Development and Application of Mass Spectrometry Based Approaches to Study Chemical Modifications of Nucleic Acids and Proteins

Peiyuan Zheng

A thesis submitted in partial fulfilment of the requirements for the degree
of
DOCTOR OF PHILOSOPHY

Department of Chemical & Biological Engineering
The University of Sheffield

November 2019

Acknowledgements

With immense gratitude, I thank my supervisor, Prof. Mark Dickman, for offering me this great PhD opportunity, and this amazing project to working on. I would also like to thank my secondary supervisor, Dr. Guillaume Hautbergue. I thank them both for their enormous guidance and support throughout my PhD.

I would like to thank Dr. Phillip Jackson, Dr. Caroline Evans, Dr. Alison Nwokeoji and Dr. Joseph Longworth, for their help and support during my study. My thanks also go to Dr. Lydia Castelli from SITraN for her support and collaboration.

I would like to thank everyone in the lab, particularly, An-Wen, Beata, Eleanor, Elizabeth, Joby and Thomas.

Special thanks to Manix Vlot from Wageningen for the collaboration on the CRISPR project.

I take this opportunity to thank Dr. Mark Heslop for offering the amazing GTA experience.

I thank my family deeply for their continued support and belief in me. Special thanks to Qiao Hou for her tremendous financial support.

Most of all, my very heartfelt thanks to Ke Ji, for all her love and support throughout the journey.

Publications from work performed in this thesis

Vlot M, Houkes J, Lochs SJA, Swarts DC, Zheng P, Kunne T, Mohanraju P, Anders C, Jinek M, van der Oost J, Dickman MJ, Brouns SJJ.Bacteriophage DNA glucosylation impairs target DNA binding by type I and II but not by type V CRISPR-Cas effector complexes. Nucleic Acids Res. 2018 Jan 25;46(2):873-885.

In preparation.

Vlot M, Zheng P, van der Oost J, Brouns SJJ. Incorporation of 2'-Deoxycytidine analogues using genome engineering in *E. coli*.

Zheng P, Hautbergue G and Dickman MJ. Studying the effects of global demethylation on the mRNA interactome.

Table of Contents

List of Figures

Chapter 1

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

List of tables

Abbreviations

AAA	amino acid analysis
ABC	ammonium bicarbonate
ACN	acetonitrile
ACR	anti-CRISPR proteins
AdOx	adenosine, periodate oxidized
AGO	Argonaute protein
ALS	amyotrophic lateral sclerosis
APS	ammonium persulfate
AQUA	absolute quantification
AUC	area under curve
BPC	base peak chromatogram
Cas	CRISPR-associated
CID	collision induced dissociation
CLIP	cross-linking immunoprecipitation
Conc.	Concentration
CRISPR	clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
crRNP	CRISPR ribonucleoprotein
dCMP	deoxycytidine monophosphate
DEPC	diethyl pyrocarbonate
dFBS	dialyzed fetal bovine serum
DM	myotonic dystrophy
DMA	dimethylarginine
DMEM	Dulbecco's Modified Eagle Medium
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
DPR	dipeptide repeat protein
dsRBM	double stranded RNA binding motif
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
eGFP	enhanced green fluorescent protein
EIC	extracted ion chromatogram
ELISA	enzyme-linked immunosorbent assay

EMSA	electrophoretic mobility shift assay
ESI	electrospray ionization
ETD	electron transfer dissociation
FA	formic acid
FDR	false discovery rate
FT	Fourier transform
FTD	frontotemporal dementia
GAR	glycine-arginine-rich
GELC-MS	in-gel digestion coupled with mass spectrometry analysis
GO	gene ontology
HA	haemagglutinin
HCD	higher-energy collisional dissociation
HDR	homology directed repair
HEK	human embryonic kidney
hmC	5-hydroxymethyl-cytosine
hmU	hydroxymethyluridine
hnRNP	heterogeneous nuclear ribonucleoproteins
HPLC	high-performance liquid chromatography
HRE	hexanucleotide repeat expansion
IA	immunoassay
IAA	iodoacetamide
ICAT	isotope coded affinity tags
iTRAQ	isobaric tags for relative and absolute quantification
KH	K-homology domain
LC-MS	liquid chromatography–mass spectrometry
LF	lethal factor
LFQ	label-free quantification
LOD	limit of detection
mA	methyladenosine
MALDI	matrix assisted laser desorption ionization
mC	methylcytidine
MGE	mobile genetic element
MGF	Mascot generic format
miRNA	micro RNA
MMA	monomethylarginine

mRNA	messenger RNA
mRNP	messenger ribonucleoprotein particles
MS	mass spectrometry
MS/MS or MS$_n$	tandem mass spectrometry
MW	molecular weight
m/z	mass-to-charge ratio
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
NLS	nuclear localization signal
OF	oedema factor
PA	protective antigen
PADs	protein arginine deiminases
PADI4	peptidyl arginine deiminase, type 4
PAGE	polyacrylamide gel electrophoresis
PAM	protospacer adjacent motif
PAR-CLIP	photoactivatable ribonucleoside–enhanced CLIP
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PGC	porous graphitic carbon
PMSF	phenylmethylsulfonyl fluoride
pre-crRNA	precursor CRISPR RNA
PTM	post-translational modification
RAN	repeat-associated non-ATG (codon)
RBD	RNA binding domain
RBM	RNA binding motif
RBP	RNA binding protein
RM	restriction modification
RNA	ribonucleic acid
RNAi	RNA interference
RNP	ribonucleoprotein
rPA	recombinant protective antigen
RPLC	reversed phase liquid chromatography
RRM	RNA recognition motif
rRNA	ribosomal RNA
SCA	spinocerebellar ataxia

SD	standard deviation
SDS	sodium dodecyl sulfate
sgRNA	single guide RNA
SILAC	stable isotope labelling with amino acids in cell culture
SIS	stable isotope standards
snRNA	small nuclear RNA
TBE	Tris/Borate/EDTA
TBS	Tris-buffered saline
TEMED	tetramethylethylenediamine
TFA	trifluoroacetic acid
TIC	total ion current
TMT	tandem mass tags
TOF	time-of-flight
tracrRNA	trans-activating crRNA
tRNA	transfer RNA
UHPLC	ultra-high performance liquid chromatography
UHR	ultra-high resolution
UDPG	uridine diphosphate glucose
UTR	untranslated region
UV	ultraviolet
WB	western blot
WCE	whole cell extract
XIC	extracted-ion chromatogram
Znf	zinc fingers
4-SU	4-thiouridine
6-SG	6-thioguanosine

Note: For protein name/gene name not included in this list, refer to Uniprot website (www.uniprot.org).

Thesis abstract

Nucleic acids and proteins are the most important biomolecules essential to all known life forms on Earth. Understanding the interactions between nucleic acids and proteins is of crucial importance, as they play a central role in a wide range of fundamental cellular processes such as DNA replication, DNA recombination, DNA repair, RNA transcription, RNA processing and translation. Chemical modifications of nucleic acids (e.g. post-transcriptional RNA modifications) in conjunction with chemical modifications of proteins (post-translational modifications) enable the cell to regulate these interactions and therefore control a range of important biological phenomena.

Mass spectrometry based methods are powerful tools that can be utilised to analyse biomolecules including their chemical modifications. In this thesis I have developed and optimised mass spectrometry based methods to study the effects of chemical modifications of both nucleic acids and proteins and provide further insight into a number of important biological systems.

To study the effect of DNA modifications on CRISPR/Cas systems, a range of DNA substrates with different chemical modifications were synthesised and subsequently validated using liquid chromatography in conjunction with mass spectrometry. Prepared DNA were tested by collaborators and results showed that DNA glucosylation can interfere with type I-E and type II-A CRISPR-Cas systems activities but not type V-A.

In addition, quantitative proteomic approaches were used to study the effects of protein post-translational modifications on both the mRNA interactome *in vivo* and the proteins that bind to GGGGCC RNA repeats *in vitro*. Stable isotope labelling with amino acids in cell culture (SILAC) in conjunction with mass spectrometry based approaches were successfully used to identify candidate proteins that interact with mRNA *in vivo*, or with RNA GGGGCC repeats *in vitro*, are altered upon the addition of the global demethylase AdOx.

Lastly, a new mass spectrometry based absolute quantification method (AQUA) was developed and applied to determine the concentration of PA/rPA in anthrax vaccine products.

Chapter 1: Introduction

1.1 Nucleic acid-protein interactions

Nucleic acids, including deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), together with proteins, are essential to all known life forms on Earth. Nucleic acids are the carriers of genetic information, while proteins are the regulators involved in all biological processes. Their interactions underpin most cellular processes including DNA replication, DNA recombination, DNA repair, transcription of RNA, RNA processing and translation. Understanding of these nucleic acid-protein interactions has become a major goal for biologists, as such interactions play a central role in a wide range of fundamental cellular processes.

1.2 Principles of nucleic acid protein interactions

1.2.1 Physical forces for nucleic acid protein interactions

Proteins can interact with nucleic acids through a number of physical forces, including hydrogen bonding interactions (or dipolar interactions), salt bridges (or electrostatic interactions), hydrophobic effects (or entropic effects), and van der Waals interactions (including stacking interactions) (Figure 1.1). These forces contribute to different types of nucleic acid protein interactions, either specific or non-specific. More details regarding these forces are described below.

Hydrogen bonding interactions are caused by an electrostatic field formed between a hydrogen atom and a highly electronegative atom such as oxygen, fluorine and nitrogen. Although energetically weak, hydrogen bonds are the main cause for sequence-specific interactions. Sequence-specific interactions are non-covalent interactions between nucleic acids and proteins where a specific nucleic acid sequence motif or a specific nucleotide composition such as GC-rich sites are involved. Hydrogen bonding between nucleic acids and proteins are usually formed between nucleic acid bases and polar amino acid side chains (Figure 1.1A), or between nucleic acid bases and aromatic rings of amino acids (π-hydrogen bonds, weaker and less favoured). Apart from amino acid side chains, protein amide backbone groups (C=O or N-H) can also form hydrogen bonds. For amino acids, both the oxygen atoms of nucleic acid phosphate backbones and those from the ribose sugar rings are also common hydrogen bonding acceptors.

Salt bridge refers to non-covalent electrostatic interactions between opposite charged groups, and usually represents the combination force of hydrogen bonding and ionic bonding. As in nucleic acid protein interactions, salt bridges are usually formed between oxygen atoms of nucleic acid backbones and the guanidinium moiety of arginine (Figure 1.1B), or other charged moieties such as the ε-amino group of lysine and the imidazole ring of histidine (Donald et al., 2011). Salt bridges may contribute to the stability of protein-nucleic acid complexes (Hendsch and Tidor, 1994; Sindelar, Hendsch and Tidor, 1998). Compared to hydrogen bonding activities where alignment of groups are required to obtain optimal strength, salt bridges are not directional. The strength of salt bridges depends on the distance between charged groups and dielectric constant (Barril et al., 1998).

The hydrophobic effect is a short-range entropic effect caused by the disruption of dynamic hydrogen bonds between water molecules at a non-polar surface (Silverstein, 1998). When this effect is applied to proteins and nucleic acids, they tend to interact with others in a way that non-polar parts of their external surface exposed to the aqueous medium are kept to a minimum (Figure 1.1C).

Van der Waals interactions arise owing to the electronic charge distribution variations over time. The transient asymmetry of the charge distribution at any instant of an atom interacts electrostatically with neighbouring atoms, resulting a complementary asymmetry electron distribution (Figure 1.1D). Van der Waals interactions are weak (2-4 kJ/mol) compared to other interactions described previously, and sensitive to structural fluctuations. The force (attraction) is increased when atoms get closer until they reach the van der Waals contact distance, after which the repulsive force grows. Considering the size of a typical biomolecule interfaces, the cumulative contribution of van der Waals interactions can be significant. Stacking interactions are one form of van der Waals interactions that take place between aromatic rings containing π-bonds. In nucleic acid-protein interactions, stacking interactions can take place between, for example, the side chain of phenylalanine and nucleic acid bases.

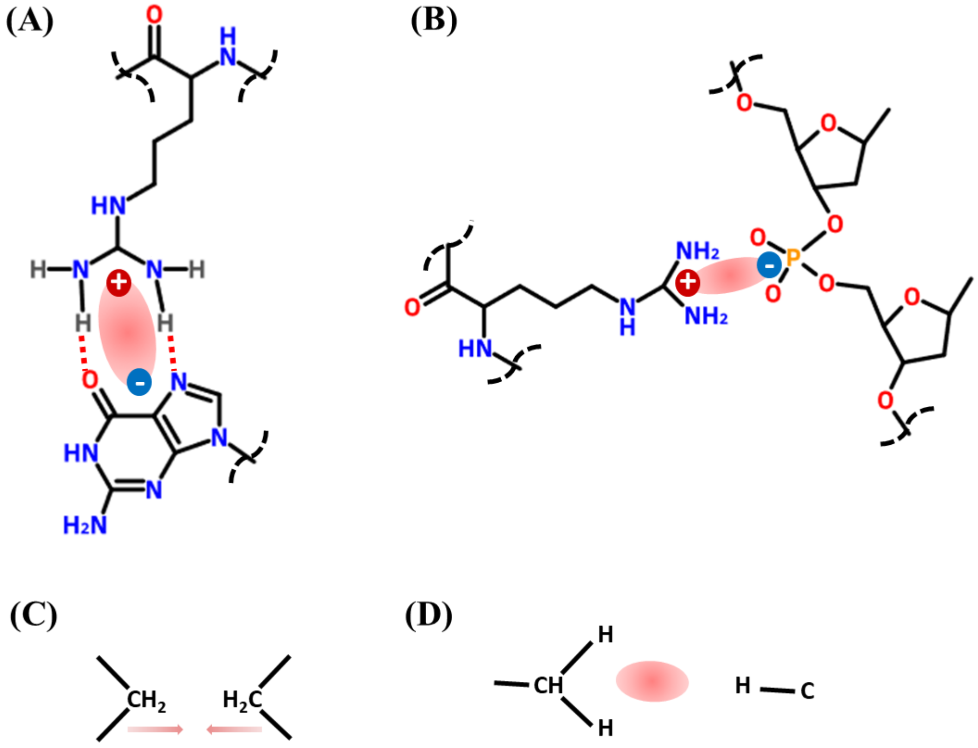


Figure 1.1 | Summary of physical forces used in nucleic acid-protein interactions. (A) Hydrogen bonding example showing the interactions between amino acid side chains (arginine) and nucleic acid base (guanine). Hydrogen bonds are indicated by dashed red lines. **(B)** Salt bridge between amino acid residue and nucleic acid backbone. **(C)** Hydrophobic effect between non-polar groups, arrows show the force directions. **(D)** Van der Waals Interactions.

1.2.2 Protein-DNA interactions

Double stranded DNA (dsDNA) is composed of a negatively charged sugar-phosphate backbone on the outside and stacked base pairs on the inside. The double helix structure of DNA contains grooves (major and minor). In the grooves, the outer edges of the nitrogenous bases are exposed with atoms available for hydrogen bonding, making them the binding sites for proteins. The chemical distinction between A-T and G-C pair surfaces of DNA is an important feature for proteins to recognise and bind to specific sequences. In B-DNA (Figure 1.2A), the biologically predominant form in cells (Richmond and Davey, 2003), the major grooves are deep and wide, which facilitate the access of secondary protein structures such as α-helices and β-sheets (Figure 1.2B). With the base pairs of certain sequences exposed in the grooves as “signatures”, particular sequence can be recognized by proteins (Rohs et al., 2010). Physical forces including hydrogen bonds and salt bridges contribute to the specific base sequence recognition and interaction. This nucleic acid-protein interaction can also be enhanced by oligomerisation and multi-complex formation of proteins, for example, the general control protein GCN4 (Figure 1.2B right panel). A number of different DNA binding domains found in proteins are discussed in the following section.

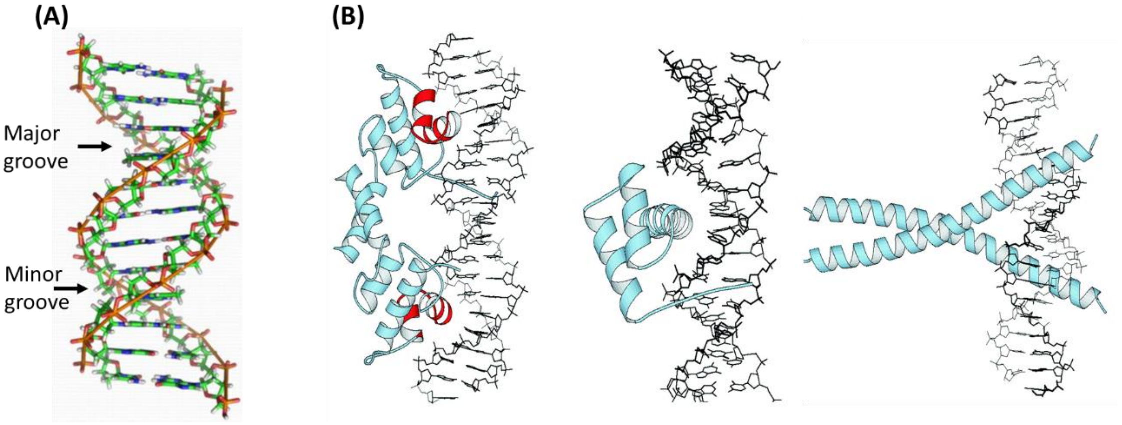


Figure 1.2 | The structure of DNA and protein-DNA interactions. (A) Structure of B-DNA showing the major and minor grooves. **(B)** Examples of DNA binding to protein motifs. From left to right showing: helix-turn-helix, engrailed, and helix zipper. Image reproduced from Garvie and Wolberger, 2001, with permission.

Among the structural elements used as DNA-binding domains, the most frequent form is the α-helix, which in most cases, interacts with major grooves of DNA (Rohs et al., 2010). The helix-turn-helix is a widely occurring DNA-binding motif structure that found in most type of life forms (Alberts et al., 2002). The helix-turn-helix is composed of two α-helices and an amino acid strand in between (Figure 1.2B left panel). The two α-helices are for DNA recognition and binding, although it is suggested that in some cases only one of the helices is included in the recognition, while the other is mainly to stabilize the interaction (Matthews et al., 1982). This structure is often found in gene expression regulation proteins such as the engrailed homeodomain of *Drosophila melanogaster* and the lambda repressor of phage lambda. Another similar structural motif is the helix-loop-helix, where two α-helices are connected by a loop. It is a DNA-binding motif found in transcriptional regulatory proteins (Massari and Murre, 2000). The leucine zipper (or leucine scissors) is another transcription regulator with the α-helices (Figure 1.2B right panel). The zipper protein can interact with DNA by inserting two α-helices into the major grove of DNA. It is named leucine zipper because within the dimerization domain of 60-80 amino acids, leucine occurs repeatedly every seven amino acids, and these leucines intermesh along the helical pair axis like a zipper. One typical example of this domain is the transcription factor bZIP (Vinson et al., 1989). It can recognize and bind to various sequences such as ACGT motifs, GCN4 motifs and palindromic sequences (Juhász et al., 2011; Nijhawan et al., 2008; Shimizu et al., 2005).

Zinc fingers refer to secondary protein structure (α-helix and β-sheet) with one or more coordinated zinc ions (Zn^{2+}) that hold them together (usually by the coordination of zinc ion and the hydrophobic core, including the participation of cysteine sulphur atoms and histidine imidazole nitrogen atoms) (Miller et al., 2001). Zinc fingers are usually in tandem repeats when interact with target molecules. In eukaryotes, although zinc finger domains are small motifs in size, they contribute to the largest groups of DNA-binding proteins. Zinc fingers can interact with DNA, RNA, proteins and other small molecules (Laity et al., 2001).

1.2.3 Protein-RNA interactions

Compared to DNA, RNA structures are more complex, thus the protein motifs for RNA reorganization are also likely to be complicated. With the secondary and tertiary structure of RNA, additional mechanisms are used for sequence specific bindings. In general, proteins bind to RNA either using a groove binding mode, or a β-sheet binding mode. In the groove binding mode, similar to DNA-protein interactions, α-helix is the main structure for binding; while in the β-sheet binding mode, a sheet that targeting unpaired RNA bases is often included (Draper, 1999). Sequence-specific binding of RNA with proteins relies largely on aromatic residues, which stack on unpaired bases.

For single-stranded RNA, a common binding domain is RNA recognition motif (RRM), which is composed of approximately 90 amino acids (Dreyfuss et al., 1988). RRMs have a large structural versatility with numerous biological functions. A typical RRM structure is composed of 4 anti-parallel β-sheets and 2 or 3 α-helices (Birney et al., 1993). These structures are arranged in a certain manner where side chains can stack with RNA bases (Figure 1.3).

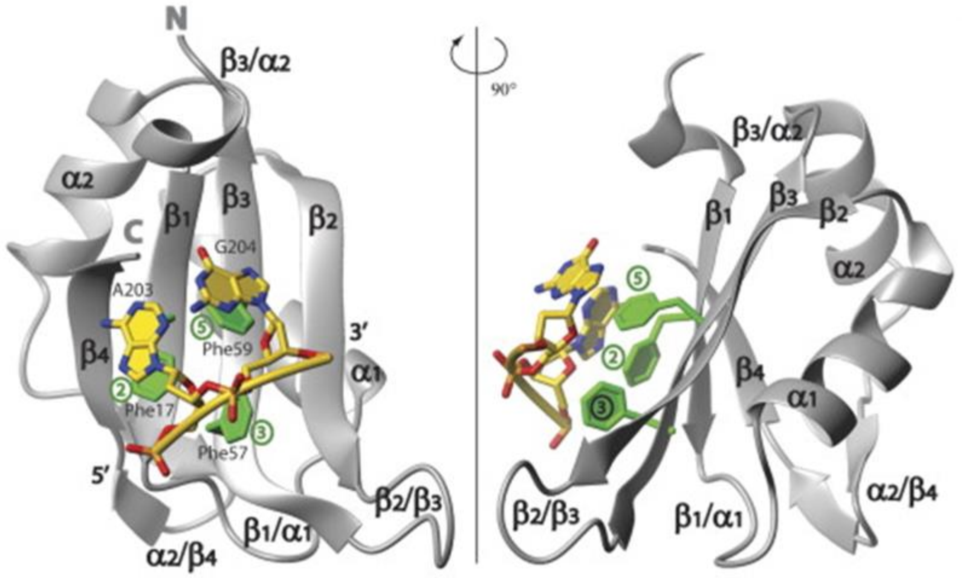


Figure 1.3 | Interactions between RRM and RNA. The model shows the binding of a single stranded nucleic acid to RRM β-sheets. The aromatic side-chains (shown in green, position shown as circled numbers 2, 3 and 5) of RRM β-sheets (β_1 and β_3) stack on the bases of nucleotides (shown in yellow). RRM is shown in grey. N: amino terminus; C: carboxyl terminus; α: α-helices; β: β-sheets. Image reproduced from (Cléry et al., 2008) with permission.

For double-stranded RNA, zinc-fingers are involved in recognition and binding, similar to DNA-protein interactions. An example for this is the zinc finger motif in transcription factor TFIIIA which interacts with ribosomal 5S RNA. Double stranded RNA binding motif (dsRBM) is another RNA-binding domain found to be of important roles including RNA processing, localization and translational repression (Tian et al., 2004; Chang and Ramos, 2005).

Protein/nucleic acid complexes are formed by proteins binding to either DNA or RNA. Some proteins such as transcription factors and histones have DNA-binding domains which can interact with the groove structure of DNA (Pabo and Sauer, 1984; Leblanc and Rodrigue, 2015). DNA-protein interactions can be either non-specific (structural proteins such as histones) or sequence dependent (such as transcription factors). As for RNA binding proteins, or RBPs, refer to proteins that bind to either single or double stranded RNA to form ribonucleoprotein (RNP) complexes. RBPs exist in both the cytoplasm and the nucleus, and have various structural motifs including zinc finger, dsRNA binding domain (dsRBD), RNA recognition motif (RRM), S1 domain and others (Lunde et al., 2007). RBPs have several crucial functions, such as RNA processing, transport and localization (Dreyfuss et al., 2002).

When RBPs bind to messenger RNA (mRNA), the complexes are named messenger ribonucleoprotein particles (mRNPs). RBPs bind to mRNA throughout the entire mRNA life cycle, which serve as structural elements and modify gene expression output (Müller-Mcnicoll and Neugebauer, 2013). Specifically, for RBPs that bind to pre-mRNA in nucleus, the formed complexes are known as heterogeneous nuclear ribonucleoproteins (hnRNPs) particles. In human cells (HeLa), the hnRNP family is composed of at least 20 major (abundant) hnRNPs, named from hnRNP A1 to hnRNP U, together with some minor (less abundant) hnRNPs such as Aly/REF, which bind to RNA in a less stable manner, or associated with only specific RNA (Piñol-Roma and Dreyfuss, 1993) (Chaudhury et al., 2010). hnRNP contains unique RNA binding domains (RBDs), including RRM (the most common type of RBD), quasi-RRM, arginine/glycine-rich box (or RGG box) and KH-K homology domain (or KH domain) (Geuens et al., 2016). With different domains, hnRNPs show diverse functions, depending on their location in cells (Piñol-Roma and Dreyfuss, 1993). In order to regulate hnRNP shuttle between nucleus and the cytoplasm, most hnRNP proteins contain a nuclear localization signal (NLS), a conventional short amino acid sequence used as a tag for nuclear import (Han et al., 2010). NLS are typically composed of exposed, positively charged arginine and lysine, making NLS prone to post-translational modifications including methylation, ubiquitination, phosphorylation and sumoylation (Chaudhury et al., 2010). These modifications lead to subcellular localisation and biological activity changes (Geuens et al., 2016).

1.3 Chemical modifications of nucleic acids and proteins

The known biological living systems on Earth are exclusively composed of macromolecules including proteins and nucleic acids. Modifications of these molecules, which fundamentally change their composition and structure, play key roles of regulation in biological systems. Understanding these modifications is of crucial importance in Biology. In this section, modifications that are generally associated with DNA, RNA and proteins will be described.

1.3.1 DNA modifications and prokaryotic defence systems

The selective pressures on prokaryotes mainly come from viruses. Bacteria need to protect themselves against intense threat of phage attack. For this purpose, diverse defence systems that target different stages of a phage life cycle have been developed in bacteria. Chemical modifications are usually involved in these defending strategies. Restriction modification (RM) system, for example, is a well-known bacteria defence system that utilises DNA modifications.

Phage, on the other hand, also chemically modify their DNA. The chemical modification of bacteriophage DNA can be found described in the 1950s (Luria and Human, 1952). There are over 10 known types of modification in different phage DNA (Warren, 1980), some examples are shown in Figure 1.4. Predominantly these modifications are employed by phage to protect their DNA from host enzymes that aim to degrade them. There are also some DNA modifications with functions yet to be discovered. For T4 phage (a phage with double-stranded DNA that infects *Escherichia coli*), all cytosine residues are replaced by 5-hydroxymethyl-cytosine (hmC), which can further be glucosylated into glucosyl hmC (ghmC) (Josse and Kornberg, 1962). Other covalent modifications in phage DNA include cytosine methylation (mC), thymine replaced by uracil (U) and hydroxymethyluridine (hmU) (Warren, 1980). Phosphorothioate modification of DNA (phosphate backbone variants with non-bridging oxygen atom replaced by sulphur) exists in a wide range of bacteria and archaea that have dnd genes (Wang et al., 2007) and was suggested as a potential host defence strategy. These variants are of interest as they can be delivered (as therapeutic oligonucleotides) by passing lipid bilayer barrier, and with prolonged lifetime in the exonuclease environment (Gryaznov, 2010).

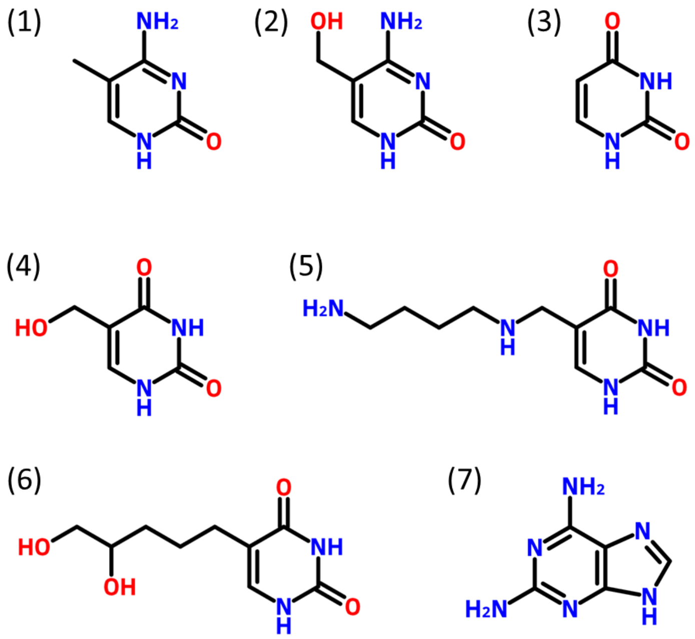


Figure 1.4 | Examples of modified bases in phage DNA. (1) 5-methylcytosine (mC); 2) 5-hydroxymethylcytosine (hmC); (3) uracil (U); (4) 5-hydroxymethyluracil (hmU); (5) α-putrescinylthymine (putT); (6) 5-dihydroxypentyluracil (dhpU); (7) 2-aminoadenine (nA).

In the long period of evolution competition against their host, bacteriophage have developed mechanisms to survive the host nuclease activity by modifying their own DNA. There are different types of modified bases observed in phage DNA, depending on the phage species. For example, cytosine bases are replaced by 5-hydroxymethylcytosine in some T4 phage that infect *Escherichia coli*, while thymine bases are replaced by 5-hydroxymethyluracil or uracil in some phage that infect *Bacillus subtilis*. Phage DNA bases can be modified fully or partially. In T4 phage DNA, cytosine bases are 100% replaced by 5-hydroxymethylcytosine, and thymine bases are 41% replaced by 5-dihydroxypentyluracil (Warren, 1980). The modification of bases changes the DNA structure, thus changes the DNA physical properties such as thermal transition temperature, which is a consequence of charge status alteration caused by modified group interacting with phosphate group of DNA (Warren, 1980). The biogenesis of modified bases is complicated and involves a series of enzymes such as dCMP hydroxymethylase and dCMP deaminase (Weigele and Raleigh, 2016).

1.3.2 CRISPR systems

The CRISPR systems (clustered regularly interspaced short palindromic repeats) are inheritable immune system recently discovered in about half of bacteria and most archaea genomes (Grissa et al., 2007). Unlike the restriction modification system (RM), another bacteria defend system that uses endonuclease to destroy invasive DNA with incorrect modifications, CRISPR systems deal with viral predation in a way analogous to RNAi in eukaryotes. CRISPR systems generally include three process stages (Figure 1.5), which will also be discussed in more details later in this section. CRISPR systems are adaptive immune systems. Foreign DNA sequences are incorporated into the host chromosome, so the invasion can be recorded inherited to younger generations.

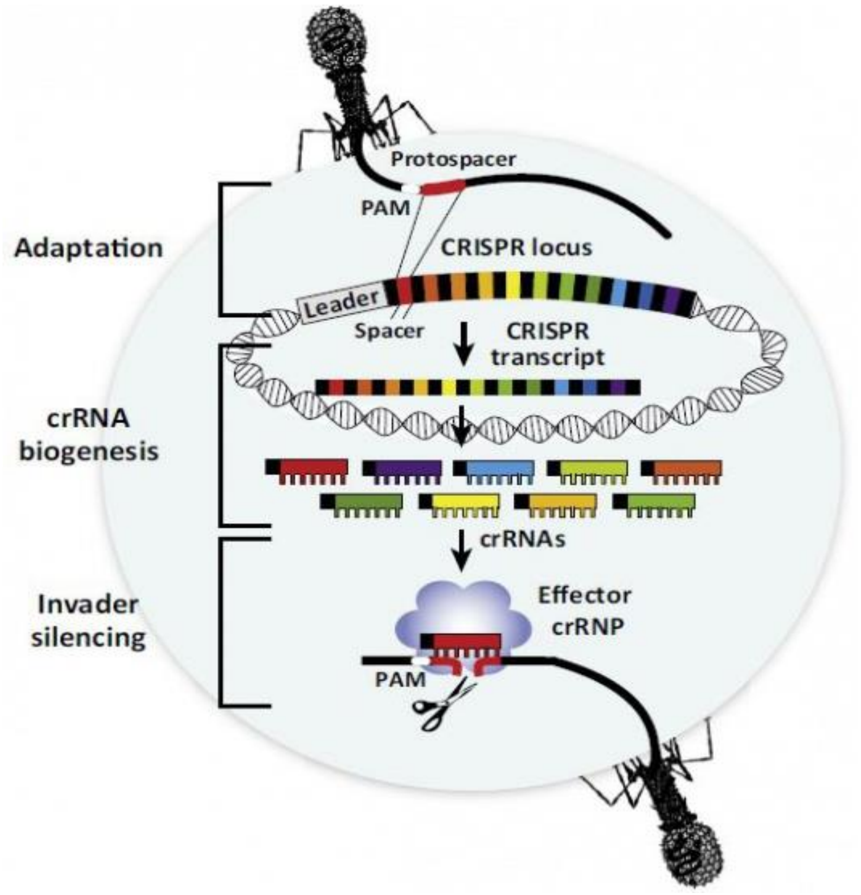


Figure 1.5 | Overview of the CRISPR system. A three stage process is described. In stage one/adaption, an invasive DNA fragment is incorporated into the host gene from the leader side, which is used for phage DNA recognition. In stage two/crRNA biogenesis, CRISPR locus is transcribed and processed into CRISPR RNAs (crRNAs) composed of repeat sequence (black) and spacer sequence (multiple colours). crRNA then associates with proteins to become crRNP, and silences the invader (stage three). Figure reprinted from Terns and Terns, 2014, with permission.

The DNA loci of CRISPR consist of cas (CRISPR-associated) genes, a leader sequence and a series of repeat-spacer array (Sorek et al., 2013). Cas genes are usually located adjacent to the repeat-spacer arrays, encoding functional proteins such as DNA helicase, nuclease and polymerase. Over 45 different genes are discovered related to or functioning as cas genes (Haft et al., 2005). Among these genes, six are considered crucial, which are named from cas1 to cas6, and only cas1 and cas2 are universally existed (Haft et al., 2005). The adenine and thymine rich leader sequence is believed critical for CRISPR RNA (crRNA) expression and spacer acquisition, as it contains promoter elements and regulatory protein binding sites (Pul et al., 2010; Yosef et al., 2012). The acquisition of new sequences as spacers is believed to take place at the leader end of the CRISPR array, because of the diversity of leader-end (Pourcel et al., 2005). The repeat-spacer structure is the feature part of CRISPR system. Both the repeat sequences and the spacer sequences are of the similar sizes, normally 20-50bp in length (Sorek et al., 2013). Although the repeat sequences are conservative, they can vary in different CRISPR loci, and the palindromic sequences gave these repeats the second structure of hairpins (Kunin et al., 2007).

The diversity of CRISPR-Cas systems is likely a consequence of fast evolution caused by severe competence against invasive mobile genetic elements (Stern and Sorek, 2011). The differences in CRISPR systems lie mainly in three aspects: 1) repeat sequences, 2) cas gene sequences and 3) cas operon architectures (Van Der Oost et al., 2014). Based on these differences, CRISPR-Cas systems are classified into 3 main types: type I, II and III. Each main type include different subtypes, referred to as for example: I-A, I-B and so forth. Cas proteins, coded by cas genes, have four major functions: 1) to serve as nuclease/recombinases for acquiring new spacers, 2) as ribonclease to guide the process of crRNA, 3) as part of CRISPR ribonucleoprotein (crRNP) complexes for target surveillance, and 4) as nucleases to degrade target nucleic acids (Van Der Oost et al., 2014). Major types of CRISPR-Cas system is usually recognized by their featured cas proteins.

Type I CRISPR systems are commonly existed in bacteria and archaea. All 6 subtypes of type I systems from A to F include a cas3 (in some cases known as cas6e or casE) gene as a hallmark (Sorek et al., 2013). Cas3 is a protein with a conserved phosphohydrolase domain and a helicase domain to cleave and unwind double stranded DNA targets (Makarova et al., 2011). However, cas3 on its own cannot provide immunity, as it does not identify invasive nucleic acids. The surveillance is provided by crRNA-guided crRNP complexes. In case of type I systems, this crRNP complex is known as Cascade (CRISPR-associated complex for antiviral defence). Cascade in *Escherichia coli* K12 (type I-E) is a sea-horse-shaped crRNP complex composed of cas protein subunits and a crRNA (see Figure 1.6). The 3D structure of Cascade was determined using cryo-electron microscopy (Wiedenheft et al., 2011), which was further confirmed by mass spectrometry analysis (van Duijn et al., 2012).

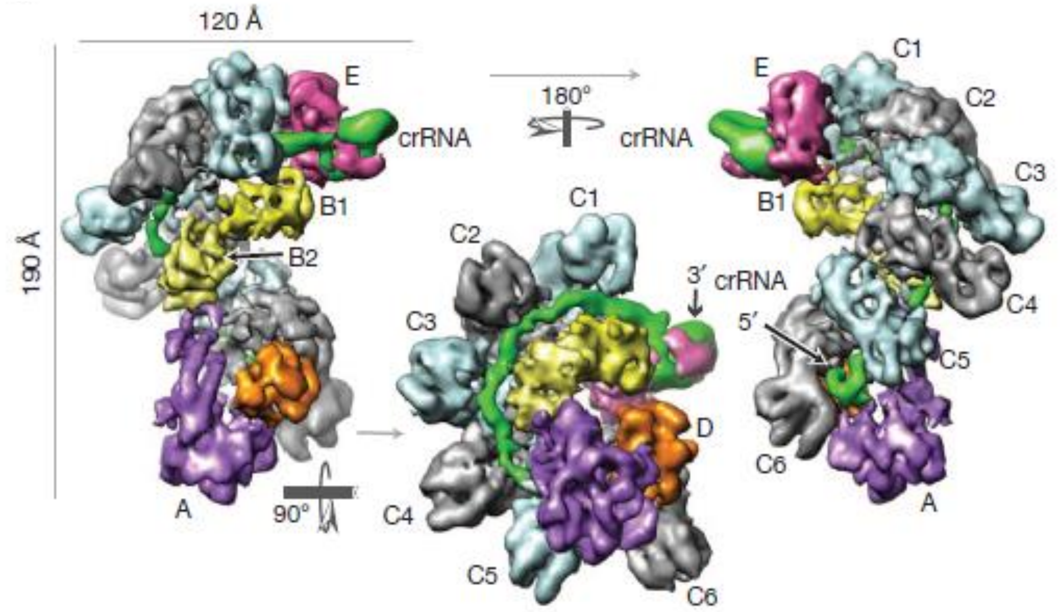


Figure 1.6 | The structure of crRNP complex (Cascade) from *E.coli*. The complex is composed of a helical backbone formed by six casC proteins (C1-C6). The crRNA lies in the groove of casC helix marked as green. casE (or cas3/cas6e) binds to the stem-loop of the crRNA as the head of the sea-horse (magenta). A dimer of two casB lies in the crRNA-CasC spine, connecting casA and casE (yellow). casA at the tail of the complex (purple) is hooked by crRNA, while casD (orange) lies adjacent to casA and interacts with casC at the tail side. Figure reprinted from Wiedenheft et al., 2011 with permission.

Cas3 not only cleaves target DNA, as described previously, but also processes the long crRNA into small mature crRNAs, a process known as crRNA biogenesis, which will be discussed in a later section.

Type II systems are different in both phylogeny and structure compared to Type I and III systems, and are only found in bacteria (Fonfara et al., 2014). Type II systems feature a cas9 gene, which encodes multifunctional proteins actively involved in different stages of this system. There are 3 subtypes: II-A, II-B and II-C. The crRNP of this type is named cas9 complex (Konermann et al., 2015). The process of crRNA for type II systems is unique. It requires the binding of trans-activating crRNA (tracrRNA) to the repeat section of pre-crRNA, so the complex can be recognized and processed into mature crRNA by enzymes such as RNase III (Sorek et al., 2013). It is suggested that both the existences of crRNA and tracrRNA are required for target DNA cleavage (Karvelis et al., 2013). Cas9 is known as an endonuclease guided by crRNA. The cleavage of target DNA in type II systems is believed due to the HNH domain and the RuvC-like domain of cas9 (Jinek et al., 2014a).

Type III systems have some similarity compared to type I systems, and can be found in both bacteria and archaea. There are two subtypes: type III-A and III-B. The crRNP of type III-A is known as the csm complexes, and for type III-B is the cmr complexes (Staals et al., 2014). In type III systems, cas10 is suggested related to target interference, while cas6, a type of endoribonuclease, exists in all type III systems (reviewed in Sorek et al., 2013). Type III-A targets DNA (Marraffini and Sontheimer, 2008), and Type III-B targets RNA (Hale et al., 2009; Cocozaki et al., 2012).

Mechanism of CRISPR interference

CRISPR-Cas systems use a three-step interference mechanism to protect themselves against invasive nucleic acids.

Stage 1: spacer acquisition

When invaded by foreign DNA, the host cells use CRISPR systems for target DNA recognition and fragmentation. It is suggested that the fragmentation process is firstly done by RM systems, and the fragmented DNA is subsequently used for spacer acquisition by CRISPR systems (Dupuis et al., 2013). The spacers are chosen by the recognition of protospacer adjacent motif (PAM), which is a sequence of 2-5 nucleotides flanked the protospacer of invading DNA (Bolotin et al., 2005). When recognized, fragmented DNA is integrated into the leader end of the CRISPR locus. Cas1 and cas2 are required for spacer acquisition (Figure 1.7). Cas1 catalyses the cleavage of DNA, while cas2 functions either as endoribonuclease or deoxyribonuclease (Babu et al., 2011; Beloglazova et al., 2008). Cas1 and cas2 form a complex proved essential for *in vivo* spacer acquisition in *E. coli* (Nuñez et al., 2014). Spacer acquisition also requires other cas proteins such as csn2, cas3, as well as some housekeeping proteins (Van Der Oost et al., 2014; Babu et al., 2011b). Detailed mechanism is still unclear.

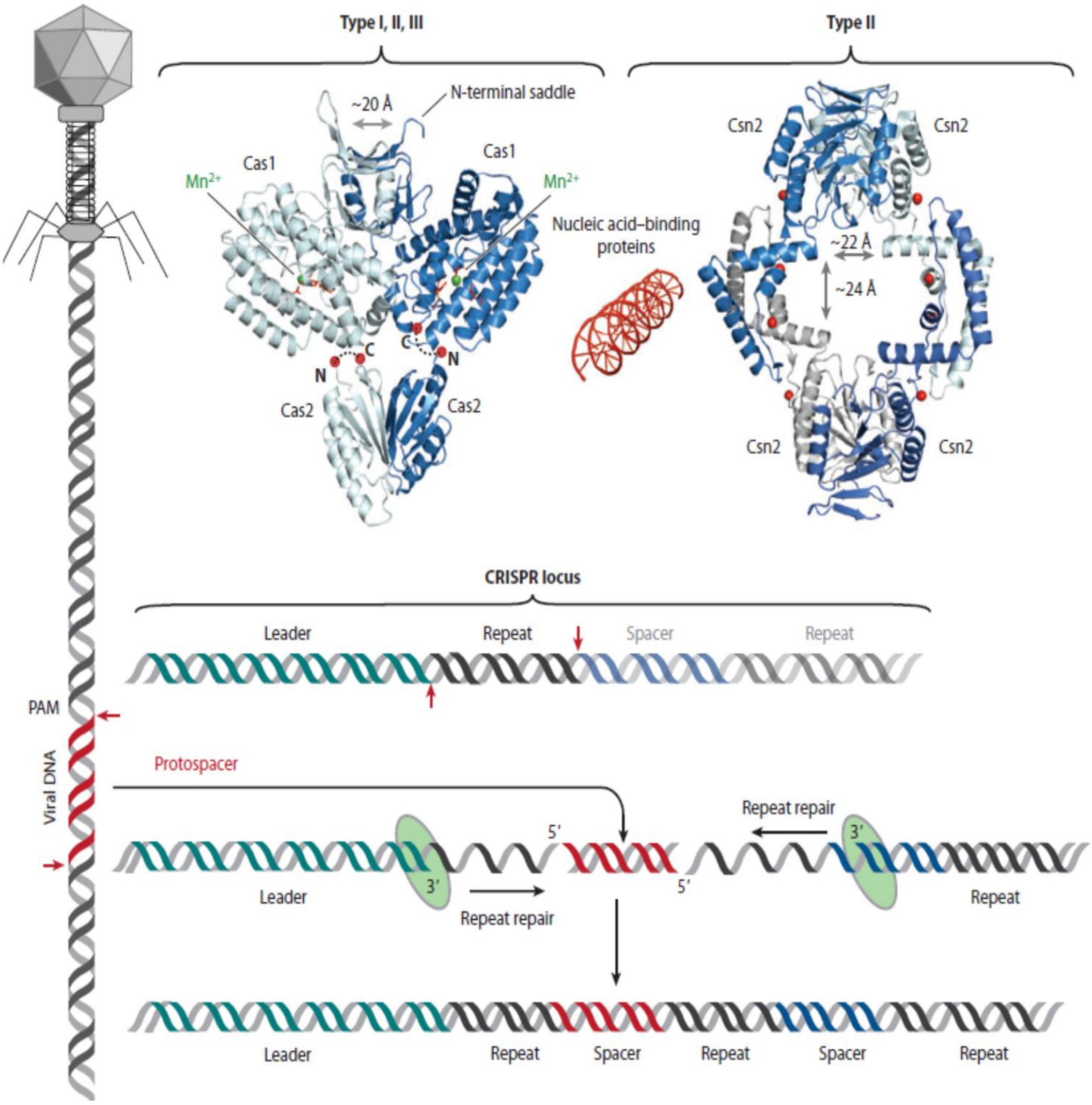


Figure 1.7 | Hypothetical mechanism of spacer sequence acquisition. The structures of cas1 and cas2 are shown on top. Selected foreign DNA for integration is known as a protospacer (red). Protospacers are flanked by PAM. Red arrows show the cleavage of the foreign DNA, which is integrated into the leader end of the CRISPR locus. The precise mechanism of this process is unknown. It is suggested that the leader-proximal repeat sequence is duplicated during the process, and cellular DNA repair proteins (green ovals) are required. Figure reprinted from Sorek et al., 2013, with permission.

Stage 2: CRISPR RNA biogenesis

After spacer acquisition, the CRISPR locus is transcribed into precursor CRISPR RNA (pre-crRNA), which is then processed into mature crRNA by the cleavage at the repeat sequences (Sorek et al., 2013), a step known as crRNA biogenesis. The endoribonucleolytic cleavage is done by cas6 homologues in type I and type III CRISPR systems, or by RNase III in type II systems (Brouns et al., 1993, Deltcheva et al., 2011). Cas6-like nucleases feature two RAMP domains, where pre-crRNAs are processed by breaking the phosphodiester bonds of the repeat regions (Wang and Li, 2012). The cas6 processed crRNA contains a 5′ handle and a 3′ hairpin structure. For type II systems, the initial stage of the crRNA biogenesis process include a tracrRNA binding to the repeat region of the pre-crRNA to form a double stranded structure - a complex can be recognized and processed by RNase III (Deltcheva et al., 2011). In this process, cas9 is also required to help the positioning for binding. The crRNA-tracrRNA complex is then trimmed by unknown enzymes to form the mature crRNA (as a hybrid) (Jinek et al., 2012). Mature crRNAs then associate with cas proteins to form complexes named crRNPs (Sorek et al., 2013). In type I and III CRISPR systems, crRNPs are composed of multiple cas proteins, whereas in type II CRISPR systems, only one cas protein (cas9) is involved. As discussed previously, the crRNP complex in type I CRISPR systems (Cascade) is composed of multiple cas protein subunits and a crRNA. Type III crRNP complex is similar to type I, and structure studies have shown homology between the two types (Van Der Oost et al., 2014). For type II systems, the studies have shown that there are two lobs in cas9: an alpha helical recognition lobe and a nuclease lobe. The former is for the coordination of guide RNA, while the latter is for PAM recognition and DNA cleavage (Jinek et al., 2014b).

Stage 3: target surveillance and interference

When invaded by non-self mobile genetic elements, crRNP complexes bind to the invasive DNA which has a protospacer sequence complementary to its own crRNA sequence. The recognition of non-self DNA sequences is possibly done through a PAM-based mechanism (Sashital et al., 2012; Gasiunas et al., 2012). When an invasion is detected, the crRNP complex base pairs with the protospacer (the 7-8 nucleotides seed region of the crRNA is exactly complementary to the target sequence) and extends, so the crRNP firmly binds to the invasive DNA (Semenova et al., 2011). This hybridization also triggers the conformational change of crRNP complex, which is thought to be a signal for nucleases to take action (Wiedenheft et al., 2011). In type I systems, the surveillance is largely done by nonspecific interactions at early stages. As long as PAM are recognized, cas proteins such as cse1 destabilize target DNA and facilitate the crRNA binding. The binding which is close to PAM is essential for CRISPR systems. Cas3 nuclease-helicase is recruited when signals from crRNP-target DNA complex are received (Gong et al., 2014).

For type III systems, detailed mechanism for target surveillance and interference remains largely unknown, but is expected of some similarities to type I systems based on structural information (Van Der Oost et al., 2014). For type II systems, the interference mechanism are unique, which involve a complex composed of a cas9 protein, a crRNA and a tracrRNA. The mechanism of PAM-based DNA recognition in type II systems is similar to that of type I. The cleavage of target DNA is done by nuclease lobe located in cas9 protein, as discussed previously.

CRISPR-Cas system is of great interests to researchers and has been widely studied over the past decade. The understanding and application of CRISPR systems are of particular interests to both industry and academia, including: (a) to help protect bacteria used in industry. Industries that based on bacteria fermentation such as yogurt and cheese production suffer greatly on phage caused losses (Brüssow, 2001). Understanding the defend systems of bacteria and using phage resistant lactic acid bacteria in starter cultures can save money; (b) to use phage as antibacterial drugs. As the resistance of bacteria to antibiotics becomes more common, phage therapies can potentially be used as an alternative; (c) applications as genetic tools. CRISPR-Cas systems can be used as novel tools for gene editing and gene expression regulation (Wiedenheft et al., 2012). CRISPR-Cas systems can be manipulated to incorporate the required specific sequences as spacers, and edit precisely at the point of interest (Sternberg et al., 2014). This feature makes the CRISPR-Cas systems very attractive. In fact, Type II CRISPR/Cas9 systems related tools are already commercially available, which have revolutionised genetic engineering approaches in mammalian cells (Cong et al., 2013; Mali et al., 2013).

DNA modifications

DNA methylation in prokaryotes is generally associated with DNA-protein interaction regulation. For bacteria and archaea, DNA methylation has been discovered on different sites of cytosine to form 4-methylcytosine (m^4C), 5-methylcytosine (m^5C) or 6-methyladenosine (m^6A) (Korlach and Turner, 2012). One known function of nucleic acid modification in prokaryotes is to protect themselves against foreign DNA. Restriction-modification (RM) is a well-studied defend system that can be found in prokaryotic organisms. In RM, a restriction endonuclease and a DNA methyltransferase work together. Restriction endonuclease degrades invasive DNA, and methyltransferase modifies and protects the host DNA, which otherwise can be targeted and cleaved by the endonuclease. Another important function of methylation in prokaryotes is for genome regulation (Casadesús and Low, 2006). For example, the timing of DNA replication and gene expression in Gammaproteobacteria is regulated by N6-methyladenine on GATC sequence by Dam methyltransferase (Marinus and Casadesus, 2009), and the cell cycle progression of Alphaproteobacteria is regulated by CcrM methyltransferase (Collier et al., 2007).

1.3.3 RNA modifications

RNA modifications are changes to the chemical composition of ribonucleic acid (RNA) after RNA been synthesised (post-transcriptional). RNA modifications are highly conserved across all known organisms (Li and Mason, 2014). Also, RNA modifications occur pervasively in almost all types of RNA, namely, messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), non-coding RNA (ncRNA), sometimes micro RNA (miRNA) and small nuclear RNA (snRNA) (Sun et al., 2016). Over 110 types of RNA modifications have been discovered, which can take place on all 4 regular ribonucleotides: A, U, G and C. In ribosomes, 60% of yeast and 95% of *E. coli* rRNA modifications occur in functionally important areas (Decatur and Fournier, 2002). In general, the changes to the chemical composition can alter the function and stability of RNA. The functions of RNA modifications remain largely uncharted and novel function are continuing to be revealed.

RNA modifications can be categorized into reversible and non-reversible (Li and Mason, 2014). Non-reversible modifications include mRNA post-transcriptional modifications such as 5′ and 3′ processing, splicing, and intron retention, while reversible modifications are usually associated with regulation, such as ribose methylation.

For mRNA modifications, six types of base modified nucleosides have been discovered so far (Figure 1.8), including N6-methyladenosine (m^{6}A), N1-methyladenosine (m^{1}A), inosine, 5-methylcytidine (m^{5}C), 5-hydroxymethylcytidine (hm^{5}C), inosine and pseudouridine, as reviewed in (Harcourt et al., 2017).

m^{6}A

m^{6}A is an abundant type of mRNA modification that has been discovered in most eukaryotes and some viruses (Niu et al., 2013; Aloni et al., 1979). This modification is catalysed by methyltransferase complex (METTL3, METTL14, WTAP) and may have important function as a gene expression regulator (Niu et al., 2013). More details regarding to m^{6}A are reviewed in chapter 5.2.

m^{1}A

m^{1}A is a reversible modification with a positively charged base and the structure fully blocks the formation of Watson-Crick base pairs, but enhances the RNA-protein interactions and the formation of RNA secondary structures. This feature makes m^{1}A a common modification in tRNA and rRNA. For mRNA, m^{1}A is likely involved in translation initiation (Harcourt et al., 2017).

m^{5}C

For m^{5}C, the role remains largely unknown. It is suggested that NSUN2 or TRDMT1 is involved in the conversion of cytosine to m^{5}C (Zhang et al., 2012; Thiagarajan et al., 2011).

hm^{5}C

Although the discovery of hm^{5}C in mRNA is relatively new, their existences are not uncommon. hm^{5}C is converted from m^{5}C by TET dioxygenases (Fu et al., 2014), and may be related to basic cellular process and development (Delatte et al., 2016).

Inosine

Inosine modification includes the conversion of adenine to inosine, which results in the alteration of base-pairing properties. Inosine can form base pairs with adenine, cytosine, and uracil (wobble base pairs). Adenosine deaminases ADAR1 and ADAR2 are responsible for mRNA inosine modification (Lehmann and Bass, 2000). As inosine modification alters base pairing preference, this modification is related to the change of encoded amino acids (Sommer et al., 1991).

Pseudouridine

As an isomer of uridine, pseudouridine is the most prevalent form of modified nucleosides in RNA. The conversion of uridine to pseudouridine is through PUS enzymes (Hamma and Ferré-D'Amaré, 2006). The function of pseudouridine is not clear. Possible roles include the regulation of gene expression and mRNA stability (Wang et al., 2013; Schwartz et al., 2014).

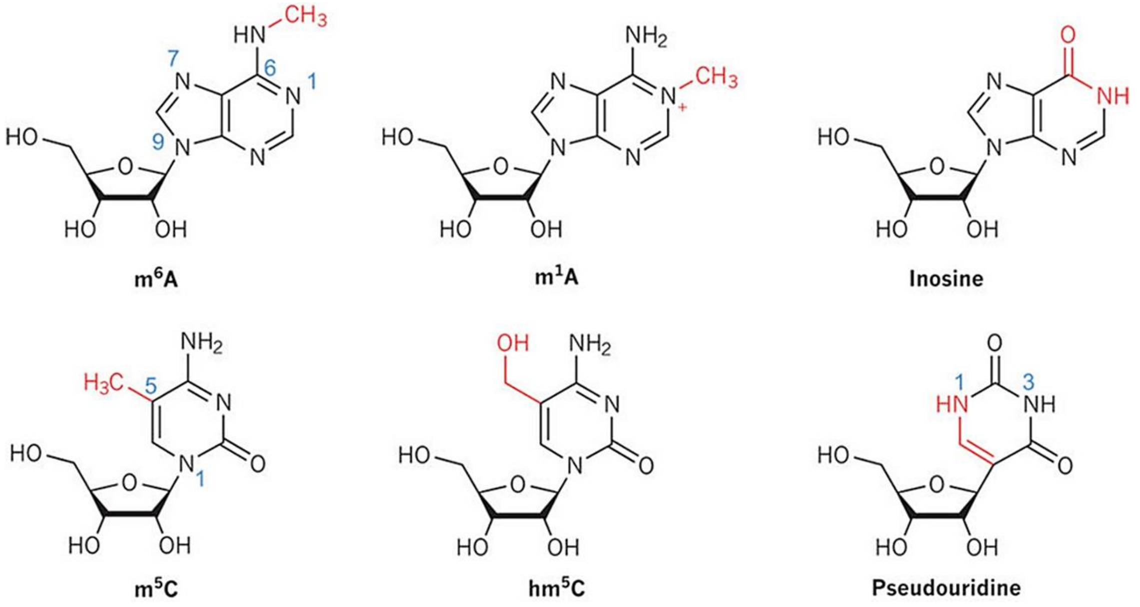


Figure 1.8 | Nucleoside base modifications in mRNA. Modification sites are coloured in red. Figure taken form Harcourt et al., 2017, with permission.

Post-transcriptional RNA modifications can be dynamic and might have functions beyond fine-tuning the structure and function of RNA. Understanding these RNA modification pathways and their functions may allow researchers to identify new layers of gene regulation at the RNA level.

1.3.4 Protein post-translational modifications

Post-translational modifications (PTMs) are the chemical modification of proteins on amino acid residues, which lead to the increase of diversity on proteome level. PTMs serve as a mechanism for cells to react to both internal and external changes. Over 200 types of PTMs have been identified (Mann and Jensen, 2003), and can be categorised into a) protein subunit proteolytic cleavage, b) the addition/removal of chemical groups with specific functions by covalent bonds and c) whole protein degradation. Based on the nature of modifications, some PTMs are reversible, such as methylation, ubiquitination, and glycosylation, while some are irreversible, such as deamidation and proteolysis (Beck-Sickinger and Mörl, 2006). PTMs usually lead to protein structure changes, which subsequently affect protein functions, such as mediating protein-protein interactions, or, in some cases, modified residues are specifically recognized for binding (Felder et al., 1993). PTMs are known to have important roles in a wide range of cellular processes including DNA repair, RNA processing, protein activity regulation, cellular localization, and signalling pathways (Walsh et al., 2005). These modifications occur almost at any point of a protein's life cycle. In some cases, PTMs are consequences of oxidative stress and are related to protein aggregation, thus PTMs are also believed to have roles in ageing and age-related disease, including arthritis, cardiovascular complications, respiratory disease, kidney disorders, neurodegenerative disorders (eg., Parkinson's and Alzheimer's disease) and cancer (Baraibar et al., 2012; Gaki and Papavassiliou, 2014; Santos and Lindner, 2017). The knowledge of PTMs can benefit the finding of therapies for these diseases, and have the potential to uncover more details about aging so the life expectancy can be extended.

PTMs are performed by modifying enzymes such as proteases, transferases, phosphatases, kinases and lipids, which count approximately 5% of the proteome (Beck-Sickinger and Mörl, 2006). Some widely studied PTMs are summarised in Table 1.1. More details will be discussed regarding arginine methylation and citrullination.

Table 1.1 | Summary of common protein PTMs.

PTM	Residue modified	Associated enzymes	Example functions
Phosphorylation	Ser, Thr, Tyr	Kinase and Phosphatase	Cellular process and signal transduction pathway regulation
Glycosylation	Asn (N-linked) Ser/Thr (O-linked)	Glycosyltransferase, Glycosidase	Protein folding, signal transduction, protein solubility
Ubiquitination	Lys	E1s, E2s and E3s	Signal destruction
Acetylation	Ala/Gly/Met/ Ser/Thr/Val (N-terminal), Lys	NATs (N-terminal)	Protein nucleic acid interactions, stability
Methylation	Arg, Lys	Methyltransferase	Gene expression regulation
Deamidation	Asn, Gln	Non-enzymatic process	Regulating interaction between protein-protein and protein-ligand
Hydroxylation	Lys, Pro	Hydroxylase	Regulating interaction between protein and ligand, protein stability
Citrullination	Arg	Protein-arginine deiminase	Affect protein hydrophobicity, folding and structure

Arginine methylation

Arginine methylation is prominent post-translational modification that can take place in both nuclear and cytoplasm. This modification is involved in many important biological processes including signal transduction, transcription regulation, subcellular localization, nucleus RNA/protein transportation and protein-protein interactions (McBride and Silver, 2001). Also, arginine methylation has been identified on histones affecting gene expression, and on hnRNPs affecting pre-mRNA processing and transportation (Fronz et al., 2008; Martin et al., 2010).

Arginine is positively charged with 5 interactive hydrogen bond donors, thus can be methylated in different ways. In eukaryotes, arginine residues are usually methylated into three forms: N^{G}-monomethylarginine (MMA), $N^{G}N'^{G}$ (symmetric) dimethylarginine (sDMA) and $N^{G}N^{G}$-(asymmetric) dimethylarginine (aDMA) (Figure 1.9A). Here G indicates that modifications take place at guanidino nitrogen atoms. Arginine methylation is carried out by a group of enzymes called protein arginine N-methyltransferase, or PRMTs, which use S-adensoyl-L-methionine (AdoMet) as methyl group donor. During the methylation process, arginine is first converted into MMA, which can be further converted into dimethylarginine with either asymmetric structure (by type I PRMTs, including PRMT1, 2, 3, 4, 6 and 8), or symmetric (by type II PRMTs, including PRMT5 and 9) (Yang et al., 2015; Blanc and Richard, 2017). Type III PRMT (PRMT7) only converts arginine to the form of MMA, and is known to use only histones as substrates (Feng et al., 2013).

PRMTs predominantly recognize glycine and arginine rich motifs, or GAR, which contain RGG/RG motifs (Thandapani et al., 2013). RGG/RG motifs are found in over 1000 proteins in humans, and are often involved in mediating protein-nucleic acid interactions. There are also PRMTs that recognize arginine neighbouring PGM (proline glycine methionine) rich motifs, such as PRMT4 (Yang and Bedford, 2013), or RxR sites in lysine/arginine rich regions, such as PRMT7 (Feng et al., 2013).

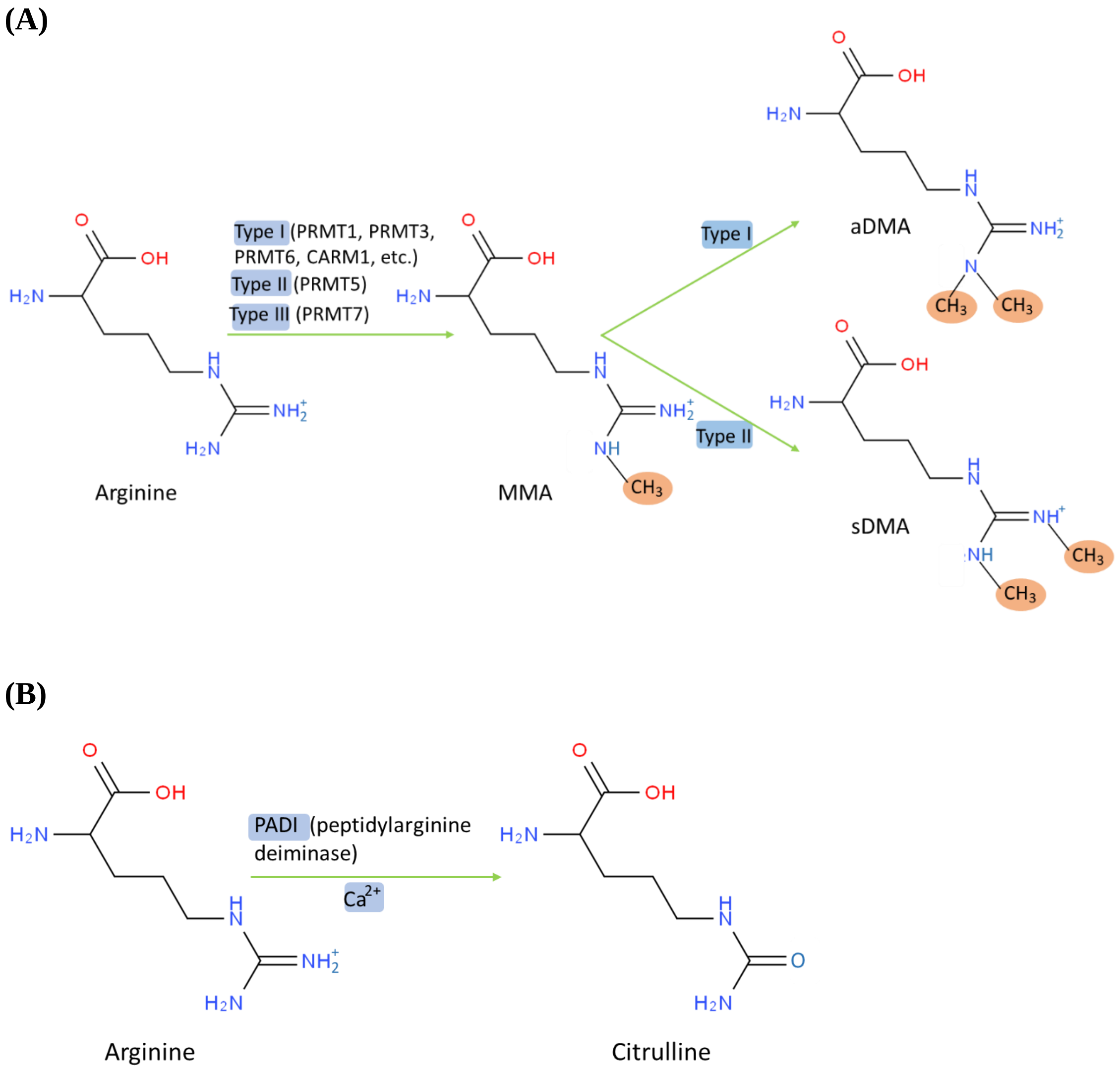


Figure 1.9 | Illustration of chemical modifications of arginine. (A) Arginine methylation **(B)** Arginine citrullination.

Citrullination

Citrullination, or deamination, is a post-translational modification where arginine residue is converted to the non-coded amino acid citrulline (Vossenaar et al., 2003). This modification is carried out by a family of enzymes called peptidylarginine deiminases (PADIs) (Wang and Wang, 2013). In this modification, ketamine group is replaced by ketone group (Figure 1.9B). Compared to arginine, citrulline has one less positively charge site (the site becomes neutral), which alters the pattern of interaction with other chemical groups. The function of citrullination remains largely unclear. Evidence has shown that citrullination may cause autoimmune diseases like rheumatoid arthritis (Coenen et al., 2007), as citrullinated proteins can be the target of immune systems

(Trouw et al., 2013). Citrullination related diseases may also include neurodegenerative disorders (e.g. Alzheimer’s disease), prion diseases, thrombosis and cancer (Martinod et al., 2013; Wang and Wang, 2013). Citrullination on histones is related to DNA damage response and transcriptional regulation (Cuthbert et al., 2004; Wang et al., 2004).

Among the family of PADIs, PADI4 is the only subtype with nuclear localisation signals and localized mainly in the nucleus (Asaga et al., 2001). PADI4 can usually be found in white blood cells (Vossenaar et al., 2004). In cancer cells PADI4 tends to be overexpressed, suggesting that it may be tumour related (Chang and Han, 2006). PADI4 plays an important role in nuclear functions, as it can citrullinate arginine (or monomethylated arginine) in various nuclear proteins, including histone 2A, 3 and 4. Upon citrullination, the effects of arginine methylation are turned off (Kouzarides, 2007; Wang and Wang, 2013).

Histone modifications

Histones are chromosomal proteins involved in gene expression regulations. A diverse array of PTMs can take place on histones include methylation, phosphorylation, acetylation, ubiquitination, crotonylation, sumoylation and more (Tan et al., 2011; Cubeñas-Potts and Matunis, 2013). The 4 core histones (H2A, H2B, H3 and H4) feature a long, highly conserved N-terminal tail, where most histone modifications take place (Davie, 1998). Residues such as arginine, lysine and serine are the common sites for histone PTMs. The modifications of histones have impacts on gene expression not only by altering chromatin structures, but also by recruiting effector proteins. The “histone code” hypothesis describes the combination of various histone PTMs through which different functions can apply accordingly. Some known functions of histone modifications include transcription regulation, chromosome condensation (Wilkins et al., 2014) and DNA repair (Li et al., 2013).

Given the importance of chemical modifications of nucleic acids and proteins, there are significant demands for high throughput, sensitive analytical methods that enable the identification and characterisation of chemical modifications of DNA/RNA and protein PTMs. The development of analytical methods to study protein-nucleic acids is also of significant interest in biology due to the importance of such interactions in a wide range of important cellular processes. Mass spectrometry is a powerful analytical tool that can identify, characterise and quantify chemical modifications of both nucleic acids

(DNA/RNA) and proteins. Furthermore, there are a number of applications of MS used to study protein-nucleic acids interactions, which will be discussed.

1.4 Mass spectrometry

1.4.1 Principle of mass spectrometry

Mass spectrometers measure the mass-to-charge ratio (m/z) of ions in the gas phase and have been widely used to analyse biomolecules including proteins and nucleic acids. A typical MS process would normally include the following steps. First, a MS instrument needs to generate gas phase ions from samples, a process known as sample ionization. Ionized molecules are then fragmented to smaller ions, which enter the analyser and are separated based on their m/z. Differentiated ions are then detected, with the signal intensity in proportion with abundance. The signals are then recoded, sorted by computer software and mass spectra are generated. The general scheme of a mass spectrometer is shown in Figure 1.10.

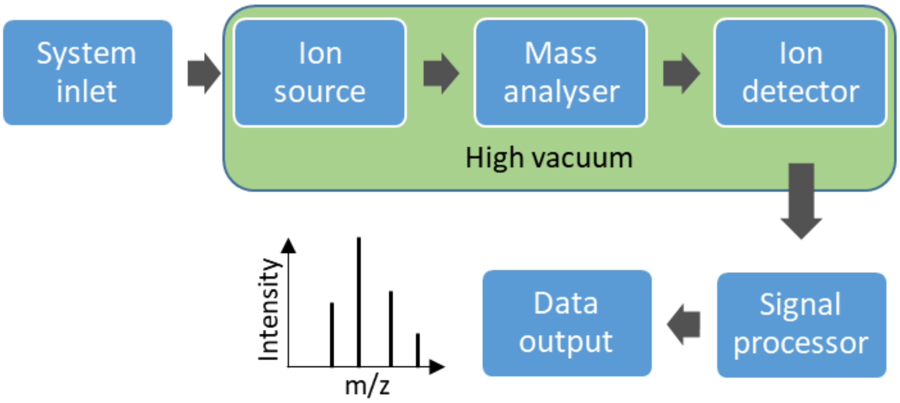


Figure 1.10 | General components and work flow of a mass spectrometer.

Sample introduction

Ionization is important, as the efficiency of ionization directly affects the sensitivity of MS. For biomolecules, two ionization techniques are commonly used: electrospray ionization (ESI) and matrix assisted laser desorption ionization (MALDI).

In ESI, sample (in solution) goes through a spray capillary, and ions are created by placing a high voltage (2-6 kV) between the capillary tip and the MS inlet (Figure 1.11A). Under this strong electric field, liquid sample that come out of the capillary sprayer is nebulized into fine mist of charged droplets (Figure 1.11B). Sample for MS analysis usually contains buffer with salt and detergent, which are introduced during sample preparation. Although ESI is tolerant to certain level of impurities, ionization results can be affected. This issue can be improved by using a tandem separation device (usually an HPLC), which is an advantage of the ESI technique (Bruins et al., 1987).

MALDI is ideal for the analysis of large biomolecules due to the minimal fragmentation of analytes and a high tolerance of contaminants (Jackson et al., 2005). In MALDI, samples are firstly mixed with matrix materials (e.g., 3, 5-dimethoxy-4-hydroxycinnamic acid) to be co-crystalized. Then a laser is applied to irradiate the matrix/sample which subsequently enters into gas phase through vaporization (Lewis et al., 2000). Most ions created by MALDI are singly charged, so clear mass spectra corresponding to the molecular weight of analytes can be generated (Hillenkamp et al., 1991).

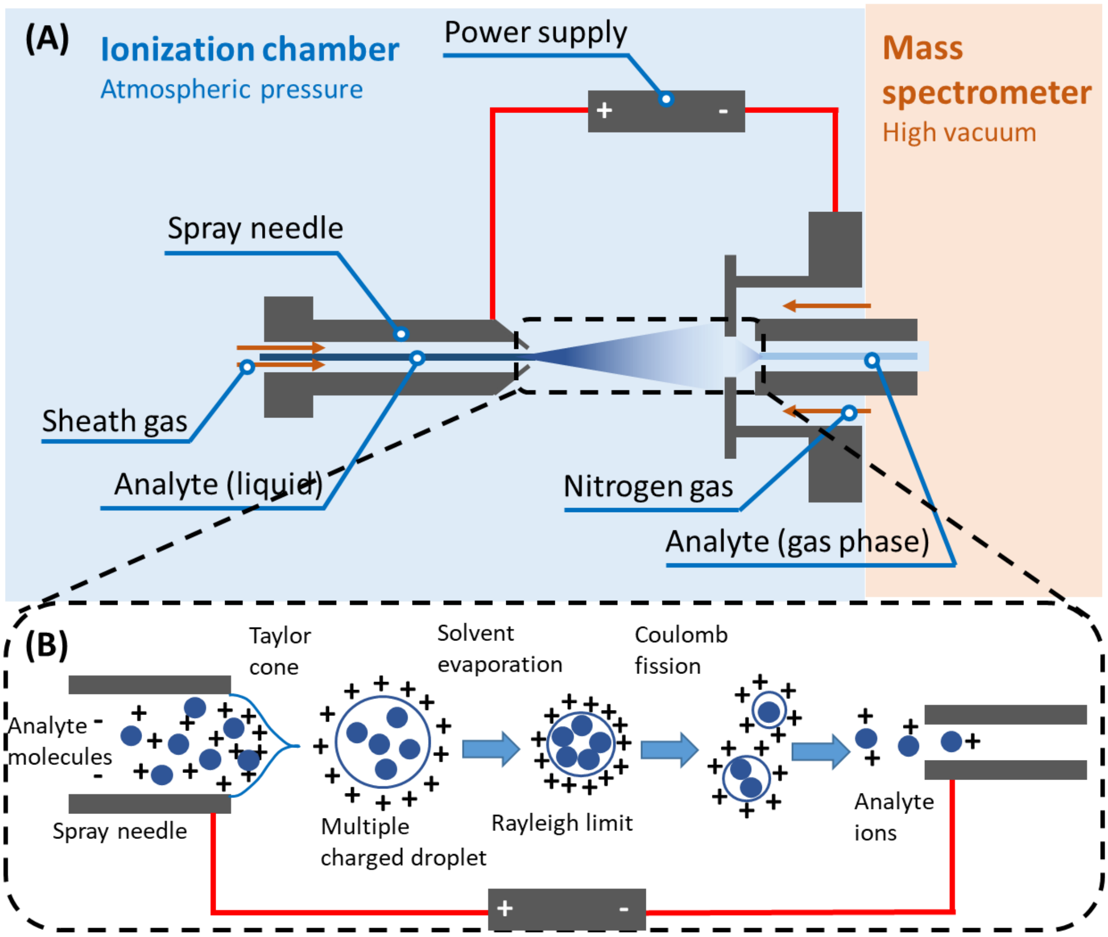


Figure 1.11 | Schematic of ionization process of an ESI instrument. (A) General composition of an ESI ion source. **(B)** Enlarged section (dash line area of **A**) showing the formation of gas phase ions.

Types of mass analysers

Mass analyser is a key component which determines the performance and resolution of a mass spectrometer. Different mass analysers can also be combined for improved performance. Here, three commonly used analyser types of mass spectrometer: quadrupole ion trap, time-of-flight, and orbitrap, are introduced.

Quadrupole ion trap mass analyser is composed of 4 parallel cylindrical rods, which are used for sample ion filtering based on m/z. When sample ions travel down the electric field created by the quadrupole, only those with certain m/z can pass (depend on voltage applied), while others with unstable trajectories will collide (Figure 1.12). By continuously altering the voltage, a range of m/z can be scanned (Hoffmann and Stroobant, 2007).

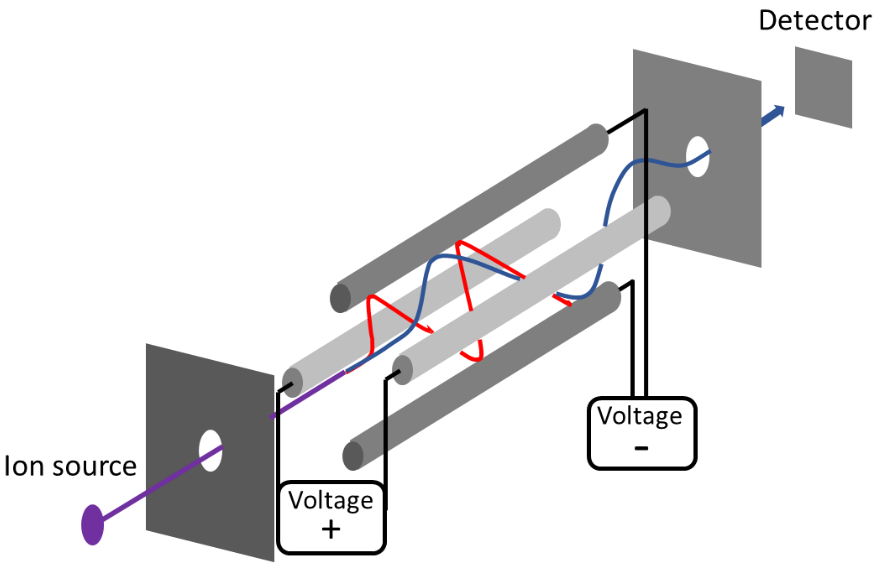


Figure 1.12 | Schematic of quadrupole ion trap mass analyser. Depend on the voltage applied to 2 parallel electrodes, resonant ions can pass and be detected (shown as blue), while non-resonance ions will not be detected (shown as red).

Time-of-flight analyser (TOF) is another widely used type of mass analyser. In TOF, desorbed and ionized sample molecules are accelerated to the same kinetic energy by the electrostatic filed applied, and forced to travel through a vacuum flight tube (Figure 1.13). As the accelerating voltage is constant, the time used for ions to travel to detector depends on their m/z. With the ions separated, a spectrum is recorded. In TOF, a reflectron (mirror used to reflect ions) is often used, which can increase the length of flight path without increasing the dimension of the analyser. More importantly, the use of reflectron can make the analyse results considerable more accurate, because a) reflectron focuses ions with the same m/z values, so the time ions used to reach the reflectron is fine tuned to assure the ions with the same m/z reach the detector at the same time, and b) kinetic energy is adjusted during the decelerate and re-accelerate processes on the reflectron (Bonk and Humeny, 2001; Wieser et al., 2012). Theoretically, TOF can analyse unlimited range of mass, but it is actually limited by the ion acceleration stage (Lee et al., 2011).

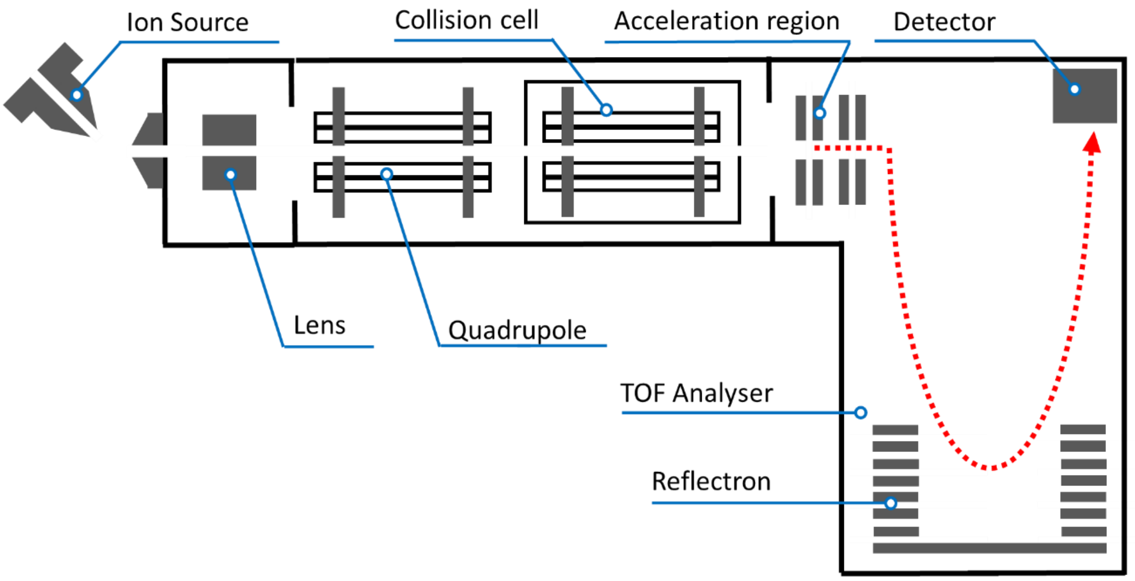


Figure 1.13 | Schematic of time-of-flight (TOF) mass analyser. Selected ions are fragmented and accelerated before entering the TOF analyser. A reflectron is usually used to increase the length of flight path. The travelling time of ions is measured and spectra are generated.

An Orbitrap mass analyser is a small electrostatic device with a spindle-shaped electrode in the centre (Figure 1.14). When ions of high energy are injected into this device, they orbit around the electrode, and form an axial motion current image which can be picked up by detector and then Fourier transformed (FT) into a mass spectrum (Makarov, 2000; Hardman and Makarov, 2003). In this type of mass analyser, m/z is determined by cyclotron frequency of ions. For Orbitrap systems, an external ion storage device is required prior to the analyser (Demartini, 2013). For this reason, a C-trap is often used as an external injection device (Makarov and Scigelova, 2010). Compared to other mass analysers, the Orbitrap has many advantages in terms of mass accuracy, linear dynamic range and resolution (Zubarev and Makarov, 2013). More recently, MS instruments with Orbitrap analyser in combination of a quadrupole mass filter (Q Exactive) have been developed, as shown in Figure 1.14.

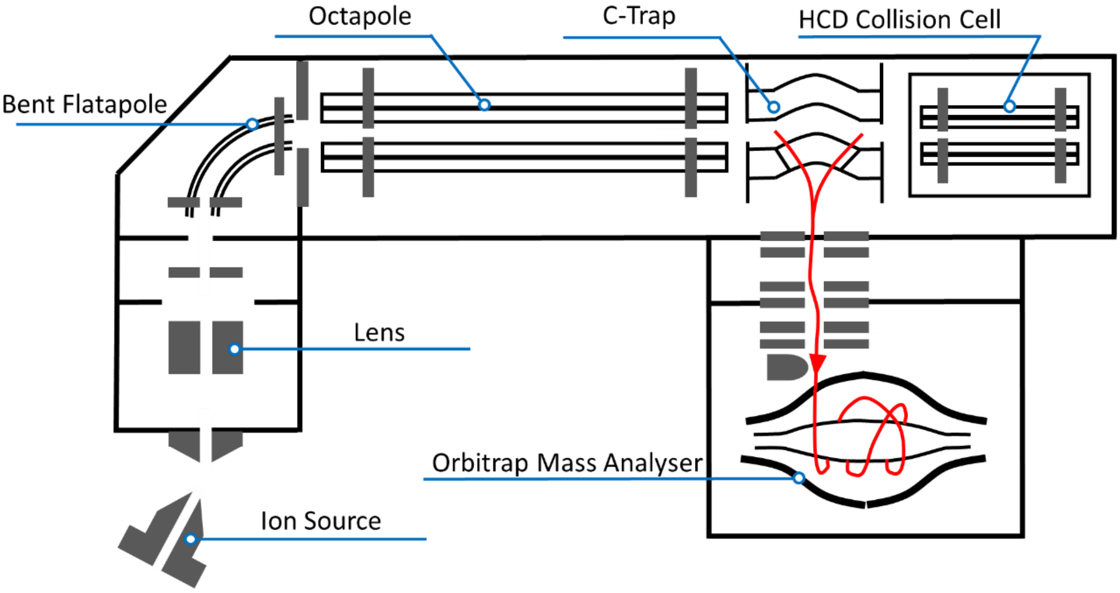


Figure 1.14 | Schematic of Q-Exactive HF mass spectrometer. This Exactive platform based instrument incorporates a selective multipole (e.g., quadruple or octapole) and an optional higher energy collision dissociation (HCD) collision cell interfaced to the C-trap. The drawing is not to scale. Red line shows the pathway of ions to the Orbitrap mass analyser.

1.4.2 Tandem mass spectrometry and peptide sequencing

Tandem MS (MS/MS, or MS^n), is a technique used to fragment selected precursor ions into smaller products ions. In case of protein analysis, selected peptides are fragmented (Figure 1.15). With the detailed information from MS/MS, peptide sequences, as well as post-translational modification sites can be identified.

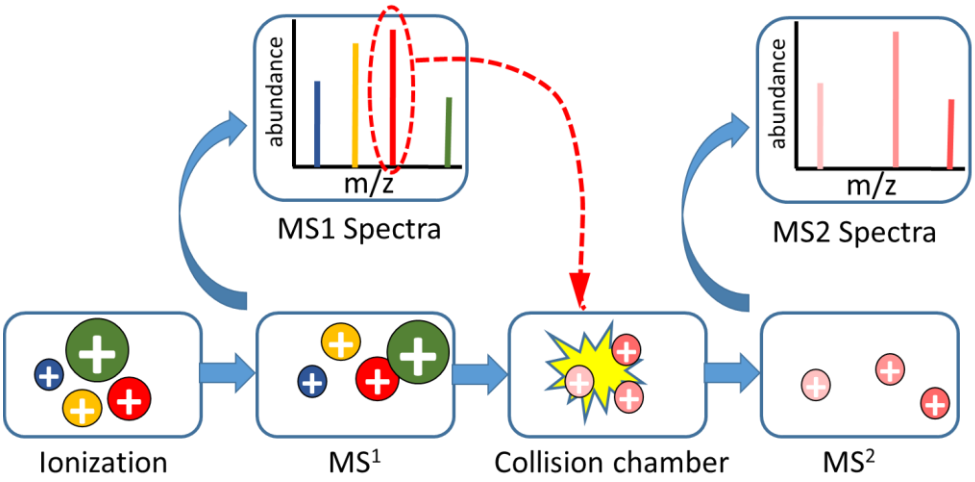


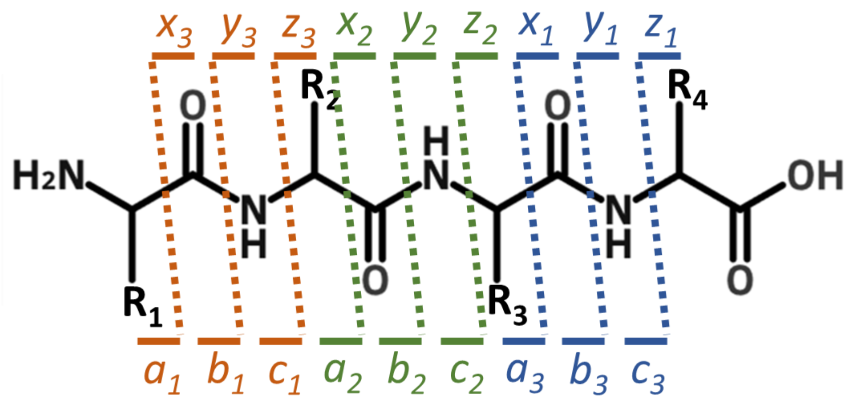
Figure 1.15 | Schematic of tandem MS. Sample ions are analysed in MS^1, then selected ions (as precursor ions, shown as red in this figure) from MS^1 are fragmented and analysed (MS^2).

Precursor ion fragmentation is an important step for proteomic analysis. Here, two commonly used fragmentation methods: collision induced dissociation (CID) and electron transfer dissociation (ETD) are discussed.

In CID, inert gas such as nitrogen, helium or argon, is used to break apart precursor ions by energetic collisions (Palzs and Suhal, 2005). For peptide analysis, amide bonds of peptide backbones are usually cleaved and protonated with CID fragmentation (Palzs and Suhal, 2005; Sleno and Volmer, 2004). As a result, complementary b ions and y ions are created (Palumbo et al., 2011). For ion type definitions, see Figure 1.16. In general, CID is effective for peptides with less than 20 amino acid residues and with no more than 4 charges (Mikesh et al., 2006). Different levels of collision energy can be applied for CID. High-energy collisional dissociation (HCD) is performed with collision energy of keV range, which is usually used in the C-trap of Orbitrap MS systems (Olsen et al., 2007), while low-energy CID can be performed with collision energy less than 100eV. Different collision energy gives different spectra. Appropriate collision energy level should be carefully determined for different applications.

ETD is a fragmentation method widely used in polymer analysis, peptide sequencing, and protein post-translational modification analysis (Brodbelt, 2016; Coon et al., 2005). In ETD, a radical anion extracted from a negative chemical ionization source is added to protonated peptides (charged with 2^+ or greater) where peptide backbone is disturbed, and cleavages take place at the N-Cα bonds (Cα is the carbon where the side chain is attached) (Kim and Pandey, 2012). Because of the position where cleavages take place, post-translational modifications of peptides can be preserved (Wiesner et al., 2008). Unlike CID, from which b ions and y ions are produced, in ETD c ions and z ions are produced. Compared to CID, ETD has no selectivity over the range of m/z. However, multiply charged ions are preferred for greater speed of react. For this reason, longer peptides with the potential of gaining more charges give better results when fragmented with ETD. Fragmentation results from ETD are predominantly generated based on the precursor ions with the highest charge status.

To use the information from fragmented peptides by tandem MS to interpret the precursor peptide sequences, a peptide sequencing nomenclature is required. A widely used nomenclature is the Roepstorff and Fohlman nomenclature (Roepstorff and Fohlman, 1984), as shown in Figure 1.16. The a, b and c represent the product ions with charge retained by the amino-terminal (or N-terminus) after fragmentation, while the x, y and z represent the product ions with charge retained by the carboxy-terminal (or C-terminus). The numbers after abc or xyz indicate how many amino acid R groups the product ions have, which are labelled from the original N-terminus of the precursor peptide (for abc), or from the C-terminus of the precursor peptide (for xyz). With this peptide sequencing information, proteins can be identified with different approaches such as using the Sequest algorithm, which features a score based autocorrelation technique with overlaps between theoretical spectra and experimental spectra mathematically determined (Eng et al., 1994), or the Mascot search engine, a probability based matching algorithm which uses a top-down method to match the predicted fragments with the experimental fragments (Perkins et al., 1999; Steen and Mann, 2004).



x_3 y_3 z_3 x_2 y_2 z_2 x_1 y_1 z_1
H_2N
R_1
R_2
R_3
R_4
O
N
H
OH
a_1 b_1 c_1 a_2 b_2 c_2 a_3 b_3 c_3

Figure 1.16 | Roepstorff and Fohlman nomenclature for peptide fragmentation.

1.4.3 Mass spectrometry for quantitative proteomics

In addition to the identification of peptides and PTMs, MS offers a powerful approach to generate quantitative data. In general there are two applications for MS based protein quantification: to determine the amount of proteins in individual samples (absolute quantification), and to compare protein amounts between multiple samples (relative quantification). Absolute quantification is to determine the amount of unknown proteins in samples based on the same proteins which amounts are known. Usually 3 or more peptides are to be compared in order to obtain justifiable quantification results. A more accurate way to do absolute quantification is to use isotope labelled peptides as internal standards. To do this, isotope labelled peptides are mixed with the unknown peptides and processed together prior to MS analysis, in which way the peptides can be directly related and compared. For relative quantification on the other hand, as no standard is required, the process is usually more cost-efficient and less time consuming. A number of quantitative proteomic workflows are discussed below.

Label-free quantification

Label-free quantification (LFQ) is a technique used to determine protein amount by comparing two or more samples directly without using any type of labelling. This technique requires samples prepared and analysed separately but with the same protocol. Protein amounts are estimated based on either the intensity of feature peptide spectra from the target proteins, or the count of peptide MS/MS spectra (Ciborowski and Silberring, 2016). In most cases, LFQ is only reliable when measuring samples with a large difference. However, compared to other quantification approaches, LFQ has advantages that a) samples are not interfered by tags or labels, and b) low costs for sample preparation. Also, with the improvement of instrument performance and data analysis software, LFQ is becoming more reliable.

Stable isotope labelling

Stable isotope labelling techniques track the passage of isotopes. For protein quantifications two labelling methods are often involved: *in vivo* metabolic labelling and *in vitro* chemical labelling.

For *in vivo* metabolic labelling, stable isotopes are added to the culture media so the microorganisms or cells can incorporate while they grow. For mammalian cell studies, a widely used labelling strategy is the stable isotope labelling by amino acids in cell culture (SILAC) developed by Mann and colleagues (Ong et al., 2002). SILAC is effective and accurate for MS based proteome quantification studies. In SILAC, amino acid of heavy labelled isotopes in cell culture media were incorporated metabolically into the whole cell proteome during protein synthesis. Compared to cells grown in non-labelled media (or light media), peptides digested from heavy isotope show a mass shift when analysed by MS. Based on the corresponding peak areas or intensities of the pair of heavy and light isotopes, relative peptide and protein quantification can be achieved (Figure 1.17). Commonly used isotopes for SILAC labelling include ^{2}H, ^{13}C and ^{15}N, or the combination of these. Advantages with SILAC quantification include a) the labelling process is relatively easy compared to other labelling methods, as the isotopic incorporation usually takes only 5-8 passages of cells to complete (Ong et al., 2002), b) no tags are added which may affect MS sensitivity, and c) quantification results are not prone to process errors, as samples are mixed beforehand and treated as one during the sample processing steps.

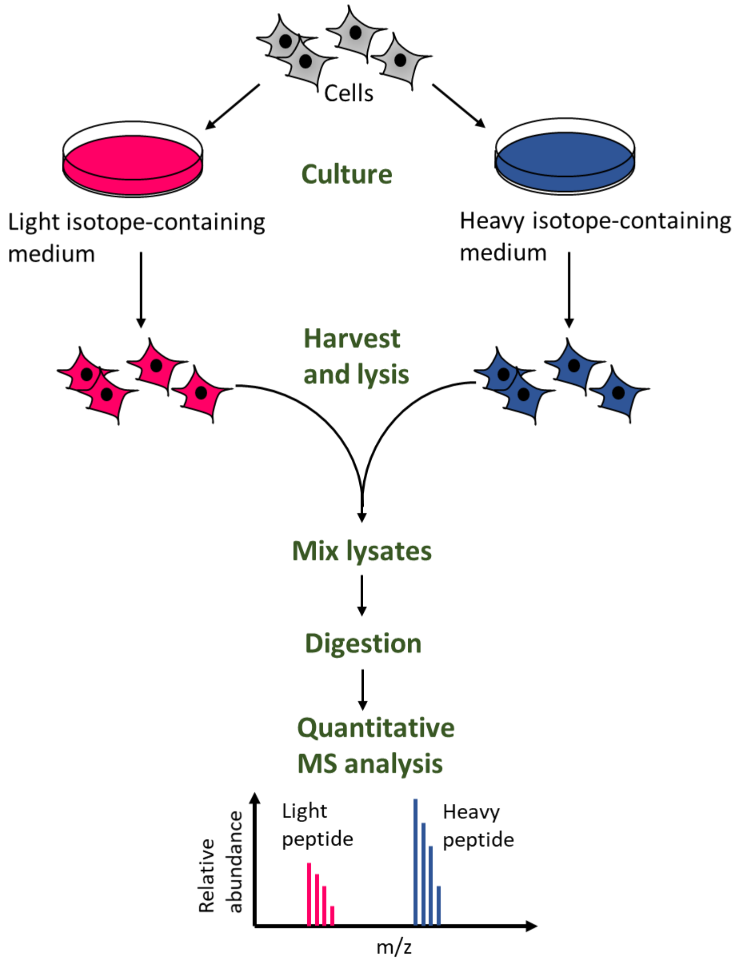


Figure 1.17 | A schematic representation of SILAC. Cells are grown in either normal medium or heavy isotope-containing medium, with one of which treated with the required experimental condition. After harvest, cells are lysed and mixed (usually with 1:1 protein amount), and then digested. When analysed by LC-MS, both the light and heavy versions of the peptides with the same sequence will co-elute, and quantifications can be made based on the ratio of the relative abundance of peptides.

For *in vitro* chemical labelling, post-harvest protein samples are labelled before or after proteolysis. Different labelling techniques can be used, for example, by targeting the thiol groups of cysteine residues using isotope coded affinity tags (ICAT) (Gygi et al., 1999), or by directly targeting the amino acid termini of peptides using isobaric tags for relative and absolute quantification (iTRAQ) (Ross, 2004), or using tandem mass tags (TMT) (Thompson et al., 2003). iTRAQ can be used to investigate multiple experimental conditions within a single experiment – usually four (4-plex) or eight (8-plex) conditions.

The principle of iTRAQ is the addition of isobaric mass labels at the amino termini or the lysine side chains of tryptic peptides in a digest mixture. The reagents are labelled in such a way that all labelled peptides are isobaric (hence the name) and have the same chemical properties, therefore indistinguishable during liquid chromatography separations (Ross, 2004). Labelled peptides yield so-called ‘reporter ions’ when fragmented (usually by CID) and can be used to quantify individual proteins within different experimental conditions. Similarly to iTRAQ, TMT labelling also uses isobaric mass tags. The reagents for TMT are comprised of the following regions: an amino acid tag linked to protein reactive groups, a mass normalizer, a cleavable linker and a mass reporter. The tags are designed in such a way that following fragmentation, TMT fragments are released to give rise to an ion with specific mass-to-charge ratio (Thompson et al., 2003). The advantage of using amino acid group targeting approaches such as iTRAQ or TMT over ICAT is that, in theory, every observable peptide can be labelled.

1.5 Aims

Chemical modifications of nucleic acids and proteins play important roles in regulating their interactions in the cell. In this Thesis I have developed and optimised mass spectrometry based methods to study the effect of chemical modifications of both nucleic acids and proteins in a number of important biological systems including:

1) Studying the effect of DNA modifications on CRISPR/Cas systems

2) Studying the effect of protein arginine methylation and citrullination on ribonucleoprotein complexes using quantitative proteomics.

In Chapter 3, the aim was to develop and optimise methods using PCR in conjunction with modified dNTPs to synthesise a range of target dsDNAs with various modifications. With these modified dsDNAs, *in vitro* assays were to be performed in collaboration with partners in Wageningen University in order to study the effects of DNA modifications on different types of CRISPR-Cas systems.

In Chapter 4, the aim was to develop analytical methods for the analysis of nucleoside modifications. Reverse phase HPLC coupled with either ultraviolet detection (UV) or mass spectrometry analysis (LC-MS) were to be used for nucleoside separation and identification. The method was to be applied on the validation of *in vitro* synthesised DNA in Chapter 3, as well as to characterise the modifications in other types of DNA including phage DNA and *E. coli* plasmid DNA.

In Chapter 5, the aim was to study the effects of protein methylation and RNA methylation on the function of mRNA interactome using *in vivo* mRNP capture assays. Stable isotope labelling with amino acids in cell culture (SILAC) in conjunction with mass spectrometry based quantification strategy was to be used to study the protein abundance differences under different experimental conditions. In addition, the optimisation of both the mRNP capture assay and the SILAC labelling were planned for this chapter.

In Chapter 6, the effects of protein methylation and arginine citrullination on proteins binding to the disease related expansion of GGGGCC repeats were to be studied. Large scale *in vitro* pulldown assays in conjunction with SILAC were to be used to determine the effects of different modification conditions on RNA GGGGCC repeat binding.

In Chapter 7 (additional chapter), the aim was to develop a mass spectrometry based absolute quantification method (AQUA) for anthrax vaccine products.

Chapter 2: Materials and Methods

2.1 Design and synthesis of dsDNA with modifications

Target dsDNAs containing a 98 bp spacer 8 (sp8) sequence with one of the following type of modifications: mC, hmC, U, hmU and phosphorothioate bonds, as well as the unmodified, were synthesised using PCR. The sp8 target sequence is underlined:

5'-**TTCATTTGATGCTCGATGAG**TTTTTCTAAAAGCTGACGACCGGGTCTCC
GCAAGTGGCACTTTTCCATGACCAAAAT**CCCTTAACGTGAGTTTTCGTT**-3'

Primers are highlighted in bold. The primer sequences designed for this work are:

Forward 5'-TTCATTTGATGCTCGATGAG-3' (BG8415), and
Reverse 5'-AACGAAAACTCACGTTAAGGGA-3' (BG8416).

Primers were synthesized by Eurofins Genomics. DNA oligos (both unmodified and modified) were synthesised using either *Taq* DNA polymerase or Q5 high-fidelity DNA polymerase (New England Biolabs, NEB), with specified reaction buffer included. dNTPs (A, T, C and G) were purchased from Promega, mdCTP from NEB, hmdCTP from Bioline, dUTP from INTEGRA biosciences, hmUTP and thio-dNTPs from TriLink Biotechnologies. DNA template used for the first round of PCR to synthesis unmodified dsDNA was the 937 bp plasmid with the sp8 sequence. The product dsDNA was then diluted 1 in 100, and used as the template for the subsequent PCR. *Taq* polymerase was used to incorporate dUTP, Q5 high-fidelity polymerase was used for the synthesis of dsDNA oligos with other modifications (including the unmodified one). Different reaction conditions for PCR were attempted to increase product yields, as described in Chapter 3. Optimised conditions for PCR in this study were as follows: total reaction volume of 50 µl with final concentration of dNTPs (or modified dNTPs) 200 µM, primers 0.5 µM, and 100 ng of DNA template. For *Taq* polymerase (used 1.25 U per 50 µl reaction), the following parameters were used: denaturation at 95 °C for 30 s followed by 30 cycles of reaction at 95 °C for 30 s, 46 °C for 60 s, 68 °C for 60 s, and the final extension at 68 °C for 3 min. For Q5 high-fidelity DNA polymerase (used 1 U per 50 µl reaction), the following parameters were used: denaturation at 98 °C for 30 s, followed by 30 cycles of reaction at 98 °C for 30 s, 61 °C for 30 s, 72 °C for 20 s, and the final extension at 72 °C for 2 min. DNA products were further concentrated by Concentrator Plus (Eppendorf) at 30 °C. Unmodified dsDNA was concentrated from 200 µl concentrated to 50 µl; dsDNA with mC from 250 µl concentrated to 50 µl; dsDNA with hmC from 300 µl concentrated to 50 µl; dsDNA with U from 500 µl concentrated to 50 µl; dsDNA with hmU from 600 µl concentrated to 50 µl; dsDNA with thio dNTPs from 600 µl concentrated to 50 µl. Concentrated dsDNA oligos were purified using QIAprep Spin Miniprep Kit (Qiagen), according to the manufacturer's protocol. The concentration of the purified samples was determined using NanoDrop 2000C (Thermo Scientific), then samples were analysed by native PAGE (see section 2.2).

dsDNA oligos with glucosylated hydroxymethylcytosine (ghmC) were enzymatically prepared by glucosylating the purified hmC dsDNA oligos as follows: T4 Phage β-glucosyltransferase (NEB) was incubated with 1 µg of 5-hmC dsDNA and 40 µM UDP-glucose in 1 × NEBuffer 4 (DTT 1mM, potassium 50 mM, magnesium acetate 10mM and Tris acetate 20mM, pH 7.9) at 37 °C for 16 h. The DNA products were subsequently purified and analysed by native PAGE in the same way as the other dsDNA oligos.

2.2 Native PAGE for DNA analysis

Native polyacrylamide gel electrophoresis (PAGE) was performed using a Mini-PROTEAN gel casting system (BioRad). A 10% gel was prepared by mixing 5 ml Acrylamide/Bis 19:1, 40% (w/v) solution (Thermo Fisher Scientific), 2 ml 10 × TBE and 13 ml dH_2O. 1 ml 10% ammonium persulfate (APS) and 40 µl Tetramethylethylenediamine (TEMED) were added to the prepared solution and mixed just before pouring the gel. 10 × TBE (1 L, pH 8.4) was prepared using 108g Tris base (Fisher Scientific), 55 g Boric acid (Sigma-Aldrich) and 9.31g Na_2EDTA (Thermo Fisher Scientific). APS and TEMED were purchased from Sigma-Aldrich. Electrophoretic buffer used was 1 × TBE. DNA loading buffer was purchased from NEB. The electrophoresis was operated using constant voltage of 180 V for approximately 30 min until loading dye reached the end of the gel. After running, the gel was rinsed in dH_2O and stained by up to 5 µl (10 mg/ml) of ethidium bromide (Sigma-Aldrich).

2.3 DNA digestion and dephosphorylation

DNA samples for HPLC or LC-MS analysis were first digested into nucleosides by adding 1 µl (5 units) DNA degradase (Zymo Research) into 44 µl DNA sample and 5 µl 10 × DNA degradase reaction buffer, mixed well and incubated at 37°C for 120 minutes. After incubation, 3 µl (3 units) shrimp alkaline phosphatase (NEB) was added and incubated at 37°C for another 60 minutes. The samples were then purified using a Nanosep Centrifugal Devices with Omega Membrane 3K, gray (Pall Corporation), and centrifuged at 5000g for 15 min and flow-through was collected. Nucleosides (A, T, G and C) were prepared by dNTP dephosphorylation and used as standards, described as follows: 50 µl reaction were prepared by adding 1µl 1mM dNTP (Bioline), 1 µl (1 unit) shrimp alkaline phosphatase (NEB), 5 µl 10 × NEBuffer 4 (NEB) and 43 µl water, mixed well and incubated at 37°C for 60 minutes. After incubation, samples were purified the same way as for the digested DNA samples.

2.4 HPLC analysis of nucleosides

Nucleoside analysis was performed on a Dionex UltiMate 3000 HPLC system (Thermo Scientific). For Accucore C30 column (50 × 2.1 mm, 2.6 µm particle size; Thermo Scientific). 5 µl samples were injected using a 20 µl sample loop. A 4 min gradient from 1% to 5% buffer B (40% acetonitrile, Fisher Scientific), then 9 min gradient to 50% buffer B, and then 1 min to 90% buffer B, and a 1 min gradient back to 99% buffer A (5 mM ammonium acetate, Sigma-Aldrich), was used. The flow rate was set at 0.2 ml/min. For Hypercarb porous graphitic carbon column (100 × 2.1 mm, 3 µm particle size; Thermo Scientific), a flow rate of 0.2 ml/min was used. 5 µl sample was loaded for each injection mixed with buffer A (10 mM ammonium acetate, pH 4.5, Sigma) and buffer B (40% acetonitrile, Fisher Scientific, with 10 mM ammonium acetate, pH 4.5, Sigma), with linear gradient of buffer B from 20% to 80% in 15 minutes, then switched to 100% buffer C (95% methanol, Fisher Scientific) for 3 minutes, then back to 20% B and 80% A in 1.1 minutes. For both columns, the temperature of column compartment was set at 30°C, and UV detector was set for 260 nm absorbance. All buffer percentages are v/v.

2.5 LC-MS analysis of nucleosides

The online LC coupled to MS was an UltiMate 3000 capillary LC system (Dionex), with different columns used for nucleoside separations. For Accucore C30 column (50 × 2.1 mm, 2.6 µm particle size, Thermo Scientific), buffer A (5 mM ammonium acetate, pH 4.5, Sigma) and buffer B (40% acetonitrile, Fisher Scientific) were used, with a linear gradient of buffer B from 3% to 20% in 17 minutes, then to 80% in 1 minute, at a constant flow rate of 100 µl/min. For Hypercarb porous graphitic carbon column (100 × 2.1 mm, 3 µm particle size, Thermo Scientific), a gradient elution was performed using buffer A (10 mM ammonium acetate, pH 4.5, Sigma) and buffer B (40% acetonitrile, Fisher Scientific, and 10 mM ammonium acetate, pH 4.5, Sigma), with a linear gradient of buffer B increase form 20% to 80% in 15 minutes, then switched to 100% buffer C (95% methanol, Fisher Scientific) for 3 minutes, then back to 20% B and 80% A in 1.1 minutes, at a constant flow rate of 100 µl/min. MS analysis was conducted on a maXis UHR TOF mass spectrometer (Bruker) equipped with an ESI nano sprayer source, Electrospray needle voltage was set to 4500V. Scan range of MS1 profile was set to m/z 100-800 in positive ion mode. Spectra data were analysed using Compass DataAnalysis software (version 4.1.359.0). All buffer percentages are v/v.

2.6 Hela cell nuclear extracts

Hela cell nuclear extract was obtained from Cilbiotech (CC-01-20-50) and was treated with Complete protease inhibitor cocktail (Roche). Extract was stored at -80 °C in aliquots before use.

2.7 Cell lines and culture conditions

SILAC culture media for HEK 293T cells were prepared using the SILAC Dulbecco’s Modified Eagle’s Medium (DMEM) (Thermo Fisher Scientific), supplied with 10% (v/v) dialyzed fetal bovine serum (dFBS) purchased from Gibco, Thermo Fisher Scientific. For light media, 0.46 mM L-Lysine and 0.47 mM L-Arginine (both from Sigma-Aldrich) were added. For heavy media (heavy isotope labelled), 0.46 mM $^{13}C_6$ $^{15}N_2$ L-Lysine-2HCl (Lys8) and 0.47 mM $^{13}C_6$ $^{15}N_4$ L-Arginine-HCl (Arg10) (both from Cambridge Isotope Laboratories, Inc.) were added. For both light and heavy media, 1.7 mM L-Proline (Sigma-Aldrich) was added as an option to reduce the conversion of arginine to proline.

All media were sterile-filtered using a 0.22 µm vacuum filtration unit (Rapid-Flow filters MF 75, Nalgene).

HEK 293T cells were cultured in 10 cm culture plates, with estimated 10 × 10^6 cells per plate. For mRNP capture assays, 8 × 10 cm plates of cells were used for each assay. For the SILAC mRNP capture assays, 8 × light plus 8 × heavy (a total of 16 × 10 cm plates) were used. For *in vitro* oligo RNA $(GGGGCC)_5$ pulldown assays, 4 × 10 cm plates of cells were used for each assay. For the SILAC *in vitro* oligo RNA $(GGGGCC)_5$ pulldown assays, 4 × light plus 4 × heavy (a total of 8 × 10 cm plates) were used. HEK 293T cells were cultured in collaboration with Sheffield Institute for Translational Neuroscience (SITraN).

2.8 Isolation of mRNA-binding proteins

After cell culture, media were removed from plates and cells were washed with 10 ml ice cold 0.1% DEPC treated PBS per plate. After wash, cells were crosslinked directly in plates at UV 254 nm (0.3 J/cm^2) on ice using a TL-2000 Ultraviolet Translinker (UVP). After, cells were lysed by adding 500 µl ice cold lysis buffer per plate. Lysis buffer was composed of 50 mM Tris HCl (pH7.5), 100 mM NaCl, 2 mM $MgCl_2$, 1 mM EDTA (pH8.0), 0.5% Igepal Ca-630, 0.5% Na-deoxycholate, RiboSafe RNase inhibitor (Bioline), 2 mM PMSF (Sigma-Aldrich), 1 × Complete protease inhibitor cocktail (Roche). All solutions were filter-sterilized (0.22 µm filter, Sigma-Aldrich) prior to use. After lysis, cells were collected using a rubber policeman into 1.5 ml tubes, and went through 10 × shear force on ice using a thin syringe needle. Cell debris was removed by centrifuging at 4°C, 13200 rpm for 5 min, and the supernatant was collected. Protein concentration was measured using the Bradford assay (Bio-Rad), as per manufacturer's instructions. Supernatant containing 1-2 mg proteins was used for each mRNP capture assay. Protein samples were denatured by adding an equal volume of 2 × binding buffer (20 mM Tris HCl pH 7.5, 1.0 M NaCl, 1% (w/v) SDS and 0.2 mM EDTA pH8.0), and then oligo $d(T)_{25}$ magnetic beads (100 µl packed-bed volume for each precipitation, NEB) were added into samples. Beads were equilibrated in 1 × binding buffer before use. The mixture with beads was then incubated at room temperature for 2 h on a rotating wheel. After binding, beads were collected on a magnetic rack and washed three times with 1 × binding buffer, while the supernatant was saved for a second and a third round of oligo (dT) precipitation. Complexes were then eluted using 65 µl elution buffer (10 mM Tris

HCl pH 7.5 and 1 mM EDTA pH8.0) for each pulldown, and treated with 10 µg RNase A at room temperature for 30 min with gentle agitation. Different batch of eluates were pooled together and subjected to SDS-PAGE, MS, or western blotting analysis. For SILAC experiments, heavy and light whole cell lysate were mixed 1:1 (estimated based on Bradford assay) before affinity purification.

2.9 Isolation of RNA repeat $(GGGGCC)_5$ binding proteins

After cell culture, cells were washed as previously described. After washing, cells were lysed with 500 µl lysis buffer per plate. Lysis buffer was composed of 1 × PBS, 0.1% Triton, 2 mM PMSF (Sigma-Aldrich) and 1 × Complete protease inhibitor cocktail (Roche). After lysis, cells were collected using a rubber policeman into 1.5 ml tubes, and went through 10 × shear force on ice using a thin syringe needle. Cell debris was then removed by centrifuging at 4°C, 13200 rpm for 5 min, and the supernatant was collected. Protein concentration was measured using the Bradford assay (Bio-Rad), as per manufacturer’s instructions. Supernatant containing 1-2 mg proteins was used for each pulldown assay, which was mixed with 10 µl (1 µg/µl) $(GGGGCC)_5$ RNA oligos with 3' biotin modification, 1 × PBS with 0.1% Triton, 1 µl RiboSafe RNase inhibitor (Bioline), 2 mM PMSF, and 1 × Complete protease inhibitor cocktail (Roche) in a 1 ml reaction, then incubated at room temperature for 15 min, and on ice for another 15 min. Mixture was then transferred to a 6 cm petridish and crosslinked on ice at UV 254 nm (0.3 J/cm^2) using a TL-2000 Ultraviolet Translinker (UVP). Mixtures were then applied to 50 µl streptavidin sepharose beads (GE healthcare) which were blocked overnight in solution containing 1 × PBS, 0.1% Triton and 1% bovine serum albumin (BSA) on a rotating wheel at 4°C. For the binding process, the mixture with beads was incubated at 4 °C for 2 h on a rotating wheel. After binding, beads were washed three times with 1 × PBS plus 0.1% Triton, and then washed twice with 1 × PBS. For the elution process, 65 µl 1 × PBS for each pulldown was used, with 10 µg RNase A added and incubated at room temperature for 30 min with gentle agitation. Eluates were analysed by SDS-PAGE, followed by MS or western blotting analysis.

For the SILAC experiments, the same amount of cell lysates for heavy and light went through the pulldown process individually, and mixed together prior to analysis using SDS-PAGE.

2.10 Cell treatment for different modification status

Adenosine, periodate oxidized (AdOx) was purchased from Sigma-Aldrich. For AdOx treatment, AdOx was added to media with concentration of 20 µM for HEK 293T cell growth 48h prior to harvest. AdOx (20 µM) was also added to lysis buffer and binding buffer to maintain the desired modification environment.

For N^6-methyladenosine (m^6A) study, an m^6A deficient HEK 293T cell line was obtained from the laboratory of Prof. S Wilson, University of Sheffield. The cell line was generated with the knockdown of Wilms tumour 1 associated protein (WTAP) using RNAi, aiming to reduce the level of RNA m^6A methylation.

For induced PADI4 overexpression, HEK 293T cells were transfected with HA-PADI4 48 hours prior to harvest.

2.11 Photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation

The procedure for photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) is the same as regular crosslinking and immunoprecipitation (CLIP), except the following: 4-Thiouridine (4-SU) (Sigma-Aldrich) was added to cell culture 16 h prior to harvest, at a final concentration of 100 µM. 4-SU treated cells were crosslinked at UV 365 nm (0.2 J/cm^2) on ice using a TL-2000 Ultraviolet Translinker (UVP).

2.12 Total RNA extraction

Whole cell lysate used for RNA extraction was obtained as described in previous materials and methods section. RNA extractions were performed using TRIzol reagent (Invitrogen) as per manufacturer’s instructions. Briefly, TRIzol reagent was added (3 times to the volume of the cell lysate) and left at room temperature for 10 minutes, then chloroform was added 1/5 the total volume of the sample and TRIzol mixture, and shook for 20 seconds before left setting down at room temperature for 3 minutes. The sample was subsequently centrifuged at 11000 rpm for 10 minutes at 4°C, the upper layer containing RNA was transferred into a fresh tube, and 1 µl glycogen added. An equal volume of isopropanol (equal to sample volume) was added and 3 M sodium acetate (1/10 sample volume) added, incubated at -20 °C overnight. After incubation, pellet was washed with 150 µl 70% ethanol in DEPC treated water, vortexed and centrifuged at 13000 rpm for 5 minutes at 4 °C, then air dried for 15 minute at room temperature. RNA was stored at -20 °C. For further analysis, RNA was re-suspended in 20 µl nuclease free water and the concentration was determined using NanoDrop 2000C (Thermo Scientific).

2.13 SDS-PAGE for protein analysis

Protein separation with 1D sodium dodecyl sulphate – polyacrylamide gel electrophoresis (SDS-PAGE) was carried out using the Mini-PROTEAN Tetra Vertical Electrophoresis Cell (Bio-Rad) gel casting system. 12% gels were used in this work and showed good resolution. The gel preparation steps are as follows: For the resolving part of the gel, 3.5 ml deionized water, 2.5 ml 4 × lower buffer (1.5 M Tris base, 0.4% SDS, adjusted pH to 8.8 with HCl, filter sterilised), 4 ml 30% acrylamide, 50 µl freshly prepared ammonium persulfate (APS), and 20 µl tetramethylethylenediamine (TEMED) were mixed and poured into a Bio-Rad gel apparatus. APS and TEMED were purchased from Sigma-Aldrich. Gels were layered with 100 µl isopropanol on top and allowed 15-30 minutes for settling (polymerisation). Then the isopropanol on top of the gel was discarded prior to pouring the stacking gel mixture to the top of the settled resolving gel. The stacking gel mixture was composed of 5.8 ml deionized water, 2.5 ml 4 × upper buffer (0.5 M Tris base, 0.4% SDS, adjusted pH to 6.8 with HCl, filter sterilised), 1.7 ml 30% acrylamide, 50 µl freshly prepared APS and 20 µl TEMED. A comb was inserted to form the loading wells, and let settle for another 15-30 minutes. When settled, gel was assembled into the gel casting system filled with 1 × SDS running buffer (15 mM Tris, pH 8.3, 1.5 M glycine, 0.1% SDS). Before loading to the gel, protein samples were mixed with 4× loading buffer (200 mM Tri-HCl pH 6.8, 8% SDS, 0.08% bromophenol blue, 50mM EDTA, 4% β-mercaptoethanol, 40% glycerol) and denatured at 94 °C for 5 min. In general cases, 16 µg sample were loaded to each well of the gel, together with 5 µl protein marker. Gels were run with voltage 80 V for stacking (10-15 min), and 180 V for resolving (until loading dye reach the bottom of the gel). After running, gels were rinsed and stained with InstantBlue (Expedeon) overnight, and distained for 4 hours using distilled water.

2.14 Western blotting

The concentration of protein in samples was determined using Bradford assay prior to SDS-PAGE analysis. The proteins on the gel were then transferred to a nitrocellulose membrane (Whatman) using an electroelution (or electrophoretic) method, described briefly as follows: filter papers, nitrocellulose membrane and gel were soaked with transfer buffer (40 mM glycine, 50 mM Tris, 0.04% SDS and 20% methanol) and a sandwich was made from top to bottom: filter paper layers, gel, nitrocellulose membrane, filter paper layers. Proteins were transferred from gel to nitrocellulose membrane using a Biometra Fastblot device (Analytikjena) at 15 mAMPS per gel. After transfer, membranes were blocked at room temperature for 1 hour in blocking buffer containing Tris-buffered saline (TBS), 0.1% Tween-20 and 5% fat-free milk (Marvel). Membranes were then incubated in blocking buffer with primary antibodies for 1 h at room temperature or overnight at 4°C. After, membranes were washed 3 times (10 minutes each) in TBS with 0.1% Tween-20 before incubation with secondary antibodies in block buffer for 1 hour at room temperature. After incubation, membranes were washed and prepared for chemiluminescent signal detection with SuperSignal West Pico Chemiluminescent substrate (Thermo Fisher Scientific) following manufacturer's instructions. Signals were detected with a Gbox imaging system (SynGene). Antibodies and dilutions used were: CHTOP antibody 1/2500, HA antibody 1/2500, α-tubulin 1/10000, anti-mouse-HRP 1/4000.

2.15 Dot blotting for total RNA

The concentration of extracted total RNA of wild type HEK 293T cells and the WTAP gene knockdown HEK 293T cells were measured by NanoDrop 2000C (Thermo Scientific) before dot blotting. For each dot, 2 µl (approximately 180 ng) of total RNA extract were dropped on an Immobilon-P PVDF Membrane (0.45 µm pore size, Sigma-Aldrich), and dots were left air dried at room temperature for 15 min, followed by another 2 µl (approximately 180 ng) of the same sample dropped to the same position. Dots were then left at room temperature and air dried for another 1 hour. After, the membrane was blocked in Tris-buffered saline (TBS) with 5% fat-free milk (Marvel) and 0.1% Tween-20 for 1 hour. The membrane was then incubated with a primary antibody (rabbit anti α-m^{6}A 1:500) in blocking buffer at room temperature for 1 hour. After incubation, the membrane was washed 3 times (10 min each) in TBS with 0.1% Tween-20, then incubated with a secondary antibody (anti rabbit 1:10000) in blocking buffer overnight at 4°C. After washing, SuperSignal West Pico Chemiluminescent substrate (Thermo Scientific) was used for the membrane signal detection, following manufacturer's instructions. Signals were detected using a G:BOX imaging system (Syngene).

2.16 Trypsin digestion

In-gel protein digestion

Gel lanes with proteins were sliced into 1 × 1 mm pieces, which were collected into 6-8 tubes and de-stained in 200 mM ammonium bicarbonate (ABC, Sigma-Aldrich) with 40% acetonitrile (ACN, Fisher Scientific) for 30 minutes at 37 °C twice, then dried using centrifugal evaporator (Eppendorf). After, gel pieces were reduced with 10 mM dithiothreitol (DTT, Sigma-Aldrich) in 50 mM ABC for one hour at 56 °C, and then alkylated with 55 mM iodoacetamide (IAA, Sigma-Aldrich) in 50 mM ABC in the dark for 30 minutes. Then gel pieces were washed with 50 mM ABC at room temperature for 15 minutes, twice, and then washed a third time with 50 mM ABC in 50% ACN at 37 °C for 15 minutes. Afterward, gel pieces were dried using Concentrator Plus (Eppendorf), and then digested with 0.4 µg trypsin (Sigma-Aldrich) per reaction in 40 mM ABC, 9% ACN overnight at 37 °C. Supernatants containing digested peptides were collected, additional extraction of peptides was performed by the addition of acetonitrile and incubated at 37 °C for 15 minutes, prior to the addition of 5% formic acid (FA, Fisher Scientific), incubated at 37 °C for 15 minutes, supernatants were collected. A third recovery step was performed with 50% ACN in 5% FA, incubated at 37 °C for 30 minutes, supernatants were collected. Supernatant collected from the three-step recovery were pooled together and concentrated to dryness using Concentrator Plus (Eppendorf) at 30 °C, vacuum-aqueous (V-AQ) mode, and stored at -20 °C. For LC-MS/MS analysis, samples were re-suspended in 0.1% trifluoroacetic acid (TFA, Sigma-Aldrich) in HPLC grade H_2O (Fisher Scientific).

In solution protein digestion

For 50 µl reactions, protein samples were added to 50 mM ammonium bicarbonate (Sigma-Aldrich), 200 ng (final concentration 4ng/µl) of trypsin (Sigma-Aldrich) together with 15 µg (1.5 µl, 10 µg/µl) ProteaseMAX Surfactant, Trypsin Enhancer (Promega). Samples were incubated at 37 °C overnight. Reaction were stopped by adding 5 µl 1% (v/v) trifluoroacetic acid (TFA) (Sigma-Aldrich). Samples were then concentrated to dryness using Concentrator Plus (Eppendorf) at 30 °C, vacuum-aqueous (V-AQ) mode. Dried samples were either stored at -20 °C, or re-suspended in 10 µl of 0.1% (v/v) TFA for LC-MS analysis.

2.17 Protein LC-MS analysis

For experiments using HEK293 cell cultured in non-SILAC media or SILAC media, digested protein samples (in-gel digested or in-solution digested) were analysed using an online liquid chromatography tandem mass spectrometry (LC-MS/MS). A nano-flow UltiMate 3000 ultra-high performance liquid chromatography (UHPLC) (Thermo Fisher Scientific) and a Q Exactive HF hybrid Quadrupole Orbitrap mass spectrometer (Thermo Fisher Scientific) were used. Samples were first re-suspended in 0.1% TFA and eluted from a 50 cm × 75 µm Easy-spray PepMap C18 analytical column (Thermo Fisher Scientific), with the column temperature maintained at 40 °C. Mobile phase A used was water with 3% acetonitrile (ACN) plus 0.1% formic acid (FA), and mobile phase B was water with 80% ACN plus 0.1% formic acid FA. Loading buffer was 0.1% TFA and 3% ACN. All mobile phase components were HPLC grade and were purchased from Fisher Scientific. All buffer percentages are v/v. Flow rate for peptide elution was set at 300 nL/min with a gradient of 3% to 40% buffer B over 60 minutes. For peptide ionisation, a spray voltage of 2.1 kV was used, with S-lens RF level set at 60, and heated capillary at 250 °C. For scanning, MS1 resolution of 120,000 at m/z 200 and MS2 resolution of 15,000 at m/z 200 were used. Full scan target was set at 3E6 with maximum fill time of 100 ms. A mass range of 375-1500 was chosen. A value of 5E4 was set as target value for fragment scans, with intensity threshold kept at 5E4. Isolation width was set at 1.2. Fixed first mass was set 100, and a value of 28 was set for the normalized collision energy. The ‘preferred’ option was used for peptide match, with isotope exclusion mode on. Positive mode (profile) was used for data acquisition.

For samples analysed using the amaZon or the maXis mass spectrometer, see section 2.20.

2.18 MS data processing and analysis

For data generated from the amaZon or the maXis mass spectrometer (data file type .d), data were processed using Compass DataAnalysis software (Version 4.1, build 359, Bruker Daltonik GmbH) and Mascot generic format (MGF) files were generated using the default ion trap processing method. MGF files were then processed using Mascot Daemon software and searched against the SwissProt database with the Homo sapiens (human) taxonomy selected. Carbamidomethyl (C) was set as fixed modification and oxidation (M) as variable modification. Decoy database was selected, and max missed cleavage was set to 1. The option of peptide charge 2^+, 3^+ and 4^+ was selected, with peptide tolerance of ±0.6 Da.

MS data acquired from the Q Exactive HF (data file type .raw) were processed by MaxQuant (version 1.5.3.30) with built-in Andromeda search engine, and searched against the Uniprot human database containing 159615 protein sequence entries (downloaded from www.uniprot.org on 12/08/2017). In MaxQuant, trypsin/P (cleaves after c-terminuses after lysine and arginine, including which proline follows) was selected as the digestion enzyme option, with a max missed cleavages of two. Carbamidomethylcysteine was set as fixed modification, while methionine oxidation and N-terminal acetylation were set as variable modification. A minimum of one unique peptide was required for protein identification, or a non-unique peptide can be assigned to the protein group of the highest peptide number (known as Occam's razor peptide) for identification. Peptides with less than 7 amino acids, or with mass over 4600 Da, were not considered for protein identification or quantification. The peptide tolerance was set at 4.5 ppm, and threshold was set at 500.

For samples prepared from in-gel digestion, the "match between run" option was selected for a combined search/identification between different gel bands of the same gel lane for the non-sequenced or non-identified peptides.

For the SILAC samples, multiplicity was set as 2, with the heavy label Arg10 and Lys8 option selected. A maximum of three arginine or lysine labelled amino acids per peptide were allowed. Pro6 (Proline 6) was added as a variable modification for the arginine to proline conversion. Up to five modifications per peptide were allowed. Unmodified unique or razor peptides were used for quantification, taking consideration of peptides with modifications including methionine oxidation, N-terminal acetylation and proline 6. The re-quantify option was selected for the missing SILAC pair identifications. For other settings the default options were used. SILAC ratios were normalized to the median of peptide levels where the $\log_2$ ratios equal to zero.

For data type generated from the Q Exactive HF system (.raw) to be analysed by Mascot, the files were firstly converted to file type (.mgf) using software ProteoWizard (http://proteowizard.sourceforge.net/downloads.shtml) and then loaded to Mascot.

2.19 Further MS data processing and bioinformatics analysis

Further proteomics data processing (post-MaxQuant) and bioinformatics analysis were carried out using Perseus software (www.perseus-framework.org, version 1.5.2.6). MaxQuant generated data files “proteinGroups.txt” were loaded to Perseus. In Perseus, protein groups were firstly filtered, with irrelevant groups including “only identified by site”, “reverse” and “potential contaminants” removed. For the rows matching, gene name was used. Gene ontology analysis was performed using the Panther classification system (pantherdb.org).

For data from SILAC experiments, data were automatically normalised to adjust shifts of protein ratio (H/L or L/H) distribution. Unless stated otherwise, SILAC data of ratios presented in this work are all normalised. One-sample t-tests were performed within Perseus. The P-value threshold was set as 0.05, and the P-values were given as –log10. Significantly changed protein groups were then filtered based on t-test difference ≤ -0.585 or ≥ 0.585 (1.5 fold change).

2.20 Absolute quantification for anthrax vaccine

The preparation of stable isotope labelled standard peptides

Stable isotope labelled standard peptides (SIS peptides) with sequence of DLNLVER, NNIAVGADESVVK and NQTLATIK were synthesised by Abingdon Health. SIS peptides (powder) were weighted, dissolved in distilled water with additional guanidine hydrochloride. The concentration of the SIS peptides was determined by amino acid analysis (AAA). AAA was performed by Protein and Nucleic Acid Chemistry Facility, University of Cambridge.

Anthrax vaccine sample preparation

Anthrax vaccine samples (including two types of samples: rPA and filtrate) received from Porton Biopharma were either diluted with distilled water, or concentrated using a 3K Nanosep Centrifugal Devices with Omega Membrane (PALL Corporation). The concentration was performed as per manufacturer's instructions, briefly, 500 µl sample was added to the filter and centrifuge at 13000 × g for 12.5 minutes, flow through was discarded and sample retained on the membrane was collected by pipetting. After dilution/concentration, anthrax vaccine samples were mixed with the required amount of SIS peptides and digested with trypsin. For each 50 µl reaction in 50 mM ammonium bicarbonate (Sigma-Aldrich), 200 ng (final concentration 4 ng/µl) trypsin (Sigma-Aldrich) together with 15 µg (1.5 µl, 10 µg/µl) ProteaseMAX Surfactant, Trypsin Enhancer (Promega) were added. Samples were incubated at 37 °C for a desired length of time. Digestion was stopped by adding 5 µl 1% trifluoroacetic acid (TFA) (Sigma-Aldrich). Samples were then concentrated to dryness using a speedvac (Concentrator Plus, Eppendorf) at 30 °C, in vacuum-aqueous (V-AQ) mode. Dried samples were either stored at -20 °C, or re-suspended in 10 µl of 0.1% TFA for LC-MS analysis.

LC-MS analysis

Prepared samples were loaded to a U3000 HPLC (Thermo Fisher Scientific) online with the amaZon ETD Ion Trap or the maXis UHR-TOF mass spectrometer (both from Bruker Daltonics) with 5 µl injection amount. Mobile phase A was water with 3% Acetonitrile (ACN) plus 0.1% formic acid (FA), and mobile phage B was water with 97% ACN plus 0.1% FA. Components of mobile phase were all HPLC grade and were purchased from Fisher Scientific. All buffer percentages are v/v. For the amaZon system, samples were loaded on a LC Packings µ-precolumn holder with 2 × connecting tubing 30 µm I.D. × 15 cm at a flow rate of 30 µl/min of 97% mobile phase A for 5 minutes, and eluted from a C18 column (Acclaim PepMap 100 C18, 3 µm, 100Å, 75 µm × 15 cm, Thermo Fisher Scientific) with gradient of 35 minutes from 3% to 40% mobile phase B at a flow rate of 30 µl/min. The temperature was set at 35 °C. Electrospray ionisation was used to introduce peptides into the mass spectrometer, with ion polarity set to positive. Capillary voltage was set to 4200 V, with end plate offset voltage 500 V. Dry gas flow rate was 4.0 L/min, dry temperature was 250 °C. MS data were acquired in enhanced resolution mode, with a scan range of 300-1350 m/z. For the maXis system, samples were loaded on a Dionex Nano-Trap Column with Dionex nanoViper Fingertight Fittings, 30µm I.D. × 100mm (Thermo Fisher Scientific) at a flow rate of 30 µl/min, 97% mobile phase A for

2 minutes, and eluted from a C18 column (Acclaim PepMap RSLC C18, 2 µm, 100Å, 75 µm × 15 cm, nanoViper, Thermo Fisher Scientific) with a gradient of 20 minutes, from 3% to 40% mobile phase B at a flow rate of 0.3 µl/min. The column temperature was set at 35 °C. Electrospray ionisation (captive spray) was used to introduce peptides into the mass spectrometer, with ion polarity set to positive. Capillary voltage was set to 1400 V, with end plate offset voltage 500 V. Dry gas flow rate was 3.0 L/min, dry temperature was 150 °C. Mass scan range was set as 50-2200 m/z.

Data processing

Data were processed using Compass DataAnalysis software (Version 4.1, build 359, Bruker Daltonik GmbH). amaZon or maXis generated PA/rPA data (data file type .d) were processed using the default ion trap processing method to generate Mascot generic format (MGF) files, which were searched against SwissProt database with bacteria taxonomy selected. Peptide abundance was calculated based on area under curve (AUC) of the extracted ion chromatogram (EIC) with the correct m/z of each SIS peptide. The sum of AUC based on the EIC representing the correct m/z of peptide was used for peptide quantification. As a general rule, for data generated from the amaZon system, a precision range (width) of ±0.2 was used, while for data generated from the maXis system, the precision range (width) was set as ±0.01. The peaks were smoothed for 2 cycles using a Gaussian algorithm. Based on AUC, the concentration of PA/rPA in the anthrax vaccine was calculated. Numerical data were processed using Microsoft Excel 2013 software.

Chapter 3: Studying the Effects of DNA Modifications on the Interference of CRISPR-Cas Systems

3.1 Abstract

CRISPR-Cas systems and restriction modifications (RM) are two host defence systems used by prokaryotes to provide protection against mobile genetic elements (MGEs). RM and CRISPR-Cas systems work by targeting specific sequences of invading DNA. One known strategy that phage use to counter-attack these defence systems of their host is to modify their own DNA. In this study, I developed and optimised methods using PCR in conjunction with modified dNTPs, to synthesise a range of target dsDNAs with various DNA modifications, including hydroxymethylcytosine and glucosylated hydroxymethylcytosine. Synthesised modified dsDNA were further tested using *in vitro* assays by collaborators at Wageningen University. Results showed that glucosylated target DNA can interfere with the activities of type I-E and type II-A CRISPR-Cas systems by lowering targeting DNA binding affinity. In contrast, for type V-A systems where Cas12a is employed, glucosylation of DNA does not interfere with binding and cleavage. For target DNA with 5-hydroxymethylated cytosine, none of the 3 types of CRISPR-Cas systems are impaired. Specifically for type I-E CRISPR-Cas systems, results also show that Cascade/Cas3 is not impeded by deoxyuridine or 5-hydroxymethyluridine containing DNA, but impeded by 5-methylcytosine and phosphorothioate linkage containing DNA.

3.2 Introduction

Bacteriophage, or phage, are the most abundant biological entities on Earth, estimated over 10^{30} (Chibani-Chennoufi et al., 2004). These viruses usually largely outnumber their host by up to 150 fold (Wigington et al., 2016), and infect bacteria 10^{25} times per second (Lima-Mendez et al., 2007). To survive under this type of pressure, bacteria have developed several defence strategies on different levels, such as entrance block, which prevents phage adsorption or phage DNA injection (Samson et al., 2013; Stern and Sorek, 2011), and abortive infection (Abi), which triggers a suicide mechanism of infected bacteria (Forde and Fitzgerald, 1999). A further strategy used is restriction modification systems (RM). RM systems are used by bacteria to protect their own DNA by using chemical modifications (e.g., cytosine methylation). Invading DNA without the same modification will be cleaved and destroyed by endonucleases (Luria, 1953; Luria and Human, 1952). Another DNA level defence system is clustered regularly interspaced short palindromic repeat with CRISPR-associated proteins, or CRISPR-Cas systems. These are adaptive immune systems, which include the integration of nucleic acid sequences (known as spacers) from invaders into the host chromosome. These sequences are subsequently transcribed and processed into small CRISPR RNAs (crRNAs), and assembled into functional RNA-nuclease complexes (Terns and Terns, 2014). CRISPR-Cas systems have been characterized into different types. In *E. coli* K12 (type I-E system), a ribonucleoprotein complex named Cascade (CRISPR-associated complex for antiviral defence) is described (Brouns et al., 1993; Jore et al., 2011). This type of CRISPR-Cas systems uses the nuclease Cas3 to bind target DNA and mediate cleavage. For type II CRISPR-Cas systems, CRISPR/Cas9 systems are the most studied. Cas9 not only serves as an endonuclease but also facilitates crRNA mediated DNA binding (Makarova et al., 2015). CRISPR/Cas9 systems have been wildly used in gene editing. Similar to type II systems, a more recently characterized CRISPR-Cas system, type V-A also uses single effector protein, Cas12a (Zetsche et al., 2015). This feature of type V-A systems makes them potential alternatives to CRISPR/Cas9 as gene-editing tools (see Chapter 1.3.2 for more information regarding gene-editing using CRISPR).

Phage, on the other hand, also developed a number of strategies to cope with the prokaryotic defence systems as part of the long biological arms race between the two. For example, for bacterial defence strategies that target specific DNA sequences (including CRISPR-Cas systems), phage can either mutate target sequences to escape restriction endonuclease activities (McGrath et al., 1999; Deveau et al., 2008), or nullify the surveillance of defending systems by recombining their own genomes (Andersson and Banfield, 2008; Paez-Espino et al., 2015). Phage can also express inhibitory proteins that target cofactors and restriction sites of RM systems (Samson et al., 2013). For CRISPR-Cas systems, similar types of inhibitory proteins coded by phage, named anti-CRISPR proteins (ACR), have been discovered recently (Pawluk et al., 2016). These ACRs can bind to Cas proteins and interfere with their functions.

Another strategy phage have developed to survive the host defense systems is to modify their own DNA, which has been described in T-even phage (Weigele and Raleigh, 2016). Depending on species, phage have different types of DNA modifications (Figure 3.1A-D). For T4 phage that infect *E. coli*, cytosine is replaced by 5-hydroxymethylcytosine (hmC). For some phage that infect *Bacillus subtilis*, thymine is replaced by 5-hydroxymethyluracil (hmU) or uracil. The percentage of modification also varies. In T4 phage, 100% of cytosine bases are replaced by hmC, and 41% of thymine bases are replaced by 5'-dihydroxypentyluracil (Warren, 1980). Base modifications can alter the structure of DNA, leading to changes of DNA physical properties such as thermal transition temperature (Warren, 1980). The biogenesis of modified bases involves a series of enzymes such as dCMP hydroxymethylase and dCMP deaminase (Mathews et al., 1979). For T4 phage to bypass RM systems of *E. coli*, the modification of hmC in their DNA is not enough, as *E. coli* can specifically target DNA with hmC with McrBC systems (Raleigh and Wilson, 1986). Thus for phage to survive, they further glucosylate their DNA with hmC converted into glucosyl-5-hydroxymethylcytosine (ghmC) (Figure 3.1E) through phage-coded glucosyl transferases, in conjunction with uridine diphosphate glucose (UDPG) as a source of glucose (Lehman and Pratt, 1960). GhmC modification is believed an effective strategy against most RM systems.

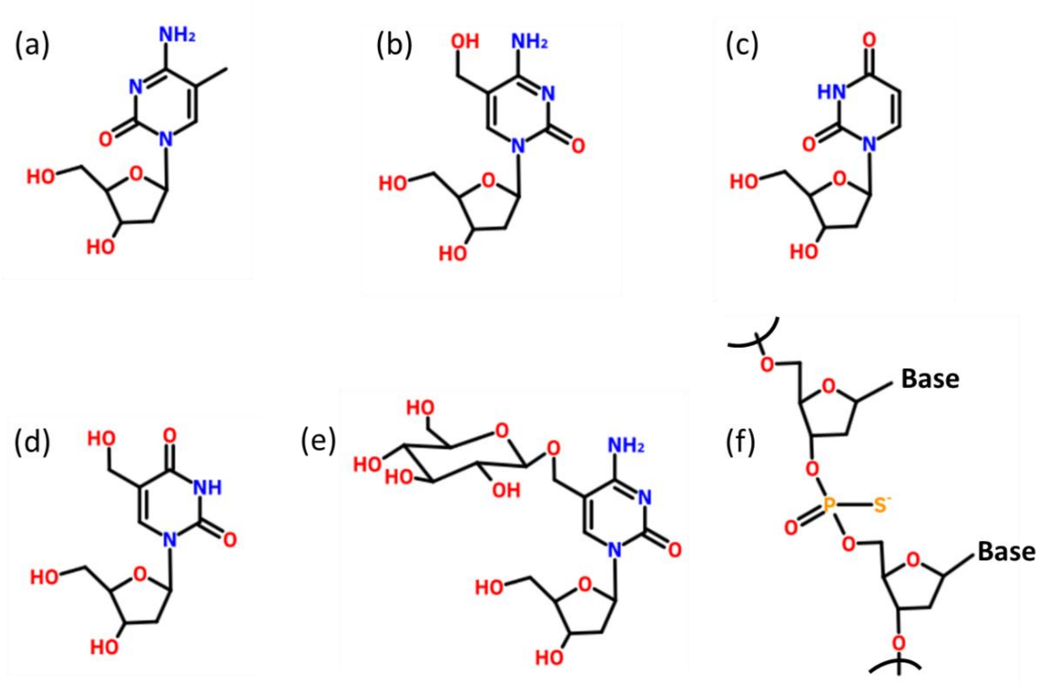


Figure 3.1 | Examples of phage DNA modifications. (a) 5-methylcytosine (mC). (b) 5-hydroxymethylcytosine (hmC). (c) 2'-deoxyuridine (dU). (d) 5-hydroxymethyl-2'-deoxyuridine (hmU). (e) Glucosyl-5-hydroxymethylcytosine (ghmC). (f) Phosphorothioate linkages.

As the defence mechanisms of CRISPR-Cas systems are different compared to RM systems, the way they act on modified target DNA is also likely to be different. In order to find out how CRISPR-Cas systems interact with modified target DNA, in this chapter, a series of modified DNA, including 5-methylcytosine (mC); 5-hydroxymethylcytosine (hmC); 2'-deoxyuridine (dU); (d) 5-hydroxymethyl-2'-deoxyuridine (hmU); (e) Glucosyl-5-hydroxymethylcytosine (ghmC); and (f) Phosphorothioate linkages (Thio) (see Figure 3.1) were synthesised, followed by *in vitro* CRISPR-Cas interaction studies carried out by collaborators at Wageningen University.

3.3 Design and synthesis of modified dsDNA

3.3.1 PCR template preparation for substrate dsDNA synthesis

To study the effect of modifications on CRISPR-Cas defence systems, target dsDNA (98 bp) with spacer 8 sequence (sp8) were required for subsequent *in vitro* Cascade analysis. Initial work was therefore focussed on the optimisation of the synthesis of modified DNA templates using PCR. Templates used for first round of PCR were 937 bp plasmids containing the sp8 sequence. PCR products (98 bp dsDNA) were synthesised using Q5 high-fidelity polymerase, with BG8415 and BG8416 as forward and reverse primers, respectively. Results are shown in Figure 3.2 (lane 1). Product dsDNA were then 1:100 and 1:1000 diluted with nuclease free water. Diluted dsDNA products were used as templates for a second round of PCR. Synthesised dsDNA are shown in Figure 3.2 (lanes 2 and 3). For best results, the 1:100 diluted first round 98 bp PCR products were chosen as templates for future dsDNA synthesis in this study.

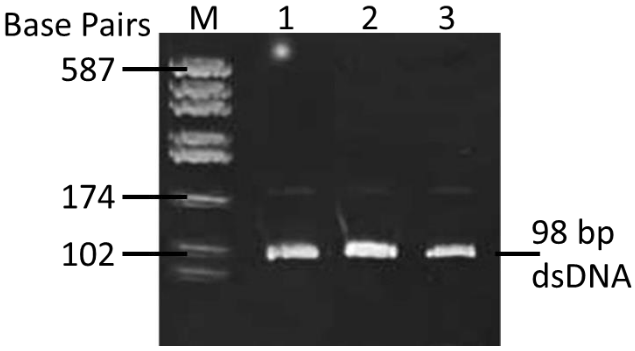
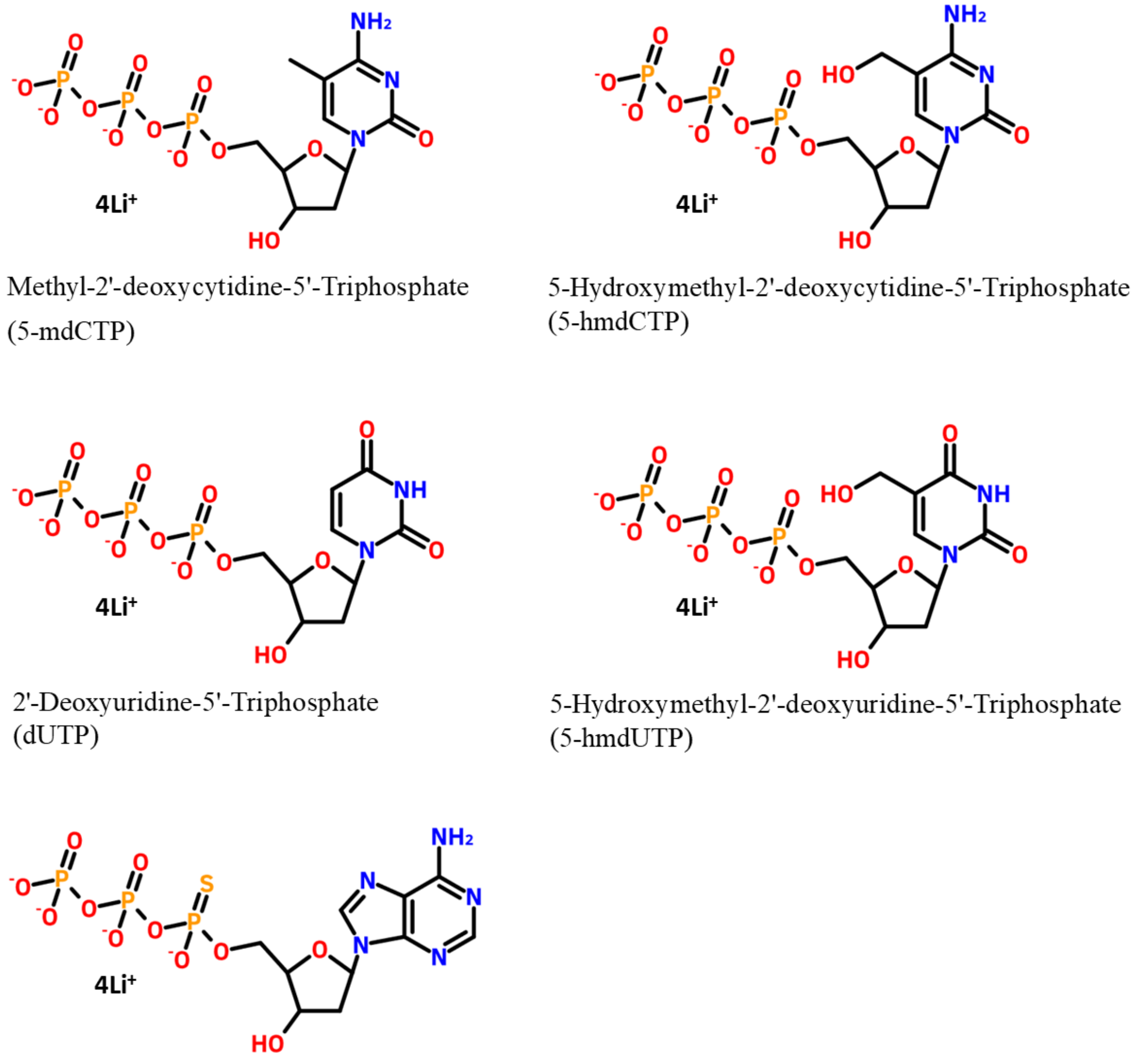


Figure 3.2 | PAGE result of PCR template preparation. A pUC 18 DNA Hae III digest (Sigma Aldrich) was used as marker (M). Lane 1: PCR synthesised dsDNA using 937 bp plasmid DNA as template. Lane 2: PCR synthesised dsDNA using 1:100 diluted product from lane 1, as template. Lane 3: PCR synthesised dsDNA using 1:1000 diluted product from lane 1, as template.

3.3.2 Optimisation of incorporation of modified dNTPs using PCR

The incorporation efficiency of modified dNTPs is often lower than unmodified dNTPs during PCR. Therefore, to successfully synthesise sufficient amounts of target DNA using PCR, a suitable DNA polymerase needs to be determined. In this study, 2 different types of DNA polymerases were tested: *Taq* DNA polymerase and Q5 high-fidelity DNA polymerase (with processivity-enhancing Sso7d DNA binding domain fused novel polymerase, ultra-low error rates).

For *Taq* polymerase, a range of different PCR reactions were performed, using 1) unmodified dNTPs (A, T, G and C), 2) dmCTP replacing dCTP, 3) dhmCTP replacing dCTP, 4) dUTP replacing dCTP and 5) α-thio-dNTPs. A summary of the modified dNTPs used in this study are shown in Figure 3.3.



Methyl-2'-deoxycytidine-5'-Triphosphate (5-mdCTP)

5-Hydroxymethyl-2'-deoxycytidine-5'-Triphosphate (5-hmdCTP)

2'-Deoxyuridine-5'-Triphosphate (dUTP)

5-Hydroxymethyl-2'-deoxyuridine-5'-Triphosphate (5-hmdUTP)

2'-Deoxyadenosine-5'-O-(1-Thiotriphosphate) (α-Thio-dATP)

Figure 3.3 | Molecular structure of modified dNTPs used in this study.

Results from the PCRs are shown in Figure 3.4. The results show that 5mdC, 5hmdC and dU were successfully incorporated into DNA (lanes 1-3), with slightly reduced efficiency compared to unmodified dNTP incorporation (lane 6). In general, the efficiency of *Taq* polymerase is relatively low even for incorporating unmodified dNTPs (comparing lane 6 to the marker bands, where approximately 61 ng was loaded). Practically a yield of 500 ng/µl dsDNA product is achievable after 35 cycles of amplification using 100 bp templates. PCR reactions using thio-dNTPs (all 4 dNTPs replaced by α-thio-dNTPs), no product was detected (lane 4).

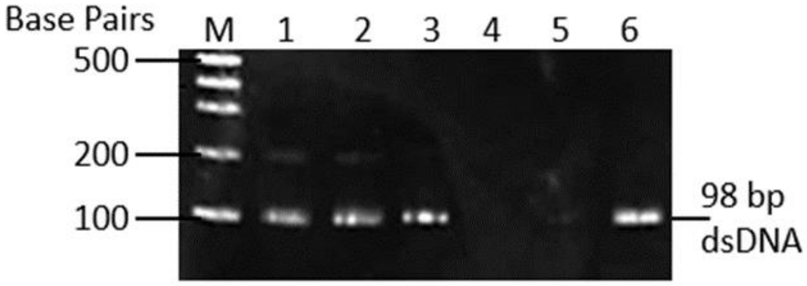


Figure 3.4 | PAGE results of modified dNTPs incorporation by PCR using *Taq* polymerase. The approximate loading amount of marker (M) is 61 ng. Lane 1-4 are PCR products with incorporated mdC, hmdC, dU, and thio-dNTPs, respectively. Lane 5 is no template control (NTC). Lane 6 is PCR product with normal dNTPs.

Since full replacement of dNTPs with α-thio-dNTPs leads to unsuccessful PCR incorporation, the next attempt was to partially replace dNTPs with α-thio-dNTPs. α-thio-dNTPs were mixed with normal dNTPs at a range of different ratios and the results are shown in Figure 3.5 (A). From lanes 1-6, with the proportion of α-thio-dNTPs increased, product yields decreased. When the amount of α-thio-dNTPs increases to above 90%, almost no product is detected (Figure 3.5A, lanes 4-6). It is worth noting that shifts of migration were observed in gel electrophoresis for DNA products using α-thio-dNTPs (Figure 3.5A, lanes 1-4). The higher percentage of thio-dNTP replacement, the shorter distance migrated on the gel. This alteration in migration is likely caused by the replacement of phosphodiesters with the phosphorothioates, consistent with α-thio-dNTPs incorporation. To further study the incorporation of α-thio-dNTPs, an experiment was performed using up to three types of α-thio-dNTPs but with full replacement of the corresponding unmodified dNTPs. Results are shown in Figure 3.5A (lanes 7-8) and Figure 3.5B (lanes 2-5). These results show that using *Taq* polymerase enabled incorporation of one or two out of the four dNTPs replaced by α-thio-dNTPs. However, even with the replacement of one type of the unmodified dNTP with the corresponding α-thio-dNTP, the incorporation efficiency was greatly reduced (compare lane 7 to lane 1 in Figure 3.5A). In experiments with any combination of three α-thio-dNTPs, PCR is unsuccessful as no product was detected (Figure 3.5B, lane 2-5.).

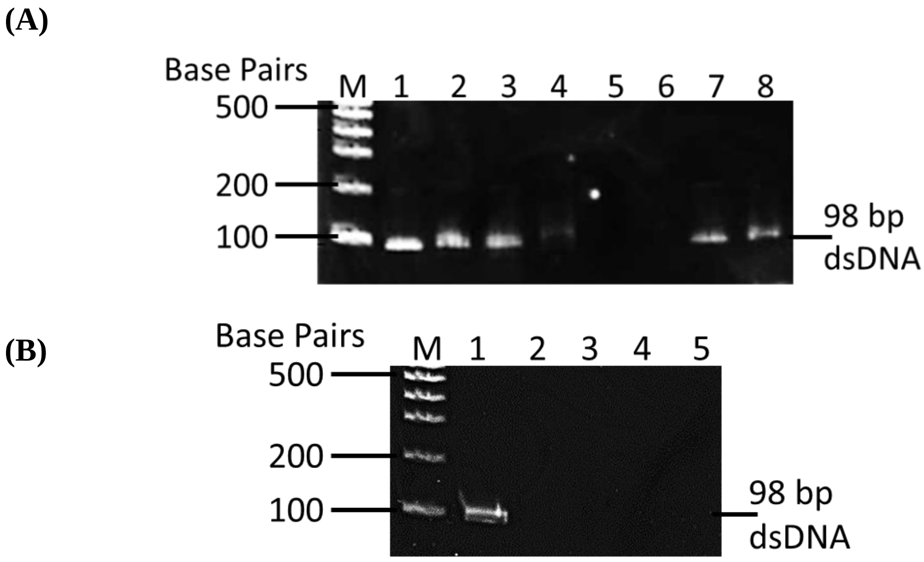


Figure 3.5 | Optimisation of α-thio-dNTP incorporation by PCR using *Taq* polymerase. (A) Using an increasing proportion of α-thio-dNTP mixtures (A, T, G and C) to replace normal dNTP mixtures (A, T, G and C) in PCR. From lane 1 to lane 6: 0%, 75%, 82.5%, 90% and 97.5% of α-thio-dNTP mixture, respectively. Lane 7: dCTP fully replaced by α-thio-dCTP in PCR (dA, dT and dG not replaced); lane 8: dCTP and dGTP fully replaced by α-thio-dCTP and α-tho-dGTP in PCR (dA and dT not replaced). **(B)** Lane 1: control with unmodified dNTPs; lanes 2-5: full replacement of three α-thio-dNTPs with one normal dNTP (normal dC, dG, dT, dA for lanes 2-5, respectively).

Following analysis of incorporation efficiency of modified dNTPs using *Taq* polyermase, further studies were performed using Q5 high-fidelity DNA polymerase for modified dNTP incorporations. Results are shown in Figure 3.6. Q5 high-fidelity DNA polymerase show good efficiency incorporating mdC and hmdC (compared to the incorporation efficiency with normal dNTPs). For dU incorporation using Q5 high-fidelity DNA polymerase, no product was observed. This is due to the high-fidelity feature of this enzyme, which recognizes the dU incorporation as an error, a proof reading ability provided by its 3'-5' exonuclease activity. However, hydroxymethyluridine (hmdU) can be incorporated by this polymerase with reduced efficiency. For α-thio-dNTP incorporations, full replacement of two out of four α-thio-dNTPs (dC and dG in this case) was proven successful, but not possible with all four dNTPs, similar to *Taq* polymerase.

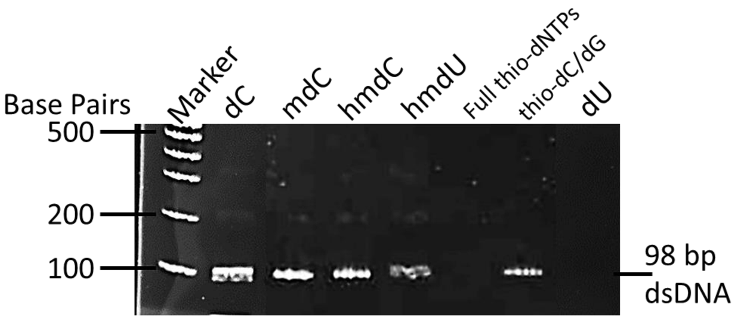


Figure 3.6 | PAGE results of modified dNTPs incorporation by PCR using Q5 high-fidelity DNA polymerase.

3.3.3 Optimisation of PCR product yield using modified dNTPs

Although both *Taq* and Q5 high-fidelity DNA polymerase were shown to be able to incorporate modified dNTPs, the yields (including unmodified dNTPs incorporation) were relatively low. For this reason, further optimizations were carried out. Different reaction conditions including the denaturing and annealing temperature, the reaction time and the concentration of reagents used (including Mg^{2+}, primers, DNA polymerases and dNTPs) were tested. Mg^{2+} as a cofactor is already supplied in the reaction buffer that comes with the kit. For *Taq* polymerase, the concentration of Mg^{2+} in 1× standard reaction buffer is 1.5 mM. For *Taq* polymerase incorporating unmodified dNTPs, the results show that increased amounts of additional Mg^{2+} up to 2.5 mM have positive effect on product yields (Figure 3.7A). Other conditions tested including the increased amount of dNTPs (from 0.1 mM to 0.5 mM), the increased amount of primers (0.1 µM to 1 µM), or prolonged denaturing time (from 30 seconds to 45 seconds), do not have significant impact on product yields (see Figure 3.7 B, C and D).

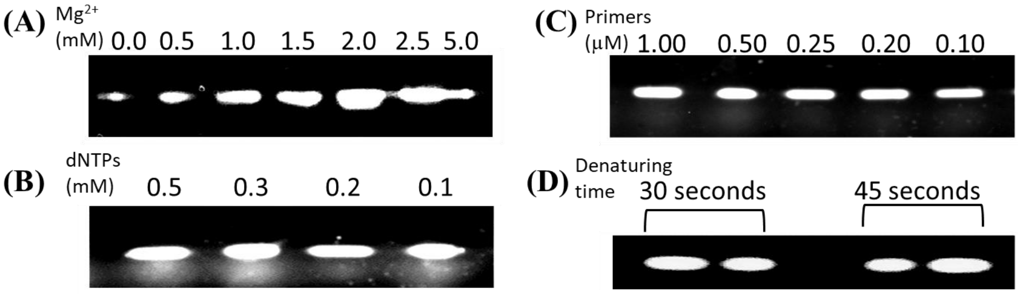


Figure 3.7 | Optimisation of unmodified dNTPs incorporation using *Taq* polymerase. Different conditions were tested for product yields optimisation purposes. **(A)** Additional amount of Mg^{2+}, **(B)** dNTPs concentration, **(C)** primer amount (forward and reverse, each), and **(D)** denaturation time length for each cycle.

Annealing temperature is a critical factor for PCR that may affect product yield. Here, different annealing temperatures were tested for *Taq* polymerase incorporating modified and unmodified dNTPs. Of the two annealing temperatures tested, 46 °C is the recommend annealing temperature calculated based on primer sequences. Results show no significant differences between the two annealing temperature conditions (Figure 3.8).

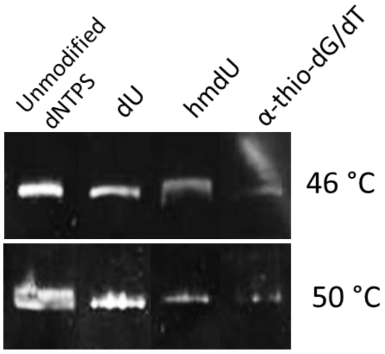


Figure 3.8 | Effect of annealing temperature on product yield using *Taq* polymerase. Incorporation of both modified and unmodified dNTPs were tested.

Hot start strategy in PCR can reduce nonspecific formation of primer dimers and increase product yields (D'aquila et al., 1991). Here, a PCR reaction was performed where *Taq* polymerase and primers were added to a preheated (to reaction temperature) mixture. As shown in Figure 3.9, using the hot start strategy with *Taq* polymerase had a small effect on product yield. For modified dNTP incorporations with lower efficiency (i.e., dU, hmU and α-thio dNTPs), hot start strategy was proved of limited beneficial effect for this study (lane 3 to lane 5).

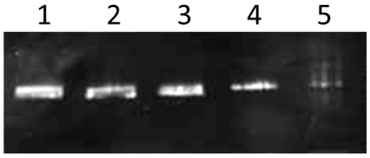


Figure 3.9 | Efficiency test of hot start PCR strategy with *Taq* polymerase. Lane 1 used hot start PCR procedure, lane 2 used normal PCR procedure, both for unmodified dNTP incorporations. Lanes 3-5 used hot start PCR for the incorporation of dU, hmU and α-thio-dG/dT, respectively.

Similar optimization was also applied to Q5 high-fidelity DNA polymerase. First, the effect of Mg^{2+} was tested, with the results shown in Figure 3.10. Unlike *Taq* polymerase, for both modified and unmodified dNTPs incorporation tested, additional Mg^{2+} (3 mM) reduced yields. These results suggest that the reaction buffer for Q5 high-fidelity DNA polymerase is already optimized for the amount of Mg^{2+} required.

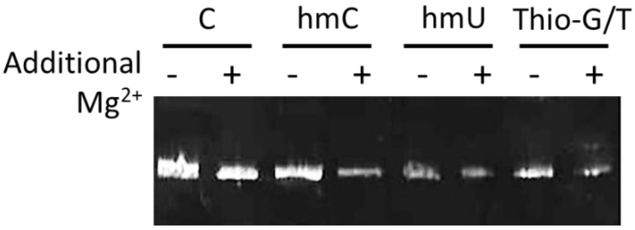


Figure 3.10 | Effect of additional Mg^{2+} on Q5 high-fidelity DNA polymerase. Tests were performed incorporating unmodified dNTPs, hmdC, hmdU and α-thio-dG/dT, with 3 mM Mg^{2+} (shown as "+") or without additional Mg^{2+} (shown as "-").

Next, the effect of denaturing temperature, annealing temperature and numbers of thermo cycle on Q5 high-fidelity DNA polymerase performance were tested for the incorporation of unmodified dNTPs, hmdU and α-thio-dG/dT, results are shown in Figure 3.11. With additional 10 thermocycles of synthesis, neither the increase in denaturing temperature (98 °C to 99.9 °C) nor the increase of annealing temperature (61 °C to 63 °C) showed beneficial effect on product yields. On the contrary, the incorporation efficiency of α-thio-dG/dT was decreased under these conditions. However, it is worth noting that higher annealing temperatures (61 °C or 63 °C) showed improved yields compared to 50 °C for Q5 high-fidelity DNA polymerase (see Figure 3.5, where 50 °C annealing temperature was used).

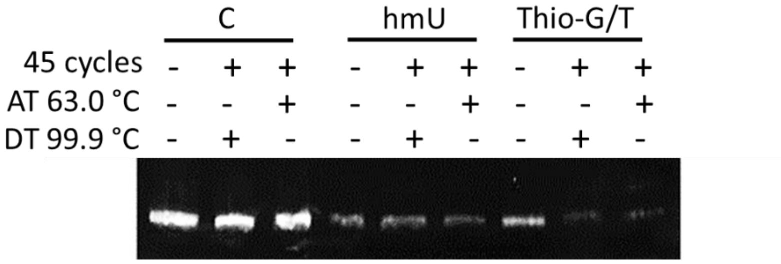

C
hmU
Thio-G/T
45 cycles - + + - + + - + +
AT 63.0 °C - - + - - + - - +
DT 99.9 °C - + - - + - - + -

Figure 3.11 | Effect of denaturing temperature (DT), annealing temperature (AT) and thermo cycles on the incorporation of hmC, hmU and α-Thio-dG/dT using Q5 high-fidelity DNA polymerase. As controls, samples were run on default conditions (DT 98 °C, AT 61 °C, 35 cycles) marked as "-".

3.3.4 Preparation of modified dsDNA substrates for interference of CRISPR-Cas Systems

Based on the PCR yield optimization results, Q5 high-fidelity DNA polymerase was chosen to synthesise dsDNA targets, except for dU, which was performed using *Taq* polymerase. The dsDNA substrates with sp8 spacer sequence (for sp8 sequence see section 2.1) were subsequently used for downstream studies of CRISPR-Cas systems (see section 3.4). The optimised final conditions used for PCR with both polymerases, are described in Chapter 2.1. For *Taq* polymerase, additional Mg^{2+} (2.5 mM) was used. Due to the fact that optimization attempts did not significantly increase product yields, reactions were carried out on a larger scale, and products were purified and concentrated to obtain the titre required for the downstream application. This is especially necessary for low efficiency incorporations such as hmU and α-thio-dNTPs. Concentrated and purified dsDNA with different modifications were then analysed on native PAGE, the final products are shown in Figure 3.12. The yields of the DNA products were determined using UV spectrophotometry (NanoDrop 2000C) and are summarised in Table 3.1, which are sufficient for further *in vitro* CRISPR-Cas studies. A number of the non-specific products including primer dimers were also generated in the purified products (Figure 3.12). It is interesting to note that the profiles of non-specific products are different for different dNTP incorporations. For example, the double bands for dU incorporation at approximately 180 bp, significant amount of small (possibly similar size as primers) non-specific products for hmC and hmU incorporations. The reasons for this are not yet understood.

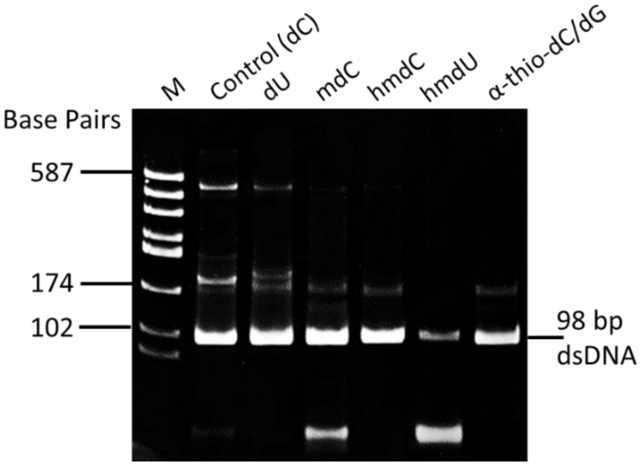


Figure 3.12 | PAGE analysis of oligo sp8 dsDNA products with various modifications. All samples were concentrated and purified before loading to the gel.

Table 3.1 | Concentration of final PCR products given by NanoDrop. Modified bases are in bold. Phosphorothioate linkages are shown as "ps".

DNA Products	Conc. (ng/µl)*
d(AGCT) (control)	61.0
d(AGC**U)**	94.0
d(AG**mCT)**	102.7
d(AG**hmC**T)	60.1
d(AGC**hmU)**	53.9
d(A**GpsCps**T)	48.2

* Note: Concentration of DNA products are based on QIAGEN kit purified samples. Actual concentration of dsDNA products are lower, as values shown include non-specific products.

3.3.5 Synthesis of glucosylated 5-hydroxymethylcytosine DNA

Purified DNA substrates with hmC incorporated (see Figure 3.12) were then enzymatically glucosylated using T4 beta-glucosyltransferase. Glucosylated dsDNA substrates were then concentrated to approximately 50 µl and analysed on PAGE (Figure 3.13). Compared to dsDNA with hmC, a small shift in migration of the glucosylated DNA on PAGE is observed, indicating the glucosylation process was successful.

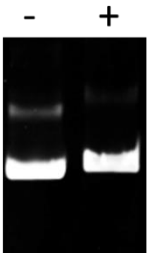


Figure 3.13 | Glucosylated hmC DNA (+) and hmC DNA (-).

3.3.6 Preparation of phosphate labelled dsDNA

Downstream CRISPR-Cas studies required the use of 5'-end and 3'-end ^{32}P labelled dsDNA. To enable selective ^{32}P phosphorylation of the dsDNA strands, in this work samples were prepared using phosphate-labelled primers. Either 5'-phosphate-labelled forward primers or reverse primers were used to label one of the two strands of dsDNA. PCR reactions were performed as described in the previous section and the results are shown in Figure 3.14. Incorporation efficiency is not affected in a significant way when phosphate-labelled primers are used (Figure 3.14A). The final dsDNA products (after concentration and purification) are shown in Figure 3.14 (B) and (C). The product concentrations were determined by NanoDrop and are listed in Table 3.2 and 3.3.

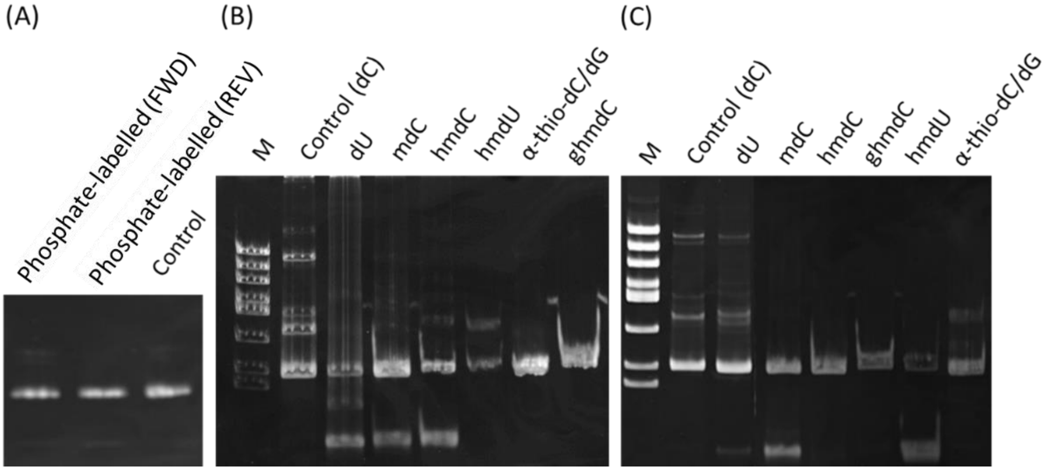


Figure 3.14 | PAGE analysis of phosphate-labelled dsDNA. (A) Incorporation efficiency of using phosphate-labelled primers compared to regular primers (control). FWD: forward primers, REV: reverse primers. **(B)** Oligo sp8 dsDNA products with various modifications using phosphate-labelled forward primers. **(C)** Oligo sp8 dsDNA products with various modifications using phosphate-labelled reverse primers. Marker used in **(B)** and **(C)** are pUC DNA Hae III digests.

Table 3.2 | Concentration of PCR products (with 5'-phosphate-labelled forward primers) by NanoDrop test. Modified bases are in bold. Phosphorothioate linkages are shown as "ps".

Products	Conc. (ng/ µl)*
d(AGCT) (control)	66.7
d(AGC**U**)	69.2
d(AG**mC**T)	120.9
d(AG**hmC**T)	83.6
d(AG**ghmC**T)	69.1
d(AGC**hmU**)	57.6
d(A**GpsCps**T)	37.9

* Note: Concentration of DNA products are based on QIAGEN kit purified samples. Actual concentration of dsDNA products are lower, as values shown include non-specific products

Table 3.3 | Concentration of PCR products (with 5'-phosphate-labelled reverse primers) by NanoDrop test. Modified bases are in bold. Phosphorothioate linkages are shown as “ps”.

Products	Conc. (ng/ µl)*
d(AGCT) (control)	33.3
d(AGC**U**)	59.2
d(AG**mC**T)	105.0
d(AG**hmC**T)	68.8
d(AG**ghmC**T)	80.3
d(AGC**hmU**)	77.3
d(A**GpsCps**T)	62.4

* Note: Concentration of DNA products are based on QIAGEN kit purified samples. Actual concentration of dsDNA products are lower, as values shown include non-specific products

To further verify the successful incorporations of various modifications into the dsDNA, samples were digested into nucleosides and analysed by HPLC-UV and LC-MS, which are discussed in chapter 4.

3.4 Studying the Effects of DNA Modifications on the Interference of CRISPR-Cas Systems

3.4.1 *In vitro* Cascade interaction

Following the synthesis and validation (for validation see Chapter 4) of the 98 bp dsDNA (containing the sp8 spacer sequence) with various DNA modifications, *in vitro* binding/cleavage reactions of the type I-E CRISPR system from *E. coli* was performed. The *in vitro* binding and cleavage analysis was performed by Marnix Vlot in collaboration with Wageningen University. Cascade complex was overexpressed in *E. coli* in conjunction with a crRNA complimentary to the sequence of target dsDNA substrate. The enzyme responsible for cleaving the target dsDNA (Cas3) was also expressed. Complexes containing a crRNA with non-matching spacer eGFP was used as negative control. *In vitro* cleavage assays were carried out under different conditions (with or without Cascade/Cas3) on DNA targets containing the modified nucleosides. For target dsDNA with unmodified cytosine, hmC and ghmC, the *in vitro* Cascade interaction results are shown in Figure 3.15. The results show that the Cascade/Cas3 RNP complex successfully degrades target dsDNA with unmodified cytosine and hmC, but not DNA containing ghmC (Figure 3.15B). To determine whether the above results were caused by the inhibition Cascade binding or inhibition of Cas3 cleavage, electrophoretic mobility shift assay (EMSA) was performed (Figure 3.15C). The results show that for dsDNA with hmC, sequence-specific binding is reduced compared to dsDNA with unmodified cytosine. For ghmC containing dsDNA, no binding is observed, indicating that the bulky structure of glucosylated DNA may prevent Cascade from binding, and could be one of the reasons ghmC containing DNA escape the CRISPR-Cas systems.

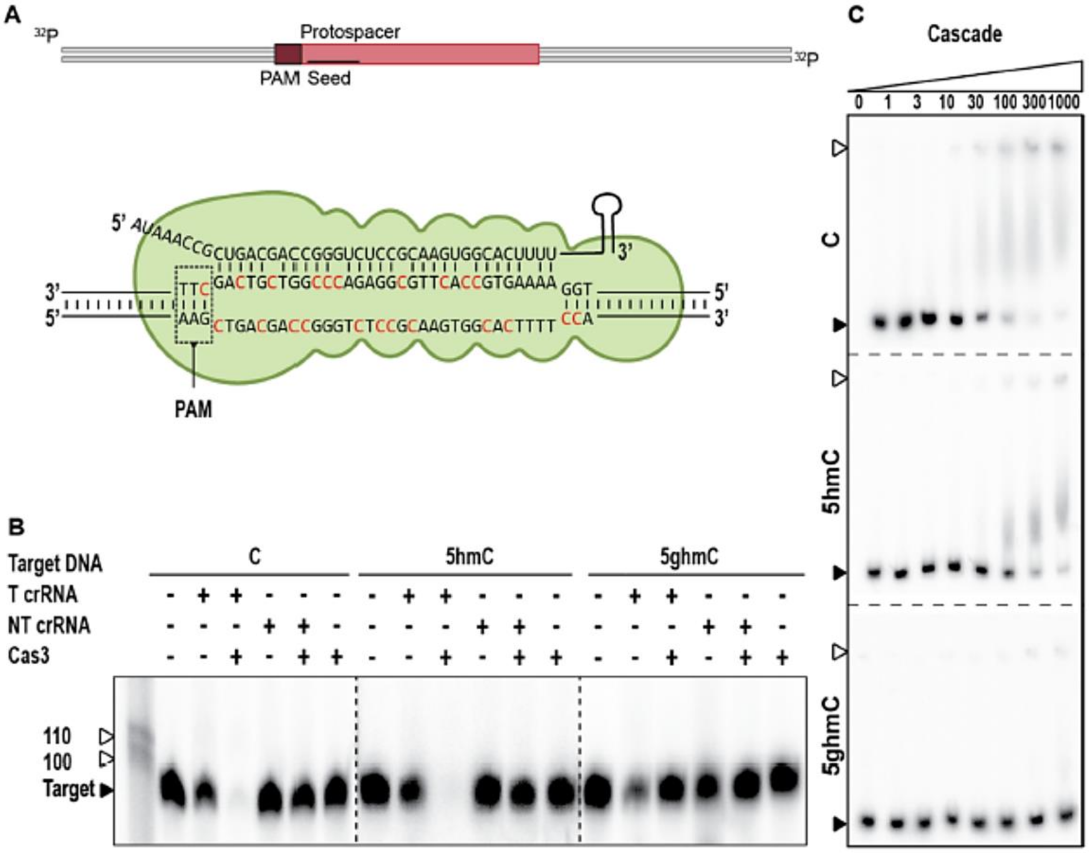

A
32P
Protospacer
32P
PAM Seed
5' AUAAACCG CUGACGACCGGGUCUCCGCAAGUGGCACUUUU 3'
3' TTC GACTGCTGGCCCAGAGGCGTTCACCGTGAAAA GGT 5'
5' AAG CTGACGACCGGGTCTCCGCAAGTGGCACTTTT CCA 3'
PAM
B
Target DNA
T crRNA
NT crRNA
Cas3
C
5hmC
5ghmC
110
100
Target
C
Cascade
0 1 3 10 30 100 300 1000
C
5hmC
5ghmC

Figure 3.15 | *In vitro* Cascade interaction with modified oligo T4 phage dsDNA (hmC and ghmC). (A) Schematic diagram of DNA targeting by Cascade and the formation of R-loop. Cytosine residue modifications are indicated in red. **(B)** Cleavage assay of Cas3 in conjunction with Cascade on 98 bp dsDNA targets with different modifications (black arrow). Marker (bp) is shown with white arrows. T crRNA indicates targeting crRNA, while NT crRNA indicates non-targeting crRNA, either of which are loaded with Cascade effector complex. Restrictions product length (bp) are undefined. **(C)** Electrophoretic Mobility Shift Assay (EMSA) of Cascade performed on unmodified dsDNA, dsDNA with hmC, or ghmC (black arrows) at increasing Cascade concentrations (nM). White arrows indicate bound target fractions. Different gels are separated by dotted lines. Figure taken from Vlot et al., 2017, with permission.

Following identification that hmC does not interfere with Cas3 cleavage in contrast to ghmC, a number of other DNA modifications were evaluated in DNA cleavage assays. *In vitro* Cascade/Cas3 cleavage assays were performed by Marnix Vlot in collaboration with Wageningen University on target DNA containing mC, dU, hmU and DNA containing phosphorothioate linkages (see Figure 3.16). The results show that target dsDNA with dU or hmU are cleaved, while target dsDNA with mC or phosphorothioate containing DNA (C/G) are not cleaved. Since EMSA was not performed for DNA with these modifications, it is not clear whether the surviving mechanism of mC or phosphorothioate containing target dsDNA are due to Cas3 inhibition or Cascade binding inhibition. Either way, it is unexpected the modification of mC granted resistance to

Cascade/Cas3 systems *in vitro*, where hmC did not. Further experiments are required to confirm this. In the case of phosphorothioate containing target dsDNA it is proposed that Cas3 inhibition is the likely mechanism here as phosphorothioate bonds are typically more resistant to nucleases compared to phosphodiester bonds.

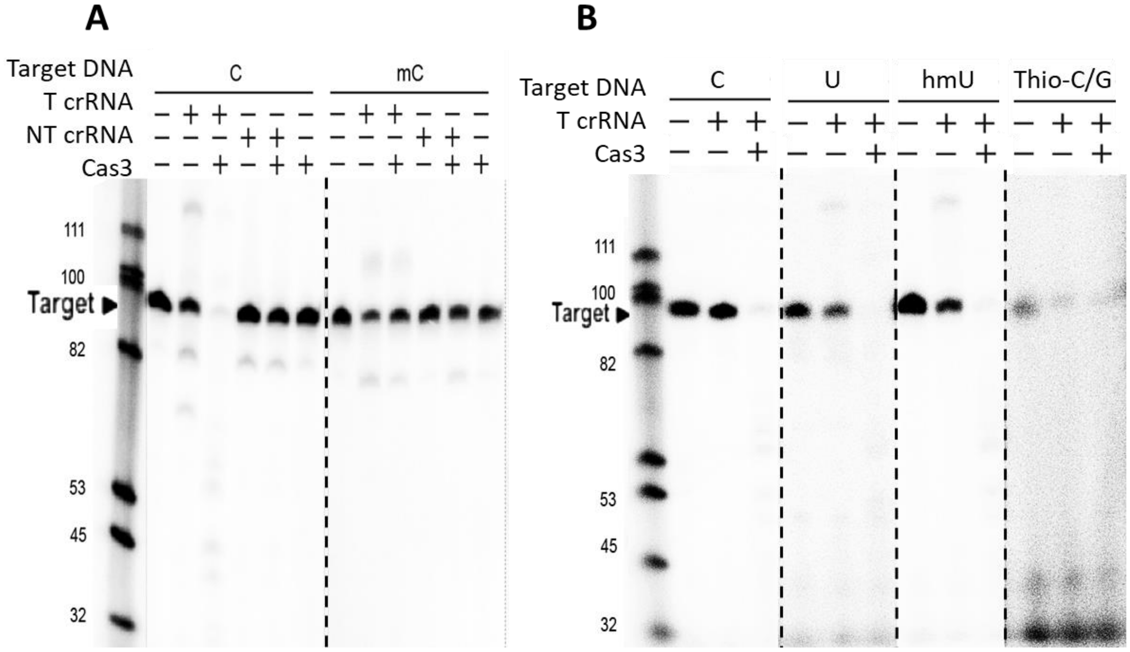


Figure 3.16 | *In vitro* Cascade interaction with modified dsDNA (mC, U, hmU and Thio-C/G = phosphorothioate containing). (A) Cleavage assay of Cas3 in conjunction with Cascade on 98 bp dsDNA target with mC (black arrow). Marker (bp) is shown as numbers. T crRNA indicates targeting crRNA, while NT crRNA indicates non-targeting crRNA, either of which are loaded with Cascade effector complex. **(B)** Cleavage assay of Cas3 in conjunction with Cascade on 98 bp dsDNA targets with U, hmU and Thio-C/G (black arrow). Marker (bp) is shown as numbers. T crRNA indicates targeting crRNA, which is loaded with Cascade effector complex. Different gels are separated by dotted lines.

3.4.2 *In vitro* Cas9 interaction

Further analysis of the effects of DNA modifications on a different CRISPR-Cas system was studied. *In vitro* binding/cleavage analysis of the type II-A CRISPR was performed by Marnix Vlot in collaboration with Wageningen University. Target dsDNA with different modifications (hmC and ghmC) were incubated with reconstituted complex containing Cas9 and single guide RNA (sgRNA). The results are shown in Figure 3.17. Similar to Casade/Cas3 systems, Cas9 with sgRNA can degrade target dsDNA with unmodified cytosine and hmC, but not ghmC (Figure 3.17B). The binding status for target DNA with different modifications were tested using EMSAs, where catalytically inactive Cas9 (dCas9) were used (Figure 3.17C). Interestingly, for hmC, sequencing specific binding is increased compared to unmodified target DNA, in contrast to previous binding studies of hmC and unmodified cytosine with Cascade/Cas3 systems. For ghmC, the binding of target DNA is observed (unlike Cascade/Cas3, where the binding is almost abolished), but with reduced affinity compared to unmodified target DNA. These results indicate that glucosylation can protect target DNA from Cas9-sgRNA systems by reducing the binding affinity. Also, glucosylation may provide scissile bond protections and lowering the interference effect even further.

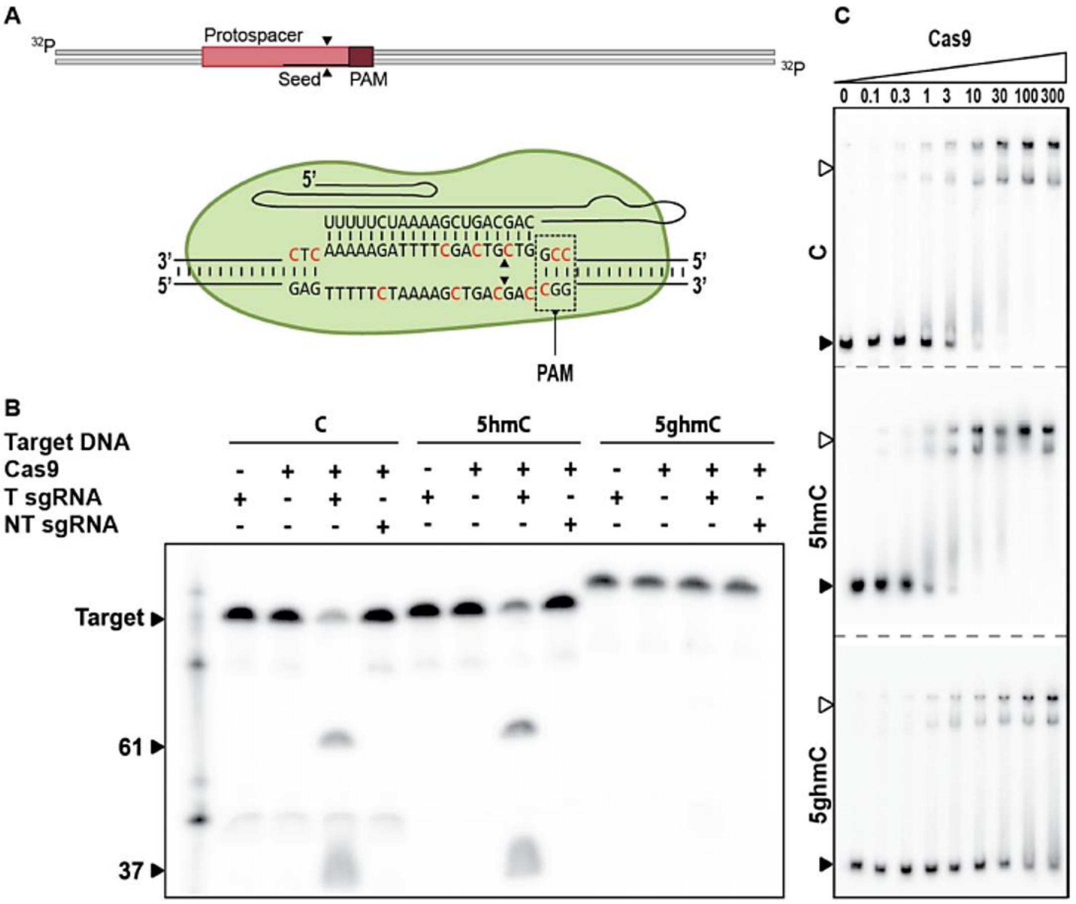


Figure 3.17 | *In vitro* Cas9-sgRNA interaction with modified dsDNA (hmC and ghmC). (A) Schematic diagram of DNA targeting by Cas9. Cytosine residue modifications are indicated in red, and black arrows indicated cleavage sites. **(B)** Cleavage assay of Cas9 on 98 bp dsDNA targets with different modifications (black arrow). Marker (bp) is shown with white arrows. T sgRNA indicates targeting sgRNA, while NT sgRNA indicates non-targeting sgRNA, either of which are loaded with Cas9. Black arrows at 61 bp and 37 bp indicate restriction products **(C)** EMSA of Cas9 performed on unmodified dsDNA, dsDNA with hmC, or ghmC (black arrows) at increasing Cas9 concentrations (nM). White arrows indicate bound target fractions. Different gels are separated by dotted lines. Figure taken from Vlot et al., 2017, with permission.

3.4.3 *In vitro* Cas12a interaction

To study the effects of the DNA modifications, another CRISPR-Cas system, type V-A Cas12a, which was derived from *Francisella novicida*, was studied. *In vitro* binding/cleavage analysis was performed by Marnix Vlot in collaboration with Wageningen University. Target dsDNA with different modifications were incubated with Cas12a and crRNA. The results are shown in Figure 3.18. Unlike type I-E and type II-A CRISPR-Cas systems discussed previously, in type V-A Cas12a systems, together with unmodified cytosine and hmC, ghmC also cannot protect target DNA from cleavage (Figure 3.18B). EMSAs conducted with catalytically inactive Cas12a (or dCas12a, mutation E1006A and R1218A) (Swarts et al., 2017) further confirmed that, binding affinity of ghmC containing target DNA were similar to unmodified and hmC DNA (Figure 3.18C). These results indicate that Cas12a-crRNA complex activities against target DNA are not impeded by DNA modifications including hmC or ghmC.

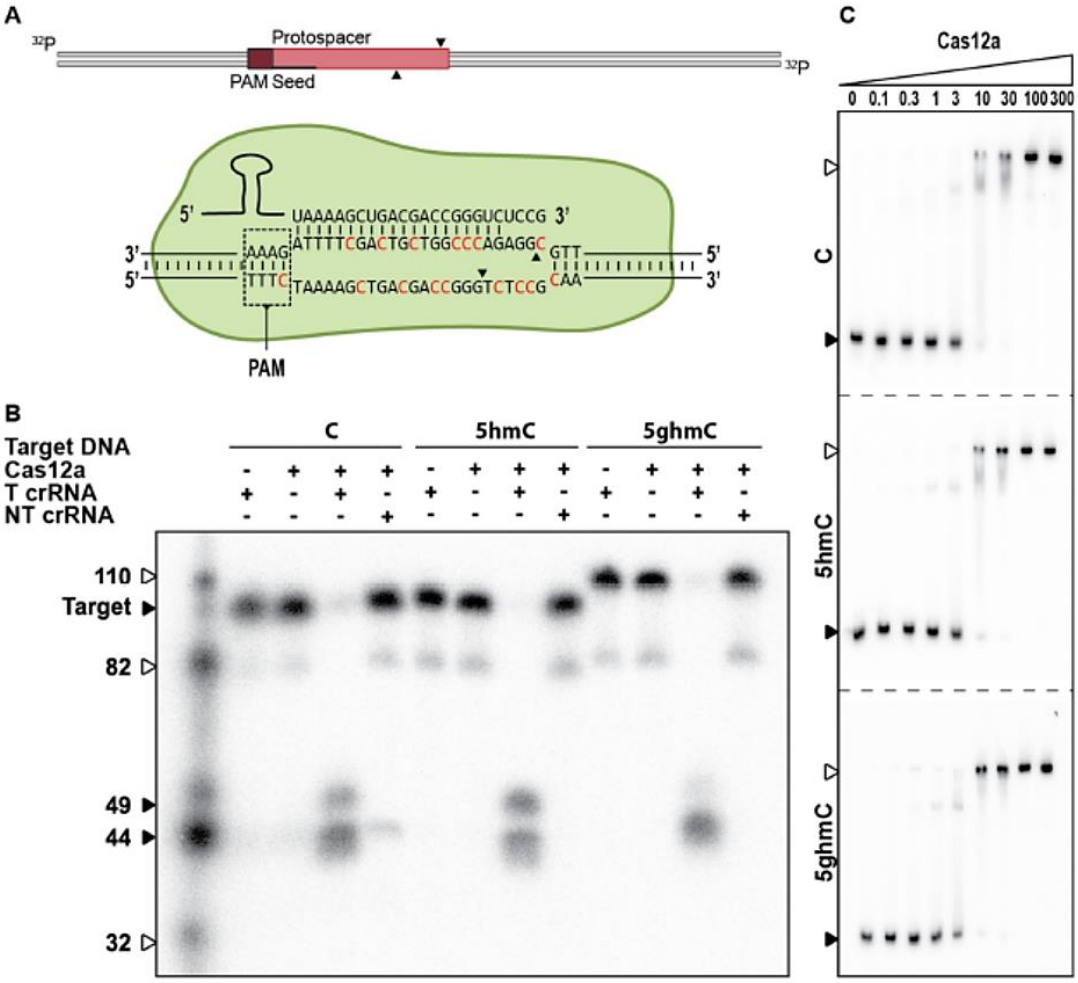


Figure 3.18 | *In vitro* type V-A CRISPR-Cas-sgRNA interaction with modified dsDNA (hmC and ghmC). (A) Schematic diagram of DNA targeting by Cas12a. Cytosine residue modifications are indicated in red, and black arrows indicated cleavage sites. **(B)** Cleavage assay of Cas12a on 98 bp dsDNA targets with different modifications (black arrow). Marker (bp) is shown with white arrows. T crRNA indicates targeting crRNA, while NT crRNA indicates non-targeting crRNA, either of which are loaded with Cas12a. Black arrows at 49 bp and 44 bp indicate restriction products **(C)** EMSA of Cas12a performed on unmodified dsDNA, dsDNA with hmC, or ghmC (black arrows) at increasing Cas12a concentrations (nM). White arrows indicate bound target fractions. Different gels are separated by dotted lines. Figure taken from Vlot et al., 2017, with permission.

Multiple strategies are used for phage to survive against CRISPR-Cas systems, and phage DNA modifications could be one of them. The *in vitro* assays performed in this study show that certain types of CRISPR-Cas systems can indeed, be affected by DNA modifications. Here, the results show glucosylated DNA severely impairs type I-E and II-A CRISPR-Cas systems *in vitro* (largely by reducing target DNA binding affinity), but have no effect on type V-A CRISPR-Cas systems. Apart from the results shown in this study, Dupuis et al., (2013) have proved that in *S. thermophiles* (type II CRISPR-Cas system), adenine methylations of phage DNA in 5'-GATC-3' sequences do not impair CRISPR interference or spacer acquisition. Yaung et al., (2014) combined bioinformatics search and experimental evidence, and demonstrated that the CRISPR-Cas system of *Streptococcus pyogenes* (type II) is not impeded when phage modify their DNA by adenine/cytosine methylation, cytosine hydroxymethylation or hmC glucosylation. These results indicate different CRISPR-Cas systems can have different interference outcome when dealing with modified DNA. Also, it is suggested that the degree of inhibition by CRISPR-Cas systems depends to target DNA sequence, or the positions of modifications, phage life style, as well as the state of phage DNA (Strotskaya et al., 2017).

Among the types of modifications, glucosylation is of great interests, as it is commonly used in phage DNA. Inhibition caused by glucosylation can be explained as steric hindrance effect, meaning the glucose group hinder the effector complexes from interacting. In addition, glucosylation can change DNA structure and stability, and can form hydrogen bonds between glucosyl moiety of side groups and neighbouring bases (Hunter, 1996). As a consequence, properties such as base pair angles are altered and specific structure build for effector actions are changed (El Hassan and Calladine, 1996). Further studies can be carried out regarding to the difference of α-glucosylated DNA and β-glucosylated DNA. In this study all DNA templates were prepared via enzymatic glucosylation using T4 β-glucosylase resulting in β-glucosylated DNA, whereas in Phage DNA a mixture of α -glucosylated DNA and β-glucosylated DNA are present. As for methylated cytosine, the inhibition of Cascade/Cas3 activity is unexpected, and can yet be explained. Further replicate experiments are needed to confirm this observation. In addition, *In vivo* studies may be used in conjunction with mutant phage DNA with mC to further confirm these results. For DNA containing phosphorothioates, the initial results of cleavage resistance are promising, as they are expected to be strong against nuclease activity, which made them popular candidates for therapeutic oligonucleotides (since they can remain intact when delivered).

The co-operation of RM and CRISPR-Cas systems can increase the survival rate of bacteria against phage invasion. However, the defence system is still far from perfect. Both phage and bacteria use multiple attack/counterattack strategies, in order to survive the strong and continuous selective pressure imposed by nature. Practically, CRISPR-Cas systems are already used as powerful gene-editing tools, and can be further manipulated/ engineered for their efficiency and accuracy. Designate DNA site modifications can be a tool to help achieve this. Specifically, the unique resistance of type V-A CRISPR-Cas systems against glucosylated DNA could prove useful in gene editing.

3.5 Conclusions

Chemical modifications of DNA are one possible strategy used by phage to evade CRISPR-Cas systems. In this study, the effect of DNA modifications on 3 different types of CRISPR-Cas systems were studied: a) type I-E (Cascade/Cas3) of *E. coli*; b) type II-A with Cas9; and c) type V-A with Cas12a from *Francisella novicida*.

Initial work focused on the optimisation of the synthesis of modified dsDNA substrates using PCR. DNA substrates were synthesised containing the following modifications: 5-methylcytosine, 5-hydroxymethylcytosine, 5-glucosylhydroxymethylcytosine, 2'-deoxyuridine, 5-hydroxymethyl-2'-deoxyuridine and phosphorothioate linkages. The efficiency incorporating modified dNTPs using PCR depends on the type of modified dNTP used. It is particularly challenging for the synthesis of hmU and phosphorothioate containing dsDNA. For α-thio-dNTPs, up to 2 types of α-thio-dNTPs can be incorporated to generate the required yield of DNA. Following the synthesis of 5-hydroxymethylcytosine modified DNA, glucosylated dsDNA was enzymatically synthesised using T4 beta glucosyltransferase.

In vitro binding/cleavage reactions of range of different CRISPR-Cas systems were performed by Marnix Vlot in collaboration with Wageningen University using the synthesised DNA substrates. Glucosyl modification of 5-hydroxymethylated cytosines in DNA interferes with type I-E and type II-A CRISPR–Cas systems. In contrast, the CRISPR–Cas type V-A system cleaves glucosyl-5-hydroxymethylated cytosine bases in DNA. Glucosylation of DNA, specifically, was found to prevent target DNA from cleavage by lowing the binding affinity in both type I-E and type II-A CRISPR-Cas systems, but not prevent from cleavage in type V-A CRISPR-Cas systems, where binding affinity remains unaffected. In general, these results offer some exciting prospects that DNA modifications used by phage to overcome restriction modification system of bacteria, can indeed affect interference of the CRISPR/Cas systems. These findings have important implications that could be exploited for genome engineering applications.

Statement

In Chapter 3, the preparation of DNA oligos with various modifications was carried out by myself (the author of this thesis), and the *In vitro* binding/cleavage reactions of CRISPR-Cas systems were performed by Marnix Vlot in collaboration with Wageningen University.

Chapter 4: Development and application of HPLC and LC-MS methods for the analysis of nucleic acid modifications

4.1 Abstract

In this Chapter I have developed and applied analytical methods for the analysis of nucleoside modifications. Analytical methods using reverse phase HPLC coupled with either ultraviolet detection (UV) or mass spectrometry analysis (LC-MS) were developed and optimised to separate and identify nucleosides (dA, dT, dG, dC and dU) as well as additional modifications including: 5-Methyl-2'-deoxycytidine (mC), 5-Hydroxymethyl-2'-deoxycytidine (hmC), 5-Hydroxymethyl-2'-deoxyuridine (hmU) and glucosyl-5-hydroxymethylation of cytosine (ghmC). For HPLC separation, two types of columns were employed: a) Superficially porous particles in conjunction with a C30 stationary phase (Accucore) and b) Porous graphitic carbon stationary phase (Hypercarb).

Following optimisation of the HPLC and LC-MS methods, I applied these approaches to identify and quantify DNA modifications in a range of biological systems including DNA generated *in vitro* using PCR, DNA extracted from different strains of T4 phage and DNA extracted from *E. coli* engineered to incorporate modified nucleosides. With HPLC-UV, *in vitro* synthesised DNA molecules were successfully varified with the incorporation of mC, hmC, U and hmU. For phage DNA modifications, HPLC-UV identification results are inconclusive, as dC in phage DNA tends to be glucosylated, and I was unable to detect dC in these samples. For engineered *E. coli* plasmid DNA, hmC was identified and quantified. In addition, LC-MS was also used, and good sensitivity was shown for the identification of nucleosides, specifically for ghmC, which could not be detected using HPLC-UV alone. Using LC-MS, ghmC was identified in all types of DNA samples.

4.2 Introduction

High performance liquid chromatography (HPLC) is a reliable and sensitive analytical technique used in compound identification, separation and quantification. For nucleic acid analysis, reversed phase liquid chromatography (RPLC) and ion-exchange chromatography are the most commonly used techniques (Tomiya et al., 2001; Meynial et al., 1995). RPLC is a technique that separates according to reversible adsorption of analytes to the stationary phase based on their hydrophobicity. Non-polar solvents (such as acetonitrile, methanol and isopropanol) that decrease the polarity of the mobile phase are used for elution. During elution, nucleosides that are more hydrophilic are weakly retained on the stationary phase, while more hydrophilic nucleosides exhibit stronger binding to the stationary phase, and therefore elute later in a gradient when increased amounts of an organic solvent are used. Being flexible with high ligand density and high peak capacity, RPLC is widely used as a separation method for small molecules such as nucleosides. However, nucleoside analysis with RPLC can be challenging, as these molecules are relatively hydrophilic, hence poorly retained on, for example, a traditional C18 column. The use of ion-pairing reagents is one way to increase retention, but will also add complexity, increase column equilibration time and are not always compatible with MS. By using lower organic content in the mobile phase can increase the retention of more hydrophilic nucleosides, but can sacrifice the separation of more hydrophobic analytes.

In typical RPLC columns, stationary phases usually consist of hydrophobic alkyl chains such as octyl (C8), octadecyl (C18) and phenyl as ligands, which are covalently attached to inert non-polar particles (usually silica) (Figure 4.1). Increased hydrophobicity of the functional groups on the stationary phase means better retention ability for non-polar analytes. Commercially available columns include Accucore C18 and C30 from Thermo Scientific.

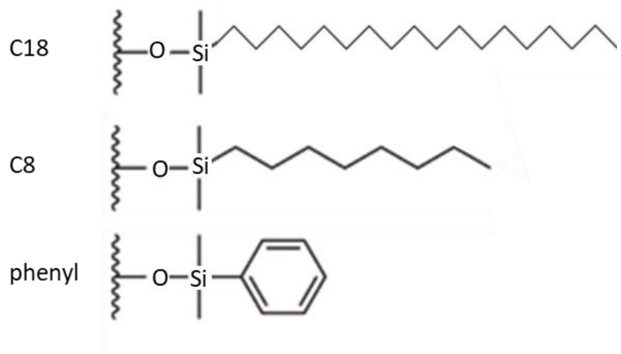


Figure 4.1 | Examples of commonly used surface groups as stationary phases in reversed phase chromatography.

Accucore C30 is a commercially available reversed phase column with superficially porous particles (or Core Enhanced Technology particles), which is composed of a solid silica core coated by a porous silica shell (Figure 4.2A). The particle size of Accucore C30 columns is 2.6 µm in diameter, with a pore of 150 Å. The superficially porous phases give several benefits, including high efficiency with reduced resistance to mass transfer and minimised eddy diffusion (Figure 4.2). Also, the back pressure generated by superficially porous columns is significantly lower compared to alternative ultra-high performance liquid chromatography (UHPLC) columns.

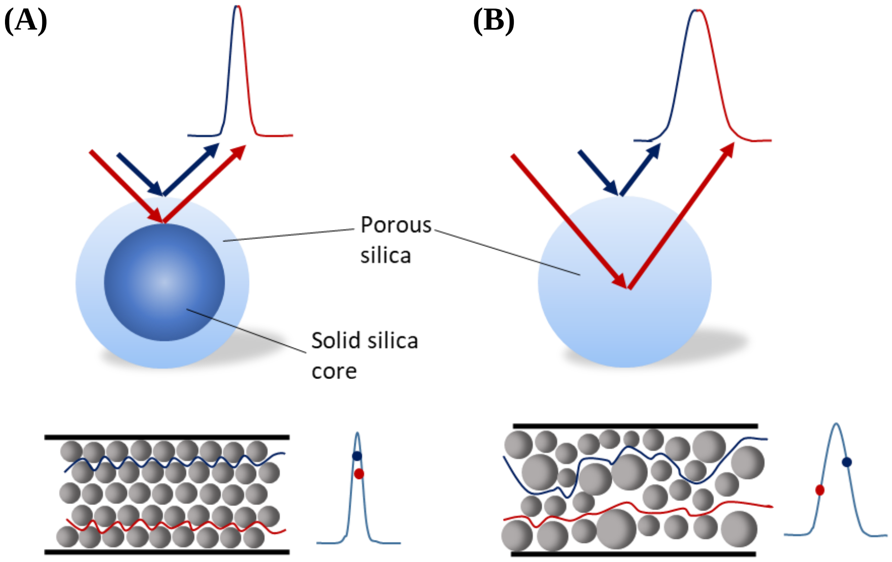


Figure 4.2 | Accucore C30 with Core Enhanced Technology. (A) Uniform packed Accucore C30 superficially porous particles, compared to **(B)** random packed totally porous particles. In Accucore C30, mass transfer resistance is reduced, with limited analyte diffusional path due to porous layer depth. Also, Eddy diffusion is minimized (lower panel).

Another type of RPLC column that been widely used in nucleotide/nucleoside analysis is porous graphitic carbon (PGC), which utilizes a special form of carbon introduced in 1979 by Knox and Gilbert (Gilbert et al., 1982). Similar to alkyl chain based columns, PGC columns are effective for the retention of non-polar analytes, but using a completely different mechanism. PGC is composed of large flat carbon atom layers, where carbon atoms are trigonal hybridized (sp^2) and hexagonally arranged with covalent bonds. Compared to true three-dimensional graphite (Figure 4.3A), in which successive layers are regularly oriented, PGC is actually two-dimensional graphite structured (or so called "turbostratic graphite") with a greater d-spacing value (>0.34 nm, Figure 4.3B). These layers are kept intact through Van der Waals forces, which provides mechanical stability and rigidity (West et al., 2010). The retention mechanism of PGC is due to the extensive layers of carbon atoms with π electrons (delocalized) and high polarizability. More planar analytes can align closer to the surface of graphite because of more interaction points, resulting in increased retention ability. At the moment, commercially available PGC particles used in HPLC columns are usually 5-7 µm in diameter, with a specific surface area of 120 m^2g^{-1}. Pore structure of these particles are sized around 25 nm, with a porosity of 75%. In this study, the PGC column used is a Hypercarb column, manufactured by Thermo Scientific. This column is stable in a wide range of solvents, with a working range of pH from 0 to 14.

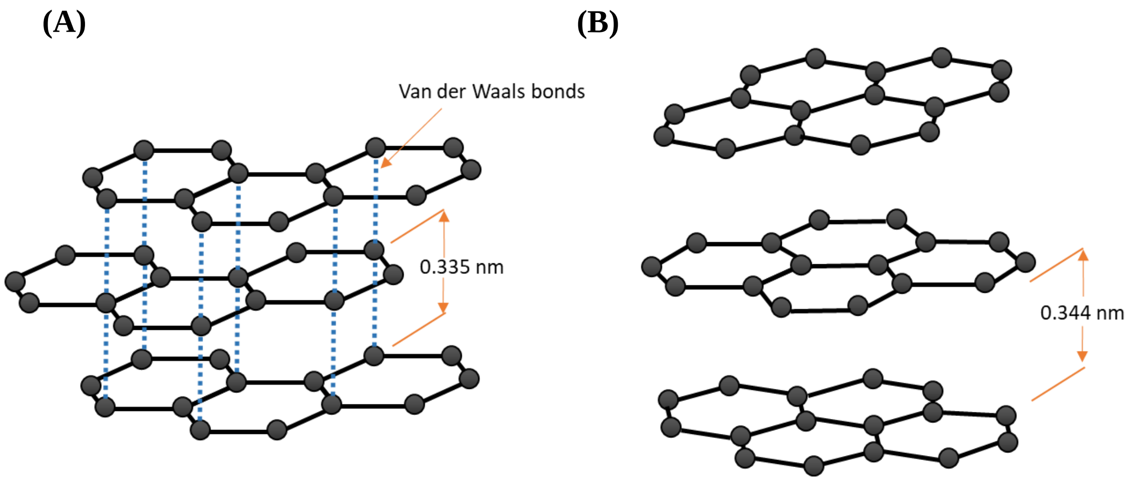


Figure 4.3 | Atomic structures of graphite. (A) 3D graphite structure with ABAB layer registration. (B) 2D turbostratic graphite structure in PGC, without layer registration.

Mass spectrometry (MS) is widely used for the characterization of nucleic acid components including nucleotides, nucleosides and nucleobases (Dudley and Bond, 2014). MS provides accurate mass-to-charge ratio (m/z) of the analytes, which often enables unambiguously identification of the corresponding analyte. For nucleoside analysis, positive ionization is often used, due to the proton accepting ability of nitrogen in nucleobase (Yen et al., 1996). Electrospray ionization (ESI) is generally used, as this soft ionization method gives less fragmentation and more compound adduct information, for example, $[M+H]^+$, $[M+Na]^+$ and $[M+K]^+$, which are useful for compound interpretation. Further fragmentation data can be generated using tandem MS (MS^n).

In this study, a HPLC based assay in conjunction with MS was developed in order to identify and quantify modifications in nucleic acids. For sample preparation, nucleic acids were enzymatically hydrolysed to release nucleosides including the modified ones. For nucleoside separation, both superficially porous silica particles (C30) and porous graphitic carbon were used on modified DNA samples including PCR synthesised dsDNA, phage DNA and *E. coli* plasmid DNA. Detection of the nucleosides was performed using both a UV detector and in conjunction with MS. The workflow is described in Figure 4.4.

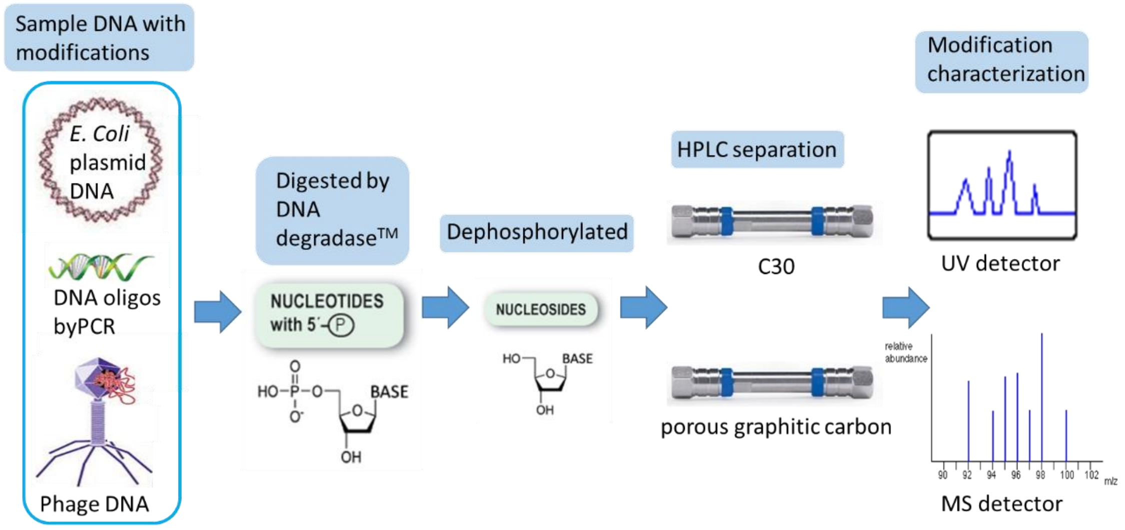


Figure 4.4 | Workflow of nucleic acid analysis using HPLC-UV and LC-MS. Nucleic acids samples are digested into nucleosides, which are then separated using HPLC with different types of columns, and analysed by UV detector or mass spectrometer.

4.3 Results and discussion

4.3.1 Assay validation for nucleoside analysis

Initial work focussed on the optimisation of HPLC methods for the analysis of nucleosides using two types of stationary phase: superficially porous particles in conjunction with a C30 stationary phase (Accucore), and porous graphitic carbon column (PGC) (Hypercarb), both from Thermo Fisher Scientific. A schematic workflow used in this study is shown in Figure 4.4. A mixture of dephosphorylated dNTPs (A, T, G and C) was used as the standard for assay optimisation and validation. For Accucore C30 columns, 5 mM ammonium acetate, pH 4.5 was used as buffer A, and 40% acetonitrile (v/v) was used as buffer B. For PGC Hypercarb columns, a three buffer system was used: buffer A with 10 mM ammonium acetate (pH 4.5), buffer B with 40% acetonitrile (v/v) with 10 mM ammonium acetate, pH 4.5, and buffer C with 95% methanol (v/v). The gradients used are shown in Figure 4.5 and Figure 4.7 for Accucore C30 and PGC Hypercarb (for detailed information see Materials and Methods, section 2.4). Both UV detection and MS analysis were used for peak identification. For C30 Accucore (Figure 4.5), with 5 µl sample loaded (approximately 200 pmol of each nucleoside), nucleoside dC, dG, dT and dA were eluted at 1.7 min, 5.5 min, 6.1 min and 11.3 min, respectively. Under these conditions, dC shows relatively weak retention as it was eluted at an early stage. However, since it was separated from other solvents/salts, the early elution is deemed acceptable. With this method, all four nucleosides were successfully separated from each other. The 4 individual nucleoside peaks were further confirmed in conjunction with MS analysis (see Figure 4.6). The majority of the nucleosides were observed as the $[M+H]^+$, $[M+Na]^+$ or $[M+K]^+$ ions.

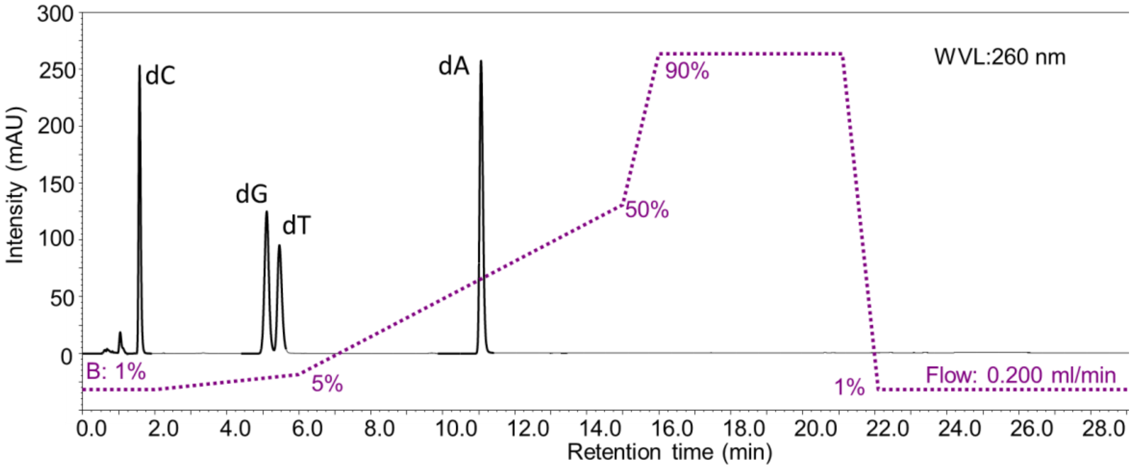


Figure 4.5 | Accucore C30 chromatogram of dephosphorylated dNTP mixture. The corresponding nucleosides are highlighted. Gradient is shown as the percentage of buffer B (dotted line).

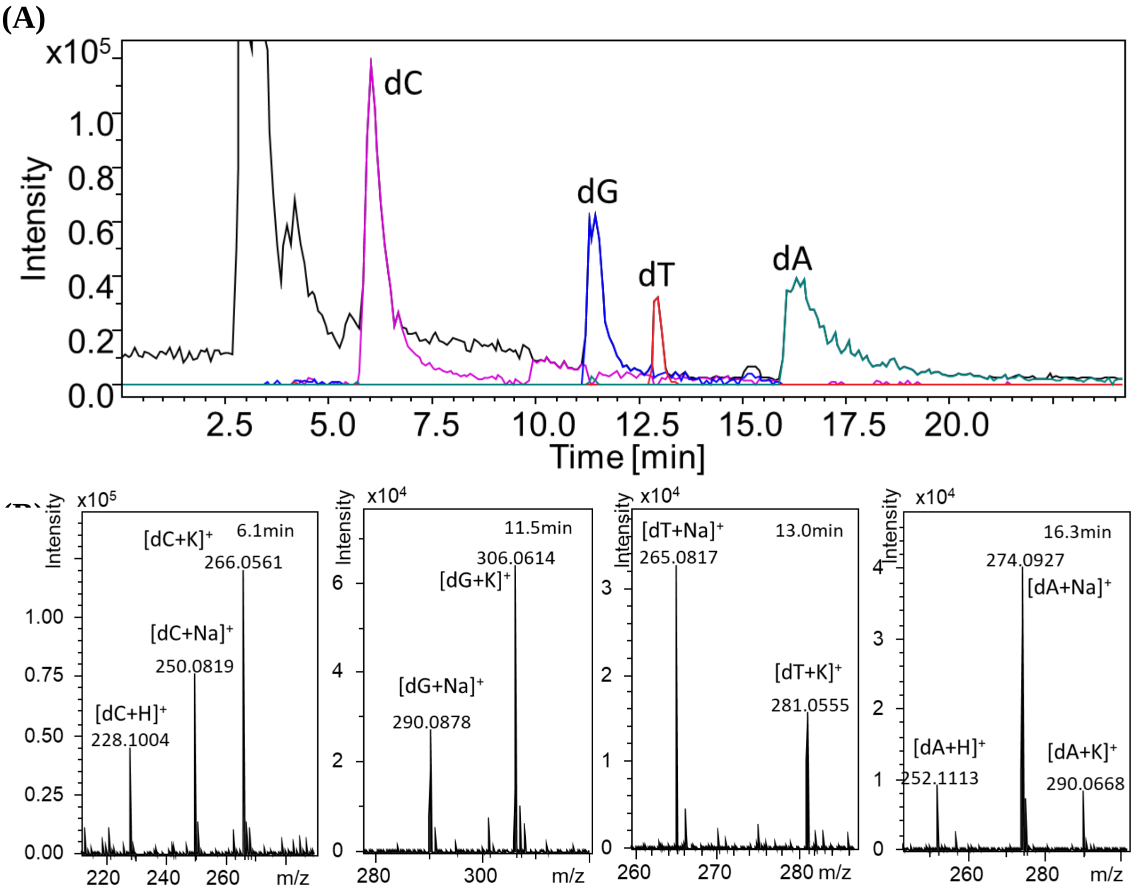


Figure 4.6 | LC-MS analysis of a dephosphorylated dNTP mixture using Accucore C30 column. (A) Extracted ion chromatograms of dA dT dG and dC (in colour) overlaid to base peak chromatogram (BPC) (black). **(B)** MS1 spectrums as evidence for individual dA, dT, dG, and dC nucleosides.

Following the analysis using the C30 Accucore column I also investigated an alternative stationary phase (porous graphitic) for the analysis of nucleosides. The validation approach for this Hypercarb column is similar to that performed on C30 Accucore column, the results are shown in Figure 4.7. Nucleoside dC, dT, dG and dA were eluted at 3.5 min, 7.6 min, 8.6 min and 10.2 min, respectively. The peaks were further confirmed by LC-MS (see Figure 4.8). The majority of nucleosides were observed as the $[M+H]^+$ and $[M+Na]^+$ ions.

In this comparison, the PGC Hypercarb column shows better retention for dC, as well as better separation of dT/dG using the gradients employed. Note that for PGC Hypercarb, dT elutes before dG. In contrast, on the Accucore C30 column dT elutes after dG. In general, PGC as stationary phase has better retaining ability for nucleosides, as all four nucleosides were eluted at approximately 50% buffer B, whilst for Accucore C30 the last nucleoside peak was observed at about 30% buffer B (buffer B for both column systems were similar). As for LC-MS analysis, Accucore C30 identified a large number of ions in $[M+Na]^+$ and $[M+K]^+$ forms, compared to PGC Hypercarb, where $[M+H]^+$ and $[M+Na]^+$ were the dominant forms. In this test, despite the good separation of dT and dG using the PGC Hypercarb column in conjunction with UV detection, dT and dG co-eluted using the gradient for the LC-MS analysis in the example shown (see Figure 4.8A). Despite the performance differences of the two columns, both the Accucore C30 and the PGC Hypercarb demonstrate the ability to separate nucleosides successfully under the conditions used.

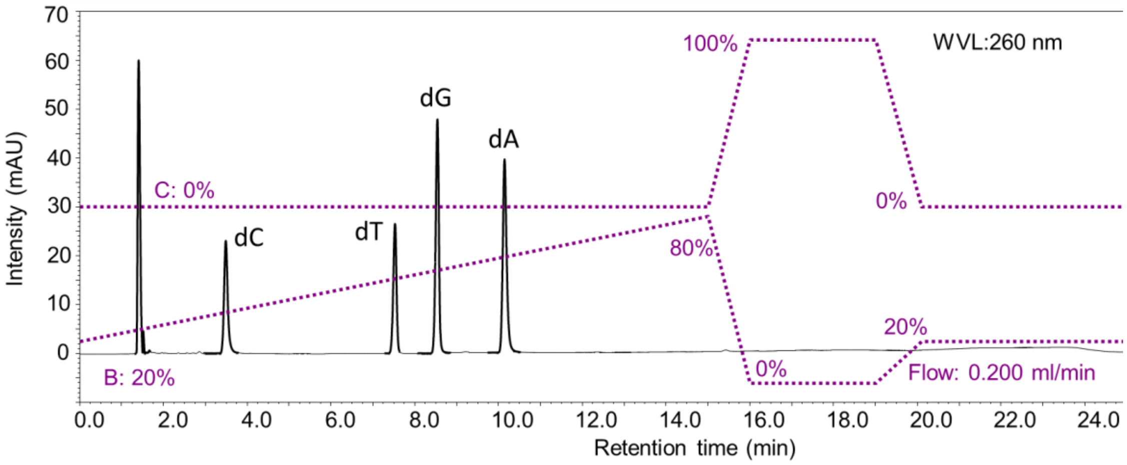


Figure 4.7 | PGC Hypercarb chromatogram of dephosphorylated dNTP mixture. Gradient is shown as the percentage of buffer B and C (dotted lines).

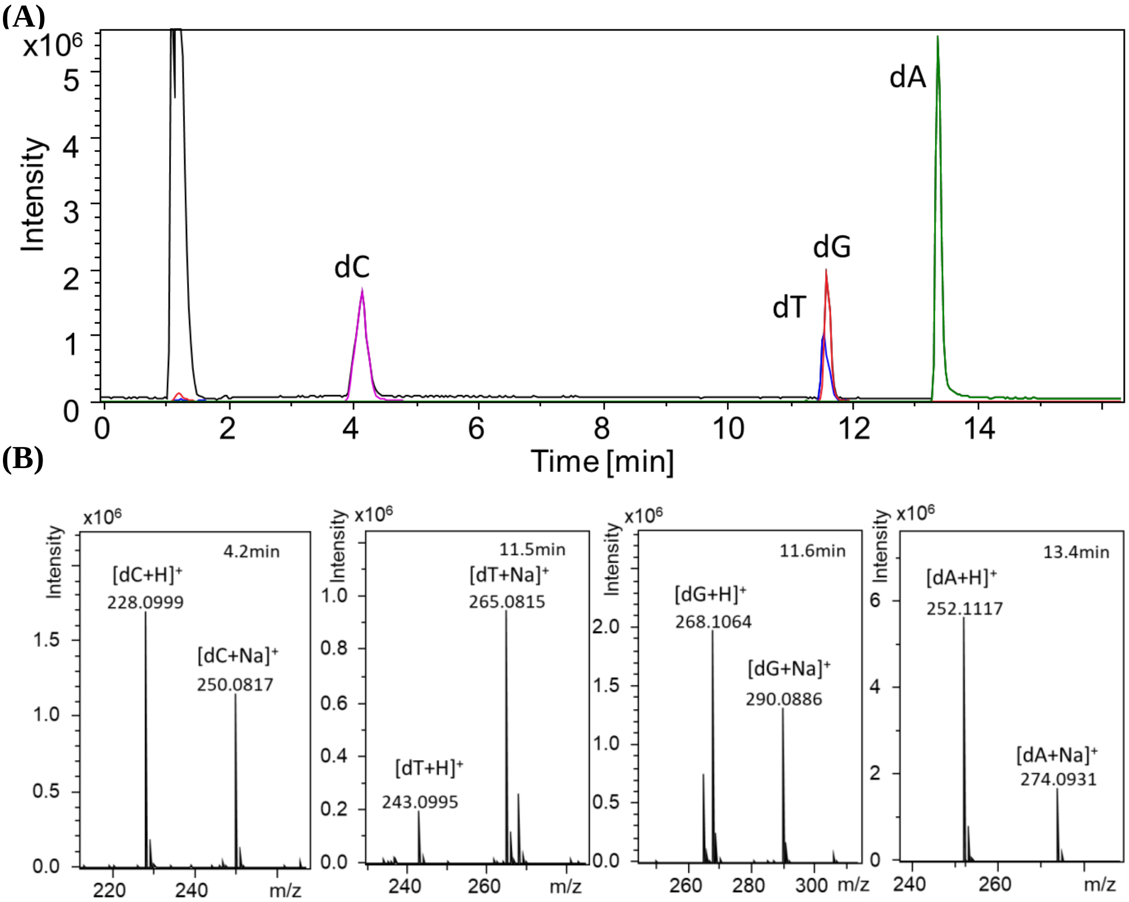


Figure 4.8 | LC-MS analysis of dephosphorylated dNTP mixture using PGC Hypercarb column. (A) Extracted ion chromatograms of dA, dT, dG and dC (in colour) overlaid to base peak chromatogram (BPC) (black). **(B)** MS1 spectrums as evidence for individual dA, dT, dG, and dC nucleosides.

4.3.2 Detection of nucleoside modifications in dsDNA generated using PCR

Following the validation of HPLC-UV/MS methods for nucleoside analysis, these approaches were then applied to identify modified nucleosides present in dsDNA generated by PCR (see Chapter 3.3). A wide range of dsDNA products were synthesised in conjunction with chemically modified nucleoside triphosphates to generate dsDNA products containing different chemical modifications including mC, hmC, dU, hmU and ghmC (see Figure 3.1). The aim is to determine if the modified dNTPs were successfully incorporated into the DNA by PCR, or successfully enzymatically synthesised (ghmC containing dsDNA). Prior to analysis, all dsDNA samples were purified using a DNA binding silica membrane to remove excess nucleoside triphosphates. The purified dsDNA substrates (containing mC, hmC, U or hmU) were then digested and dephosphorylated into nucleosides, and analysed using HPLC with UV detection at 260 nm. PCR synthesised dsDNA with unmodified dC was used as control. Samples were analysed on the Accucore C30, (see Figure 4.9 and 4.10). 9. The results show that, for mC containing DNA, a peak at 3.7 min was observed (Figure 4.9, line b), compared to unmodified dC at 1.6 min (Figure 4.9 line a). The shift in retention time reflects the more hydrophobic properties of mC, due to the additional methyl group. Note that a small peak for unmodified dC still exists in sample with mC, which is due to the presence of small amounts of dC in the primer sequences used in PCR to generate the dsDNA. For samples containing hmC, a peak was observed at 1.9 min (Figure 4.9, line c). The retention time shift compared to dC peak was due to the increase in hydrophobicity of hmC. Similarly, a small peak for dC was detected due to the presence of small amounts of dC in the primer sequences. A peak at 4.3 min was detected in all digested samples using DNA degradase but not standard dephosphorylated dNTPs (Figure 4.9 compared to Figure 4.5), indicating the peak was a contaminant from the DNA degradase buffer. In summary, the HPLC analysis demonstrates that mC and hmC were successfully incorporated into dsDNA products by PCR and therefore validate the presence of the modified nucleosides in these samples.

For the characterization of nucleosides dU and hmU, the same assay was used. Results are shown in Figure 4.10. A new peak for dU appeared at 2.4 min, and the peak for dT was reduced (remaining dT peak is due to the presence of dT in primers). For hmU, the retention time shifted to 2.6 min, indicating the hydrophobicity change caused by the additional hydroxymethyl group compared to dU. Again, these results suggest that dU and hmU were successfully incorporated into the dsDNA products generated by PCR.

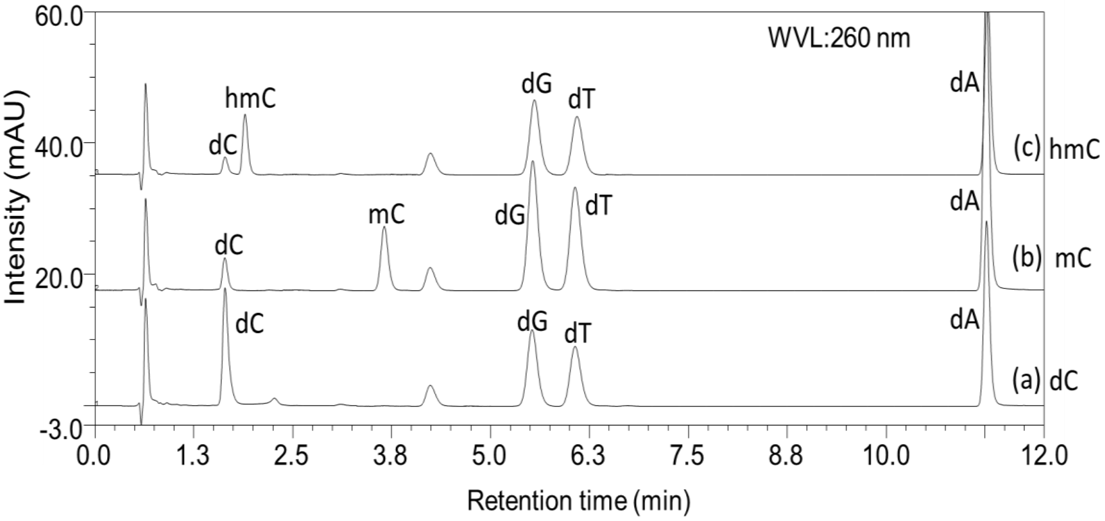


Figure 4.9 | HPLC-UV characterization of dsDNA (synthesised by PCR) containing nucleoside modifications. Chromatograms of digested dsDNA synthesised using unmodified dC (a), mC (b) and hmC (c).

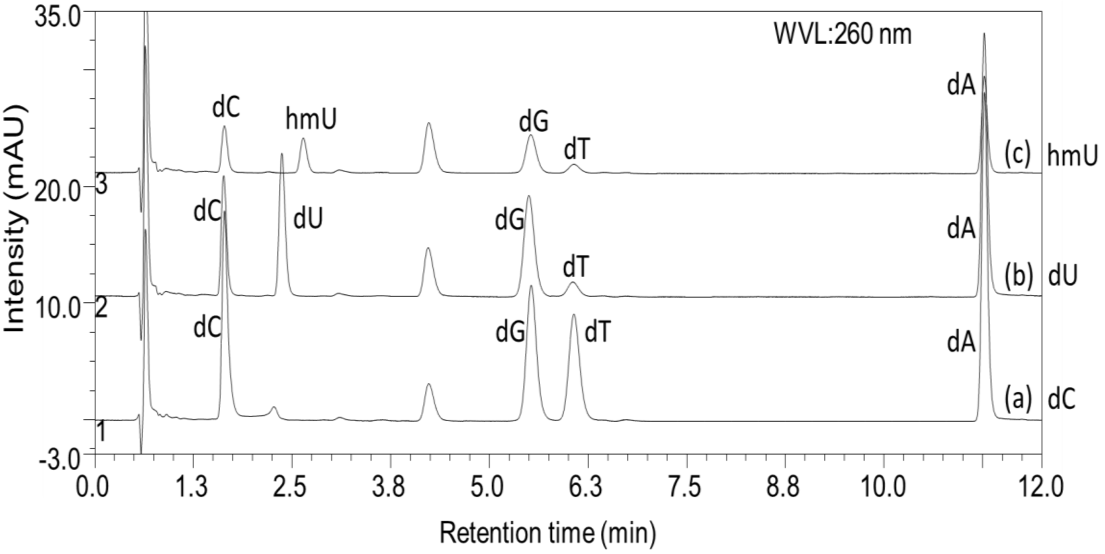


Figure 4.10 | HPLC-UV characterization of dsDNA (synthesised by PCR) containing nucleoside modifications. Chromatograms of digested dsDNA synthesised using unmodified dC (a), dU (b) and hmU (c).

4.3.3 Detection and verification of glucosylated dsDNA using LC-MS

Following the synthesis of hmC containing DNA using PCR and subsequent enzymatic glucosylation using T4 Phage β-glucosyltransferase (see Figure 3.13), verification of the glucosylated DNA was performed using HPLC and LC-MS methods. To do this, glucosylated product dsDNA was digested and dephosphorylated into nucleosides prior to analysis using the PGC Hypercarb column (Figure 4.11). From the UV chromatogram a new peak at approximately 7.1 min was detected (shown by the arrow) which possibly corresponds to ghmC.

In order to confirm the new peak identified in Figure 4.11 is actually ghmC, further characterisation using mass spectrometry was performed. The same sample was analysed using HPLC (with PGC Hypercarb column) in conjunction with a maXis (UHR-TOF) instrument. The results are shown in Figure 4.12. Previous LC-MS studies have shown that for ghmC, diagnostic ions at m/z 420 and m/z 304 representing $[M+H]^+$ and losses of deoxyribose are observed for ghmC (Liu et al., 2014) (see Figure 4.12C). In this study, the analysis of the sample of ghmC showed abundant ions with m/z 420/304 at 4.9 min, confirming the presence of ghmC (the ions of m/z 420.16 and m/z 304.12 are for $[ghmC+H]^+$ and ghmC without the 2'-deoxyribose group, respectively).

From these results, it is also worth noting that a lack of evidence for hmC (m/z 258.11) was observed in the glucosylated DNA sample, indicating the PCR products with hmC were largely converted by T4 Phage β-glucosyltransferase. In addition, a dC peak (approximately 3.6 min, Figure 4.11, also see Figure 4.7 for comparison) was detected using both LC-UV and LC-MS, due to the presence of small amounts of dC in the primer sequences used in PCR to generate the dsDNA.

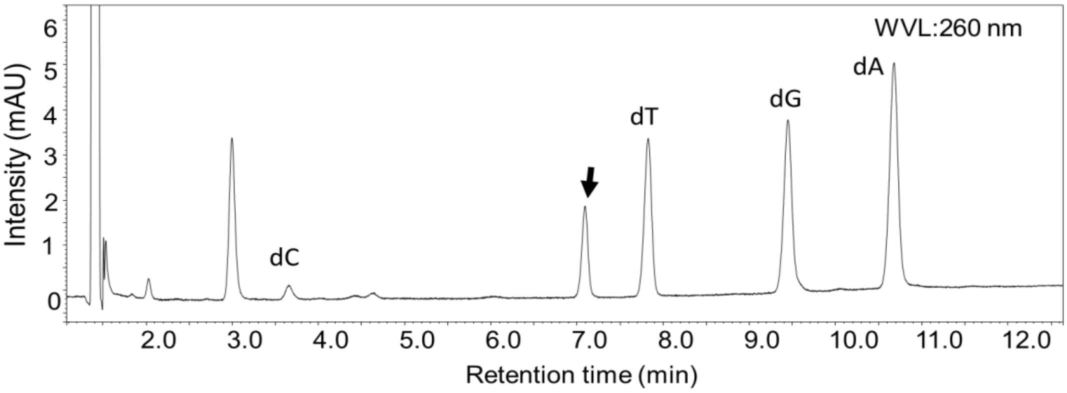


Figure 4.11 | HPLC-UV characterization of dsDNA containing ghmC. DNA digest was analysed using PGC Hypercarb column. The peak at approximately 7.1 min (highlighted with an arrow) possibly belongs to ghmC.

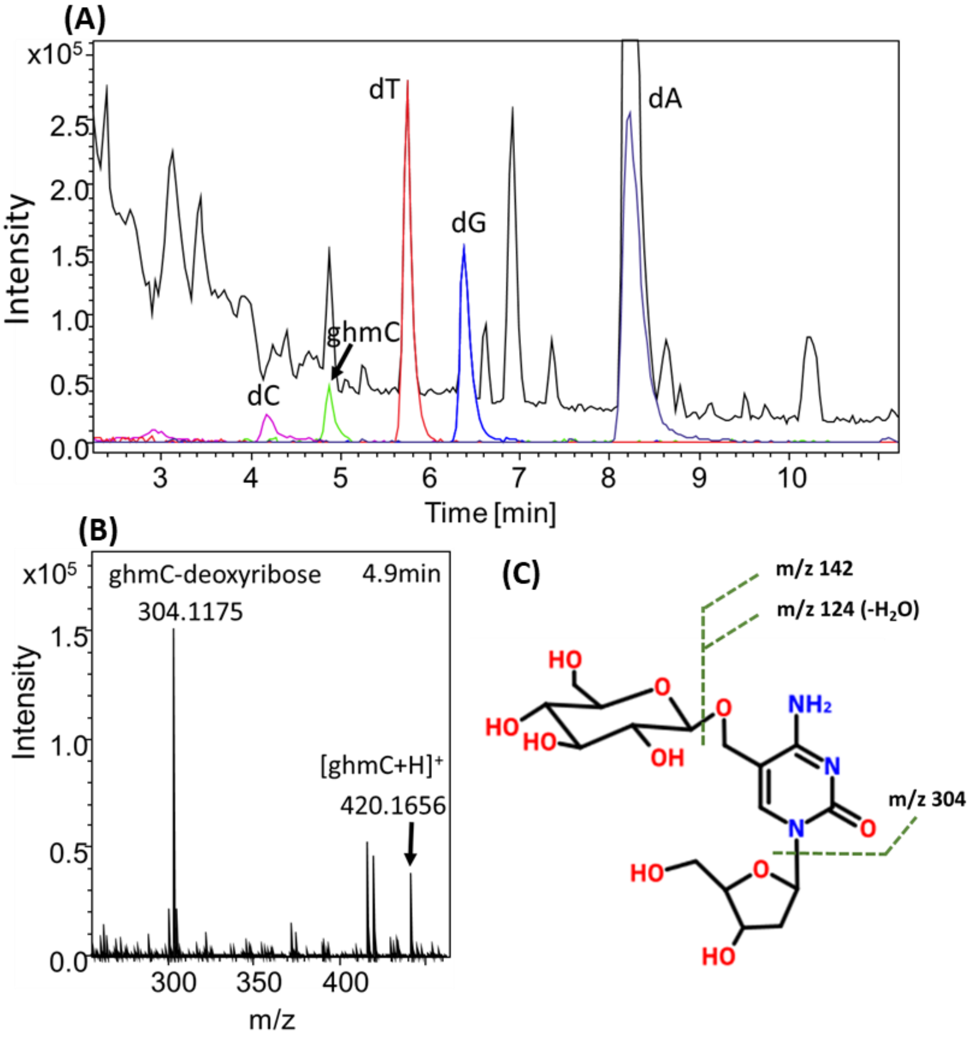


Figure 4.12 | Identification of ghmC containing dsDNA using PGC Hypercarb chromatography. (A) MS chromatogram of DNA digest containing ghmC. Extracted ion chromatograms of dA, dT, dG, dC and ghmC (in colour) are overlaid to base peak chromatogram (BPC) (black). ghmC feature peak was observed at 4.9 min. **(B)** Spectrum showing m/z evidence for ghmC. **(C)** Typical fragment ions observed for ghmC using LC-MS. For ghmC expected m/z is 420 in conjunction with fragmentation across the glycosidic bond resulting in the fragment ion of m/z 304. Fragment ions of m/z 142 and 124 are also generated in MS^2.

In comparison, the same dsDNA product with ghmC was also analysed using the Accucore C30 column. The results are shown in Figure 4.13. The UV chromatogram, however, is inconclusive for the identification of the ghmC peak, so further characterisation was carried out using HPLC in conjunction with a maXis (UHR-TOF) instrument. The extracted ion chromatogram for each of the ions corresponding to dC, dA, dT, dG and ghmC are shown in Figure 4.14. Using the diagnostic ions for $[ghmC+H]^+$ (m/z 420.16 and 304.12), a peak was observed at 11 min confirming the presence of ghmC, as discussed previously.

Similar to analysis using PGC Hypercarb, no evidence for hmC was observed with UV detector (compare Figure 4.13 to Figure 4.9) or MS (m/z 258.11) in the glucosylated DNA sample, indicating the full conversion of hmC to ghmC. A small dC peak was detected in both LC-UV and LC-MS, showing the existence of dC in primers used for PCR synthesis of dsDNA. In general, for ghmC characterisation, the results are consistent using both the PGC Hypercarb column and the Accucore C30 column. LC-MS was used for all subsequent analysis of ghmC due to difficulties in analysis via LC-UV.

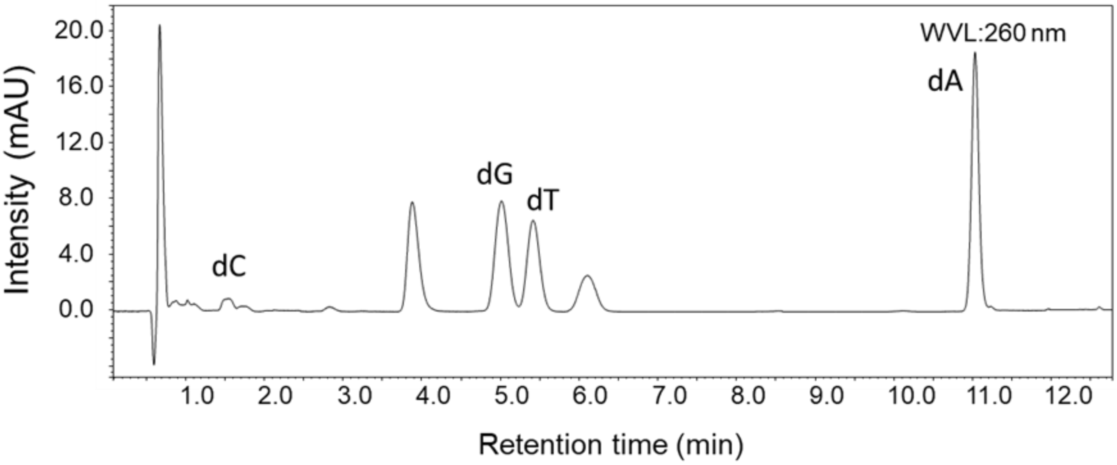


Figure 4.13 | HPLC-UV characterization of dsDNA presumably have ghmC. DNA digest was analysed using the Accucore C30 column. Unknown peaks which may belong to ghmC are to be determined using MS.

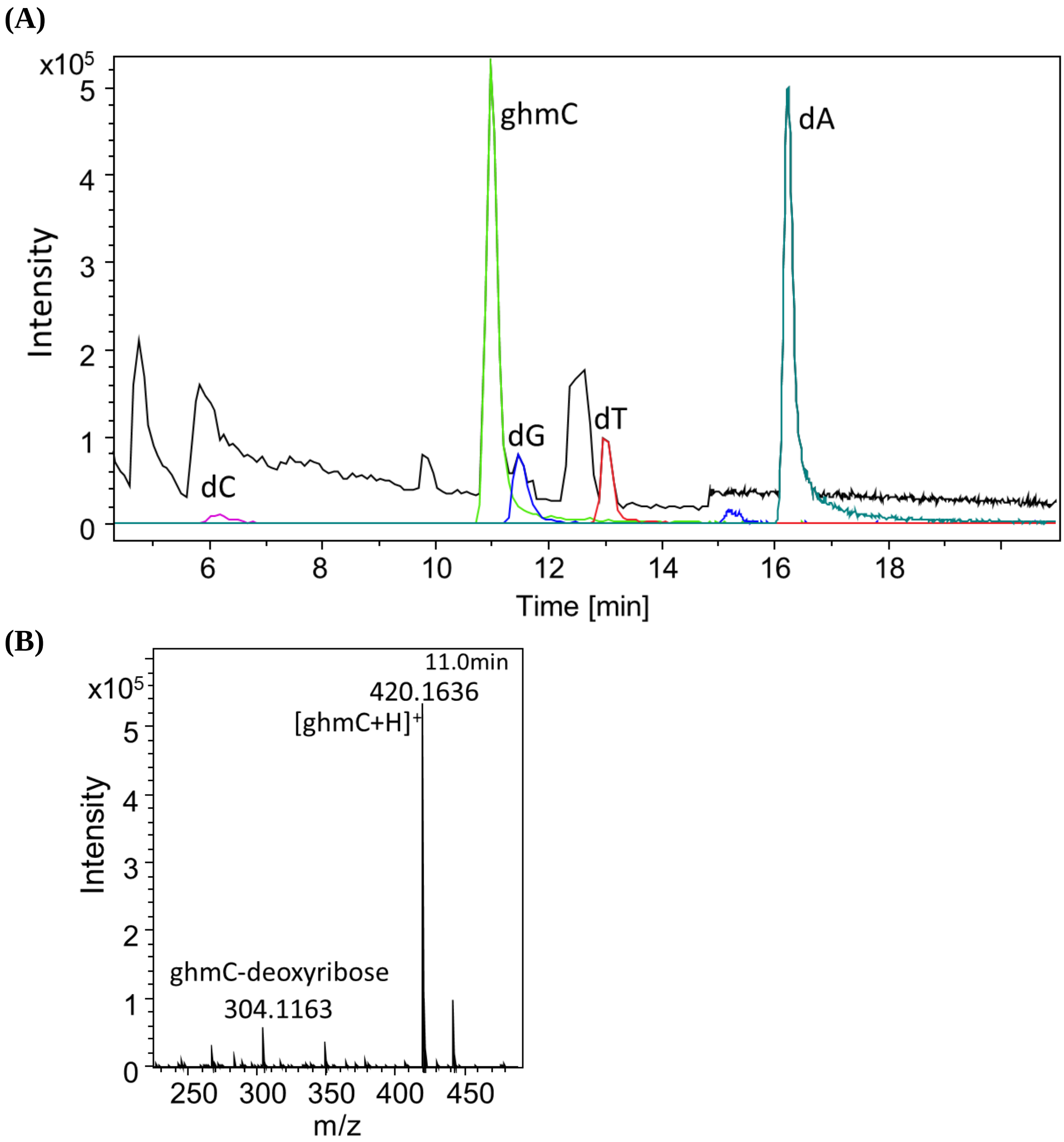


Figure 4.14 | Identification of ghmC containing dsDNA using Accucore C30 column. (A) MS chromatogram of DNA digest presumably have ghmC. Extracted ion chromatograms of dA, dT, dG, dC and ghmC (in colour) are overlaid to base peak chromatogram (BPC) (black). ghmC feature peak was observed at 4.9 min. **(B)** Spectrum showing m/z evidence for ghmC.

In summary, I have used LC-UV and LC-MS methods to validate the presence of a range of chemical modifications in PCR synthesised dsDNA. For the DNA digestion during the sample preparation, DNA degradase showed similar activity on the modified dsDNA including mC, hmC, U and hmU, compared to unmodified dsDNA (comparing the relative peak intensities in Figure 4.9 and 4.10, taking consideration of dsDNA concentration in samples). However, for the digest of ghmC containing DNA, the efficiency is reduced (comparing the relative peak intensities in Figure 4.9c and Figure 4.13, where similar copies of dsDNA with hmC or ghmC were digested, respectively).

This is expected, as glucosylation is one strategy living systems use to protect against the activity of nucleases. In this study, the nucleoside modifications mC, hmC, U, hmU were successfully detected with distinguishable retention times using an Accucore C30 column via UV 260 nm absorbance. For glucosylated samples, LC-MS was used and confirmed the existence of ghmC. Thus, HPLC-UV combined with LC-MS is a powerful technique for the analysis of modifications in DNA. These approaches were successfully used to validate a range of modifications in the DNA substrates used for CRISPR-Cas studies in Chapter 3.

4.3.4 Identification of DNA modifications present in T4 phage DNA

As a countermeasure to block endonucleases, T4 phage contain ghmC in their 169 kbp dsDNA genome. Here, the assays developed for the identification of nucleoside modifications were applied to analyse DNA extracted from three different strains of T4 phage, in order to identify and confirm the cytosine modifications present. The resulting modified T4 phage strains were to be utilised in downstream studies to study the effects of chemical modifications on CRISPR-Cas systems. Three types of T4 phage DNA samples acquired from collaborators of Wageningen University are described in Table 4.1.

Table 4.1 | Phage DNA samples acquired from collaborators of Wageningen University for modification identification

Strain	Description	Modification predicted	Concentration (ng/µl)
Phage B	T4c phage grown in *E. coli* B834	Normal dC	20
Phage CR	T4c phage grown in *E. coli* CR63 (amber suppressor)	ghmC	20
Phage WT	Wild type phage grown in B834	ghmC	120

Phage DNA samples were digested and dephosphorylated into nucleosides and analysed by HPLC using an Accucore C30 column. The results are shown in Figure 4.15. Since wild type phage DNA contains ghmC, the digest of enzymatically synthesised dsDNA with ghmC were used as standard (see previous section, Figure 4.13). From the UV chromatograms (Figure 4.15), the peaks corresponding to dG, dT and dA are observed. For dC, no peak at the expected retention time is observed, possibly due to the low retention of this hydrophilic nucleoside. Under these conditions the dC was likely to have eluted in the non-retained fraction at the injection peak, therefore making conclusions on the presence or absence of dC difficult. Therefore I focused on analysing the presence of hmC and ghmC in the phage DNA samples. The peaks marked with "*" at 1.5 min on phage B (c) and ghmC DNA(d), but not on phage WT (a) and phage CR (b), may belong to dC, as the retention time is in accordance with previous dC standards. However LC-MS analysis did not detect dC and therefore was possibly below the limit of detection. For ghmC, the MS analysis confirmed the identification of ghmC in all three types of phage DNA with the identification of the diagnostic ions for ghmC (m/z 420 and 304, see Figure 4.16). The ghmC peak intensities for phage WT and phage CR are high, consistent with the expectation they have ghmC in their DNA. For phage B, the peak intensity for ghmC is significantly lower, suggesting a smaller relative percentage of ghmC in their DNA. The LC-MS results confirm that DNA extracted from both WT phage and phage CR contain ghmC in their DNA. For phage B, which presumably has dC in its DNA, is likely composed of a mix of dC and ghmC.

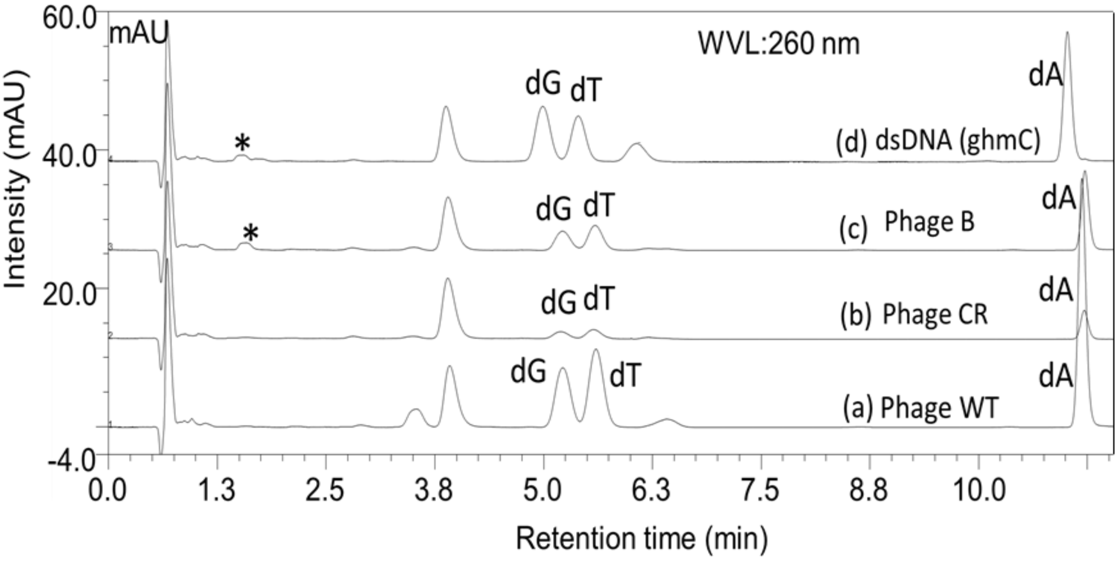


Figure 4.15 | Characterization of modifications in phage DNA with HPLC-UV. Chromatogram of (a) digested phage WT DNA, (b) digested phage CR DNA, (c) digested phage B DNA and (d) digested enzymatically synthesised dsDNA with ghmC from PCR product as standard. The peaks marked with "*" are potential peaks for dC.

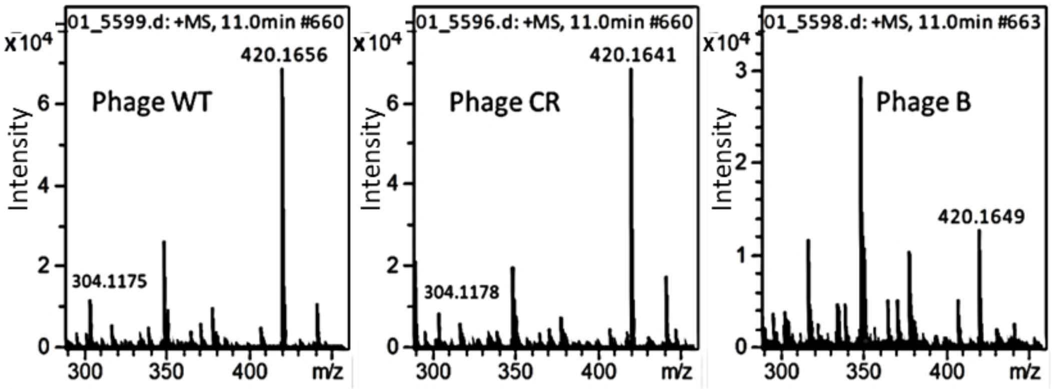


Figure 4.16 | LC-MS identification of ghmC in different phage DNA. The presence of ghmC was observed in all three types of phage DNA. The diagnostic ions for ghmC (m/z 420.16 and m/z 304.12) are shown. The peak intensities of m/z 420.16 indicate the amount of ghmC in phage B is significantly lower (1.3×10^4) compared to phage WT and CR (6.8×10^4).

4.3.5 Identification of DNA modifications in *E. coli* engineered to synthesise and incorporate modified nucleosides

The analysis of modified nucleosides was further applied to identify DNA modifications in *E. coli (cloni* 10G, Lucigen, relevant genotype: *endA1 recA1 mcrA Δ(mrr-hsdRMS-mcrBC) galU galK*) that was engineered to synthesise and incorporate modified nucleosides. *E. coli* plasmid DNA was obtained from the Laboratory of Microbiology, Department of Agrotechnology and Food Sciences, Wageningen University. Work performed at Wageningen University aimed to engineer biosynthetic pathways in *E coli*, aiming to replace a large fraction of cytosine with modifications including 5hmC and ghmC representing the known modifications present in T4 phage DNA. Such approaches will be useful to further examine the roles of these modifications and exploit the ability of *E. coli* to generate DNA with a variety of DNA modifications. Recent research using a similar approach already achieved approximately 75% of thymidine replacement by 5hmU, 63% cytidine replacement by 5hmC and about 20% cytidine replacement by ghmC, respectively (Mehta et al., 2016). In this study, attempts were made to verify and quantify the incorporation 5hmC and ghmC in *E. coli* using HPLC and LC-MS methods.

For modification identification, two batches of samples, each with three types of *E. coli* plasmid DNA, namely, pMK0 (control with dC), pHmC (engineered to incorporate hmC) and pGhmC (engineered to incorporate ghmC) were acquired from collaborators at Wageningen University. Samples were digested into nucleosides as described previously (see Figure 4.4), and analysed using HPLC-UV. Analysis was performed using both the Accucore C30 and the Hypercarb columns.

The results of the analysis on the Accucore C30 column are shown in Figure 4.17, with all 4 unmodified nucleosides clearly present. Analysis of the plasmid digest samples revealed a number of peaks that are not present in the analysis of the PCR synthesised dsDNA digest (see Figure 4.9 and 4.10 for comparison), which therefore complicated the analysis. Although differences can be observed, there is a lack of evidence showing the existence of hmC and ghmC based solely on the LC-UV results. Further analysis was carried out using LC-MS. For the MS analysis of plasmid pHmC, no peak for hmC (m/z 258.1) was detected, possibly due to the very low amount of hmC incorporated in plasmid pHmC, or it was completely absent. For the characterization of ghmC in plasmid pGhmC, the LC-MS results are shown in Figure 4.18. Identification of the corresponding diagnostic ions (m/z 420.16 and m/z 304.12 for ghmC at 11.4 min) are shown, confirming the presence of ghmC (Figure 4.18). This peak for ghmC is only observed in plasmid pGhmC.

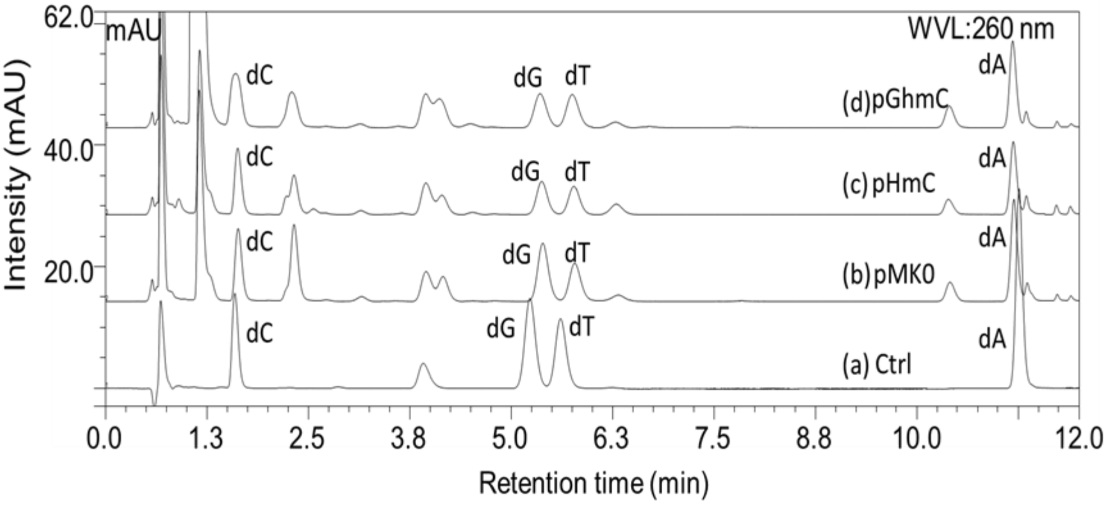


Figure 4.17 | Characterization of modifications in *E. coli* plasmid DNA with HPLC-UV. Accucore C30 chromatogram of nucleosides from *E. coli* plasmid DNA (batch 1) digestion: (a) PCR synthesised dsDNA digest (as control), (b) plasmid pMK0 digest, (c) plasmid pHmC digest, and (d) plasmid pGhmC digest.

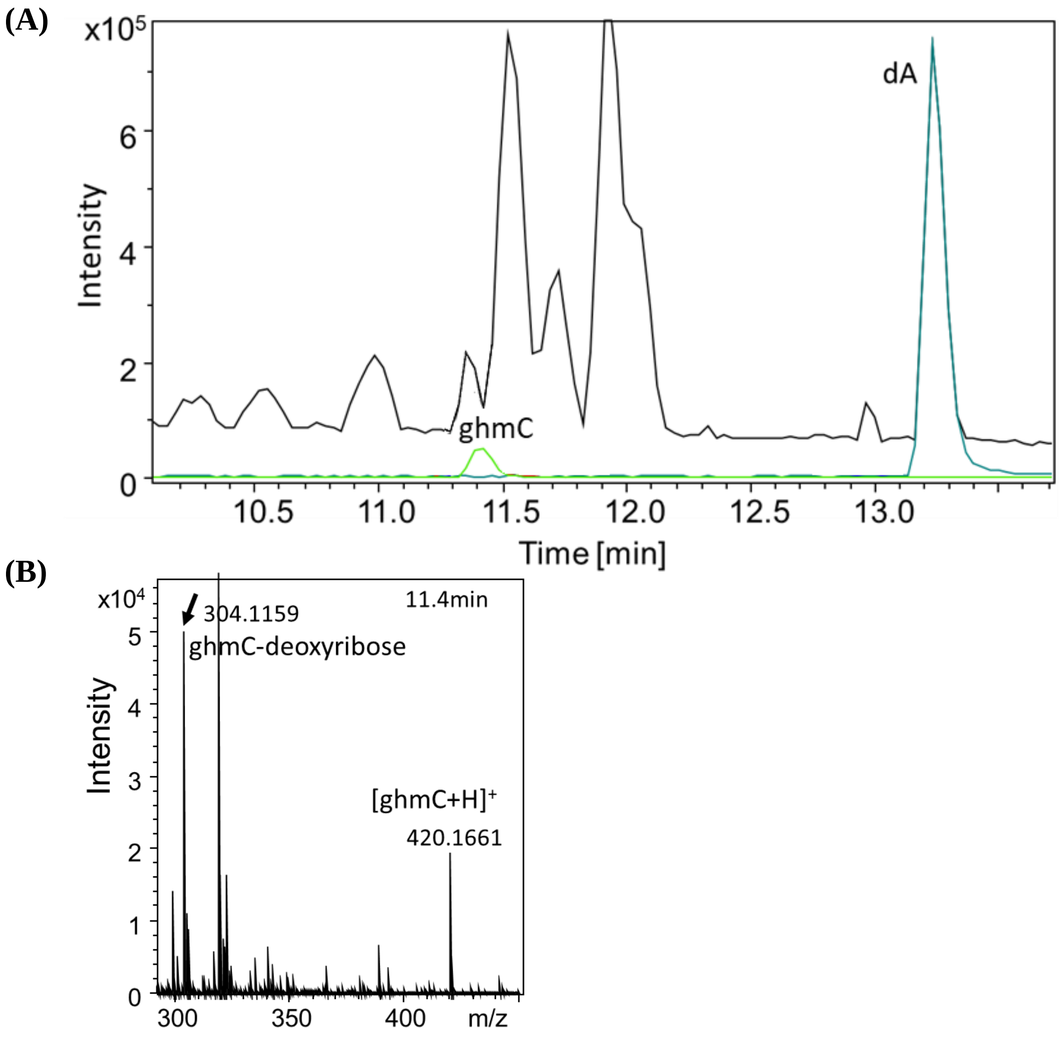


Figure 4.18 | LC-MS characterization of ghmC in *E. coli* plasmid pGhmC. (A) Based on the Accucore C30 column results (plasmid batch 1), extracted ion chromatogram shows the peak for ghmC (green peak at 11.4 min). **(B)** MS spectrum showing the evidence for ghmC with m/z 420.2 and 304.1.

Alternatively, analysis of replicate plasmid samples was performed using a PGC Hypercarb column. The results are shown in Figure 4.19. The Hypercarb column demonstrates increased retention of dC and dC analogues, as previously discussed. The HPLC-UV analysis in conjunction with the hmC standard (of which retention time at approximately 4.6 min, Figure 4.19a) was used to confirm the presence of hmC in plasmid pHmC (Figure 4.19c).

For plasmid pGhmC (Figure 4.19d), the peak at 3.7 min corresponding to dC is reduced compared to plasmid pMK0 and pHmC, indicating the amount of unmodified dC in plasmid ghmC is lower. For ghmC characterization, by comparing to the digest of the dsDNA with ghmC (Figure 4.19e), peaks at the same retention time (7.1 min) may belong to ghmC. Further evidence was obtained using LC-MS analysis of the dsDNA containing ghmC, where ghmC was confirmed eluting prior to dT and dG (see Figure 4.12). A number of unassigned peaks from the HPLC-UV analysis are also highlighted with "*" (Figure 4.19) in which were observed in the plasmid DNA digests which are possibly dC analogues.

Based on the HPLC-UV results, all three *E. coli* plasmid DNA contain unmodified cytosine base, the relative amount is high in plasmid pMK0 and pHmC, and low in plasmid pGhmC. The analysis of plasmid pHmC enabled the verification of the presence of peak corresponding to hmC, therefore indicating the successful incorporation of the modification into the *E. coli* plasmid DNA. For ghmC, HPLC-UV characterization is more difficult, as peak intensities are generally low. However LC-MS analysis confirmed the presence of ghmC in plasmid pGhmC.

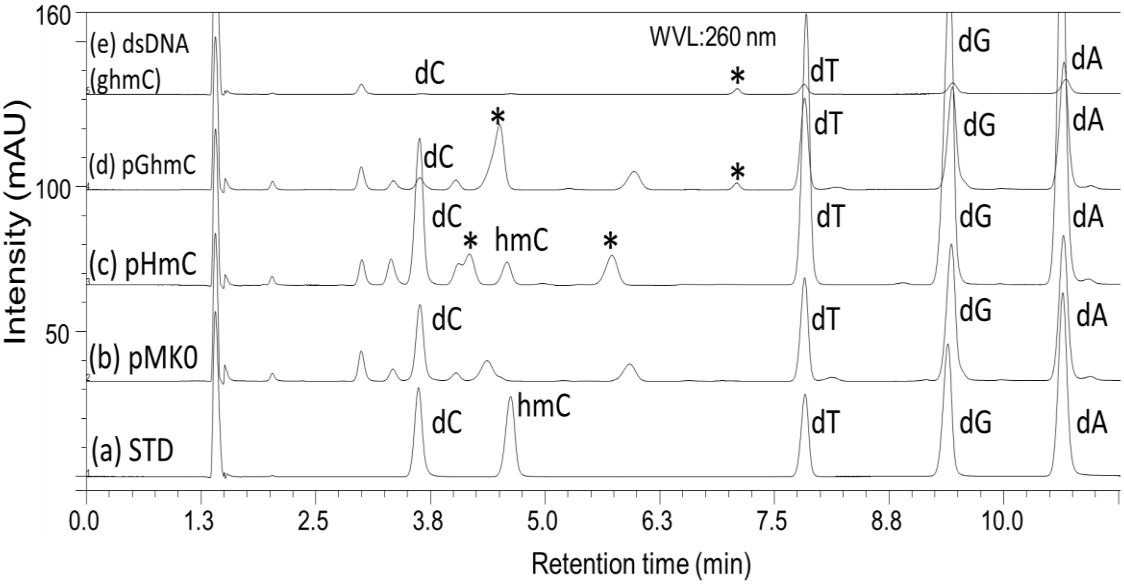


Figure 4.19 | Characterization of modifications in *E. coli* plasmid DNA with HPLC-UV. Hypercarb chromatogram of nucleosides from *E. coli* plasmid (batch 2) DNA digestion. (a) Equal amount of dephosphorylated dNTPs (unmodified) and hmC mixture as standard, (b) plasmid pMK0 digest, (c) plasmid pHmC digest, (d) plasmid pGhmC digest, and (e) oligo sp8 dsDNA digest with ghmC. Peaks marked as "*"in (d) and (e) at 7.1 min are potential ghmC peaks. Other "*" marked peaks are uncharacterized peaks which are unique in the sample.

4.3.6 Quantification of DNA modifications in *E. coli* plasmid DNA

Following the validation of hmC and ghmC incorporation into the DNA of engineered *E. coli* strains, further work was performed aiming to determine the incorporation rates, i.e., the percentage of modified bases in plasmid DNA. Initial work focussed on using the relative peak areas from the HPLC-UV chromatogram. Analysis was performed on the Hypercarb column, as the hmC peak is well separated from other peaks (Figure 4.19c). The chromatogram also shows the presence of multiple peaks that are not yet assigned or identified. Further work is required to analyse the unidentified peaks using LC-MS. To use the UV chromatogram for accurate quantification it is important to take into account differences in the extinction coefficients of the different nucleosides when measuring the area under the peaks (or area under curve, AUC), therefore equal amounts of hmC and dC standards were analysed. Results show the AUC of hmC and dC are slightly different (AUC of dC 3.1596 compared to AUC of hmC 3.3207, peaks shown in Figure 4.19a). For unbiased and accurate quantification, a normalisation factor of 0.95 was used when calculating the AUC of hmC peak. The relative quantification of hmC (as percentage) in the engineered *E. coli* plasmid DNA (pHmC) is calculated using the following equation in conjunction with the data obtained in Figure 4.19.

$$\%hmC = \left[\frac{hmC(AUC) \times 0.95}{dC(AUC) + hmC(AUC) \times 0.95}\right] \times 100$$

Using this approach, the calculated incorporation rate of hmC in plasmid pHmC is 13.8%. Replicate analysis using increased loadings on the HPLC was also performed, resulting in an incorporation of hmC at 13.4%.

4.4 Conclusions

In this Chapter I have developed and applied HPLC-UV in conjunction with LC-MS methods for the analysis of nucleoside modifications in DNA. Two types of reverse phase columns were predominantly used: superficially porous particles (C30) and porous graphitic carbon. Following optimisation of the HPLC approaches, these methods were applied to analyse DNA containing nucleoside modifications in a range of different biological systems, including PCR synthesised dsDNA, phage DNA and *E. coli* engineered to incorporate modified nucleosides.

Initial work focussed on using superficially porous particles in conjunction with a hydrophobic C30 stationary phase (Accucore C30) and a porous graphitic carbon Hypercarb. Results showed that high resolution separations of the nucleosides were obtained with increased retention of dC and dC analogues observed on the Hypercarb column. HPLC-UV was successfully used to verify the presence of the mC, hmC, U and hmU in dsDNA synthesised *in vitro* using PCR in conjunction with nucleotide triphosphate analogues which were prepared for further CRISPR-Cas studies. However, the HPLC-UV was not conclusive in the verification of the presence of ghmC. Unambiguous identification of glucosylation in dsDNA was achieved using LC-MS resulting in the detection of ghmC by virtue of the diagnostic ions (m/z 304.1 and 420.2).

The LC-MS methods were also used to examine DNA modifications of T4 phage and T4 phage mutants. The LC-MS results demonstrated that ghmC was present in all three different types of phage (WT, CR and B), with the relative amount of ghmC lowest in phage B.

In addition, I analysed DNA extracted from *E. coli* engineered to incorporate modified nucleosides to both identify and quantify the abundance of modified nucleosides hmC and ghmC in DNA. The results confirmed the presence of hmC and ghmC in in these plasmid DNA extracted from *E. coli* engineered to incorporate these modifications. The percentage incorporation rate of hmC was further determined by analysing the peak areas of dC/hmdC from the HPLC-UV chromatograms.

In summary, the HPLC-UV in conjunction with LC-MS assays developed in this chapter are proved accurate and reliable, which can be applied to study the modifications in nucleic acids.

Chapter 5: Studying the effect of arginine methylation and RNA methylation on the mRNA interactome

5.1 Abstract

In this study I have optimised *in vivo* mRNP capture assays to study the effects of both protein methylation and RNA methylation on mRNA binding. Large scale messenger ribonucleoprotein (mRNP) capture assays were performed using UV crosslinking to capture RNA-protein interactions in a human embryonic kidney 293T cell line prior to purification of the mRNA using oligo-d(T) and in-gel digestion coupled with mass spectrometry analysis (GeLC-MS). This approach identified approximately 900 proteins, the majority of which are recognized as nucleic acid binding proteins.

Further quantitative proteomic analysis was performed using mRNP capture assays in conjunction with stable isotope labelling with amino acids in cell culture (SILAC) to determine the effects of the global methylation inhibitor adenosine dialdehyde (AdOx) and the effects of RNA methylation (N^6-methyladenosine) on mRNA binding. The results of the mRNP capture assay performed on cells grown with and without N^6-methyladenosine RNA identified approximately 500 proteins, of which approximately 400 were quantified. However, no differences in the mRNA interactome were observed. Further analysis revealed that the RNAi knockdown of the pre-mRNA-splicing regulator gene (WTAP) was not efficient in the cell line used, resulting in no significant alterations of RNA methylation and therefore the mRNA binding was not altered.

The analysis of the effects of the global methylation inhibitor AdOx identified over 600 proteins of which over 500 were quantified in the SILAC label-swap experiment. 24 proteins showed increased mRNA binding and 50 showed decreased mRNA binding (> 1.5 fold change in protein abundance). A wide number of identified proteins contain sites of arginine methylation which are of particular interest. These results are the first global quantitative analysis of the methylation effects on mRNA binding and highlight a number of interesting RNA binding proteins as candidates for further studies to examine the role of arginine methylation and RNA binding.

5.2 Introduction

RNA biology is largely determined by the interplay of RNAs with RNA binding proteins (RBPs) within dynamic ribonucleoproteins (RNPs) (Glisovic et al., 2008). Both the RBP repertoire and RBP activities of cells respond to external stimuli and events within the cell. Many RBPs interact with messenger RNAs (mRNAs) via a limited set of modular RNA-binding domains (RBDs), including the RNA recognition motif (RRM), K-homology domain (KH), zinc fingers (Znf), etc. (Lunde et al., 2007). Based on the analysis of RNA interaction domains, over 600 RBPs in mammalian genomes have been annotated (Müller-Mcnicoll and Neugebauer, 2013).

To study the mRNA-bound proteome, or mRNA interactome, many research groups have used UV crosslinking in conjunction with oligo-d(T) for mRNA-binding protein isolation, taking advantage of the polyadenylation of mRNA (Castello et al., 2012; Baltz et al., 2012). With this approach, hundreds of potential mRNA interacting proteins were identified. With the proteins identified, further bioinformatics analysis of RBDs can be performed to describe the relationships between proteins and RNA in biological processes.

Arginine is an amino acid with a guanidine group, which is protonated and positively charged under physiological environment. The positive charge can be easily delocalized due to the conjugated nitrogen lone pairs and the double bond. As a result, arginine can be altered into different forms when interacting with other molecules, and is also involved in post-translational modifications (PTM). Arginine modifications have been known to affect key biological functions including transcriptional regulation and RNA processing (Pahlich et al., 2006; Paik et al., 2007). Arginine methylation alters its overall hydrophobicity and can cause steric hindrance, which in turns affects protein-protein, as well as protein-nucleic acid interactions (McBride and Silver, 2001; Liu and Dreyfuss, 1995). Specifically, arginine and glycine rich motifs of proteins, often referred to as GAR domains or RGG boxes, are the second most common RBD in human genome (Gerstberger et al., 2014; Rajyaguru and Parker, 2012; Ozdilek et al., 2017). Moreover, GAR domains are also the sites for arginine methylation (Thandapani et al., 2013). However, the binding properties of GAR domains remain poorly understood. The conformational plasticity and adaptability of GAR domains make them capable of targeting a range of different RNAs (Järvelin et al., 2016). For proteins such as HNRNPU, GAR domains are the only identified form of RBD (Kiledjian and Dreyfuss, 1992). GAR domains are found associated with neurodegenerative diseases, such as amyotrophic lateral sclerosis (ALS), fragile X mental retardation syndrome, and cancer (Thandapani et al., 2013).

N^6-mehyladenosine (m^6A) is an abundant type of mRNA modification, with multi-protein complexes involved (Figure 5.1). The modification is reversible, the forward process (methylation) is catalysed by WTAP, METTL3 and METTL14, and its reverse (demethylation) involves FTO and ALKBH5. The modification m^6A often takes place near stop codons and in 3' untranslated regions (UTRs), and is involved in mRNA splicing, degradation and protein expression regulation (Harcourt et al., 2017). The mechanism of m^6A modification is still not fully understood. The addition of the methyl group has a structural effect blocking base pairing, and the base stacking effect is enhanced (Roost et al., 2015). RNA with m^6A are known to bind to YTH domain proteins (such as YTHDC1 and YTHDF2) with high affinity, due to the aromatic hydrophobic pockets in the YTH domains (Dominissini et al., 2012). In the nucleus of cells, the m^6A modification can act as a protein binding switch by altering the RNA structure. As shown in Figure 5.1, the m^6A modification facilitates the reorganisation and binding of heterogeneous nuclear ribonucleoprotein C (HNRNPC, an abundant nuclear protein mediating alternative pre-mRNA splicing) to RNA (Liu et al., 2015). In the cytoplasm, several proteins (e.g., YTHDF1, YTHDF2 and eIF3) can recognize and bind to the m^6A for different purposes, including the regulation of RNA translation and stability (Wang et al., 2015).

In this study, an m^6A deficient HEK-293T cell line (with WTAP gene knockdown) was used to study the effect of mRNA m^6A modification on mRNA-protein interactions.

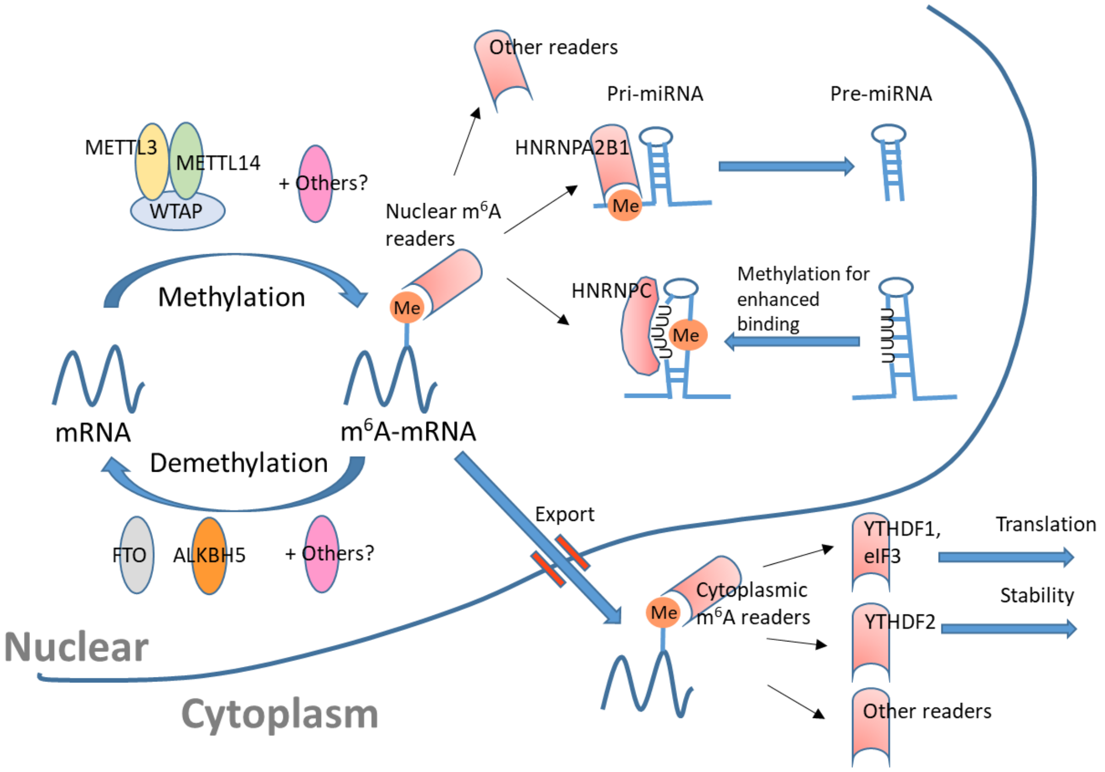


Figure 5.1 | m^{6}A modification of mRNA and its functions. For details see text.

For the alteration of protein arginine methylation in this work, a HEK-293T cell line was treated with adenosine-2', 3'- dialdehyde (also known as Adenosine periodate oxidized, or AdOx). AdOx is a global indirect methyltransferase inhibitor widely used in cell culture for *in vivo* and *in vitro* protein methylation studies. When AdOx is present, activity of S-adenosyl-L-homocysteine hydrolase is inhibited (Hoffman, 1979), which leads to increased amounts of S-adenosyl-L-homocysteine (AdoHcy) (Bartel and Borchardt, 1984). AdoHcy acts as a product inhibitor of those methyltransferases (such as protein arginine methyltransferases, or PRMTs) which use S-adenosyl-L-methionine (AdoMet) as the methyl donor (Johnson et al., 1993) (Figure 5.2). In cell culture, the addition of AdOx keeps proteins and nucleic acids in a hypomethylated state (Chen et al., 2004).

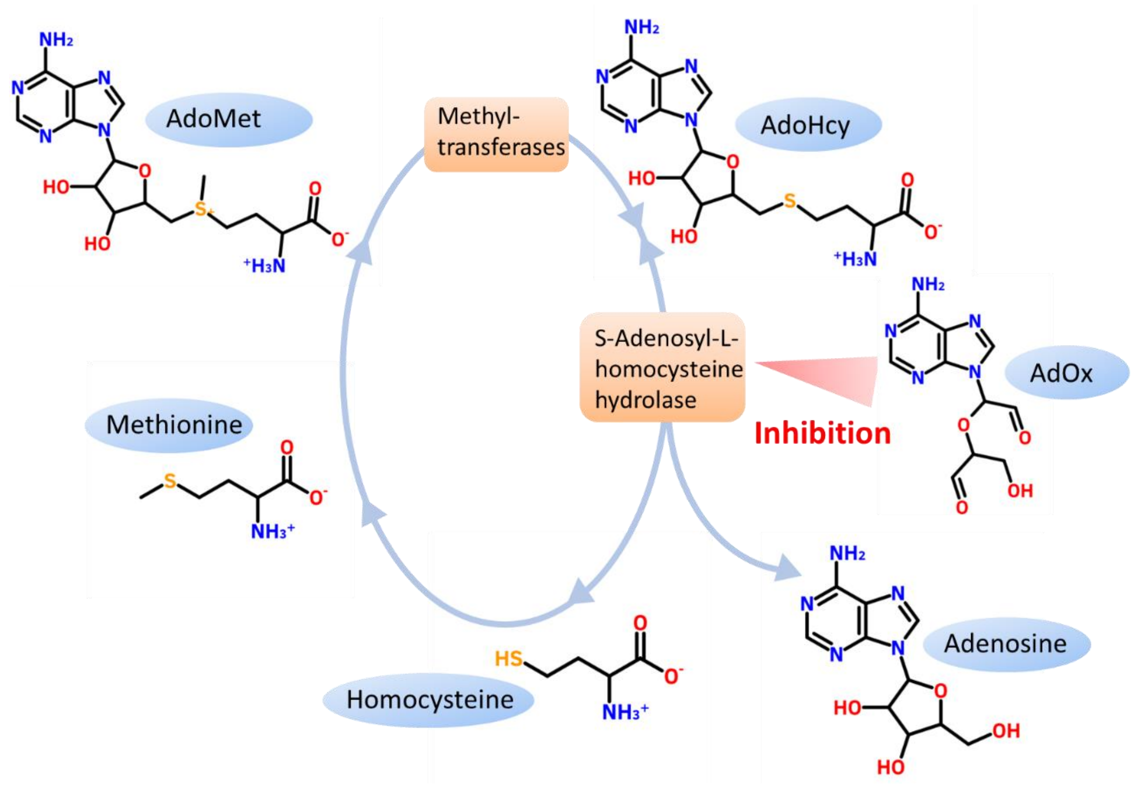


Figure 5.2 | Metabolism of adenosine, showing the role of AdOx as inhibitor of S-adenosyl-L-homocysteine hydrolase. Since the conversion of S-adenosyl-L-homocysteine (AdoHcy) to adenosine and homocysteine is restricted, the accumulation of AdoHcy, which is a general S-adenosylmethionine (AdoMet) based methyltransferase inhibitor, leads to decreased methylation of products.

Although studies have been carried out on how different modifications affect mRNA and protein interaction, few are focused on the effect of arginine methylation/ citrullination or mRNA with N^6-methyladenosine (m^{6}A) on the global scale mRNA-bound proteome. The aim of this work is to use a SILAC labelling strategy in conjunction with UV crosslinking, oligo-d(T) enrichments of the mRNA bound proteome in conjunction with mass spectrometry to quantify as much mRNA-binding proteins as possible. More importantly, this work aims to identify those mRNA-binding proteins of which the binding activities are significant affected by different modification status. These proteins may potentially have key functions in certain biology pathways, and the knowledge can help in understanding the mechanisms involved in, for example, a certain disease.

It is proposed that protein post-translational modifications and chemical modifications of RNA affect RNA-protein interactions. Therefore the work designed in this Chapter aims to study the effect(s) of protein methylation and RNA methylation (N^6-methyladenosine) on the mRNA interactome using mass spectrometry methods in conjunction with SILAC.

5.3 Results and Discussion

5.3.1 Development and optimisation of the mRNP capture assay in conjunction with mass spectrometry analysis

In order to study the effects of chemical modifications on the mRNA interactome, mRNP capture assays were performed on a human embryonic kidney 293T cell line (HEK 293T) in this study. For *in vivo* mRNA capture assays, a UV crosslinking strategy was used. During the crosslinking process, irreversible covalent bonds can form between nucleic acids and proteins in close proximity (Chodosh, 2001). After the crosslinking, the cells were lysed and mRNA and binding proteins were isolated using oligo-(dT) beads to base-pair with the mRNA poly (A) tails. The beads were then washed under denaturing conditions and the unbound proteins were removed. Finally the mRNA-binding proteins were eluted after the digestion of mRNA using RNase (see Figure 5.3).

Generally there are two ways to do the UV crosslinking. For regular crosslinking and immunoprecipitation (CLIP), cells are crosslinked under short-wavelength UV light (254 nm) (Figure 5.3A left panel). While for an improved mRNP capture assay termed as photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) (Hafner et al., 2010), a photo-reactive nucleoside analogue, such as 4-thiouridine (4SU) or 6-thioguanosine (6SG), was added to media and metabolically incorporated by cells into their RNA during culturing. Living cells are then crosslinked at a longer wavelength (UV 365 nm) with increased efficiency (Figure 5.3A right panel). 4SU is predicted to crosslink with aromatic amino acids, and the predicted site for crosslinking is shown in Figure 5.3B.

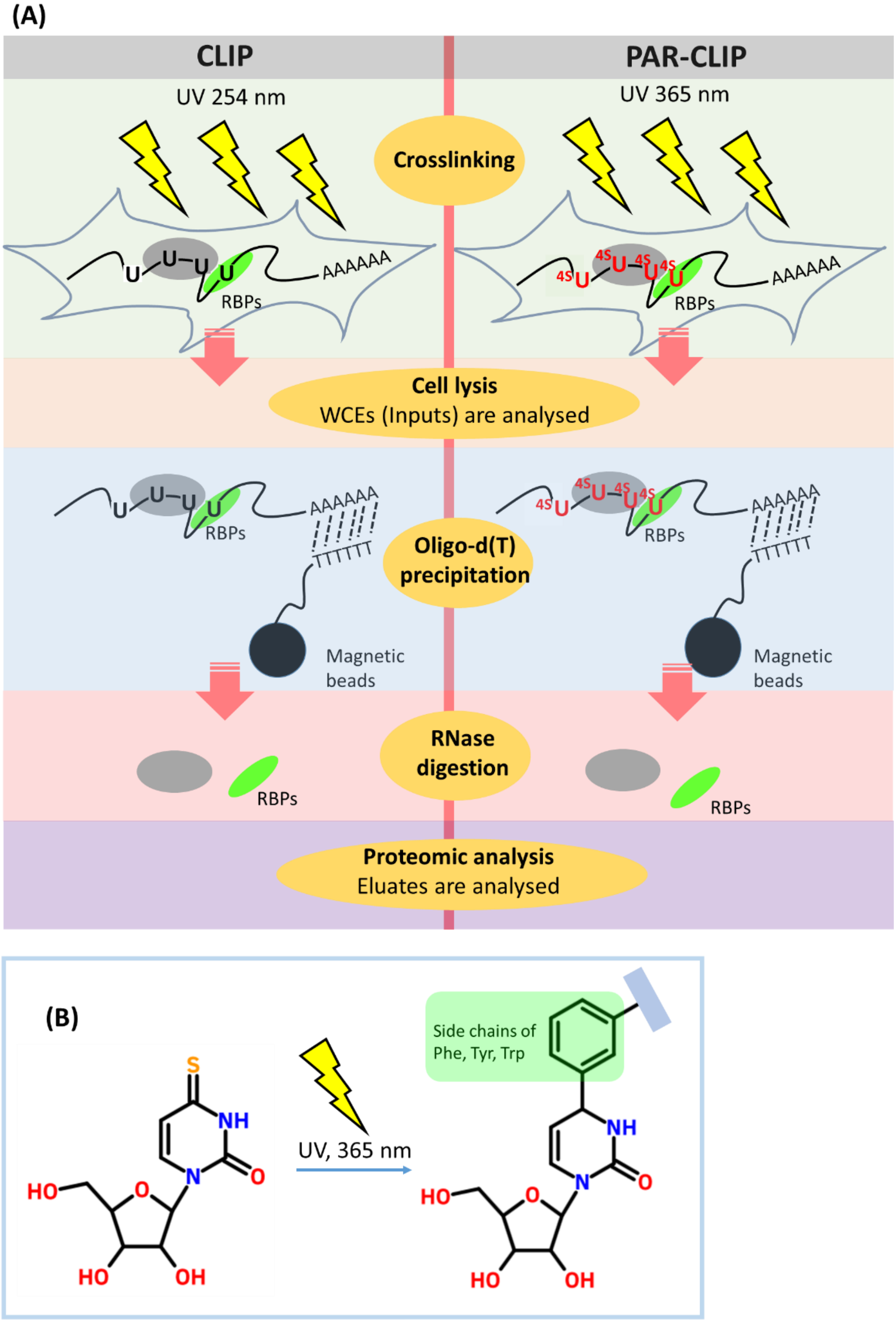


Figure 5.3 | Schematic of mRNP capture assay using CLIP and PAR-CLIP. (A) A comparison between the experimental setups of CLIP and PAR-CLIP. WCE: whole cell extract. **(B)** Predicted site for the crosslinking of 4SU to an aromatic amino acid side chain when irradiated at UV 365 nm.

5.3.2 Optimization of crosslinking efficiency

Initial work focussed on optimising the mRNP capture assay in conjunction with MS analysis to identify the mRNA interactome. The chemistries of crosslinking using CLIP and PAR-CLIP are distinct (Wetzel and Soöll, 1977; Greenberg, 1979). Under PAR-CLIP conditions, only 4SU containing mRNA can crosslink to proteins (Ascano et al., 2012). In order to compare the crosslinking efficiency, HEK 293T cells were cultured with or without 4SU treatment. For 4SU treated cells, crosslinking was performed at UV 365 nm, 0.2 J/cm^2, or at a mix of UV 365 nm and UV 254 nm at the same time (using two different UV light sources), 0.25 J/cm^2. For non-4SU treated cells, crosslinking was performed at UV 254nm, 0.3 J/cm^2. Roughly 10 mg of protein lysate were used for the precipitation of mRNA-binding proteins using oligo-d(T) beads. Both protein from input lysates and isolated mRNA-binding protein in eluates were resolved by SDS-PAGE (see Figure 5.3A). The results (Figure 5.4) show the use of 4SU does not increase the recovery of proteins. Instead, cells grown in normal conditions with irradiation at UV 254 nm show better mRNP capture yield in this work. Interestingly, cells treated with 4SU but crosslinked at a mixed source of UV 254 nm and UV 365 nm show a good amount of mRNA-binding proteins captured.

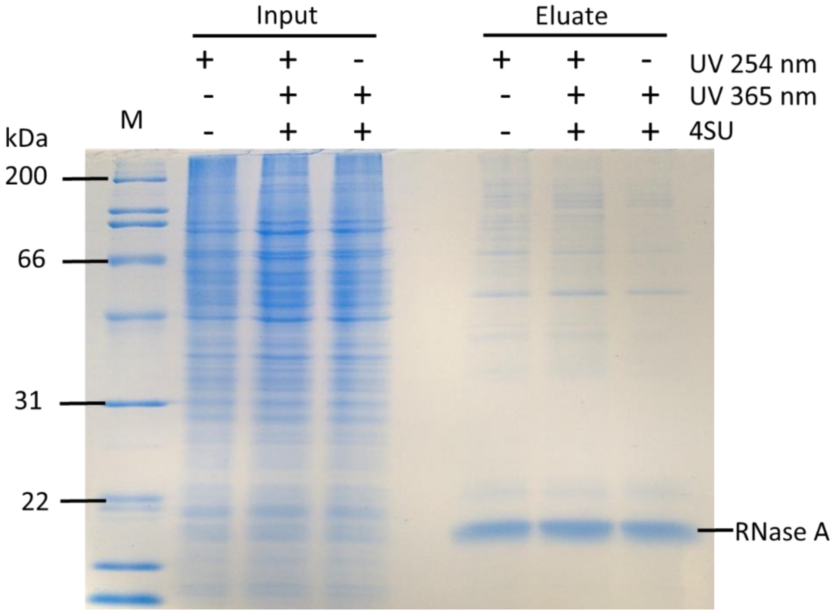


Figure 5.4 | Comparison of mRNP capture assay using CLIP and PAR-CLIP. Non 4SU treated cells were crosslinked at 254 nm, 0.3 J/cm^2, while 4SU treated cells were crosslinked at either a mixed UV source (2 × 254 nm bulbs and 3 × 365 nm bulbs), 0.25 J/cm^2, or at 365 nm only, 0.2 J/cm^2. Proteins were visualised by Coomassie blue staining. For inputs, equal amount of protein (based on Bio-Rad assay) were loaded. M: protein marker.

To further examine the mRNP interactome in the eluates of different crosslinking conditions, GeLC-MS analysis was used in conjunction with the Q Exactive HF mass spectrometer (see Chapter 2.17). For PAR-CLIP eluate (4SU treated cells crosslinked at UV 365 nm, 0.2 J/cm^2), 691 protein groups (proteins that cannot be identified by unique peptides but with shared peptides are grouped as one group) were identified, compared to 909 protein groups identified in regular CLIP (UV 254nm, 0.3 J/cm^2) (Figure 5.5A). Gene Ontology (GO) annotations of these protein groups identified show similarity based on molecular function and protein class categorisation (Figure 5.5, B and C), indicating that regardless of the crosslinking chemistries, similar mRNA-protein binding profiles are seen using either CLIP or PAR-CLIP. However, regular CLIP did show better crosslinking outcome in this work, which is in contradiction to the claim that PAR-CLIP can significantly increase mRNA-protein binding activity. This may be due to the cell lines used in this work cannot incorporate 4SU effectively. Since with long-wavelength UV (above 310 nm) natural nucleotides do not crosslink as efficiently, the amount of 4SU containing mRNAs were limited for protein binding. This also explains why 4SU treated cells crosslinked under both short-wavelength and long-wavelength UV resulted in increased protein crosslinking. Another possibility for reduced crosslink efficiency is the reduced energy for PAR-CLIP crosslinking (0.2 J/cm^2) compared to non-4SU crosslinking at 0.3 J/cm^2, or mixed UV wavelength crosslinking at 0.25 J/cm^2. However, according to PAR-CLIP described in other groups work, a low crosslinking energy (either 0.2 J/cm^2 or 0.15 J/cm^2) was suggested for best results (Spitzer et al., 2014; Baltz et al., 2012). Indeed, PAR-CLIP was also carried out in this experiment with crosslinking energy of 0.3 J/cm^2, no significant increase of protein amount was observed based on SDS-PAGE (data not shown). In summary, since PAR-CLIP resulted in decreased crosslinking efficiency, normal CLIP strategy was used for crosslinking in the following experiments.

Further validation of the mRNP interactome generated in the above experiments was performed by comparison with recent studies by Baltz et al., 2012. They detected 797 proteins using PAR-CLIP and oligo-(dT) precipitation with HEK 293T cells. These proteins were identified in 3 biological replicates thus with good confidence. Here a combined search of four mRNP capture replicates in this work identified over 900 protein groups (protein groups were counted if identified in at least one assay). These protein groups were compared with Baltz et al., 2012, and results are shown in Figure 5.6. The results show approximately 65% of proteins identified by Baltz's group were also identified in this work. Moreover, from GO annotation analysis (Figure 5.6B and C), the type of proteins identified are consistent. Specifically, for protein categorisation based on molecular function, the two most abundant categories are binding and catalytic activity; for protein class classification, nucleic acid binding proteins are the most abundant type, which are as expected. In general, more proteins were identified with the mRNP capture assay in this work compared to Baltz's group. This may partly due to the more stringent criteria they used. In summary the results show that mRNP capture was optimised using standard UV crosslinking resulting in the identification of over 1000 proteins in the mRNA interactome. The resulting GO ontology analysis revealed nucleic acid binding was the most abundant category, consistent with previous mRNA interactome studies (Baltz et al., 2012).

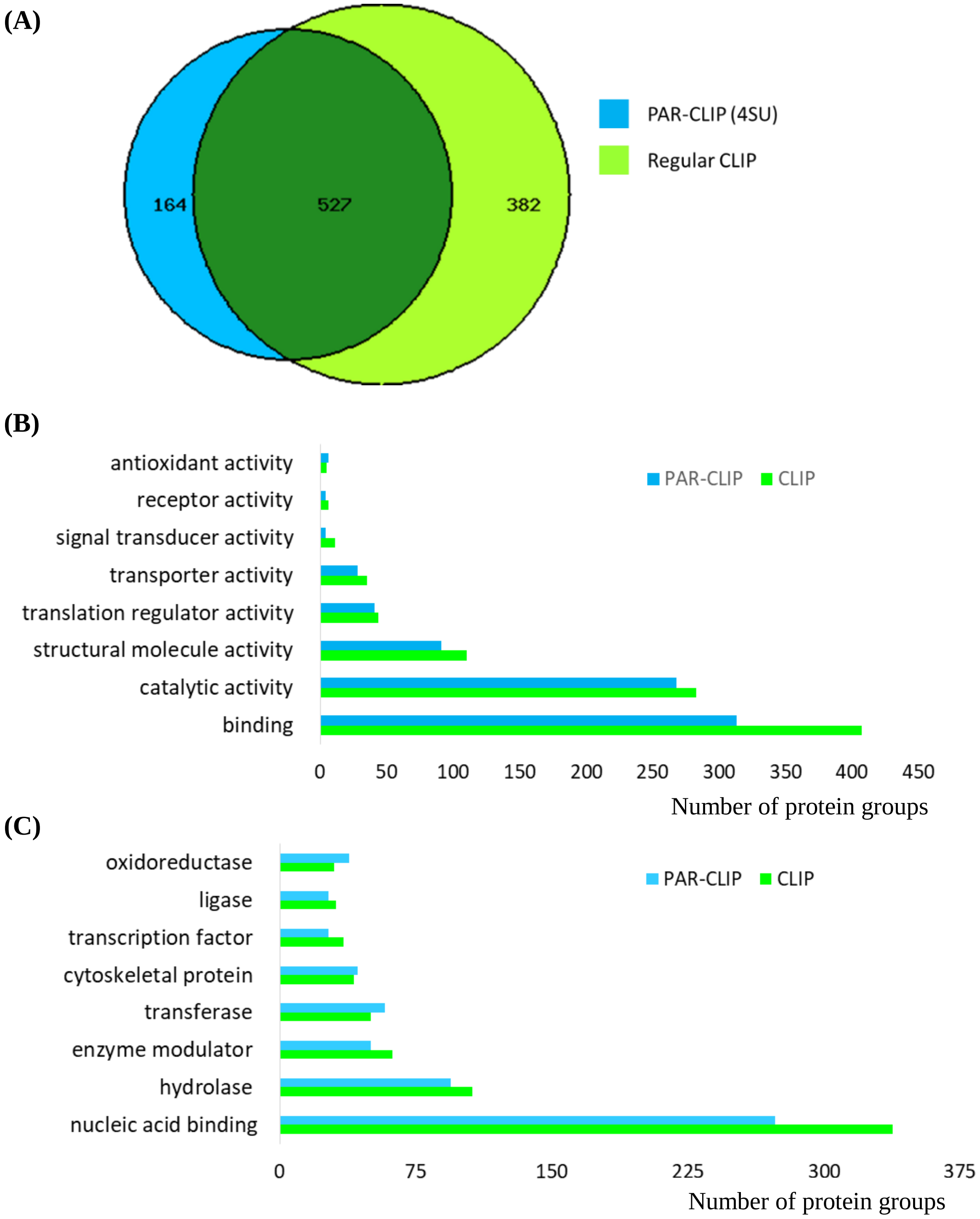


Figure 5.5 | Comparison of CLIP and PAR-CLIP based on mass spectrometry identification results. **(A)** Venn diagram comparing the number of protein groups identified using CLIP and PAR-CLIP. **(B)** Distribution of protein groups identified by the two crosslinking methods based on GO molecular function classification. **(C)** Categorisation (by class) of protein identified by the two methods. GO annotations were performed using the PANTHER classification system.

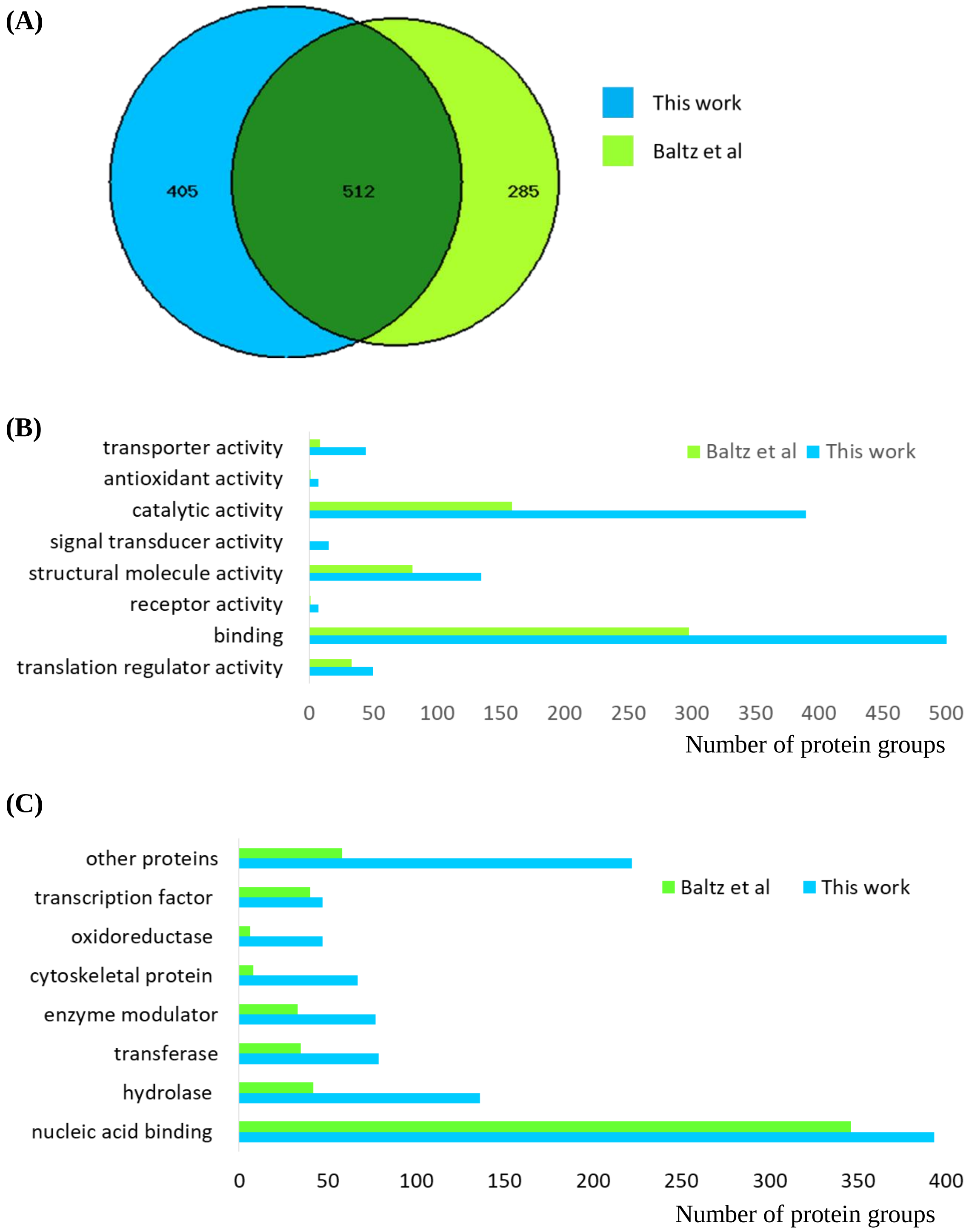


Figure 5.6 | Cross reference of protein groups identified in this work to Baltz et al., 2012. (A) Venn diagram comparing the number of proteins or protein groups identified in this work and in the work of Baltz's group. **(B)** GO molecular function distribution of protein or protein groups identified in this work and from Baltz's group. **(C)** Categorisation (by class) of protein identified in this work and from Baltz's group. GO annotations were performed using the PANTHER classification system.

5.3.3 Optimisation of SILAC for quantitative proteomic analysis of the mRNA interactome

Following optimisation and validation of the mRNA capture assay, quantitative MS analysis of the mRNA interactome is required to demonstrate that protein binding to mRNA is perturbed under different conditions. To perform quantitative proteomic analysis, stable isotope labelling by amino acids in cell culture (SILAC) was used. Initial studies were performed using HEK 293T cells in conjunction with SILAC labelling to determine the labelling efficiency, as well as to optimise the cell growth in SILAC media.

First, the incorporation rate of heavy isotope was assessed. In this work, heavy labelled arginine ($^{13}C_6$ $^{15}N_4$ L-Arginine-HCl, or Arg10) and lysine ($^{13}C_6$ $^{15}N_2$ L-Lysine-2HCl, or Lys8) were used. Initial SILAC tests were performed in heavy medium only using HEK 293T cells. Cells were cultured for 4 passages, harvested on day 14, then lysed and fractionated by SDS-PAGE. A single fraction (gel band) was selected and in-gel digested, then analysed using MS (maXis). To assess the labelling efficiency, two peptide sequences (containing either lysine or arginine) were examined (Figure 5.7). For both peptide sequences MS data show low levels of light peptide signals (similar to noise level), compared to the isotope cluster corresponding to heavy peptides. Incorporation rate is calculated using equation:

$$Incorporation\ (\%) = \frac{ratio\ (\frac{H}{L})}{ratio\ \left(\frac{H}{L}\right) + 1} \times 100$$

For lysine containing peptide ETVSEESNVLCLSK, the incorporation rate is 97.3%, and for arginine containing peptide DAGTIAGLNVLR, the corporation rate is 97.7%. With this approach, the overall isotope incorporation rate (average) was checked for all SILAC experiments performed in this work.

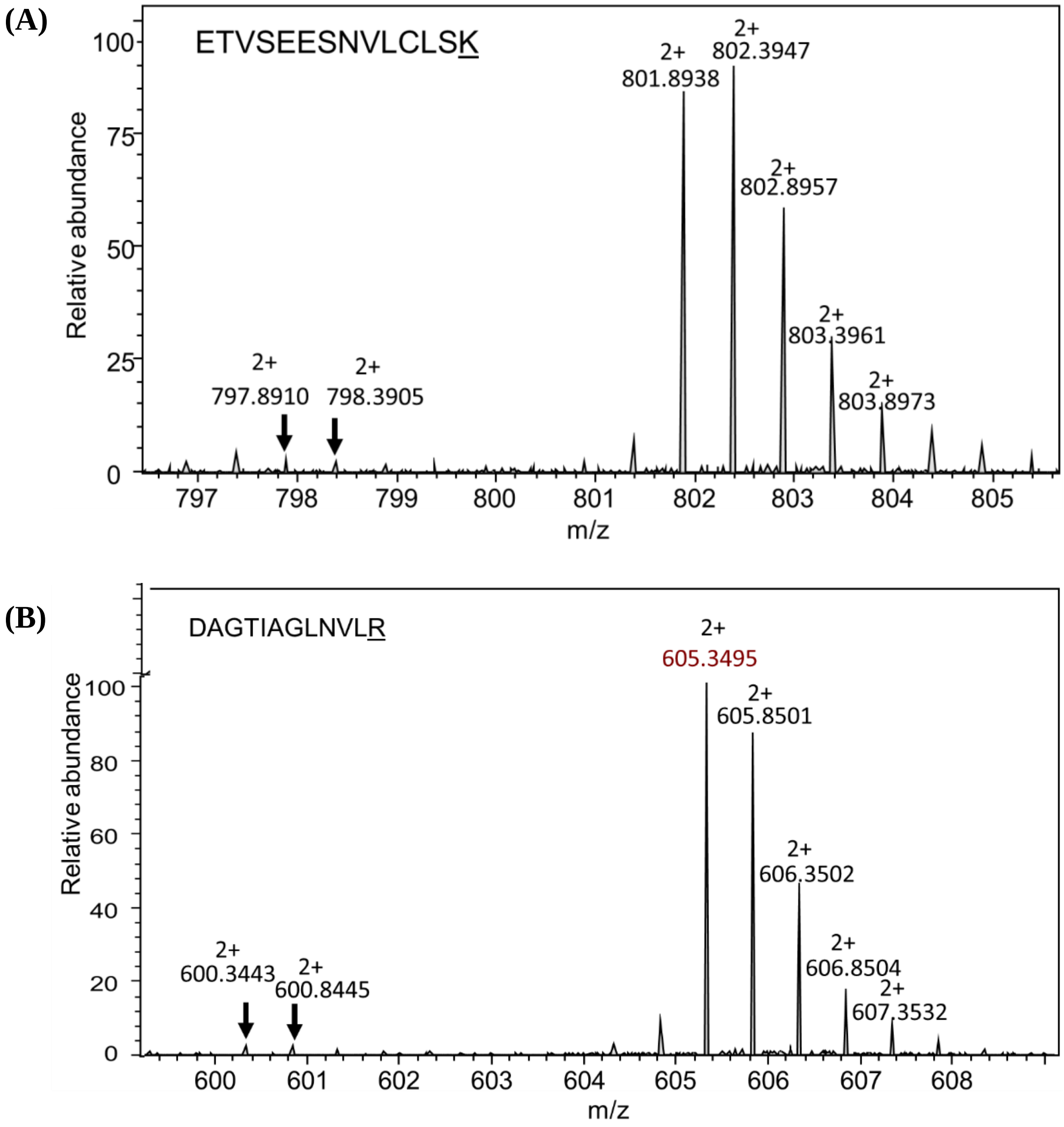


Figure 5.7 | Incorporation of ^{10}R and ^{8}K in heavy medium. MS data of the tryptic lysine containing peptide **(A)** ETVSEESNVLCLSK and arginine containing peptide **(B)** DAGTIAGLNVLR from HEK 293T cells cultured using heavy labelled SILAC medium (^{8}K^{10}R) reveal limited signals corresponding to the light version of peptides.

Despite being a widely used mass spectrometry-based quantitative method for proteomic studies, the accuracy of SILAC can be compromised due to arginine to proline metabolic conversion, where arginine is used as a precursor for proline synthesis (Bendall et al., 2008) (Figure 5.8A). The formation of heavy proline (Pro6) can reduce the signal intensity of proline containing heavy peptide at the expected m/z (Figure 5.8B, green), the proportion of which shifts to higher m/z corresponding to the heavy peptide plus proline 6 (Figure 5.8B, red). This results in an underestimation when quantifying proline containing peptides based on relative abundance.

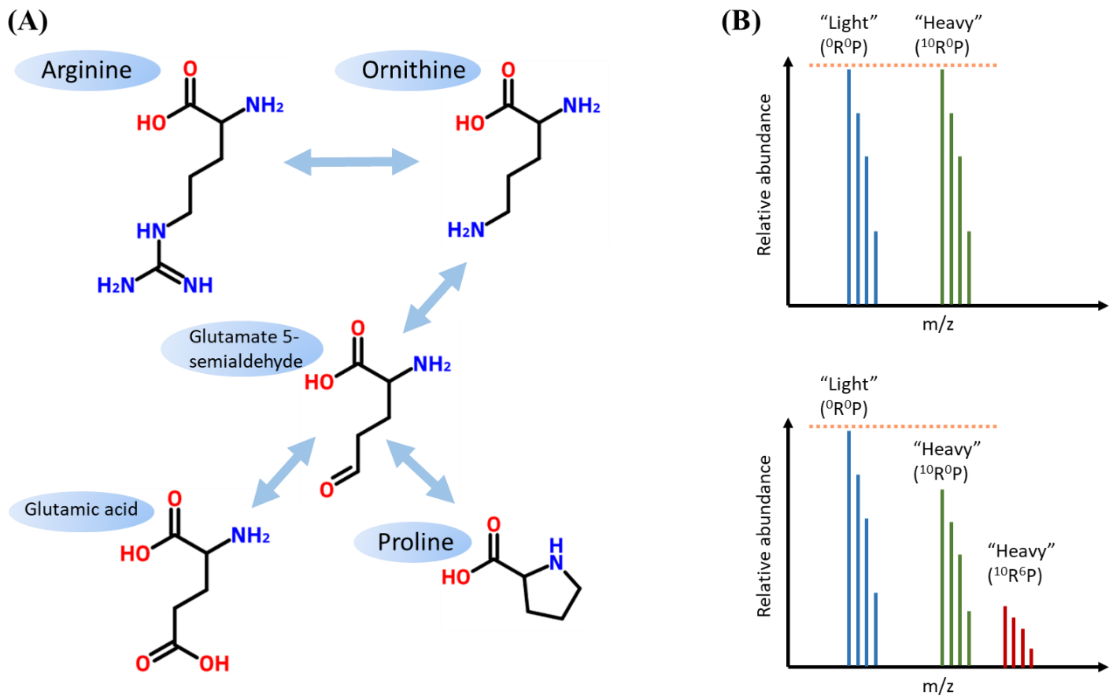


Figure 5.8 | The arginine to proline conversion issue for SILAC quantification. (A) Metabolic conversion of arginine to proline in eukaryotes. **(B)** Conceptual mass spectra show a non-proline containing peptide (upper panel) and a proline containing peptide (lower panel). The signal intensities of heavy and light peptide are expected to be equivalent, however, the conversion to heavy proline split the expected heavy peptide signal (green) into two (green and red).

Following initial optimisation, the MS analysis revealed that arginine to proline conversions were present in proteins extracted from HEK 293T cells cultured in heavy SILAC labelled media ($^{8}K^{10}R$). In Figure 5.9, two tryptic peptides (containing either a proline and an arginine, Figure 5.9A, or a proline and a lysine, Figure 5.9B) are shown as examples. The MS spectra show clear evidence of the additional isotope clusters corresponding to heavy proline. The proline conversion rate is 36.9% for peptide GVVDSEDLPLNISR, and 38.3% for peptide TWNDPSVQQDIK.

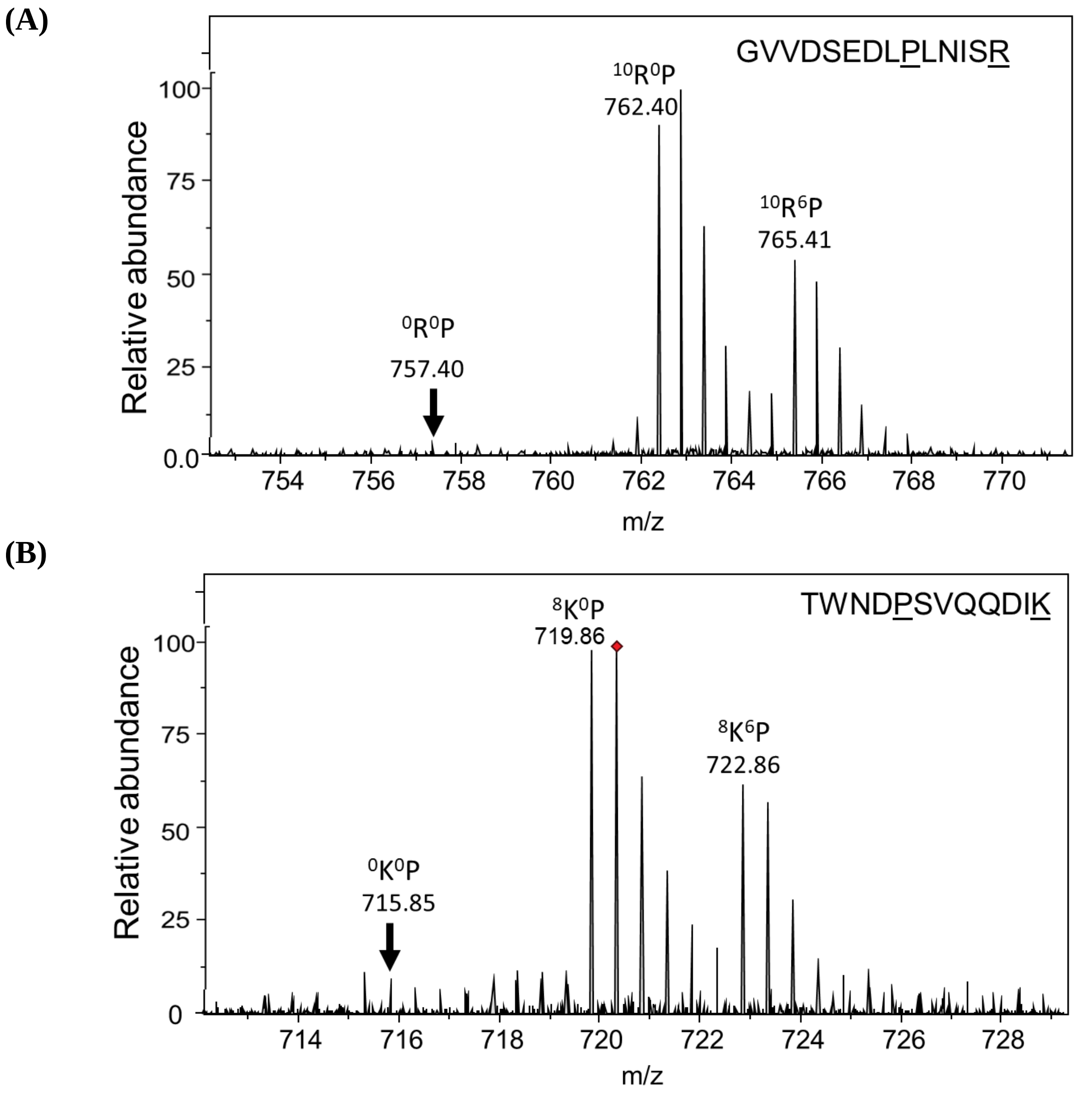


Figure 5.9 | MS spectra showing the arginine to proline conversion. Tryptic proline-containing peptides **(A)** GVVDSEDLPLNISR and **(B)** TWNDPSVQQDIK from HEK 293T cells grown on heavy labelled SILAC media (^{10}R^{8}K) reveal clear signals corresponding to heavy proline, indicating the existence of arginine-proline conversion.

This arginine to proline conversion issue can be solved in a post-quantification correction approach, for example, by normalising all proline-containing peptides, either manually or using dedicated algorithms (Gruhler et al., 2005), or by using ^{4}R in light media as an experimental correction (Van Hoof et al., 2007). Alternatively, the arginine to proline conversion can be prevented or reduced in the first place. A straightforward way to do this is to limit the amount of arginine in the media. However, arginine deficiency can affect cell behaviour and growth (Ong et al., 2003; Scott et al., 2000). Another commonly used strategy is to add unlabelled proline in both heavy and light media (Lößner et al., 2011). To use this approach, the amount of additional proline needs be carefully determined experimentally, as excessive amount of proline can lead to the conversion of proline to arginine in a reverse metabolic procedure.

To overcome the arginine to proline conversion issue in this study, additional unlabelled proline was added to the SILAC media. A final concentration of 1.7 mM proline was determined, and cells were cultured and processed following the same procedure as previously described**,** prior to analysis by MS (Q Exactive HF). Two proline containing peptides with either arginine or lysine are shown as examples in Figure 5.10. Compared to non-proline treated cells, the intensities of heavy labelled proline isotope peaks are significantly reduced. The proline conversion rate for peptide DNPGVVTCLDEAR is approximately 8.0%, and for peptide NPDDITNEEYGEFYK is below 7.1%. Although the conversion was not fully prevented, a significant reduction was achieved. In conjunction with post-quantification normalization approaches and SILAC label-swap strategies, the conversion issue that could affect downstream protein quantitation can be prevented.

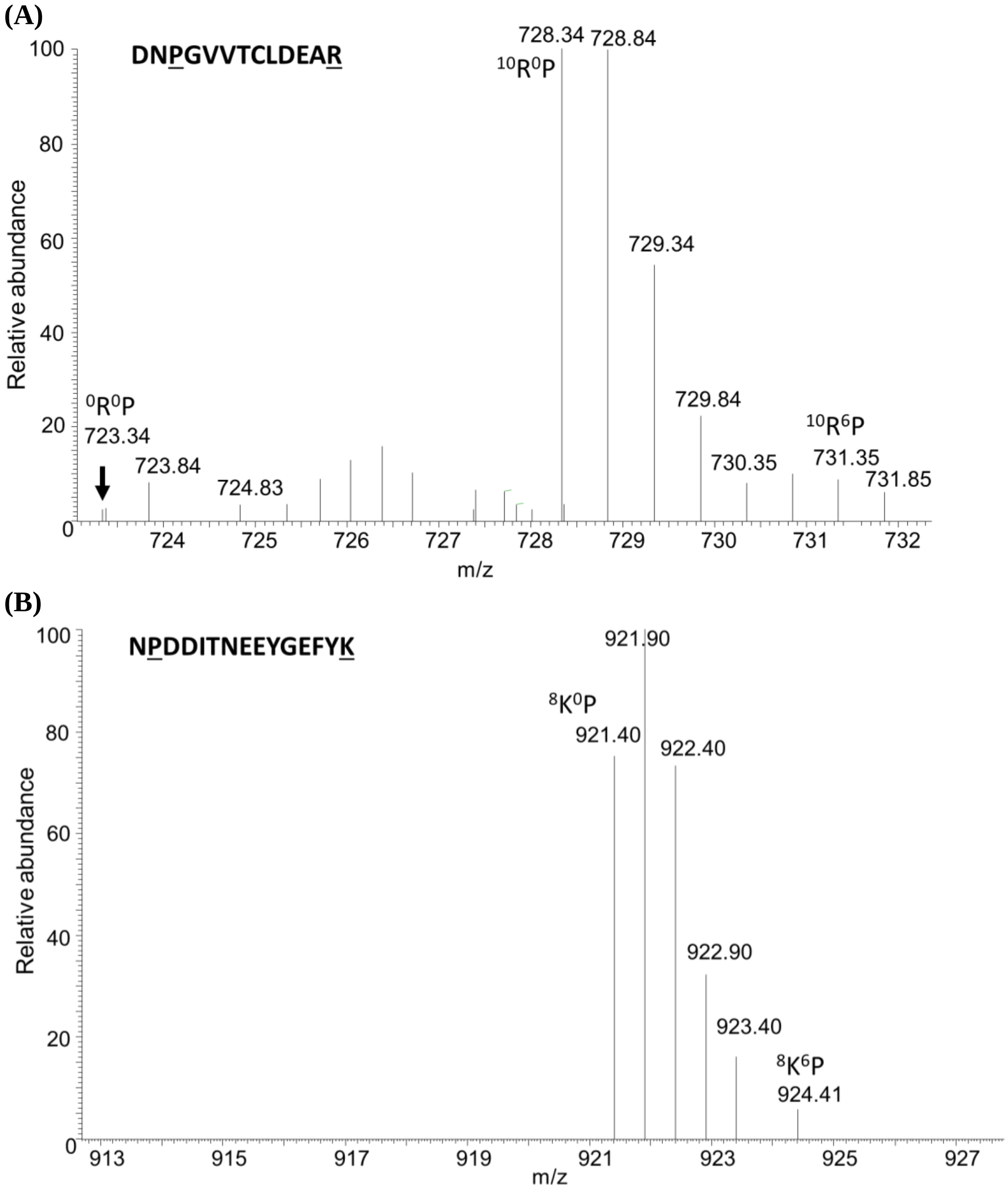


Figure 5.10 | MS spectra showing the effect of using additional proline during cell culture on the arginine to proline conversion. Tryptic proline-containing peptides **(A)** DNPGVVTCLDEAR and **(B)** NPDDITNEEYGEFYK grown on heavy labelled SILAC media (^{10}R^{8}K) reveal significant reduction of signals corresponding to the heavy proline.

In this study, I chose to use SILAC experiments in conjunction with label-swap strategy for the biological replicates. This strategy can effectively determine false positives (where changes are found in only one labelling of data), and more importantly, attenuate the inaccuracies caused by common issues associated with SILAC experiments including incomplete isotope incorporation and arginine-proline conversion, by averaging the ratios of individual replicate quantification data (Park et al., 2012). As a result, more reliable quantification of protein expression ratios can be achieved. An example of typical label-swap SILAC data set interpretation is shown in Figure 5.11.

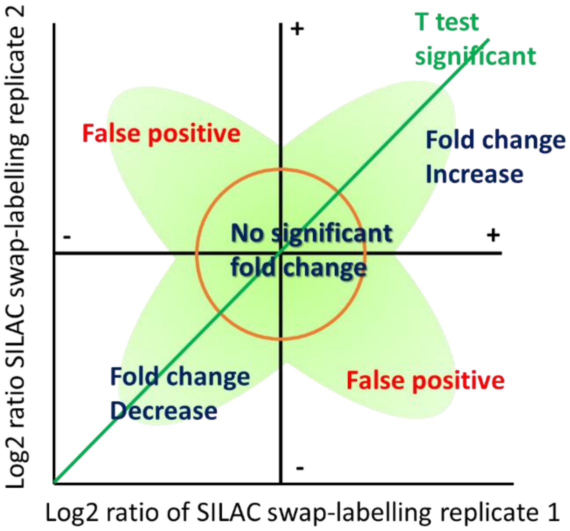

T test significant
False positive
Fold change Increase
No significant fold change
Fold change Decrease
False positive
+
-
+
-
Log2 ratio SILAC swap-labelling replicate 2
Log2 ratio of SILAC swap-labelling replicate 1

Figure 5.11 | Schematic of SILAC label-swap data set interpretation.

5.3.4 Studying the effect of methylation on the mRNA interactome

To study the effect of global methylation on mRNA-protein interactions, cells were grown in the presence and absence of the global methylation inhibitor, AdOx, in conjunction with mRNP capture assays and SILAC MS analysis. To limit the issue of arginine to proline conversion, all cells were cultured in SILAC media (both heavy and light) with additional L-proline (1.7 mM). As for crosslinking, since no improvement of efficiency was observed with the 4SU treatment to cells, normal CILP was used.

Initial work focussed on verifying of protein demethylation with the addition of AdOx. HEK 293T cells were treated with 20 µM AdOx for 48 hours for methylation inhibition. Protein extracts were then analysed by SDS PAGE and western blot using a CHTOP antibody. CHTOP is a protein with known arginine methylation (Chang et al., 2013). The results (Figure 5.12) show the existence of a hypo-methylated CHTOP band for AdOx treated cell extract, confirming methylation inhibition.

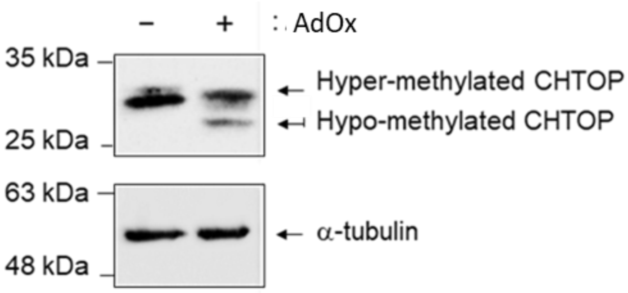


Figure 5.12 | Western blotting of AdOx treated cells and control cells. Whole cell extract (WCE) of both cell lines were analysed with CHTOP antibody. α-tubulin was shown as input control.

After the confirmation of cell methylation alteration with AdOx treatment, mRNP capture assays were performed to study the effect of methylation on the global mRNA interactome. Cells were cultured in SILAC heavy and light media with label-swap for biological replicates (using the optimised SILAC growths as previously described). The mRNP capture assays were performed as previously optimised using equal amounts of heavy and light cell lysates (10 mg each) mixed together prior to oligo-d(T) purification. Three consecutive pulldowns were performed using the same lysates. The reason for doing this is that multiple pulldowns not only increase the mRNP recover rate, but also essential to capture certain proteins such as crosslinked AGO (Baltz et al., 2012). After, three pulldowns, eluates were pooled together prior to analysis using SDS-PAGE. The results are shown in Figure 5.13.

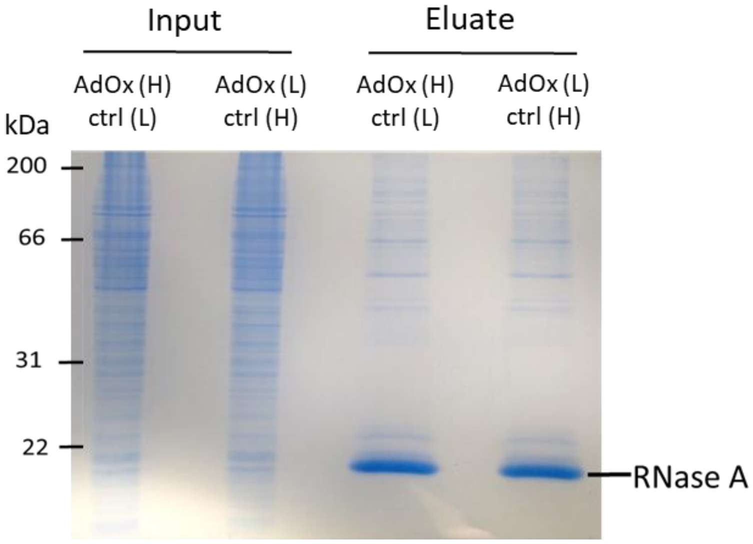
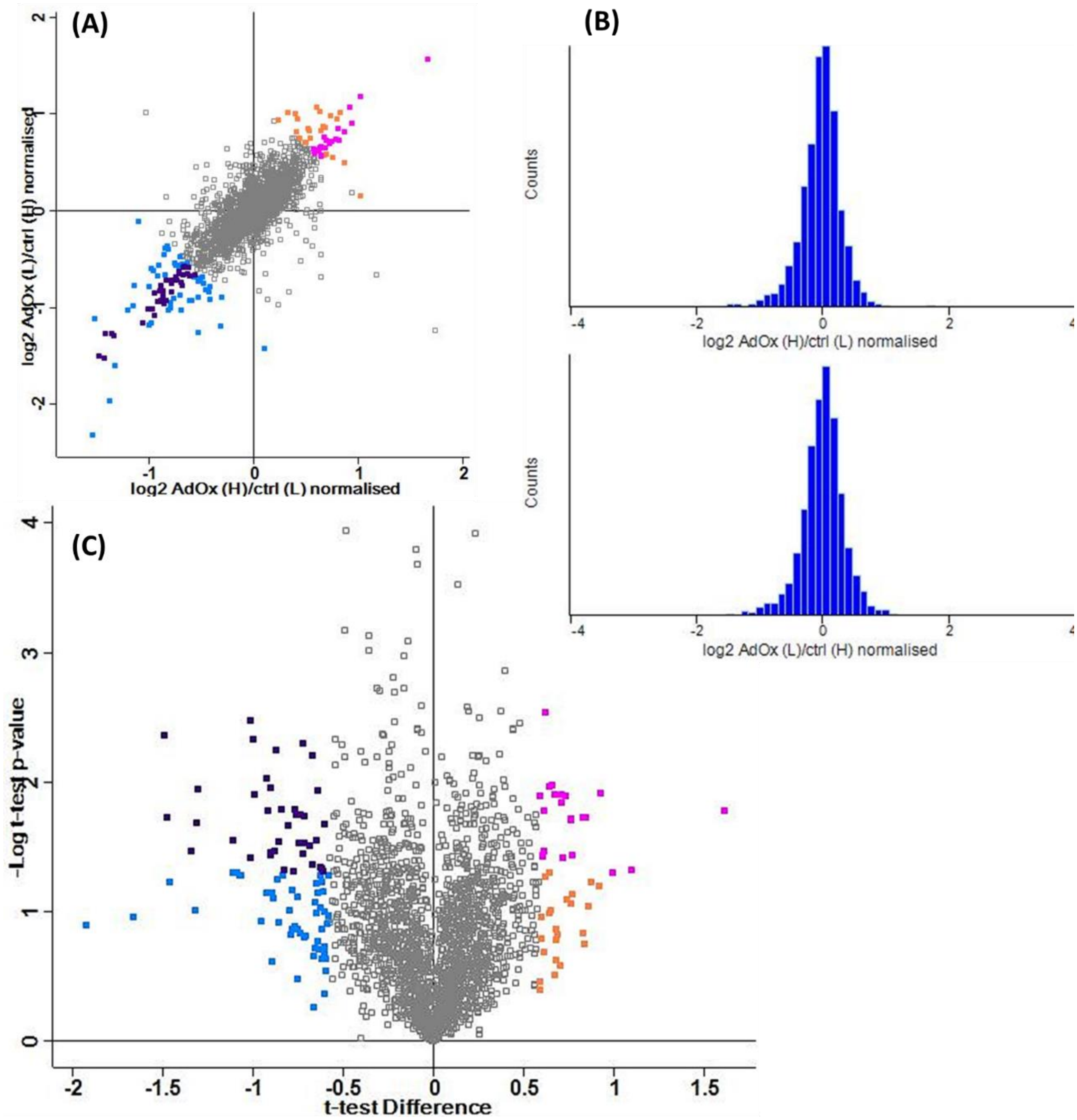


Figure 5.13 | SDS-PAGE analysis of inputs (WCE) and eluates (mRNP capture assay) from AdOx treated cell lines (AdOx) and control cell lines (ctrl). For both inputs and eluates, equal amount of heavy and light proteins were mixed prior to gel analysis. Eluates from three consecutive pulldowns were pooled together as one.

Quantitative analysis of the whole cell protein expression

Following SDS-PAGE analysis, input proteins were in-gel digested and analysed by LC-MS (Q Exactive HF). It is important to analyse and compare the proteomes from cells grown in the presence and absence of AdOx as this is likely to affect protein expression over a range of different proteins. For MS data analysis, raw data were processed using MaxQuant (Cox and Mann, 2008), and statistical analysis was performed in Perseus (perseus-framework.org). Results show close to 3000 protein groups (see Section 5.3.2 for protein group definition) were identified in each replicate of input, of which over 80% were quantified (Figure 5.14). The $\log_2$ fold change values of protein groups distribute evenly around 0 (Figure 5.14B), indicating good consistency of the label-swap replicates. Quantitative proteomic analysis revealed that 46 protein groups were identified with over 1.5 fold increase ($\log_2$ value > 0.585) on expression level, of which 22 are t-test significant (p value 0.05). On the other hand, the expression of 91 protein groups were decreased over 1.5 fold ($\log_2$ value < -0.585), of which 39 are t-test significant (Figure 5.14A and C). These results demonstrate the successful application of the SILAC workflow to identify a number of differentially expressed proteins upon the addition of the methylation inhibitor AdOx. These results will be used in conjunction with quantitative mRNP capture assays described in the following section.

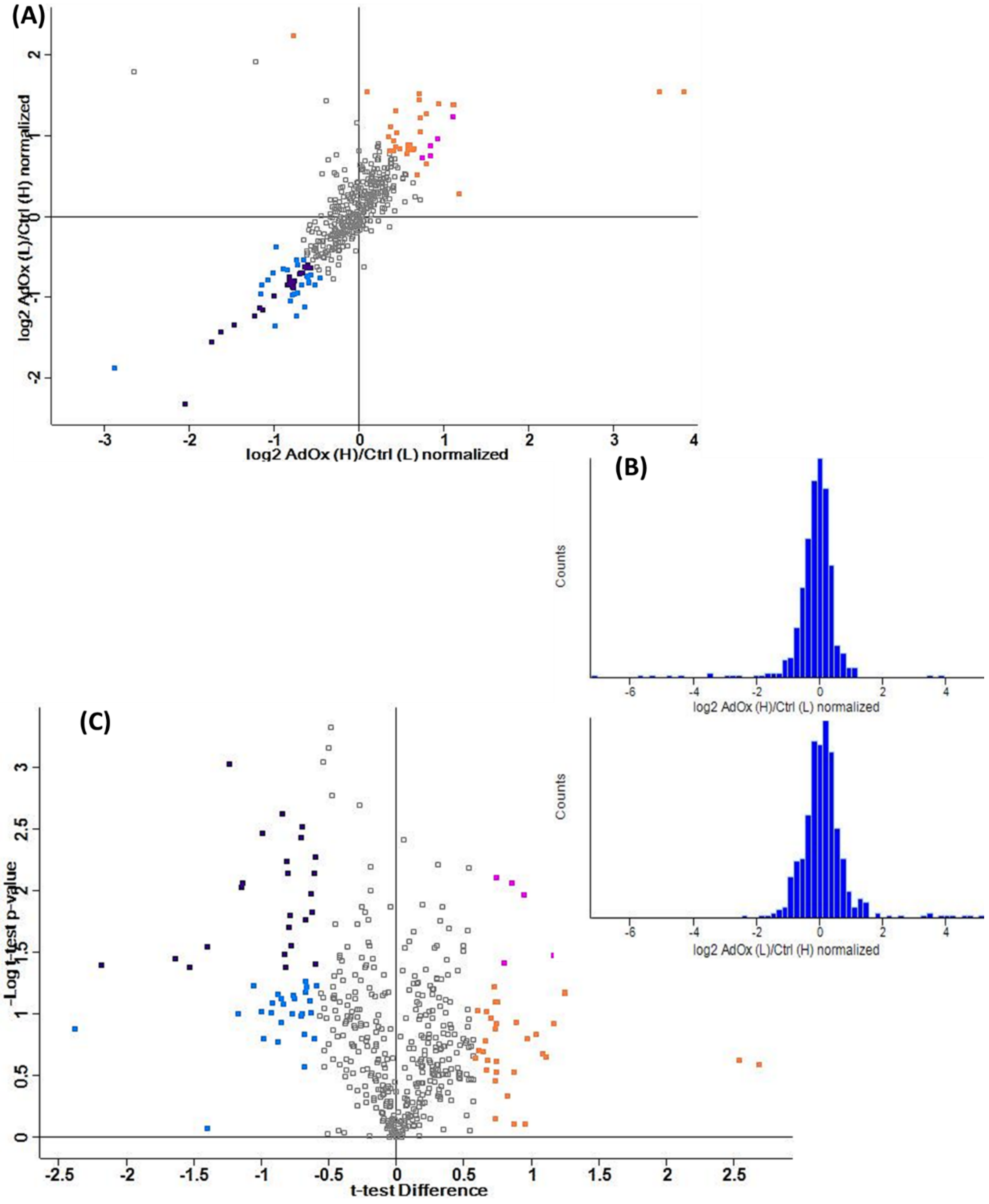


Experiment	Identified	Quantified
AdOx (H), ctrl (L)	2986	2483
AdOx (L), ctrl (H)	2931	2448

Figure 5.14 | Identification of WCE (input) proteome by quantitative mass spectrometry. (A) Scatter plot comparing the normalised $\log_2$ fold change upon AdOx treatment with label-swap biological replicates (two data sets). Each dot represents one protein group. **(B)** Histogram showing the normalised ratio distribution, with the count of each bin representing the number of protein groups. **(C)** Volcano plot showing log-fold changes in peptide intensities on the x axis and p values on the y axis. In **(A)** and **(C)**, protein groups with over 1.5 fold increase are coloured orange, of which t-test significant coloured pink; protein groups with over 1.5 fold decrease are coloured blue, of which t-test significant coloured purple.

Quantitative analysis of the mRNP capture assay

Similar to input proteins, in order to compare mRNP capture assays performed in the presence and absence of AdOx, proteins from mRNP capture assay eluates were in-gel digested and analysed using LC-MS (Q Exactive HF) following SDS-PAGE analysis. Approximately 650 protein groups were identified in each SILAC label-swap replicate, of which 80% were quantified, as summarised in Figure 5.15. Again, good consistency is seen for the two sets of data (label-swap replicates) from mRNP capture assays, as protein group $\log_2$ fold change values distribute evenly around 0 (Figure 5.15B). Quantitative proteomic analysis revealed that 37 protein groups were identified with over 1.5 fold increase ($\log_2$ value > 0.585), of which 5 are t-test significant (p value 0.05). For decreased proteins, 51 protein groups were identified over 1.5 fold change, of which 24 are t-test significant (Figure 5.15A and C).



Experiment	Identified	Quantified
AdOx (H), ctrl (L)	634	510
AdOx (L), ctrl (H)	640	532

Figure 5.15 | Effects of AdOx on mRNA interactome analysis by quantitative mass spectrometry. (A) Scatter plot comparing the normalised $\log_2$ fold change of AdOx treatment with label-swap biological replicates (two sets of data). Each dot represents one protein group. **(B)** Histogram showing the normalised ratio distribution, with the count of each bin representing the number of protein groups. **(C)** Volcano plot showing log-fold changes in peptide intensities on the x axis and p values on the y axis. In **(A)** and **(C)**, protein groups with over 1.5 fold increase are coloured orange, of which t-test significant coloured pink; protein groups with over 1.5 fold decrease are coloured blue, of which t-test significant coloured purple.

From the SILAC MS analysis, the results show that AdOx treatment not only changes protein abundance in the mRNA interactome, but also the whole cell protein expression. Thus it is necessary to confirm difference in abundance in the mRNA interactome is a consequence of alterations in the interaction with mRNA and not simply a change in the amount of protein in the cell extract used in the mRNP capture assay. To do this, the ratios (AdOx treatment over control) of proteins from the mRNP capture assay were normalised to those of WCE. However, not all proteins identified in the mRNA interactome were identified in the WCE, in which case, no normalisation was applied. The post-normalisation to WCE results are shown in Figure 5.16. In summary, the results do not alter the majority of proteins that were shown to differentially binding to mRNA with the addition of AdOx following normalisation to protein expression changes in the WCE.

Table 5.1 summarises those proteins whose mRNA binding increased with the addition of AdOx. These proteins are not necessarily t-test significant, but are generally with over 1.5 fold increase based on two label-swap replicates. Further analysis of the proteins identified was also performed based on network connections, where many proteins listed are involved in multiple connections and key functions (Figure 5.17).

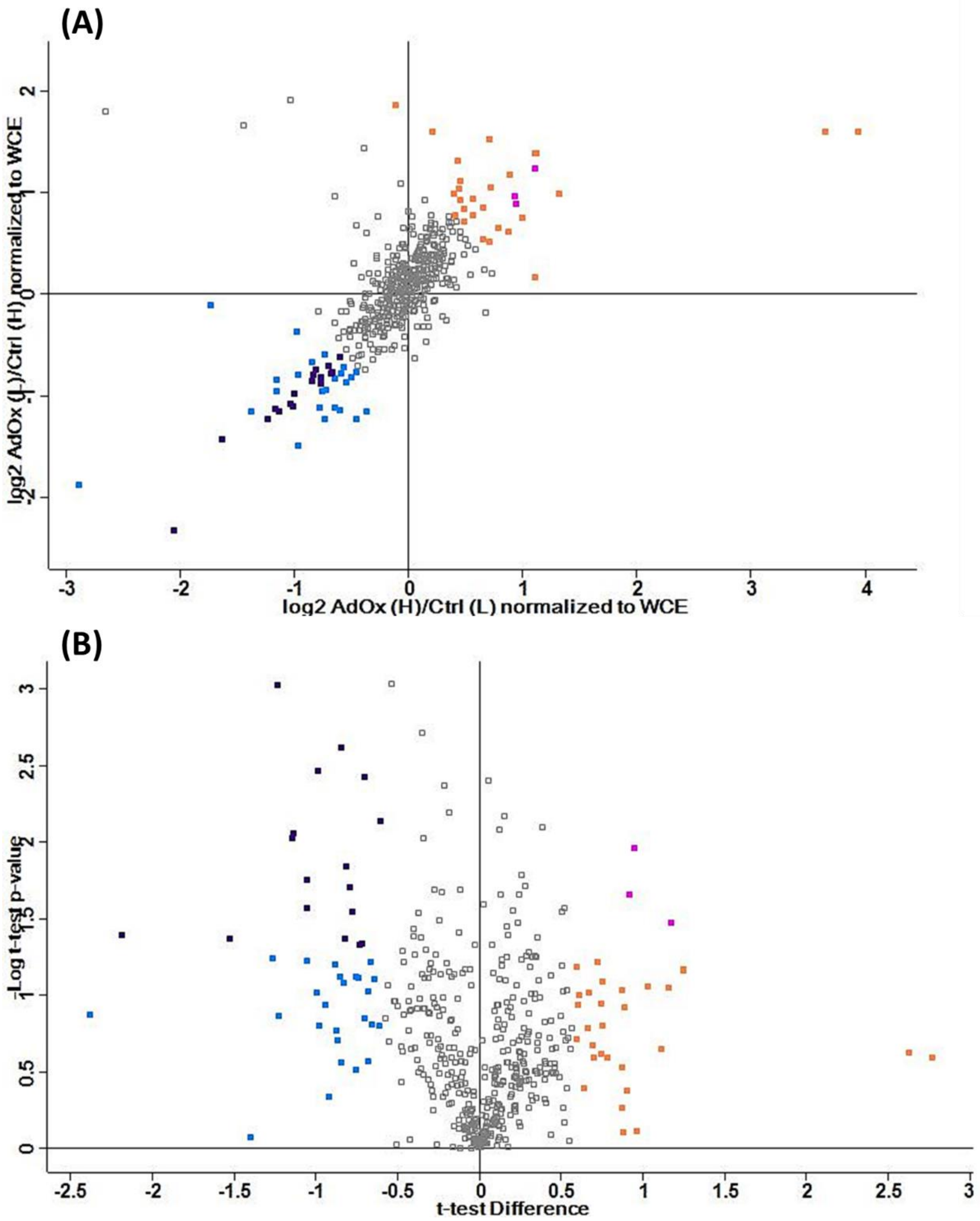


Figure 5.16 | Effects of AdOx on mRNA interactome including normalisation to expression changes observed in the cells. (A) Scatter plot comparing the normalised $\log_2$ fold change of AdOx treatment with label-swap biological replicates (2 sets of data). Each dot represents one protein group. **(B)** Volcano plot showing log-fold changes in peptide intensities on the x axis and p values on the y axis. Protein groups with over 1.5 fold increase are coloured orange, of which t-test significant coloured pink; protein groups with over 1.5 fold decrease are coloured blue, of which t-test significant coloured purple.

Table 5.1 | Proteins that showed increased mRNA binding upon treatment with AdOx.

Proteins	Log$_2$ Ratio AdOx/Ctrl (Average)	Log$_2$ Ratio normalised to WCE (Average)	Known methylation sites*		Methylation sites (UniProt)	RGG/RG sites	Cross reference **
C11orf68	0.89	0.89	R × 1			RG	Y
EEF1A1	0.74	0.52	R × 6	K × 14	G1, K165	RG	Y
LARP4B	0.74	0.64	R × 9	K × 1	R404, R419	RG	Y
NSUN2	0.86	0.59	R × 2			RG	
NYNRIN	1.11	1.11		K × 1		RG	
PABPC4	2.69	2.77	R × 13	K × 3	R419, R432, R436, R454, R530	GR	Y
PRMT1	0.41	1.03	R × 2				
PRRC2B	1.25	1.25	R × 14			RGG, RG repeats	Y
RBM33	0.75	0.75	R × 17		R470, R1028	RGG/RG	Y
SART3	0.74	0.74	R × 6	K × 1	R906	RG	Y
SNRP70	0.6	0.87	R × 4	K × 2		RGG/RG	
STAU1	0.74	0.51	R × 3			RGG/RG	Y
TRNAU1AP	1.17	1.17	R × 1			RGG	
TRUB1	0.8	1.15	R × 1			RG	
TUBB2C	0.7	0.6	R × 4	K × 1		RG	

*Based on iPTMnet database. R: arginine methylation sites; K: lysine methylation sites.
**Cross reference to the list of unique identified proteins with methylations by Geoghegan *et al.*, 2015. Y: the protein is on the list.

From Table 5.1, all proteins identified with log$_2$ fold change over 1.5 have known methylation sites, and most of them have potential RGG/RG motifs for methylations. Also, over half of the listed proteins are also recognised by other research groups including (Geoghegan et al., 2015) (cross reference shown in the list) and (Guo et al., 2014) (main source of UniProt methylation database).

Among the listed proteins, PABPC4 is identified with the largest fold increase upon AdOx treatment. In eukaryotes, PABPC4 binds to mRNA poly (A) tails, and have four RNA-binding domains (NCBI report). There are several known methylation sites in this protein. PRRC2B is also detected with large fold change. PRRC2B has multiple RGG/GRG sites, and is known involved in RNA binding and cell differentiation. However more detailed functions of this protein are not well characterised. Other interesting proteins include LARP4B (translation regulation), SART3 (mRNA processing and splicing), and NSUN2 (RNA methyltransferase).

Previous work (Hung et al., 2010) shows that the methylation of ALYREF (a mammalian mRNA export factor) reduces its RNA binding capacity. With the treatment of AdOx, a methylation inhibitor, the binding of ALYREF to mRNA is expected to increase. Indeed, in this study, ALYREF is identified with a significant increase in one set of the label-swap biological replicates ($\log_2$ fold change 0.42, 1.85 after normalisation to WCE), which is consistent with the previous results. However, no significant change is observed for this protein in the other replicate ($\log_2$ fold change -0.01, and -0.1 after WCE normalisation).

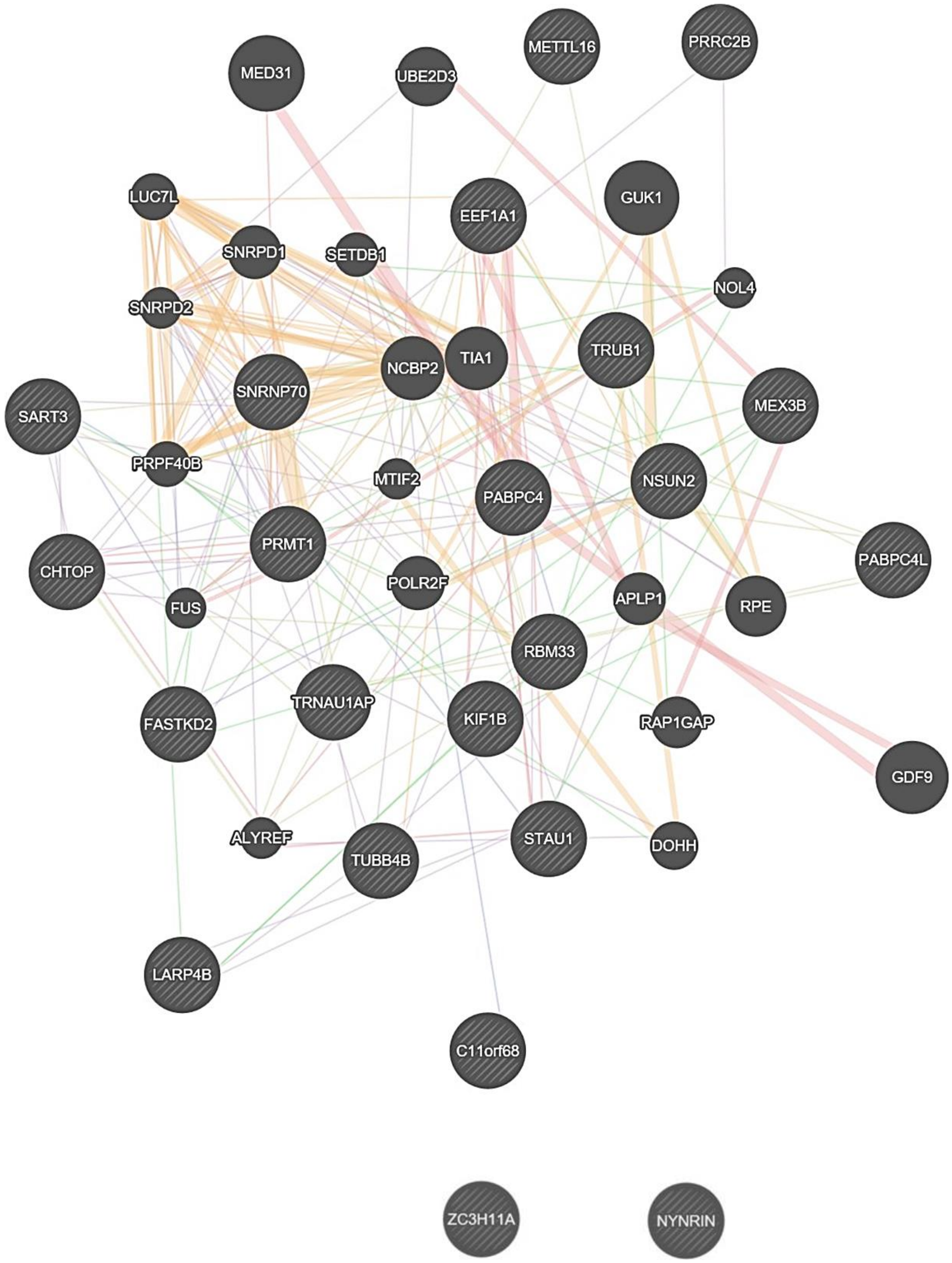


Figure 5.17 | Networks involved in proteins listed in Table 5.1. Data searched using GeneMANIA (genemania.org). Proteins listed in Table 5.1 are shown as cross-hatched circles, while solid circles are relevant proteins predicted by software. Lines represent interactions. Line thickness indicates interaction strength, and line colour indicates interaction type.

Similarly, proteins that show decreased binding to mRNA with the additional of AdOx are listed in Table 5.2, and the network connections involved in these proteins are shown in Figure 5.18.

Table 5.2 | Proteins that showed decreased mRNA binding upon treatment with AdOx.

Proteins	Log$_2$ Ratio AdOx/Ctrl (Average)	Log$_2$ Ratio normalised to WCE (Average)	Known methylation sites*		Methylation sites (UniProt)	RGG/RG sites	Cross reference **
ALG13	-1.05	-1.05				GRG/RG	
APOBEC3F	-0.7	-0.7	R × 1			RG	
ASCC3	-0.81	-0.88	R × 1	K × 3		RG	
CIRBP	-0.67	-0.36	R × 11			Multiple RGG/RG	Y
CPSF7	-0.63	-0.1	R × 1			RGG	Y
DDX21	-0.85	-0.21	R × 5	K × 1		Multiple RG	
DHX8	-0.79	-0.79	K × 1			RG	
DZIP3	-2.19	-2.19	K × 1			GRG, RG	
ESRP2	-2.38	-2.38	R × 1			RGG/RG	
FAM120C	-1.23	-1.23	R × 3			Y	Y
FUS	-0.47	-0.66	R × 28	K × 1	22 × R methylation sites	Multiple RGG, RG repeats	Y
HELZ2	-0.78	-0.78	K × 2			RG	
HNRNPLL	-0.86	-0.86	R × 1			RGG/RG	
LARP1	-0.93	-1.27	R × 6	K × 2		RGG, RG repeats	Y
LARP1B	-1.53	-1.53	R × 5			RG repeats, multiple RG	
LSM14B	-0.59	-0.29	R × 8			RGG, multiple RG	Y
NCL	-0.62	-0.72	R × 16	K × 4		Mutiple RGG/RG	Y
PPIL4	-0.92	-0.87	R × 5	K × 1		RGG	Y
PRKDC	-0.68	-0.3	R × 11	K × 6		RG	
PRPF8	-0.88	-0.4	R × 8	K × 5	K1425	RG	Y
PTBP2	-1.64	-0.92				RG	
RBM10	-1.17	-1.06	R × 5	K × 1	R902	RGG/RG	Y
RBM15B	-0.83	-0.83	R × 4			RGG, multiple RG	Y
RBM22	-0.68	-0.84	R × 1	K × 3		RG	
RBM27	-0.53	-0.68	R × 9	K × 4	R455	RG repeats, RG	Y
RBM4	-0.14	-0.74	R × 2			RG	Y

RBM4B	-0.77	-1.23	R × 1			RG	
RBM5	-0.98	-0.98	R × 1	K × 2		RG	Y
RC3H2	-0.82	-0.82	R × 1	K × 1		RG	
SF3A3	-0.7	-0.56	R × 3	K × 1		RG	
SF3B1	-0.7	-0.71	R × 6	K × 1		RGG/RG	
SF3B4	-0.54	-0.76	R × 5			Multiple RG	Y
SNW1	-0.78	-0.73	R × 4			RGG/GRG	Y
SRRT	-0.6	-0.95	R × 10		R833, R840, R850	GRG/RG	Y
XRCC6	-0.6	-0.37	R × 6	K × 4			
XRN1	-0.66	-0.66	R × 1			RG/GRG	
YTHDC2	-1.15	-1.15	R × 5	K × 3		RGG/RG	Y
ZC3H18	-0.77	-1.06	R × 3	K × 3		RGG/RG	Y
ZC3H8	-1	-1	K × 1			RG	
ZFC3H1	-0.61	-0.61	R × 9	K × 2		Multiple RGG/RG	

*Based on iPTMnet database. R: arginine methylation sites; K: lysine methylation sites.
**Cross reference to the list of unique identified proteins with methylations by Geoghegan *et al.*, 2015. Y: the protein is on the list.

Compared to the number of protein groups that increase upon AdOx treatment, more proteins are seen with over 1.5 fold decrease. The majority of proteins listed in table 5.2 have potential RGG/RG motifs and recognised methylation sites from the iPTMnet database. Specifically, proteins of particular interests are described below.

HNRNPLL is an RNA-binding protein required for alternative splicing (Oberdoerffer et al., 2008). One known arginine site is recognised for this protein, together with a RGG motif. FUS is a protein with various methylation sites. This protein contains a tri-RGG motif: RGG(X_{0-4})RGG(X_{0-4})RGG, a feature also seen in HNRNPA1 and HNRNPU (Thandapani et al., 2013). LARP1 is a RNA binding protein which possibly involved in the regulation of cell division, migration and apoptosis (Burrows et al., 2010). Six arginine methylation sites together with a few RGG motifs are recognised in this protein. LARP1 features a RG repeats (8 × RG) from the position 338. This study suggests methylation may facilitate the binding of LARP1 to mRNA. YTHDC2 is a protein with multiple methylation sites and RGG/RG motifs. This protein specifically recognises and binds to RNA with N6-methyladenosine (m^{6}A), and is proved to enhance translation efficiency and reduce target mRNA abundance (Hsu et al., 2017). ESRP2 is an mRNA splicing factor which regulates the formation of epithelial cell-specific isoforms (Warzecha et al., 2009). This protein sees the highest level of decrease upon AdOx treatment in this study. However, the number of possible methylation sites for this protein is limited. RBM10 also shows a significant decrease upon AdOx treatment. With multiple methylation sites and RGG/RG motifs, this protein is possibly involved in post-transcriptional processing.

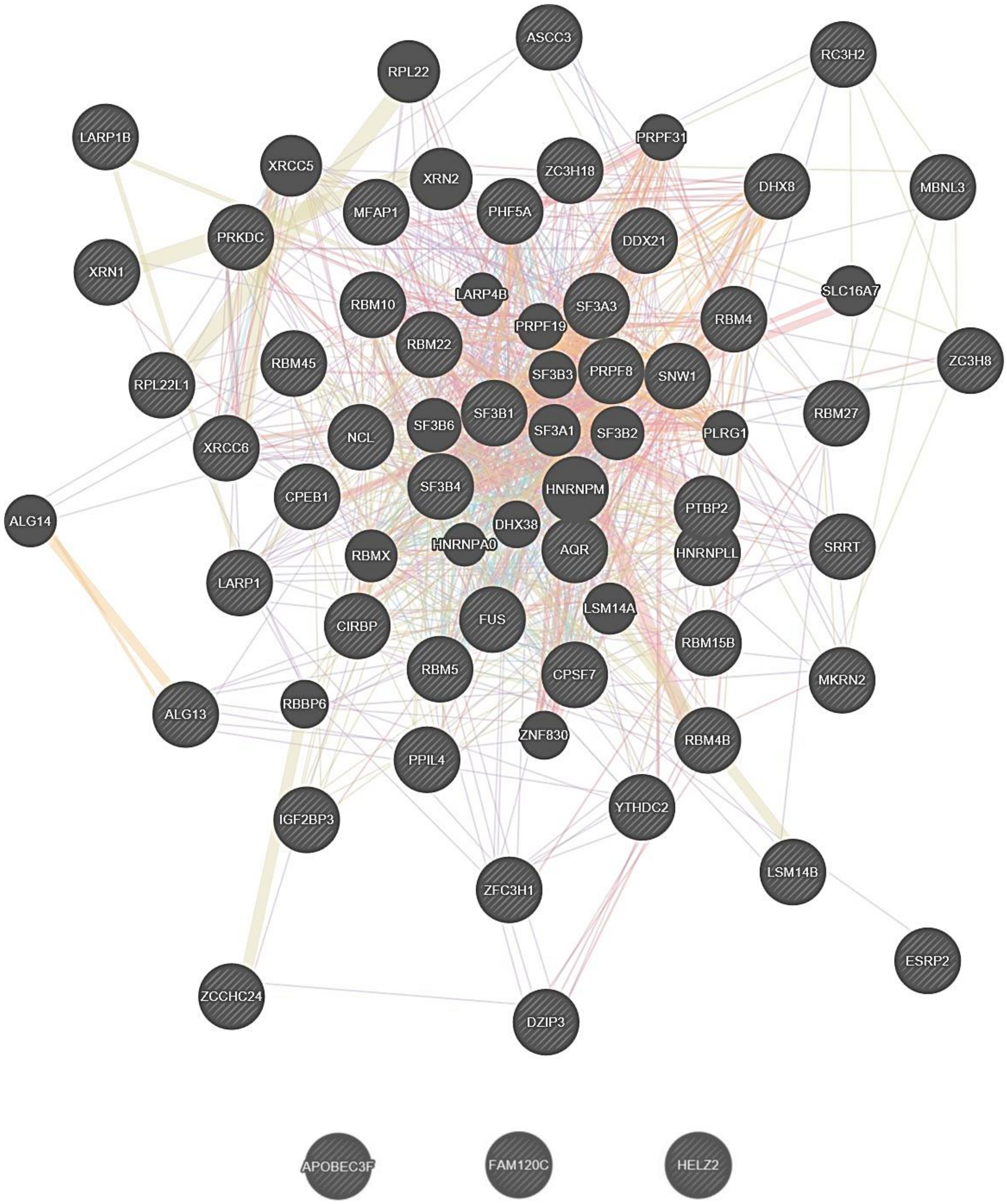


Figure 5.18 | Networks involved in proteins listed in Table 5.2. Data searched using GeneMANIA (genemania.org). Proteins listed in Table 5.2 are shown as cross-hatched circles, while solid circles are relevant proteins predicted by software. Lines represent interactions. Line thickness indicates interaction strength, and line colour indicates interaction type.

Apart from AdOx, the citrullination of arginine is another purposed experimental factor that can affect the interactions between proteins and nucleic acids. However, for the study of citrullination effect on the mRNA-binding proteome, experiments subject to further planning, due to the fact that induced PADI4 expression did not affect protein expression of both input and binding proteins (oligo GC *in vitro* pulldown assay, see Chapter 6).

5.3.5 Studying the effect of m6A on the mRNA interactome

To study the effect of N^6-methyladenosine (m^6A) of mRNA on mRNA-protein interactions, mRNP capture assays were carried out comparing a wild type cell line and an m^6A deficient cell line. The m^6A deficient cell line was generated using RNAi knockdown of Wilms tumour 1 associated protein (WTAP), a protein that recruits the m^6A methyltransferase complex (see Figure 5.1). Therefore knockdown results in reduced m^6A RNA methylation. Cells (HEK 293T) were obtained from the laboratory of Prof S Wilson, University of Sheffield. For simplification purpose, m^6A deficient cell lines are described as m^6A, and wild type cell lines as ctrl. Both the cell lines were cultured in heavy and light media for SILAC label-swap biological replicates using the optimised SILAC growth conditions as previously described. After, mRNP capture assays were performed as previously optimised, using a mix of light and heavy cell lysate (10 mg each) prior to three consecutive enrichments using oligo-d(T) beads. Isolated mRNA-binding proteins were resolved by SDS-PAGE (Figure 5.19). Here, the eluates from the first and second mRNP enrichments were loaded separately for comparison. It is clear that considerable amount of proteins can be isolated after the first round of pulldown assay.

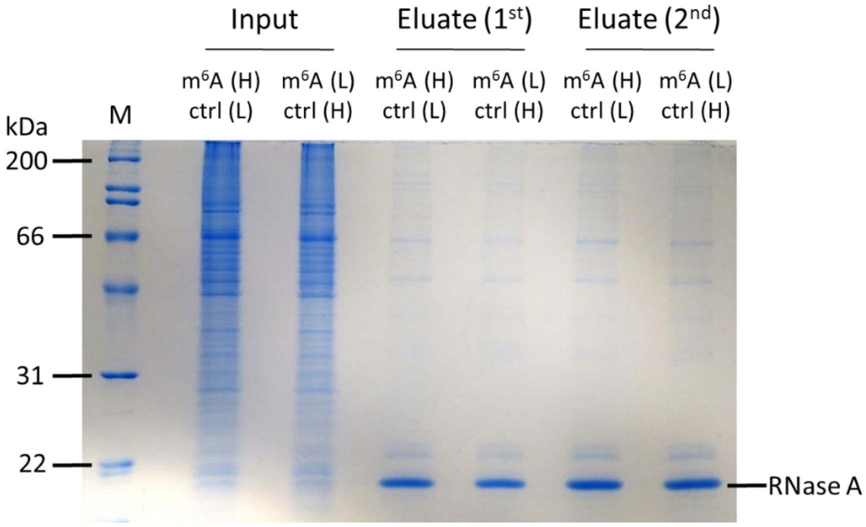
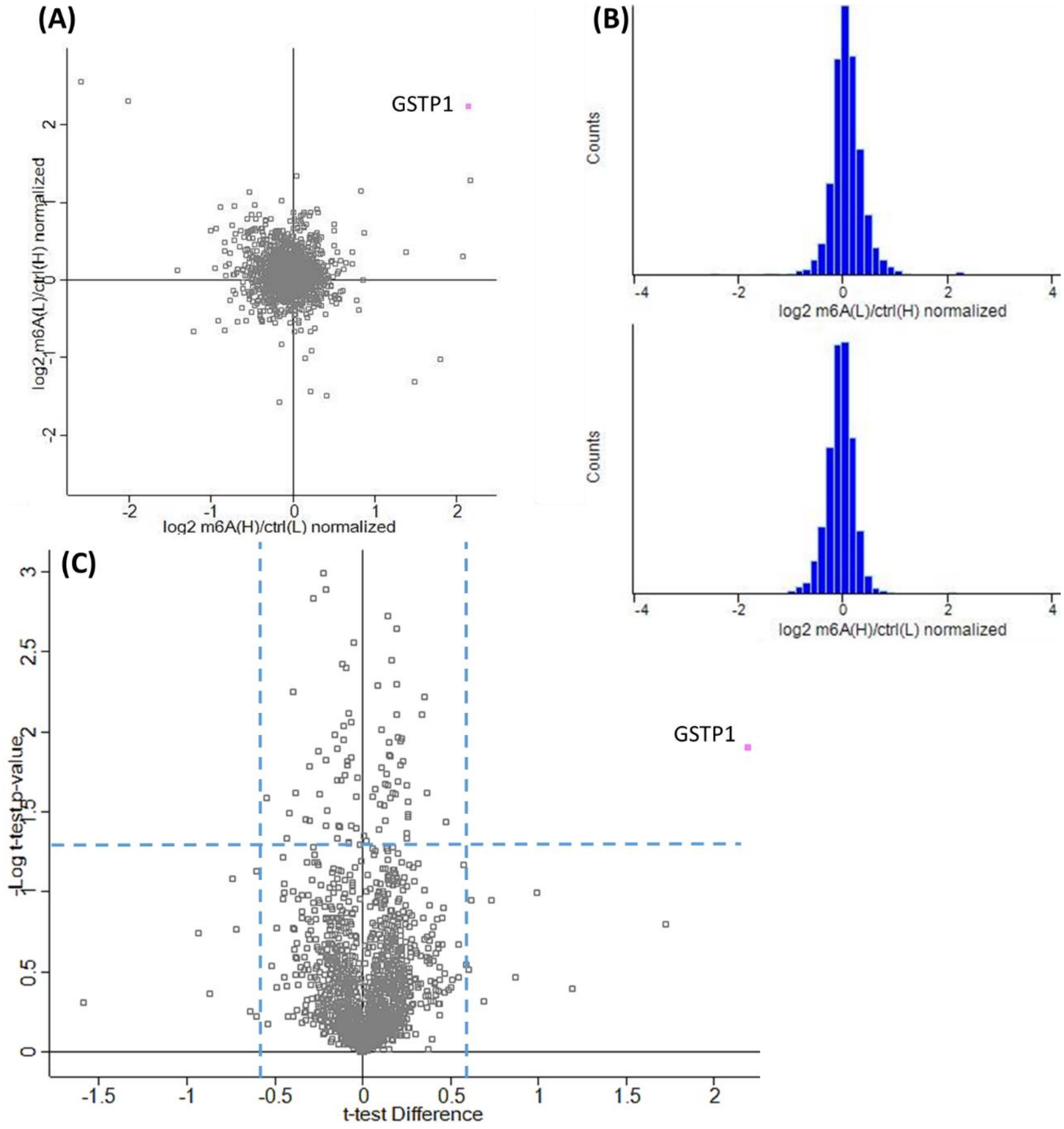


Figure 5.19 | SDS-PAGE analysis of inputs (WCE) and eluates (mRNP capture assay) from m^6A deficient cell lines (m^6A) and wild type cell lines (ctrl). Both cell lines were grown in SILAC media with label-swap biological replicates, then crosslinked at UV 254 nm upon harvest. Three consecutive pulldowns were performed, with the first two shown on gel. Proteins were visualised by Coomassie blue staining. M: protein marker.

Quantitative analysis of the whole cell protein expression

After the SDS-PAGE visualisation of mRNA-binding proteins captured, mass spectrometry based analysis was carried out to study the protein expression in the WCE under m^{6}A deficient condition, as the WTAP gene knock down may affect protein expression over a range of different proteins. To do this, input proteins were in-gel digested and analysed on a Q Exactive HF mass spectrometer. For MS data analysis, raw data were processed using MaxQuant, and statistical analysis was performed in Perseus. Results show approximately 2500 protein groups were identified in each replicate of input, of which 75% were quantified (Figure 5.20). The log$_2$ fold change values of protein groups distribute evenly around 0 (Figure 5.20B), indicating good consistency of the label-swap replicates. However, quantitative proteomic analysis showed no significant changes (either increase or decrease) on expression level (Figure 5.20C, one-sample t-test shows that no protein falls into t-test significant area with over 1.5 fold change, or log$_2$ value 0.585) except GSTP1 (pink dot, Figure 5.20A and C), which has no direct association known with m^{6}A. The results indicate the growth and protein expression of the expected WTAP gene knock down cell line (-m^{6}A) were not significantly affected. This is not expected, as the alteration of m^{6}A status is a major biological event for cell behaviour, and theoretically a list of proteins should have significant fold changes in expression level.



Experiment	Identified	Quantified
m6A (H), ctrl (L)	2359	1856
m6A (L), ctrl (H)	2479	1905

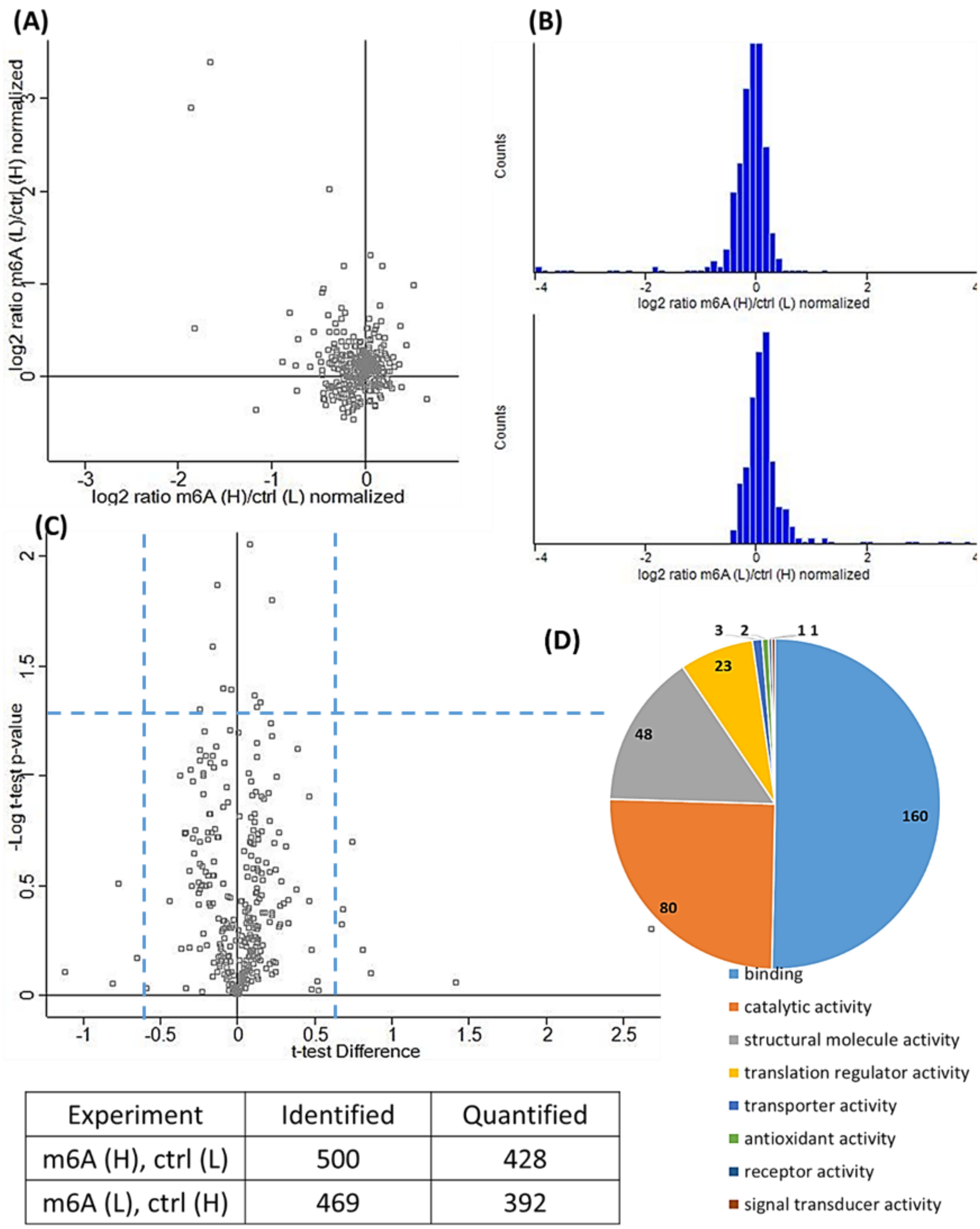
Figure 5.20 | Identification of WCE (input) proteome by quantitative mass spectrometry. (A) Scatter plot comparing the normalised log$_2$ fold change of m^{6}A deficient cell line (m^{6}A) and wild type (ctrl) with label-swap biological replicates (two data sets). Each dot represents one protein group. **(B)** Histogram showing the normalised ratio distribution, with the count of each bin representing the number of protein groups. **(C)** Volcano plot showing log-fold changes in peptide intensities on the x axis and p values on the y axis. Dotted lines (blue) mark the boundaries of t-test significance and 1.5 fold change.

Quantitative analysis of the mRNP capture assay

Following the analysis of the WCE proteome, quantitative analysis was performed on the enriched mRNA-binding proteins to determine if protein binding ability to mRNA has been altered due to m^{6}A deficiency. Proteins from mRNP capture assay eluates were in-gel digested and analysed using LC-MS (Q Exactive HF). Approximately 500 protein groups were identified in each SILAC label-swap replicate, of which 400 were quantified, as summarised in Figure 5.21. Quantified heavy and light protein groups distribute evenly around 0 (Figure 5.21B), indicating good consistency of label-swap. However, quantitative proteomic analysis showed no significant increase or decrease in protein amounts (Figure 5.21C, one-sample t-test shows that no protein falls into t-test significant area with over 1.5 fold change, or log$_2$ value 0.585). Due to the fact that a number of mRNA-binding proteins are proved sensitive to m^{6}A such as HNRNPC (see Chapter 5.2), no changes of mRNA-binding proteins in this work suggests the invalidation of experiment. Table 5.3 lists a number of proteins that are closely related to mRNA m^{6}A modification. These proteins were identified and quantified in this work, but showed no change of amounts in m^{6}A deficient cell lines. Furthermore, GO annotation analysis of proteins recovered from mRNP capture assay based on molecular function gives correct protein type classifications compared to previous experiments (Figure 5.21D), suggesting the mRNP capture assay was performed successfully. These evidences strongly suggest the m^{6}A deficient cells may not have WTAP gene properly knocked down.

Table 5.3 | list of mRNA-binding proteins quantified in this experiment with known relation to m^{6}A modification.

Gene names	Protein names	log$_2$ ratio m^{6}A(H)/ ctrl (L)	log$_2$ ratio m^{6}A (L)/ ctrl (H)
HNRNPC	Heterogeneous nuclear ribonucleoproteins C1/C2	-0.02	0.12
YTHDF2	YTH domain-containing family protein 2	-0.05	0.13
HNRNPA2B1	Heterogeneous nuclear ribonucleoproteins A2/B1	0.01	0.22
YTHDF1	YTH domain-containing family protein 1	-0.29	-0.10
YTHDF2	YTH domain-containing family protein 2	-0.05	0.13
EIF3A	Eukaryotic translation initiation factor 3 subunit A	0.26	0.06



Experiment	Identified	Quantified
m6A (H), ctrl (L)	500	428
m6A (L), ctrl (H)	469	392

Figure 5.21 | Effect of m^{6}A deficient on mRNA-bound proteome analysis by quantitative mass spectrometry. (A) Scatter plot comparing the normalised log$_2$ fold change of m^{6}A deficient cell line (m^{6}A) and wild type (ctrl) with label-swap biological replicates (two data sets). Each dot represents one protein group. **(B)** Histogram showing the normalised ratio distribution, with the count of each bin representing the number of protein groups. **(C)** Volcano plot showing log-fold changes in peptide intensities on the x axis and p values on the y axis. Dotted lines (blue) mark the boundaries of t-test significance and 1.5 fold change. **(D)** GO annotation of quantified proteins in eluates based on molecular function. GO annotation was performed using PANTHER classification system (pantherdb.org).

To test this hypothesis, immunoblotting experiments were performed on both the m^{6}A deficient cell line and the wild type cell line for comparison. For WCE, western blot was performed using a WTAP antibody. Also, total RNA extraction was performed on both cell lines, and extracted total RNA was analysed in a dot blot experiment using a N^6-methyladenosine antibody. Indeed, results show no significant difference in either test between the two cell lines (Figure 5.22). Similar amount of WTAP protein is present in the WTAP gene knockdown cell line compared to the wild type cell line, and the amount of N^6-methyladenosine present on RNA was the same in the WTAP gene knockdown cell line compared to the wild type cell line.

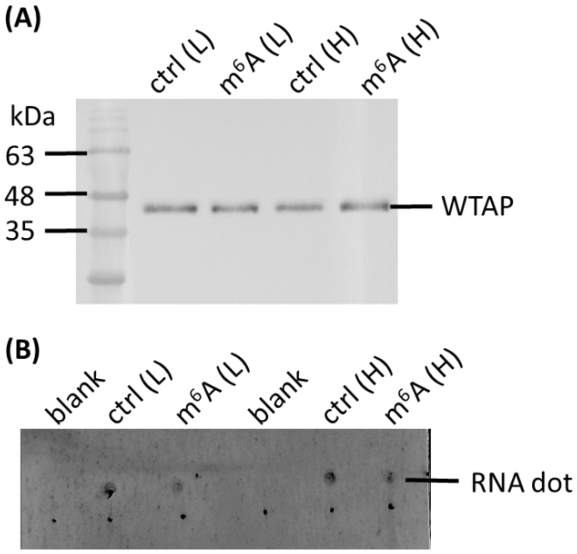


Figure 5.22 | Immunoblot tests of m^{6}A deficient cell line. (A) WCE of the m^{6}A deficient cell line (m^{6}A) and the wild type cell line (ctrl) were analysed by western blotting (WB) with a WTAP antibody. **(B)** Total RNA of both the m^{6}A and the wild type cell line was analysed by dot blotting with a N^6-methyladenosine antibody.

These results demonstrate that the RNAi knockdown was not efficient in the cell line used, resulting in no significant alteration of RNA methylation. These results are consistent with the previous proteomic analysis, which showed no significant changes in protein expression in the m^{6}A deficient cell line or the mRNA bound proteome when compared to the wild type.

5.4 Conclusions

RNA biology is largely determined by the interplay of RNAs with RNA binding proteins within dynamic ribonucleoprotein complexes in the cell. Recent studies have provided insights into the repertoire of different RNA binding proteins that bind mRNA, termed the mRNA interactome (Baltz et al., 2012; Castello et al., 2012). Protein post-translational modifications and chemical modifications of RNA can affect RNA-protein interactions. In this Chapter, I have studied the effects of both protein methylation and RNA methylation on mRNA interactome.

Initial work focussed on the optimisation of *in vivo* mRNP capture assays. Large scale mRNP capture assays were performed using CLIP and PAR-CLIP approaches to capture RNA-protein interactions in the cell prior to purification of the mRNA using oligo-d(T) and GeLC-MS analysis. Standard UV crosslinking at 254 nm proved to be the most efficient for crosslinking proteins to mRNA. The MS analysis identified over 900 proteins, the majority of which are nucleic acid binding proteins as expected.

Further quantitative proteomic analysis was performed using mRNP capture assays in conjunction with stable isotope labelling with amino acids in cell culture to determine the effects of the global methylation inhibitor AdOx and the effects of RNA methylation (N^6-methyladenosine) on mRNA binding. For the mRNP capture assays performed on wild type cells and N^6-methyladenosine RNA deficient cells, approximately 500 proteins were identified, of which 400 were quantified. However, no difference in the mRNA interactome was observed. Further analysis revealed that the RNAi knockdown was not efficient in the cell line used, resulting in no significant alteration of RNA methylation and therefore the mRNA binding was not altered.

The analysis of the effects of the global methylation inhibitor AdOx identified over 600 proteins of which over 500 were quantified in the SILAC label-swap experiment. 24 proteins showed increased mRNA binding and 50 showed decreased mRNA binding (with over 1.5 fold change in protein abundance). A wide number of identified proteins contain sites of arginine methylation which are of particular interest. These results are the first global quantitative analysis of the effect of methylation on mRNA binding, with a number of interesting RNA binding proteins highlighted as candidates for further studies to examine the role of arginine methylation on RNA binding. Identified proteins of particular interests include PABPC4, PRRC2B, LARP4B, SART3, NSUN2 (with increased mRNA binding upon AdOx treatment), and HNRNPLL, FUS, LARP1, YTHDC2, ESRP2, RBM10 (with decreased mRNA binding upon AdOx treatment). Further insight into these proteins may contribute to the understanding of mRNP cellular functions and the regulation of post-translational modifications that are included in certain biological pathways.

Chapter 6: Studying the effect of protein methylation and citrullination on the binding to $(GGGGCC)_5$ repeats

6.1 Abstract

The expansion of GGGGCC repeats in the *C9orf72* gene is believed to be the cause of both amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). The molecular pathogenesis however, is unclear. A number of proteins that interact with GGGGCC repeats have been reported. In this study, an *in vitro* pulldown assay was used to study the effects of both protein methylation and arginine citrullination on the proteins that bind to RNA GGGGCC repeats.

Quantitative proteomic analysis was performed using *in vitro* pulldown assays in conjunction with stable isotope labelling with amino acids in cell culture (SILAC) to determine the effects of the protein methylation and citrullination on the binding to GGGGCC repeats. The analysis of the effects of the global methylation inhibitor AdOx identified over 700 proteins of which over 600 were quantified in the SILAC label swap experiment. 73 proteins showed increased RNA binding and 57 showed decreased RNA binding to the GGGGCC repeats (> 1.5 fold change in protein abundance). To study the effects of arginine citrullination, a cell line with induced overexpression of PADI4 was used. Over 700 proteins were identified, of which close to 600 were quantified in the SILAC label swap experiment, similar to the AdOx experiments. However, only 9 proteins showed increased $(GGGGCC)_5$ oligo RNA binding and 2 showed decreased $(GGGGCC)_5$ oligo RNA binding (> 1.5 fold change in protein abundance).

These results are the first global quantitative analysis of the effects of methylation/citrullination on GGGGCC repeat binding and highlight a number of interesting GGGGCC repeat binding proteins as candidates for further studies to examine the role of protein PTMs in amyotrophic lateral sclerosis and frontotemporal dementia.

6.2 Introduction

$(GGGGCC)_n$ is a hexanucleotide repeat expansion (HRE) discovered recently in the non-coding region (the first intron or the promoter region) of the *C9orf72* gene, and is commonly believed to be the cause of two neurodegenerative disorders known as amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) (Mori et al., 2013; DeJesus-Hernandez et al., 2011; Renton et al., 2011). Both diseases are related to devastating symptoms including dementia, language ability deficiency, progressive muscle size decreasing and change of personalities (Josephs et al., 2011; Mackenzie et al., 2010). Apart from ALS and FTD, evidence suggests HRE of *C9orf72* is also related to other neurodegenerative diseases such as Alzheimer's and Huntington's (Rollinson et al., 2012; Hensman et al., 2014).

Compared to normal human *C9orf72* gene with GGGGCC repeat units of less than 25, gene sequencing results show several hundreds to thousands of this repeat pattern in patients with ALS or FTD (van der Zee et al., 2013). The pathology of ALS and FTD related GGGGCC repeats, as well as the correlation between the number of repeats and the progression of disease remain largely unknown (Haeusler et al., 2014). Hypothesis of the disease mechanisms are described in Figure 6.1. One theory is that the mutation interferes with the normal expression of the *C9orf72* gene, as evidence showed decreased amount of transcribed mRNA from *C9orf72* in ALS/FTD carriers (DeJesus-Hernandez et al., 2011; Gijselinck et al., 2012). However, the function of the protein encoded by *C9orf72* is unknown. The second possibility is RNA mediated toxicity. Evidence showed accumulation of RNA transcripts containing GGGGCC repeats in the spinal cord material and the frontal cortex of the patients, which could lead to toxicity (DeJesus-Hernandez et al., 2011). The expansion of GGGGCC repeats can cause sequestration and affect the protein-RNA interactions (Renoux and Todd, 2012). In fact, a number of GGGGCC repeat binding proteins were already identified, such as ribonucleoproteins hnRNPA1 and hnRNPA2B1, both of which are thought to be ALS/FTD related (Kim et al., 2013). The third hypothesis is described as repeat-associated non-ATG (RAN) translation. Unlike conventional translations, RAN-translation requires no ATG codon to initiate the translation process. This was first discovered in the expanded CAG repeats in spinocerebellar ataxia type 8 (SCA8) and myotonic dystrophy type 1 (DM1), which lead to the formation of toxic homopolymeric peptides (Zu et al., 2010). Similarly, GGGGCC repeats were also revealed to have this type of translation mode (Cleary and Ranum, 2013), through which homopolymeric dipeptide repeat proteins (DPRs) including poly Gly-Ala, poly Gly-Pro and poly Gly-Arg, can be produced across the reading frames of the repeats and accumulated in neurons of patients (Mori et al., 2013; Callister et al., 2016).

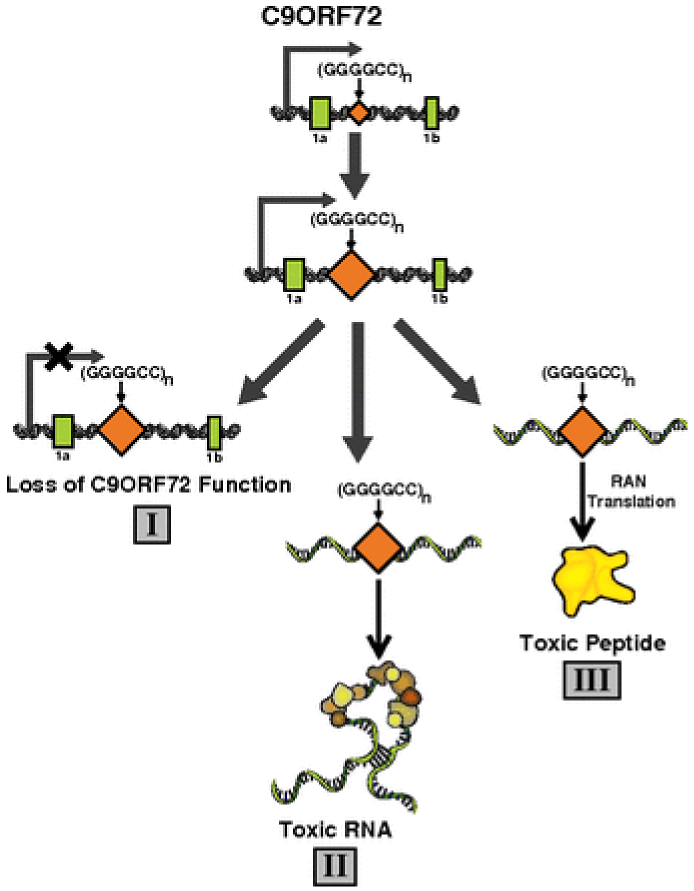


Figure 6.1 | Illustration of the proposed mechanisms of how GGGGCC repeats in *C9orf72* are associated with ALS and FTD. Figure reprinted with permission from Orr, 2013.

RNA pulldowns in conjunction with mass spectrometry have been widely used to study the GGGGCC repeat binding proteome. Several proteins have been identified to recognize and interact with GGGGCC repeats, which can be categorized based on their functions, including for example, pre-mRNA slicing factors HNRNPA1, HNRNPA2B1, FUS, EWSR1, SRSF (1-3), TAF15, heterogeneous nuclear ribonucleoproteins HNRNPK and HNRNPH/F (Cooper-Knock et al., 2015), helicase DDX21, DHX15 and DHX30, interleukins ILF2 (Mori et al., 2013), stability SAFB2, and others including PCBP2

(Haeusler et al., 2014), ALYREF (Cooper-Knock et al., 2015) and ADARB2 (Donnelly et al., 2013).

G-quadruplexes are unusual secondary structures which can be formed in G-rich DNAs and RNAs both *in vitro* and *in vivo* (Neidle, 2009; Biffi et al., 2013; Bugaut and Balasubramanian, 2012) (Figure 6.2). This structural change can affect multiple biological processes such as gene regulation and RNA processing (Lipps and Rhodes, 2009; Melko and Bardoni, 2010). Studies have shown that G-quadruplex structures can also form within the GGGGCC repeats in the *C9orf72* gene (Reddy et al., 2013). The formation of this structure not only affects RNA-RNA interactions, but is also likely to affect protein binding patterns. Some proteins preferentially bind to a G-quadruplex, such as HNRNPU and NCL, whereas other proteins including HNRNPF and RPL7 are not sensitive to structural differences (Haeusler et al., 2014).

Protein post-translational modifications can change the chemical structure and the property of proteins, which in turn affect protein-RNA interactions. In this study, protein methylation and arginine citrullination were studied for their effects on binding to GGGGCC repeats, using quantitative mass spectrometry methods in conjunction with SILAC.

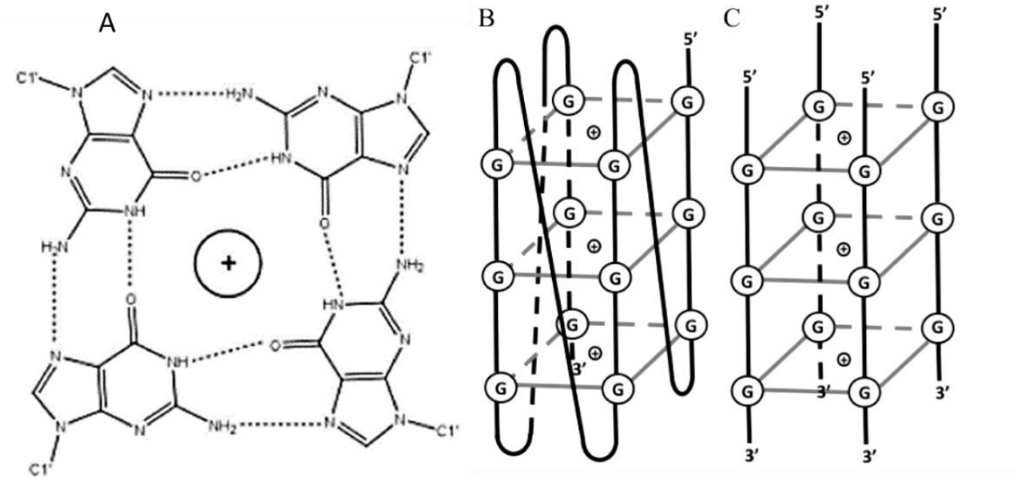


Figure 6.2 | Schematic represent the formation of G-quadruplex with GGGGCC repeats. **(A)** Interactions between four guanine residues, which are stabilized by ions such as K^+ and Na^+ in the centre. **(B)** The formation of G-quadruplex in the form where guanines interact with each other within one molecule. **(C)** The formation of G-quadruplex in the form where guanines interact with guanines from different molecules. Figure reprinted with permission from Reddy *et al.*, 2013.

6.3 Results and Discussion

6.3.1 Optimisation of an *in vitro* binding assay to study the proteins binding to GGGGCC repeats

For the identification of proteins that interact with RNA GGGGCC repeats, an *in vitro* assay was optimised in this study. A 3' biotinylated RNA oligo $(GGGGCC)_5$ with the sequence of 5'-$(GGGGCC)_5$-Bio-3' was used as a probe representing the $(GGGGCC)_n$ HRE in the *C9orf72* gene. Biotinylation is the most widely used affinity tag for oligonucleotides due to its high specific affinity to streptavidin, fast binding and high stability in various experimental conditions (Jazurek et al., 2016). For GGGGCC repeats, repeats of five were used because a minimum of four repeats for the GGGGCC repeat expansion are required to form the correct G-quadruplex structure *in vitro* (Figure 6.2), and have similar binding ability and specificity to associated proteins compared to longer ones (Mori et al., 2013).

Initial work was performed to optimise the *in vitro* RNA pulldowns using HeLa nuclear extracts. 3' biotinylated RNA oligos $(GGGGCC)_5$ were added to the protein extracts (up to 2 mg was used for each pulldown experiment) and crosslinked at UV 254 nm, 0.3 J/cm^2 prior to affinity purification with streptavidin beads . Proteins were eluted from the RNA oligos with the addition of RNase A and analysed using SDS-PAGE (Figure 6.3). The results show the successful enrichment of proteins binding to the GGGGCC repeats.

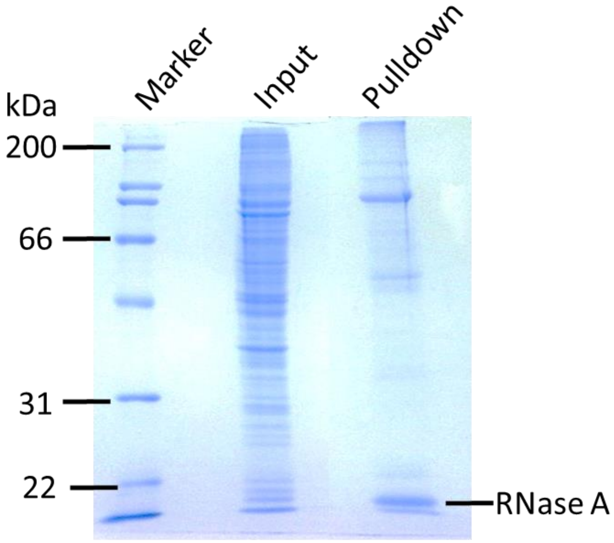


Figure 6.3 | SDS-PAGE results of input (HeLa nuclear extracts) and pulldown (eluates). Proteins were visualised by Coomassie blue staining.

Further analysis was performed using mass spectrometry (amaZon) for protein identification. In-gel digestion was performed on the pulldown lane of Figure 6.3, and data was searched against the SwissProt data base with Homo sapiens (human) taxonomy using Mascot Daemon. 155 proteins were identified (see Chapter 6 Appendices). Gene ontology (GO) annotation results based on molecular function show most protein groups are with binding functions (53%) and catalytic activity (28%) (Figure 6.4), indicating the *in vitro* pulldown assay was effective.

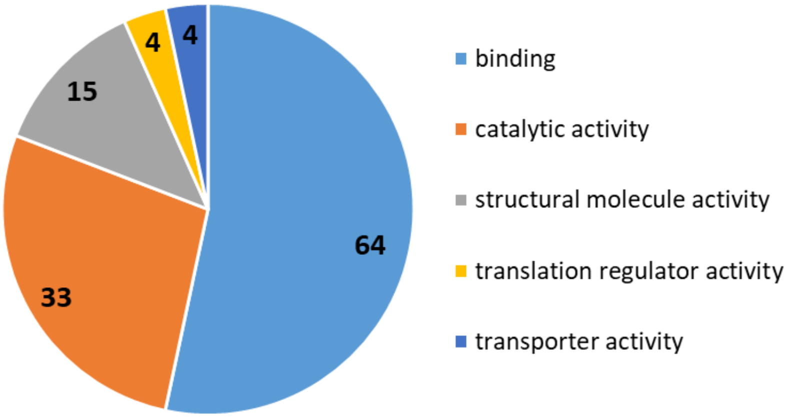


Figure 6.4 | *In vitro* pulldown assay validation based on gene ontology (GO) annotation. *In vitro* pulldown assay was performed using HeLa nuclear extracts. Identified protein groups were categorized based on molecular function.

Following optimisation and validation of *in vitro* RNA pulldowns, quantitative MS analysis of the $(GGGGCC)_5$ binding proteome is required to demonstrate that binding to RNA $(GGGGCC)_5$ sequence is perturbed under different conditions. To perform quantitative proteomic analysis, stable isotope labelling by amino acids in cell culture (SILAC) with label-swap strategy for biological replicates was used (as previously described in Chapter 5). The experimental procedure is designed as shown in Figure 6.5.

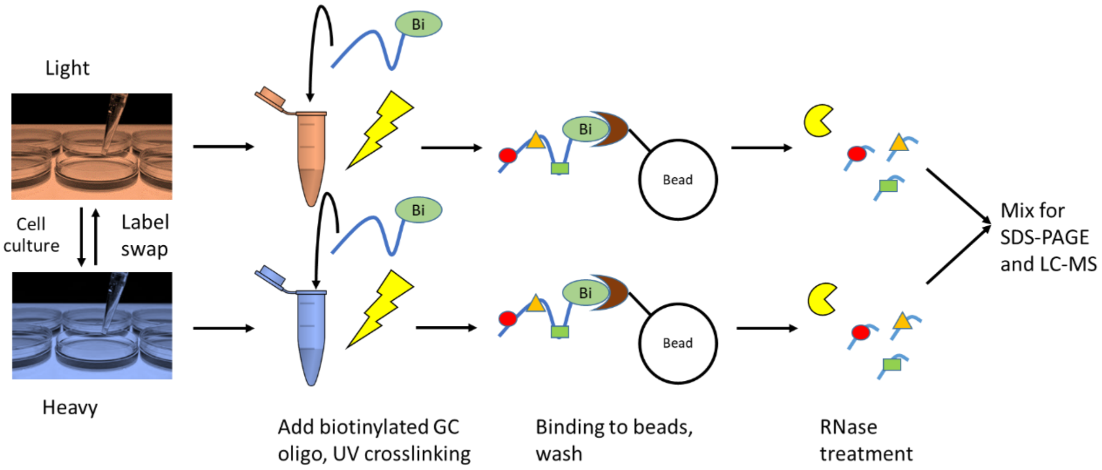


Figure 6.5 | Schematic of *in vitro* RNA pulldown assay for quantitative studies of proteins binding to GGGGCC repeats. Key steps include SILAC swap-labelling, *in vitro* UV crosslinking, streptavidin-biotin affinity precipitation and RNA-binding protein recovery.

6.3.2 Studying the effect of methylation on the GGGGCC repeat interactome

To study the effect of protein methylation on GGGGCC repeat binding, cells were grown in the presence and absence of the global methylation inhibitor AdOx in conjunction with *in vitro* RNA pulldown assays and SILAC MS analysis. For protein demethylation with the addition of AdOx, HEK 293T cells were treated with 20 µM AdOx for 48 hours for methylation inhibition. Protein demethylation in the presence of AdOx were verified by SDS PAGE and western blotting using CHTOP antibody (see Chapter 5).

For the SILAC experiments, cells were cultured in heavy and light media with the label-swap strategy for biological replicates (optimised SILAC growth conditions were used as previously described in Chapter 5). *In vitro* RNA pulldown assays were performed separately using up to 2 mg of light and heavy cell lysate each for the $(GGGGCC)_5$ oligo RNA crosslinking and affinity purification. The recovered eluates of both heavy and light were mixed (1:1 protein amount) prior to analysis on SDS-PAGE. The results shown in Figure 6.6 demonstrate the successful enrichment of proteins binding to the GGGGCC repeat oligos in the presence and absence of AdOx.

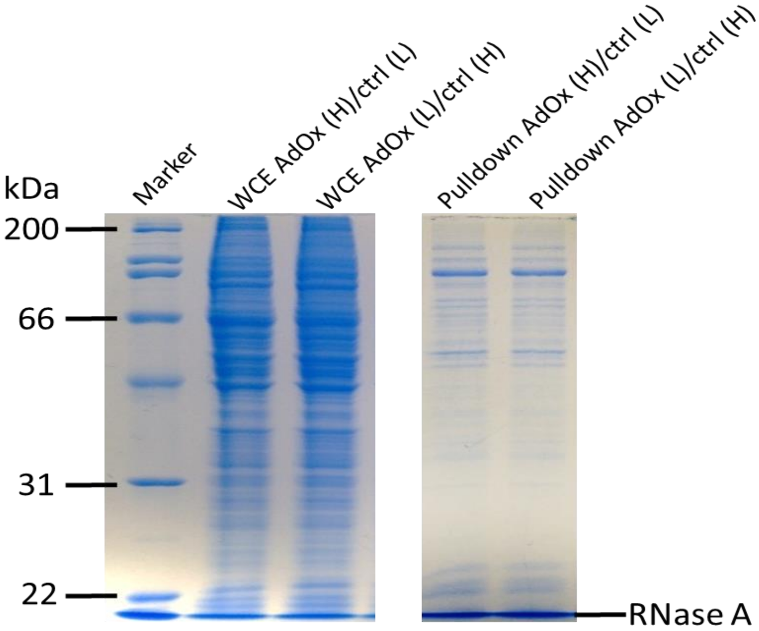
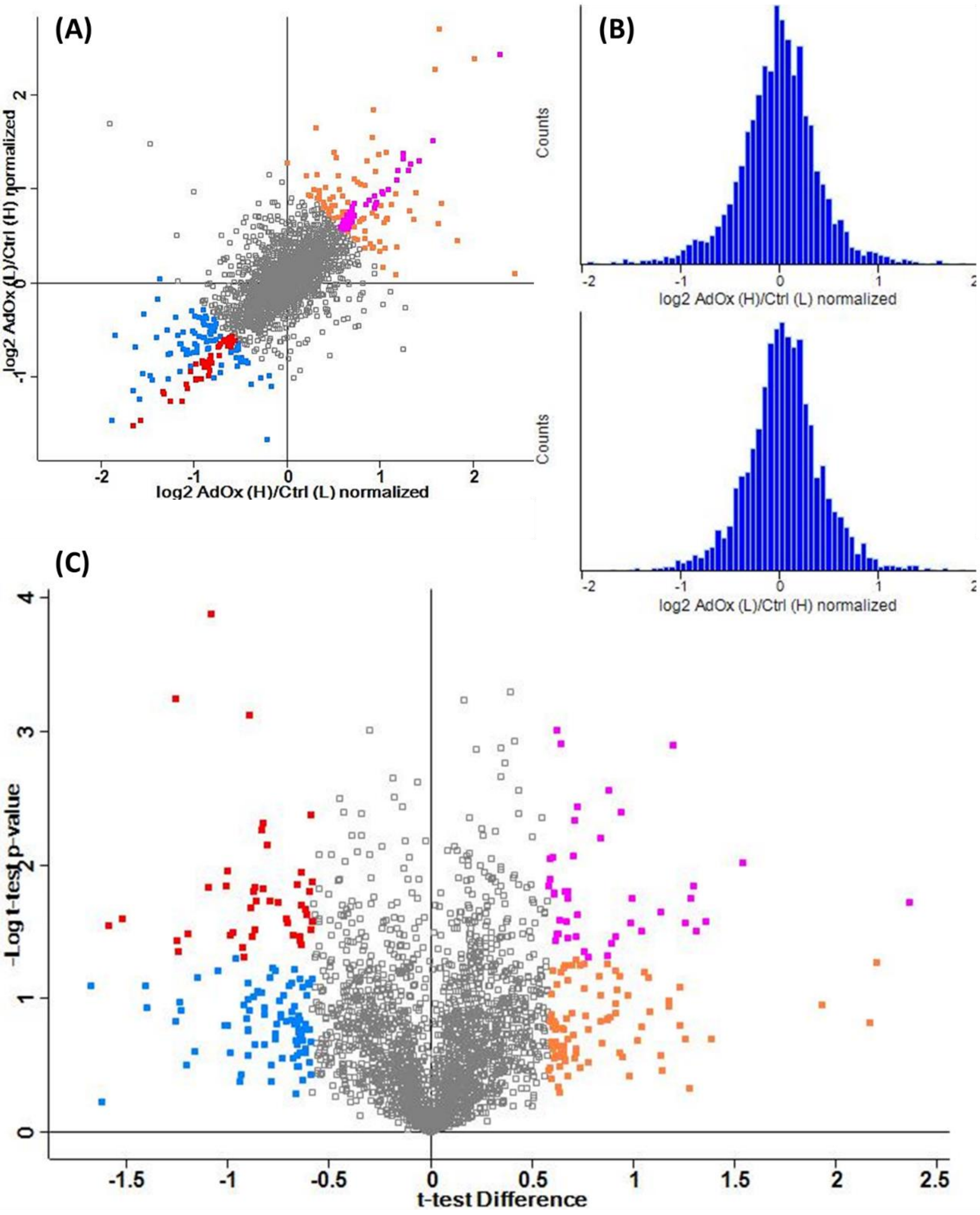


Figure 6.6 | SDS-PAGE results of whole cell extracts (WCE) and *in vitro* pulldown eluates from AdOx treated cells (AdOx) and control cells (ctrl). HEK 293T cells were grown in SILAC media with label-swap biological replicates, and crosslinked at UV 254 nm. For both WCE and pulldowns, samples were mixed with estimated protein amount of 1:1 heavy to light prior to loading on the gel.

Quantitative analysis of the whole cell protein expression

In addition to the analysis of the proteins binding to the GGGGCC repeats, it is important to analyse the whole cell protein lysates (WCE) in the presence and absence of AdOx to determine the effect of AdOx on protein expression levels (as previously described in Chapter 5). SILAC labelled WCE samples were therefore in-gel digested and analysed on the Q Exactive HF. Raw data were processed using MaxQuant, and statistical analysis was performed in Perseus.

The MS analysis identified over 3000 protein groups (see Section 5.3.2 for protein group definition) in each replicate and over 85% of which were quantified. The quantitative MS analysis is summarised in Figure 6.7. The log$_2$ fold values of protein groups distribute evenly around 0 (Figure 6.7B), indicating good consistency of the label-swap replicates. Quantitative proteomic analysis revealed 193 proteins which were identified with over 1.5 fold change (log$_2$ value > 0.585). Among these proteins 42 are t-test significant (p value 0.05). Another 193 proteins showed decreased expression with over 1.5 fold change (log$_2$ value < -0.585), of which 43 were t-test significant (Figure 6.7A and C). The results demonstrate the successful application of the SILAC workflow to identify a number of differentially expressed proteins upon the addition of the methylation inhibitor AdOx. These results will be used in conjunction with the *in vitro* RNA binding assays described in the following section.

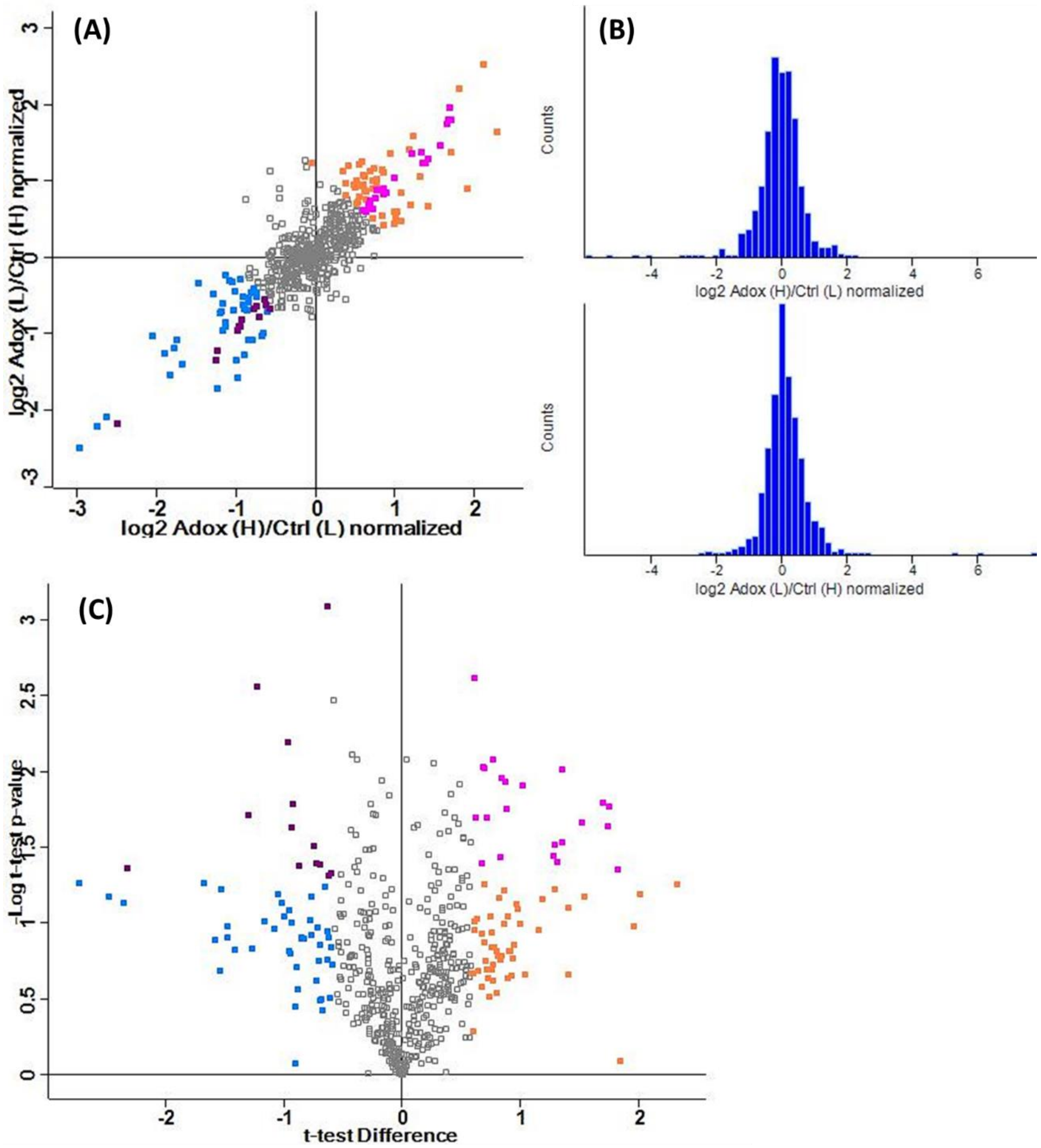


Experiment	Identified	Quantified
AdOx (H), ctrl (L)	3262	2813
AdOx (L), ctrl (H)	3129	2718

Figure 6.7 | Analysis of the effects of AdOx on the proteome by quantitative mass spectrometry. (A) Scatter plot comparing the normalised $\log_2$ fold change upon AdOx treatment with label-swap biological replicates (two data sets). Each dot represents one protein group. **(B)** Histogram showing the normalised ratio distribution, with the count of each bin representing the number of protein groups. **(C)** Volcano plot showing log-fold changes in peptide intensities on the x axis and p values on the y axis. In **(A)** and **(C)**, protein groups with over 1.5 fold increase are coloured orange, of which t-test significant coloured pink; protein groups with over 1.5 fold decrease are coloured blue, of which t-test significant coloured red.

Quantitative proteomic analysis of the *in vitro* binding to RNA GGGGCC repeats

Following SDS-PAGE analysis of the RNase eluted fraction from the *in vitro* RNA pulldown assays performed in the presence and absence of AdOx, the MS analysis identified close to 800 protein groups in each SILAC label-swap replicate, and roughly 85% of which were quantified (566 protein groups were quantified in both forward and reverse labelled samples). The results are summarised in Figure 6.8. Protein groups $\log_2$ fold values distribute evenly around 0 (Figure 6.8B), indicating good consistency of the label-swap replicates and the *in vitro* RNA pulldown assays. The SILAC label swap experiment shows a number of proteins were either enriched or depleted in the RNA GGGGCC repeat interactome upon the addition of the methylation inhibitor AdOx (Figure 6.8A). 73 protein groups were identified with increased $(GGGGCC)_5$ binding (over 1.5 fold change in protein abundance, $\log_2$ value > 0.585), and 22 of which are t-test significant (p value 0.05). The MS analysis identified 57 protein groups with decreased mRNA binding (with over 1.5 fold change in protein abundance, $\log_2$ value < -0.585), and 13 of which are t-test significant (Figure 6.8A and C).



Experiment	Identified	Quantified
AdOx (H), ctrl (L)	738	633
AdOx (L), ctrl (H)	792	666

Figure 6.8 | Effects of AdOx on *in vitro* GGGGCC repeat interactome analysis by quantitative mass spectrometry. (A) Scatter plot comparing the normalised $\log_2$ fold change of AdOx treatment with label-swap biological replicates (two sets of data). Each dot represents one protein group. **(B)** Histogram showing the normalised ratio distribution, with the count of each bin representing the number of protein groups. **(C)** Volcano plot showing log-fold changes in peptide intensities on the x axis and p values on the y axis. In **(A)** and **(C)**, protein groups with over 1.5 fold increase are coloured orange, of which t-test significant coloured pink; protein groups with over 1.5 fold decrease are coloured blue, of which t-test significant coloured purple.

In addition to the above GeLC-MS analysis, I performed in-solution digestion of the eluted proteins prior to MS analysis and compared to the previous in-gel digestions (see Figure 6.9). The results show that the number of protein groups quantified by in-gel digestion almost doubled compared to those quantified by in-solution digestion (both the results from heavy and light labelling were considered). This may partly due to the protein degradation for samples used for the in-solution digest. Despite the number of protein groups quantified, results from both the in-gel digestion and the in-solution digestion show similarities regarding to the protein abundance changes (i.e., increase or decrease under the effect of AdOx). Full protein lists with fold changes can be found in Chapter 6 Appendices.

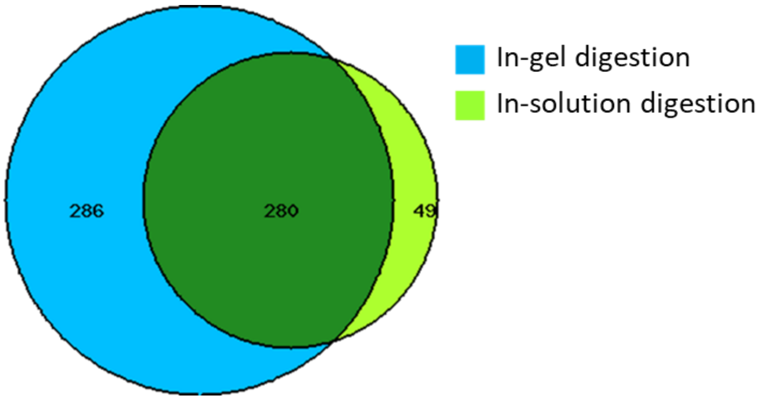


Figure 6.9 | Comparison of the number of protein groups quantified in both label-swap replicates by in-gel digestion and in-solution digestion.

From the SILAC MS analysis results, AdOx treatment not only changes protein abundance in the RNA pulldowns, but also the whole cell protein expression. Thus it is necessary to confirm difference in abundance in binding to GGGGCC repeat is a consequence of alterations in the interaction with $(GGGGCC)_5$ RNA oligos and not simply a change in the amount of protein in the cell extract used in the *in vitro* RNA pulldown assay. To do this, the ratios (AdOx treatment over control) of proteins from the RNA pulldown eluates were normalised to those of the WCE. However, not all proteins identified in the mRNA interactome were identified in the WCE, in which case, no normalisation was applied. The post-normalisation to WCE results are shown in Figure 6.10. In summary, the results do not alter the majority of proteins that were shown to differentially bind to RNA oligo $(GGGGCC)_5$ with the addition of AdOx following normalisation to protein expression changes in the WCE.

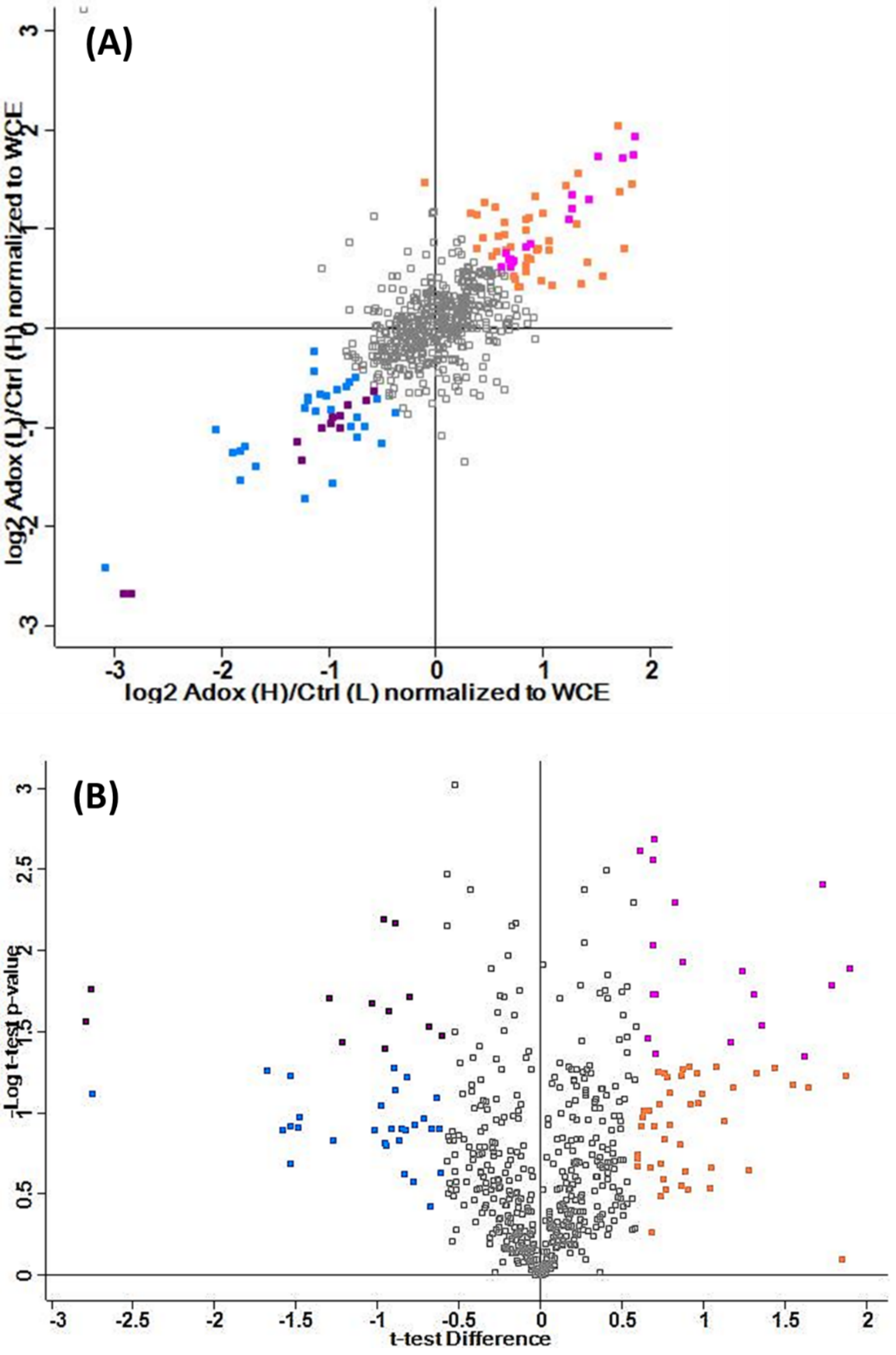


Figure 6.10 | Effects of AdOx on GGGGCC repeat interactome including normalisation to expression changes observed in the cells. (A) Scatter plot comparing the normalised $\log_2$ fold change of AdOx treatment with label-swap biological replicates (2 sets of data). Each dot represents one protein group. **(B)** Volcano plot showing log-fold changes in peptide intensities on the x axis and p values on the y axis. Protein groups with over 1.5 fold increase are coloured orange, of which t-test significant coloured pink; protein groups with over 1.5 fold decrease are coloured blue, of which t-test significant coloured purple.

Table 6.1 summarises those proteins with RNA (GGGGCC)$_5$ binding increased upon the AdOx treatment. Results from in-gel and in-solution digestion were taken into consideration at this stage. These proteins are not necessarily t-test significant, but are generally of over 1.5 fold increase based on two label-swap replicates.

Table 6.1 | Proteins that showed increased binding to GGGGCC repeats upon treatment with AdOx.

Proteins	Log$_2$ Ratio AdOx/Ctrl (Average)	Log$_2$ Ratio normalised to WCE (Average)	Known methylation sites*		Methylation sites (UniProt)	RGG/RG sites	Cross ref. **
ALKBH2	0.87	0.87				RG	
BCCIP	1.05	0.99	R × 2				
C14orf166	0.64	0.7	R × 1	K × 1		RG	
CAPRIN1	0.89	0.9	R × 12	K × 3	R626, R633, R640, R698	RGG, multiple RG	Y
CCAR1	1.05	<0.585	R × 1			RG	
CPSF6	1.06	1.08	R × 5			GR repeats, RG	Y
CRNKL1	1.96	1.04	R × 1	K × 1		RG	
DDX1	0.68	0.7	R × 2	K × 2		RG	Y
DDX17	1.17	<0.585	R × 17	K × 2	R684	Multiple RGG/RG	Y
DDX39B/A	1.25	1.22	R × 4			GRG	
EPB41	0.87	0.92	R × 2	K × 2		RG	
EWSR1	0.7	<0.585	R × 35		29 R methylation sites	Multiple RGG/RG	Y
FAM98B	0.61	0.81				RG	Y
FXR1	0.95	0.93	R × 9	K × 1		RGG, GRG, GR repeats	Y
FXR2	0.64	0.76	R × 8			GR repeats, RG	Y
G3BP1	1.06	0.83	R × 9	K × 2		RGG/RG	Y
GNB2L1	0.77	0.8	R × 4	K × 4		RG	Y
HEXIM1	1.21	<0.585	R × 4	K × 1			
HNRNPA2B1	0.75	1	R × 13	K × 1	R203, R213, R228, R238, R266, R325, R350	Multiple RGG/RG	Y

HNRNPC	1.53	1.87	R × 5	K × 3		RG	Y
HNRNPDL	0.99	1.09	R × 7	K × 2	R25	RG repeats	Y
HNRNPK	0.77	0.68	R × 26	K × 3	R316	Multiple RGG/RG	Y
HNRNPL	0.71	0.83	R × 8			RG	Y
HNRNPM	1.08	1.02	R × 13	K × 1	R496	RGG/RG	Y
HPSE	1.04	1.04				RG	
HSD17B4	1.75	1.62				RGG/RG	
KHSRP	0.98	0.88	R × 11	K × 3	R411, R413, R415, R442	Multiple RGG, GR repeats	Y
KIFC1	0.96	0.74	R × 2	K × 1		RG	
LUC7L2	1.59	1.45	R × 1			GR	
LUC7L3	1.79	1.58	R × 1			RG	
METTL3	2.32	1.79	R × 1			RG	
MPLKIP	2.05	2.05	R × 11		R57, R59, R68, R77, R117	RG	Y
NIP7	1.55	1.55	R × 2				
NOP2	1.02	<0.585	R × 1	K × 1		RG	Y
NUDT21	0.6	0.82	R × 2	K × 1	R15	RG	Y
OTUD4	0.7	0.66	R × 5	K × 4		RGG/RG	Y
PABPN1	0.82	0.87	R × 19		R259, R263	GRG/RG	Y
PAK1IP1	1.36	1.36				RG	
PRMT5	1.41	1.31	R × 1			RG	
PRPF4	1.25	0.86	R × 1				
RBM16	0.78	<0.585	R × 14		R917, R927, R938	RG	
RBM26	0.77	1.08	R × 3	K × 1		Multiple RG repeats	
RBM39	0.98	0.85	R × 7	K × 1		RG	Y
RBM4	0.88	<0.585	R × 2			RG	Y
RCC1	1.37	1.41	R × 1			RGG/RG	Y
RRP9	1.18	1.18	R × 1		R10	RG	Y
SAFB	1.74	1.73	R × 16	K × 3	R811, R868, R874, R884	RGG, GRG,RG	Y
SART1	0.6	0.96	R × 2	K × 2		RG	
SART3	2.01	1.87	R × 6	K × 1	R906	RG	Y
SCAF11	0.76	0.76	R × 10	K × 3	R1151	RG repeats	Y
SF3B1	0.91	0.99	R × 6	K × 1		RGG/RG	
SRSF10	0.81	0.81	R × 4			RG	Y
SRSF7	0.68	1.13	R × 3			RG	
SSRP1	0.94	0.79	R × 4	K × 2		RG	Y

TRA2A	0.96	0.96	R × 3			RG	
TRA2B	1.14	1.38	R × 10		R241	RGG/RG	Y
U2AF2	1.01	1	R × 2	K × 1		RG	
UBAP2L	2.21	2	R × 8	K × 1	R187, R190	GRG, RGG, RG repeats	Y
USP10	0.99	0.78				RG	
ZFP91	1.35	1.9	R × 3	K × 1		RGG/RG	

*Based on iPTMnet database. R: arginine methylation sites; K: lysine methylation sites.
**Cross reference to the list of unique identified proteins with methylations by Geoghegan et al., 2015. Y: the protein is on the list.

From Table 6.1, most proteins with over 1.5 fold increase ($\log_2 > 0.585$) were identified with RGG/RG motifs that can be potentially modified, and/or have known methylation sites based on the iPTMnet database. Proteins identified with a large fold increase from the enrichment of GGGGCC repeat binding assay upon AdOx treatment are described in the section below.

METTL3 is an essential member of the N^6-methyltransferase complex, although not likely to be self-methylated, but it is involved in the internal adenosine residue post-transcriptional methylations in eukaryotic mRNA (Yue et al., 2015). UBAP2L has an RGG/RG motif in the N terminus which can be directly methylated by PRMT1, and is required for the accurate distribution of chromosomes (Maeda et al., 2016). CRNKL1 is known involved in pre-mRNA splicing (Chung et al., 2002), but lacking methylation sites on itself. SAFB is a protein with c-terminal RG rich region and have multiple sites for methylation. Apart from the proteins discussed above, a number of heterogeneous nuclear ribonucleoproteins (hnRNPs) were also identified enriched in elutes of GGGGCC binding assay under demethylation environment, including HNRNPC, HNRNPK, HNRNPDL, HNRNPM and HNRNPA2B. Other proteins such as MPLKIP, SART3, LUC7L3 and PRMT5 may also be proved interesting.

Similarly, a number of proteins that were decreased in abundance in the oligo RNA $(GGGGCC)_5$ interactome when treated with AdOx are listed in Table 6.2.

Table 6.2 | Proteins that showed decreased binding to GGGGCC repeats upon treatment with AdOx.

Proteins	Log$_2$ Ratio AdOx/Ctrl (Average)	Log$_2$ Ratio normalised to WCE (Average)	Known methylation sites*		Methylation sites (UniProt)	RGG/RG sites	Cross ref.**
ACAD11	-1.17	>-0.585	R × 2	K × 1		RG	
ACTL6A	-1.98	-1.41	K × 1				
AGAP3	-0.95	-0.95	K × 1			RG	Y
ALYREF	-1.86	-1.86	R × 15	K × 1	R38, R58, R63, R71, R204, K235	Multiple RGG, RG repeats	Y
ANGEL2	-0.86	-0.86	R × 4	K × 1		GRG/RG	Y
BCKDK	-0.85	-0.85	K × 1			RGG	
CDKN2AIPNL	-1.58	-1.58					
CKAP4	-1.41	-0.98	R × 1	K × 1		RG	
CKMT1A	-0.62	-0.8	R × 3			RG	
CLP1	-1.27	-1.27				RG	
CMAS	-0.84	-0.84	R × 4	K × 1		RGG/RG	Y
DIMT1	-0.99	-0.99	K × 1			RG	
DISC1	-1.08	-1.22	R × 2	K × 2		RGG/RG	Y
DKC1	-0.7	-1.12					
EED	-1.48	-1.48	K × 4		K66, K197, K268, K284	RG	
EEF1G	-0.63	-0.72	R × 2	K × 1		RG	
HMMR	-1.48	-1.48				RG	
HNRNPR	-0.76	>-0.585	R × 12	K × 5		Multiple RGG, RG repeats	Y
HNRNPU	-0.68	>-0.585	R × 22	K × 5	R702, R733, R739, R755, R762	Mutiple RGG/RG	Y
HSPA1B	-1.06	-1.03	R × 1	K × 1		RG	
HSPA8	-1.3	-1.21	R × 2	K × 7	R469, K561	RG	Y
KCTD12	-1.05	-0.71	R × 1			RG	
LSM14A	-1	-1.03	R × 6	K × 1		Multiple RGG/RG repeats	Y
MAZ	-0.81	-0.81	R × 2			RG	
MPG	-1.23	-1.23				RGG/RG	
NARS	-1.21	-1.25					
NSUN2	-2.65	-2.97	R × 2			RG	
PA2G4	-2.38	-2.34	R × 3	K × 3			

PCF11	-1.3	-1.3	R × 31		R805, R820, etc, 11 R sites	RG	Y
POLDIP3	-0.73	-0.9	R × 3	K × 4	R33	RG	
PSIP1	-1.19	-0.61	K × 1			RG	
PUS1	-1.39	-1.08					
PUS7	-2.73	-1.53	R × 4			RG	Y
PXDN	-0.96	-0.96	R × 1			RG	
RCC2	-0.63	-0.82	R × 4	K × 3		RGG, GRG	Y
RPL12	-1.09	-0.8	R × 2				
SMARCA1	-1.54	-1.54				RG	
SNRPD1	-1.2	-1.21	R × 10			RGG, GR repeats	
SNRPN	-0.91	>-0.585	R × 8		R172	GRG/RG	Y
SRP14	-0.79	-0.92	K × 2				
SSB	-0.92	-1.01	R × 1	K × 1		RG	
TRMT2B	-1.53	-1.53				RG	
TRMT6	-2.7	-3.09	R × 1	K × 1		RGG/RG	
TRMT61A	-2.51	-2.94	R × 2			GRG, RG	
TROVE2	-1.05	-1.02	R × 1	K × 2		GRG/RG	
TSN	-0.93	-0.95	R × 1			RG	
XRN2	-1.22	-1.22	R × 18	K × 2	R824, R847, R851, R883, R895, R946	RGG/RG	Y
YTHDC2	-0.96	>-0.585	R × 5	K × 3		RGG/GRG/RG	Y

*Based on iPTMnet database. R: arginine methylation sites; K: lysine methylation sites.
**Cross reference to the list of unique identified proteins with methylations by Geoghegan et al., 2015. Y: the protein is on the list.

As shown in Table 6.2, most proteins with a decrease of over 1.5 fold ($\log_2 < -0.585$) were identified with RGG/RG motifs as well as known methylation sites based on the iPTMnet database. Proteins identified with a large fold change decrease from the enrichment of GGGGCC repeat binding assay under AdOx treatment are described in the section below.

Two tRNA adenine methyltransferase subunits, TRMT6, and TRMT61A, were identified amongst the largest fold decrease proteins. TRMT6 catalyses the methylation of m^1A of initiator methionyl-tRNA at position 58 (Finer-Moore et al., 2015), while TRMT61A, as part of the mRNA N1-methyltransferase complex, mediates N1 adenosine methylation of a subset of mRNAs (Safra et al., 2017). Also, a tRNA uracil methyltransferase homolog, TRMT2B was identified, which is known to catalyse the 5mU modification in tRNA, and may also play roles in tRNA stabilisation and maturation (UniProt resource). NSUN2 is another tRNA methyltransferase identified, which catalyses the m^5C methylation (Brzezicha et al., 2006). Two pseudouridylate synthase (PUS1 and PUS7) were identified among the proteins with fold decrease. These proteins convert uridine in RNA to pseudouridine, and may contribute the stabilisation of RNA secondary and tertiary structures (see chapter 1.3.3). The lack of potential methylation sites in PUS1 suggests this protein may not respond to demethylation environment directly. Two hnRNPs were identified with decreased amount, HNRNPU and HNRNPR, but the fold changes were not significant, especially after the normalisation to WCE.

Other proteins such as PA2G4, YTHDC2, PCF11, LSM14A, SNRPD1 and XRN2 may also be proved interesting. It is also worth noting that ALYREF was found amongst the decrease group of proteins ($\log_2$ fold change -1.86), which is opposite to the mRNA binding results (see chapter 5.3.4).

Following quantitative MS analysis, additional western blotting experiments were carried out on three proteins known to bind to RNA GGGGCC repeats (see Figure 6.11). Results show AdOx treatment decreased the amount of ALYREF binding to the GGGGCC, which is in line with the MS data ($\log_2$ AdOx/Ctrl -1.86). For HNRNPA1, no significant change was seen in the elutes from western blotting, which is consistent with the MS quantification where the fold change of HNRNPA1 was less than 1.5. For HNRNPK, however, no signal was detected using western blotting.

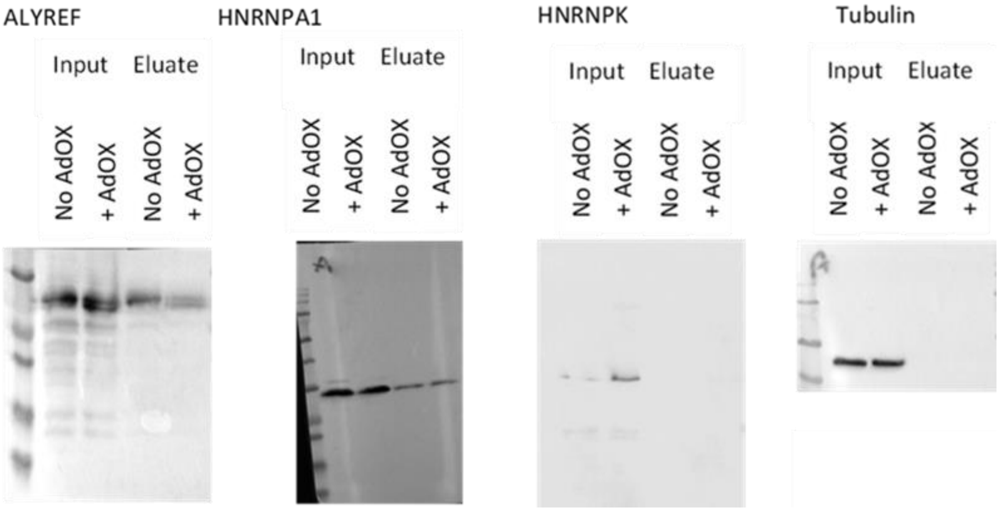


Figure 6.11 | Western blotting results of known GGGGCC repeat binding proteins in cell lysate (input) and RNA pulldown (eluate). Three proteins that known to bind to RNA GGGGCC repeats were tested, with tubulin used as input control. Proteins were from cells either treated with AdOx or without. See text for details.

6.3.3 Studying the effect of citrullination on the GGGGCC repeat interactome

To study the effect of protein citrullination on the binding to RNA GGGGCC repeats, cells were transfected with HA-PADI4, the enzyme catalyses citrullination of arginine (simplified as PADI4 treatment hereafter). Induced expression of PADI4 were verified by SDS PAGE and western blotting (see Figure 6.12).

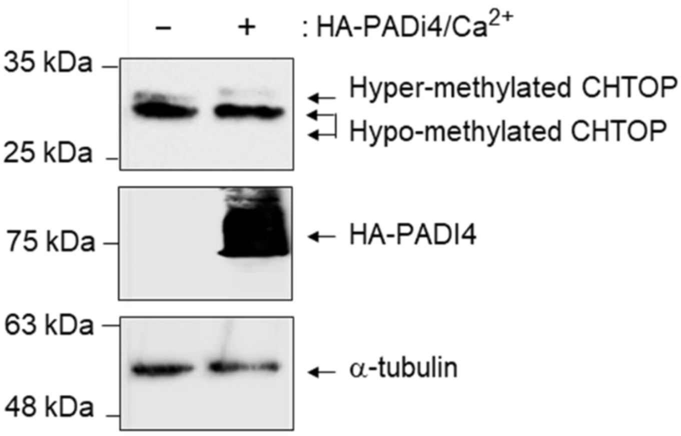


Figure 6.12 | Western blotting of the HA-PADI4 transfected cells and the control cells. WCEs of both the cell lines were analysed with CHTOP and HA antibody. A reduced level of hyper-methylated CHTOP was observed, indicating the inhibition of CHTOP methylation in the HA-PADI4 transfected cells. A significant amount of HA-PADI4 was observed in the HA-PADI4 transfected cells, which was absent in the control cells, indicating the success of the HA-PADI4 transfection. α-tubulin was used as input control.

For the SILAC experiments, cells were cultured in heavy and light media with a label-swap strategy for biological replicates (using the optimised SILAC growths as previously described). *In vitro* RNA pulldown assays were performed separately using up to 2 mg (each) of light and heavy cell lysate, prior to binding to $(GGGGCC)_5$, in conjunction with UV crosslinking and affinity purification using streptavidin beads. The recovered eluates of both heavy and light were mixed (1:1 protein amount) prior to analysis on SDS-PAGE. The results are shown in Figure 6.13, demonstrating the successful enrichment of proteins binding to the GGGGCC repeat oligos.

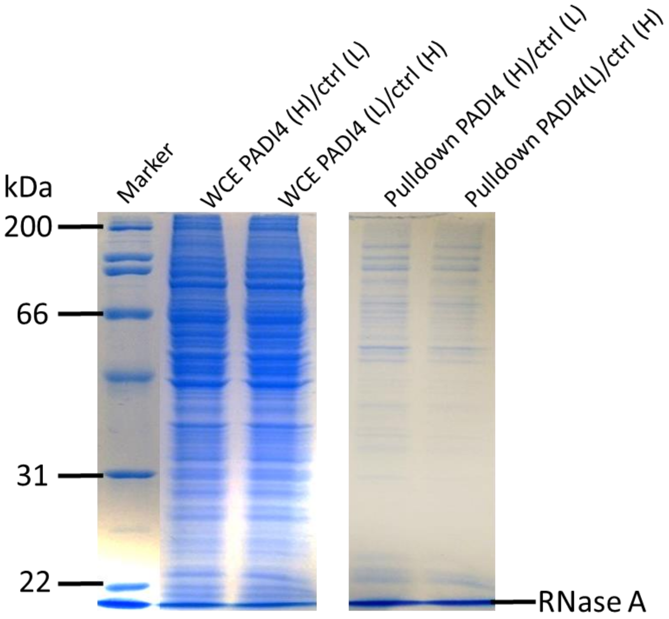
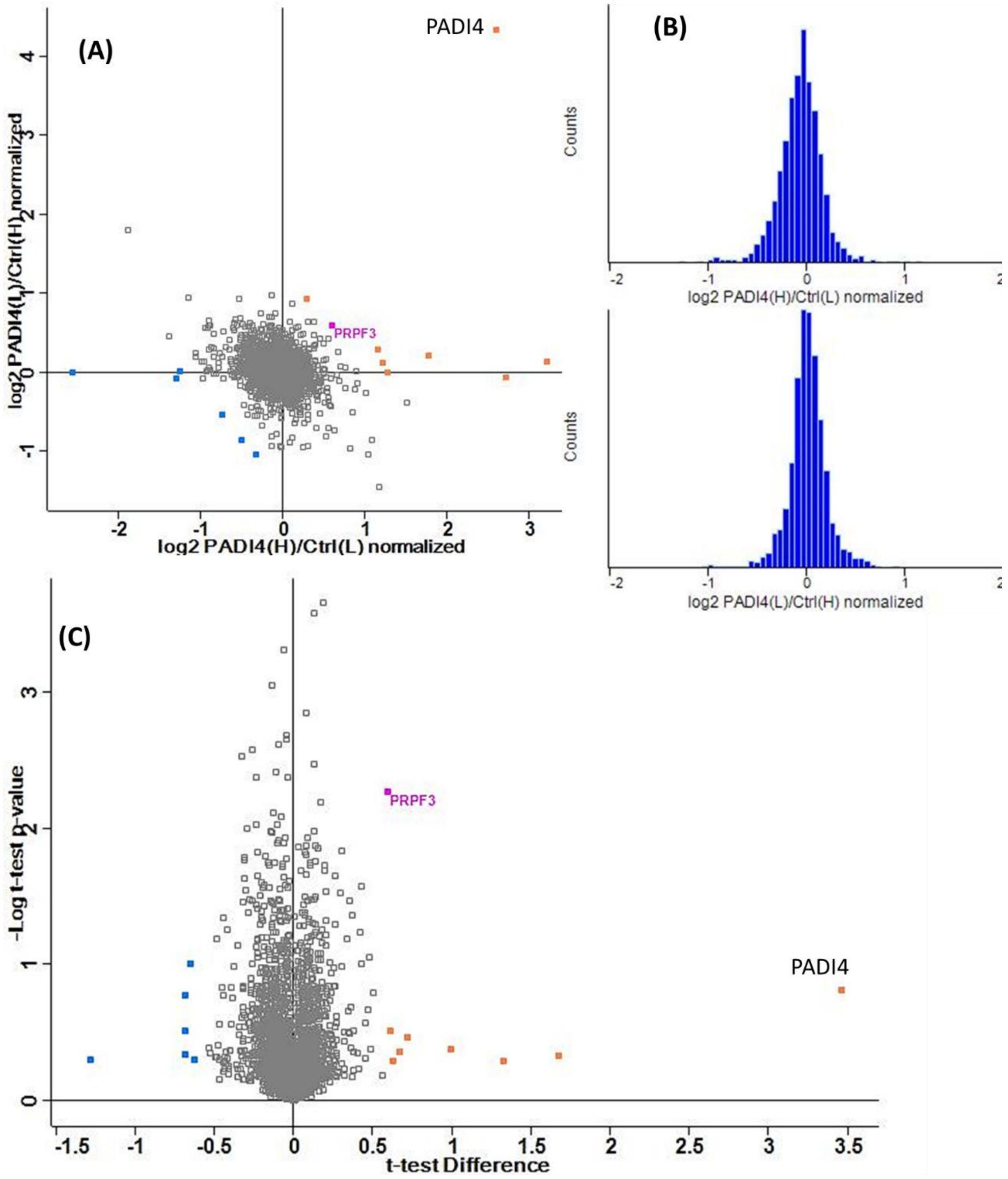


Figure 6.13 | SDS-PAGE analysis of whole cell extracts (WCE) and *in vitro* RNA pulldown eluates for the PADI4 transfected cells (PADI4) and the wild type cells (ctrl). HEK 293T cells were grown in SILAC media with label-swap biological replicates, and crosslinked at UV 254 nm. For both WCE and pulldowns, samples were mixed with estimated 1:1 protein amount heavy to light prior to loading on the gel.

Quantitative analysis of the whole cell protein expression

Following SDS-PAGE analysis, WCE were in-gel digested and analysed on a Q Exactive HF MS system to determine the effect of PADI4 on the expression of a range of different proteins (as described in previously). Raw data were processed using MaxQuant, and statistical analysis was performed in Perseus. For WCE, over 3600 protein groups were identified in each replicate and over 3200 of which were quantified. Results are shown in Figure 6.14. Protein groups $\log_2$ fold values distribute evenly around 0 (Figure 6.14B), indicating good consistency of the label-swap replicates. Quantitative proteomic analysis revealed that the majority of proteins were not affected by the PADI4 overexpression (fold change of less than 1.5, Figure 6.14A and C). These results are in contrast to the effects of AdOx on the proteome, likely due to that PADI4 is an isotype of PADs family and have limited biological roles, or limited substrates *in vivo*. Moreover, the SILAC proteomic analysis demonstrated the successful overexpression of PADI4 (>10 fold increased expression, Figure 6.14A and C). Other protein groups that are increased with the overexpression of PADI4 include SUN2, SRSF11 and PRPF3. These results will be used in conjunction with quantitative *in vitro* $(GGGGCC)_5$ pulldown assays described in the section below.



Experiment	Identified	Quantified
PADI4 (H), ctrl (L)	3640	3231
PADI4 (L), ctrl (H)	3704	3240

Figure 6.14 | Analysis of the effects of PADI4 overexpression on the proteome by quantitative mass spectrometry. (A) Scatter plot comparing the normalised $\log_2$ fold change upon PADI4 overexpression with label-swap biological replicates (two data sets). Each dot represents one protein group. **(B)** Histogram showing the normalised ratio distribution, with the count of each bin representing the number of protein groups. **(C)** Volcano plot showing log-fold changes in peptide intensities on the x axis and p values on the y axis. In **(A)** and **(C)**, protein groups with over 1.5 fold increase are coloured orange, of which t-test significant coloured pink (PRPF3); protein groups with over 1.5 fold decrease are coloured blue.

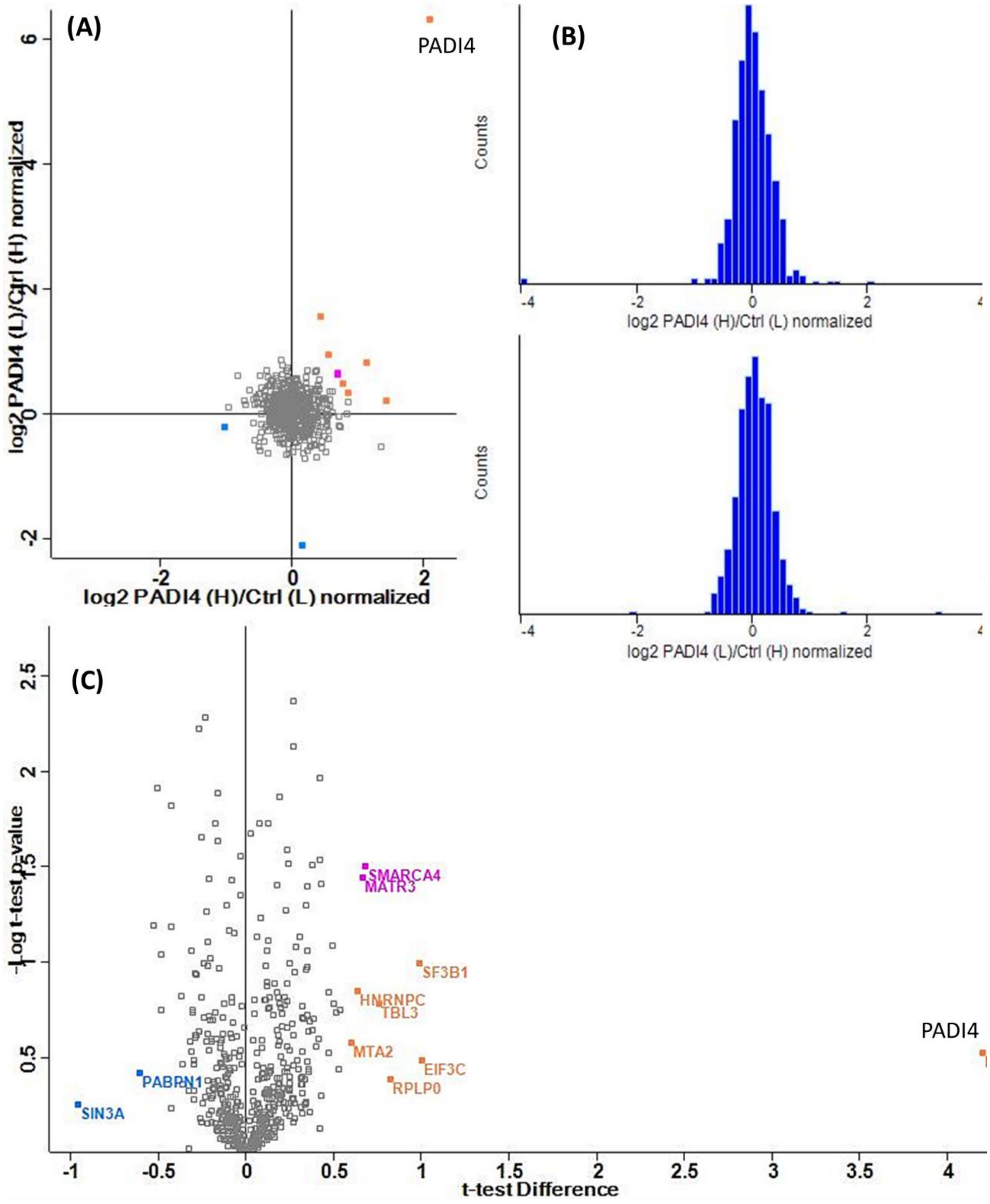
Quantitative proteomic analysis of the *in vitro* binding to RNA GGGGCC repeats

Following SDS-PAGE analysis of the eluted fraction from the *in vitro* $(GGGGCC)_5$ pulldown assays performed in both wild type and PADI4 overexpression cells, the MS analysis identified over 700 protein groups in each SILAC label-swap replicate, over 80% of which were quantified. The results are summarised in Figure 6.15. Good consistency is seen for the two sets of data (label-swap replicates) from the *in vitro* $(GGGGCC)_5$ pulldown assays, as protein group $\log_2$ fold change values distribute evenly around 0 (Figure 6.15B). The SILAC label-swap experiment showed that upon the overexpression of PADI4, the abundance of most proteins in the $(GGGGCC)_5$ pulldown were not significantly affected (Figure 6.15A and C). However, there are a small number of proteins of potential interest identified with over 1.5 fold increase in abundance ($\log_2$ value > 0.585) for $(GGGGCC)_5$ binding, which are listed in Table 6.3.

Table 6.3 | Proteins that showed increased binding to GGGGCC repeats upon the PADI4 overexpression.

Proteins	$\mathbf{Log_2}$ PADI4/Ctrl average	Normalized to WCE average
SMARCA4	0.68	0.91
MATR3	0.67	0.54
HNRNPC	0.64	0.52
EIF3C	1.00	1.13
SF3B1	0.99	1.05
MTA2	0.60	0.73
TBL3	0.76	0.76
PADI4	4.21	0.74

From table 6.3, PADI4 is of the largest fold change before normalised to WCE. When taking the amount of PADI4 in the input (WCE), the fold change is largely reduced, indicating the increase of PADI4 in pulldown is mainly due to the increase in input. All other proteins that showed increased binding to GGGGCC repeats are not specifically recognized as PADI4 substrates, thus further verifications and studies are required.



Experiment	Identified	Quantified
PADI4 (H), ctrl (L)	777	662
PADI4 (L), ctrl (H)	710	590

Figure 6.15 | Effects of PADI4 on *in vitro* GGGGCC repeat interactome analysis by quantitative mass spectrometry. (A) Scatter plot comparing the normalised $\log_2$ fold change upon PADI4 overexpression with label-swap biological replicates (two data sets). Each dot represents one protein group. **(B)** Histogram showing the normalised ratio distribution, with the count of each bin representing the number of protein groups. **(C)** Volcano plot showing log-fold changes in peptide intensities on the x axis and p values on the y axis. In **(A)** and **(C)**, protein groups with over 1.5 fold increase are coloured orange, of which t-test significant coloured pink (PRPF3); protein groups with over 1.5 fold decrease are coloured blue.

As described in previous sections, the protein abundance change in the RNA pulldown eluates may due to the amount changed in input. The amount of PADI4 for example, is largely in excess in the WCE of PADI4 induced cells compared to wild type, which may result in the fold difference observed in the RNA pulldowns. Taking this into consideration, the ratios (PADI4 treatment over control) of proteins from the *in vitro* pulldown assay were normalised to those of WCE. In case of proteins which were not identified in the WCE, no normalisation was applied. The post-normalisation results are shown in Figure 6.16, and proteins with increase over 1.5 fold ($\log_2$ value > 0.585) are shown in Table 6.3.

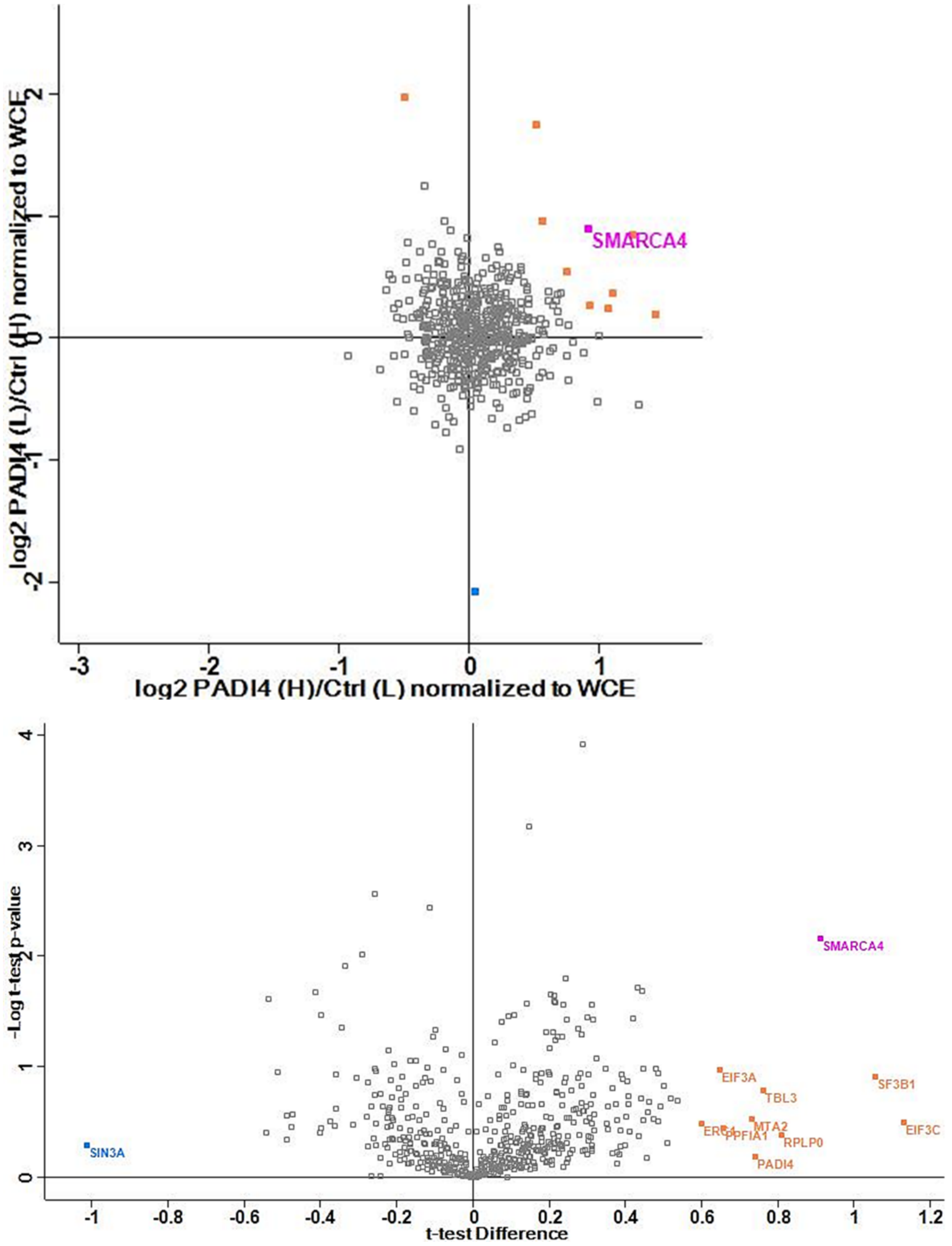
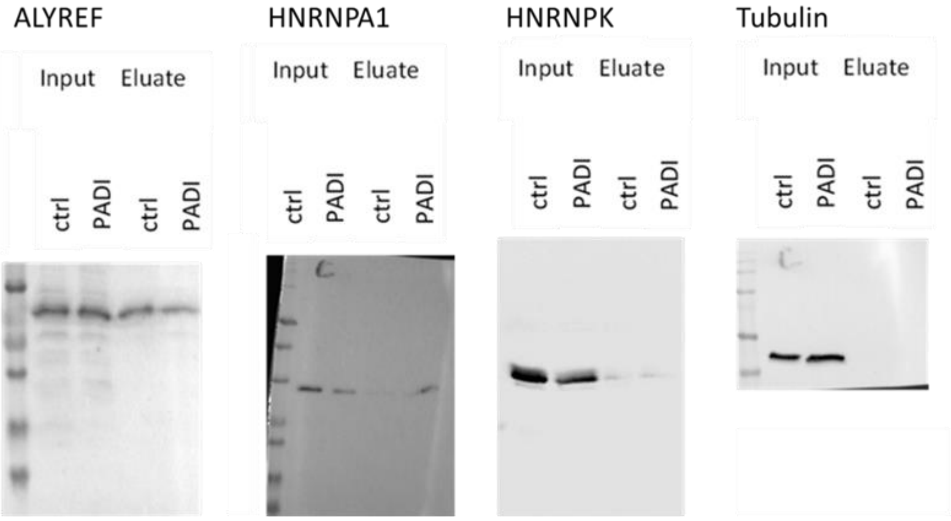
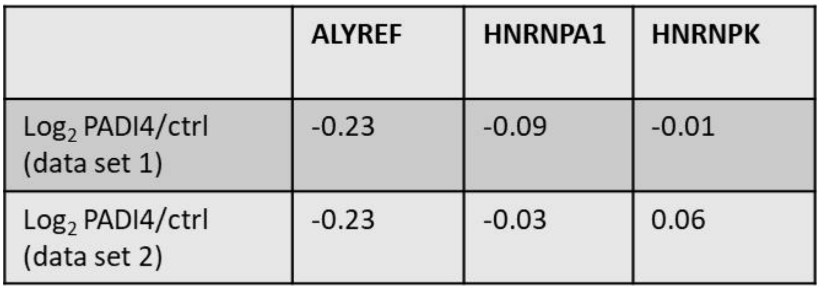


Figure 6.16 | Effects of PADI4 on GGGGCC repeat interactome including normalisation to expression changes observed in the cells. (A) Scatter plot comparing the normalised $\log_2$ fold change upon PADI4 overexpression with label-swap biological replicates (two data sets). Each dot represents one protein group. **(B)** Volcano plot showing log-fold changes in peptide intensities on the x axis and p values on the y axis. Protein groups with over 1.5 fold increase are coloured orange, of which t-test significant coloured pink; protein groups with over 1.5 fold decrease are coloured blue.

Following MS analysis, additional western blotting experiments were performed to validate the quantitative MS results (see Figure 6.17). The results show that for ALYREF, a small decrease in abundance in the eluate from the binding to GGGGCC repeats was observed upon PADI4 treatment, which is in line with the MS data ($\log_2$ ratio -0.23). Western blotting results also show a small increase in HNRNPA1 in the eluate from the binding to GGGGCC repeats with PADI4 treatment. This is in contrast to the MS analysis where no significant fold change was observed. No difference in abundance is observed for HNPNPK from both the western blotting and the MS results.





	ALYREF	HNRNPA1	HNRNPK
Log_2 PADI4/ctrl (data set 1)	-0.23	-0.09	-0.01
Log_2 PADI4/ctrl (data set 2)	-0.23	-0.03	0.06

Figure 6.17 | Western blotting analysis of selected proteins of known binding to GGGGCC repeats. The data are shown for using antibodies to analyse ALYREF, HNRNPA1 and HNRNPK in both cell lysates (input) and RNA pulldowns (eluate). Tubulin was used as input control. Table shows the quantitate MS results (two sets of data from the label-swap biological replicates) for comparison.

6.4 Conclusions

To understand the mechanism of the GGGGCC expansion repeats in amyotrophic lateral sclerosis and familial frontotemporal degeneration, quantitative proteomic experiments have previously been used to identify the binding partners to the repeats in a number of different studies. In this study the aim was to further understand how the interactions between these RNA binding proteins and GGGGCC repeats are affected by protein post-translational modifications, in particular arginine methylation and citrullination.

Large scale *in vitro* pulldown assays were performed using a 3'-biotinylated RNA $(GGGGCC)_5$ oligo, which was UV crosslinked to protein extracts from human embryonic kidney cells and affinity purified using streptavidin beads. Binding proteins were then recovered by the addition of RNase prior to GeLC-MS analysis. Further quantitative proteomic analysis was performed using *in vitro* pulldown assays in conjunction with stable isotope labelling with amino acids in cell culture to determine the effects of the global methylation inhibitor AdOx and the effects of PADI4 (arginine citrullination) on RNA oligo $(GGGGCC)_5$ binding.

To study the effects of protein methylation on $(GGGGCC)_5$ binding, GeLC-MS analysis identified over 700 proteins, of which over 600 were quantified in the SILAC label swap experiment. 73 proteins showed increased RNA binding and 57 showed decreased RNA binding (with > 1.5 fold change in protein abundance). These results are the first global study to demonstrate that protein methylation of a wide range of RNA binding proteins affects the interaction with $(GGGGCC)_5$ repeats. Moreover, further analysis revealed a wide number of identified proteins contain potential sites of arginine methylation which are of particular interest and consistent with the global quantitative analysis. Proteins with fold change increase include METTL3, UBAP2L, SAFB, MPLKIP, SART3, as well as several hnRNPs, while proteins with fold change decrease include some tRNA methyltransferase, pseudouridylate synthase, ALYREF, YTHDC2, PCF11, LSM14A, SNRPD1 and XRN2. Western blotting is required for further validation of these proteins. Meanwhile, further MS data interpretation is to be carried out to verify arginine methylation/demethylation sites of the proteins identified.

The quantitative analysis of the effects of citrullination was performed using the cell line overexpressing the enzyme PADI4. In contrast to the experiments performed in the presence of AdOx, only limited changes were observed on the proteome. In addition, only a small number of proteins were identified that altered their binding to $(GGGGCC)_5$ repeats. Potentially interesting proteins include SMARCA4, SF3B1, EIF3C and TBL3 should be further validated by western blotting. Although the overexpression of PADI4 was verified from both western blotting and SILAC MS analysis, the limited perturbation on the proteome in conjunction with only small changes in the protein binding to GGGGCC repeats might be due to inactivity of the enzyme and only limited protein citrullination in this cell line. Further work needs to be performed to verify citrullination of proteins in the cells expressing PADI4 and in particular citrullination of the above identified proteins.

These results are the first global quantitative analysis of the effect of methylation and citrullination on RNA GGGGCC repeat binding, and highlight a number of interesting proteins as candidates for further studies to examine the role of methylation/citrullination on GGGGCC binding *in vitro*. This study provides an opportunity for more detailed downstream analysis of the identified proteins which will provide further insight into the role that protein post-translational modifications play in binding to GGGGCC repeat in amyotrophic lateral sclerosis and frontotemporal dementia.

Chapter 7: Absolute quantification of recombinant protective antigen (rPA) in anthrax vaccine using stable isotope dilution LC ESI MS

7.1 Abstract

Anthrax vaccine is currently manufactured by Porton Biopharma who are the sole manufacturer of the UK's licensed anthrax vaccine. The active ingredient in the vaccine is a sterile filtrate of an alum precipitated anthrax antigen (recombinant protective antigen, rPA) in a solution for injection. The concentration of rPA in anthrax vaccine products needs to be accurately quantified, which is currently performed using ELISA based methods. However, there are a number of caveats to ELISA based quantification especially in complex mixtures which may interfere with the ELISA. Therefore, it was proposed to develop an alternative quantitative mass spectrometry method in conjunction with stable isotope dilution approaches as an alternative to ELISA analysis. In this Chapter I developed an MS based absolute quantification method (AQUA) to determine the absolute concentration of anthrax vaccine products. Absolute quantification was performed on 2 different MS instruments (Ion trap and UHR-TOF) and compared to previous quantitative analysis performed using ELISA.

7.2 Introduction

Anthrax is a serious infectious disease caused by a rod-shaped gram-positive bacterium named *Bacillus anthracis* (Dixon et al., 1999). Anthrax spores are usually found in soil, and can survive tough conditions including extreme temperature, drought, ultraviolet light, gamma radiation and many antimicrobial agents (Watson and Keir, 1994). Anthrax is spread mainly by contact with spores from soil, water and plants. Once spores reach the living environment, they develop into new organisms which cause infection. Most mammals are susceptible to this disease, especially grazing herbivores (Dixon et al., 1999). Similarly, humans who come into contact with the source of spores, e.g., infected animal products, can be infected. Human anthrax can develop in 3 forms: cutaneous (skin) anthrax, gastrointestinal anthrax and inhalation anthrax (Kamal et al., 2011). Although human infection by anthrax is relatively uncommon, systemic infection can lead to severe syndromes including death (Meselson et al., 1994). Thus prevention for this disease is necessary especially for people who are likely to be exposed to *Bacillus anthracis* bacteria.

Protective antigen (PA), a receptor-binding moiety, together with 2 effector moieties oedema factor (OF) and lethal factor (LF), are the 3 protein components of anthrax toxin secreted by *Bacillus anthracis* (Bradley et al., 2001). This tripartite toxin allows the bacterium to bypass the host immune system and cause infection. Briefly, PA can bind to cell-surface receptor (type I membrane protein), which mediates the entry of OF and LF into target cells (Milne et al., 1994; Molloy et al., 1992). To protect against this disease, vaccines are usually used. Human anthrax vaccines were first developed in the late 1930s in the Soviet Union (Shlyakhov and Rubinstein, 1994). Anthrax vaccines used currently are screened free of live *Bacillus anthracis* cells thus do not cause anthrax. The most promising vaccine for anthrax are subunit vaccines composed of purified recombinant protective antigen (rPA) (Keitel, 2006), which stimulates the immune system to produce relevant antibodies and block the toxin cell-entering pathway. In Britain, anthrax vaccines (alum precipitated sterile filtrate) containing active ingredient anthrax antigen are manufactured by Porton Biopharma Limited on behalf of the government.

Quantification of active ingredients in vaccines, as part of product quality control, is important in the vaccine manufacturing process. Available methods for analyte measurements include using either immunoassays (IAs) or mass spectrometry. For immunoassays, specific antibodies are used to recognize and bind to antigens that need to be analysed (or less commonly, use antigen to detect antibodies), meanwhile measurable signal, such as radiation, fluorescence or change of colour, are produced as a result of binding for quantification. IAs have been extensively used as diagnostic approaches for several decades. A typical IA that been widely used is enzyme-linked immunosorbent assay (ELISA). The key of an ELISA is the antibody-antigen interaction. Antigens are immobilised on a solid surface, and antibodies with linked enzyme can bind. Detection is through the measureable products produced from the activity of the conjugated enzyme. Mass spectrometric methods, on the other hand, often work together with chromatography for separation purposes, measure the amount of analytes based on mass-to-charge ratio of ionized peptides. LC-MS based quantification methods have been developed rapidly over the past decade, and can serve as golden standard for protein qualification and quantification analysis (Botelho et al., 2013). Unlike IAs which are limited by the availability of specific antibodies, mass spectrometric methods can be applied to any target protein. To get good reliability IAs need the consistent antibody properties, which is difficult. Even in commercial platform IAs from the same supplier, slight differences may occur for different batches of products. Another issue for IAs is sensitivity. When sample concentration is low, mass spectrometry is often the method of choice due to better sensitivity and accuracy (Ohlsson et al., 2013) and therefore less sample consumption. Furthermore, mass spectrometry methods can identify protein post-translational modifications, which generally are not detectable using IAs.

Here, an LC-MS based absolute quantification method (AQUA) using stable isotope labelling (Kirkpatrick et al., 2005) was developed to determine the concentration of anthrax vaccine products. A schematic illustration of the process is shown in Figure 7.1. In this method, synthetic peptides with stable isotopes incorporated (i.e., ^{13}C, ^{15}N, ^{18}O, ^{2}H) are spiked into the samples as internal standards, also termed as stable isotope-labelled standards (SIS) (Sturm et al., 2012). ^{13}C and ^{15}N are more commonly used as they do not alter peptide (not deuterated) chromatographic retention time (Kettenbach et al., 2011). Because SIS and their counterparts (proteolytic digested peptides from samples) are chemically indistinguishable with only molecular weight differences, the ratio of heavy and light peptides (H/L) given by mass spectrometry equals to ratio of amount. With the SIS amount already known, sample peptide as well as protein amount can thus be determined.

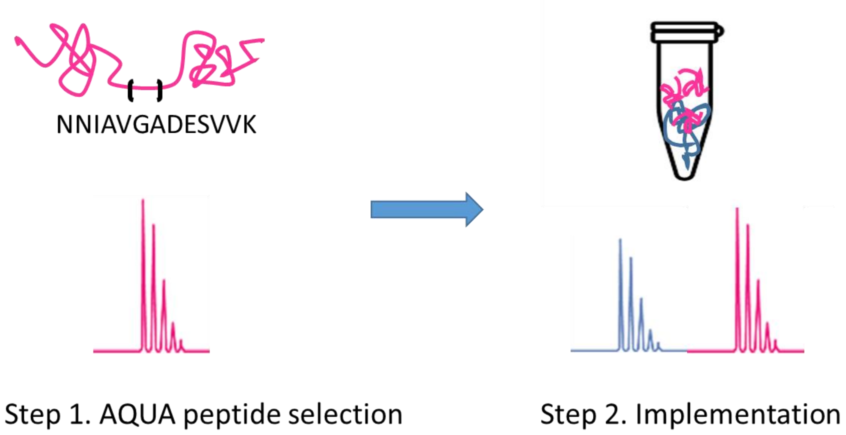


Figure 7.1 | A schematic illustration of the protein-AQUA method. In step 1, an optimal tryptic peptide featured in protein of interest was chosen and synthesised using heavy isotopes, then analytical method (LC-MS) was optimized to resolve and monitor both native and AQUA peptides. In step 2, a known amount of AQUA peptides were added to native protein sample, and digested together with trypsin. Digested sample was then analysed by LC-MS, based on which native peptides can be quantified.

7.3 Results and Discussion

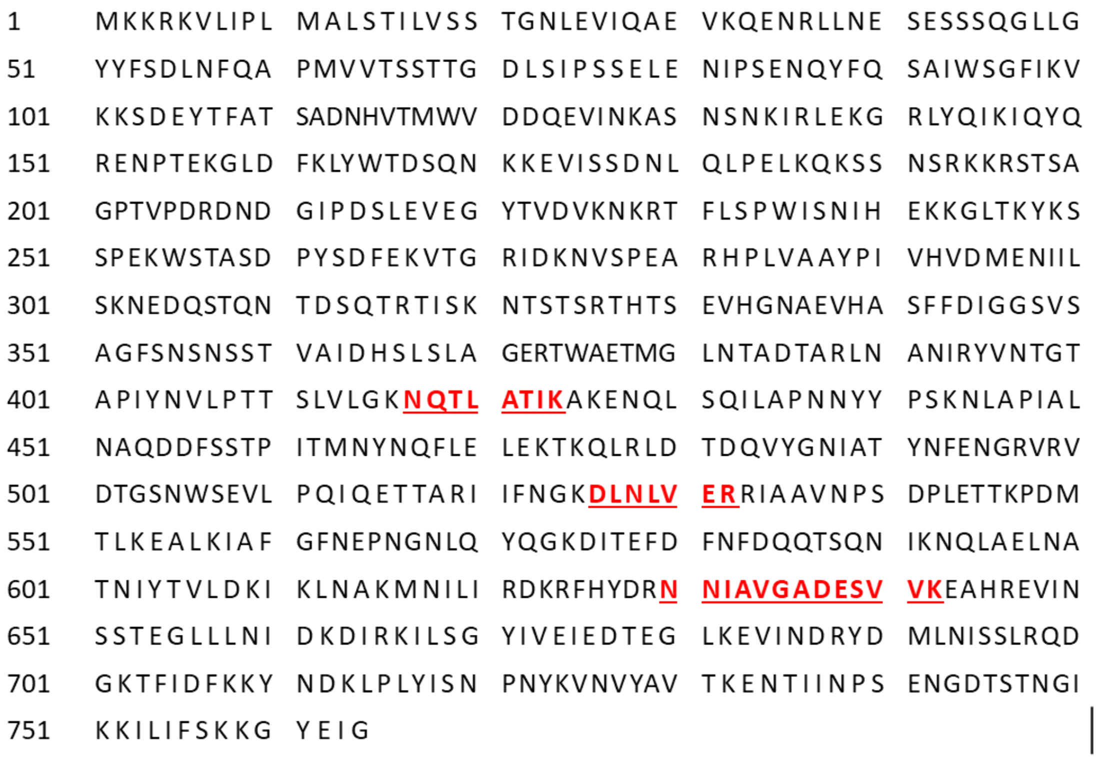
7.3.1 Preparation of proteotypic stable isotope-labelled standards

First, peptide sequences were selected for chemical synthesis of the internal standards (SIS). Rules apply for peptide selection (Kettenbach et al., 2011). Briefly, these tryptic peptides were designed in a way that they are unique to rPA, no overlapping m/z signal from peptide isoforms or modification forms, free from trypsin cleavage pattern change (avoid proline at carboxylic side or phosphorylation at amino-terminal side, avoid repeat arginine or lysine site that may cause ragged ends), no methionine, cysteine or tryptophan containing peptides which may be oxidized differently during sample preparation, and peptides chosen should be generally stable under sample preparation conditions and the mass spectrometry analysis environment. Initial studies were performed on tryptic digestion of rPA in conjunction with LC ESI MS analysis to determine potential proteotypic peptides. Based on the above parameters and the LS ESI MS data, 3 peptides (peptide A, B and C) were chosen for synthesis as stable isotope labelled peptides (see Figure 7.2 and Table 7.1).

Table 7.1 | List of peptides chosen for rPA quantification.

Peptides	Sequence	Light MW (Da) *	Heavy MW (Da) **
A	DLNL**V**ER	857.96	863.96
B	NNIA**V**GADESVVK	1315.44	1321.44
C	NQT**L**ATIK	888.02	895.02

The heavy isotope labelled amino acids are in bold.
*Molecular weights of light peptides were calculated based on sequences using GenScript website (www.genscript.com/tools/peptide-molecular-weight-calculator). **Molecular weights of heavy peptides are the theoretical value based on isotope labelling.



```
1     MKKRKVLIPL  MALSTILVSS  TGNLEVIQAE  VKQENRLLNE  SESSSQGLLG
51    YYFSDLNFQA  PMVVTSSTTG  DLSIPSSELE  NIPSENQYFQ  SAIWSGFIKV
101   KKSDEYTFAT  SADNHVTMWV  DDQEVINKAS  NSNKIRLEKG  RLYQIKIQYQ
151   RENPTEKGLD  FKLYWTDSQN  KKEVISSDNL  QLPELKQKSS  NSRKKRSTSA
201   GPTVPDRDND  GIPDSLEVEG  YTVDVKNKRT  FLSPWISNIH  EKKGLTKYKS
251   SPEKWSTASD  PYSDFEKVTG  RIDKNVSPEA  RHPLVAAYPI  VHVDMENIIL
301   SKNEDQSTQN  TDSQTRTISK  NTSTSRTHTS  EVHGNAEVHA  SFFDIGGSVS
351   AGFSNSNSST  VAIDHSLSLA  GERTWAETMG  LNTADTARLN  ANIRYVNTGT
401   APIYNVLPTT  SLVLGKNQTL  ATIKAKENQL  SQILAPNNYY  PSKNLAPIAL
451   NAQDDFSSTP  ITMNYNQFLE  LEKTKQLRLD  TDQVYGNIAT  YNFENGRVRV
501   DTGSNWSEVL  PQIQETTARI  IFNGKDLNLV  ERRIAAVNPS  DPLETTKPDM
551   TLKEALKIAF  GFNEPNGNLQ  YQGKDITEFD  FNFDQQTSQN  IKNQLAELNA
601   TNIYTVLDKI  KLNAKMNILI  RDKRFHYDRN  NIAVGADESV  VKEAHREVIN
651   SSTEGLLLNI  DKDIRKILSG  YIVEIEDTEG  LKEVINDRYD  MLNISSLRQD
701   GKTFIDFKKY  NDKLPLYISN  PNYKVNVYAV  TKENTIINPS  ENGDTSTNGI
751   KKILIFSKKG  YEIG
```

Figure 7.2 | Amino acid sequence of PA. Sequence Length: 764 amino acids, molecular mass: 85811 Da. Selected peptides as SIS for AQUA were highlighted in underlined red. Sequence data sourced from SwissProt when rPA sample was searched against Bacteria (*Bacillus anthracis*) database (accessed on December 2015).

Following chemical synthesis, SIS were verified by LC-MS for chromatographic behaviour and spectra verifications. For SIS analysis, two types of instrumentation platforms were used in this study: amaZon (ion trap) and maXis (quadrupole-TOF). First, a mixture of approximately 500 fmol of each SIS peptide (calculated based on weight) was injected and mass spectrometric analysis was performed on the amaZon. The results are shown in Figure 7.3. The chromatographic behaviours of the peptide A, B and C are good, as they were well separated and eluted in sharp peaks at 30.1, 28.9 and 27.3 minutes, respectively. MS1 spectra of SIS peptides show correct m/z signals, representing $[M+2H]^{2+}$, as calculated for the heavy labelled peptides.

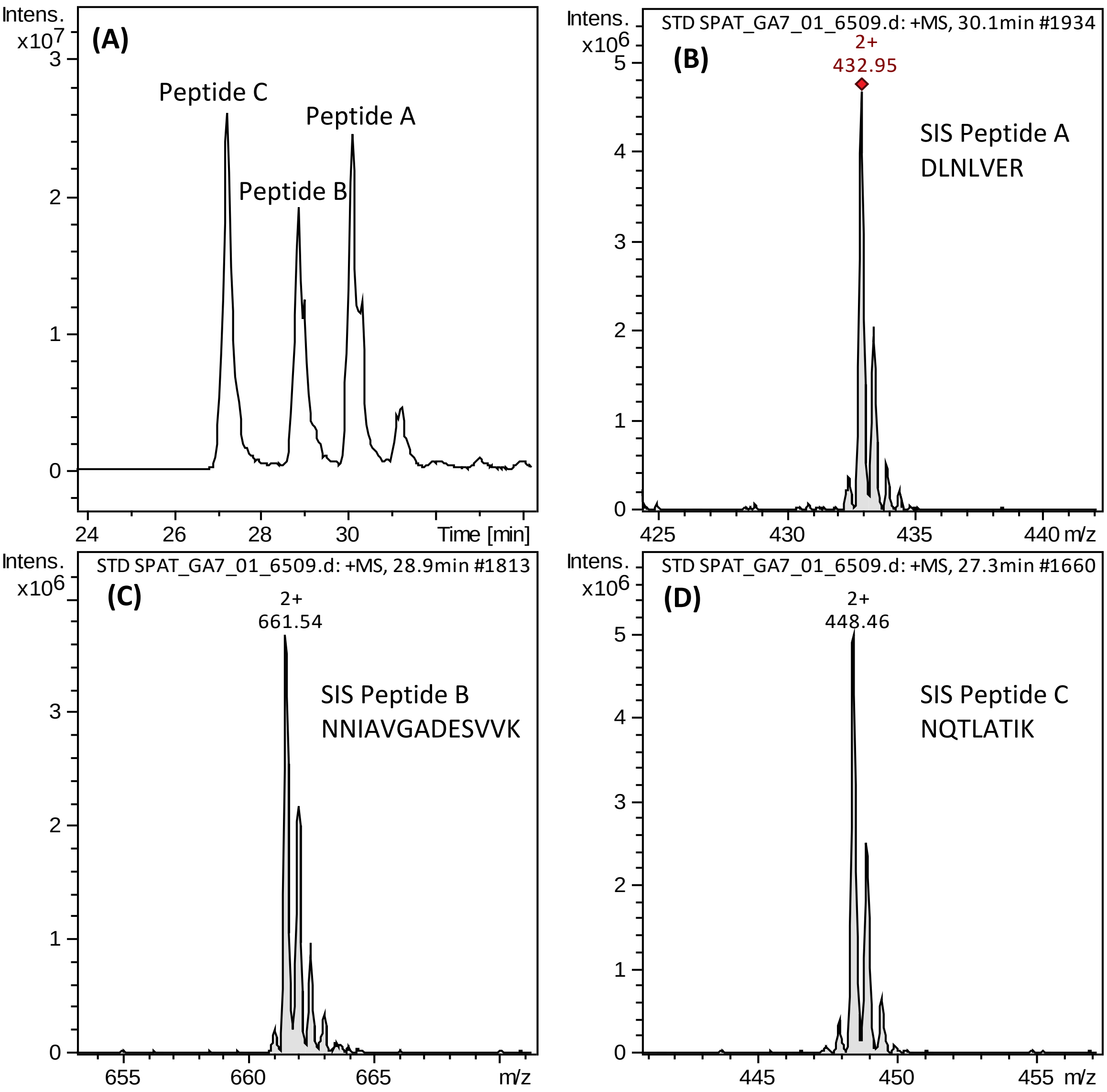


Figure 7.3 | Validation of SIS peptides, part I. (A) Base peak chromatogram (BPC) of the SIS peptides A, B and C. **(B)** to **(D)** MS1 spectrum of SIS peptides A, B and C, respectively, confirming the molecular weight of the heavy labelled peptides.

MS performances may cause departure in linearity, especially for low concentrations of analyte, when sample preparation process such as digestion and purification, as well as sample lost to tubes/tips due to adsorption, tend to have larger effects on quantification. Therefore further experiments were performed to determine the limit of detection (LOD) and assess quantification linearity using a serial dilution of each peptide on both the amaZon and the maXis MS systems. Each SIS peptide was diluted and injected with the amount of 0.04 pmol, 0.20 pmol, 0.40 pmol and 1.00 pmol. Results of calibration curves are shown in Figure 7.4. Abundance are based on area under curve (AUC) of extracted ion chromatograms (EIC) of the SIS $[M+2H]^{2+}$ m/z, which is 432.7, 661.3 and 448.2 for

SIS peptide A, B and C, respectively. All 3 SIS peptides show good linearity within the concentration range used in this experiment ($R^2 \geq 0.99$), for both MS instruments. Note that the accuracy of these results may also be affected by sample injection consistency of instruments (autosampler of HPLC), as slight different inject volumes may occur for each run.

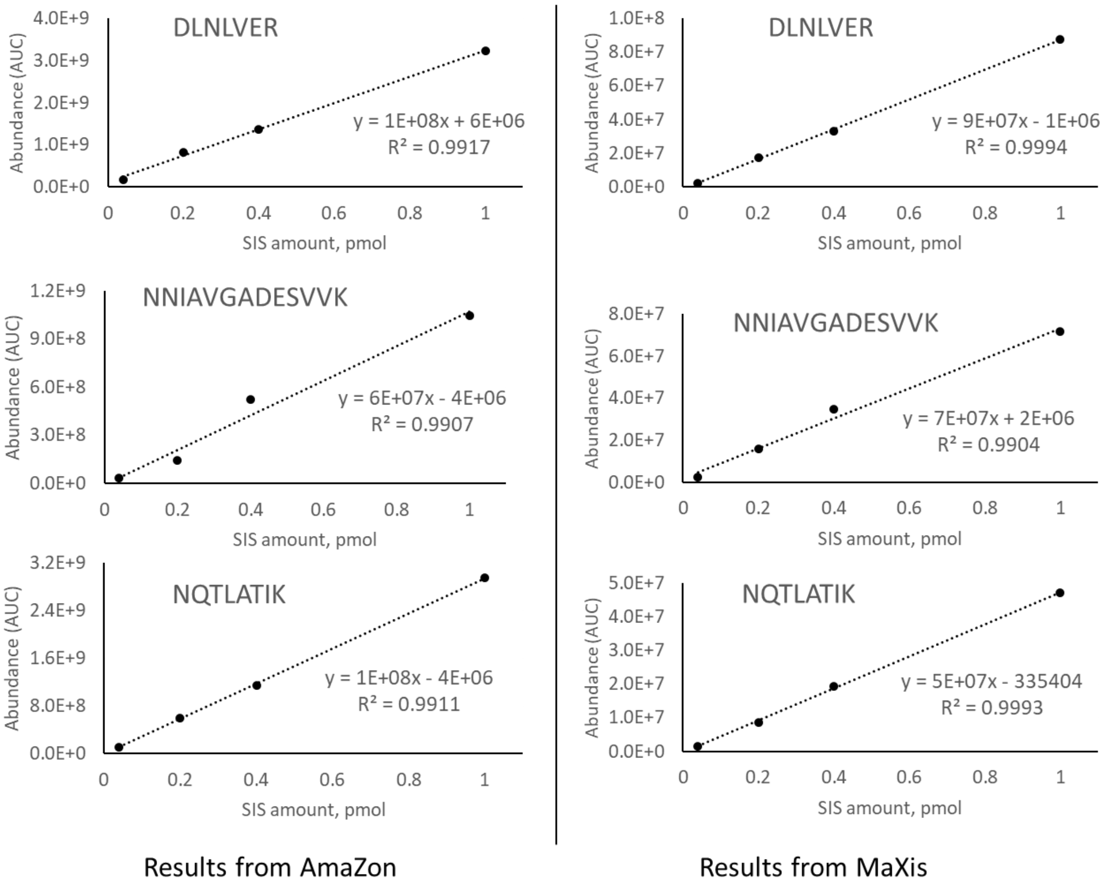


Figure 7.4 | Validation of SIS peptides, part II. Linear calibration curve for all three SIS peptides were spanning 25 fold range based on four dilutions. Serial dilutions were tested on both the amZon and the maXis MS systems. The dotted line represents the linear regression with its formula and R-squared value showed on the bottom right corner of each plot.

7.3.2 Accurate quantification of SIS peptides using amino acid analysis

Using internal standards for unknown peptide quantification, it is important that SIS peptides are accurately quantified. A straightforward quantification method would be to weight the samples directly. However, the accuracy of weighing relies heavily on personal skills and analytical balance working conditions. Also, sample loss is inevitable when handling powdered samples and making solutions. In order to determine the SIS peptide more accurately, amino acid analysis (AAA) was usually used. Briefly, peptides to be quantified are first hydrolysed (typically under acid conditions) into amino acids, which are then derivatized and separated by HPLC, and quantification performed based on internal and external standards. Practically, all 20 natural amino acids can be analysed, which gives detailed data for accurate and reliable quantifications. In general, AAA for peptide quantification has low microgram sensitivity.

In this study, synthesised SIS peptides were accurately quantified by AAA following validation. First, SIS pepitdes A, B and C (powder) were weighted and dissolved in ddH$_2$O, with additional aqueous guanidine hydrochloride (4M solution) as chaotropic agent for better solubility. Final guanidine hydrochloride concentration and approximate peptide concentration based on weighing is shown as below:

A) DLNLVER ~0.6mg/ml, Guanidine hydrochloride 2M
B) NNIAVGADESVVK ~0.5 mg/ml, Guanidine hydrochloride 1.92M
C) NQTLATIK ~0.8 mg/ml, Guanidine hydrochloride 0.8M

For AAA, all three synthesised peptides were sent to Protein and Nucleic Acid Chemistry Facility, University of Cambridge. Each analysis was performed in duplicate. The results were shown in Table 7.2 (details of AAA results are shown in Chapter 7 Appendices). Based on these concentrations, SIS were diluted to 1 pmole/µl with ddH$_2$O and ready for AQUA spike-in analysis.

Table 7.2 | Concentration of SIS peptides determined by amino acid analysis (AAA).

Peptides	Sequence	Conc.I (pmol/µl)	Conc. II (pmol/µl)	Average (pmol/µl)
A	DLNLVER	585.07	573.42	579.24
B	NNIAVGADESVVK	310.86	309.45	310.15
C	NQTLATIK	711.10	706.38	708.74

7.3.3 Optimization of rPA digestion time

For accurate quantification of rPA in conjunction with SIS peptides, it is important to ensure the rPA samples were fully digested but avoid nonspecific hydrolysis. Therefore, trypsin digestion time was carefully determined by the experiment described as follows: rPA sample (1 pmol approximately, estimated based on ELISA quantification) with SIS peptides spiked in (0.5~1 pmol) were trypsin digested for different lengths of time (2, 4, 8, or 16 hours respectively) prior to LC-MS analysis (maXis). SIS peptides were used as internal standards, with spiked amount ranging from 0.5~1 pmol. EIC of both the light and heavy version of peptides were processed, and the resulting AUC calculated as a measure of peptide abundance. An example is shown in figure 7.5 (A). Digest rate is defined as the ratio of endogenous rPA peptides (light) over spiked-in SIS peptides (heavy). Results are shown in Figure 7.5 (B). For SIS peptide A (top chart), digest was completed between 4-8 hours, where the rPA digest rate remains largely constant. After 8 hours, further non-specific digestion or potential loss of peptides occurs, causing the rPA digest rate to drop. In SIS peptide B spike-in scenario (middle chart), complete digest occurs after 8 hours of digestion, followed by a reduction in digest rate at 16 hours. However, digest rate from 4-8 hours is only marginally increased (from 3.1 to 3.5). For SIS peptide C (bottom chart), the complete digest time occurs between 4-8 hours, similar to SIS peptide A scenario. Based on these results, a digest time of 6.5 hours was used for the following rPA AQUA analysis. rPA digests were also checked on SDS-PAGE to ensure the completion of digestion after 6.5 hours (Figure 7.6).

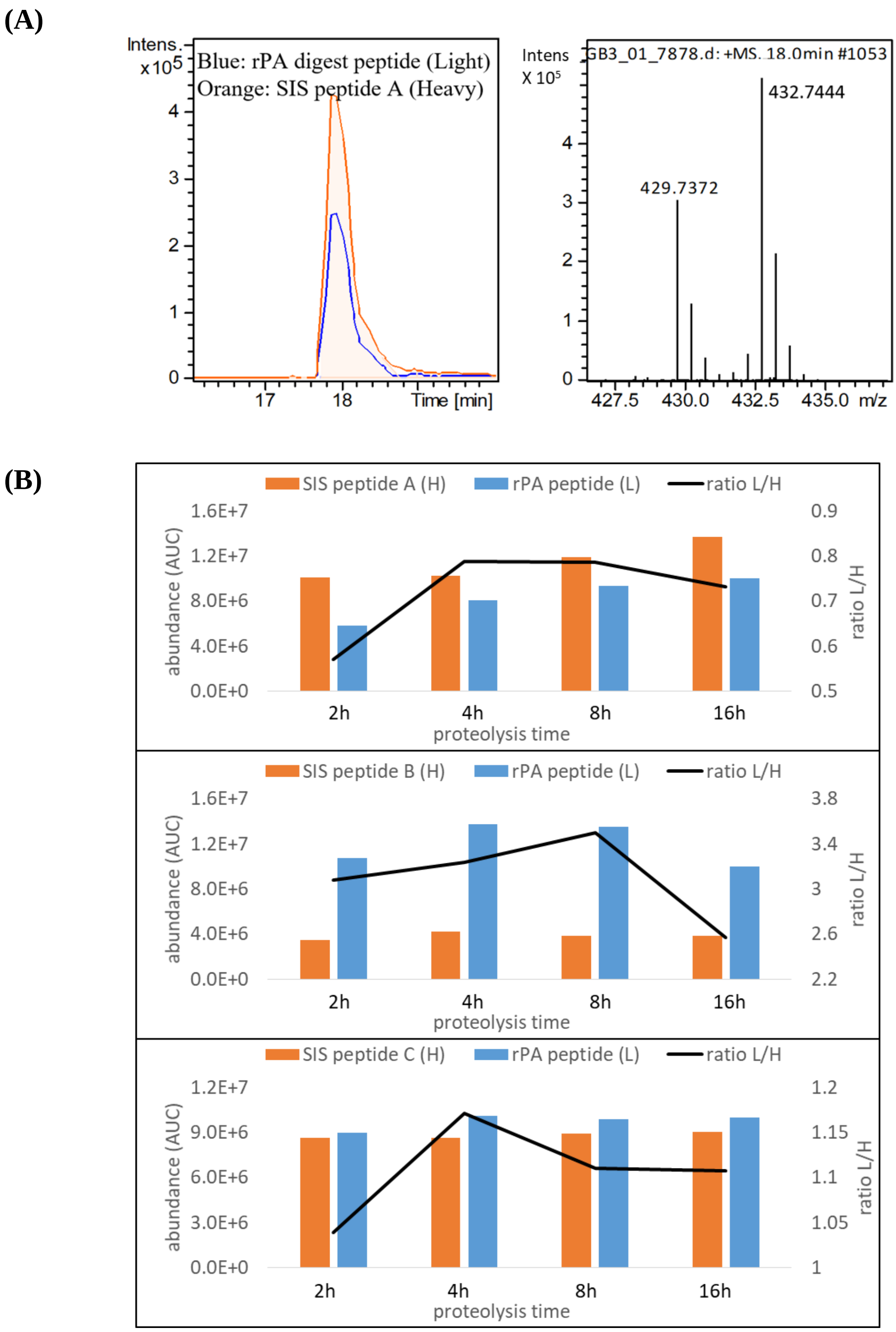


Figure 7.5 | Optimization of trypsin digestion time. rPA samples were trypsin digested for different lengths of time. SIS peptides were spiked in as internal standards. **(A)** An example showing the EIC of 2 hour rPA digest with SIS peptide A (left), AUC were calculated as abundance. Spectrum on the right shows the evidence of m/z of light and heavy versions of peptide A (DLNLVER). **(B)** rPA digest with spiked in SIS peptide, from top chart to bottom, SIS peptide A, B and C, respectively. Abundance is shown as orange bar (H). rPA digest peptide abundance is shown as blue bar (L). Black line on secondary axis shows the ratio of digested rPA peptide abundance over SIS peptide abundance (L/H), indicating relative rPA digest rate.

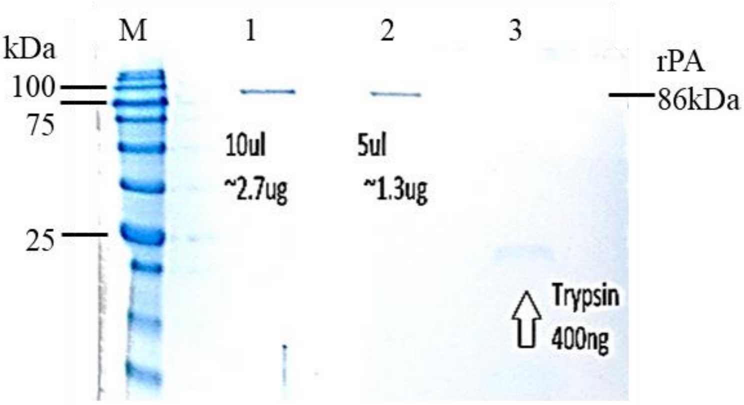


Figure 7.6 | SDS-PAGE results for rPA trypsin digest. rPA samples (1.3µg approximately) were trypsin degested for 6.5 hours before loaded on gel (lane 3). Lane 1 and 2 were non-digested rPA samples as negative controls.

7.3.4 Quantification of anthrax vaccine rPA

Following optimisation of the trypsin digest conditions and accurate quantification of the SIS peptides using AAA, LC-MS AQUA analysis of anthrax vaccine rPA was performed. The quantification was initially performed on purified rPA standard in PBS solution from Porton Biopharma with a given concentration of 6.9 µg/µl based on ELISA analysis (batch number 1C13, described as rPA hereafter). This rPA served as a standard which will further validate the AQUA assay developed in this study. To start with, rPA from stock was diluted to 30 ng/µl (approximately 0.35 pmol/µl, assuming rPA MW = 85811 Da). Analysis was performed using two different rPA amount group tests for the quantification: 1) 2.7 µl (1 pmol) with SIS peptides spiked-in at 0.25 pmol, 1.00 pmol or 4.00 pmol of each; 2) 5.4 µl (2 pmol) with SIS peptides spiked-in of 0.5 pmol, 2.00 pmol or 8.00 pmol of each. For both groups, SIS were spiked in as a mixture of 3 different peptides prior to trypsin digest. Sample preparation steps include digestion, speedvac concentration and re-suspension. No purification/desalting steps were applied to these samples. Both the amaZon and the maXis were used for AQUA analysis. Figure 7.7 shows a comparison of chromatograms from two instruments (rPA digest with 0.25 pmol SIS spiked in was chosen as an example). The results showed both consistent order of retention times and H/L ratios across the 2 instruments. Figure 7.8 shows the rPA quantification results using the amaZon. Abundances were calculated based on AUC of smoothed EIC peak, as described in previous section.

Based on L/H ratio for each SIS peptide, calculated results of rPA concentration are summarized in table 7.3 (amaZon section). Results based on SIS peptide A and B show excellent agreement with each other (85 pmol/µl, 86 pmol/µl respectively for the 2.7 µl rPA digest group, and 73 pmol/µl, 80 pmol/µl respectively for the 5.4 µl rPA digest group). SIS peptide C however, gave relatively greater quantification values with greater SD (103 pmol/µl for the 2.7 µl rPA digest group and 118 pmol/µl for the 5.4 µl rPA digest group). The total average of the two digest groups of all 3 SIS peptides was 90.8±9.3 pmol/µl (SD here represents the average SD of calculation based on three SIS peptides), compares to the expected concentration of 80.41 pmol/µl based on ELISA measurements (assuming rPA MW = 85,811 Da). Similarly, rPA quantification obtained on the maXis are shown in Figure 7.9. rPA concentration determined by SIS peptides A, B and C are summarized in table 7.3 (maXis section). Results from the maXis also show good agreement for SIS peptide A and B (71 pmol/µl, 79 pmol/µl respectively for the 2.7 µl rPA digest group, and 72 pmol/µl, 77 pmol/µl respectively for the 5.4 µl rPA digest group) but greater values for SIS peptides C (98 pmol/µl for 2.7 µl rPA digest group and 108 pmol/µl for 5.4 µl rPA digest group). The total average of the two digest groups of all 3 SIS peptides was 84.0±5.8 pmol/µl (SD here represents the average SD of calculation based on three SIS peptides). This is compared to the expected concentration of 80.41 pmol/µl based on ELISA measurements (assuming rPA MW = 85,811 Da).

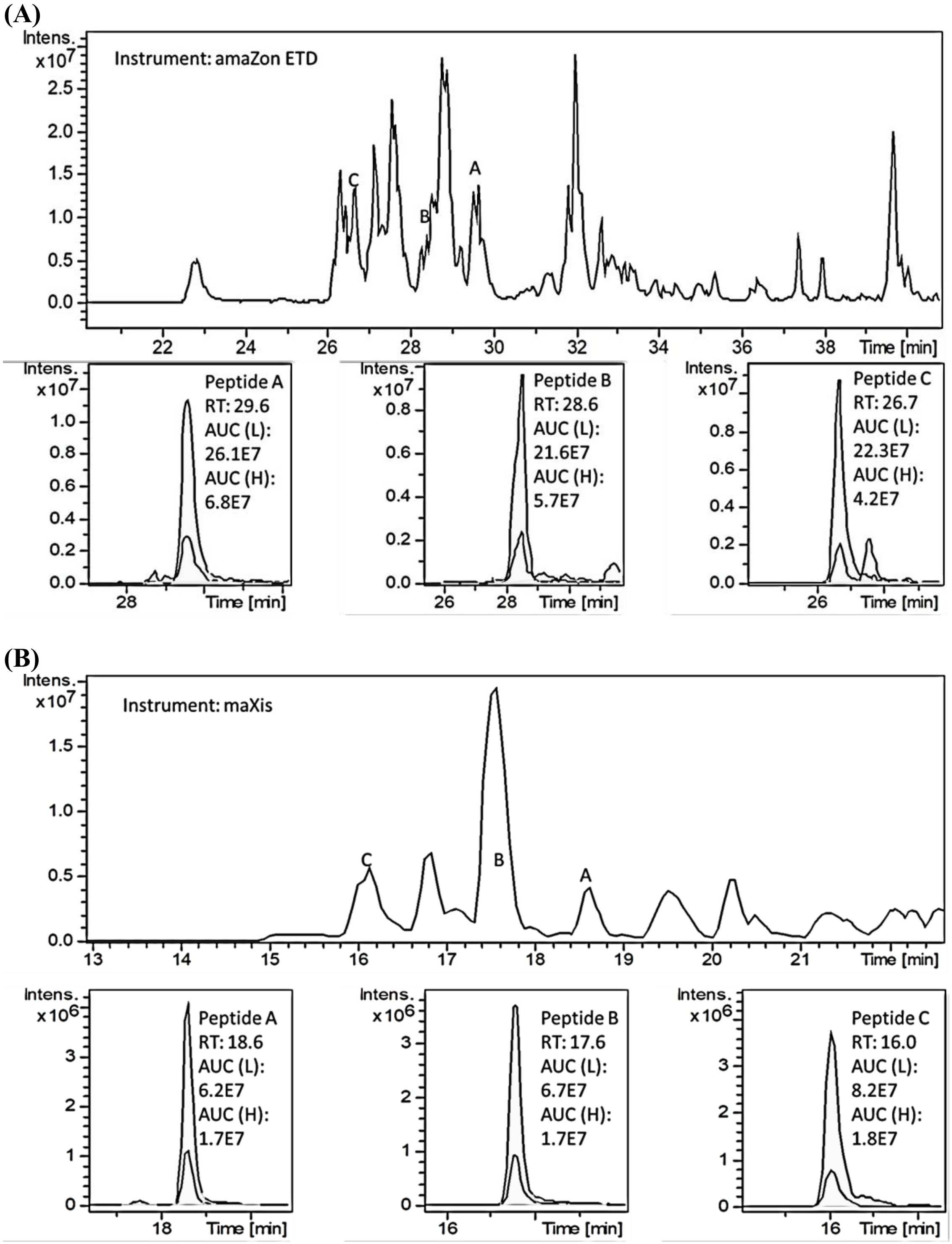


Figure 7.7 | Quantification of anthrax vaccine rPA. (A) and **(B)** An example of the AQUA workflow, showing 2.7 µl (~1 pmol) rPA digest with 0.25 pmol SIS spiked in. **(A)** Base peak chromatogram generated by amaZon ETD (top) and the extracted ion chromatograms of the light peptides from rPA digest and spiked-in heavy SIS (bottom). **(B)** Base peak chromatograms and extracted ion chromatograms of the same sample generated by maXis. RT: retention time; AUC: area under curve; L: light version of peptide; H: heavy version of peptide.

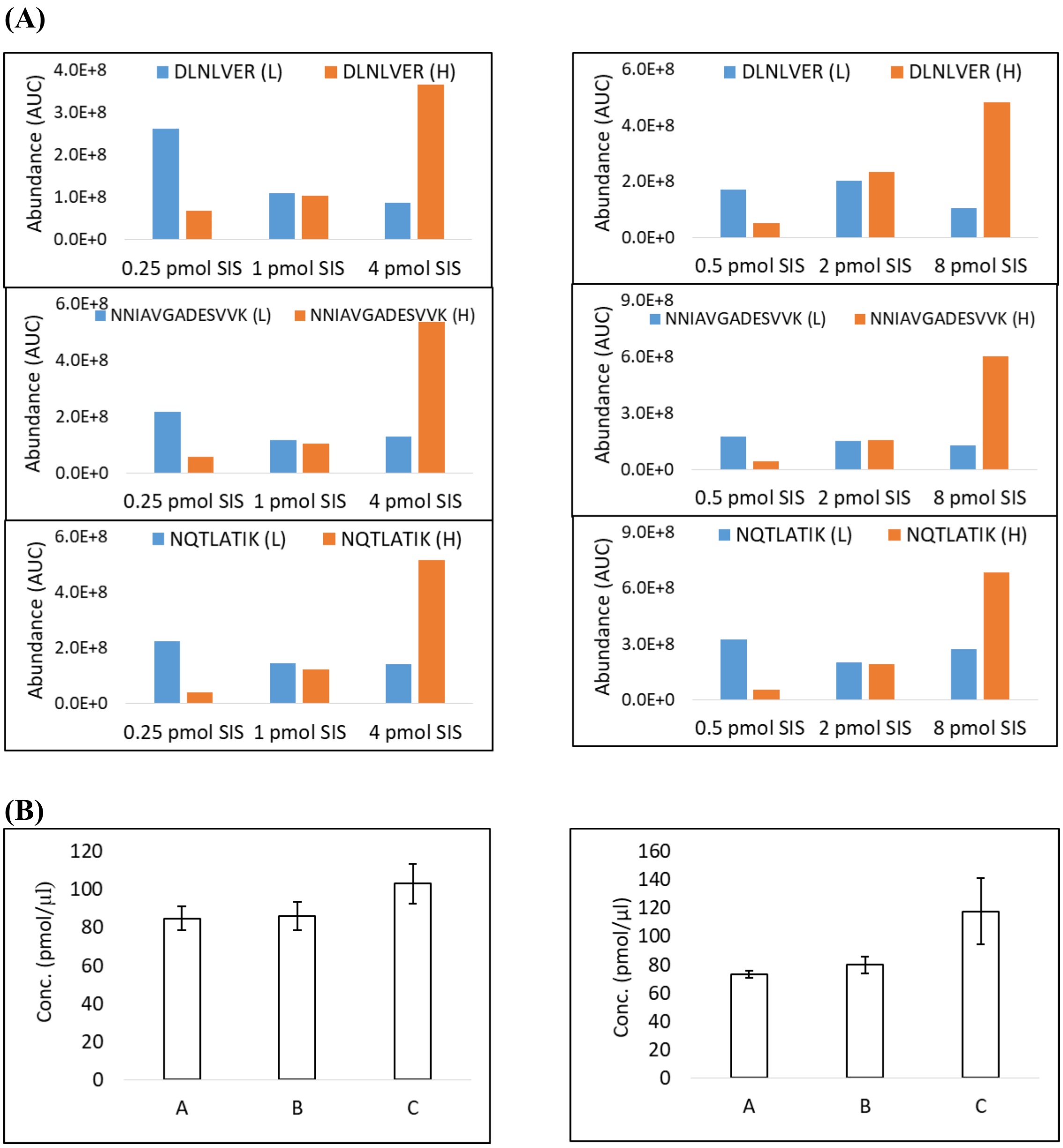


Figure 7.8 | Absolute quantification of rPA using LC ESI MS (amaZon). (A) AUC abundance of 2.7 µl (approximately 1 pmol estimated from ELISA) rPA digest (left panel) or 5.4 µl (approximately 2 pmol estimated from ELISA) rPA digest (right panel), with different SIS peptide spiked-in of different amount. (B) rPA concentration calculated based on three different SIS peptides. Left panel: results for 2.7 µl rPA digest; right panel: results for 5.4 µl rPA digest. Each experiment was performed in triplicates, the error bars represent the standard deviation.

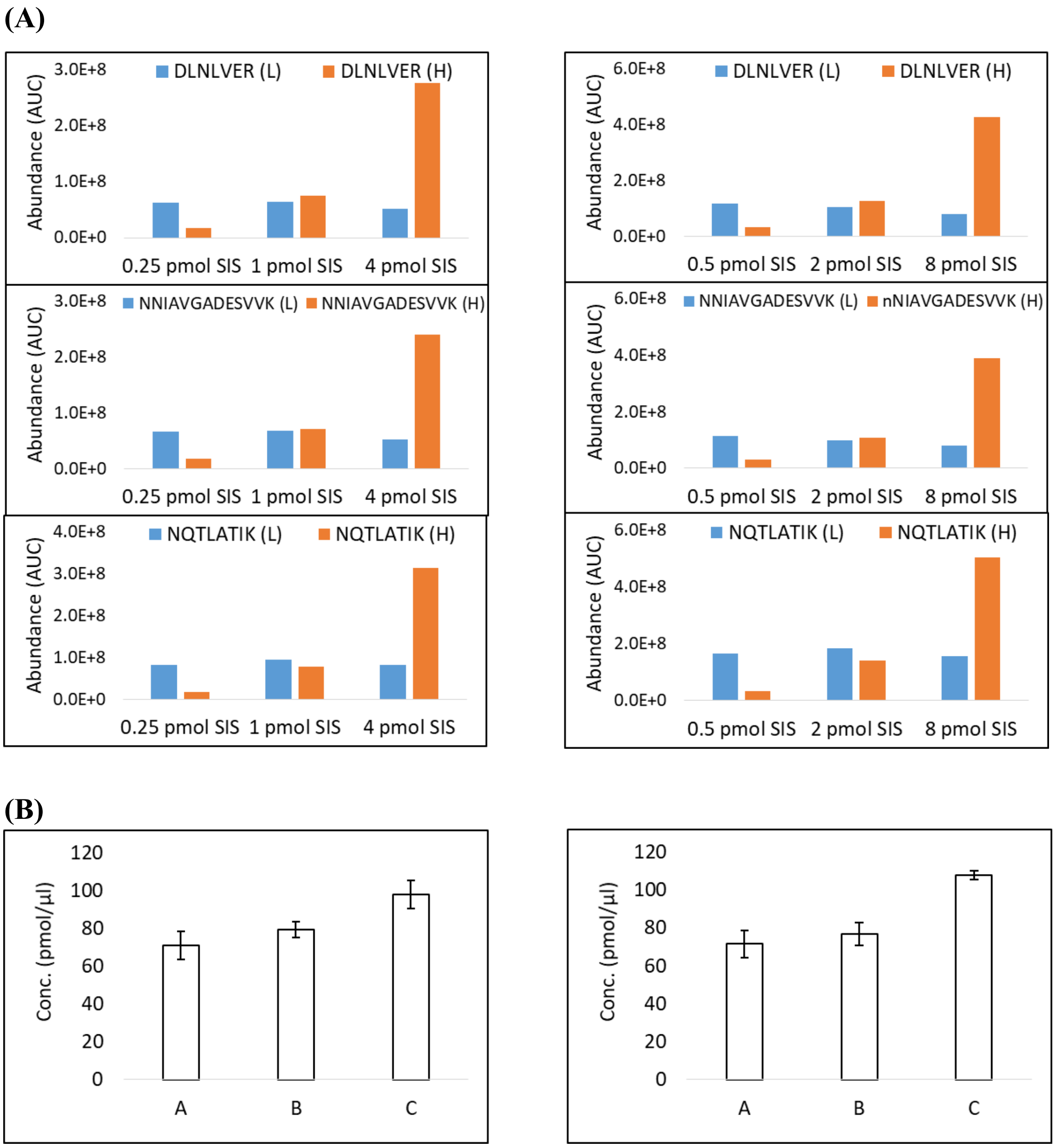


Figure 7.9 | Absolute quantification of rPA using LC ESI MS (maXis). (A) AUC abundance of 2.7 µl (approximately 1 pmol estimated from ELISA) rPA digest (left panel) or 5.4 µl (approximately 2 pmol estimated from ELISA) rPA digest (right panel), with different SIS peptides spiked-in of different amounts. (B) rPA concentration calculated based on three different SIS peptides. Left panel: results for 2.7 µl rPA digest; right panel: results for 5.4 µl rPA digest. Each experiment was performed in triplicates, the error bars represent the standard deviation.

Table 7.3 | Summary of rPA quantification using LC-MS. Calculations are based on L/H ratios of AUC generated from EIC of different amount of rPA digest and SIS peptides. Results from amaZon and maXis are compared. Quantification results are converted to original rPA stock before dilution (pmol/µl). SD: standard deviation.

		SIS amount SIS name	0.25 pmol	1 pmol	4 pmol	average	SD
Results from amaZon	2.7µl (~1 pmol) rPA digest	A	81.34	91.94	80.96	84.74	6.23
		B	81.17	94.86	82.46	86.16	7.56
		C	114.30	100.67	94.19	103.05	10.27
		average	92.27	95.82	85.87	**91.32**	8.02
	5.4µl (~2 pmol) rPA digest	SIS amount SIS name	0.5 pmol	2 pmol	8 pmol	average	SD
		A	70.30	74.71	74.85	73.28	2.59
		B	82.59	83.78	73.02	79.80	5.90
		C	126.86	91.29	134.91	117.69	23.21
		average	93.25	83.26	94.26	**90.26**	10.57
Results from maXis	2.7µl (~1 pmol) rPA digest	SIS amount SIS name	0.25 pmol	1 pmol	4 pmol	average	SD
		A	76.81	73.58	62.66	71.02	7.42
		B	82.35	81.23	74.36	79.31	4.33
		C	98.26	105.02	90.15	97.81	7.45
		average	85.80	86.61	75.72	**82.71**	6.40
	5.4µl (~2 pmol) rPA digest	SIS amount SIS name	0.5 pmol	2 pmol	8 pmol	average	SD
		A	78.33	72.11	64.21	71.55	7.08
		B	81.61	78.40	69.81	76.60	6.10
		C	106.87	110.26	105.90	107.68	2.29
		average	88.94	86.92	79.97	**85.28**	5.16

In this section, three different SIS peptides were used for the absolute quantification of rPA. From the results, the sample concentration values varied between different SIS peptides used. Generally speaking, peptide DLNLVER and NNIAVGADESVVK gave results in good agreement (relatively small differences and low SD). Calculation based on the peptide NQTLATIK resulted in higher values. Although this may affect the accuracy of quantification, the variation between calculated results based on three SIS peptides falls between 10%-30%. Within the same standard peptide (same amount of rPA digest, different amount of SIS peptides spiked-in, or different amount rPA digest groups, i.e., 2.7 µl and 5.4 µl used), the results are consistent. Comparing the two MS instruments, for the same sample, similar quantification results were obtained (differences less than 10% with few exceptions). Values calculated based on the amaZon tend to be slightly higher than those based on the maXis, in general. Further validation can be made by comparing the quantification results from the same SIS peptide of different stocks. Instrument resolution can also have influence on results. Here, results obtained on the maXis are more consistent with smaller SD compared to those obtained on the amaZon.

7.3.5 Quantification of anthrax vaccine filtrate

Data from previous section showed the ability to accurately quantify the amount of rPA using stable isotope labelling in conjunction with LC ESI MS. Data obtained were consistent with previous ELISA data using a purified recombinant protective antigen sample. Further studies were performed to use the developed LC-MS assay for the absolute quantification of the anthrax vaccine sample, which was sterile-filtered *Bacillus anthracis* Sterne strain harvest filtrate BHN 130/14/05 (filtrate for short hereafter). This Sterne strain is not virulent due to the natural loss of pXO2 plasmid thus lacking the ability to produce capsules, which this bacterium uses to protect itself from phagocytosis (Fasanella, 2013). Sterne strains are currently widely used for anthrax vaccine manufacturing because they are able to stimulate the immune response without causing infection. Compared to artificially produced rPA samples, native antigen produced and extracted from natural organisms are usually present in low concentrations and in complex compositions, which could affect the immunoassays and make the accuracy of ELISA based quantification compromised. LC-MS methods on the other hand, do not have this issue.

In this study, a vial of filtrate sample contains native PA (approximate concentration of 5 µg/ml based on ELISA quantification) was first concentrated to approximately 343 fmol/µl (sample concentrated from 500 µl to 85 µl, assuming PA MW=85811). Sample preparation steps and data analysis were exactly the same as those applied to rPA, as described in section 7.3.3. A summary of the results is shown in Figure 7.10, Figure 7.11 and Table 7.4. From both the amaZon and the maXis generated data, quantification results based on SIS peptide A are relatively low (62.7 pmol/µl on average), compared to results based on SIS peptide B and C (average of 82.9 pmol/µl and 88.6 pmol/µl, respectively, which are in good agreement with each other). Total average values of the two digest groups generated from all 3 SIS peptides with two instruments show good consistency, giving 79.9±10.3 pmol/µl from the amaZon data and 76.3±3.5 pmol/µl from the maXis data (SD represent the average SD of calculation based on three SIS peptides). This is compared to the expected concentration of 58.3 pmol/µl based on ELISA measurements (assuming rPA MW=85,811).

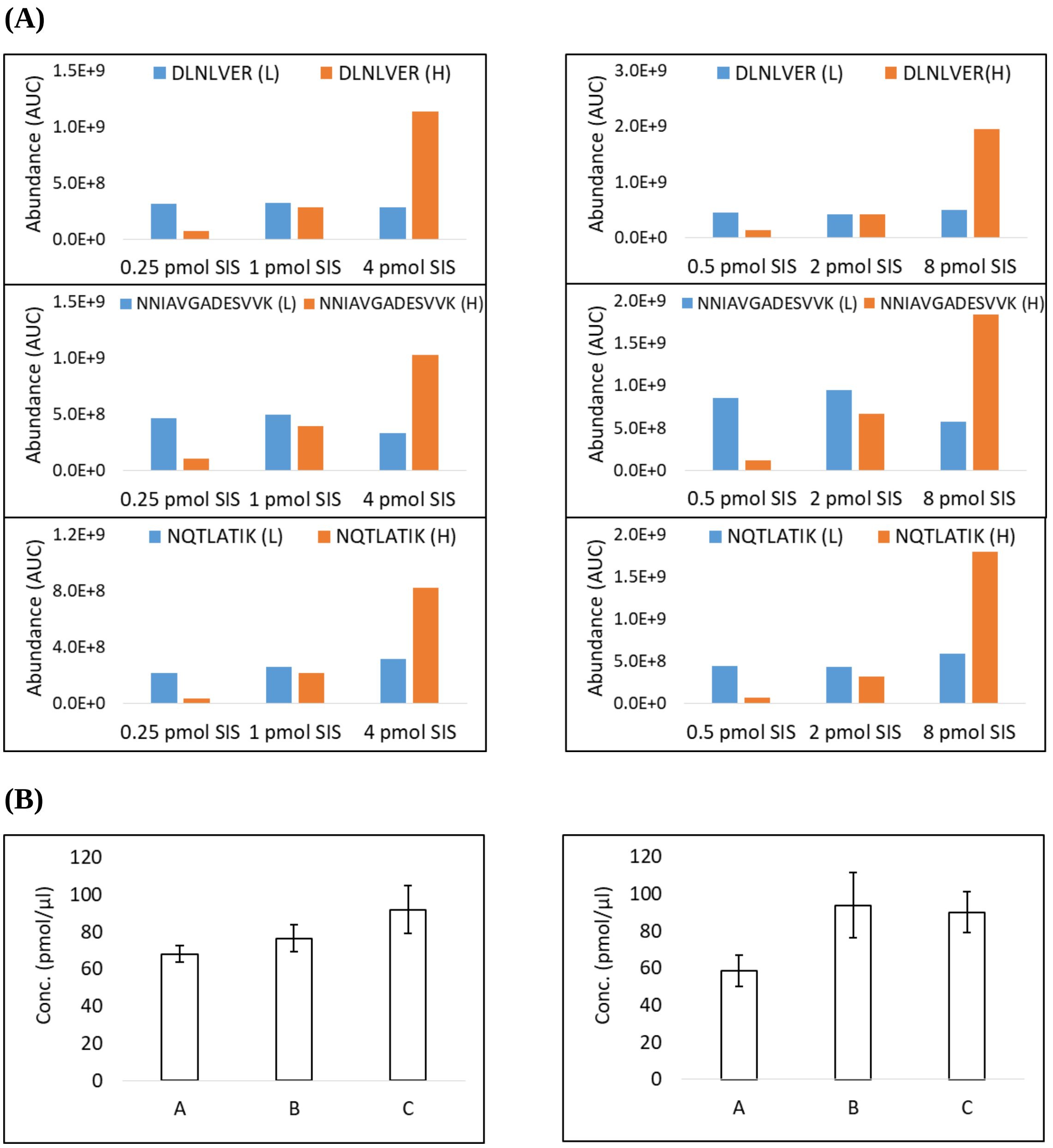


Figure 7.10 | Absolute quantification of anthrax vaccine filtrate using LC ESI MS (amaZon). (A) AUC abundance of 2.7 µl (approximately 1 pmol estimated from ELISA) rPA digest (left panel) or 5.4 µl (approximately 2 pmol estimated from ELISA) rPA digest (right panel), with different SIS peptides spiked-in of different amounts. (B) rPA concentration calculated based on three different SIS peptides. Left panel: results for 2.7 µl rPA digest; right panel: results for 5.4 µl rPA digest. Each experiment was performed in triplicates, the error bars represent the standard deviation.

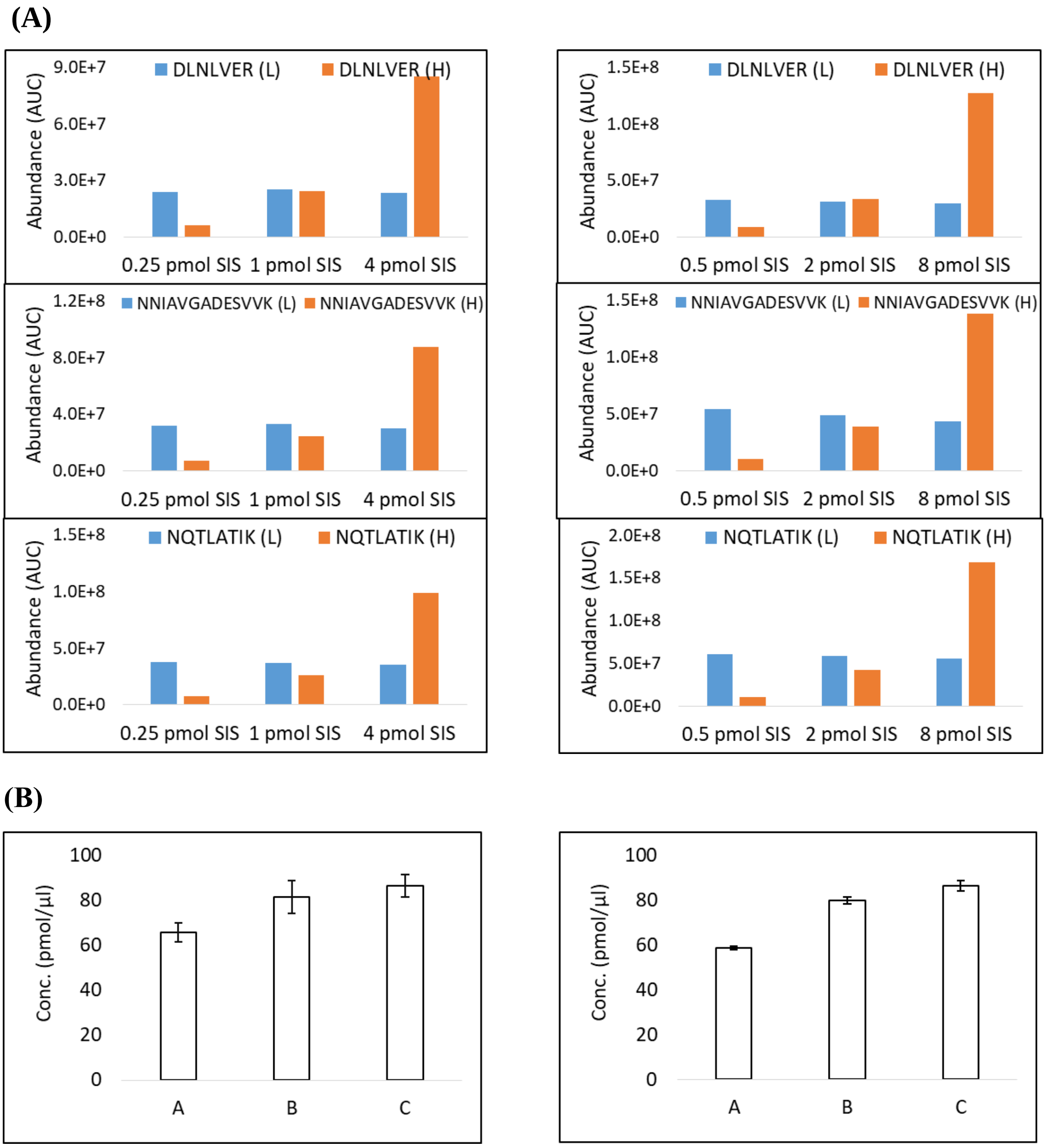


Figure 7.11 | Absolute quantification of anthrax vaccine filtrate using LC ESI MS (maXis). (A) AUC abundance of 2.7 µl (approximately 1 pmol estimated from ELISA) rPA digest (left panel) or 5.4 µl (approximately 2 pmol estimated from ELISA) rPA digest (right panel), with different SIS peptides spiked-in of different amounts. (B) rPA concentration calculated based on three different SIS peptides. Left panel: results for 2.7 µl rPA digest; right panel: results for 5.4 µl rPA digest. Each experiment was performed in triplicates, the error bars represent the standard deviation.

Table 7.4 | Summary of filtrate quantification using LC-MS. Calculations are based on L/H ratio of AUC generated from EIC of different amount of filtrate digest and SIS peptides. Results from amaZon and maXis are compared. Quantification results are converted equal to original filtrate stock before concentration (pmol/µl). SD: standard deviation.

		SIS amount SIS **name**	0.25 pmol	1 pmol	4 pmol	average	SD
Results from amaZon	2.7µl (~1 pmol) filtrate digest	A	70.09	71.13	62.93	68.05	4.47
		B	68.60	78.50	82.49	76.53	7.15
		C	101.09	77.19	97.65	91.98	12.92
		average	79.93	75.60	81.02	**78.85**	8.18
	5.4µl (~2 pmol) filtrate digest	SIS amount SIS name	0.5 pmol	2 pmol	8 pmol	average	SD
		A	48.80	62.97	64.10	58.62	8.53
		B	113.30	89.57	78.78	93.88	17.66
		C	102.81	84.54	82.90	90.08	11.05
		average	88.30	79.03	75.26	**80.86**	12.41
Results from maXis	2.7µl (~1 pmol) filtrate digest	SIS amount SIS name	0.25 pmol	1 pmol	4 pmol	average	SD
		A	61.20	66.26	69.28	65.58	4.08
		B	73.10	84.96	86.34	81.47	7.28
		C	80.60	88.71	89.48	86.26	4.92
		average	71.63	79.98	81.70	**77.77**	5.43
	5.4µl (~2 pmol) filtrate digest	SIS amount SIS name	0.5 pmol	2 pmol	8 pmol	average	SD
		A	58.17	59.42	58.18	58.59	0.72
		B	81.45	78.93	78.59	79.66	1.56
		C	87.85	87.15	83.47	86.16	2.35
		average	75.82	75.17	73.41	**74.80**	1.55

7.4 Conclusions

The aim of this chapter was to develop a robust, rapid and accurate mass spectrometry based approach for the absolute quantification of PA/rPA in anthrax vaccine. Mass spectrometry has been used as an alternative to immunoassays for many protein quantifications because of its high sensitivity, direct measurement and high-throughput. For the LC ESI MS assay development, three peptide sequences unique to PA/rPA were selected and chemically synthesised with heavy isotopes to be used as internal standards. These standards were then validated by MS (with correct m/z signals, good sensitivity and linearity), and the accurate concentration of these standard peptides was determined by amino acid analysis. For the preparation, anthrax vaccine samples were either diluted or concentrated to approximately 350 fmol/µl for efficient digestion, recovery and instrument analysis. To ensure samples were fully digested, a digestion time course experiment was carried out for rPA samples with heavy isotope labelled peptide standards spiked in. An optimized 6.5 hour digest time was determined for trypsin digestion.

LC ESI MS assays were first applied to rPA samples. Quantifications were based on ratios between native digested peptides (light) and standard peptides (heavy) using area under curve from extracted ion chromatograms. Although quantitation results based on different standard peptides varied, they are relatively consistent within individual standards. For different MS instruments, the amaZon gives comparable but slightly higher values and greater SD for rPA quantifications (90.8 ± 9.3, average value based on two rPA sample amount group tests, with three spiked SIS peptides each) compared to data obtained on the maXis (84.0 ± 5.8, average value based on two rPA sample amount group tests, with three spiked SIS peptides each). It is expected that results from the maXis are more accurate, due to its higher resolution. These results are in good consistency with the ELISA based quantification (80.4 fmol/µl, assuming rPA MW = 85811 Da), indicating the MS based quantification method developed here is reliable.

LC-MS analysis in conjunction with SIS was also used for absolute quantification of an anthrax vaccine filtrate sample. ELISA based methods are not ideal for the analysis of such samples, as the complex composition can affect the immunoassay performance. For quantification, although variations were observed using different peptide standards, the average values are comparable between different runs and instruments. The average value from the amaZon based on two filtrate sample amount group tests with three SIS peptides spiked each was 79.9 ± 10.3 fmol/µl, while from the maXis was 76.3 ± 3.5 fmol/µl. These results are comparable to the ELISA quantification result (58.3 fmol/µl, assuming PA MW = 85811 Da), the total average from this study is similar.

These results demonstrate the successful development and application of an LC ESI MS method in conjunction with SIS for the absolute quantification of anthrax vaccine PA/rPA.

Chapter 8: Final conclusions and future work

8.1 Thesis conclusions

Chemical modifications of nucleic acids and proteins play important roles in regulating their interactions in the cell. In this Thesis I have developed and optimised mass spectrometry based methods to study the effect of chemical modifications of both nucleic acids and proteins in a number of important biological systems including:

1) Studying the effect of DNA modifications on CRISPR/Cas systems (project in collaboration with Wageningen University)

2) Studying the effect of protein arginine methylation and citrullination on ribonucleoprotein complexes using quantitative proteomics

In chapter 3, I studied the effects of DNA modifications on the interference of CRISPR-Cas systems. CRISPR-Cas systems and restriction modifications (RM) are two known host defence system used by prokaryotes to provide protection against mobile genetic elements. RM and CRISPR-Cas systems work by targeting DNA specific sequences of invading DNA. One known strategy that phage use to counter-attack these defence systems of their host is to chemically modify their own DNA. However the potential role of DNA modifications in CRISPR-Cas systems is unknown. Initial work focussed on the synthesis of DNA substrates for CRISPR-Cas systems. dsDNA substrates were generated with a wide range of modifications, including 5-methylcytosine, 5-hydroxymethylcytosine, 2'-deoxyuridine, 5-hydroxymethyl-2'-deoxyuridine and phosphorothioate linkages. The polymerase chain reaction (PCR) was used to incorporate modified dNTPs. Due to the fact that modified dNTPs decrease PCR efficiency, optimisation was required. In addition to the above modifications, glucosylated dsDNA was enzymatically synthesised using dsDNA with 5-hydroxymethylcytosine in conjunction with T4 beta glucosyltransferase. Following the synthesis of modified DNA, *in vitro* binding/cleavage reactions of a range of different CRISPR-Cas systems was performed in collaboration with Wageningen University. The results showed that glucosylation can protect DNA from the action of type I-E and type II-A CRISPR-Cas systems, but not type V-A systems. In addition, the results also provided further insight into the mechanism of the inhibition in type I-E and type II-A CRISPR-Cas systems and demonstrated that the binding affinity of the protein complex was reduced, therefore inhibiting the ability to cleave the target DNA. These results highlight the effect of DNA modifications (glucosylation in particular) and their important impact on regulating protein-nucleic interactions as exemplified in the CRISPR-Cas systems. Furthermore, these results offer some exciting prospects that DNA modifications used by phage to overcome restriction modification system of bacteria, can indeed affect interference of the CRISPR/Cas systems. These findings have important implications that could be exploited for genome engineering applications.

Alongside the work performed in Chapter 3, I also developed and optimised HPLC-UV and LC-MS methods for the analysis of nucleic acid modifications. Using RP HPLC in conjunction with UV detection, different types of modifications including 5-methylcytosine, 5-hydroxymethylcytosine, 2'-deoxyuridine and 5-hydroxymethyl-2'-deoxyuridine, were confirmed in dsDNA synthesised previously using PCR in Chapter 3. Two types of reverse phase columns were evaluated including, superficially porous particles in conjunction with a C30 stationary phase (Accucore) and a porous graphitic carbon stationary phase (Hypercarb). The results showed good performances in nucleoside separation and identification using HPLC-UV. Moreover, the HPLC analysis enabled us to rapidly characterise and validate the DNA substrates used in Chapter 3 to study the effects of DNA modifications on CRISPR/Cas systems. However, the HPLC-UV analysis did not readily detect glucosylation of DNA. Therefore an LC-MS method was employed. This analytical method was also successfully applied for other biological systems including DNA from different strains of T4 phage and an engineered *E. coli* strain capable of synthesising DNA with various modifications. The method developed showed good sensitivity for the identification of glucosylated 5-hydroxymethyl-2'-deoxycytosine. In addition, quantitative analysis was also employed to determine the percentage incorporation of 5-hydroxymethyl-2'-deoxycytosine in DNA by analysing the peak areas of 2'-deoxycytosine and 5-hydroxymethyl-2'-deoxycytosine from HPLC-UV chromatograms.

In addition to the work described in Chapters 3 and 4 studying the effects of DNA modifications, research performed in Chapters 5 and 6 focused on studying the effects of protein post-translational modifications on RNA binding using quantitative proteomics. Quantitative proteomic experiments were performed using stable isotope labelling with amino acids in cell culture (SILAC). Label-swap strategies were used for the biological replicates. In chapter 5, the effects of the global methylation inhibitor AdOx and the effects of RNA methylation (N^6-methyladenosine) on global mRNA binding were studied. To do this, large scale mRNP capture assays were optimised, using UV crosslinking of cells to form the mRNP complex, oligo-d(T) affinity purification and RNase treatment for mRNA-binding protein recovery. Standard crosslinking and immunoprecipitation (CLIP) was compared to photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP), and the results showed the former was a preferred method for crosslinking in HEK293T cells. However, the results also highlighted the problem of arginine to proline conversion in the SILAC experiments, and required the use of additional proline in cell culture media to reduce the conversion to an acceptable level for accurate quantitative mass spectrometry analysis (which was further corrected by post-quantification normalisation based on label-swap replicates).

To study the effects of protein methylation on mRNA binding, cells were grown in the presence and absence of the global methylation inhibitor AdOx. SILAC label-swap experiments were performed in conjunction with mRNP capture assays to quantitatively analyse changes in mRNA binding on a global scale. The results identified over 600 proteins of which over 500 were quantified in the SILAC label-swap experiment. 24 proteins showed increased mRNA binding and 50 showed decreased mRNA binding (> 1.5 fold change in protein abundance). A wide number of identified proteins contain sites of arginine methylation which are of particular interest. These results are the first global quantitative analysis of the effect of methylation on mRNA binding and highlight a number of interesting RNA binding proteins as candidates for further studies to examine the role of arginine methylation and RNA binding.

To study the effect of N^6-methyladenosine (m^6A) of mRNA on mRNA-protein interactions, mRNP capture assays were carried out comparing wild type cell line and m^6A deficient cell line. The m^6A deficient cell line used was an RNAi knockdown of Wilms tumour 1 associated protein (WTAP), a protein that recruits the m^6A methyltransferase complex. The results of the mRNP capture assay from cells grown with and without m^6A RNA identified approximately 500 proteins of which 400 were quantified. However, no significant difference in the mRNA interactome was observed. Further analysis revealed that the RNAi knockdown was not efficient in the cell line used, resulting in no significant alteration of RNA methylation, therefore the mRNA binding was not altered.

In chapter 6, similar SILAC and quantitate MS approaches were applied to study the effects of protein methylation and citrullination binding to an RNA oligonucleotide $(GGGGCC)_5$ *in vitro*. Large scale *in vitro* RNA pulldown assays were used in conjunction with cells grown in the presence and absence of AdOx. The results identified more than 100 proteins which are with over 1.5 fold change either increased or decreased in RNA binding. To study the effects of citrullination, a cell line overexpressing the enzyme PADI4 was used in conjunction with *in vitro* RNA binding assays and SILAC MS analysis. The results identified over 700 proteins of which close to 600 were quantified in the SILAC label swap experiment, similar to the AdOx experiments. However, only 9 proteins showed increased $(GGGGCC)_5$ oligo RNA binding and 2 showed decreased (GGGGCC)5 oligo RNA binding (> 1.5 fold change in protein abundance). These results are the first global quantitative analysis of the effects of methylation/citrullination on GGGGCC repeat binding and highlight a number of interesting GGGGCC repeat binding proteins as candidates for further studies to examine the role of arginine modifications and RNA binding protein PTMs in amyotrophic lateral sclerosis and frontotemporal dementia.

The final Chapter 7 focussed on the application of quantitative MS analysis for absolute quantification of protective antigen or recombinant protective antigen in anthrax vaccines. Stable isotope dilution mass spectrometry was used in conjunction with stable isotope-labelled peptide standards. Quantitative MS analysis was performed using 2 different MS instruments (maXis and amaZon) and the results compared. The quantitative MS results were consistent across the two different instruments. Moreover the absolute quantification based on all three standard peptides was in agreement with the ELISA based quantification of rPA, demonstrating the MS based method for absolute quantification is robust, rapid and accurate. In addition, LC-MS analysis in conjunction with stable isotope-labelled peptide standards was also used for the absolute quantification of an anthrax vaccine filtrate sample. ELISA based methods are not ideal for the analysis of such samples, as the complex composition can affect the immunoassay performance. These results demonstrate the successful development and application of an LC ESI MS method in conjunction with stable isotope-labelled peptide standards for the absolute quantification of anthrax vaccine PA/rPA.

8.2 Future work

Further studies on the effects of chemical modifications of DNA in CRISPR-Cas systems could be extended to include additional modifications. *In vitro* CRISPR-Cas interference analysis focused mainly on 5-hydroxymethylcytosine and 5-glucosylhydroxymethylcytosine. DNA with phosphorothioate linkages in particular is of great interests, as preliminary results showed resistance to type I-E CRISPR-Cas systems, and it is believed a good candidate that can potentially be used as non-cleavable DNA templates in applications such as homology directed repair (HDR). The challenge remains, however, is the biological synthesis of DNA with phosphorothioate linkages. This study demonstrated that little or no PCR products with phosphorothioate linkages were synthesised when 3 or 4 thio-dNTPs were used, which means it is impossible to achieve dsDNA with fully phosphorothioated backbones using PCR. To further optimise the synthesis of fully phosphorothioated DNA, alternative DNA polymerases should be tested in conjunction with protein engineering approaches that will enable effective incorporation of thio-dNTPs during PCR. In addition, the ability of *E. coli* to generate DNA with phosphorothioate backbones is of particular interest. Further work could be aimed at engineering *E. coli* to incorporate thiophosphate or chemically modify its phosphate backbone of DNA post synthesis in the cell.

Although an analytical method was developed using HPLC-UV and LC-MS and successfully applied for modified nucleoside analysis, a few issues remain. In particular, improved methods for nucleoside quantification remain a challenge. Further studies could be aimed at utilising stable isotope labelling methods in conjunction with MS based quantification. Compared to HPLC-UV based quantifications, LC-MS with isotope labelling gives strong evidence of m/z confirmation of target compounds, unbiased signals, and can be applied for the quantification of two or more samples by mixing them together. Isotopes such as ^{13}C and ^{15}N can be used for labelling. With a known amount of standard, the unknown sample amount can be calculated based on the ratio of heavy and light signal of corresponding nucleoside.

The result presented in Chapters 5 and 6 are the first global quantitative analysis of the effect of protein methylation and citrullination on RNA binding. The results identify a number of interesting RNA binding proteins as candidates for further studies to examine the role of arginine methylation on RNA binding both *in vitro* and *in vivo*. Further studies need to be performed to validate the quantitative analysis. Due to time constraints it was not possible to perform western blotting on a number of proteins of interest, the binding of which to RNA were altered upon the addition of AdOx. In addition, further MS analysis could be performed to demonstrate reduced arginine methylation in these proteins. Further work is necessary to also validate that upon overexpression of PADI4, enzyme activity results in citrullination of substrate proteins. Again analysis of the current MS data could be used to show increased citrullination in the proteome. However, global analysis of citrullination is difficult, as the resulting mass difference is the same as deamidation of N/Q residues, therefore it is difficult to distinguish citrullination (R) from deamidation (N/Q).

Due to the issue of ineffective RNAi knockdown of WTAP in the cell line used, further experiments need to be performed. To achieve the cell line with the required feature, an alternative WTAP knockdown technique can be used in conjunction with CRISPR/Cas9 editing. CRISPR/Cas9 WTAP knockdown cell lines are commercially available, or can be obtained from other labs, which can be used to study the effects of N6 methyladenosine on mRNA binding.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. 2002. Molecular Biology of the Cell. 4th edition. New York: Garland Science; Meiosis. *New York: Garland Science*.

Aloni, Y., Dhar, R. and Khoury, G. 1979. Methylation of nuclear simian virus 40 RNAs. *Journal of virology*. **32**(1), 52–60.

Andersson, A.F. and Banfield, J.F. 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science (New York, N.Y.)*. **320**(5879), 1047–50.

Asaga, H., Nakashima, K., Senshu, T., Ishigami, A. and Yamada, M. 2001. Immunocytochemical localization of peptidylarginine deiminase in human eosinophils and neutrophils. *Journal of leukocyte biology*. **70**(1), 46–51.

Ascano, M., Hafner, M., Cekan, P., Gerstberger, S. and Tuschl, T. 2012. Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley Interdisciplinary Reviews: RNA*. **3**(2), 159–177.

Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A., Phanse, S., Joachimiak, A., Koonin, E. V., Savchenko, A., Emili, A., Greenblatt, J., Edwards, A.M. and Yakunin, A.F. 2011. A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Molecular Microbiology*. **79**(2), 484–502.

Baltz, A.G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., Wyler, E., Bonneau, R., Selbach, M., Dieterich, C. and Landthaler, M. 2012. The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular Cell*. **46**(5), 674–690.

Baraibar, M.A., Liu, L., Ahmed, E.K. and Friguet, B. 2012. Protein oxidative damage at the crossroads of cellular senescence, aging, and age-related diseases. *Oxidative Medicine and Cellular Longevity*. Article ID919832.

Barril, X., Alemán, C., Orozco, M. and Luque, F.J. 1998. Salt bridge interactions: Stability of the ionic and neutral complexes in the gas phase, in solution, and in proteins. *Proteins: Structure, Function and Genetics*. **32**(1), 67–79.

Beck-Sickinger, A.G. and Mörl, K. 2006. Posttranslational Modification of Proteins. Expanding Nature’s Inventory. By Christopher T. Walsh. *Angewandte Chemie International Edition*. **45**(7), 1020–1020.

Beloglazova, N., Brown, G., Zimmerman, M.D., Proudfoot, M., Makarova, K.S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W., Koonin, E. V., Edwards, A.M., Savchenko, A. and Yakunin, A.F. 2008. A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *Journal of Biological Chemistry*. **283**(29), 20361–20371.

Bendall, S.C., Hughes, C., Stewart, M.H., Doble, B., Bhatia, M. and Lajoie, G.A. 2008. Prevention of Amino Acid Conversion in SILAC Experiments with Embryonic Stem Cells. *Molecular & Cellular Proteomics*. **7**(9), 1587–1597.

Biffi, G., Tannahill, D., McCafferty, J. and Balasubramanian, S. 2013. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nature Chemistry*. **5**(3), 182–186.

Birney, E., Kumar, S. and Krainer, A.R. 1993. Analysis of the RNA-recognition motif and RS and RGG domains: Conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Research*. **21**(25), 5803–5816.

Blanc, R.S. and Richard, S. 2017. Arginine Methylation: The Coming of Age. *Molecular Cell*. **65**, 8–24.

Bolotin, A., Quinquis, B., Sorokin, A. and Dusko Ehrlich, S. 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*. **151**(8), 2551–2561.

Bonk, T. and Humeny, a. 2001. MALDI-TOF-MS Analysis of Protein and DNA. *The Neuroscientist*. **7**(1), 6–12.

Botelho, J.C., Shacklady, C., Cooper, H.C., Tai, S.S., Van Uytfanghe, K., Thienpont, L.M. and Vesper, H.W. 2013. Isotope-dilution liquid chromatography-tandem mass spectrometry candidate reference method for total testosterone in human serum. *Clin Chem*. **59**(2), 372–380.

Bradley, K.A., Mogridge, J., Mourez, M., Collier, R.J. and Young, J.A. 2001. Identification of the cellular receptor for anthrax toxin. *Nature*. **414**(6860), 225–229.

Brodbelt, J.S. 2016. Ion Activation Methods for Peptides and Proteins. *Analytical Chemistry*. **88**(1), 30–51.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E. V. and Van Der Oost, J. 1993. Small Crispr Rnas Guide Antiviral Defense in Prokaryotes. *Cancer Epidemiology Biomarkers and Prevention*. **2**(6), 531–535.

Bruins, A.P., Covey, T.R. and Henion, J.D. 1987. Ion spray interface for combined liquid chromatography/atmospheric pressure ionization mass spectrometry. *Analytical Chemistry*. **59**(22), 2642–2646.

Brüssow, H. 2001. Phages of Dairy Bacteria. *Annual Review of Microbiology*. **55**(1), 283–303.

Brzezicha, B., Schmidt, M., Makałowska, I., Jarmołowski, A., Pieńkowska, J. and Szweykowska-Kulińska, Z. 2006. Identification of human tRNA: m5C methyltransferase catalysing intron-dependent m5C formation in the first position of the anticodon of the pre-tRNA(CAA)Leu. *Nucleic Acids Research*. **34**(20), 6034–6043.

Bugaut, A. and Balasubramanian, S. 2012. 5′-UTR RNA G-quadruplexes: Translation regulation and targeting. *Nucleic Acids Research*. **40**(11), 4727–4741.

Burrows, C., Abd Latip, N., Lam, S.-J., Carpenter, L., Sawicka, K., Tzolovsky, G., Gabra, H., Bushell, M., Glover, D.M., Willis, A.E. and Blagden, S.P. 2010. The RNA binding protein Larp1 regulates cell division, apoptosis and cell migration. *Nucleic acids research*. **38**(16), 5542–5553.

Callister, J.B., Ryan, S., Sim, J., Rollinson, S. and Pickering-Brown, S.M. 2016. Modelling C9orf72 dipeptide repeat proteins of a physiologically relevant size. *Human Molecular Genetics*. **25**(23), 5069–5082.

Casadesús, J. and Low, D. 2006. Epigenetic gene regulation in the bacterial world. *Microbiology and molecular biology reviews : MMBR*. **70**(3), 830–856.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., Krijgsveld, J. and Hentze, M.W. 2012. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell*. **149**(6), 1393–1406.

Chang, C. Te, Hautbergue, G.M., Walsh, M.J., Viphakone, N., Van Dijk, T.B., Philipsen, S. and Wilson, S.A. 2013. Chtop is a component of the dynamic TREX mRNA export complex. *EMBO Journal*. **32**(3), 473–486.

Chang, K.Y. and Ramos, A. 2005. The double-stranded RNA-binding motif, a versatile macromolecular docking platform. *FEBS Journal*. **272**(9), 2109–2117.

Chang, X. and Han, J. 2006. Expression of peptidylarginine deiminase type 4 (PAD4) in various tumors. *Molecular Carcinogenesis*. **45**(3), 183–196.

Chaudhury, A., Chander, P. and Howe, P.H. 2010. Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: Focus on hnRNP E1's multifunctional regulatory roles. *RNA*. **16**(8), 1449–1462.

Chen, D.H., Wu, K.T., Hung, C.J., Hsieh, M. and Li, C. 2004. Effects of adenosine dialdehyde treatment on in vitro and in vivo stable protein methylation in heLa cells. *Journal of Biochemistry*. **136**(3), 371–376.

Chibani-Chennoufi, S., Bruttin, A., Dillmann, M.L. and Brüssow, H. 2004. Phage-host interaction: An ecological perspective. *Journal of Bacteriology*. **186**(12), 3677–3686.

Chodosh, L. a 2001. UV crosslinking of proteins to nucleic acids. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*. **Chapter 12**, Unit 12.5.

Chung, S., Zhou, Z., Huddleston, K.A., Harrison, D.A., Reed, R., Coleman, T.A. and Rymond, B.C. 2002. Crooked neck is a component of the human spliceosome and implicated in the splicing process. *Biochimica et Biophysica Acta - Gene Structure and Expression*. **1576**(3), 287–297.

Ciborowski, P. and Silberring, J. 2016. *Proteomic Profiling and Analytical Chemistry:*

The Crossroads: Second Edition.

Cleary, J.D. and Ranum, L.P.W. 2013. Repeat-associated non-ATG (RAN) translation in neurological disease. *Human Molecular Genetics*. **22**(R1), 45–51.

Cléry, A., Blatter, M. and Allain, F.H.T. 2008. RNA recognition motifs: boring? Not quite. *Current Opinion in Structural Biology*. **18**(3), 290–298.

Cocozaki, A.I., Ramia, N.F., Shao, Y., Hale, C.R., Terns, R.M., Terns, M.P. and Li, H. 2012. Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex. *Structure*. **20**(3), 545–553.

Coenen, D., Verschueren, P., Westhovens, R. and Bossuyt, X. 2007. Technical and diagnostic performance of 6 assays for the measurement of citrullinated protein/peptide antibodies in the diagnosis of rheumatoid arthritis. *Clinical chemistry*. **53**(3), 498–504.

Collier, J., McAdams, H.H. and Shapiro, L. 2007. A DNA methylation ratchet governs progression through a bacterial cell cycle. *Proceedings of the National Academy of Sciences*. **104**(43), 17111–17116.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., Zhang, F., Porteus, M.H., Baltimore, D., Miller, J.C., Sander, J.D., Wood, A.J., Christian, M., Zhang, F., Miller, J.C., Reyon, D., Boch, J., Moscou, M.J., Bogdanove, A.J., Stoddard, B.L., Jinek, M., Gasiunas, G., Barrangou, R., Horvath, P., Siksnys, V., Garneau, J.E., Deveau, H., Garneau, J.E., Moineau, S., Horvath, P., Barrangou, R., Makarova, K.S., Bhaya, D., Davison, M., Barrangou, R., Deltcheva, E., Sapranauskas, R., Magadán, A.H., Dupuis, M.E., Villion, M., Moineau, S., Deveau, H., Mojica, F.J., Díez-Villaseñor, C., García-Martínez, J., Almendros, C., Jinek, M., Doudna, J.A., Malone, C.D., Hannon, G.J., Meister, G., Tuschl, T., Certo, M.T., Mali, P., Carr, P.A., Church, G.M., Barrangou, R., Horvath, P., Brouns, S.J. and Guschin, D.Y. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*. **339**(6121), 819–823.

Coon, J.J., Syka, J., Shabanowitz, J., Hunt, D.F. and Others 2005. Tandem mass spectrometry for peptide and protein sequence analysis. *Biotechniques*. **38**(4), 519–521.

Cooper-Knock, J., Bury, J.J., Heath, P.R., Wyles, M., Higginbottom, A., Gelsthorpe, C., Highley, J.R., Hautbergue, G., Rattray, M., Kirby, J. and Shaw, P.J. 2015. C9ORF72 GGGGCC expanded repeats produce splicing dysregulation which correlates with disease severity in amyotrophic lateral sclerosis. *PLoS ONE*. **10**(5), e0127376.

Cox, J. and Mann, M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*. **26**(12), 1367–1372.

Cubeñas-Potts, C. and Matunis, M.J. 2013. SUMO: A Multifaceted Modifier of Chromatin Structure and Function. *Developmental Cell*. **24**(1), 1–12.

Cuthbert, G.L., Daujat, S., Snowden, A.W., Erdjument-Bromage, H., Hagiwara, T., Yamada, M., Schneider, R., Gregory, P.D., Tempst, P., Bannister, A.J. and Kouzarides, T. 2004. Histone deimination antagonizes arginine methylation. *Cell*. **118**(5), 545–553.

D’aquila, R.T., Bechtel, L.J., Videler, J.A., Eron, J.J., Gorczyca, P. and Kaplan, J.C. 1991. Maximizing sensitivity and specificity of PCR by preamplification heating. *Nucleic Acids Research*. **19**(13), 3749.

Davie, J.R. 1998. Covalent modifications of histones: Expression from chromatin templates. *Current Opinion in Genetics and Development*. **8**(2), 173–178.

Decatur, W.A. and Fournier, M.J. 2002. rRNA modifications and ribosome function. *Trends in Biochemical Sciences*. **27**(7), 344–51

DeJesus-Hernandez, M., Mackenzie, I.R., Boeve, B.F., Boxer, A.L., Baker, M., Rutherford, N.J., Nicholson, A.M., Finch, N.C.A., Flynn, H., Adamson, J., Kouri, N., Wojtas, A., Sengdy, P., Hsiung, G.Y.R., Karydas, A., Seeley, W.W., Josephs, K.A., Coppola, G., Geschwind, D.H., Wszolek, Z.K., Feldman, H., Knopman, D.S., Petersen, R.C., Miller, B.L., Dickson, D.W., Boylan, K.B., Graff-Radford, N.R. and Rademakers, R. 2011. Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron*. **72**(2), 245–256.

Delatte, B., Wang, F., Ngoc, L.V., Collignon, E., Bonvin, E., Deplus, R., Calonne, E., Hassabi, B., Putmans, P., Awe, S., Wetzel, C., Kreher, J., Soin, R., Creppe, C., Limbach, P.A., Gueydan, C., Kruys, V., Brehm, A., Minakhina, S., Defrance, M., Steward, R. and Fuks, F. 2016. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science*. **351**(6270), 282–285.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J. and Charpentier, E. 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*. **471**(7340), 602–607.

Demartini, D.R. 2013. A Short Overview of the Components in Mass Spectrometry Instrumentation for Proteomics Analyses. *Tandem Mass Spectrometry - Molecular Characterization*. **1**, 30–57.

Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P. and Moineau, S. 2008. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *Journal of Bacteriology*. **190**(4), 1390–1400.

Dixon, T.C., Meselson, M., Guillemin, J. and Hanna, P.C. 1999. Anthrax. *N Engl J Med*. **341**(11), 815–826.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., Sorek, R. and Rechavi, G. 2012. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*. **485**(7397), 201–206.

Donald, J.E., Kulp, D.W. and DeGrado, W.F. 2011. Salt bridges: Geometrically specific, designable interactions. *Proteins: Structure, Function and Bioinformatics*. **79**(3), 898–915.

Donnelly, C.J., Zhang, P.W., Pham, J.T., Heusler, A.R., Mistry, N.A., Vidensky, S., Daley, E.L., Poth, E.M., Hoover, B., Fines, D.M., Maragakis, N., Tienari, P.J., Petrucelli, L., Traynor, B.J., Wang, J., Rigo, F., Bennett, C.F., Blackshaw, S., Sattler, R. and Rothstein, J.D. 2013. RNA Toxicity from the ALS/FTD C9ORF72 Expansion Is Mitigated by Antisense Intervention. *Neuron*. **80**(2), 415–428.

Draper, D.E. 1999. Themes in RNA-protein recognition. *Journal of Molecular Biology*. **293**(2), 255–270.

Dreyfuss, G., Kim, V.N. and Kataoka, N. 2002. Messenger-RNA-binding proteins and the messages they carry. *Nature Reviews Molecular Cell Biology*. **3**(3), 195–205.

Dreyfuss, G., Swanson, M.S. and Piñol-Roma, S. 1988. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends in Biochemical Sciences*. **13**(3), 86–91.

Dudley, E. and Bond, L. 2014. Mass spectrometry analysis of nucleosides and nucleotides. *Mass Spectrometry Reviews*. **33**(4), 302–331.

van Duijn, E., Barbu, I.M., Barendregt, a., Jore, M.M., Wiedenheft, B., Lundgren, M., Westra, E.R., Brouns, S.J.J., Doudna, J. a., van der Oost, J. and Heck, a. J.R. 2012. Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in CRISPR-associated protein complexes from Escherichia coli and Pseudomonas aeruginosa. *Molecular & Cellular Proteomics*., **11**(11), 1430–1441.

Dupuis, M.È., Villion, M., Magadán, A.H. and Moineau, S. 2013. CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. *Nature Communications*. **4**, e2087.

Eng, J.K., McCormack, A.L. and Yates, J.R. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*. **5**(11), 976–89.

Fasanella, A. 2013. Bacillus anthracis, virulence factors, PCR, and interpretation of results. *Virulence*. **4**(8), 659–660.

Felder, S., Zhou, M., Hu, P., Ureña, J., Ullrich, A., Chaudhuri, M., White, M., Shoelson, S.E. and Schlessinger, J. 1993. SH2 domains exhibit high-affinity binding to tyrosine-phosphorylated peptides yet also exhibit rapid dissociation and exchange. *Molecular and cellular biology*. **13**(3), 1449–55.

Feng, Y., Maity, R., Whitelegge, J.P., Hadjikyriacou, A., Li, Z., Zurita-Lopez, C., Al-Hadid, Q., Clark, A.T., Bedford, M.T., Masson, J.Y. and Clarke, S.G. 2013. Mammalian protein arginine methyltransferase 7 (PRMT7) specifically targets RXR sites in lysine- and arginine-rich regions. *Journal of Biological Chemistry*. **288**(52), 37010–37025.

Finer-Moore, J., Czudnochowski, N., O’Connell, J.D., Wang, A.L. and Stroud, R.M.

2015. Crystal Structure of the Human tRNA m1A58 Methyltransferase-tRNA3LysComplex: Refolding of Substrate tRNA Allows Access to the Methylation Target. *Journal of Molecular Biology*. **427**(24), 3862–3876.

Fonfara, I., Le Rhun, A., Chylinski, K., Makarova, K.S., Lécrivain, A.L., Bzdrenga, J., Koonin, E. V. and Charpentier, E. 2014. Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Research*. **42**(4), 2577–2590.

Forde, A. and Fitzgerald, G.F. 1999. Bacteriophage defence systems in lactic acid bacteria *In*: *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*. **76**(1-4), 89–113.

Fronz, K., Otto, S., Kölbel, K., Kühn, U., Friedrich, H., Schierhorn, A., Beck-Sickinger, A.G., Ostareck-Lederer, A. and Wahle, E. 2008. Promiscuous modification of the nuclear poly(A)-binding protein by multiple protein-arginine methyltransferases does not affect the aggregation behavior. *Journal of Biological Chemistry*. **283**(29), 20408–20420.

Fu, L., Guerrero, C.R., Zhong, N., Amato, N.J., Liu, Y., Liu, S., Cai, Q., Ji, D., Jin, S.G., Niedernhofer, L.J., Pfeifer, G.P., Xu, G.L. and Wang, Y. 2014. Tet-mediated formation of 5-hydroxymethylcytosine in RNA. *Journal of the American Chemical Society*. **136**(33), 11582–11585.

Gaki, G.S. and Papavassiliou, A.G. 2014. Oxidative stress-induced signaling pathways implicated in the pathogenesis of Parkinson’s disease. *NeuroMolecular Medicine*. **16**(2), 217–230.

Garvie, C.W. and Wolberger, C. 2001. Recognition of specific DNA sequences. *Molecular Cell*. **8**(5), 937–946.

Gasiunas, G., Barrangou, R., Horvath, P. and Siksnys, V. 2012. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences*. **109**(39), E2579–2586.

Geoghegan, V., Guo, A., Trudgian, D., Thomas, B. and Acuto, O. 2015. Comprehensive identification of arginine methylation in primary T cells reveals regulatory roles in cell signalling. *Nature Communications*. **6**, e6758.

Gerstberger, S., Hafner, M. and Tuschl, T. 2014. A census of human RNA-binding proteins. *Nature Reviews Genetics*. **15**(12), 829–845.

Geuens, T., Bouhy, D. and Timmerman, V. 2016. The hnRNP family: insights into their role in health and disease. *Human Genetics*. **135**(8), 851–867.

Gijselinck, I., Van Langenhove, T., van der Zee, J., Sleegers, K., Philtjens, S., Kleinberger, G., Janssens, J., Bettens, K., Van Cauwenberghe, C., Pereson, S., Engelborghs, S., Sieben, A., De Jonghe, P., Vandenberghe, R., Santens, P., De Bleecker, J., Maes, G., Bäumer, V., Dillen, L., Joris, G., Cuijt, I., Corsmit, E., Elinck, E., Van Dongen, J., Vermeulen, S., Van den Broeck, M., Vaerenberg, C., Mattheijssens, M., Peeters, K., Robberecht, W., Cras, P., Martin, J.J., De Deyn, P.P., Cruts, M. and Van Broeckhoven, C. 2012. A C9orf72 promoter repeat expansion in a Flanders-Belgian cohort with disorders of the frontotemporal lobar degeneration-amyotrophic lateral sclerosis spectrum: A gene identification study. *The Lancet Neurology*. **11**(1), 54–65.

Gilbert, M.T., Knox, J.H. and Kaur, B. 1982. Porous glassy carbon, a new columns packing material for gas chromatography and high-performance liquid chromatography. *Chromatographia*. **16**(1), 138–146.

Glisovic, T., Bachorik, J.L., Yong, J. and Dreyfuss, G. 2008. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*. **582**(14), 1977–1986.

Gong, B., Shin, M., Sun, J., Jung, C.-H., Bolt, E.L., van der Oost, J. and Kim, J.-S. 2014. Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proceedings of the National Academy of Sciences*. **111**(46), 16359–16364.

Greenberg, J.R. 1979. Ultraviolet light-induced crosslinking of mRNA to proteins. *Nucleic Acids Research*. **6**(2), 715–732.

Grissa, I., Vergnaud, G. and Pourcel, C. 2007. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*. **8**,172.

Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P., Færgeman, N.J., Mann, M. and Jensen, O.N. 2005. Quantitative Phosphoproteomics Applied to the Yeast Pheromone Signaling Pathway. *Molecular & Cellular Proteomics*. **4**(3), 310–327.

Gryaznov, S.M. 2010. Oligonucleotide N3′→P5′ phosphoramidates and thio-phoshoramidates as potential therapeutic agents. *Chemistry and Biodiversity*. **7**(3), 477–493.

Guo, A., Gu, H., Zhou, J., Mulhern, D., Wang, Y., Lee, K.A., Yang, V., Aguiar, M., Kornhauser, J., Jia, X., Ren, J., Beausoleil, S.A., Silva, J.C., Vemulapalli, V., Bedford, M.T. and Comb, M.J. 2014. Immunoaffinity Enrichment and Mass Spectrometry Analysis of Protein Methylation. *Molecular & Cellular Proteomics*. **13**(1), 372–387.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat.Biotechnol.* **17**(10), 994–999.

Haeusler, A.R., Donnelly, C.J., Periz, G., Simko, E.A.J., Shaw, P.G., Kim, M.-S., Maragakis, N.J., Troncoso, J.C., Pandey, A., Sattler, R., Rothstein, J.D. and Wang, J. 2014. C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature*. **507**(7491), 195–200.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.C., Munschauer, M., Ulrich, A., Wardle, G.S., Dewell, S., Zavolan, M. and Tuschl, T. 2010. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*. **141**(1), 129–141.

Haft, D.H., Selengut, J., Mongodin, E.F. and Nelson, K.E. 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/cas subtypes exist in prokaryotic genomes. *PLoS Computational Biology*. **1**(6), 0474–0483.

Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M. and Terns, M.P. 2009. RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell*. **139**(5), 945–56

Hamma, T. and Ferré-D&apos;Amaré, A.R. 2006. Pseudouridine Synthases. *Chemistry and Biology*. **13**(11), 1125–1135.

Han, S.P., Tang, Y.H. and Smith, R. 2010. Functional diversity of the hnRNPs: past, present and perspectives. *Biochemical Journal*. **430**(3), 379–392.

Harcourt, E.M., Kietrys, A.M. and Kool, E.T. 2017. Chemical and structural effects of base modifications in messenger RNA. *Nature*. **541**(7637), 339–346.

Hardman, M. and Makarov, A.A. 2003. Interfacing the orbitrap mass analyzer to an electrospray ion source. *Analytical Chemistry*. **75**(7), 1699–1705.

El Hassan, M.A. and Calladine, C.R. 1996. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *Journal of Molecular Biology*. **259**(1), 95–103.

Hendsch, Z.S. and Tidor, B. 1994. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Science*. **3**(2), 211–226.

Hensman, D.J., Poulter, M., Beck, J., Hehir, J., Polke, J.M., Campbell, T., Adamson, G., Mudanohwo, E., McColgan, P., Haworth, A., Wild, E.J., Sweeney, M.G., Houlden, H., Mead, S. and Tabrizi, S.J. 2014. C9orf72 expansions are the most common genetic cause of Huntington disease phenocopies. *Neurology*. **82**(4), 292–299.

Hillenkamp, F., Karas, M., Beavis, R.C. and Chait, B.T. 1991. Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry of Biopolymers. *Analytical Chemistry*. **63**(24), 1193A-1203A.

Hoffmann, E. De and Stroobant, V. 2007. *Mass Spectrometry - Priniples and Applications.*

Hsu, P.J., Zhu, Y., Ma, H., Guo, Y., Shi, X., Liu, Y., Qi, M., Lu, Z., Shi, H., Wang, J., Cheng, Y., Luo, G., Dai, Q., Liu, M., Guo, X., Sha, J., Shen, B. and He, C. 2017. Ythdc2 is an N6 -methyladenosine binding protein that regulates mammalian spermatogenesis. *Cell Research*. **27**(9), 1115–1127.

Hung, M.L., Hautbergue, G.M., Snijders, A.P.L., Dickman, M.J. and Wilson, S.A. 2010. Arginine methylation of REF/ALY promotes efficient handover of mRNA to TAP/NXF1. *Nucleic Acids Research*. **38**(10), 3351–3361.

Hunter, C.A. 1996. Sequence-dependent DNA structure. *BioEssays*. **18**(2), 157–162.

Jackson, S.N., Wang, H.Y.J., Woods, A.S., Ugarov, M., Egan, T. and Schultz, J.A.

2005. Direct tissue analysis of phospholipids in rat brain using MALDI-TOFMS and MALDI-ion mobility-TOFMS. *Journal of the American Society for Mass Spectrometry*. **16**(2), 133–138.

Järvelin, A.I., Noerenberg, M., Davis, I. and Castello, A. 2016. The new (dis)order in RNA regulation. *Cell Communication and Signaling*. **14**(1), 9.

Jazurek, M., Ciesiolka, A., Starega-Roslan, J., Bilinska, K. and Krzyzosiak, W.J. 2016. Identifying proteins that bind to specific RNAs - Focus on simple repeat expansion diseases. *Nucleic Acids Research*. **44**(19), 9050–9070.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*. **337**(6096), 816–821.

Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A.T., Charpentier, E., Nogales, E. and Doudna, J.A. 2014. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science*. **343**(6176), 1247997.

Jore, M.M., Lundgren, M., Van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, Ü., Wurm, R., Wagner, R., Beijer, M.R., Barendregt, A., Zhou, K., Snijders, A.P.L., Dickman, M.J., Doudna, J.A., Boekema, E.J., Heck, A.J.R., Van Der Oost, J. and Brouns, S.J.J. 2011. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Structural and Molecular Biology*. **18**(5), 529–536.

Josephs, K.A., Hodges, J.R., Snowden, J.S., MacKenzie, I.R., Neumann, M., Mann, D.M. and Dickson, D.W. 2011. Neuropathological background of phenotypical variability in frontotemporal dementia. *Acta Neuropathologica*. **122**(2), 137–153.

Josse, J. and Kornberg, A. 1962. Glucosylation of deoxyribonucleic acid. III. alpha- and beta-Glucosyl transferases from T4-infected Escherichia coli. *The Journal of biological chemistry*. **237**, 1968–1976.

Juhász, A., Makai, S., Sebestyén, E., Tamás, L. and Balázs, E. 2011. Role of conserved non-coding regulatory elements in lMW glutenin gene expression. *PLoS ONE*. **6**(12), e29501

Kamal, S.M., Rashid, A.M., Bakar, M.A. and Ahad, M.A. 2011. Anthrax: An update. *Asian Pacific Journal of Tropical Biomedicine*. **1**(6), 496–501.

Karvelis, T., Gasiunas, G., Miksys, A., Barrangou, R., Horvath, P. and Siksnys, V. 2013. crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. *RNA Biology*. **10**(5), 841–851.

Keitel, W. a 2006. Recombinant protective antigen 102 (rPA102): profile of a second-generation anthrax vaccine. *Expert review of vaccines*. **5**, 417–430.

Kettenbach, A.N., Rush, J. and Gerber, S.A. 2011. Absolute quantification of protein and post-translational modification abundance with stable isotope–labeled synthetic peptides. *Nature Protocols*. **6**(2), 175–186.

Kiledjian, M. and Dreyfuss, G. 1992. Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. *The EMBO journal*. **11**(7), 2655–2664.

Kim, H.J., Kim, N.C., Wang, Y.-D., Scarborough, E.A., Moore, J., Diaz, Z., MacLea, K.S., Freibaum, B., Li, S., Molliex, A., Kanagaraj, A.P., Carter, R., Boylan, K.B., Wojtas, A.M., Rademakers, R., Pinkus, J.L., Greenberg, S.A., Trojanowski, J.Q., Traynor, B.J., Smith, B.N., Topp, S., Gkazi, A.-S., Miller, J., Shaw, C.E., Kottlors, M., Kirschner, J., Pestronk, A., Li, Y.R., Ford, A.F., Gitler, A.D., Benatar, M., King, O.D., Kimonis, V.E., Ross, E.D., Weihl, C.C., Shorter, J. and Taylor, J.P. 2013. Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature*. **495**(7442), 467–473.

Kim, M.-S. and Pandey, A. 2012. Electron transfer dissociation mass spectrometry in proteomics. *PROTEOMICS*. **12**(4–5), 530–542.

Kirkpatrick, D.S., Gerber, S.A. and Gygi, S.P. 2005. The absolute quantification strategy: A general procedure for the quantification of proteins and post-translational modifications. *Methods*. **35**(3 SPEC.ISS.), 265–273.

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., Nureki, O. and Zhang, F. 2015. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*. **517**(7536), 583–588.

Korlach, J. and Turner, S.W. 2012. Going beyond five bases in DNA sequencing. *Current Opinion in Structural Biology*. **22**(3), 251–261.

Kouzarides, T. 2007. SnapShot: Histone-Modifying Enzymes. *Cell*. **131**(4). 822

Kunin, V., Sorek, R. and Hugenholtz, P. 2007. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biology*. **8**(4), R61.

Laity, J.H., Lee, B.M. and Wright, P.E. 2001. Zinc finger proteins: New insights into structural and functional diversity. *Current Opinion in Structural Biology*. **11**(1), 39–46.

Leblanc, B.P. and Rodrigue, S. 2015. *DNA-protein interactions: Principles and protocols: Fourth Edition*.

Lee, J., Chen, H., Liu, T., Berkman, C.E. and Reilly, P.T.A. 2011. High resolution time-of-flight mass analysis of the entire range of intact singly-charged proteins. *Analytical Chemistry*. **83**(24), 9406–9412.

Lehman, I.R. and Pratt, E.A. 1960. On the structure of the glucosylated hydroxymethylcytosine nucleotides of coliphages T2, T4, and T6. *The Journal of biological chemistry*. **235**, 3254–3259.

Lehmann, K.A. and Bass, B.L. 2000. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. *Biochemistry*. **39**(42), 12875–12884.

Lewis, J.K., Wei, J. and Siuzdak, G. 2000. Matrix-assisted Laser Desorption / Ionization Mass Spectrometry in Peptide and Protein Analysis. *Encyclopedia of Analytical Chemistry*. 5880–5894.

Li, F., Mao, G., Tong, D., Huang, J., Gu, L., Yang, W. and Li, G.M. 2013. The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. *Cell*. **153**(3), 590–600.

Li, S. and Mason, C.E. 2014. The Pivotal Regulatory Landscape of RNA Modifications. *Annual Review of Genomics and Human Genetics*. **15**(1), 127–150.

Lima-Mendez, G., Toussaint, A. and Leplae, R. 2007. Analysis of the phage sequence space: The benefit of structured information. *Virology*. **365**(2), 241–249.

Lipps, H.J. and Rhodes, D. 2009. G-quadruplex structures: in vivo evidence and function. *Trends in Cell Biology*. **19**(8), 414–422.

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M. and Pan, T. 2015. N6-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature*. **518**(7540), 560–564.

Liu, Q. and Dreyfuss, G. 1995. In vivo and in vitro arginine methylation of RNA-binding proteins. *Molecular and cellular biology*. **15**(5), 2800–2808.

Liu, S., Ji, D., Cliffe, L., Sabatini, R. and Wang, Y. 2014. Quantitative mass spectrometry-based analysis of β-D-glucosyl-5-hydroxymethyluracil in genomic DNA of trypanosoma brucei. *Journal of the American Society for Mass Spectrometry*. **25**(10), 1763–1770.

Lößner, C., Warnken, U., Pscherer, A. and Schnölzer, M. 2011. Preventing arginine-to-proline conversion in a cell-line-independent manner during cell cultivation under stable isotope labeling by amino acids in cell culture (SILAC) conditions. *Analytical Biochemistry*. **412**(1), 123–125.

Lunde, B.M., Moore, C. and Varani, G. 2007. RNA-binding proteins: Modular design for efficient function. *Nature Reviews Molecular Cell Biology*. **8**(6), 479–490.

Luria, S.E. 1953. Host-induced modifications of viruses. *Cold Spring Harbor symposia on quantitative biology*. **18**, 237–244.

Luria, S.E. and Human, M.L. 1952. A nonhereditary, host-induced variation of bacterial viruses. *Journal of bacteriology*. **64**(4), 557–569.

Mackenzie, I.R.A., Rademakers, R. and Neumann, M. 2010. TDP-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia. *The Lancet Neurology*. **9**(10), 995–1007.

Maeda, M., Hasegawa, H., Sugiyama, M., Hyodo, T., Ito, S., Chen, D., Asano, E., Masuda, A., Hasegawa, Y., Hamaguchi, M. and Senga, T. 2016. Arginine methylation of ubiquitin-associated protein 2-like is required for the accurate distribution of chromosomes. *FASEB Journal*. **30**(1), 312–323.

Makarov, A. 2000. Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*. **72**(6), 1156–1162.

Makarov, A. and Scigelova, M. 2010. Coupling liquid chromatography to Orbitrap mass spectrometry. *Journal of Chromatography A*. **1217**(25), 3938–3945.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., van der Oost, J. and Koonin, E. V. 2011. Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology*. **9**(6), 467–477.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., Horvath, P., Moineau, S., Mojica, F.J.M., Terns, R.M., Terns, M.P., White, M.F., Yakunin, A.F., Garrett, R.A., Van Der Oost, J., Backofen, R. and Koonin, E. V. 2015. An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology*. **13**(11), 722–736.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M. 2013. RNA-guided human genome engineering via Cas9. *Science*. **339**(6121), 823–826.

Mann, M. and Jensen, O.N. 2003. Proteomic analysis of post-translational modifications. *Nature Biotechnology*. **21**(3), 255–261.

Marinus, M.G. and Casadesus, J. 2009. Roles of DNA adenine methylation in host-pathogen interactions: Mismatch repair, transcriptional regulation, and more. *FEMS Microbiology Reviews*. **33**(3), 488–503.

Marraffini, L.A. and Sontheimer, E.J. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*. **322**(5909), 1843-1845

Martin, G., Ostareck-Lederer, A., Chari, A., Neuenkirchen, N., Dettwiler, S., Blank, D., Ruegsegger, U., Fischer, U. and Keller, W. 2010. Arginine methylation in subunits of mammalian pre-mRNA cleavage factor I. *RNA*. **16**(8), 1646–1659.

Martinod, K., Demers, M., Fuchs, T.A., Wong, S.L., Brill, A., Gallant, M., Hu, J., Wang, Y. and Wagner, D.D. 2013. Neutrophil histone modification by peptidylarginine deiminase 4 is critical for deep vein thrombosis in mice. *Proceedings of the National Academy of Sciences*. **110**(21), 8674–8679.

Massari, M.E. and Murre, C. 2000. Helix-Loop-Helix Proteins: Regulators of Transcription in Eucaryotic Organisms. *Molecular and Cellular Biology*. **20**(2), 429–440.

Mathews, C.K., North, T.W. and Prem Veer Reddy, G. 1979. Multienzyme complexes in DNA precursor biosynthesis. *Advances in Enzyme Regulation*. **17**(C), 133–156.

Matthews, B.W., Ohlendorf, D.H., Anderson, W.F. and Takeda, Y. 1982. Structure of the DNA-binding region of lac repressor inferred from its homology with cro repressor. *Proceedings of the National Academy of Sciences*. **79**(5), pp.1428–1432.

McBride, A.E. and Silver, P.A. 2001. State of the Arg: Protein methylation at arginine comes of age. *Cell*. **106**(1), 5–8.

McGrath, S., Seegers, J.F.M.L., Fitzgerald, G.F. and Van Sinderen, D. 1999. Molecular characterization of a phage-encoded resistance system in Lactococcus lactis. *Applied and Environmental Microbiology*. **65**(5), 1891–1899.

Mehta, A.P., Li, H., Reed, S.A., Supekova, L., Javahishvili, T. and Schultz, P.G. 2016. Replacement of 2′-Deoxycytidine by 2′-Deoxycytidine Analogues in the E. coli Genome. *Journal of the American Chemical Society*. **138**(43), 14230–14233.

Melko, M. and Bardoni, B. 2010. The role of G-quadruplex in RNA metabolism: Involvement of FMRP and FMR2P. *Biochimie*. **92**(8), 919–926.

Meselson, M., Guillemin, J., Hugh-Jones, M., Langmuir, A., Popova, I., Shelokov, A. and Yampolskaya, O. 1994. The Sverdlovsk anthrax outbreak of 1979. *Science*. **266**(5188), 1202–1208.

Meynial, I., Paquet, V. and Combes, D. 1995. Simultaneous Separation of Nucleotides and Nucleotide Sugars Using an Ion-Pair Reversed-Phase HPLC: Application for Assaying Glycosyltransferase Activity. *Analytical Chemistry*. **67**(9), 1627–1631.

Mikesh, L.M., Ueberheide, B., Chi, A., Coon, J.J., Syka, J.E.P., Shabanowitz, J. and Hunt, D.F. 2006. The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et Biophysica Acta - Proteins and Proteomics*. **1764**(12), 1811–1822.

Miller, J., McLachlan, A.D. and Klug, A. 2001. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *Journal of Trace Elements in Experimental Medicine*. **14**(2), 157–169.

Milne, J.C., Furlong, D., Hanna, P.C., Wall, J.S. and Collier, R.J. 1994. Anthrax protective antigen forms oligomers during intoxication of mammalian cells. *Journal of Biological Chemistry*. **269**(32), 20607–20612.

Molloy, S.S., Bresnahan, P.A., Leppla, S.H., Klimpel, K.R. and Thomas, G. 1992. Human furin is a calcium-dependent serine endoprotease that recognizes the sequence Arg-X-X-Arg and efficiently cleaves anthrax toxin protective antigen. *Journal of Biological Chemistry*. **267**(23), 16396–16402.

Mori, K., Lammich, S., Mackenzie, I.R.A., Forné, I., Zilow, S., Kretzschmar, H., Edbauer, D., Janssens, J., Kleinberger, G., Cruts, M., Herms, J., Neumann, M., Van Broeckhoven, C., Arzberger, T. and Haass, C. 2013. HnRNP A3 binds to GGGGCC repeats and is a constituent of p62-positive/TDP43-negative inclusions in the hippocampus of patients with C9orf72 mutations. *Acta Neuropathologica*. **125**(3), 413–423.

Mori, K., Weng, S.-M., Arzberger, T., May, S., Rentzsch, K., Kremmer, E., Schmid, B., Kretzschmar, H.A., Cruts, M., Van Broeckhoven, C., Haass, C. and Edbauer, D. 2013. The C9orf72 GGGGCC Repeat Is Translated into Aggregating Dipeptide-Repeat Proteins in FTLD/ALS. *Science*. **339**(6125), 1335–1338.

Müller-Mcnicoll, M. and Neugebauer, K.M. 2013. How cells get the message: Dynamic assembly and function of mRNA-protein complexes. *Nature Reviews Genetics*. **14**(4), 275–287.

Neidle, S. 2009. The structures of quadruplex nucleic acids and their drug complexes. *Current Opinion in Structural Biology*. **19**(3), 239–250.

Nijhawan, A., Jain, M., Tyagi, A.K. and Khurana, J.P. 2008. Genomic survey and gene expression analysis of the basic leucine zipper transcription factor family in rice. *Plant physiology*. **146**(2), 333–350.

Niu, Y., Zhao, X., Yang, Y. and Wang, X. 2013. REVIEW N 6 -methyl-adenosine (m6A) in RNA : An Old Modification with A Novel Epigenetic Function. *Genomics, Proteomics & Bioinformatics*. **11**(1), 8–17.

Nuñez, J.K., Kranzusch, P.J., Noeske, J., Wright, A. V., Davies, C.W. and Doudna, J.A. 2014. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nature Structural and Molecular Biology*. **21**(6), 528–534.

Oberdoerffer, S., Moita, L.F., Neems, D., Freitas, R.P., Hacohen, N. and Rao, A. 2008. Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. *Science*. **321**(5889), 686–691.

Ohlsson, C., Nilsson, M.E., Tivesten, Å., Ryberg, H., Mellström, D., Karlsson, M.K., Ljunggren, Ö., Labrie, F., Orwoll, E.S., Lee, D.M., Pye, S.R., O'Neill, T.W., Finn, J.D., Adams, J.E., Ward, K.A., Boonen, S., Bartfai, G., Casanueva, F.F., Forti, G., Giwercman, A., Han, T.S., Huhtaniemi, I.T., Kula, K., Lean, M.E.J., Pendleton, N., Punab, M., Vanderschueren, D., Wu, F.C.W., Vandenput, L., Petrone, L., Corona, G., Borghs, H., Slowikowska-Hilczer, J., Walczak-Jedrzejowska, R., Silman, A., Steer, P., Lage, M., Castro, A.I., Földesi, I., Fejes, I., Korrovitz, P. and Jiang, M. 2013. Comparisons of immunoassay and mass spectrometry measurements of serum estradiol levels and their influence on clinical association studies in men. *Journal of Clinical Endocrinology and Metabolism*. **98**(6), E1907-1102

Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S. and Mann, M. 2007. Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*. **4**(9), 709–712.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A. and Mann, M. 2002. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular & Cellular Proteomics*. **1**(5), 376–386.

Ong, S., Kratchmarova, I. and Mann, M. 2003. Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). *Journal of proteome research*. **2**, 173–181.

Van Der Oost, J., Westra, E.R., Jackson, R.N. and Wiedenheft, B. 2014. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature Reviews Microbiology*. **12**(7), 479–492.

Orr, H.T. 2013. Toxic RNA as a driver of disease in a common form of ALS and dementia. *Proceedings of the National Academy of Sciences of the United States of*

America. **110**(19), 7533–7534.

Ozdilek, B.A., Thompson, V.F., Ahmed, N.S., White, C.I., Batey, R.T. and Schwartz, J.C. 2017. Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. *Nucleic Acids Research*. **45**(13), 7984–7996.

Pabo, C.O. and Sauer, R.T. 1984. Protein-DNA Recognition. *Annual Review of Biochemistry*. **53**(1), 293–321.

Paez-Espino, D., Sharon, I., Morovic, W., Stahl, B., Thomas, B.C., Barrangou, R. and Banfielda, J.F. 2015. CRISPR immunity drives rapid phage genome evolution in streptococcus thermophilus. *mBio*. **6**(2), 1–9.

Pahlich, S., Zakaryan, R.P. and Gehring, H. 2006. Protein arginine methylation: Cellular functions and methods of analysis. *Biochimica et biophysica acta*. **1764**(12), 1890–1903.

Paik, W.K., Paik, D.C. and Kim, S. 2007. Historical review: the field of protein methylation. *Trends in Biochemical Sciences*. **32**(3), 146–152.

Palumbo, A.M., Smith, S.A., Kalcic, C.L., Dantus, M., Stemmer, P.M. and Reid, G.E. 2011. Tandem mass spectrometry strategies for phosphoproteome analysis. *Mass Spectrometry Reviews*. **30**(4), 600–605.

Palzs, B. and Suhal, S. 2005. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews*. **24**(4), 508–548.

Park, S.-S., Wu, W.W., Zhou, Y., Shen, R.-F., Martin, B. and Maudsley, S. 2012. Effective correction of experimental errors in quantitative proteomics using stable isotope labeling by amino acids in cell culture (SILAC). *Journal of proteomics*. **75**(12), 3720–3732.

Pawluk, A., Staals, R.H.J., Taylor, C., Watson, B.N.J., Saha, S., Fineran, P.C., Maxwell, K.L. and Davidson, A.R. 2016. Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nature Microbiology*. **1**(8), 16085.

Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. **20**(18), 3551–67.

Piñol-Roma, S. and Dreyfuss, G. 1993. hnRNP proteins: localization and transport between the nucleus and the cytoplasm. *Trends in Cell Biology*. **3**(5), 151–155.

Pourcel, C., Salvignol, G. and Vergnaud, G. 2005. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*. **151**(3), 653–663.

Pul, Ü., Wurm, R., Arslan, Z., Geißen, R., Hofmann, N. and Wagner, R. 2010. Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS. *Molecular Microbiology*. **75**(6), 1495–1512.

Rajyaguru, P. and Parker, R. 2012. RGG motif proteins: Modulators of mRNA functional states. *Cell Cycle*. **11**(14), 2594–2599.

Raleigh, E.A. and Wilson, G. 1986. Escherichia coli K-12 restricts DNA containing 5-methylcytosine. *Proceedings of the National Academy of Sciences of the United States of America*. **83**(23), 9070–9074.

Reddy, K., Zamiri, B., Stanley, S.Y.R., Macgregor, R.B. and Pearson, C.E. 2013. The disease-associated r(GGGGCC)n repeat from the C9orf72 gene forms tract length-dependent uni- and multimolecular RNA G-quadruplex structures. *Journal of Biological Chemistry*. **288**(14), 9860–9866.

Renoux, A.J. and Todd, P.K. 2012. Neurodegeneration the RNA way. *Progress in Neurobiology*. **97**(2), 173–189.

Renton, A.E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J.R., Schymick, J.C., Laaksovirta, H., van Swieten, J.C., Myllykangas, L., Kalimo, H., Paetau, A., Abramzon, Y., Remes, A.M., Kaganovich, A., Scholz, S.W., Duckworth, J., Ding, J., Harmer, D.W., Hernandez, D.G., Johnson, J.O., Mok, K., Ryten, M., Trabzuni, D., Guerreiro, R.J., Orrell, R.W., Neal, J., Murray, A., Pearson, J., Jansen, I.E., Sondervan, D., Seelaar, H., Blake, D., Young, K., Halliwell, N., Callister, J.B., Toulson, G., Richardson, A., Gerhard, A., Snowden, J., Mann, D., Neary, D., Nalls, M.A., Peuralinna, T., Jansson, L., Isoviita, V.M., Kaivorinne, A.L., Hölttä-Vuori, M., Ikonen, E., Sulkava, R., Benatar, M., Wuu, J., Chiò, A., Restagno, G., Borghero, G., Sabatelli, M., Heckerman, D., Rogaeva, E., Zinman, L., Rothstein, J.D., Sendtner, M., Drepper, C., Eichler, E.E., Alkan, C., Abdullaev, Z., Pack, S.D., Dutra, A., Pak, E., Hardy, J., Singleton, A., Williams, N.M., Heutink, P., Pickering-Brown, S., Morris, H.R., Tienari, P.J. and Traynor, B.J. 2011. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*. **72**(2), 257–268.

Richmond, T.J. and Davey, C.A. 2003. The structure of DNA in the nucleosome core. *Nature*. **423**, 145–150.

Roepstorff, P. and Fohlman, J. 1984. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biological Mass Spectrometry*. **11**(11), 601.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B. and Mann, R.S. 2010. Origins of Specificity in Protein-DNA Recognition. *Annual Review of Biochemistry*. **79**, 233–269

Rollinson, S., Halliwell, N., Young, K., Callister, J.B., Toulson, G., Gibbons, L., Davidson, Y.S., Robinson, A.C., Gerhard, A., Richardson, A., Neary, D., Snowden, J., Mann, D.M.A. and Pickering-Brown, S.M. 2012. Analysis of the hexanucleotide repeat in C9ORF72 in Alzheimer's disease. *Neurobiology of Aging*. **33**(8), 1846.e5–1846.e6.

Roost, C., Lynch, S.R., Batista, P.J., Qu, K., Chang, H.Y. and Kool, E.T. 2015. Structure and thermodynamics of N^6-methyladenosine in RNA: A spring-loaded base modification. *Journal of the American Chemical Society*. **137**(5), 2107–2115.

Ross, P.L. 2004. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Molecular & Cellular Proteomics*.

3(12), 1154–1169.

Safra, M., Sas-Chen, A., Nir, R., Winkler, R., Nachshon, A., Bar-Yaacov, D., Erlacher, M., Rossmanith, W., Stern-Ginossar, N. and Schwartz, S. 2017. The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature*. **551**(7679), 251–255.

Samson, J.E., Magadán, A.H., Sabri, M. and Moineau, S. 2013. Revenge of the phages: Defeating bacterial defences. *Nature Reviews Microbiology*. **11**(10), 675–687.

Santos, A.L. and Lindner, A.B. 2017. Protein Posttranslational Modifications: Roles in Aging and Age-Related Disease. *Oxidative Medicine and Cellular Longevity*. 5716409.

Sashital, D.G., Wiedenheft, B. and Doudna, J.A. 2012. Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. *Molecular Cell*. **46**(5), 606–615.

Schwartz, S., Bernstein, D.A., Mumbach, M.R., Jovanovic, M., Herbst, R.H., León-Ricardo, B.X., Engreitz, J.M., Guttman, M., Satija, R., Lander, E.S., Fink, G. and Regev, A. 2014. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell*. **159**(1), 148–162.

Scott, L., Lamb, J., Smith, S. and Wheatley, D.N. 2000. Single amino acid (arginine) deprivation: rapid and selective death of cultured transformed and malignant cells. *British journal of cancer*. **83**(6), 800–10.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.J. and Severinov, K. 2011. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences*. **108**(25), 10098–10103.

Shimizu, H., Sato, K., Berberich, T., Miyazaki, A., Ozaki, R., Imai, R. and Kusano, T. 2005. LIP19, a basic region leucine zipper protein, is a fos-like molecular switch in the cold signaling of rice plants. *Plant and Cell Physiology*. **46**(10), 1623–1634.

Shlyakhov, E.N. and Rubinstein, E. 1994. Human live anthrax vaccine in the former USSR. *Vaccine*. **12**(8), 727–730.

Silverstein, T.P. 1998. The Real Reason Why Oil and Water Don’t Mix. *Journal of Chemical Education*. **75**(1), 116.

Sindelar, C. V, Hendsch, Z.S. and Tidor, B. 1998. Effects of salt bridges on protein structure and design. *Protein Science*. **7**(9), 1898–1914.

Sleno, L. and Volmer, D.A. 2004. Ion activation methods for tandem mass spectrometry. *Journal of Mass Spectrometry*. **39**(10), 1091–1112.

Sommer, B., Köhler, M., Sprengel, R. and Seeburg, P.H. 1991. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell*. **67**(1), 11–19.

Sorek, R., Lawrence, C.M. and Wiedenheft, B. 2013. CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. *Annual Review of Biochemistry*. **82**(1), 237–266.

Spitzer, J., Hafner, M., Landthaler, M., Ascano, M., Farazi, T., Wardle, G., Nusbaum, J., Khorshid, M., Burger, L., Zavolan, M. and Tuschl, T. 2014. PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): A Step-By-Step Protocol to the Transcriptome-Wide Identification of Binding Sites of RNA-Binding Proteins. *Methods in Enzymology*. **539**, 113–161.

Staals, R.H.J., Zhu, Y., Taylor, D.W., Kornfeld, J.E., Sharma, K., Barendregt, A., Koehorst, J.J., Vlot, M., Neupane, N., Varossieau, K., Sakamoto, K., Suzuki, T., Dohmae, N., Yokoyama, S., Schaap, P.J., Urlaub, H., Heck, A.J.R., Nogales, E., Doudna, J.A., Shinkai, A. and vanderOost, J. 2014. RNA Targeting by the Type III-A CRISPR-Cas Csm Complex of Thermus thermophilus. *Molecular Cell*. **56**, 518–530.

Steen, H. and Mann, M. 2004. The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*. **5**(9), 699–711.

Stern, A. and Sorek, R. 2011. The phage-host arms race: Shaping the evolution of microbes. *BioEssays*. **33**(1), 43–51.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C. and Doudna, J.A. 2014. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*. **507**(7490), 62–67.

Strotskaya, A., Savitskaya, E., Metlitskaya, A., Morozova, N., Datsenko, K.A., Semenova, E. and Severinov, K. 2017. The action of Escherichia coli CRISPR-Cas system on lytic bacteriophages with different lifestyles and development strategies. *Nucleic acids research*. **45**(4), 1946–1957.

Sturm, R., Sheynkman, G., Booth, C., Smith, L.M., Pedersen, J.A. and Li, L. 2012. Absolute quantification of prion protein (90-231) using stable isotope-labeled chymotryptic peptide standards in a lc-mrm aqua workflow. *Journal of the American Society for Mass Spectrometry*. **23**(9), 1522–1533.

Sun, W.J., Li, J.H., Liu, S., Wu, J., Zhou, H., Qu, L.H. and Yang, J.H. 2016. RMBase: A resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Research*. **44**(D1), D259–D265.

Swarts, D.C., van der Oost, J. and Jinek, M. 2017. Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. *Molecular Cell*. **66**(2), 221-233.e4.

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J.S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., Lu, Z., Ye, Z., Zhu, Q., Wysocka, J., Ye, Y., Khochbin, S., Ren, B. and Zhao, Y. 2011. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*. **146**(6), 1016–1028.

Terns, R.M. and Terns, M.P. 2014. CRISPR-based technologies: Prokaryotic defense weapons repurposed. *Trends in Genetics*. **30**(3), 111–118.

Thandapani, P., O'Connor, T.R., Bailey, T.L. and Richard, S. 2013. Defining the RGG/RG Motif. *Molecular Cell*. **50**(5), 613–623.

Thiagarajan, D., Dev, R.R. and Khosla, S. 2011. The DNA methyltranferase Dnmt2 participates in RNA processing during cellular stress. *Epigenetics*. **6**(1), 103–113.

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T. and Hamon, C. 2003. Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*. **75**(8), 1895–1904.

Tian, B., Bevilacqua, P.C., Diegelman-Parente, A. and Mathews, M.B. 2004. The double-stranded-RNA-binding motif: Interference and much more. *Nature Reviews Molecular Cell Biology*. **5**(12), 1013–1023.

Tomiya, N., Ailor, E., Lawrence, S.M., Betenbaugh, M.J. and Lee, Y.C. 2001. Determination of Nucleotides and Sugar Nucleotides Involved in Protein Glycosylation by High-Performance Anion-Exchange Chromatography: Sugar Nucleotide Contents in Cultured Insect Cells and Mammalian Cells. *Analytical Biochemistry*. **293**(1), 129–137.

Trouw, L.A., Huizinga, T.W.J. and Toes, R.E.M. 2013. Autoimmunity in rheumatoid arthritis: different antigens—common principles. *Annals of the Rheumatic Diseases*. **72**(suppl 2), ii132–ii136.

Van Hoof, D., Pinkse, M.W.H., Oostwaard, D.W. Van, Mummery, C.L., Heck, A.J.R. and Krijgsveld, J. 2007. An experimental correction for arginine-to-proline conversion artifacts in SILAC-based quantitative proteomics. *Nature Methods*. **4**(9), 677–678.

Vinson, C.R., Sigler, P.B. and McKnight, S.L. 1989. Scissors-grip model for DNA recognition by a family of leucine zipper proteins. *Science (New York, N.Y.)*. **246**(4932), 911–916.

Vlot, M., Houkes, J., Lochs, S.J.A., Swarts, D.C., Zheng, P., Kunne, T., Mohanraju, P., Anders, C., Jinek, M., van der Oost, J., Dickman, M.J. and Brouns, S.J.J. 2017. Bacteriophage DNA glucosylation impairs target DNA binding by type I and II but not by type V CRISPR–Cas effector complexes. *Nucleic Acids Research*. **46**(2), 873–885.

Vossenaar, E.R., Radstake, T.R.D., Van Der Heijden, A., Van Mansum, M.A.M., Dieteren, C., De Rooij, D.J., Barrera, P., Zendman, A.J.W. and Van Venrooij, W.J. 2004. Expression and activity of citrullinating peptidylarginine deiminase enzymes in monocytes and macrophages. *Annals of the Rheumatic Diseases*. **63**(4), 373–381.

Vossenaar, E.R., Zendman, A.J.W., Van Venrooij, W.J. and Pruijn, G.J.M. 2003. PAD, a growing family of citrullinating enzymes: Genes, features and involvement in disease. *BioEssays*. **25**(11), 1106–1118.

Walsh, C.T., Garneau-Tsodikova, S. and Gatto, G.J. 2005. Protein posttranslational modifications: The chemistry of proteome diversifications. *Angewandte Chemie - International Edition*. **44**(45), 7342–7372.

Wang, L., Chen, S., Xu, T., Taghizadeh, K., Wishnok, J.S., Zhou, X., You, D., Deng, Z. and Dedon, P.C. 2007. Phosphorothioation of DNA in bacteria by dnd genes. *Nature Chemical Biology*. **3**(11), 709–710.

Wang, R. and Li, H. 2012. The mysterious RAMP proteins and their roles in small RNA-based immunity. *Protein Science*. **21**(4), 463–470.

Wang, S. and Wang, Y. 2013. Peptidylarginine deiminases in citrullination, gene regulation, health and pathogenesis. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. **1829**(10), 1126–1135.

Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., Ren, B., Pan, T. and He, C. 2013. N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*. **505**(7481), 117–120.

Wang, X., Zhao, B.S., Roundtree, I.A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K., Shi, H. and He, C. 2015. N6-methyladenosine modulates messenger RNA translation efficiency. *Cell*. **161**(6), 1388–1399.

Wang, Y., Wysocka, J., Sayegh, J., Lee, Y.H., Pertin, J.R., Leonelli, L., Sonbuchner, L.S., McDonald, C.H., Cook, R.G., Dou, Y., Roeder, R.G., Clarke, S., Stallcup, M.R., Allis, C.D. and Coonrod, S.A. 2004. Human PAD4 regulates histone arginine methylation levels via demethylimination. *Science*. **306**(5694), 279–283.

Warren, R.A.J. 1980. Modified Bases in Bacteriophage DNAs. *Annual Review of Microbiology*. **34**(1), 137–158.

Warzecha, C.C., Sato, T.K., Nabet, B., Hogenesch, J.B. and Carstens, R.P. 2009. ESRP1 and ESRP2 Are Epithelial Cell-Type-Specific Regulators of FGFR2 Splicing. *Molecular Cell*. **33**(5), 591–601.

Watson, A. and Keir, D. 1994. Information on which to Base Assessments of Risk from Environments Contaminated with Anthrax Spores. *Epidemiology and Infection*. **113**, 479–490.

Weigele, P. and Raleigh, E.A. 2016. Biosynthesis and Function of Modified Bases in Bacteria and Their Viruses. *Chemical Reviews*. **116**(20), 12655–12687.

West, C., Elfakir, C. and Lafosse, M. 2010. Porous graphitic carbon: A versatile stationary phase for liquid chromatography. *Journal of Chromatography A*. **1217**(19), 3201–3216.

Wetzel, R. and Soöll, D. 1977. Analogs of methionyl-tRNA synthetase substrates containing photolabile groups. *Nucleic Acids Research*. **4**(5), 1681–1694.

Wiedenheft, B., Lander, G.C., Zhou, K., Jore, M.M., Brouns, S.J.J., Van Der Oost, J., Doudna, J.A. and Nogales, E. 2011. Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature*. **477**(7365), 486–489.

Wiedenheft, B., Sternberg, S.H. and Doudna, J.A. 2012. RNA-guided genetic silencing systems in bacteria and archaea. *Nature*. **482**(7385), 331–338.

Wieser, A., Schneider, L., Jung, J. and Schubert, S. 2012. MALDI-TOF MS in microbiological diagnostics-identification of microorganisms and beyond (mini review). *Applied Microbiology and Biotechnology*. **93**(3), 965–974.

Wiesner, J., Premsler, T. and Sickmann, A. 2008. Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. *Proteomics*. **8**(21), 4466–4483.

Wigington, C.H., Sonderegger, D., Brussaard, C.P.D., Buchan, A., Finke, J.F., Fuhrman, J.A., Lennon, J.T., Middelboe, M., Suttle, C.A., Stock, C., Wilson, W.H., Wommack, K.E., Wilhelm, S.W. and Weitz, J.S. 2016. Re-examination of the relationship between marine virus and microbial cell abundances. *Nature Microbiology*. **1**(3), 15024.

Wilkins, B.J., Rall, N.A., Ostwal, Y., Kruitwagen, T., Hiragami-Hamada, K., Winkler, M., Barral, Y., Fischle, W. and Neumann, H. 2014. A cascade of histone modifications induces chromatin condensation in mitosis. *Science*. **343**(6166), 77–80.

Yang, Y. and Bedford, M.T. 2013. Protein arginine methyltransferases and cancer. *Nature Reviews Cancer*. **13**(1), 37–50.

Yang, Y., Hadjikyriacou, A., Xia, Z., Gayatri, S., Kim, D., Zurita-Lopez, C., Kelly, R., Guo, A., Li, W., Clarke, S.G. and Bedford, M.T. 2015. PRMT9 is a Type II methyltransferase that methylates the splicing factor SAP145. *Nature Communications*. **6,** 6428.

Yaung, S.J., Esvelt, K.M. and Church, G.M. 2014. CRISPR/Cas9-mediated phage resistance is not impeded by the DNA modifications of phage T4. *PLoS ONE*. **9**(6), e98811.

Yen, T.Y., Charles, M.J. and Voyksner, R.D. 1996. Processes that affect electrospray ionization-mass spectrometry of nucleobases and nucleosides. *Journal of the American Society for Mass Spectrometry*. **7**(11), 1106–1108.

Yosef, I., Goren, M.G. and Qimron, U. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Research*. **40**(12), 5569–5576.

Yue, Y., Liu, J. and He, C. 2015. RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes and Development*. **29**(13), 1343–1355.

van der Zee, J., Gijselinck, I., Dillen, L., Van Langenhove, T., Theuns, J., Engelborghs, S., Philtjens, S., Vandenbulcke, M., Sleegers, K., Sieben, A., Bäumer, V., Maes, G., Corsmit, E., Borroni, B., Padovani, A., Archetti, S., Perneczky, R., Diehl-Schmid, J., de Mendonça, A., Miltenberger-Miltenyi, G., Pereira, S., Pimentel, J., Nacmias, B., Bagnoli, S., Sorbi, S., Graff, C., Chiang, H.H., Westerlund, M., Sanchez-Valle, R., Llado, A., Gelpi, E., Santana, I., Almeida, M.R., Santiago, B., Frisoni, G., Zanetti, O., Bonvicini, C., Synofzik, M., Maetzler, W., vom Hagen, J.M., Schöls, L., Heneka, M.T., Jessen, F., Matej, R., Parobkova, E., Kovacs, G.G., Ströbel, T., Sarafov, S., Tournev, I., Jordanova, A., Danek, A., Arzberger, T., Fabrizi, G.M., Testi, S., Salmon, E., Santens, P., Martin, J.J., Cras, P., Vandenberghe, R., De Deyn, P.P., Cruts, M., Van Broeckhoven, C., Ramirez, A., Kurzwelly, D., Sachtleben, C., Mairer, W., Firmo, C., Antonell, A., Molinuevo, J., Forsell, C., Lillius, L., Kinhult Ståhlbom, A., Thonberg, H., Nennesmo, I., Börjesson-Hanson, A., Bessi, V., Piaceri, I., Helena Ribeiro, M., Oliveira, C., Massano, J., Garret, C., Pires, P., Danel, A., Ferrari, S. and Cavallaro, T. 2013. A Pan-European Study of the C9orf72 Repeat Associated with FTLD: Geographic Prevalence, Genomic Instability, and Intermediate Repeats. *Human Mutation*. **34**(2), 363–373.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., Van Der Oost, J., Regev, A., Koonin, E. V. and Zhang, F. 2015. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell*. **163**(3), 759–771.

Zhang, X., Liu, Z., Yi, J., Tang, H., Xing, J., Yu, M., Tong, T., Shang, Y., Gorospe, M. and Wang, W. 2012. The tRNA methyltransferase NSun2 stabilizes p16 INK4 mRNA by methylating the 3′2-untranslated region of p16. *Nature Communications*. **3**, 712.

Zu, T., Gibbens, B., Doty, N.S., Gomes-pereira, M., Huguet, A. and Stone, M.D. 2010. Non-ATG – initiated translation directed by microsatellite expansions. *Pnas*. **108**(1), 260–265.

Zubarev, R.A. and Makarov, A. 2013. Orbitrap Mass Spectrometry. *Anal Chem*. **85**(11), 5288–5296.

Appendices

Chapter 4 Appendices

Plasmid sequences

```
>Operon_minus_B-gt_(PvuI/PacI)_in_pMK-RQ Confirmed by sequencing
CTAAATTGTAAGCGTTAATATTTTGTTAAAATTCGCGTTAAATTTTTGTTAAATCAGCTCATTTTTTAACC
AATAGGCCGAAATCGGCAAAATCCCTTATAAATCAAAAGAATAGACCGAGATAGGGTTGAGTGGCCGC
TACAGGGCGCTCCCATTCGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGTTTCGGTGCGGGCCTCTT
CGCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTT
TCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAGCGCGACGTAATACGACTCACTATAGGGCGAAT
TGAAGGAAGGCCGTCAAGGCCGCATATGCATGCATGAATTCGATGCATCCAGGCATCAAATAAAACGA
AAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCTCTACTAGAGTC
ACACTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATAGAATTCAACCATGGAAGATCTTTTAAGAAG
GAGATATACATATGGCTCACTTTAATGAATGTGCTCATTTGATCGAAGGTGTTGATAAAGCTCAAAATG
AATACTGGGATATTCTCGGTGATGAAAAAGATCCGCTGCAAGTTATGCTTGATATGCAGCGGTTTTTAC
AGATTCGTTTGGCTAATGTCCGCGAATACTGCTATCATCCAGATAAATTAGAAACTGCCGGTGATGTTG
TTTCTTGGATGCGTGAACAAAAAGACTGTATTGATGATGAATTTCGCGAACTTCTGACTTCTCTTGGTGA
AATGTCACGTGGTGAAAAAGAAGCTTCTGCTGTATGGAAAAAATGGAAAGCACGTTATATTGAAGCGC
AAGAAAAACGCATTGATGAAATGTCCCCCGAAGACCAGCTCGAAATTAAATTTGAGCTTGTGGATATA
TTTCATTTCGTATTAAATATGTTTGTTGGCCTTGGAATGAATGCGGAAGAAATCTTTAAACTTTATTATC
TGAAGAACAAACATAATTTTGAACGTCAAGATAATGGATATTAACAATTGGTCGACAAATAATAAAAA
AGCCGGATTAATAATCTGGCTTTTTATATTCTCTCGTACGCGATTAAGGCGCGCCTTATAGTACCTTTAG
TGTATTTTTAATTTTAGAAAAAAGTTCTTCAAGAGAACCATCGTTTGTAATTACTAAATCGCCATCACGA
ATTGGCAATCCAGCTTCTGTAATATGTGTATCATTGGATTTTTGACCAGGACGAACTACATGAATTACTG
TAGCACCCATCGCCCTAGCCGCATCCATTTCATGATCTTGACGGGTATCAGGAACGATATAATAATCAT
AACCTGAGTTAAATTTATCAAGATAATCTAAAGCAAATAATTTTACCCAGTACATGCGGTCGAAGTTAT
TAACAATCAAATCCGTACCTAGGGCTTGCATCAGACGACGGACTGACCATTGATCTTCAATATTATTTA
TAACGTCAGTAATCTTATTAAATGCTACGAAATTAACTGATTCTTTTCCTTCGTCATCAAAAACAAACAC
ACCTTTAATTGGGCTTTTACCATTAAGATAACAAAATGCTTGTTCCATAATCGTGATTACTTCTAATTTA
GTCAGATTTAAATTAGTCTCACGATCATAGTCAATTCCTTCAAACTCTTTACGAGTTAAGCAAGGATAG
TCAGTGTTTGCTGCAAATACTCCCCATGCATAAGCCAATGCATCCTTAATGGGACCAGCAAGTTGGTAT
TTAACTGCAGAATAATTGCTCATGATAAAATCAGCAGTAGTATCTTTTCCACTACGCTTTACACCGCTTA
AAAAGATTAGTTTCATATGTATATCTCCTTCTACGCGTCCCGGGTTAAGCGTATTTTCCTACATAATCTT
TTTTCGAAATATGTGTTTCGCCAGTTTTCCACCAATGATCAACCAAATAAAAATGACGAGAATACACAT
GTAGGCTTCCAACATTCCATATAATGGAACCTGCTTTATACTGGCGAGTTGAATCACCTGCATTCAAAT
CAGATACTAATTTATCTAATACGTATTTTTGCCATGCATAATCATTACGGAATCCGAAGACCACGTCATT
TGAGCGCATGTTAACAACCGCATTGATTTTCTTGTCACGAATCAGGTATTGTACTGTATTCGTGCACATG
AAATCTGACATACCATCTTTATTATAGTCAAACTGCATAGATGGACGAGTATAAATCATGATACCACGT
CGAGAATCAGGATTTTGACCAAGTTCAGCTAAACACATGTCATACTGAGCATAGTTATCTTCTGACCAG
ATAGCCCAACCATAATTCGAGTTAATTTCACCTTTAGAAGATGCTACTTGTTGCCAAATCTTCGGTGTTT
CACCCGGAATATCTTTAACGAACAAGCTTTTAGATTTATACCATTCAAGTTCACGCTGAATGTATTCATC
ATTAAGAGCGCCAAAAATAAACGGTTCATCTGCTACAAATGATGCGCCAATAATTTCAATAGTTTTAAC
ACCTGTTTTATCAACTACGAAATCTTTTTCTTTTAATGCAAGCCCCAAATGAAGACGGATTTCTTCAACT
GTCATAGAGTCACTAATCATATGTATATCTCCTTCTTCCGGAGAACTTTTCAATTCGCGTTAAACAAAAT
TATTACTAGAGGGAAACCGCCAGGGTCTCCCTACGACCAGTCTAAAAAGCGCCTCAATTCGCGACCTTC
TCGTTACTGACAGGAAAATGGGCCATTGGCAACCAGGGAAAGATGAACGTGATGATGTTCACAATTTG
CTGAATTGTGGTGGACGAATTTGGATCCCTGGGCCTCATGGGCCTTCCTTTCACTGCCCGCTTTCCAGTC
GGGAAACCTGTCGTGCCAGCTGCATTAACATGGTCATAGCTGTTTCCTTGCGTATTGGGCGCTCTCCGCT
TCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGGTAAAGCCTGGGGTGCCTAATGAGCAAAAGGCC
AGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGAC
GAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGG
CGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGC
CTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTC
GTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAAC
TATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATT
AGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAG
AAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTG
ATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAA
AAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACG
TTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAG
TTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTATTAGAAAAATTCATCCAGCAG
ACGATAAAACGCAATACGCTGGCTATCCGGTGCCGCAATGCCATACAGCACCAGAAAACGATCCGCCC
ATTCGCCGCCCAGTTCTTCCGCAATATCACGGGTGGCCAGCGCAATATCCTGATAACGATCCGCCACGC
```

CCAGACGGCCGCAATCAATAAAGCCGCTAAAACGGCCATTTTCCACCATAATGTTCGGCAGGCACGCA
TCACCATGGGTCACCACCAGATCTTCGCCATCCGGCATGCTCGCTTTCAGACGCGCAAACAGCTCTGCC
GGTGCCAGGCCCTGATGTTCTTCATCCAGATCATCCTGATCCACCAGGCCCGCTTCCATACGGGTACGC
GCACGTTCAATACGATGTTTCGCCTGATGATCAAACGGACAGGTCGCCGGGTCCAGGGTATGCAGACG
ACGCATGGCATCCGCCATAATGCTCACTTTTTCTGCCGGCGCCAGATGGCTAGACAGCAGATCCTGACC
CGGCACTTCGCCCAGCAGCAGCCAATCACGGCCCGCTTCGGTCACCACATCCAGCACCGCCGCACACG
GAACACCGGTGGTGGCCAGCCAGCTCAGACGCGCCGCTTCATCCTGCAGCTCGTTCAGCGCACCGCTCA
GATCGGTTTTCACAAACAGCACCGGACGACCCTGCGCGCTCAGACGAAACACCGCCGCATCAGAGCAG
CCAATGGTCTGCTGCGCCCAATCATAGCCAAACAGACGTTCCACCCACGCTGCCGGGCTACCCGCATGC
AGGCCATCCTGTTCAATCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCAT
GAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAA
AAGTGCCAC

>pMK-0_SfiI_2 null (Circularized)
GGGCCTTCCTTTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAACATGGTCATAG
CTGTTTCCTTGCGTATTGGGCGCTCTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGGTA
AAGCCTGGGGTGCCTAATGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCT
GGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCG
AAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCC
GACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCA
CGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTT
CAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCG
CCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTT
GAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGT
TACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTT
TGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGG
GTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTT
CACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTC
TGACAGTTATTAGAAAAATTCATCCAGCAGACGATAAAACGCAATACGCTGGCTATCCGGTGCCGCAA
TGCCATACAGCACCAGAAAACGATCCGCCCATTCGCCGCCCAGTTCTTCCGCAATATCACGGGTGGCCA
GCGCAATATCCTGATAACGATCCGCCACGCCCAGACGGCCGCAATCAATAAAGCCGCTAAAACGGCCA
TTTTCCACCATAATGTTCGGCAGGCACGCATCACCATGGGTCACCACCAGATCTTCGCCATCCGGCATG
CTCGCTTTCAGACGCGCAAACAGCTCTGCCGGTGCCAGGCCCTGATGTTCTTCATCCAGATCATCCTGAT
CCACCAGGCCCGCTTCCATACGGGTACGCGCACGTTCAATACGATGTTTCGCCTGATGATCAAACGGAC
AGGTCGCCGGGTCCAGGGTATGCAGACGACGCATGGCATCCGCCATAATGCTCACTTTTTCTGCCGGCG
CCAGATGGCTAGACAGCAGATCCTGACCCGGCACTTCGCCCAGCAGCAGCCAATCACGGCCCGCTTCG
GTCACCACATCCAGCACCGCCGCACACGGAACACCGGTGGTGGCCAGCCAGCTCAGACGCGCCGCTTC
ATCCTGCAGCTCGTTCAGCGCACCGCTCAGATCGGTTTTCACAAACAGCACCGGACGACCCTGCGCGCT
CAGACGAAACACCGCCGCATCAGAGCAGCCAATGGTCTGCTGCGCCCAATCATAGCCAAACAGACGTT
CCACCCACGCTGCCGGGCTACCCGCATGCAGGCCATCCTGTTCAATCATACTCTTCCTTTTTCAATATTA
TTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACA
AATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTAAATTGTAAGCGTTAATATTTTGTTAAAA
TTCGCGTTAAATTTTTGTTAAATCAGCTCATTTTTTAACCAATAGGCCGAAATCGGCAAAATCCCTTATA
AATCAAAAGAATAGACCGAGATAGGGTTGAGTGGCCGCTACAGGGCGCTCCCATTCGCCATTCAGGCT
GCGCAACTGTTGGGAAGGGCGTTTCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGA
TGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCC
AGTGAGCGCGACGTAATACGACTCACTATAGGGCGAATTGAAGGAAGGCCG

Original data for the calculation of pHmC hmC incorporation (peaks for hmC and dC are highlighted)

Data set 1

No.	Ret.Time	Area	Amount	Type	Height	Rel.Area	Resolution
	min	mAU*min	n.a.		mAU	%	
75	3.631	5.5952	n.a.	BM	50.779	10.41	n.a.
76	3.912	0.0017	n.a.	M	0.114	0	n.a.
77	4.056	0.7039	n.a.	M	7.43	1.31	n.a.
78	4.171	1.3833	n.a.	MB	10.577	2.57	n.a.
79	4.391	0	n.a.	BMB	0.013	0	4.72
80	4.406	0	n.a.	BMB	0.011	0	1.87
81	4.585	0.9404	n.a.	BMB	7.843	1.75	n.a.
82	4.85	0.0005	n.a.	BM	0.051	0	n.a.
83	4.858	0.0001	n.a.	Ru	0.011	0	n.a.

Data set 2

No.	Ret.Time	Area	Amount	Type	Height	Rel.Area	Resolution
	min	mAU*min	n.a.		mAU	%	
53	3.583	14.5705	n.a.	M	93.835	11.54	1.92
54	3.751	0.8113	n.a.	Rd	11.068	0.64	n.a.
55	4.013	3.8038	n.a.	M	27.827	3.01	n.a.
56	4.173	0.0088	n.a.	M	1.477	0.01	n.a.
57	4.361	2.3697	n.a.	M	18.274	1.88	n.a.
58	4.578	0.0004	n.a.	M	0.067	0	n.a.
59	4.585	0.0007	n.a.	MB	0.05	0	n.a.

Chapter 5 Appendices

Digital version of this thesis includes the protein group lists below (Excel files):

5-1: pulldowns mRNP capture (PAR-CLIP)

5-2: pulldowns mRNP capture (CLIP)

5-3: mRNA binding pulldown with AdOx treatment (AdOx heavy/ Ctrl light)

5-4: mRNA binding pulldown with AdOx treatment (AdOx light/ Ctrl heavy)

5-5: WCE with AdOx treatment (AdOx heavy/ Ctrl light)

5-6: WCE with AdOx treatment (Ctrl heavy/ AdOx light)

5-7: mRNA binding pulldown with m^6A treatment (m^6A heavy/ Ctrl light)

5-8: mRNA binding pulldown with m^6A treatment (m^6A light/ Ctrl heavy)

Chapter 6 Appendices

Mascot identified proteins in HeLa nuclear extracts

NUCL_HUMAN
HNRH1_HUMAN
HNRH2_HUMAN
HNRPF_HUMAN
K2C1_HUMAN
HSP7C_HUMAN
PARP1_HUMAN
TBB5_HUMAN
TBA1C_HUMAN
HNRPU_HUMAN
RBBP4_HUMAN
TBB4B_HUMAN
TBB4A_HUMAN
RBBP7_HUMAN
CCAR1_HUMAN
HSP72_HUMAN
YBOX1_HUMAN
HS71L_HUMAN
K1C10_HUMAN
ACL6A_HUMAN
TBA4B_HUMAN
HS71A_HUMAN
LA_HUMAN
CHD4_HUMAN
TBB6_HUMAN
HSP76_HUMAN
YBOX3_HUMAN
YBOX2_HUMAN
GRP78_HUMAN
PUF60_HUMAN
RN213_HUMAN
RYR2_HUMAN
IMDH2_HUMAN
RBM47_HUMAN
E2F8_HUMAN
TBB8_HUMAN
K22E_HUMAN
CHD3_HUMAN
SF3A1_HUMAN
TBB1_HUMAN
MTA2_HUMAN
P66A_HUMAN
K2C6B_HUMAN
FUS_HUMAN
U2AF2_HUMAN
HNRL2_HUMAN
RUVB2_HUMAN
K2C1B_HUMAN
NOLC1_HUMAN
EF1A1_HUMAN
TF3C4_HUMAN
GRSF1_HUMAN
RUVB1_HUMAN
CHD5_HUMAN
HNRPK_HUMAN
P66B_HUMAN
SF3B1_HUMAN
K1C9_HUMAN
TF3C1_HUMAN
HDAC2_HUMAN
ALBU_HUMAN
DDX1_HUMAN
SUZ12_HUMAN
HDAC1_HUMAN
SF3B3_HUMAN
THUM1_HUMAN
SMCA4_HUMAN
MBD2_HUMAN
EF1A2_HUMAN
IF2GL_HUMAN
HNRPM_HUMAN
RTCB_HUMAN
RCC1_HUMAN
TIF1B_HUMAN
DDX21_HUMAN
CSRP1_HUMAN
RCC2_HUMAN
TF3C3_HUMAN
SMCA5_HUMAN
SMCA1_HUMAN
DHX36_HUMAN
RBP56_HUMAN
IQGA1_HUMAN
XRCC6_HUMAN
SRSF6_HUMAN
DDX5_HUMAN
ANXA3_HUMAN
SMRC2_HUMAN
XRN2_HUMAN
MTA1_HUMAN
MTA3_HUMAN
LMNA_HUMAN
ASPM_HUMAN
SRSF4_HUMAN
CDC5L_HUMAN
ZNF35_HUMAN
PCDB5_HUMAN
K2C4_HUMAN
GFAP_HUMAN
K2C8_HUMAN
K2C80_HUMAN
K2C7_HUMAN
SFPQ_HUMAN
RPA49_HUMAN
FAKD3_HUMAN
DDX23_HUMAN
NSUN2_HUMAN
SF3B2_HUMAN
TF3C2_HUMAN
K1C14_HUMAN
K1C17_HUMAN
XRCC5_HUMAN
DX39B_HUMAN
SNUT1_HUMAN
PRP19_HUMAN
K1C27_HUMAN
K1C25_HUMAN
K1C28_HUMAN
SMRC1_HUMAN
G3ST4_HUMAN
MMS22_HUMAN
AT11C_HUMAN
AT11A_HUMAN
LARP1_HUMAN
EWS_HUMAN
PLOD1_HUMAN
HNRH3_HUMAN
GLTP_HUMAN
EFCB5_HUMAN
SMCA2_HUMAN
RPA1_HUMAN
ABCF1_HUMAN
HNRL1_HUMAN
SF3A3_HUMAN
IF16_HUMAN
RPA2_HUMAN
ZN207_HUMAN
CTSRG_HUMAN
HCFC1_HUMAN
OBSCN_HUMAN
ADDG_HUMAN
SETD4_HUMAN
UIMC1_HUMAN
BAIP2_HUMAN
PIGP_HUMAN
TRHDE_HUMAN
G3BP2_HUMAN
KRA53_HUMAN
ANKZ1_HUMAN
HIRA_HUMAN
KI21B_HUMAN
WDR78_HUMAN
O11A1_HUMAN
DYH8_HUMAN
SH3R1_HUMAN

Digital version of this thesis includes protein group lists (Excel files):

6-1 : GGGGCC repeat binding pulldown with AdOx treatment (AdOx heavy/ Ctrl light)

6-2 : GGGGCC repeat binding pulldown with AdOx treatment (AdOx light /Ctrl heavy)

6-3 : WCE (input for *in vitro* pulldown) with AdOx treatment (AdOx heavy/ Ctrl light)

6-4 : WCE (input for *in vitro* pulldown) with AdOx treatment (Ctrl heavy/ AdOx light)

6-5 : GGGGCC repeat binding pulldown with PADI4 treatment (PADI4 heavy/ Ctrl light)

6-6 : GGGGCC repeat binding pulldown with PADI4 treatment (PADI4 light /Ctrl heavy)

6-7 : WCE (input for *in vitro* pulldown) with PADI4 treatment (PADI4 heavy/ Ctrl light)

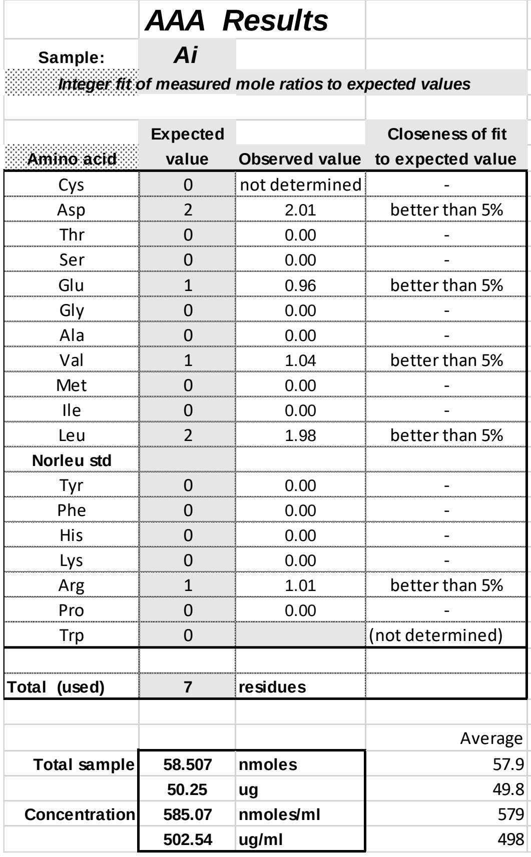
6-8 : WCE (input for *in vitro* pulldown) with PADI4 treatment (PADI4 heavy/ AdOx light)

6-9 : In-solution digest, GGGGCC repeat binding pulldown with AdOx treatment (AdOx heavy/ Ctrl light)

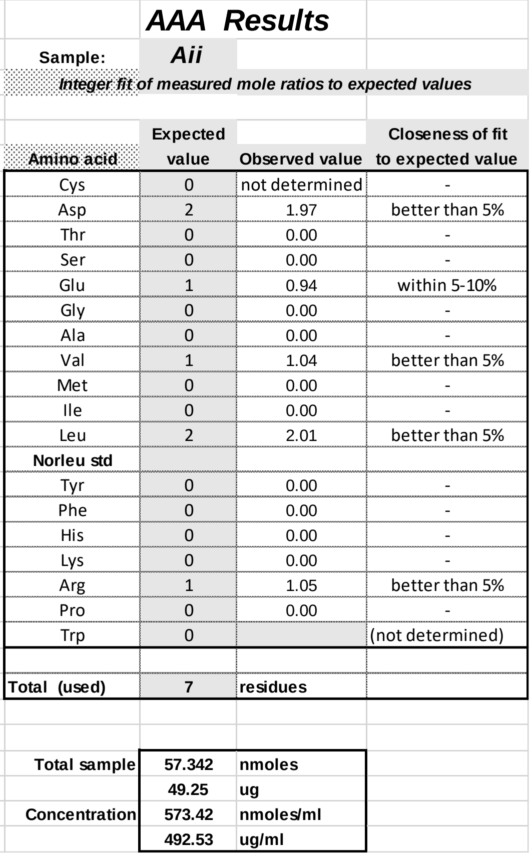
6-10 : In-solution digest, GGGGCC repeat binding pulldown with AdOx treatment (AdOx light /Ctrl heavy)

Chapter 7 Appendices

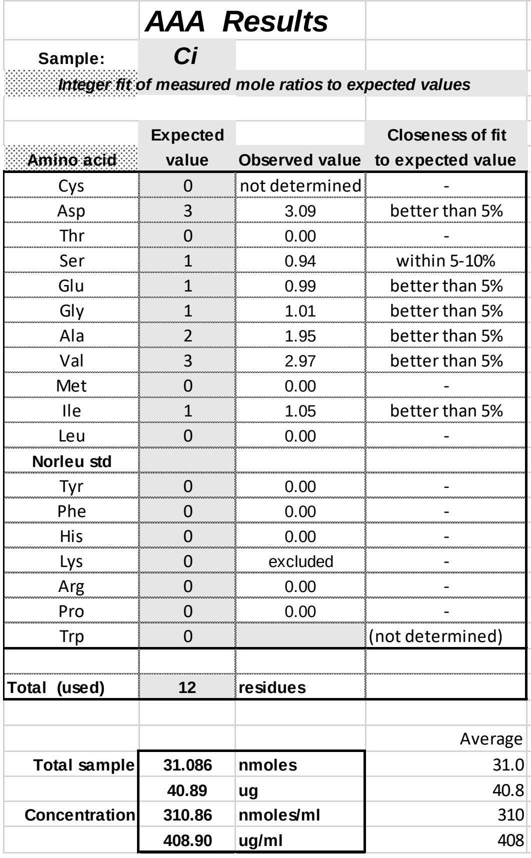
Amino acid analysis results of SIS peptides.



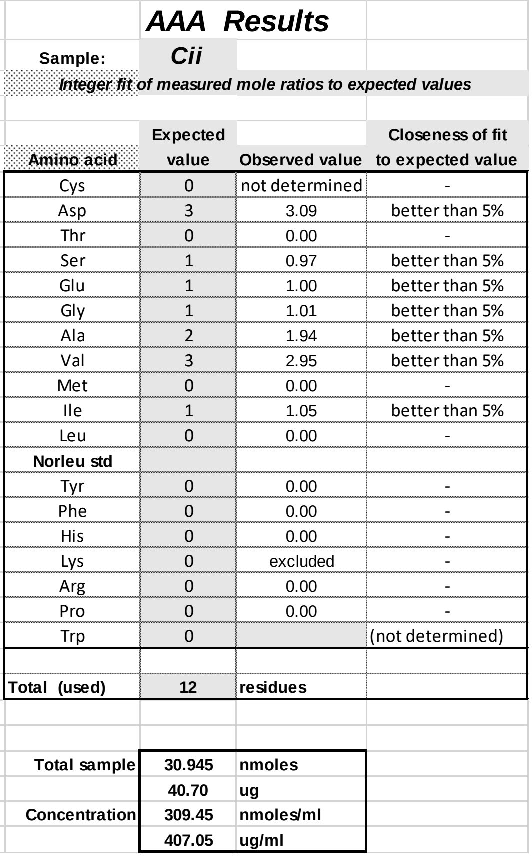
	AAA Results		
Sample:	***Ai***		
Integer fit of measured mole ratios to expected values			
	Expected		**Closeness of fit**
Amino acid	**value**	**Observed value**	**to expected value**
Cys	0	not determined	-
Asp	2	2.01	better than 5%
Thr	0	0.00	-
Ser	0	0.00	-
Glu	1	0.96	better than 5%
Gly	0	0.00	-
Ala	0	0.00	-
Val	1	1.04	better than 5%
Met	0	0.00	-
Ile	0	0.00	-
Leu	2	1.98	better than 5%
Norleu std			
Tyr	0	0.00	-
Phe	0	0.00	-
His	0	0.00	-
Lys	0	0.00	-
Arg	1	1.01	better than 5%
Pro	0	0.00	-
Trp	0		(not determined)
Total (used)	**7**	**residues**	
			Average
Total sample	**58.507**	**nmoles**	57.9
	50.25	**ug**	49.8
Concentration	**585.07**	**nmoles/ml**	579
	502.54	**ug/ml**	498



	AAA Results		
Sample:	***Aii***		
Integer fit of measured mole ratios to expected values			
	Expected		**Closeness of fit**
Amino acid	**value**	**Observed value**	**to expected value**
Cys	0	not determined	-
Asp	2	1.97	better than 5%
Thr	0	0.00	-
Ser	0	0.00	-
Glu	1	0.94	within 5-10%
Gly	0	0.00	-
Ala	0	0.00	-
Val	1	1.04	better than 5%
Met	0	0.00	-
Ile	0	0.00	-
Leu	2	2.01	better than 5%
Norleu std			
Tyr	0	0.00	-
Phe	0	0.00	-
His	0	0.00	-
Lys	0	0.00	-
Arg	1	1.05	better than 5%
Pro	0	0.00	-
Trp	0		(not determined)
Total (used)	**7**	**residues**	
Total sample	**57.342**	**nmoles**	
	49.25	**ug**	
Concentration	**573.42**	**nmoles/ml**	
	492.53	**ug/ml**	



	AAA Results		
Sample:	***Ci***		
Integer fit of measured mole ratios to expected values			
	Expected		**Closeness of fit**
Amino acid	**value**	**Observed value**	**to expected value**
Cys	0	not determined	-
Asp	3	3.09	better than 5%
Thr	0	0.00	-
Ser	1	0.94	within 5-10%
Glu	1	0.99	better than 5%
Gly	1	1.01	better than 5%
Ala	2	1.95	better than 5%
Val	3	2.97	better than 5%
Met	0	0.00	-
Ile	1	1.05	better than 5%
Leu	0	0.00	-
Norleu std			
Tyr	0	0.00	-
Phe	0	0.00	-
His	0	0.00	-
Lys	0	excluded	-
Arg	0	0.00	-
Pro	0	0.00	-
Trp	0		(not determined)
Total (used)	**12**	**residues**	
			Average
Total sample	**31.086**	**nmoles**	31.0
	40.89	**ug**	40.8
Concentration	**310.86**	**nmoles/ml**	310
	408.90	**ug/ml**	408



	AAA Results		
Sample:	***Cii***		
Integer fit of measured mole ratios to expected values			
	Expected		**Closeness of fit**
Amino acid	**value**	**Observed value**	**to expected value**
Cys	0	not determined	-
Asp	3	3.09	better than 5%
Thr	0	0.00	-
Ser	1	0.97	better than 5%
Glu	1	1.00	better than 5%
Gly	1	1.01	better than 5%
Ala	2	1.94	better than 5%
Val	3	2.95	better than 5%
Met	0	0.00	-
Ile	1	1.05	better than 5%
Leu	0	0.00	-
Norleu std			
Tyr	0	0.00	-
Phe	0	0.00	-
His	0	0.00	-
Lys	0	excluded	-
Arg	0	0.00	-
Pro	0	0.00	-
Trp	0		(not determined)
Total (used)	**12**	**residues**	
Total sample	**30.945**	**nmoles**	
	40.70	**ug**	
Concentration	**309.45**	**nmoles/ml**	
	407.05	**ug/ml**	

AAA Results

Sample: *Di*

Integer fit of measured mole ratios to expected values

Amino acid	Expected value	Observed value	Closeness of fit to expected value
Cys	0	not determined	-
Asp	0	excluded	-
Thr	2	1.93	better than 5%
Ser	0	0.00	-
Glu	0	excluded	-
Gly	0	0.00	-
Ala	1	0.94	within 5-10%
Val	0	0.00	-
Met	0	0.00	-
Ile	1	1.02	better than 5%
Leu	1	1.04	better than 5%
Norleu std			
Tyr	0	0.00	-
Phe	0	0.00	-
His	0	0.00	-
Lys	1	1.07	within 5-10%
Arg	0	0.00	-
Pro	0	0.00	-
Trp	0		(not determined)
Total (used)	**6**	**residues**	

			Average
Total sample	**71.110**	**nmoles**	70.9
	63.15	**ug**	62.9
Concentration	**711.10**	**nmoles/ml**	709
	631.46	**ug/ml**	629

AAA Results

Sample: *Dii*

Integer fit of measured mole ratios to expected values

Amino acid	Expected value	Observed value	Closeness of fit to expected value
Cys	0	not determined	-
Asp	0	excluded	-
Thr	2	1.92	better than 5%
Ser	0	0.00	-
Glu	0	excluded	-
Gly	0	0.00	-
Ala	1	0.95	within 5-10%
Val	0	0.00	-
Met	0	0.00	-
Ile	1	1.02	better than 5%
Leu	1	1.04	better than 5%
Norleu std			
Tyr	0	0.00	-
Phe	0	0.00	-
His	0	0.00	-
Lys	1	1.08	within 5-10%
Arg	0	0.00	-
Pro	0	0.00	-
Trp	0		(not determined)
Total (used)	**6**	**residues**	

Total sample	**70.638**	**nmoles**
	62.73	**ug**
Concentration	**706.38**	**nmoles/ml**
	627.27	**ug/ml**

AUC raw data for anthrax vaccine AQUA

rPA, amaZon

2.7 µl rPA from: 5ul diluted to 1150ul(6.9mg/ml to 30ng/ul)

Abundance data	Area under curve		
	0.25 pmol SIS	1 pmol SIS	4 pmol SIS
DLNLVER (L)	260970464	110919864	86823921
DLNLVER	68328922	102773052	365438016
NNIAVGADESVVK (L)	216167200	117704212	129995502
NNIAVGADESVVK (H)	56717988	105704103	537164224
NQTLATIK (L)	223245664	145230288	143049696
NQTLATIK (H)	41594568	122894056	517512000

5.4 µl rPA from: 5ul diluted to 1150ul(6.9mg/ml to 30ng/ul)

Abundance data	Area under curve		
	0.5 pmol SIS	2 pmol SIS	8 pmol SIS
DLNLVER (L)	172146976	204548848	106481012
DLNLVER (H)	52151768	233225936	484754994
NNIAVGADESVVK (L)	174924192	152491424	128980650
NNIAVGADESVVK (H)	45107324	155042672	601835984
NQTLATIK (L)	324284240	202689755	271940832
NQTLATIK (H)	54437650	189143928	686841664

Filtrate (PA) amaZon

2.7 µl Filtrate concentrated from 500ul to 85ul using PALL 3k filter

Abundance data	Area under curve		
	0.25 pmol SIS	1 pmol SIS	4 pmol SIS
DLNLVER (L)	313689056	319812224	284551972
DLNLVER (H)	70443072	283095584	1138871680
NNIAVGADESVVK (L)	465213280	498458656	337467641
NNIAVGADESVVK (H)	106745542	399826848	1030283444
NQTLATIK (L)	218976656	263639936	317464128
NQTLATIK (H)	34097064	215055200	818770560

5.4 µl Filtrate concentrated from 500ul to 85ul using PALL 3k filter

Abudnance data	Area under curve		
	0.5 pmol SIS	2 pmol SIS	8 pmol SIS
DLNLVER (L)	450720800	418667232	497555996
DLNLVER(H)	145381872	418633824	1954779136
NNIAVGADESVVK (L)	852881344	942887232	575489326
NNIAVGADESVVK (H)	118491810	662767616	1839847528
NQTLATIK (L)	447634240	435566016	591700036
NQTLATIK (H)	68536880	324383616	1797512832

rPA maXis

2.7 µl rPA from: 5ul diluted to 1150ul(6.9mg/ml to 30ng/ul)

Abundance data	Area under curve		
	0.25 pmol SIS	1 pmol SIS	4 pmol SIS
DLNLVER (L)	62297404	64407388	50606272
DLNLVER (H)	17273158	74564160	275207040
NNIAVGADESVVK (L)	66965860	68473712	52264348
NNIAVGADESVVK (H)	17317730	71810488	239496656
NQTLATIK (L)	82199920	95108272	83088576
NQTLATIK (H)	17816412	77143936	314065664

5.4 µl rPA from: 5ul diluted to 1150ul(6.9mg/ml to 30ng/ul)

Abundance data	Area under curve		
	0.5 pmol SIS	2 pmol SIS	8 pmol SIS
DLNLVER (L)	117993800	105856048	80590176
DLNLVER (H)	32080354	125053368	427692544
NNIAVGADESVVK (L)	113911384	98483912	79610304
nNIAVGADESVVK (H)	29727072	107010688	388597696
NQTLATIK (L)	164430144	183730464	157040560
NQTLATIK (H)	32766274	141948432	505276768

Filtrate (PA) maXis

2.7 µl Filtrate concentrated from 500ul to 85ul using PALL 3k filter

Abundance data	Area under curve		
	0.25 pmol SIS	1 pmol SIS	4 pmol SIS
DLNLVER (L)	23889822	25430236	23407346
DLNLVER (H)	6144884	24164500	85093976
NNIAVGADESVVK (L)	32139584	33098752	30012660
NNIAVGADESVVK (H)	6920506	24528512	87549024
NQTLATIK (L)	37942648	36595188	35249180
NQTLATIK (H)	7410402	25973960	99214960

5.4 µl Filtrate concentrated from 500ul to 85ul using PALL 3k filter

Abundance data	Area under curve		
	0.5 pmol SIS	2 pmol SIS	8 pmol SIS
DLNLVER (L)	33171210	31455618	29469278
DLNLVER (H)	8976668	33333534	127578440
NNIAVGADESVVK (L)	54218596	48863892	43149284
NNIAVGADESVVK (H)	10477907	38978396	138282256
NQTLATIK (L)	60270208	58774952	55762612
NQTLATIK (H)	10798835	42460904	168244144