S/T-Y phosphorylation

The cross-talk between R methylation and S/T-Y phosphorylation has already been described in the literature and linked to the subcellular localization of proteins [149] and the promotion of stem-like properties in cancer [150]. Interestingly, phosphorylation and methylation share some features, such as both being localized in disordered protein regions and both being functionally linked to the modulation of protein-RNA interaction and LLPS [151-153]. To study this aspect more in-depth, we determined a 15-amino acid window, centred on each R-methyl-site annotated in ProMetheusDB, then we downloaded the list of annotated phosphorylation sites from the Phosphosite Plus database [154] and assessed if phosphorylated residues are preferentially enriched within the methyl-R-centred sequence windows. We found that 706 (47.8%) of the 1478 R-methyl-sites in our dataset present an S, T or Y phosphorylation site within their respective 15-amino acid window (Figure 38). By repeating this analysis on 1500 R residues randomly selected from the human proteome, we observed that the proportion of sites featuring a phosphorylation site in the same 15-amino acid window was halved (356 sites, corresponding to 23.7%). This result indicates a statistically significant overrepresentation of phosphosites in the proximity of methyl-R compared to randomly selected R sites, and supports the idea of co-occurrence of these two PTMs (Figure 41A). We then expanded the analysis of possible cross-talk of R methylation with other PTMs and found that R-methyl-sites seem to significantly anti-correlate with the presence of a nearby K ubiquitination site, compared to the dataset of randomly selected R-sites (Figure 41B). Instead, the associations between R methylation and K acetylation or sumoylation are not statistically significant (Figure 41C-D). However, these observations shall be taken with caution because the numbers of acetylation, ubiquitination and sumoylation sites annotated in Phosphosite Plus is one order of magnitude lower than that of phosphosites thus introducing a potential bias in our analysis. Still, the co-presence of phosphorylation and methylation seems to be specific and not generically applicable to all annotated PTMs.

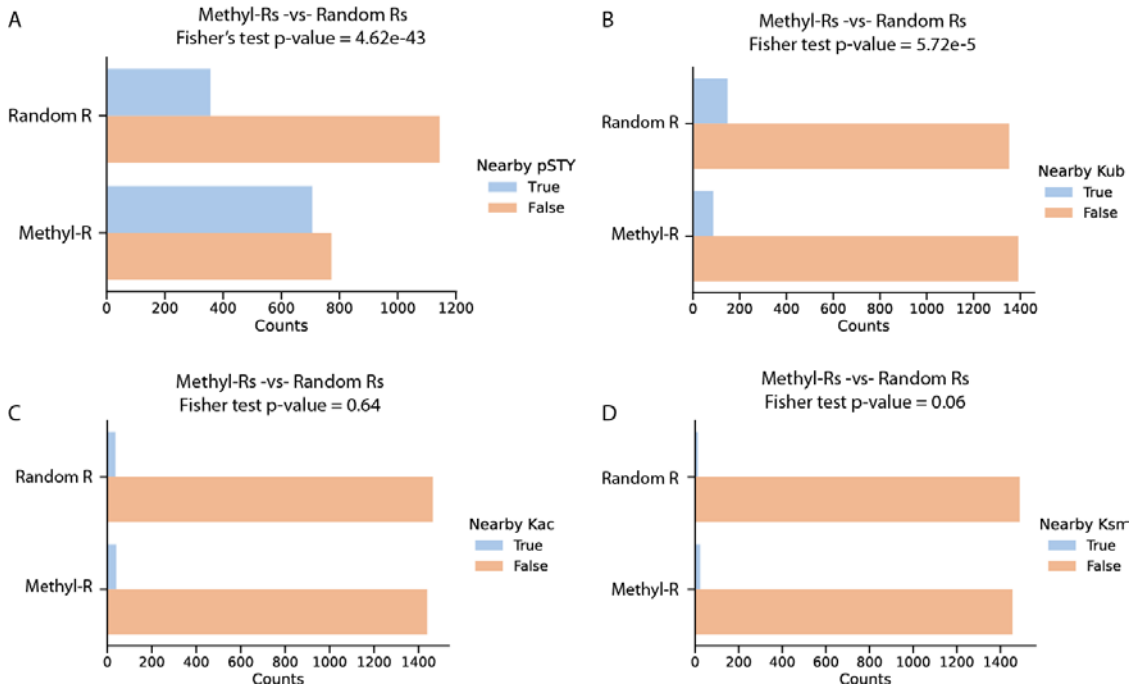


Figure 41. Counts of R-methyl-sites that occur in the proximity of a phosphorylation (A), ubiquitination (B), acetylation(C) or sumoylation (D) site. *As a control, counts of randomly sampled R sites from the human proteome are also shown (p-values calculated by Fisher's exact test).*

To understand whether this co-presence of methylation and phosphorylation was linked to a specific biological process, we performed GO analysis on the proteins showing a statistically significant co-presence of R-methyl-sites and phospho-S/T-Y sites. Overall, upon performing a functional enrichment using the entire list of proteins in ProMetheusDB as background, we found these proteins are even more strongly linked to RNA binding and splicing (Figure 42).

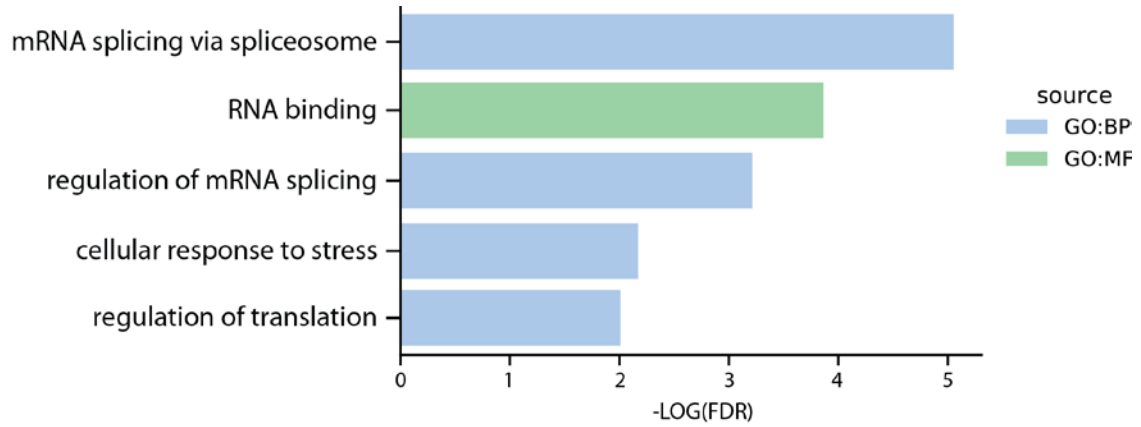


Figure 42. Functional terms enriched from proteins that display proximal R methylation and S/T-Y phosphorylation sites. *(GO:BP = Gene ontology biological processes; GO:CC = gene ontology cellular component; GO:MF = Gene ontology molecular function; REAC = Reactome pathways; KEGG = KEGG pathways; CORUM = CORUM protein complexes.)*

Next, we asked whether the R-methyl-sites co-occurring with phosphosites are subject to more change in response to a biological stimulus than those that are not associated. We intersected the results of the cross-talk analysis with the dynamic information of the SILAC experiments, where we profiled methylation changes in response to type I PRMTs inhibition by MS023, cisplatin (CDDP) treatment and PRMT1 knock-down (KD) or overexpression (OE). Despite being very different stimuli, we expected to observe an effect on PRMT1 methylation targets in all cases, as it is the most active type I enzyme (and thus more affected by MS023 than other PRMTs) and CDDP treatment causes relocalization of PRMT1 to chromatin (therefore inducing a decrease in the methylation levels of cytosolic proteins). Interestingly, we found that phospho-S and phospho-Y sites seem to occur significantly more frequently in the proximity of methyl-sites that are MS023-regulated than of the non-regulated ones (Figure 43A-B). We also observed that the R-methyl-sites unchanging upon either CDDP treatment or PRMT1 knock-down tend to present a phospho-T site in their surrounding 15-amino acid sequence window (Figure 43C-D). The different responses we observe to these stimuli suggest that there are multiple ways these PTMs can affect each other; these different mechanisms imply that non-histone proteins could be regulated by a “code” of modification similar to the one extensively described on histones.

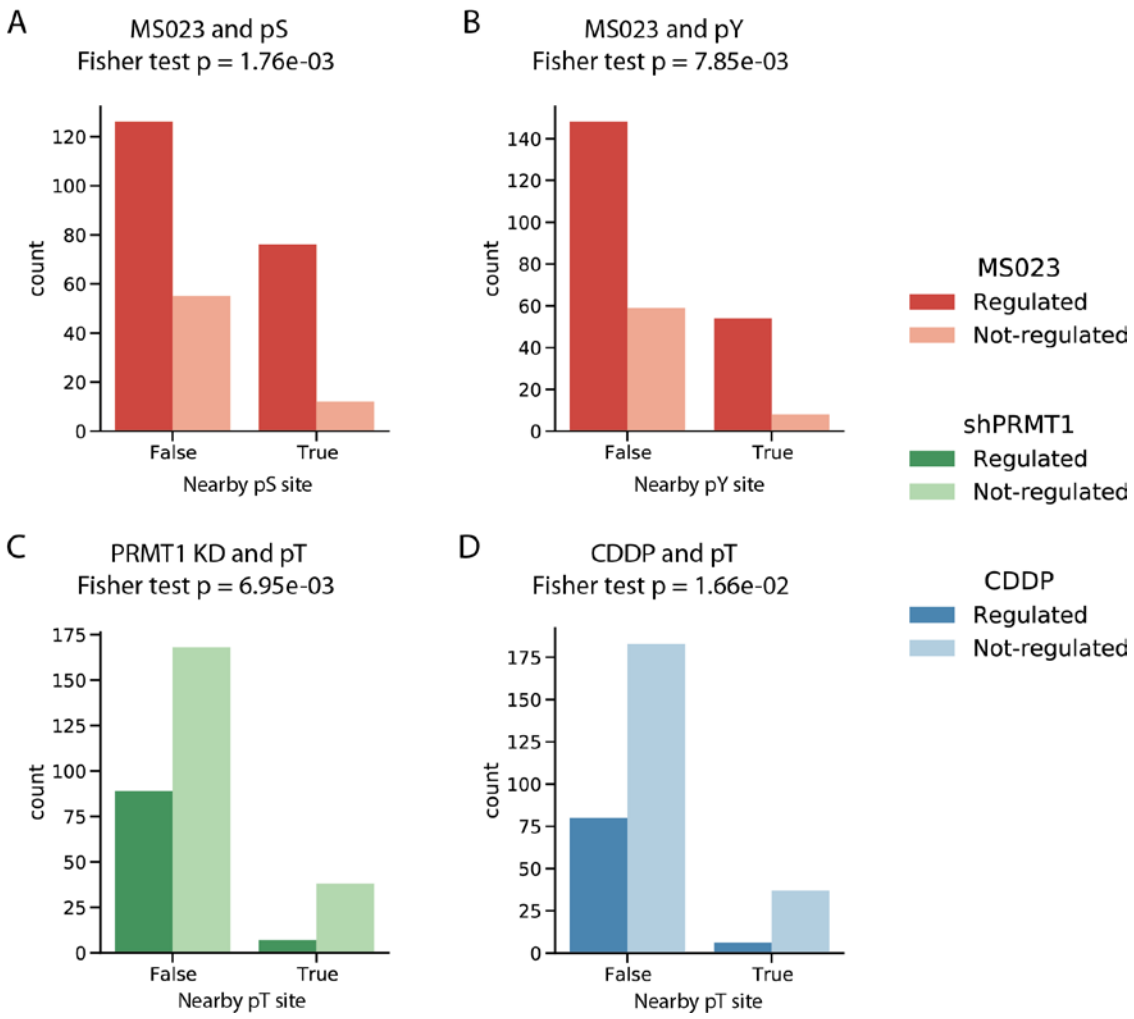


Figure 43. Cross-talk between regulated R-methyl-sites and phosphorylation sites. *The bar plots show counts of R-methyl-sites classified based on their regulation state upon different stimuli and their proximity to S/T-Y phosphorylation sites (p-values calculated by Fisher's exact test).*

5.7. Methylation beyond K and R: hmSILAC-based detection of non-canonical methylation sites

Protein methylation has been observed not only on K and R but also on residues such as Aspartate (D), Glutamate (E), Asparagine (N), Glutamine (Q), Serine (S) threonine (T) and Histidine (H). However, besides H [16], a systematic study on these non-canonical methylated residues by MS is lacking, due not only to the high FDR that already plagues R methylation studies but also to the difficulty in pinpointing the PTM to its exact residue when multiple

putative methyl-sites are present on the same peptide. Within this context, we aimed to assess to what extent the hmSILAC strategy could help to address this issue and fill this gap of knowledge. Intending to fully exploit the hmSILAC strategy and the large set of MS data available to us, we set to re-analyze all our hmSILAC-labelled, non-affinity enriched samples with the Andromeda search engine of MaxQuant, allowing mono-methylation to occur not on D, E, N, Q, S, T and H in addition to K and R, under the rationale that all enzymatically-driven methylation should use SAM as the universal methyl-group donor.

MaxQuant output data were processed with hmSEEKER 2.0 using the same criteria employed for the assignment of the high-quality R-methyl-sites (i.e. peptide score > 25; PTM localization probability > 0.75; use of ML model to identify true doublets) to produce a list of orthogonally-validated, non-canonical protein methylation sites. We identified a total of 111 methylation sites, 42 of which occurring on non-canonical residues (Figure 44A). Although most proteins only present one type of methylated residue, combinations of up to 5 different methyl-residues appear on a few proteins (Figure 44B), such as TAF15 (experimentally found to be methylated at R, K, S, D and Q), HSPA8 (methylated at K, D, E and N) and EEF1A1 (methylated at K, S, D, Q). The methylated residues that co-occur most frequently appear to be R and S, which coexist on five proteins (HNRNPD, HNRNPK, SFPQ, TAF15 and VIRMA). In addition, we performed functional enrichment analysis on the proteins bearing non-canonical methylated residues; this analysis does not highlight novel functional categories or pathways but suggests that non-canonical methylations might share with R methylation a role in the regulation of splicing and RNA binding (Figure 44C).

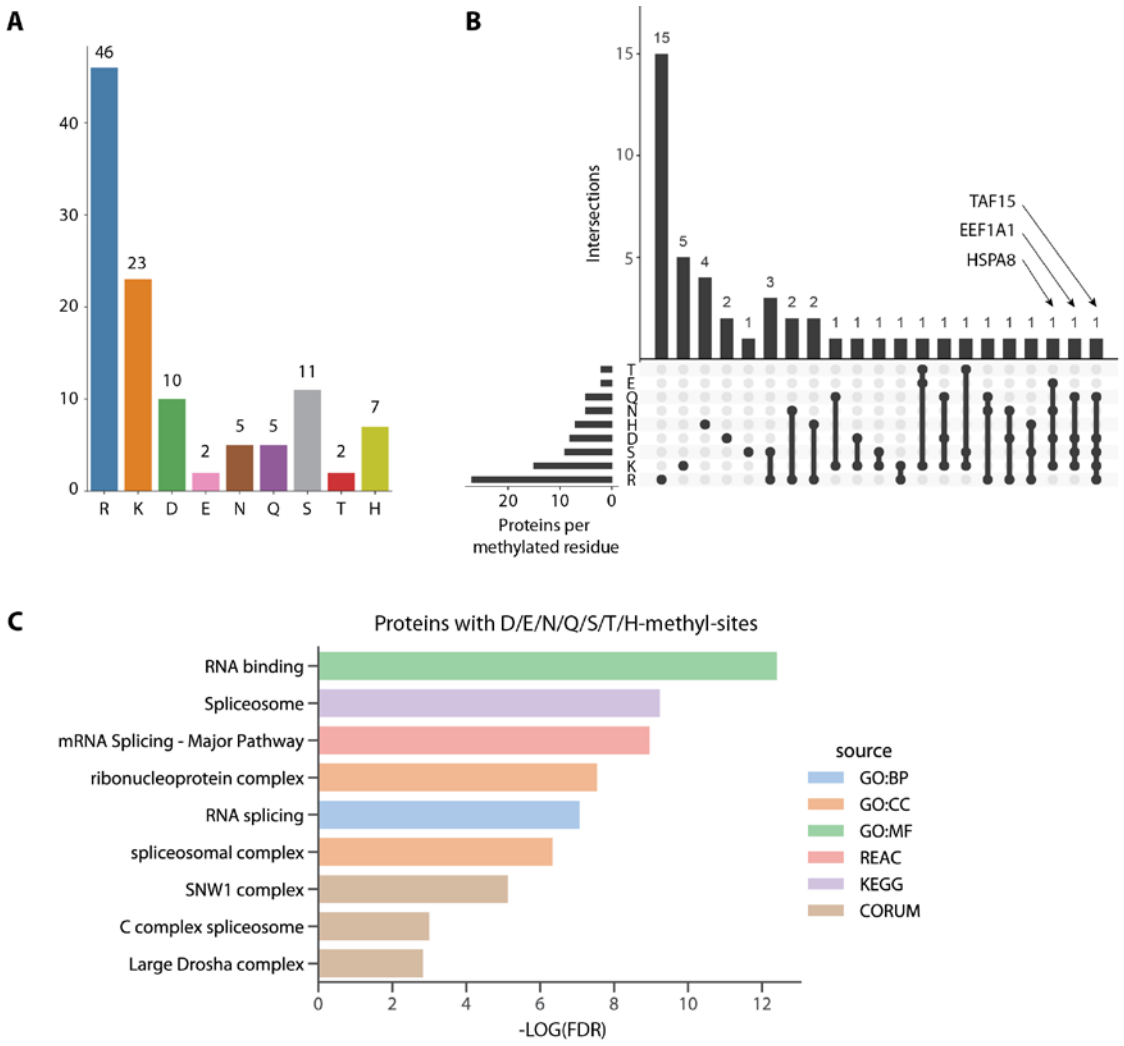


Figure 44. Re-analysis of hmSILAC MS data to identify methylations sites beyond K and R. ***A)*** *Number of methyl-sites detected upon re-analysis of the Input MS data, grouped by residue.* ***B)*** *UpSets plot representation of the co-occurrence of methylated residues on different proteins.* ***C)*** *Gene Ontology terms, pathways and complexes significantly enriched among proteins bearing methylations on non-K/R residues.*

This exploratory analysis also revealed the presence of two unconventional novel methyl-sites on H3, on S28 (H3S28me) and T32 (H3T32me). This initial observation prompted us to focus on non-canonical methylation on histones upon two main considerations: first, histone methylation is widely regarded as a core component of the histone code, hence the detection of novel methyl-sites may lead to a better understanding of this molecular barcode in the gene expression regulation; second, since standard database search engines produce suboptimal results when multiple PTMs are searched in a large protein database, we reasoned that limiting

the analysis to histones would be a reasonable trade-off to expand our methylation search to non-canonical methyl-residues while avoiding the issues that arise from the expansion of search space (i.e. high number of false positives and false negatives).

Thus, we applied an optimized biochemical and analytical pipeline to a set of hmSILAC-labelled histone samples for the in-depth identification of methylation sites on these proteins; from a biochemical point of view, we took advantage of multiple proteases (i.e. Trypsin, LysC, ArgC and LysargiNase) digestion of histones to generate overlapping peptides and maximise protein coverage (Figure 45).

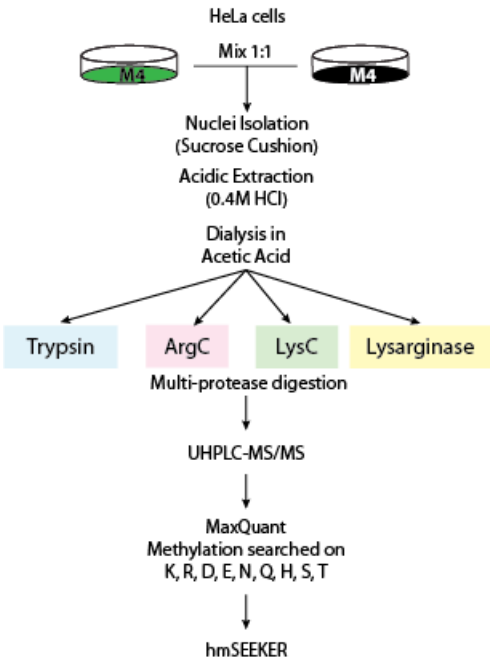


Figure 45. Schematic representation of the histone samples preparation protocol. *Four different proteases were used in parallel to digest histones to generate overlapping peptides and maximize sequence coverage.*

From the computational point of view, the acquired MS data were searched with MaxQuant using a filtered version of the UniProt Human protein database that included only histone proteins, using the same search parameters selected for the global methylation search. The subsequent analysis with hmSEEKER v2.0 for methyl-pair matching and orthogonal validation of novel methyl-sites led to the unambiguous annotation of mono-methylation at the following sites: H3S28, H3T32, H1F0K125, H1F0S131.

Both S28 and T32 are found on the H3 27-40 peptide, which includes two important sites of epigenetic modifications, K27 and K36. Our experimental data showed that the S28me mark could coexist with K27me and T32me, while peptides carrying simultaneously K27me and T32me or bearing S28me or T32me in combination with K36me were not detected (Figure 46).

Additional novel methylations were found on the linker histone variant H1F0 (Figure 47), which is found in fully differentiated cells or cells with have low rates of division [155]; the observation that K125 of H1F0 was annotated as modified by all methylation degrees (mono-, di- and tri-methylated) supports the idea that this is a naturally occurring methylation regulated in vivo, whose function and associated enzymes are however still unknown.

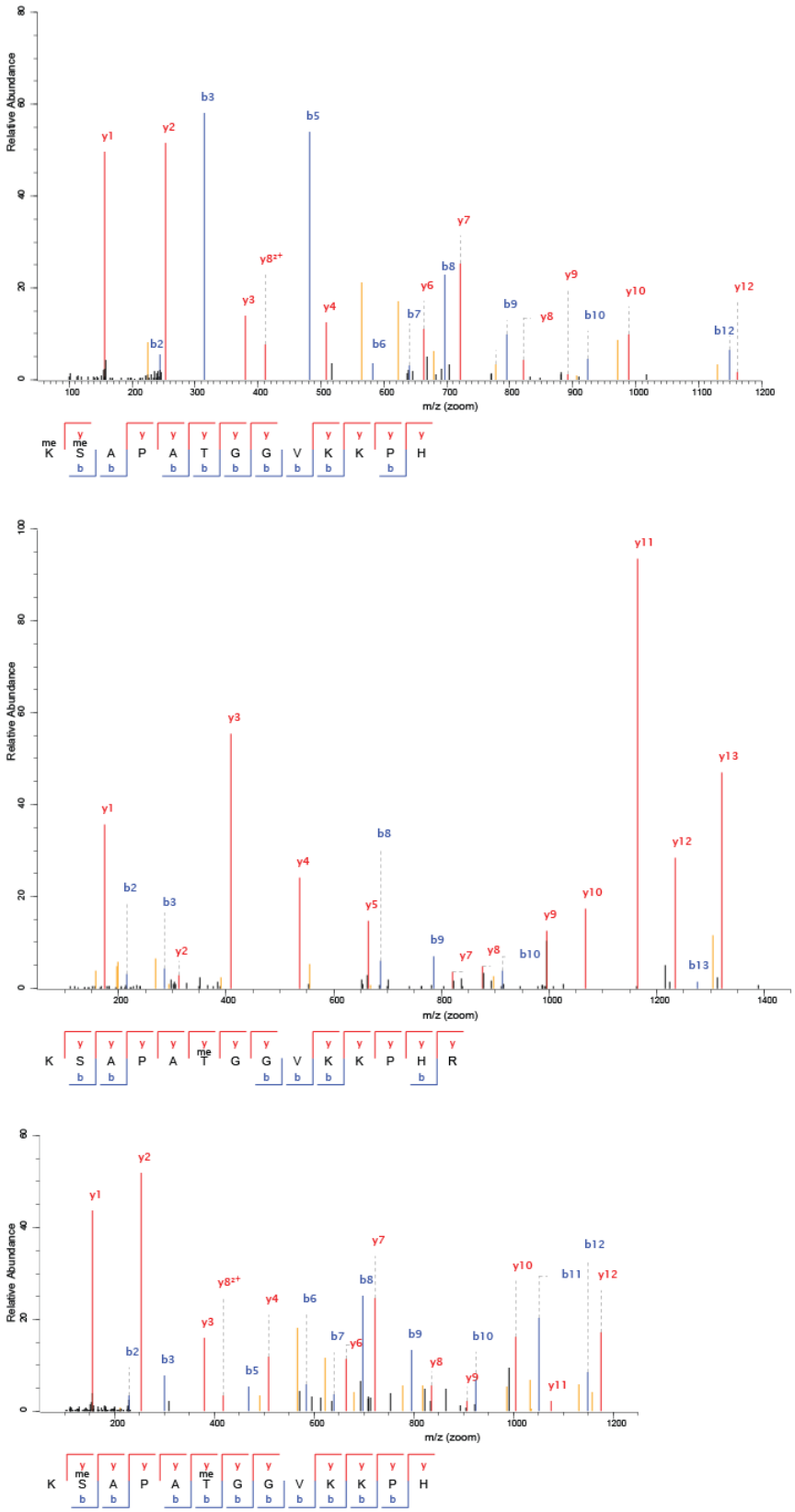


Figure 46. MS/MS spectra of histone H3 27-40 peptide methylated on S28 (top), T32 (middle) or both (bottom).

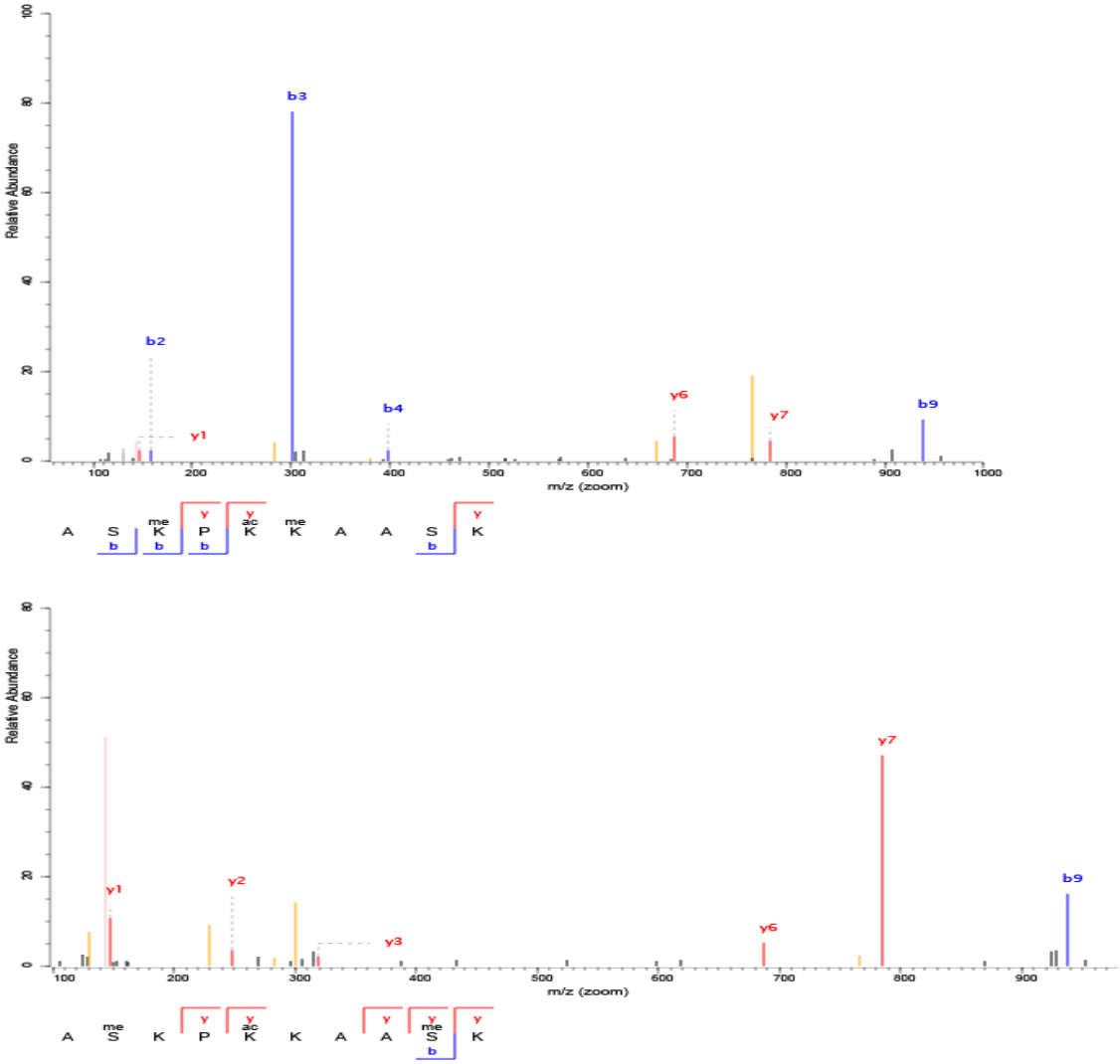


Figure 47. MS/MS spectra of histone H1 peptides. *The fragmentation spectra permit the localization of methylation on K125 (top) and S131 (bottom) of histone H1.*

5.8. Unrestricted analysis of Histone PTMs

5.8.1. MS-based identification of histone PTMs by Open Modification Search

Having expanded the annotation of methyl-sites on non-canonical residues of histones, we attempted to expand our analysis to other hPTMs. To do so, we analysed histone MS data with ionbot [77], a machine learning-based database search engine that predicts the intensities of fragment ions in MS/MS spectra, making it much more accurate in the identification of modified peptides. More importantly, ionbot is able to perform an "open PTM search", a strategy that allows the detection of virtually any mass shift on an amino acid residue.

To expand the annotation of hPTMs, histone samples from MDA-MB-46 triple negative breast cancer (TNBC) cell lines, treated or not with the histone deacetylase inhibitor Panobinostat, were analyzed by MS. Three different sample preparation protocols were used: in-solution ArgC digestion and in-gel Trypsin digestion after chemical derivatization with either D6-acetic anhydride or propionic anhydride and phenyl isocyanate (Pro-PIC). Each experiment was done in triplicate, for a total of 18 samples (Figure 48).

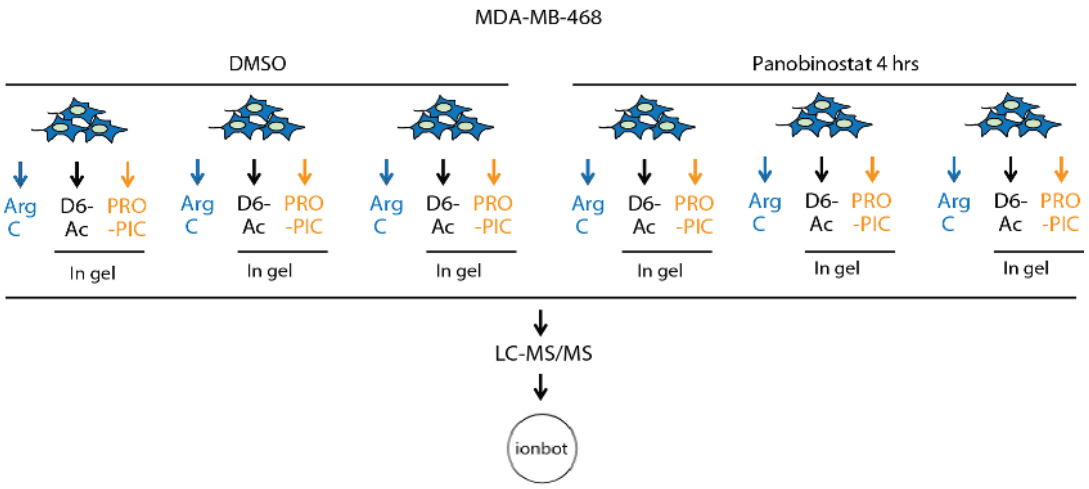


Figure 48. Summary of the biochemical and analytical workflow adopted to analyse histone PTMs with open search.

The ionbot open search identified a total of 180 modifications sites on histones, bearing a total of 299 modifications (Figure 49). Among them, the most abundant were, as expected, K acetylation (n=58), K methylation (n=52) and R methylation (n=28). Notably, we identified all

the well-annotated modification sites on histones H3 (K4, K9, K14, K18, K23, K27, K36, K37, K79) and H4 (K5, K8, K12, K16, K20). Regarding other PTMs, we found (among others) 4 methyl-S-sites, 10 K butyrylation sites, 7 T acetylation sites and 3 S acetylation sites. Although we did not directly detect phosphorylation, we identified a total of 20 dehydrated S/T/Y residues, which could result from the loss of a phosphate group by a phosphorylated S/T/Y residue.

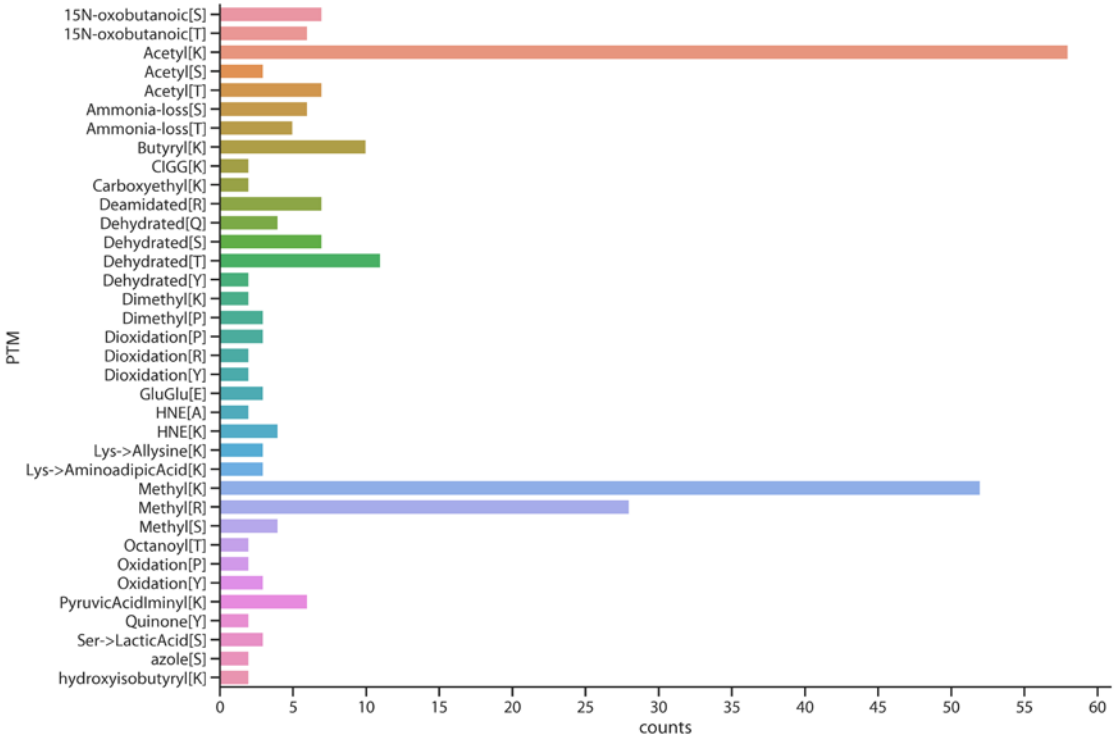


Figure 49. Counts of different PTM sites identified on histones by ionbot.

5.8.2. Histone mutations detected by Open Search

Interestingly, in addition to PTMs, the unbiased MS-based analysis of histones by ionbot identified 350 amino acid substitutions (Figure 50). This finding is interesting because mutant histones (often called “oncohistones”) have been found in a wide range of cancer types, including head and neck cancers, gliomas, sarcomas, and carcinosarcomas. The most well-known examples of these oncohistone mutations are H3K27M, H3K36M and H3G34R/V, which are present in paediatric gliomas and disrupt the deposition of methylation marks on H3 [156]. Notably, these mutations were not detected in the TNBC samples analysed here, suggesting that our analysis was reliable.

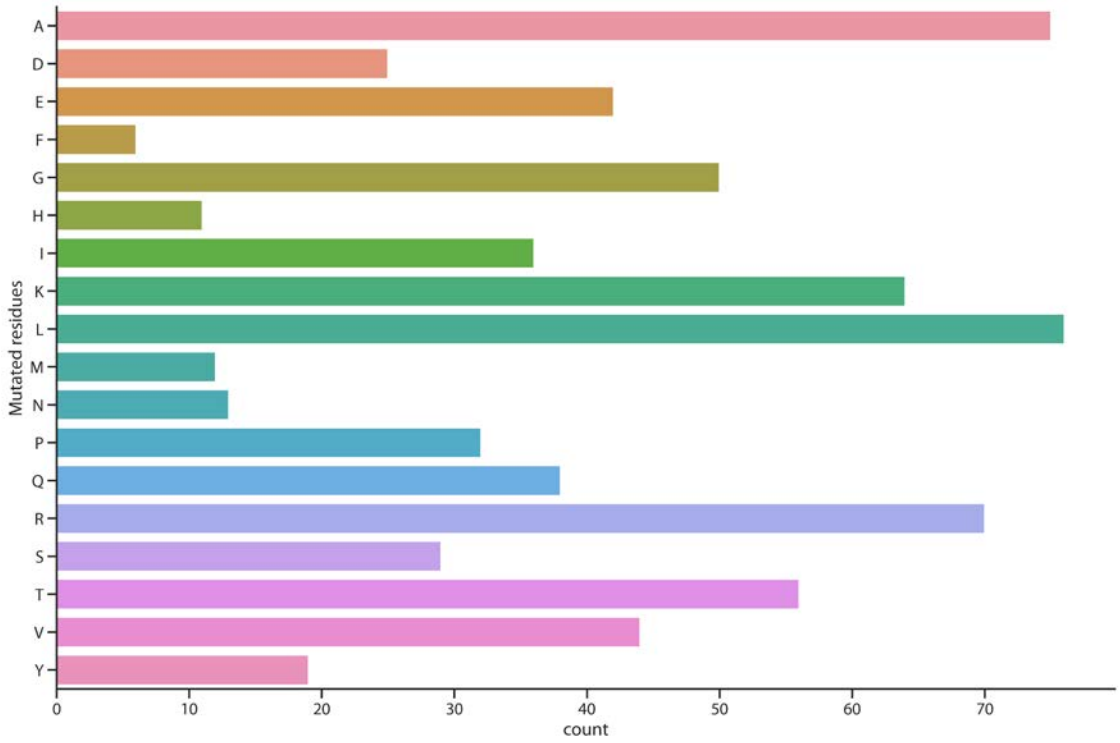


Figure 50. Counts of mutated residues identified on histones by the ionbot open search.

To try and understand their biological relevance, we mapped mutations onto histones 3D crystal structures the web-based tool Mechismo, in order to identify mutations that occurred at sites of protein-protein interaction and predict the effect that these mutations could have on the binding of other proteins. The majority of mutated residues is involved in interactions between histones and the most represented non-histone protein is DAXX (Figure 51), with 131 interactions involving mutated residues (62 with histone H3 and 69 with histone H4): 58 of these were predicted by Mechismo to negatively affect the interaction.

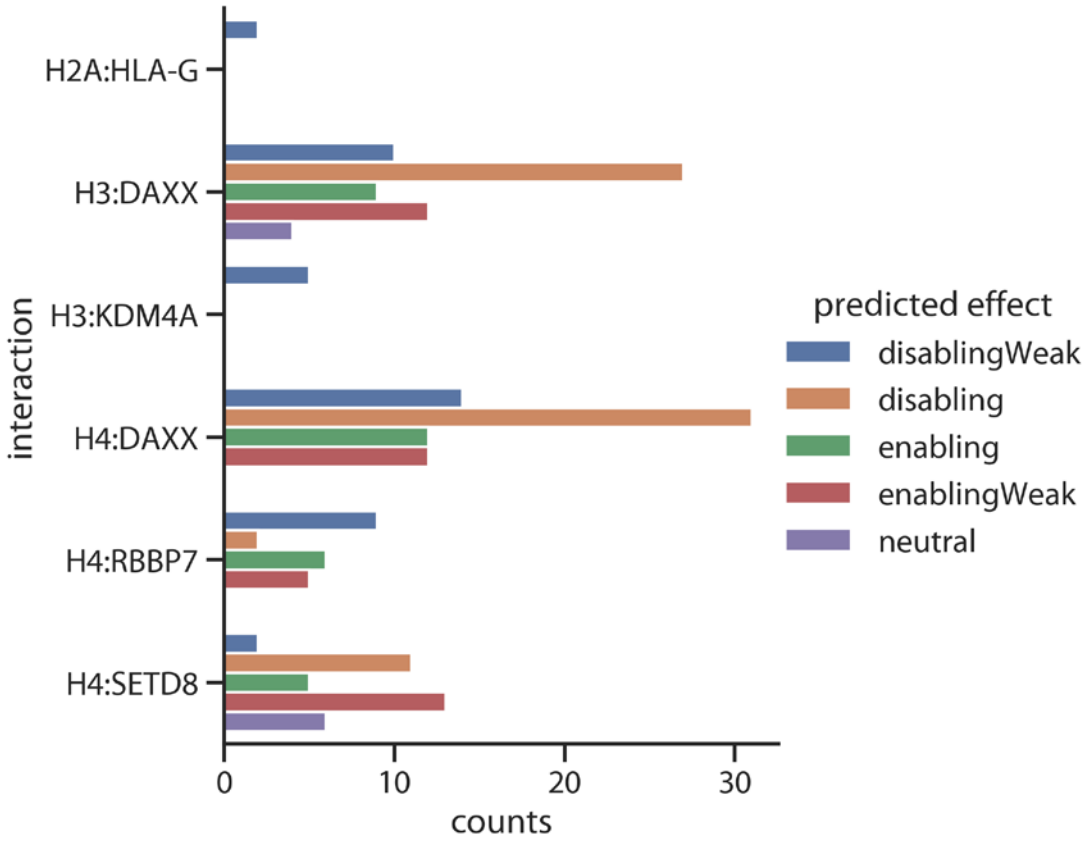


Figure 51. Counts of physical interactions between histone and non-histone proteins that are affected by mutations. *Interactions are classified by Mechismo based on their predicted effect. Several mutation were predicted to impair the interaction between histone H3/H4 and the histone chaperone DAXX.*

6. DISCUSSION

In this study, we employed MS-based proteomics to study protein post-translational modifications (PTMs) from different perspectives. Our initial focus was the analysis of one specific PTM, i.e. methylation, at the level of the whole human proteome. The study of protein methylation by MS is made challenging by the high false discovery rate of this modification, which stems from the fact that amino acid substitutions and chemical artefacts can be erroneously interpreted as methylations. Thus, we first generated a dataset of orthogonally-validated protein methylations through the analysis of heavy methyl SILAC (hmSILAC) MS data with an updated version hmSEEKER, an in-house developed bioinformatics tool. Briefly, doublets of light and heavy methyl-peptides were evaluated by an ML model that was trained on doublets generated by M-containing peptides, which, unlike proper methyl-peptides, can be easily identified by database search engines and therefore provide a reliable positive control to our model. After training the model, we observed that the Retention Time difference (dRT) parameter of the doublets was the least important for discriminating true and false peptide pairs. On the one hand, this is in contrast with the first iteration of hmSEEKER, where dRT and Mass Error (ME) were the main predictors of methyl-peptide doublets, but on the other hand, this choice of features further differentiates our ML model from the one adopted by MethylQuant [157], a similar tool that scores putative hmSILAC doublets based on the isotope distribution and elution profile correlation of the two peaks.

Second, we studied changes in R methylation, which over the last decade was found to be a key player in a wide variety of biological processes. We used a SILAC setup to profile R methylation dynamics in different cell lines in response to genetic modulation of the major type I PRMT, PRMT1, and upon treatment with the chemotherapic drug cisplatin (CDDP), the type I PRMTs inhibitor MS023 and the PRMT5 inhibitor GSK591. Briefly, within the context of CDDP response, we found that this drug triggers the relocalization of PRMT1 to chromatin, leading to a reduction in the methylation levels of cytosolic proteins and the increase of the H4R3me2a histone mark at the promoters of Senescence-Associated Secretory Phenotype (SASP) genes. Activation of the SASP genes induces the arrest of the cell cycle and protects cancer cells from apoptosis, eventually making them resistant to chemotherapy. In addition, we were able to define the sequence specificity of PRMT5, showing that this PRMT preferentially methylates R sites that are located between two G (i.e. GRG motifs), as opposed

to PRMT1, which preferentially methylates RG/RGG motifs. This finding was strengthened by our investigation of R methylation changes in Acute Myeloid Leukemia cells treated with either GSK591 or MS023, which showed that PRMT1 and PRMT5 can methylate different sites on a common set of protein substrates. Finally, by analysing R methylation changes on components of the Large Drosha Complex (LDC) upon knock-down or overexpression of PRMT1, we were able to link R methylation to the process of miRNA biogenesis.

Next, to perform a deeper and more comprehensive analysis of the human methyl-proteome, results from the hmSILAC and SILAC experiments were combined to produce the ProMetheus database (ProMetheusDB), which, at the time of writing, is the largest repository of high-confidence methylation sites. The functional analysis performed on R-methyl-sites revealed that proteins carrying either multiple R-methyl-sites and/or dynamically regulated ones are more strongly interconnected with other proteins in our Reactome-derived network, thus representing potential hubs of protein:protein interactions. Interestingly, the protein clusters emerging from the network analysis suggest a potential role of R methylation in the immune response, as exemplified by the “FCGR-dependent phagocytosis” cluster. While it is known that PRMT1-mediated H4R3me2a promotes the expression of PPARgamma, a transcription factor regulating monocyte differentiation into anti-inflammatory macrophages [158], our results expand the role of R methylation in immunity beyond mere transcriptional regulation, showing that several methyl-proteins are involved in antigen processing and exogenous DNA/RNA sensing pathways (such as CUL1 and G3BP1, respectively). Other clusters that might be connected to this biological process are the ones related to cytoskeleton dynamics and lipid metabolism. Anti-inflammatory macrophages metabolize lipids as a source of energy, while inflammatory ones use fatty acids to produce prostaglandins and leukotrienes, which act as hormone-like signal molecules to regulate the inflammatory response [159,160]. The impact of R methylation on proteins of the cytoskeleton (such as Actin) has been already described in neuronal development, where PRMTs regulate the formation of both the axon and the dendrites [161]; it is possible PRMTs control cytoskeleton dynamics also in macrophages and potentially other cell types.

By mapping the R-methyl-sites annotated in ProMetheusDB onto protein structures, we confirmed that this PTM mostly occurs in unstructured, low-complexity regions, which are characterized by short linear motifs well-established as important mediators of protein:RNA and protein:protein interactions (PPIs) [162]. Our observation that R methylation most

frequently occurs in IDRs, together with the notion that heavily R methylated proteins are also central nodes within functional and physical PPI networks, is indicative of a functional link between R methylation and PPI modulation. Indeed, by inspecting available 3D complexes, we found a small number of sites where R methylation could modulate protein:protein interactions by increasing or reducing the hydrophobicity and hydrogen-bonding capability of R situated at protein interfaces.

Our data and a recent study by Xiang-Bo Wan and co-workers [163] identify NONO R256 and R251, respectively, as methylation sites. Both residues are predicted by Mechismo to interact with apolar residues of PSPC1 (L171 and L222, respectively). Moreover, the amino acid substitution of NONO R256 with an Isoleucine (I) is a mutation associated with colorectal cancer in ActiveDriverDB [164] and is also predicted by Mechismo to enhance the NONO:PSPC1 interaction; similarly, a hypothetical R251I substitution is also predicted to strengthen the binding. Taken together, these observations suggest that PRMT1 may elicit an oncogenic effect (at least in colorectal cancer) by modulating the interaction between NONO and PSPC1 through the methylation of NONO R251 and/or R256.

By contrast, we hypothesize that methylation of SRSF1 on R154 might inhibit the interaction between SRSF1 and SRPK1. SRPK1-mediated phosphorylation of SRSF1 regulates alternative splicing and promotes the expression of protein variants that have anti-apoptotic and pro-angiogenic properties [165]. In this context, Jacky Chi Ki Ngo and collaborators showed that blocking the interaction of SRPK1 and SRSF1 with a PPI inhibitor could reduce SRSF1 phosphorylation and thus suppress angiogenesis [166]. Our data suggest that a similar result could be achieved by regulating SRSF1 methylation levels, although more in-depth mechanistic studies are needed to confirm this hypothesis.

An additional interesting interaction suggested by Mechismo involves the R264 of SAM Synthase (MAT2A), which is located at the interface between the subunits that form the active dimer of the enzyme, where it contacts E57, A281 and K285 of the other MAT2A monomer. This residue also interacts with the substrate ATP and the metal ions serving as cofactors for the reaction. The observation that the enzyme that synthesizes SAM, the biological methyl-group donor, can also be methylated at a site relevant for its catalytic activity may suggest the existence of a feedback loop controlling SAM levels in the cell. It has already been reported that the RNA methyltransferase METTL16 binds and methylates the 3'-UTR region of

MAT2A mRNA to prevent its translation. When SAM levels are low, the mRNA of MAT2A cannot be methylated due to lack of the methyl donor and the protein is translated [167]. Therefore, MAT2A protein methylation at R264 could represent an additional layer of post-translational regulation of the enzyme.

Still, our structural analysis suffers from two limitations: first, as already discussed, methylation often occurs in disordered regions which cannot be studied through classical crystallography approaches; second, not all the structures of proteins that can be crystallized have necessarily been solved. However, thanks to the advancements in the field of AI, the scientific community now has access to AlphaFold, which can predict a protein structure from its primary sequence with unprecedented speed and accuracy [168]. Therefore, as a future perspective, it could be interesting to expand our analysis using AI-predicted protein structures.

Phosphorylation and methylation sites co-localize on disordered SRGG motifs of some proteins, where the two PTMs are mutually exclusive because the presence of negative charges inhibits the binding of PRMTs to their recognition motifs [149]. By overlapping ProMetheusDB with the Phosphosite Plus phosphoproteomics dataset, we confirmed that methylated R-sites are significantly more likely to occur in the proximity of a phospho-S but also expand this observation to phospho-T and -Y residues. Furthermore, the analysis of co-occurrence of these two PTMs in the context of dynamic regulation suggests the existence of a link between the regulation of an R-methyl-site and its proximity to a phosphosite. Acute stimuli such as treatment with MS023 (which inhibits type I PRMTs, with a preference for PRMT1 at the experimental condition used in our studies) cause a change in the methylation of Rs that are close to phosphosites; instead, modulation of PRMT1 expression and the treatment with the chemotherapeutic drug cisplatin affect methyl-Rs that are distant from phosphorylation sites. This result highlights the importance, in the future, of developing experimental pipelines enabling the simultaneous profiling of these two (or even more) PTMs, something that is currently still at a pioneering stage [151]. Our analysis, for instance, focused exclusively on methyl-sites, whereas the datasets of phosphorylation, acetylation, ubiquitination and sumoylation sites were mostly derived from other studies focused on one individual PTM at a time. Profiling differently modified isoforms of a peptide (e.g. unmodified, methylated, phosphorylated, co-modified) within a single experiment would be

very informative to experimentally assess which PTMs can truly co-exist and which are mutually exclusive, both a basal state but also during transition or in response to external cues.

In the second part of this work, we extended the range of our analysis beyond R methylation. This was prompted by the observation that the hmSILAC biochemical and analytical pipeline could be used to annotate methyl-sites also on amino acids that are traditionally excluded from MS-based analysis, such as D, E, N, Q, S, T and H. To be able to search methylation on all these residues, however, we chose to focus our attention on histone proteins, as histone PTMs (hPTMs) are well-known to be involved in crucial biological processes like DNA transcription, replication and repair.

Our analysis of non-canonical histone methylation sites identified two hPTMs of particular interest: H3S28me and H3T32me. Both occur on the peptide 27-40 of histone H3 and attracted our attention for their potential cross-talk with well-known modifications on K27 and K36. For instance, methylation of H3K27 is a repressive mark [169], whereas methylation of H3K36 plays a role in transcriptional regulation, whereby H3K36me2 counteracts gene silencing by blocking the recruitment of PRC2 complexes; however, when genes are transcribed, this mark is replaced by H3K36me3 to prevent transcription initiation from intragenic regions [170]. Our experimental data showed that the S28me mark could coexist with K27me and T32me, while peptides carrying simultaneously K27me and T32me, or bearing S28me/T32me in combination with K36me were not detected. A more in-depth analysis of this peptide and its numerous differentially modified isoforms is needed to corroborate this observation: measuring the relative abundance of the K27me/S28me and S28me/T32me peptide isoforms could allow developing some hypotheses on their biological role and cross-talk with other epigenetic marks. Similarly, it would be interesting to confirm whether S28me/T32me are mutually exclusive with K36 methylation or to find how these novel methyl marks are associated with neighbouring acetylation and phosphorylation.

We argue that our analysis of non-canonical methylation sites could be linked to our previous analysis on PTMs cross-talk. A recent paper from the Eyers' group [171] explored the field of non-canonical phosphorylations and identified several phosphosites on R, K, D, E, H and C residues. Interestingly, these residues overlap with the putative non-canonical methyl-residues we have investigated here: it is, therefore, possible to envisage that the cross-talk of methylation and phosphorylation is not limited to proximal sites, but that may also occur

through the physical competition for the same substrates/residues. Two possible case-studies emerging here are H3S28 and R55 of Keratin type I cytoskeletal 18 (KRT18R55). H3S28 is a known phosphorylation site that we identify as methylated in our hmSILAC histone dataset. Phosphorylation of H3S28 can occur in response to extracellular stress and is proposed to override repressive epigenetic marks to temporarily express genes that would normally be silenced [172]. A reasonable hypothesis is that methylation of H3S28 may cooperate with methylation of H3K27 in silencing genes [169], which would explain why we did not detect H3S28 in combination with the aforementioned active transcription marks H3K36me2/me3 and H3K27ac [173]. As a future perspective, since H3S28ph is also necessary for chromatin condensation [174], we could investigate how this residue is modified specifically during mitosis.

KRT18R55 is a methyl-site that was first identified in [175] and orthogonally validated in our hmSILAC experiments. The site is also present in the list of non-canonical phosphosites published by Eyers and co-workers, thus could be another example of competition between PTMs. Unfortunately, there is not enough data in the literature to make a hypothesis on the function of this modification site.

In the final section of this work, we explored the possibility of further expanding the hPTMs analysis by making use of open modification search tools. In fact, one limitation of proteomics studies is the ability to profile many PTMs at the same time; in the context of histones, this is critical because histones are hyper-modified proteins and their PTMs show extensive cross-talk. As such, it is not the individual histone modification marks but the way they combine with each other that defines the functional state of chromatin regions.

Histone MS data were analyzed with ionbot, a novel search engine developed by the group of Lennart Martens at Ghent University [77]. The analysis re-identified all the most well-known modifications sites on histones H3 and H4 but also highlighted the presence of many low-abundance PTMs such as short-chain acylations, which are closely linked to cell metabolism and could represent interesting biomarkers or therapeutic targets.

Intriguingly, the open search identified not only PTMs but also several amino acid substitutions, of which over 100 are predicted by Mechismo to destabilize the interaction of DAXX with histones H3 and H4. We believe this finding to be noteworthy because DAXX is a histone chaperone with a wide range of biological functions and can act as either an

oncogene or a tumour suppressor; while its overexpression has been observed in several cancers [176], its depletion can activate transposable elements (such as endogenous retroviruses) and lead to genomic instability [177]. This behaviour is influenced by its interaction with the H3.3:H4 histone dimer, which stabilizes DAXX: in Hoelper et al. [177], the authors showed that the reduction of DAXX levels can be caused by its inability to bind histone H3.3. and performed point mutations on DAXX yet did not investigate histone mutations that could also disrupt the interaction. In fact, mutations of H3.3 are much more common in cancers compared to those of DAXX [156]. Therefore, investigating mutations that prevent the formation of the DAXX:H3.3:H4 complex could lead to the discovery of novel cancer biomarkers.

Some major points remain to be addressed within the field of PTMs. In the methyl-proteomics field, there is a need for more efficient workflows that allow to annotate R-methyl-proteomes through sensibly smaller scale experiments, as the prerequisite for the investigation of this PTM in more relevant model systems, such as primary cells, tissues, organoids, similarly to what was made possible in the phospho-proteomics field [178]. The implementation of Data-Independent Acquisition (DIA) methods could also prove beneficial by reducing the identification bias towards the most abundant proteins and thus increase the depth of coverage of the methyl-proteome [179]. Better strategies for the enrichment of di-methyl-R-peptides are also needed to address the bias introduced by the use of the CST anti-MMA kit. Although quantification of intracellular methyl-R derived from degradation of methylated proteins showed that ADMA is the most abundant of the three [180], most MS-based methyl-proteomics studies have so far reported a significantly larger number of MMA sites than ADMA/SDMA sites. Our group already discussed this issue, suggesting that the anti-pan-MMA antibodies traditionally used to enrich methyl-peptides outperform anti-pan-ADMA and SDMA ones thus generating a bias in methyl-proteomics studies [34]. Along the same line, there is a strong need for efficient antibodies for the enrichment of K-methyl-peptides to map the non-histone K-methyl-proteome. The availability of such reagents would allow researchers to finally address the question whether K methylation is widespread beyond histone, similarly to K acetylation and R methylation, regulating cellular processes beyond chromatin-based transcriptional regulation.

Besides the annotation of the steady-state methyl-proteome, we recognize a need to accelerate the acquisition of dynamic data. The triple SILAC strategy we adopted for some experiments

is overall laborious and limited in its multiplexing capabilities. In the near future it would be useful to assess the feasibility of applying isobaric mass tags (TMT) to MS-based methyl-proteomics profiling, to overcome this limitation and perform multiplexed experiments where methylation changes can be profiled across multiple experimental conditions.

From the analytical point of view, some future developments can also be conceived: first, spectral library searching [181] is faster than conventional database searching (which can take several weeks when non-canonical methyl-sites are considered), however its application to methyl-proteomics was limited by the low number of high-confidence methyl-peptides spectra available. Within this context, the MS/MS spectra of methyl-peptides that were orthogonally validated by hmSILAC could be used to build a spectral library to be used as reference, thus speeding up the analysis of new data. Second, it is crucial to separately analyse ADMA and SDMA, which can be distinguished by searching for their diagnostic neutral loss ions within the MS/MS spectra [31,123]. However, at the moment, a systematic method for the automatic annotation of these characteristic neutral losses is still lacking; while general purpose tools can be adapted to perform this task [182], the aforementioned lack of high-confidence spectra makes this task difficult. For these reasons, it is still recommended to visually inspect the MS/MS spectra of di-methyl-R-peptides, which is impractical in the case of large datasets such as ProMetheusDB.

Finally, our unbiased analysis of hPTMs is one of the earliest applications of an open search algorithm to MS data as complex as those obtained from hyper-modified histone proteins. Because open search produces large amounts of data, it is indispensable to develop an effective way to filter the peptide identifications and narrow down few hPTMs that can later be validated through in vitro or in vivo experiments. At the moment, we are experimenting with intersecting acetylated and methylated peptides identified by open search with those identified by MaxQuant to produce a set of positive controls, then using these control peptides to train machine learning-based methods that select PSMs that have the highest probability of being correct. As a future perspective, we might attempt to score PSMs based on the **“**generating function approach” rather than the traditional target-decoy approach, which is used by MaxQuant and ionbot. The generating function approach was proposed by Kim et al. in 2008 [183] and is supposed to eliminate the requirement of a “decoy” database by calculating the statistical significance of individual PSMs instead of the global FDR of the results. This approach has been already implemented by MS-GF+ [73] and StarGF [184], a

database search and a spectra search engine respectively, both of which outperform state-of-the-art tools in their respective category of search engines.

Overall, we believe that, despite its current limitations, our unbiased analysis of hPTMs could serve as a starting point to develop and optimize ad hoc bioinformatics methods to study the histone code more in depth. Moreover, it is likely that advances in the identification of hyper-modified histone peptides will also benefit the open search of PTMs on non-histone proteins: from this point of view, it would be interesting to carry out a systematic analysis of non-canonical methylations and phosphorylations with open search tools and compare its results to those obtained by traditional methods.

7. Authored articles

- E. Massignani, A. Cuomo, D. Musiani, S. Jammula, G. Pavesi, T. Bonaldi, **hmSEEKER: Identification of hmSILAC Doublets in MaxQuant Output Data.** Proteomics. 19, e1800300 (2019).

- J. Y. Fong, L. Pignata, P.-A. Goy, K. C. Kawabata, S. C.-W. Lee, C. M. Koh, D. Musiani, E. Massignani, A. G. Kotini, A. Penson, C. M. Wun, Y. Shen, M. Schwarz, D. H. Low, A. Rialdi, M. Ki, H. Wollmann, S. Mzoughi, F. Gay, C. Thompson, T. Hart, O. Barbash, G. M. Luciani, M. M. Szewczyk, B. J. Wouters, R. Delwel, E. P. Papapetrou, D. Barsyte-Lovejoy, C. H. Arrowsmith, M. D. Minden, J. Jin, A. Melnick, T. Bonaldi, O. Abdel-Wahab, E. Guccione, **Therapeutic Targeting of RNA Splicing Catalysis through Inhibition of Protein Arginine Methylation.** Cancer Cell. 36, 194-209.e9 (2019).

- D. Musiani, J. Bok, E. Massignani, L. Wu, T. Tabaglio, M. R. Ippolito, A. Cuomo, U. Ozbek, H. Zorgati, U. Ghoshdastider, R. C. Robinson, E. Guccione, T. Bonaldi, **Proteomics profiling of arginine methylation defines PRMT5 substrate specificity.** Sci Signal. 12, eaat8388 (2019).

- D. Musiani, R. Giambruno, E. Massignani, M. R. Ippolito, M. Maniaci, S. Jammula, D. Manganaro, A. Cuomo, L. Nicosia, D. Pasini, T. Bonaldi, **PRMT1 Is Recruited via DNA-PK to Chromatin Where It Sustains the Senescence-Associated Secretory Phenotype in Response to Cisplatin.** Cell Rep. 30, 1208-1222.e9 (2020).

- V. Spadotto, R. Giambruno, E. Massignani, M. Mihailovich, M. Maniaci, F. Patuzzo, F. Ghini, F. Nicassio, T. Bonaldi, **PRMT1-mediated methylation of the microprocessor-associated proteins regulates microRNA biogenesis.** Nucleic Acids Res. 48, 96–115 (2020).

- D. Musiani, E. Massignani, A. Cuomo, A. Yadav, T. Bonaldi, **Biochemical and Computational Approaches for the Large-Scale Analysis of Protein Arginine Methylation by Mass Spectrometry.** Curr Protein Pept Sci. 21, 725–739 (2020).

- M. Maniaci, F. L. Boffo, E. Massignani, T. Bonaldi, **Systematic Analysis of the Impact of R-Methylation on RBPs-RNA Interactions: A Proteomic Approach.** Front. Mol. Biosci. 8 (2021), doi:10.3389/fmolb.2021.688973.

- Enrico Massignani, Roberto Giambruno, Marianna Maniaci, Luciano Nicosia, Avinash Yadav, Alessandro Cuomo, Francesco Raimondi, Tiziana Bonaldi, **ProMetheusDB: an in-depth analysis of the high-quality human methyl-proteome.** *Paper under submission.*

8. Tables

Cell line	Fraction	Enrichment	Antibody	AA specificity	PTM specificity	Separation Method	Protease
HeLa	Ncl	Protein IP	APO 9328PU-N	K	Pan-methyl	SDS-PAGE	Trypsin
HeLa	Cyt	Protein IP	APO 9328PU-N	K	Pan-methyl	SDS-PAGE	Trypsin
HeLa	Ncl	Protein IP	LS-C60093/28979	K	Pan-methyl	SDS-PAGE	Trypsin
HeLa	Cyt	Protein IP	LS-C60093/28979	K	Pan-methyl	SDS-PAGE	Trypsin
HeLa	Ncl	Protein IP	Ab23366	K	me1, me2	SDS-PAGE	Trypsin
HeLa	Cyt	Protein IP	Ab23366	K	me1, me2	SDS-PAGE	Trypsin
HeLa	Ncl	Protein IP	ICP0501/ICP0601	K	me1, me2, me3	SDS-PAGE	Trypsin
HeLa	Cyt	Protein IP	ICP0501/ICP0601	K	me1, me2, me3	SDS-PAGE	Trypsin
HeLa	Ncl	Protein IP	Ab76118	K	me3	SDS-PAGE	Trypsin
HeLa	Cyt	Protein IP	Ab76118	K	me3	SDS-PAGE	Trypsin
HeLa	Ncl	Protein IP	D5A12	R	MMA (R*GG)	SDS-PAGE	Trypsin
HeLa	Cyt	Protein IP	D5A12	R	MMA (R*GG)	SDS-PAGE	Trypsin
HeLa	Ncl	Protein IP	Me-R4-100	R	MMA	SDS-PAGE	Trypsin
HeLa	Cyt	Protein IP	Me-R4-100	R	MMA	SDS-PAGE	Trypsin
HeLa	Ncl	Protein IP	Rabbit 2 (SYM)	R	SDMA	SDS-PAGE	Trypsin
HeLa	Cyt	Protein IP	Rabbit 2 (SYM)	R	SDMA	SDS-PAGE	Trypsin
HeLa	Ncl	Protein IP	SYM10	R	SDMA	SDS-PAGE	Trypsin
HeLa	Cyt	Protein IP	SYM10	R	SDMA	SDS-PAGE	Trypsin
HeLa	Ncl	Protein IP	Rabbit 1 (ASYM)	R	ADMA	SDS-PAGE	Trypsin
HeLa	Cyt	Protein IP	Rabbit 1 (ASYM)	R	ADMA	SDS-PAGE	Trypsin
HeLa	Ncl	Protein IP	ASYM24	R	ADMA	SDS-PAGE	Trypsin
HeLa	Cyt	Protein IP	ASYM24	R	ADMA	SDS-PAGE	Trypsin
HeLa	Ncl	-	-	-	-	SDS-PAGE	Trypsin
HeLa	Cyt	-	-	-	-	SDS-PAGE	Trypsin
HeLa	Ncl	-	-	-	-	SDS-PAGE	Trypsin
HeLa	Cyt	-	-	-	-	SDS-PAGE	Trypsin
HeLa	Ncl	-	-	-	-	IEF	Trypsin
HeLa	Cyt	-	-	-	-	IEF	Trypsin
HeLa	Ncl	-	-	-	-	IEF	Trypsin
HeLa	Cyt	-	-	-	-	IEF	Trypsin
HeLa	Ncl	-	-	-	-	IEF	Trypsin
HeLa	Cyt	-	-	-	-	IEF	Trypsin
SK-OV-3	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13563, PTMScan Kit #13474	R	MMA, SDMA, ADMA	-	Trypsin
HeLa	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13563, PTMScan Kit #13474	R	MMA, SDMA, ADMA	-	Trypsin
NB4	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13474	R	MMA, ADMA	HpH-RP (before enrichment)	Trypsin
NB4	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13474	R	MMA, ADMA	HpH-RP (before enrichment)	Lysarginase
HeLa	WCE	Protein IP	DGCR8 (Abcam ab90579), FUS (Bethyl A300-293A), DDX5 (Abcam ab126730), Drosha (Santa Cruz sc-33778),	-		SDS-PAGE	Trypsin
U2OS	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13563, PTMScan Kit #13474	R	MMA, SDMA, ADMA	HpH-RP (before enrichment)	Trypsin
SK-OV-3	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13563, PTMScan Kit #13474	R	MMA, SDMA, ADMA	-	Trypsin
HeLa	WCE	-	-	-	-	-	Trypsin
NB4	WCE	-	-	-	-	-	Trypsin
NB4	WCE	-	-	-	-	-	Lysarginase
HeLa	WCE	-	-	-	-	-	Trypsin
U2OS	WCE	-	-	-	-	SDS-PAGE	Trypsin

Table 1. Summary of hmSILAC experiments analyzed. *MMA = Mono-methyl-arginine; SDMA = Symmetric di-methyl-arginine; ADMA = Asymmetric di-methyl-arginine.*

Cell line	Fraction	Enrichment	Antibody	AA specificity	PTM specificity	Separation method	Protease	Stimulus
NB4	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13474	R	MMA, SDMA	HpH-RP	Trypsin	GSK591 -vs- UT
NB4	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13563, PTMScan Kit #13474	R	MMA, SDMA, ADMA	HpH-RP	Trypsin	MS023 -vs- UT
SK-OV-3	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13474	R	MMA, ADMA	HpH-RP	Trypsin	CDDP -vs- UT / shPRMT1+CDDP -vs- UT / shPRMT1+CDDP -vs- CDDP
SK-OV-3	Nuclear	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13474	R	MMA, ADMA	HpH-RP	Trypsin	CDDP -vs- UT / MS023+CDDP -vs- UT / MS023+CDDP -vs- CDDP
SK-OV-3	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13563, PTMScan Kit #13474	R	MMA, SDMA, ADMA	-	Trypsin	CDDP -vs- UT
SK-OV-3	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13563, PTMScan Kit #13474	R	MMA, SDMA, ADMA	-	Trypsin	shPRMT1 -vs- UT
HeLa	WCE	Peptide IP	PTMScan Kit #12235, PTMScan Kit #13563	R	MMA, SDMA	HpH-RP	Trypsin	GSK591 -vs- UT
HeLa	Nuclear	Protein IP	DGCR8 (Abcam ab90579), FUS (Bethyl A300-293A), DDX5 (Abcam ab126730), Drosha (Santa Cruz sc-33778),	-	-	SDS-PAGE	Trypsin	shPRMT1 -vs- UT / oePRMT1 -vs- UT

Table 2. Summary of SILAC experiments analyzed. *MMA = Mono-methyl-arginine; SDMA = Symmetric di-methyl-arginine; ADMA = Asymmetric di-methyl-arginine.*

9. References

1. Jensen ON (2006) Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol 7: 391-403.
2. Lothrop AP, Torres MP, Fuchs SM (2013) Deciphering post-translational modification codes. FEBS Lett 587: 1247-1257.
3. Wu Z, Huang R, Yuan L (2019) Crosstalk of intracellular post-translational modifications in cancer. Arch Biochem Biophys 676: 108138.
4. Karve TM, Cheema AK (2011) Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease. J Amino Acids 2011: 207691.
5. Tan YT, Lin JF, Li T, Li JJ, Xu RH, et al. (2021) LncRNA-mediated posttranslational modifications and reprogramming of energy metabolism in cancer. Cancer Commun (Lond) 41: 109-120.
6. Xu H, Wang Y, Lin S, Deng W, Peng D, et al. (2018) PTMD: A Database of Human Disease-associated Post-translational Modifications. Genomics Proteomics Bioinformatics 16: 244-251.
7. Diallo I, Seve M, Cunin V, Minassian F, Poisson JF, et al. (2019) Current trends in protein acetylation analysis. Expert Rev Proteomics 16: 139-159.
8. Hyun K, Jeon J, Park K, Kim J (2017) Writing, erasing and reading histone lysine methylations. Exp Mol Med 49: e324.
9. Musiani D, Giambruno R, Massignani E, Ippolito MR, Maniaci M, et al. (2020) PRMT1 Is Recruited via DNA-PK to Chromatin Where It Sustains the Senescence-Associated Secretory Phenotype in Response to Cisplatin. Cell Rep 30: 1208-1222 e1209.
10. Fong JY, Pignata L, Goy PA, Kawabata KC, Lee SC, et al. (2019) Therapeutic Targeting of RNA Splicing Catalysis through Inhibition of Protein Arginine Methylation. Cancer Cell 36: 194-209 e199.
11. Geoghegan V, Guo A, Trudgian D, Thomas B, Acuto O (2015) Comprehensive identification of arginine methylation in primary T cells reveals regulatory roles in cell signalling. Nat Commun 6: 6758.
12. Afjehi-Sadat L, Garcia BA (2013) Comprehending dynamic protein methylation with mass spectrometry. Curr Opin Chem Biol 17: 12-19.

13. Tessarz P, Santos-Rosa H, Robson SC, Sylvestersen KB, Nelson CJ, et al. (2014) Glutamine methylation in histone H2A is an RNA-polymerase-I-dedicated modification. Nature 505: 564-568.
14. Wilkinson AW, Diep J, Dai S, Liu S, Ooi YS, et al. (2019) SETD3 is an actin histidine methyltransferase that prevents primary dystocia. Nature 565: 372-376.
15. Biterge B, Richter F, Mittler G, Schneider R (2014) Methylation of histone H4 at aspartate 24 by protein L-isoaspartate O-methyltransferase (PCMT1) links histone modifications with protein homeostasis. Sci Rep 4: 6674.
16. Kapell S, Jakobsson ME (2021) Large-scale identification of protein histidine methylation in human cells. NAR Genom Bioinform 3: lqab045.
17. Murray B, Antonyuk SV, Marina A, Van Liempd SM, Lu SC, et al. (2014) Structure and function study of the complex that synthesizes S-adenosylmethionine. IUCrJ 1: 240-249.
18. Lorton BM, Shechter D (2019) Cellular consequences of arginine methylation. Cell Mol Life Sci 76: 2933-2956.
19. Yamaguchi A, Kitajo K (2012) The effect of PRMT1-mediated arginine methylation on the subcellular localization, stress granules, and detergent-insoluble aggregates of FUS/TLS. PLoS One 7: e49267.
20. Stopa N, Krebs JE, Shechter D (2015) The PRMT5 arginine methyltransferase: many roles in development, cancer and beyond. Cell Mol Life Sci 72: 2041-2059.
21. Yang Y, Bedford MT (2013) Protein arginine methyltransferases and cancer. Nat Rev Cancer 13: 37-50.
22. Chang B, Chen Y, Zhao Y, Bruick RK (2007) JMJD6 is a histone arginine demethylase. Science 318: 444-447.
23. Zhang J, Jing L, Li M, He L, Guo Z (2019) Regulation of histone arginine methylation/demethylation by methylase and demethylase (Review). Mol Med Rep 19: 3963-3971.
24. Zheng Q, Osunsade A, David Y (2020) Protein arginine deiminase 4 antagonizes methylglyoxal-induced histone glycation. Nat Commun 11: 3241.
25. Wu Q, Schapira M, Arrowsmith CH, Barsyte-Lovejoy D (2021) Protein arginine methylation: from enigmatic functions to therapeutic targeting. Nat Rev Drug Discov 20: 509-530.

26. Luo M (2018) Chemical and Biochemical Perspectives of Protein Lysine Methylation. Chem Rev 118: 6656-6705.

27. Wysocka J, Allis CD, Coonrod S (2006) Histone arginine methylation and its dynamic regulation. Front Biosci 11: 344-355.

28. Vadnais C, Chen R, Fraszczak J, Yu Z, Boulais J, et al. (2018) GFI1 facilitates efficient DNA repair by regulating PRMT1 dependent methylation of MRE11 and 53BP1. Nat Commun 9: 1418.

29. Sanchez G, Bondy-Chorney E, Laframboise J, Paris G, Didillon A, et al. (2016) A novel role for CARM1 in promoting nonsense-mediated mRNA decay: potential implications for spinal muscular atrophy. Nucleic Acids Res 44: 2661-2676.

30. Szewczyk MM, Ishikawa Y, Organ S, Sakai N, Li F, et al. (2020) Pharmacological inhibition of PRMT7 links arginine monomethylation to the cellular stress response. Nat Commun 11: 2396.

31. Musiani D, Bok J, Massignani E, Wu L, Tabaglio T, et al. (2019) Proteomics profiling of arginine methylation defines PRMT5 substrate specificity. Sci Signal 12.

32. Cheng D, Cote J, Shaaban S, Bedford MT (2007) The arginine methyltransferase CARM1 regulates the coupling of transcription and mRNA processing. Mol Cell 25: 71-83.

33. Guccione E, Richard S (2019) The regulation, functions and clinical relevance of arginine methylation. Nat Rev Mol Cell Biol 20: 642-657.

34. Spadotto V, Giambruno R, Massignani E, Mihailovich M, Maniaci M, et al. (2020) PRMT1-mediated methylation of the microprocessor-associated proteins regulates microRNA biogenesis. Nucleic Acids Res 48: 96-115.

35. Maniaci M, Boffo FL, Massignani E, Bonaldi T (2021) Systematic Analysis of the Impact of R-Methylation on RBPs-RNA Interactions: A Proteomic Approach. Front Mol Biosci 8.

36. Hwang JW, Cho Y, Bae GU, Kim SN, Kim YK (2021) Protein arginine methyltransferases: promising targets for cancer therapy. Exp Mol Med 53: 788-808.

37. Portela A, Esteller M (2010) Epigenetic modifications and human disease. Nat Biotechnol 28: 1057-1068.

38. Couto ESA, Wu CY, Citadin CT, Clemons GA, Possoit HE, et al. (2020) Protein Arginine Methyltransferases in Cardiovascular and Neuronal Function. Mol Neurobiol 57: 1716-1732.

39. So HK, Kim S, Kang JS, Lee SJ (2021) Role of Protein Arginine Methyltransferases and Inflammation in Muscle Pathophysiology. Front Physiol 12: 712389.

40. Smith E, Zhou W, Shindiapina P, Sif S, Li C, et al. (2018) Recent advances in targeting protein arginine methyltransferase enzymes in cancer therapy. Expert Opin Ther Targets 22: 527-545.

41. Yi SJ, Kim K (2020) New Insights into the Role of Histone Changes in Aging. Int J Mol Sci 21.

42. Stachecka J, Kolodziejski PA, Noak M, Szczerbal I (2021) Alteration of active and repressive histone marks during adipogenic differentiation of porcine mesenchymal stem cells. Sci Rep 11: 1325.

43. Duan Q, Chen H, Costa M, Dai W (2008) Phosphorylation of H3S10 blocks the access of H3K9 by specific antibodies and histone methyltransferase. Implication in regulating chromatin dynamics and epigenetic inheritance during mitosis. J Biol Chem 283: 33585-33590.

44. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, et al. (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40: 897-903.

45. Nakanishi S, Lee JS, Gardner KE, Gardner JM, Takahashi YH, et al. (2009) Histone H2BK123 monoubiquitination is the critical determinant for H3K4 and H3K79 trimethylation by COMPASS and Dot1. J Cell Biol 186: 371-377.

46. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315-326.

47. Chan JC, Maze I (2020) Nothing Is Yet Set in (Hi)stone: Novel Post-Translational Modifications Regulating Chromatin Function. Trends Biochem Sci 45: 829-844.

48. Zheng Q, Maksimovic I, Upad A, David Y (2020) Non-enzymatic covalent modifications: a new link between metabolism and epigenetics. Protein Cell 11: 401-416.

49. Sadakierska-Chudy A, Filip M (2015) A comprehensive view of the epigenetic landscape. Part II: Histone post-translational modification, nucleosome level, and chromatin regulation by ncRNAs. Neurotox Res 27: 172-197.

50. Fraga MF, Ballestar E, Villar-Garea A, Boix-Chornet M, Espada J, et al. (2005) Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. Nat Genet 37: 391-400.

51. Hamamoto R, Furukawa Y, Morita M, Iimura Y, Silva FP, et al. (2004) SMYD3 encodes a histone methyltransferase involved in the proliferation of cancer cells. Nat Cell Biol 6: 731-740.

52. Kondo Y, Shen L, Suzuki S, Kurokawa T, Masuko K, et al. (2007) Alterations of DNA methylation and histone modifications contribute to gene silencing in hepatocellular carcinomas. Hepatol Res 37: 974-983.

53. Vire E, Brenner C, Deplus R, Blanchon L, Fraga M, et al. (2006) The Polycomb group protein EZH2 directly controls DNA methylation. Nature 439: 871-874.

54. Urdinguio RG, Sanchez-Mut JV, Esteller M (2009) Epigenetic mechanisms in neurological diseases: genes, syndromes, and therapies. Lancet Neurol 8: 1056-1072.

55. Seligson DB, Horvath S, McBrian MA, Mah V, Yu H, et al. (2009) Global levels of histone modifications predict prognosis in different cancers. Am J Pathol 174: 1619-1628.

56. de Oliveira DT, Guerra-Sa R (2020) Uncovering epigenetic landscape: a new path for biomarkers identification and drug development. Mol Biol Rep 47: 9097-9122.

57. Nitsch S, Zorro Shahidian L, Schneider R (2021) Histone acylations and chromatin dynamics: concepts, challenges, and links to metabolism. EMBO Rep 22: e52774.

58. Tsiatsiani L, Heck AJ (2015) Proteomics beyond trypsin. FEBS J 282: 2612-2626.

59. Huesgen PF, Lange PF, Rogers LD, Solis N, Eckhard U, et al. (2015) LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification. Nat Methods 12: 55-58.

60. Shi Y, Xiang R, Horvath C, Wilkins JA (2004) The role of liquid chromatography in proteomics. J Chromatogr A 1053: 27-36.

61. Michalski A, Damoc E, Hauschild JP, Lange O, Wieghaus A, et al. (2011) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol Cell Proteomics 10: M111 011015.

62. Banerjee S, Mazumdar S (2012) Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte. Int J Anal Chem 2012: 282574.

63. Douglas DJ (2009) Linear quadrupoles in mass spectrometry. Mass Spectrom Rev 28: 937-960.

64. Deng L, Handler DCL, Multari DH, Haynes PA (2021) Comparison of protein and peptide fractionation approaches in protein identification and quantification from

Saccharomyces cerevisiae. J Chromatogr B Analyt Technol Biomed Life Sci 1162: 122453.

65. Nagaraj N, D'Souza RC, Cox J, Olsen JV, Mann M (2010) Feasibility of large-scale phosphoproteomics with higher energy collisional dissociation fragmentation. J Proteome Res 9: 6786-6794.

66. Zubarev RA, Horn DM, Fridriksson EK, Kelleher NL, Kruger NA, et al. (2000) Electron capture dissociation for structural characterization of multiply charged protein cations. Anal Chem 72: 563-573.

67. Olsen JV, Macek B, Lange O, Makarov A, Horning S, et al. (2007) Higher-energy C-trap dissociation for peptide modification analysis. Nat Methods 4: 709-712.

68. Scigelova M, Makarov A (2006) Orbitrap mass analyzer--overview and applications in proteomics. Proteomics 6 Suppl 2: 16-21.

69. Verheggen K, Raeder H, Berven FS, Martens L, Barsnes H, et al. (2020) Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. Mass Spectrom Rev 39: 292-306.

70. Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5: 976-989.

71. Hirosawa M, Hoshida M, Ishikawa M, Toya T (1993) MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming. Comput Appl Biosci 9: 161-167.

72. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, et al. (2004) Open mass spectrometry search algorithm. J Proteome Res 3: 958-964.

73. Kim S, Pevzner PA (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5: 5277.

74. Craig R, Beavis RC (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom 17: 2310-2316.

75. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, et al. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10: 1794-1805.

76. Kumar D, Yadav AK, Dash D (2017) Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data. Methods Mol Biol 1549: 17-29.

77. Degroeve S, Gabriels R, Velghe K, Bouwmeester R, Tichshenko N, et al. (2021) ionbot: a novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification. bioRxiv: 2021.2007.2002.450686.

78. Colaert N, Degroeve S, Helsens K, Martens L (2011) Analysis of the resolution limitations of peptide identification algorithms. J Proteome Res 10: 5555-5561.

79. Shteynberg DD, Deutsch EW, Campbell DS, Hoopmann MR, Kusebauch U, et al. (2019) PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. J Proteome Res 18: 4262-4272.

80. Bantscheff M, Lemeer S, Savitski MM, Kuster B (2012) Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Anal Bioanal Chem 404: 939-965.

81. Fenselau C (2007) A review of quantitative methods for proteomic studies. J Chromatogr B Analyt Technol Biomed Life Sci 855: 14-20.

82. Ong SE, Mann M (2007) Stable isotope labeling by amino acids in cell culture for quantitative proteomics. Methods Mol Biol 359: 37-52.

83. Mertins P, Qiao JW, Patel J, Udeshi ND, Clauser KR, et al. (2013) Integrated proteomic analysis of post-translational modifications by serial enrichment. Nat Methods 10: 634-637.

84. Blagoev B, Ong S-E, Kratchmarova I, Mann M (2004) Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. Nat Biotechnol 22: 1139-1145.

85. Ji F, Zhou M, Zhu H, Jiang Z, Li Q, et al. (2021) Integrative Proteomic Analysis of Posttranslational Modification in the Inflammatory Response. Genomics Proteomics Bioinformatics.

86. 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: 1367-1372.

87. Gupta N, Bandeira N, Keich U, Pevzner PA (2011) Target-decoy approach and false discovery rate: when things may go wrong. J Am Soc Mass Spectrom 22: 1111-1120.

88. Sticker A, Martens L, Clement L (2017) Mass spectrometrists should search for all peptides, but assess only the ones they care about. Nat Methods 14: 643-644.

89. Erce MA, Pang CN, Hart-Smith G, Wilkins MR (2012) The methylproteome and the intracellular methylation network. Proteomics 12: 564-586.

90. Perez-Burgos L, Peters AH, Opravil S, Kauer M, Mechtler K, et al. (2004) Generation and characterization of methyl-lysine histone antibodies. Methods Enzymol 376: 234-254.

91. Blum G, Islam K, Luo M (2013) Bioorthogonal profiling of protein methylation (BPPM) using an azido analog of S-adenosyl-L-methionine. Curr Protoc Chem Biol 5: 45-66.

92. Guo H, Wang R, Zheng W, Chen Y, Blum G, et al. (2014) Profiling substrates of protein arginine N-methyltransferase 3 with S-adenosyl-L-methionine analogues. ACS Chem Biol 9: 476-484.

93. Hart-Smith G, Yagoub D, Tay AP, Pickford R, Wilkins MR (2016) Large Scale Mass Spectrometry-based Identifications of Enzyme-mediated Protein Methylation Are Subject to High False Discovery Rates. Mol Cell Proteomics 15: 989-1006.

94. Ong SE, Mittler G, Mann M (2004) Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. Nat Methods 1: 119-126.

95. Caslavka Zempel KE, Vashisht AA, Barshop WD, Wohlschlegel JA, Clarke SG (2016) Determining the Mitochondrial Methyl Proteome in Saccharomyces cerevisiae using Heavy Methyl SILAC. J Proteome Res 15: 4436-4451.

96. Garcia BA, Mollah S, Ueberheide BM, Busby SA, Muratore TL, et al. (2007) Chemical derivatization of histones for facilitated analysis by mass spectrometry. Nat Protoc 2: 933-938.

97. Maile TM, Izrael-Tomasevic A, Cheung T, Guler GD, Tindell C, et al. (2015) Mass spectrometric quantification of histone post-translational modifications by a hybrid chemical labeling method. Mol Cell Proteomics 14: 1148-1158.

98. Noberini R, Savoia EO, Brandini S, Greco F, Marra F, et al. (2021) Spatial epi-proteomics enabled by histone post-translational modification analysis from low-abundance clinical samples. Clin Epigenetics 13: 145.

99. Noberini R, Robusti G, Bonaldi T (2021) Mass spectrometry-based characterization of histones in clinical samples: applications, progresses, and challenges. FEBS J.

100. El Kennani S, Crespo M, Govin J, Pflieger D (2018) Proteomic Analysis of Histone Variants and Their PTMs: Strategies and Pitfalls. Proteomes 6.

101. Pevzner PA, Mulyukov Z, Dancik V, Tang CL (2001) Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. Genome Res 11: 290-299.

102. Na S, Paek E (2015) Software eyes for protein post-translational modifications. Mass Spectrom Rev 34: 133-147.

103. Hernandez P, Gras R, Frey J, Appel RD (2003) Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics 3: 870-878.

104. Na S, Paek E (2009) Prediction of novel modifications by unrestrictive search of tandem mass spectra. J Proteome Res 8: 4418-4427.

105. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJ, et al. (2010) TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res 9: 1716-1726.

106. Na S, Bandeira N, Paek E (2012) Fast multi-blind modification search through tandem mass spectrometry. Mol Cell Proteomics 11: M111 010199.

107. Savitski MM, Nielsen ML, Zubarev RA (2006) ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. Mol Cell Proteomics 5: 935-948.

108. Bandeira N, Tsur D, Frank A, Pevzner PA (2007) Protein identification by spectral networks analysis. Proc Natl Acad Sci U S A 104: 6140-6145.

109. Yu F, Teo GC, Kong AT, Haynes SE, Avtonomov DM, et al. (2020) Identification of modified peptides using localization-aware open search. Nat Commun 11: 4065.

110. Rappsilber J, Mann M, Ishihama Y (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc 2: 1896-1906.

111. Massignani E, Cuomo A, Musiani D, Jammula S, Pavesi G, et al. (2019) hmSEEKER: Identification of hmSILAC Doublets in MaxQuant Output Data. Proteomics 19: e1800300.

112. O'Shea JP, Chou MF, Quader SA, Ryan JK, Church GM, et al. (2013) pLogo: a probabilistic approach to visualizing sequence motifs. Nat Methods 10: 1211-1212.

113. Tareen A, Kinney JB (2020) Logomaker: beautiful sequence logos in Python. Bioinformatics 36: 2272-2274.

114. Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H (2020) gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. F1000Res 9.

115. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, et al. (2014) The Reactome pathway knowledgebase. Nucleic Acids Res 42: D472-477.

116. Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, et al. (2021) The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 49: D344-D354.

117. Betts MJ, Lu Q, Jiang Y, Drusko A, Wichmann O, et al. (2015) Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. Nucleic Acids Res 43: e10.

118. Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B (2004) PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics 4: 1551-1561.

119. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.

120. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, et al. (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods 9: 345-350.

121. Parca L, Truglio M, Biagini T, Castellana S, Petrizzelli F, et al. (2020) Pyntacle: a parallel computing-enabled framework for large-scale network biology analysis. Gigascience 9.

122. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, et al. (2012) Fiji: an open-source platform for biological-image analysis. Nat Methods 9: 676-682.

123. Brame CJ, Moran MF, McBroom-Cerajewski LD (2004) A mass spectrometry based method for distinguishing between symmetrically and asymmetrically dimethylated arginine residues. Rapid Commun Mass Spectrom 18: 877-881.

124. Kahles A, Lehmann KV, Toussaint NC, Huser M, Stark SG, et al. (2018) Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell 34: 211-224 e216.

125. Bremang M, Cuomo A, Agresta AM, Stugiewicz M, Spadotto V, et al. (2013) Mass spectrometry-based identification and characterisation of lysine and arginine methylation in the human proteome. Mol Biosyst 9: 2231-2247.

126. Reichholf B, Herzog VA, Fasching N, Manzenreither RA, Sowemimo I, et al. (2019) Time-Resolved Small RNA Sequencing Unravels the Molecular Principles of MicroRNA Homeostasis. Mol Cell 75: 756-768 e757.

127. Araya N, Hiraga H, Kako K, Arao Y, Kato S, et al. (2005) Transcriptional down-regulation through nuclear exclusion of EWS methylated by PRMT1. Biochem Biophys Res Commun 329: 653-660.

128. Tradewell ML, Yu Z, Tibshirani M, Boulanger MC, Durham HD, et al. (2012) Arginine methylation by PRMT1 regulates nuclear-cytoplasmic localization and toxicity of FUS/TLS harbouring ALS-linked mutations. Hum Mol Genet 21: 136-149.

129. Allegra D, Bilan V, Garding A, Dohner H, Stilgenbauer S, et al. (2014) Defective DROSHA processing contributes to downregulation of MiR-15/-16 in chronic lymphocytic leukemia. Leukemia 28: 98-107.

130. Wei H-H, Fan X-J, Hu Y, Tian X-X, Guo M, et al. (2021) A systematic survey of PRMT interactomes reveals the key roles of arginine methylation in the global control of RNA splicing and translation. Science Bulletin 66: 1342-1357.

131. Geuens T, Bouhy D, Timmerman V (2016) The hnRNP family: insights into their role in health and disease. Hum Genet 135: 851-867.

132. Liu ZS, Cai H, Xue W, Wang M, Xia T, et al. (2019) G3BP1 promotes DNA binding and activation of cGAS. Nat Immunol 20: 18-28.

133. Reineke LC, Lloyd RE (2015) The stress granule protein G3BP1 recruits protein kinase R to promote multiple innate immune antiviral responses. J Virol 89: 2575-2589.

134. Rollins DA, Coppo M, Rogatsky I (2015) Minireview: nuclear receptor coregulators of the p160 family: insights into inflammation and metabolism. Mol Endocrinol 29: 502-517.

135. Michel JJ, Xiong Y (1998) Human CUL-1, but not other cullin family members, selectively interacts with SKP1 to form a complex with SKP2 and cyclin A. Cell Growth Differ 9: 435-449.

136. Oppikofer M, Bai T, Gan Y, Haley B, Liu P, et al. (2017) Expansion of the ISWI chromatin remodeler family with new active complexes. EMBO Rep 18: 1697-1706.

137. Tikhanovich I, Zhao J, Olson J, Adams A, Taylor R, et al. (2017) Protein arginine methyltransferase 1 modulates innate immune responses through regulation of peroxisome proliferator-activated receptor gamma-dependent macrophage differentiation. J Biol Chem 292: 6882-6894.

138. Cho JH, Lee R, Kim E, Choi YE, Choi EJ (2018) PRMT1 negatively regulates activation-induced cell death in macrophages by arginine methylation of GAPDH. Exp Cell Res 368: 50-58.

139. Zhao J, O'Neil M, Vittal A, Weinman SA, Tikhanovich I (2019) PRMT1-Dependent Macrophage IL-6 Production Is Required for Alcohol-Induced HCC Progression. Gene Expr 19: 137-150.

140. Musselman CA, Kutateladze TG (2021) Characterization of functional disordered regions within chromatin-associated proteins. iScience 24: 102070.

141. Bah A, Forman-Kay JD (2016) Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications. J Biol Chem 291: 6696-6705.

142. Darling AL, Uversky VN (2018) Intrinsic Disorder and Posttranslational Modifications: The Darker Side of the Biological Dark Matter. Front Genet 9: 158.

143. Piovesan D, Tabaro F, Paladin L, Necci M, Micetic I, et al. (2018) MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. Nucleic Acids Res 46: D471-D476.

144. Valverde R, Edwards L, Regan L (2008) Structure and function of KH domains. FEBS J 275: 2712-2726.

145. Fic W, Bastock R, Raimondi F, Los E, Inoue Y, et al. (2021) RhoGAP19D inhibits Cdc42 laterally to control epithelial cell shape and prevent invasion. J Cell Biol 220.

146. Fulton MD, Brown T, Zheng YG (2019) The Biological Axis of Protein Arginine Methylation and Asymmetric Dimethylarginine. Int J Mol Sci 20.

147. Dhar S, Vemulapalli V, Patananan AN, Huang GL, Di Lorenzo A, et al. (2013) Loss of the major Type I arginine methyltransferase PRMT1 causes substrate scavenging by other PRMTs. Sci Rep 3: 1311.

148. Hartel NG, Chew B, Qin J, Xu J, Graham NA (2019) Deep Protein Methylation Profiling by Combined Chemical and Immunoaffinity Approaches Reveals Novel PRMT1 Targets. Mol Cell Proteomics 18: 2149-2164.

149. Smith DL, Erce MA, Lai YW, Tomasetig F, Hart-Smith G, et al. (2020) Crosstalk of Phosphorylation and Arginine Methylation in Disordered SRGG Repeats of Saccharomycescerevisiae Fibrillarin and Its Association with Nucleolar Localization. J Mol Biol 432: 448-466.

150. Liu N, Yang R, Shi Y, Chen L, Liu Y, et al. (2020) The cross-talk between methylation and phosphorylation in lymphoid-specific helicase drives cancer stem-like properties. Signal Transduct Target Ther 5: 197.

151. Hamey JJ, Nguyen A, Wilkins MR (2021) Discovery of Arginine Methylation, Phosphorylation, and Their Co-occurrence in Condensate-Associated Proteins in Saccharomyces cerevisiae. J Proteome Res 20: 2420-2434.

152. Owen I, Shewmaker F (2019) The Role of Post-Translational Modifications in the Phase Transitions of Intrinsically Disordered Proteins. Int J Mol Sci 20.

153. Schisa JA, Elaswad MT (2021) An Emerging Role for Post-translational Modifications in Regulating RNP Condensates in the Germ Line. Front Mol Biosci 8: 658020.

154. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, et al. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res 43: D512-520.

155. Di Liegro CM, Schiera G, Di Liegro I (2018) H1.0 Linker Histone as an Epigenetic Regulator of Cell Proliferation and Differentiation. Genes (Basel) 9.

156. Schwartzentruber J, Korshunov A, Liu XY, Jones DT, Pfaff E, et al. (2012) Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. Nature 482: 226-231.

157. Tay AP, Geoghegan V, Yagoub D, Wilkins MR, Hart-Smith G (2018) MethylQuant: A Tool for Sensitive Validation of Enzyme-Mediated Protein Methylation Sites from Heavy-Methyl SILAC Data. J Proteome Res 17: 359-373.

158. Tikhanovich I, Zhao J, Bridges B, Kumer S, Roberts B, et al. (2017) Arginine methylation regulates c-Myc-dependent transcription by altering promoter recruitment of the acetyltransferase p300. J Biol Chem 292: 13333-13344.

159. Yan J, Horng T (2020) Lipid Metabolism in Regulation of Macrophage Functions. Trends Cell Biol 30: 979-989.

160. Batista-Gonzalez A, Vidal R, Criollo A, Carreno LJ (2019) New Insights on the Role of Lipid Metabolism in the Metabolic Reprogramming of Macrophages. Front Immunol 10: 2993.

161. Qualmann B, Kessels MM (2021) The Role of Protein Arginine Methylation as Post-Translational Modification on Actin Cytoskeletal Components in Neuronal Structure and Function. Cells 10.

162. Tompa P, Davey NE, Gibson TJ, Babu MM (2014) A million peptide motifs for the molecular biologist. Mol Cell 55: 161-169.

163. Yin X-K, Wang Y-L, Wang F, Feng W-X, Bai S-M, et al. (2021) PRMT1 enhances oncogenic arginine methylation of NONO in colorectal cancer. Oncogene 40: 1375-1389.

164. Krassowski M, Pellegrina D, Mee MW, Fradet-Turcotte A, Bhat M, et al. (2021) ActiveDriverDB: Interpreting Genetic Variation in Human and Cancer Genomes Using Post-translational Modification Sites and Signaling Networks (2021 Update). Front Cell Dev Biol 9: 626821.

165. Nowak DG, Woolard J, Amin EM, Konopatskaya O, Saleem MA, et al. (2008) Expression of pro- and anti-angiogenic isoforms of VEGF is differentially regulated by splicing and growth factors. J Cell Sci 121: 3487-3495.

166. Li Q, Zeng C, Liu H, Yung KWY, Chen C, et al. (2021) Protein-Protein Interaction Inhibitor of SRPKs Alters the Splicing Isoforms of VEGF and Inhibits Angiogenesis. iScience 24: 102423.

167. Pendleton KE, Chen B, Liu K, Hunter OV, Xie Y, et al. (2017) The U6 snRNA m(6)A Methyltransferase METTL16 Regulates SAM Synthetase Intron Retention. Cell 169: 824-835 e814.

168. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, et al. (2021) Highly accurate protein structure prediction for the human proteome. Nature 596: 590-596.

169. Wiles ET, Selker EU (2017) H3K27 methylation: a promiscuous repressive chromatin mark. Curr Opin Genet Dev 43: 31-37.

170. Huang C, Zhu B (2018) Roles of H3K36-specific histone methyltransferases in transcription: antagonizing silencing and safeguarding transcription fidelity. Biophys Rep 4: 170-177.

171. Hardman G, Eyers CE (2020) High-Throughput Characterization of Histidine Phosphorylation Sites Using UPAX and Tandem Mass Spectrometry. Methods Mol Biol 2077: 225-235.

172. Sawicka A, Seiser C (2012) Histone H3 phosphorylation - a versatile chromatin modification for different occasions. Biochimie 94: 2193-2201.

173. Zhang T, Zhang Z, Dong Q, Xiong J, Zhu B (2020) Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. Genome Biol 21: 45.

174. Sawicka A, Hartl D, Goiser M, Pusch O, Stocsits RR, et al. (2014) H3S28 phosphorylation is a hallmark of the transcriptional response to cellular stress. Genome Res 24: 1808-1820.

175. Guo A, Gu H, Zhou J, Mulhern D, Wang Y, et al. (2014) Immunoaffinity enrichment and mass spectrometry analysis of protein methylation. Mol Cell Proteomics 13: 372-387.

176. Mahmud I, Liao D (2019) DAXX in cancer: phenomena, processes, mechanisms and regulation. Nucleic Acids Res 47: 7734-7752.

177. Hoelper D, Huang H, Jain AY, Patel DJ, Lewis PW (2017) Structural and mechanistic insights into ATRX-dependent and -independent functions of the histone chaperone DAXX. Nat Commun 8: 1193.

178. Lindhorst PH, Hummon AB (2020) Proteomics of Colorectal Cancer: Tumors, Organoids, and Cell Cultures-A Minireview. Front Mol Biosci 7: 604492.

179. Bekker-Jensen DB, Bernhardt OM, Hogrebe A, Martinez-Val A, Verbeke L, et al. (2020) Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. Nat Commun 11: 787.

180. Davids M, Teerlink T (2013) Plasma concentrations of arginine and asymmetric dimethylarginine do not reflect their intracellular concentrations in peripheral blood mononuclear cells. Metabolism 62: 1455-1461.

181. Lam H (2011) Building and searching tandem mass spectral libraries for peptide identification. Mol Cell Proteomics 10: R111 008565.

182. Kelstrup CD, Frese C, Heck AJ, Olsen JV, Nielsen ML (2014) Analytical utility of mass spectral binning in proteomic experiments by SPectral Immonium Ion Detection (SPIID). Mol Cell Proteomics 13: 1914-1924.

183. Kim S, Gupta N, Pevzner PA (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. J Proteome Res 7: 3354-3363.

184. Guthals A, Boucher C, Bandeira N (2015) The generating function approach for Peptide identification in spectral networks. J Comput Biol 22: 353-366.