Carbamylation across the Mammalian Proteome

LAURA EMMA HEATH

Carbamylation across the Mammalian Proteome

A thesis submitted for the degree of

Doctor of Philosophy

Emma Heath

Department of Biosciences

2024

Abstract

Carbon dioxide (CO_2) is a ubiquitous bioactive gas involved in regulating multiple mammalian biochemical pathways. CO_2 regulates diverse biological processes through various sensing mechanisms, including the carbamylation post-translational modification (PTM). Carbamylation in mammals has been implicated in the ventilatory response by haemoglobin, cellular conductance by connexin 26 and proteolysis by ubiquitin (Ub). CO_2-protein interactions remain an active area of research, with emerging technologies profiling these modifications across the proteome.

This thesis presents a mammalian proteome screen and identifies carbamylation sites across the HEK293 cell line. Proteome coverage was increased 7-fold using a fractionation-based workflow, and a method for carbamate validation was developed. Nine novel carbamylation sites were identified across two distinct screening datasets. Following the identification of carbamylated Ub K48 and Histone H3K79, the biological relevance of carbamylation at these targets was investigated.

Proteolysis targeting chimeras (PROTACs) are an emerging therapeutic technology which relies on the ubiquitin-proteasome system to target specific proteins for degradation. A luminescence assay was used to assess the degradation efficiencies of seven bromodomain containing 4 (BRD4) targeting PROTACs in HEK293 and nine SWI/SNF related, matrix-associated, actin dependent regulator of chromatin, subfamily a, member 2 (SMARCA2) targeting PROTACs in NCIH838 when exposed to normoxic and hypercapnic CO_2. The data revealed that the PROTACs' dose-response is independent of CO_2 concentration.

Across mass spectrometry (MS) datasets, eleven novel histone carbamylation sites were identified, and carbamylation at H3K79 was selected for further study. H3K79 methylation is a well-characterised histone PTM that results in DNA transcription activation. An *in-vitro* and *in-cellulo* assay assessed H3K79 methylation under elevated CO_2. The data indicated that H3K79 methylation is

enhanced under elevated inorganic carbon and that H3K79 methylation mediated transcriptional change is dependent on the partial pressure of CO_2 (PCO_2).

Declaration

The research within this thesis was carried out from October 2019 to February 2024 at Durham University and AstraZeneca. The work presented is my own unless otherwise indicated by a statement or citation. Collaborative work was carried out in the optimisation stages of Chapter 3. As indicated in Chapter 5, the RNA sequencing work was supported by Cambridge Genomic Services. None of the material has previously been submitted for any other qualification.

Statement of Copyright

Acknowledgements

Thank you to all the people who have supported me with the work presented in this thesis. Firstly, I would like to thank my supervisors, Martin Cann, Jarrod Walsh, and Finn Holding. This endeavour would not have been possible without Martin's optimism and perseverance. Thank you, Martin, for the opportunity to conduct this work and for helping me grow as a scientific researcher. Thank you, Jarrod and Finn, for your mentorship, meeting with me regularly, introductions to AZ contacts and propelling the vision of this project. Thanks to both AZ and the BBSRC for funding this project and the industrial opportunity provided.

Special thanks to those who helped me in the lab. Firstly, Victoria who laid the foundation for this research, Lysnay for support with bioinformatics and Rachel for supporting my research at HTS. I am thankful to Adrian for his guidance on mass spectrometry.

Thank you to everyone in lab 234, office 231 and HTS for the interesting scientific conversations over cake. I am particularly grateful to Dori and Yasmine for relating to the PhD experience and for the fun adventures we had throughout the past four years. Thank you to Kathryn, for weekend trips along the train line and showing me Edinburgh's best vegan pancake spot. To Rachael for being there for me since school and beyond.

To my family for all your support throughout my life and in this academic endeavour. In particular, thanks to my mum for always being there for me and your reassurance. To my dad for displaying the value of hard work and perseverance no matter the circumstance. To my granny and grandad for all the special moments shared and thought-provoking conversations.

Finally, to my husband Matthew, my biggest supporter, favourite adventure partner, and best friend. Thanks for sharing in both the joys and challenges of research and joining me on the pet-sitting trips during thesis writing. I could not have done this without you.

Table of Contents

List of Abbreviations

AF	ALL1-Fused gene from chromosome protein
AGE	Advanced glycosylation end product
ALL	Acute lymphocytic leukaemia
AML	Acute myeloid leukaemia
AMP	Adenosine monophosphate
AMPK	AMP kinase
ANOVA	Analysis of variance
APC	Adenomatosis polyposis coli
AQP	Aquaporin
ARDs	Acute respiratory distress syndrome
ARP	Assay ready plates
ATP	Adenosine triphosphate
AZ	AstraZeneca
BAM	Binary alignment map
BAT	Brown adipose tissue
BCA	Bicinchoninic acid
bcl	Binary base call
BD	Bromodomain
BKCa	Calcium-activated potassium channels
BME	β-mercaptoethanol
bp	Base pair
BRD4	Bromodomain containing protein 4
BSA	Bovine serum albumin
C	Contaminant
C18	Carbon 18
CA	Carbonic anhydrase
CAA	Number of amino acids in common
CAF1	Chromatin assembly factor 1

cAMP	Cyclic adenosine monophosphate
CAND1	Cullin-associated and neddylation-dissociated 1
Car4	Carbonic anhydrase 4
cdc34	Cell division cycle 34
CHAPS	3-((3-cholamidopropyl) dimethylammonio)-propane sulfonate)
CHIP-seq	Chromatin immunoprecipitation sequencing
Ci	Inorganic Carbon
CI	Confidence interval
CIC2	Chloride channel 2
CID	Collision induced dissociation
CK1α	Casein kinase 1α
CMG	Compound Management Group
CML	Carboxymethyl-lysine
c-MYC	Cellular myelocytomatosis
COPD	Chronic obstructive pulmonary disease
Cq	Quantitation cycle number
CR	Concentration response
CRBN	Cereblon
CREB	cAMP response binding element
CRL	Cullin
CryoEM	Cryogenic electron microscopy
CTD	C terminal domain
CV	Column Volumes
D	DMSO
Da	Daltons
DDB1	DNA–binding protein 1
DEG	Differentially expressed gene
DMEM	Dulbecco's modified medium
DMSO	Dimethyl sulfoxide
DNAse	Deoxyribonuclease

DOT1L	Disruptor of telomeric silencing 1-like
DotCom	DOT1L complex
DRC	Dose Response Curve
dsDNA	Double stranded DNA
DTT	Dithiothreitol
E. coli	*Escherichia coli*
EGFR	Epidermal growth factor receptor
ELISA	Enzyme-linked immunosorbent assay
endA1	Endonuclease A1
ENL	Eleven Nineteen-leukaemia
EPZ-5676	Pinometostat
ES	Embryonic Stem Cells
ESI	Electrospray ionisation
FBS	Foetal bovine serum
FDR	False Discovery Rate
FoxO3a	Forkhead box O3
FRET	Fluorescence resonance energy transfer
Fzd9	Frizzled class receptor
GC-D	Guanylyl cyclase
gDNA	genomic DNA
GO	Gene ontology
GSK3	Glycogen synthase kinase 3
Gtf	Gene transfer format
h	Hours
H1	Histone H1
H2A	Histone H2A
H2B	Histone H2B
H3	Histone H3
H3K79	Histone H3 lysine 79
H4	Histone H4
HA	Hydroxylamine

Hb	Haemoglobin
HECT	Homologous to E6AP C terminus
HEK293	Human Embryonic Kidney 293
HiBiT	High bioluminescence tag
HIF1α	Hypoxia Inducible Factor 1 Subunit Alpha
HILLIC	Hydrophilic interaction liquid chromatography
HMTs	Histone methyltransferases
Hoxa9	Homeobox A9
HPLC	High performance liquid chromatography
Hrp	Horseradish peroxidase
HTS	High Throughput Screening
IC_{50}	Half-maximal inhibitory concentration
ICAM1	Intercellular adhesion molecule-1
ID	Identification
IDH2	Isocitrate dehydrogenase 2
IgG	Immunoglobulin G
IKK	Inhibitory-kB kinase
IL-8	Interleukin-8
IPA	Isopropyl alcohol
IPTG	Isopropyl β-D-1-thiogalactopyranoside
K	Lysine
Kav	Partition coefficient
K_d	Binding affinity
kDa	Kilodaltons
K_M	Michaelmas constant
LacI	Lac repressor
LB	Lysogeny broth
LCMSMS	Liquid chromatography tandem mass spectrometry
LgBiT	Large bioluminescence tag
logFC	Log fold change
LPS	Lipopolysaccharide

m/z	Mass to charge ratio
M	Molar
MAD	Median absolute deviation
mAu	milli absorbance
MCTs	Multiple comparison tests
Meis1	Myeloid ecotropic viral integration site 1
mg	Milligrams
mL	Millilitres
mM	Millimolar
MLL	Mixed lineage leukaemia
Mod	Modification
mRNA	Messenger RNA
MTase	Methyltransferase
MTT	3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide
MURF1	Muscle-specific Ring Finger Protein 1
Mw	Molecular weight
NAPL1	Nucleosome assembly protein-like 1
NC	Without contaminants
NCIH838	Lung cell line from a white male with stage 3B adenocarcinoma
NCP	Nucleosome core particle
NDK	Nucleoside diphosphate kinase
Nedd8	Neuronal precursor cell-expressed developmentally down-regulated protein 8
NF-κB	Nuclear factor kappa B
ng	Nanograms
NIK	NF-kB inducing kinase
Nm	Nanometres
NMR	Nuclear magnetic resonance
NTP	Nuclear triphosphate pools

nTPM	normalised number of transcripts
OD_{600}	Optical density at 600 nanometers.
OH	Hydroxyl
ORF	Open reading frame
ori	Origin of replication
P	Pinometostat
PA	Propionic acid
PBS	Phosphate-buffered saline
pBS	Phagemid
PC	Principal component
PCA	Principal component analysis
PCO_2	Partial pressure carbon dioxide
Pct	Percentage
PDB	Protein data bank
pET	Protein expression vector
PHD	Prolyl hydroxylase domain
PI	Isoelectric point
pK_b	Base dissociation constant
POI	Protein of interest
PP2A	Protein phosphatase 2A
PPM	Parts per million
PROTAC	Proteolysis targeting chimera
PSM	Peptide to spectrum match
PTM	Post-translational modification
Q	Quaternary amino group
qPCR	Quantitative polymerase chain reaction
Q-TOF	Quadrupole-Time of flight
RING	Really Interesting New Gene
RBR	RING between RING
Rbx	RING box protein
RE	Restriction enzyme

recA	Recombinase A
Rh	Rhesus
RIL	Arginine, isoleucine, leucine
rlog	regularised log
RNApol	RNA polymerase
RNAse	Ribonuclease
RNA-seq	RNA sequencing
RPL19	Ribosomal protein L19
rpm	Revolutions per minute
RPTPγ	Receptor protein tyrosine phosphatase γ
RT	Reverse transcriptase
RUBICON	Rubicon Autophagy Regulator
RZ'	Robust Z’
s	Seconds
S: B	signal-to-background ratio
SAH	S-adenosyl-L-homocysteine
SAM	S-adenosyl methionine
SAR	Structure-activity relationship
SBS	Sequencing by Synthesis
SCF	Skp1-cullin 1-F-box
SDS-PAGE	Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis
SIR3	Silent Information Regulator complex
SMARCA2	SWI/SNF related, matrix-associated, actin-dependent regulator of chromatin, subfamily a, member 2
S.O.C.	Super optimal broth
SP	Sulphopropyl
Spry2	Sprouty2
STV	Sequence temperature values
SWI/SNF	Switch/Sucrose Non-Fermentable
TCA	Tricarboxylic acid
TEAB	Triethylammonium bicarbonate

TEO	Triethyloxonium tetrafluoroborate
TFA	Trifluoracetic acid
TLR	Toll-like receptor
TMM	Trimmed median of means
TRIM	Tripartite Motif Containing
TW	Triton wash
U	Enzyme Units
Ub	Ubiquitin
ubcH5	Ubiquitin-conjugating enzyme E2 H5
µM	Micromolar
Upc1	Uncoupling protein 1
UPS	Ubiquitin proteasome system
v/v	volume by volume
Vc	Geometric column volume
Ve	Elution volume
VHL	Von Happel Landau
Vmax	Maximum velocity
Vo	Void volume
w/v	Weight by volume
Wnt	Wingless-related integration site pathway
Wnt7a	Wnt family member 7A
Z'	Z prime
°C	Degrees Celsius
-dF/dT	Rate of change in fluorescence

List of Schemes

List of Equations

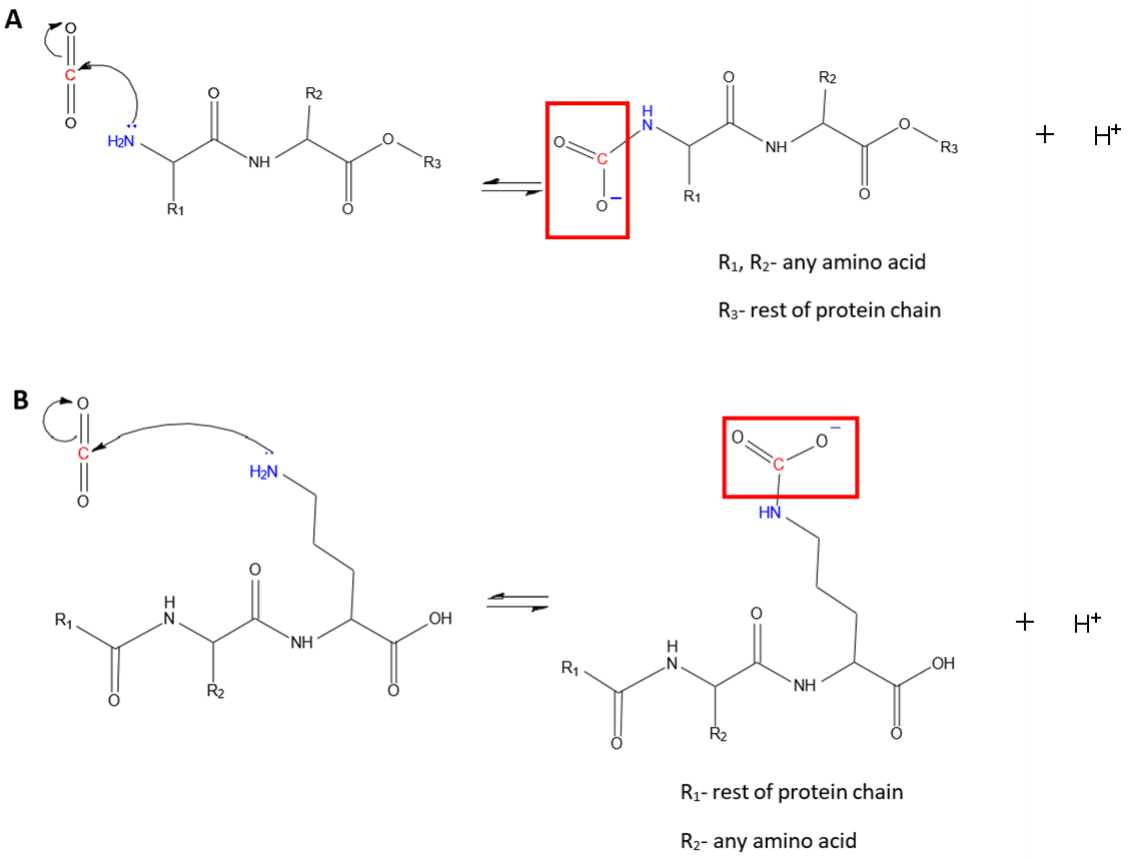
List of Tables

List of Figures

1. Introduction

1.1 Overview

This thesis investigates the post-translational modification (PTM) of CO_2 onto mammalian proteins. This spontaneous PTM, known as carbamylation, occurs on deprotonated neutral amines at physiological pH, as shown in Scheme 1-1.



Scheme 1-1 The mechanism of carbamylation involves the reversible nucleophilic addition of carbon dioxide onto a neutral amine, which includes the (A) N-terminal amino group and (B) lysine sites on a protein. The CO_2 modification corresponds to a mass shift of 44.01 Da, as shown by the group highlighted in red.

Before 2018, the carbamate PTM had been overlooked mainly due to the challenge posed by its reversibility. Linthwaite *et al.* addressed this problem by developing a trapping methodology for mass spectrometry identification of carbamates.[1] This thesis uses Linthwaite *et al.'s* work as a foundation to further elucidate the role of carbamylation in mammalian systems. Firstly, a proteome screen was conducted to identify carbamate hits in a HEK293 lysate. Subsequently, two mammalian carbamate sites were selected to illustrate the biological relevance of the carbamate PTM. This chapter provides an overview of the biological regulation mediated by CO_2 in mammalian systems, carbamate identification strategies and the importance of studying the carbamate modification.

1.2 CO_2 Physiology

The carbon cycle balances the processes of photosynthesis [2] and respiration [3] which exchange CO_2 as a substrate and product, respectively, to sustain life on Earth. Furthermore, the importance of carbon dioxide regulation is highlighted throughout evolution, whereby carbon dioxide sensing mechanisms are employed across species, from bacteria to mammals.[4] Organisms sense acute changes in CO_2, which activates a distinct and adaptive physiological response to combat environmental stress.[4] In addition, chronically elevated CO_2 levels are also known to alter gene expression, ultimately leading to disease progression.[5] This section covers the CO_2 dissolution, transport, chemosensing, and the CO_2-sensitive transcriptional response in mammalian systems.

1.2.1 CO_2 Dissolution and Carbonic Anhydrase.

CO_2 is a metabolic by-product [6] dissolves in water to form carbonic acid. Carbonic acid swiftly dissociates into bicarbonate ions and protons, as shown by Equation 1-1; therefore, CO_2 is critical to intracellular pH homeostasis.[7] Chemical equilibrium processes regulate the balance of the four inorganic carbon (Ci) species: carbon dioxide (CO_2), carbonic acid (H_2CO_3), bicarbonate (HCO_3^-) and carbonate ions (CO_3^{2-}).

$$CO_2 + \underset{1}{H_2O \rightleftarrows} H_2CO_3 \underset{2}{\rightleftarrows} HCO_3^- + H^+ \underset{3}{\rightleftarrows} CO_3^{2-} + H^+$$

Equation 1-1 The dissolution of carbon dioxide into carbonic acid (1), which spontaneously dissociates into a bicarbonate ion and proton (2). Furthermore, a bicarbonate ion can dissociate into a carbonate ion and proton (3).

The equilibrium in Equation 1-1 is pH-dependent, and the equilibrium constant is defined by Henry's law.[8] Figure 1-1 shows the speciation of the Ci species in equilibrium across varying pH. At physiological pH 7.4, HCO_3^- is the dominant Ci species, followed by CO_2. A major branch of CO_2 research focuses on detecting subtle partial pressure CO_2 /HCO_3^- fluctuations in various cellular environments using biochemical, bioimaging and molecular biology techniques.[9]

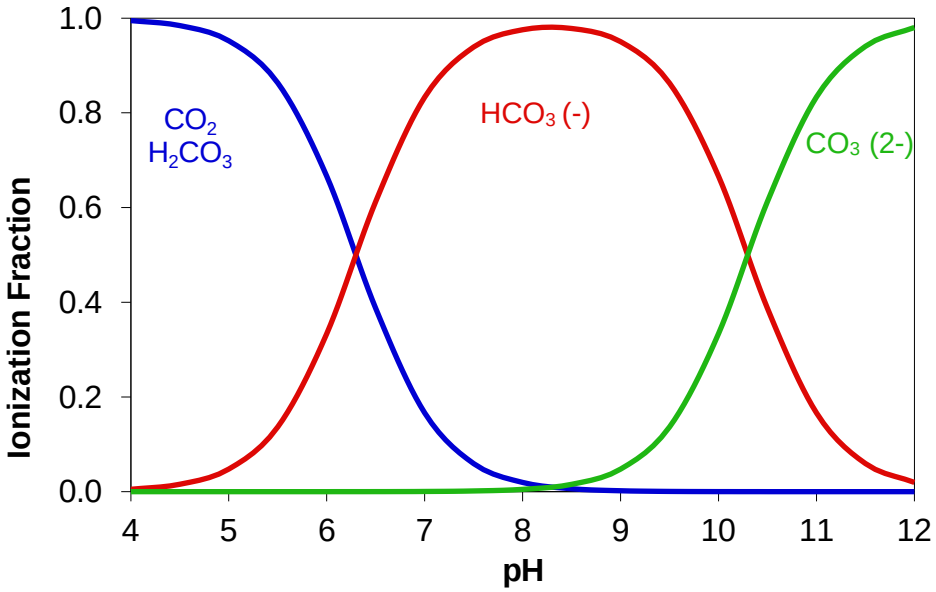


Figure 1-1 The relative speciation of inorganic carbon in water versus the pH. This is adapted from the cited reference. [10]

The rate-limiting step in Equation 1-1 is the hydration of CO_2 in step 1, which is catalysed by carbonic anhydrase (CA).[11] CAs play an important role in pH and CO_2 homeostasis. Equation 1-1 highlights high CO_2 partial pressures (PCO_2) are associated with increased acidosis in the cellular environment.

The active site of CA contains a zinc ion, three histidine residues and a catalytically important hydrogen bond. This hydrogen bond is located between the hydroxyl group of a threonine residue and zinc-bound water, enhancing water's potential for nucleophilic attack. CO_2 is bound into a hydrophobic pocket containing valine, and leucine resides.[12,13] CAs are found in both prokaryotes and eukaryotes, emphasizing their importance across biology.[14]

1.2.2 CO_2/ HCO_3^- Transport

Before the discovery of specialized CO_2 transport channels, the sole mechanism for CO_2 transmembrane transport was thought to be passive diffusion.[15] Membrane permeability for CO_2 diffusion is influenced by cholesterol content,[16] the number of CO_2 impermeable proteins, and steric interactions. [17] However, the permeability of cell membranes towards CO_2 remains an active area of research with differing opinions.[18] Specialized CO_2 transport channels found in vertebrates include aquaporins (AQPs) and rhesus (Rh) proteins are primarily localized to erythrocytes and have been expertly reviewed elsewhere.[19]

In brief, AQPs consist of a central hydrophobic pore surrounded by a tetramer composed of four hydrophilic monomeric pores. Alkaline surface pH shifts and ^{18}O isotope mass spectrometry determined AQP's role in gas exchange.[14] Using these techniques, erythrocytes lacking AQP1 reduced the transport of CO_2 by 60% compared to wild-type erythrocytes.[15] Geyer *et al*. assessed the permeability of mammalian AQP channels semi-quantitatively by monitoring the alkaline pH shift associated with CO_2 influx in oocyte models. This permeability is governed by cellular aerobic metabolism with a localized and time-dependent sensitivity range.[20]

Human erythrocytes are effectively the only differentiated cells that express Rhesus proteins. Rh proteins have a few forms, and their nomenclature relates to their glycosylation state and the surface antigen expressed. Trimeric Rh proteins comprise a hydrophobic central pore [21] surrounded by a combination of individual Rh subunits comprising a hydrophilic pore.[18] The human erythrocyte membrane HCO_3^- exchanger, termed Band 3, coimmunoprecipitates with the Rh complex.[22] Furthermore, treatment with a specific Band 3 inhibitor markedly decreased HCO_3^- and CO_2 transport.[19] These results indicate coupling between these channels to mediate rapid bidirectional transport of CO_2 in response to various conditions.[23]

At physiological pH, most Ci is found as bicarbonate ions. Therefore, specialised channels exist to transport HCO_3^-. The solute carriers, SLC4 and SLC26 gene families, encode the cotransporters

Na^+/HCO_3^- and anion exchangers Cl^-/HCO_3^-, which act throughout the body as acid load or extruders.[24,25] These HCO_3^- transporters are associated with CAs, suggesting this enzyme acts as a pH coupling protein.[26] In addition, SLC4 and SLC26 are regulated kinetically by the extracellular pH and a range of receptors. They are extensively expressed, highlighting the diverse functionality of these channels and modulation of many key biological functions in both the secretory [22,27,28]and non-secretory organs.[29]

1.2.3 CO_2 Regulation in Mammals

To retain physiological levels of CO_2, mammalian systems rely on chemosensing and adaptive response mechanisms. Detection of abnormal CO_2 levels results in the elicitation of an adaptive response signalling pathway. These response mechanisms are classified into acute and chronic, relating to physiological and sustained cellular responses via gene expression, respectively.[4] Three categories of chemosensors exist, including proxy CO_2 sensors that indirectly sense pH or HCO_3^- and direct CO_2 sensors.

The acute physiological response to elevated PCO_2 primarily relies on specialised central and peripheral chemosensors that transduce neuronal signals to alter the ventilation rate.[30,31] The CO_2 chemosensing regions of the brain are covered in detail in the cited review.[32] Chemoreception of elevated CO_2 levels results in an increase in intracellular calcium due to potassium ion channel closure. Activation of Ca^{2+} signalling increases the ventilatory drive by muscle contraction, cell motility and synaptic signalling.[33]

The ventilation rate is the major acute adaptative response pathway triggered by hypercapnia. However, several other mammalian chemosensors elicit a physiological response to altered CO_2 levels with various biological roles. For example, taste reception by carbonic anhydrase 4 (Car4), olfactory sensing by guanylyl cyclase D (GC-D) in rodents, cyclic adenosine monophosphate (cAMP) signalling by adenyl cyclases and secretion via HCO_3^- sensing by the receptor protein tyrosine phosphatase γ (RPTPγ). Figure 1-2 illustrates several mammalian chemosensors that sense pH, HCO_3^- and CO_2.

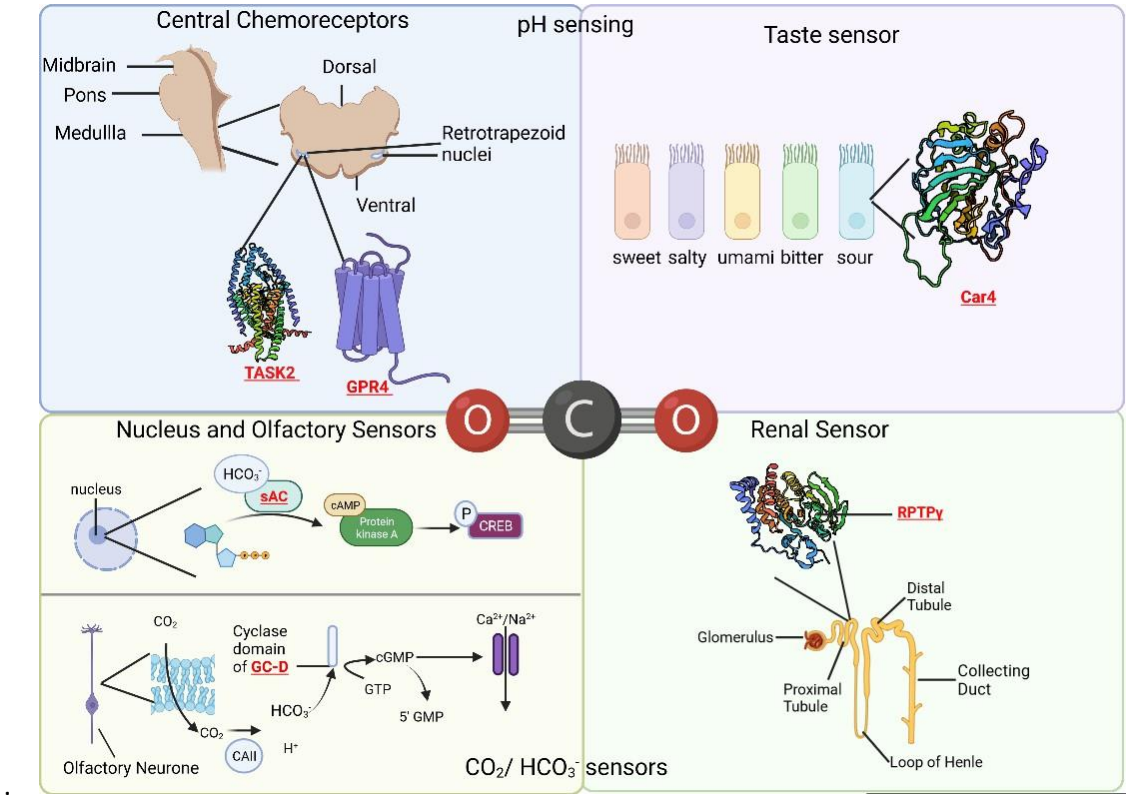


Figure 1-2 CO_2 chemosensing throughout mammalian physiology. Created using BioRender and cited references [34–42]

The exact mechanism for the chronic adaptive response to abnormal CO_2 levels remains an active area of research. The mechanism that mediates the CO_2-dependent transcriptional response has not been identified to date.[32] However, several mammalian signalling pathways have been implicated in the chronic transcriptional response to elevated CO_2, including immunosuppression, muscle atrophy and obesity.[43] The transcription factors involved in these phenotypes include nuclear factor kappa B (NF-kB), Forkhead box O3 (FoxO3a), and cAMP response binding element (CREB).

The NF-κB pathway is the master regulator for the immune and inflammatory response. NF-κB activation has been characterised into two branches: the canonical and non-canonical signalling pathways, which have distinct regulatory functions.[44] Both NF-κB pathways mediate temporal gene regulation by forming multiple homodimers and heterodimers comprised of NF-κB transcription factors, which translocate to the nucleus and bind specific DNA elements.[45] As summarised in the cited

reviews, NF-κB proteins from both signalling axes are differentially expressed during the cellular response to hypercapnia, which alters downstream gene transcription.[43,46] In the context of the canonical NF-kB pathway, hypercapnia has been shown to reduce the expression of the inflammatory genes markedly, intercellular adhesion molecule-1 (*ICAM1*) and interleukin-8 (*IL-8*).[47] Moreover, upstream from the NF-κB pathway is the toll-like receptor 4 (TLR4) protein [48] which binds lipopolysaccharide (LPS). LPS is a common stimulus for canonical NF-kB activation, and researchers have found that TLR4 expression is reduced under hypercapnic conditions.[49] The specific mechanism of how hypercapnia alters the complex NF-κB pathways remains an active area of research.

Hypercapnia results in muscle atrophy due to an increased muscle cell proteasomal degradation rate. Under hypercapnic conditions, phosphorylated AMP-activated protein kinase (AMPK) levels increase rapidly. AMPK is a metabolic sensor that regulates anabolic and catabolic pathways.[50] Activated AMPK mediates the phosphorylation of the transcription factor, FoxO3a which translocates to the nucleus and increases the expression of the ubiquitin ligase Muscle-specific ring finger protein 1 (MURF1), which drives proteasomal degradation.[51]

Phosphorylation of the transcription factor CREB via protein kinases leads to induced expression of various target genes in a context-specific manner. A study using pre-adipocytes showed that hypercapnia elevates the DNA binding activity of CREB in a cAMP-dependent manner, which stimulates the expression of proadipogenic transcription factors, leading to increased adipogenesis.[47]

Chronic exposure to elevated CO_2 has also been linked to mitochondrial dysfunction, suppressing O_2 consumption, ATP production and inhibiting cell proliferation due to downregulation of the tricarboxylic acid (TCA) cycle enzyme isocitrate dehydrogenase 2 (IDH2).[52]

Finally, a previous gene expression analysis study in human airway epithelial cells identified downregulation of genes involved with nucleosome assembly under hypercapnic conditions. In particular, the expression of Histones H2A, H2B, H1 and the nucleosome assembly protein-like 1

(NAPL1) were downregulated. This finding links hypercapnia to nucleosome-mediated gene transcription, genome stability and the antibacterial response.[53]

1.2.4 Abnormal CO_2 Levels and Disease.

Clinically, physiological levels of CO_2 have been defined as 35-45 mmHg.[54] Fluctuations outside this range occur in both health and disease, and a change as low as 10 mmHg can elicit an acute CO_2-responsive mechanism.[55] Levels above 45 mmHg PCO_2 are considered hypercapnic, and levels below 35 mmHg PCO_2 are considered hypocapnic. Hypercapnia is associated with increased acidosis and contrastingly, hypocapnia is associated with increased alkalosis.

Hypercapnia is the retention of carbon dioxide in the body, which can cause pathological changes and is involved in a range of diseases, for example, skeletal muscular atrophy, [50] cystic fibrosis, [56] obesity,[57] acute respiratory distress syndrome (ARDs), [58] and chronic obstructive pulmonary disease (COPD).[59] Remarkably, despite the clear contribution of elevated CO_2 to poor prognosis in sleep apnoea, obesity, and COPD, there is almost no knowledge of its biological targets. Surprisingly, high levels of carbon dioxide have been recognised as an agent with protective effects and have previously been used in a therapy called permissive hypercapnia.[60] However, the exact impacts are unknown, cause concern and are heavily debated within the field.[61]

Hypocapnia is a risk factor for neonatal mortality and neurodevelopment deficits.[62] Hypocapnia is associated with alkalosis, which can cause problems for dialysis patients.[63] These disease states indicate that carbon dioxide levels must be tightly regulated to maintain healthy cells. Therefore, identifying new mammalian CO_2 targets could help to explain the cell disease states.

1.3 Interaction of CO_2 with Proteins.

To address rising levels of CO_2 due to the climate crisis, researchers have considered expediting the efficiency of natural CO_2 sequesters such as, RuBisCo and CAs, which catalyse carboxylation and dissolution, respectively. Due to this research interest, computational approaches using bioinformatics and molecular modelling have been used to consider a broader perspective of how CO_2 interacts with proteins. A possible mechanism of CO_2 interaction with proteins is via ligand-based hydrogen bonding, and *in silico* approaches have been used to assess the properties of CO_2 binding sites in research led by Drummond *et al.* [64,65] and Cundari *et al.* [66] Both researchers found that the accessible hydrogen bond donors on proteins had the highest affinity for CO_2.

An alternative interaction is the direct covalent modification of CO_2 on proteins under physiological conditions, as shown in Scheme 1-1.[67] This reaction involves a neutral amine's spontaneous, reversible covalent modification via a nucleophilic addition reaction with CO_2 termed carbamylation. The prevalence of this anionic PTM on a proteomic scale has previously been neglected due to its liability.

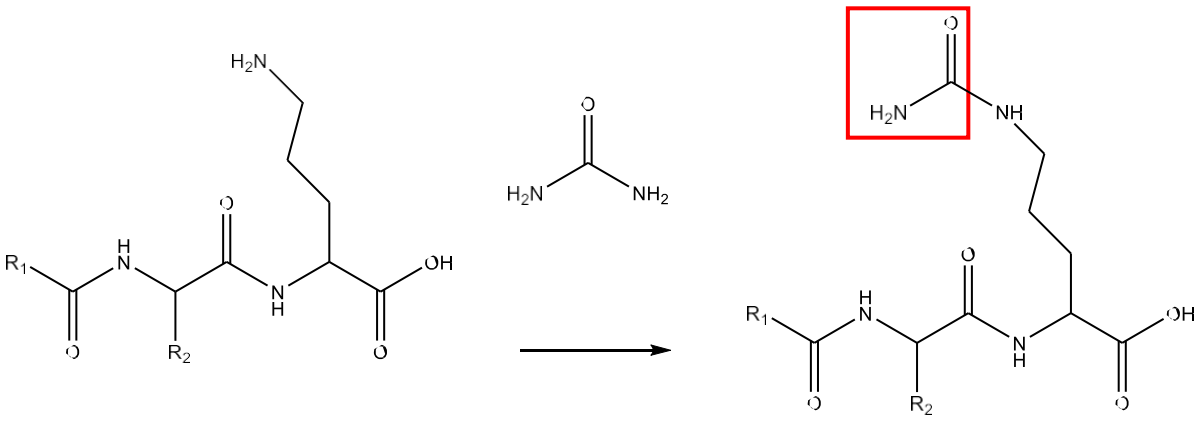
Three neutral amines are found across the 20 amino acids, which are the building blocks of proteins. These residues include arginine, histidine, and lysine, with sidechain pKa's of ~ 12, 6 and 10, respectively.[68] Both the basicity and percentage of deprotonated amine found at pH 7 are governed by the pKa. Therefore, the principal targets for carbamylation are lysine's ε-amino group and α - N terminal residues.

The probability of carbamate formation depends on the surrounding pH. The pH alters the speciation of Ci into CO_2 (Figure 1-1) [69] and the percentage of lysine ionization [70] with these two parameters having an inverse relationship. For example, at pH 7.4, the percentage of CO_2 is below 10% of the total Ci content, and the protonated form of lysine dominates; thus, carbamate formation is unfavourable. Moreover, this process is non-enzymatic and is thought to occur in structurally and pH-privileged environments.

Predictions from computational studies indicate that carbamylation could be used by at least 1.3% of large proteins.[71] However, until recently, [1,72] there were no methods for profiling carbamylation under physiological conditions across the proteome. The specifics of the trapping [1] and chemical biology [72] proteomic approaches are discussed in Chapter 3, section 3.2.

1.4 An Alternative Form of Carbamylation

As an aside, it is important to note that the terms, carbamoylation and carbamylation have both been used interchangeably to describe the modification of neutral amines with urea or isocyanate.[73] Despite having the same nomenclature, the properties of each PTM type are distinct. The carbamylation PTM arising from urea is irreversible, has a mass of 43.006 Da (Scheme 1-2) and using urea base digestion methods is undesirable due to this PTM affecting MS protein identification and quantification.[74] The carbamylation PTM discussed in this thesis relates to the reversible modification arising from CO_2 (Scheme 1-1).



Scheme 1-2 Carbamylation PTM by urea or isocyanate on lysine. The urea-based modification corresponds to a mass shift of 43.006 Da, as shown by the group highlighted in red.

1.5 Relevance and Detection of Carbamylation

The first carbamate sites identified were the N-terminal valine of haemoglobin [75] and an active site lysine of RuBisCo.[76] The biochemical activities mediated by these proteins were directly linked to environmental CO_2. Therefore, in 1983, Lorimer hypothesized that carbamate modification plays a key role in widespread biological regulation.[67] The properties of CO_2 that support this hypothesis include its widespread presence, reversibility of binding, and potential to form intramolecular interactions that alter the protein's structure in the bound state.[77]

Since the identification of haemoglobin and RuBisCo as carbamylation targets, nuclear magnetic resonance (NMR), [78] computational studies, [71] X-ray crystallography [79–81] and mutational analyses [82] were used to identify carbamates. However, these methods are limited and not suited to widespread proteome identification. Therefore, the proteomic-based strategies [1,72] were developed to offer a physiologically relevant and site localisation detection method for carbamates. For an in-depth discussion on the discovery and relevance of literature-reported carbamates, the reader is directed to Blake *et al.'s* review.[83] Furthermore, the three literature reported mammalian carbamates include haemoglobin, connexin 26, and ubiquitin are detailed in section 3.2.3.

Carbamate target sites previously reported are found in pH-privileged environments or as stabilized bonding interactions within the protein's structure across various species. The lysine modification sites on haemoglobin, [75] ubiquitin, [84] alanine racemase [79] and β-lactamase [81] have lowered pKa values on the modified amine to favour carbamylation.[85] Whilst the RuBisCo, [76] urease, [80] alanine racemase, [79] β-lactamase, [81] and connexin 26 [86] carbamates have metalloenzyme, amino acid residue, hydrogen bonding or salt bridge stabilizing interactions, respectively. Interestingly, carbamylation has been identified across the kingdom of life with diverse biological roles in the ventilatory response, facilitation of enzyme catalysis, cellular signalling, and proteolysis.

1.6 Experimental Considerations for Studying Carbamylation.

The biological significance of MS and NMR carbamate hits can be verified by antibody techniques,[84] methods which have identified CO_2 sensors [87] and target-specific assays. The techniques discussed in this section (1.6) offer a generic foundation for identifying the downstream signalling and physiological effects of carbamylation targets. PCO_2 levels can be controlled during cell culture or by adding sodium bicarbonate ($NaHCO_3^-$) in a biochemical assay. pH is important in any carbamylation study to delineate CO_2 effects from acidosis. To ensure physiologically relevant data, extracellular pH must be buffered to account for PCO_2 levels in cell culture. In biochemical assays, the buffer must be strong enough to resist pH changes introduced by elevated Ci levels, and the total anionic concentration should remain constant across a dataset. These methods for pH control under elevated CO_2 can be referred to as buffered hypercapnia. Similarly, an intracellular pH control has been identified to account for intracellular acidosis, as described by Cook *et al*.[88] In mutagenesis studies, naturally occurring glutamate mimics carbamylation and can be used as a positive control.[1]

1.7 Motivation for Investigation

Carbamylation has previously been overlooked due to being a liable PTM with limited experimental approaches for site identification. Historically reported target carbamate sites are generally stable enough to be identified by X-ray crystallography, and previous *in-silico* approaches predict new sites based on stable buried carbamylation sites. However, the advent of the physiologically relevant proteomic trapping methodology [1] offers an exciting avenue for novel research. This investigation focused on carbamylation in mammalian systems due to the implications of CO_2 on health and signalling pathways, as outlined in sections 1.2 and 1.3.

1.8 Aims and Hypotheses

The two primary hypotheses underlining this investigation are that carbamylation is widespread in the mammalian proteome and that carbamylation is of biological importance to biochemical pathways. This investigation is split into three overarching aims, each detailed in chapters 3-5. The first aim was to profile the CO_2-sensitive HEK293 cell line for carbamylation sites across the mammalian proteome, as discussed in Chapter 3. The second aim was to expand carbamylation research into a pharmaceutical setting using Ub K48 carbamylation in the context of PROTACs, as discussed in Chapter 4. The third aim was to assess the biological significance of histone carbamylation on DNA transcription, as discussed in Chapter 5. Finally, the results are summarised in Chapter 6 and future directions for this work are clearly outlined.

2. Materials and Methods

The methods described in sections 2.3-2.5 correspond to Chapters 3-5. In certain cases, additional method details are provided within the relevant result chapters.

2.1 Materials and Equipment

All materials were purchased from Merck unless otherwise stated. Continuous cell culture reagents and biomolecule quantification kits were purchased from ThermoFisher Scientific. Samples were dried in a vacuum concentrator (Eppendorf), and peptides were separated on an ekspert nano LC 425 (Eksigent), coupled to a duospray electrospray ioniser with a TripleTOF 6600 Q-TOF (Sciex). pH measurements were recorded by the TIM856 pH Stat Titration Manager (Radiometer Analytical) calibrated by pH standards (Fisher). All cultured cell lines were maintained in a CO_2 incubator (Sayno) and were counted by a hemacytometer or Vi-CELL XR Cell Counter (Beckman Coulter). Spectrophotometry measurements were determined using a PHERAstar FSX (BMG) or Synergy H4 (Biotek). Protein purifications were run on the AKTA PURE FPLC and AKTA start (Cytiva). RNA sequencing was performed on an Illumina sequencer.

2.1.1 Cell lines

The human embryonic kidney (HEK293) cell line was used for proteomic screening. HEK293 endo HiBiT BRD4 and non-small cell lung cancer, NCI-H838 HiBiT SMARCA2 strains were used for PROTAC dosing obtained from the global cell bank at AstraZeneca Alderley Park. Rosetta Turner (DE3), MAX Efficiency™ DH5α (ThermoFisher Scientific) and BL21 - Codon Plus (DE3) – RIL (Agilent) competent cells were used for protein expression.

2.2 Biomolecule Quantification

2.2.1 Bradford Assay

A dilution series of bovine serum albumin (BSA) standards in a buffer relevant to the experiment was made with concentrations ranging between 25-1500 µg/ml. Standards and samples (5 µl) were loaded onto a 96-well plate with Bradford reagent (180 µl). Plates were mixed (30 s), incubated at 37 °C for 15 minutes, cooled to room temperature, and the absorbance was read at 595 nm. Sample concentration was determined by using the linear response BSA standard curve.

2.2.2 Bicinchoninic acid (BCA) Assay

This assay was used when buffer components were incompatible with the Bradford reagent. A dilution series of BSA standards in a buffer relevant to the experiment was made with concentrations ranging between 25-2000 µg/ml. Standards and samples (25 µl) were loaded onto a 96-well plate, and BCA reagent (200 µl) was added. Plates were mixed (30 s), incubated at 37 °C for 30 minutes, cooled to room temperature, and the absorbance read at 562 nm. Sample concentration was determined by using the BSA standard curve.

2.2.3 Peptide Assay

The Pierce Quantitative Colorimetric Peptide Assay was used to quantify peptides. A dilution series of the peptide standard was made with concentrations ranging between 0-1000 µg/ml. Standards and samples (20 µl) were loaded onto a 96-well plate, and the working reagent (180 µl) was added. Plates were mixed (30 s), incubated at 37 °C for 15 minutes, cooled to room temperature, and the absorbance read at 480 nm. Sample concentration was determined by using the peptide standard curve.

2.3 Proteomics

2.3.1 HEK 293 Cell Culture and Harvesting

HEK293 cells were grown (37 °C) and incubated with carbon dioxide in Dulbecco's modified medium (DMEM) supplemented with fetal bovine serum (FBS, 10%), streptomycin/penicillin (1%) in Nunc™EasYFlask™ Cell Culture Flasks, T-75 to a confluency of 80%. To maintain HEK293 cells, they were washed with phosphate-buffered saline, followed by incubation with Trypsin-EDTA (0.25%) to lift cells which were reseeded into fresh DMEM media. Alternatively, cells were harvested in phosphate-buffered saline (PBS, 4 ml) and pelleted by centrifugation (4,000 rpm, 5 minutes).

2.3.2 Carbon Dioxide Incubation

Cell lines were exposed to varying carbon dioxide levels throughout this project, including 5%, 10% and 20%. DMEM media without phenol red was buffered with HEPES pH7.5 (12.5 mM), and sodium bicarbonate at 20, 40, and 60 mM was added to the media for 5%, 10% and 20% CO_2 incubations, respectively.

2.3.3 Sample Preparation

Datasets were processed using cells with passage numbers 19 and 18 for cells grown at 5% and 10% CO_2, respectively. Cell pellets were resuspended in phosphate buffer (pH,7.4, 100 mM), sonicated at 10%, four times for 30 seconds each time followed by 30 seconds resting on ice, homogenised for each dataset, and centrifuged (16,000 rpm, 10 minutes). The resulting lysate was stirred at room temperature with varying sodium bicarbonate concentrations (0, 20, 50 mM). Protein amounts were estimated before trapping with a Bradford assay (described in 2.2.1, 3- 4 mg). Trapped samples were prepared by the stepwise addition of triethyloxonium tetrafluoroborate (TEO, Et_3OBF_4) (240 mg, 1.47 mmol) in phosphate buffer whilst pH was maintained at 7.4 via the responsive addition of sodium hydroxide (NaOH, 1 M) by the automatic burette. Untrapped samples were prepared the same way, without adding TEO. The reaction was left for 1 h to ensure all the TEO was hydrolysed. Samples were dialysed against Milli-Q water (16 h), dried, and stored (-20 °C).

2.3.4 Protein Digestion using the S-trap Protocol

All reagents used in this process were LCMS reagent grade. Samples were resuspended in a minimum volume of S-trap lysis buffer (5% SDS, 50 mM triethylammonium bicarbonate, TEAB, pH 7.55). Protein amounts were estimated using a BCA (described in 2.2.2). Disulfides were reduced using dithiothreitol (Melford Biostores, DTT, 20 mM, 5 minutes, 95 °C) and alkylated with iodoacetamide in the dark (40 mM, 30 minutes). Samples were clarified by centrifugation (13,000 g, 5 minutes). The S-trap protocol was used for digestion without modifications. In brief, samples were applied to mini or midi S-trap columns (ProtiFi) depending on the protein amount present. Trypsin gold (Promega) was added to the protein with a ratio of 1:20 and incubated overnight (37 °C, 18 h). Peptides were eluted and dried. Peptides were resuspended in LC-MS water and quantified with the Pierce Quantitative Colorimetric Peptide Assay (described in 2.2.3).

2.3.5 Hydrophilic Interaction Liquid Chromatography (HILLIC)

Dried peptides were resuspended in 3% acetonitrile and 0.1% formic acid in 0.3M ammonium formate pH 3 and clarified to remove precipitant. Acetonitrile was added to a final concentration of 85%. The HILLIC column was equilibrated in releasing solution followed by binding solution by 10 column volumes of each. Releasing solution is 5% acetonitrile, 30 mM ammonium formate pH 3.0 and the binding solution is 85% acetonitrile, 30 mM ammonium formate pH 3.0. The sample was added to the column and washed using 5 column volumes of binding solution. The analyte was eluted with 2 column volumes of releasing solution and peptides were dried and stored prior to LCMSMS injection.

2.3.6 Peptide Fractionation

To increase protein group identification (ID), peptides (500 µg) were fractionated using an increasing acetonitrile gradient (0 - 40% acetonitrile over 40 minutes, 40 - 80% acetonitrile over 5 minutes) on a reverse phase HPLC column (C18). Samples were collected at 20 s intervals to reflect the bandwidth of a peptide and combined into 16 fractions. Any suspect peaks in the UV spectra were not combined into the final peptide fractions. Acetonitrile was removed, and peptides dried.

2.3.7 Liquid Chromatography-Tandem Mass Spectrometry (LCMSMS)

Dried peptides were resuspended in 2% acetonitrile and 0.1% formic acid pH 3 to prepare for injection. Before nanoflow chromatographic separation of peptides (1 μg or 5 μg for purified protein digests and complex peptide mixtures, respectively) were concentrated by trap-and-elute reverse phase chromatography using a Triart C18 capillary guard 1/32", S-3 μm, 5 × 0.5 mm (YMC) trap column. Peptides were separated on a Triart C18 capillary guard 1/32", S-5μm, 150 × 0.3 mm (YMC) resolving column. The separation gradient started with 0.1% formic acid in water (A) with an increasing organic phase of 0.1% formic acid in acetonitrile (B). Online chromatographic separation was performed (45 or 90 minutes for purified protein digests and complex peptide mixtures, respectively) on the nanoLC at a flow rate of 5 μl/min using a low micro gradient flow module with a linear gradient of 3 – 30% B (60 minutes), then to 80% B over (19 minutes), held (3 minutes) before returning to 3% B and re-equilibrated in this buffer (27 minutes).

Data-dependent top-30 MS-MS acquisition started immediately upon gradient initiation and lasted 80 minutes. MS acquisition was performed in the positive mode, starting at 0.5 minutes for 80.5 minutes with a cycle time of 1.996 s. Throughout this period, precursor-ion scans (400 to 1600 m/z) with an accumulation time of 250 ms enabled the selection of up to 30 multiply charged ions for collision-induced dissociated (CID) fragmentation. The switch criteria to trigger MS2 were ions 400-1600 m/z, charge state 2 to 5, > 500 cps, and 15 second rolling exclusion to limit multiple fragmentation of the same peptide. MS/MS spectrum acquisition (m/z 100-1500) for 30 ms with rolling collision energy in high sensitivity mode.

2.3.8 Data Processing

The acquired mass spectra information was formatted as raw and wiff files, which were converted to mgf files using MS convert, Proteowizard. The mgf file was run on PEAKs X Studio 10.5 to obtain peptide sequences matching the ion fragmentation patterns. Peptides matched to the mammalian database (UP000005640), alongside the contaminant database and de-novo sequencing

peptides were listed. The digest mode was semi-specific, with max missed cleavages set to 3. PTMs searched for included carbamidomethylation (57.02 Da), ethylation (28.03 Da), oxidation (methionine; 15.99 Da), acetylation (N-terminus; 42.01 Da), carboxyethyl (72.02 Da). A 1% false discovery rate (FDR) threshold was set for peptides and for proteins the threshold was -10logP ≥ 20 and ≥2 unique peptides.

Raw wiff files were run in Protein Pilot, which converts the file to mgf, matches peptides to the protein database and filters based on a preset FDR. The inbuilt parameters of Protein Pilot were used with the addition of a carboxyethyl modification set, which included the PTMs detailed in the above paragraph. Carboxyethyl modification sites were identified by analysis of mass chromatogram in both PEAKs and Protein Pilot. The hits that were present in both were assigned high confidence.

2.4 Pharmaceutical Screening

2.4.1 Cell Culture

Crispr Cas9 knock-in was used to express a HiBiT tagged protein of interest from an endogenous promoter. Both HiBiT-tagged cell lines were maintained in DMEM high glucose and Roswell Park Memorial Institute (RPMI) 1640 media for HEK 293 and NCI-H838 cell lines, respectively, supplemented with FBS (10%) and glutaMAX (1%). Cells were passaged as described previously when confluency reached 80%.

2.4.2 Assay Plates

The Compound Management Group (CMG) at AstraZeneca (AZ) prepared assay-ready plates (ARP) from the AZ compound collection using the Echo555 liquid handler (Labcyte). ARPs were made in concentration-response using PROTAC tool compounds with a concentration range of 60 µM - 0.33 nM for BRD4 and 60 µM - 6.67 nM for SMARCA2.

2.4.3 Nano-Glo HiBiT Lytic Detection Assay

Cells were detached, pelleted, and resuspended in media relevant for the 5% or 10% CO_2 incubation. Cells were seeded into 1536 well ARPs using a Multidrop dispenser (ThermoFisher) at 2,500 or 8,000 cells per well for NCI-H838 and HEK293, respectively. Cells dosed with PROTACs were incubated overnight (18 h) in a humidified incubator at 37 °C and 5% or 10% CO_2.

The Nano-Glo Lytic detection system (Promega) includes the lytic buffer, lytic substrate and LgBiT protein. The lysis buffer was prepared as stated by the manufacturer with 1X lytic buffer, 1:50 lytic substrate, and 1:100 LgBit diluted in PBS.

Before lysis, media was removed from the plates using a Blue Washer (BlueCatBio, 1400 rpm, 5s). Lysis buffer (5 µl) was added using an angle headed Certus Flex (Gyger). Plates were covered and incubated (20 minutes), and luminescence was recorded using a LUM plus module with 0.02 s measurement interval time.

2.4.4 Statistical Analysis

The output files from the FSX were in csv format. The raw values were directly uploaded onto Genedata Screener (v17). Data was normalised against neutral controls, and concentration response (CR) curves were plotted to assess protein degradation. AZ curve categorisation was assigned to each CR curve using the Hill coefficient, curve bottom, top and span, and confidence interval. The statistics that verify the reproducibility and feasibility of an assay in the AZ high throughput screening department (HTS) include the Z factor, Z', RZ', signal/background and the coefficient of variation %. Concordance between replicates was determined using Spotfire.

2.4.5 Western Blot Protocol

Cell lysates were sonicated (10s x2) and centrifuged (15,000 rpm, 2 minutes, 30 s). Each lysate (12 µg) was denatured, reduced, and separated on a 20% SDS PAGE gel. The gel was transferred to a nitrocellulose membrane (2 h, 30 minutes). Membranes were blocked with milk (1 h) cut in half, one side incubated with anti-vinculin and the other with anti-BRD4 (Abcam, 1/1,500 and 1/1,000 respectively, 16 h, 4 °C). Following washing, the membrane was incubated with horseradish peroxidase conjugated (Hrp) goat anti-mouse and anti-rabbit immunoglobulin G (IgG) conjugate, respectively (Abcam, 1/1,000, 1 h, 4 °C). Detection was facilitated by ECL™ Prime Western blot detection reagent (Merck). The blot was visualised using automatic exposure mode on the Chemidoc (Biorad). Blots were quantified using ImageJ's integrated density of bands subtracted from the background.[89]

2.5 Nucleosome Preparation and Assays

2.5.1 Isolation of Native Nucleosomes.

HEK293 cells were harvested and pelleted by centrifugation (1000 rpm, 5 minutes, 4 °C). The EpiQuik™ Total Histone Extraction Kit (Epigentek) was used without modifications. In brief, a pre-lysis buffer was used to release nuclei. Following this, histones were extracted using an acidic buffer, and finally, a pH balancing buffer was added to the extracts where the final pH was 6-7.

2.5.2 Nucleosome Preparation for Mass Spectrometry

Trapping was performed on native nucleosomes or recombinant histone octamers (250 µg) as described previously in section 2.4.3. Processing of samples proceeded straight to protease digest, or peptides were modified by propionylation, as discussed in section 2.5.3. Proteases used at a ratio of 1:20 included trypsin, Arg-C and Lys-C.

2.5.3 Propionylation of Nucleosomes

The dried histone pellet was resuspended in TEAB (1M) and propionic anhydride reagent in isopropyl alcohol (IPA) at a ratio of 1:79. Histone proteins were incubated (room temperature, 30 minutes). An equal volume of water was added to reduce the propionic acid concentration before incubation (37°C, 30 minutes). Samples were dried, and hydroxylamine (0.5 M) was added with ammonium hydroxide to adjust the pH to 12. This reaction proceeded for 20 minutes at room temperature; next, the pH was adjusted to pH 3 using formic acid. Protein samples were dried, trypsin digested, and another round of propionylation was performed on the peptides.

2.5.4 Recombinant Nucleosome Purification

pET28a_Synthetic_Human_H2A.1, pET28a _Human_H2B.1, pET28a _Human_H3.1, pET28a_Human_H4 were gifts from Joe Landry (Addgene plasmid # 42634, 42630, 42631, 42633). pET24_sfCherry- Histone H2B type1 – (C/E/F/G/I.) and pET24_sfCherry_Kana backbone- Histone H4 were gifts from Sarah Caswell and Miguel Rodrigo from AZ protein science. pENTR223_12x widom 601

DNA was a gift from Scot Wolfe (Addgene plasmid # 114358). The competent cells used were a combination of Tuner and Rosetta DE3 cells for H2A.1 and H3.1. For H2B.1 and H4, BL21-CodonPlus-RIL (Agilent) were used. For pENTR2_12 widom 601 DNA, max efficiency DH5α (Thermo Fisher Scientific) was used.

2.5.4.1 Transformation

Competent cells were thawed on ice (10 minutes, 20 µL), and plasmid DNA (1 µl, 50 ng/µL) was added. The components were mixed, left on ice (30 minutes), and exposed to heat shock (42 °C for 10 s). Tubes were placed back on ice (5 minutes), and S.O.C. medium was added to the mixture. The culture was incubated with shaking (37 °C, 60 minutes, 250 rpm). Serial dilutions of the S.O.C culture were obtained and streaked out onto selection plates with the appropriate antibiotic (cell line and plasmid specific), and plates were incubated (18 h, 37 °C). Single colonies were picked and incubated in Lysogeny Broth (LB, 18 h, 37 °C) to make glycerol stocks.

2.5.4.2 Small-Scale Protein Test Expression

Transformed Rosetta Tuner (DE3) *E. coli* cells containing H2A.1/H3.1 were grown in LB supplemented with 50 µg/ml kanamycin and 25 µg/ml chloramphenicol at 37 °C with shaking. (18 h, 140 rpm) Transformed BL21-CodonPlus (DE3)-RIL *E. coli* cells with H2B.1/H4 and were grown in LB medium supplemented with 50 µg/ml kanamycin, 25 µg/ml chloramphenicol and 10 µg/mL tetracycline at 37 °C with shaking (18 h, 140 rpm). Overnight bacterial cultures (1 mL) were reseeded into fresh LB (25 mL). Cultures were grown to an OD_{600} of 0.4 at 37 °C and induced with Isopropyl β-D-1-thiogalactopyranoside (IPTG, 200 µM). A sample of pre-induction culture was collected for the Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE) gel. Induced cultures were placed at 25 or 18 °C, and samples were collected at various time points post-induction. Equal amounts of culture were harvested using the OD_{600} reading, centrifuged (4000 rpm, 5 minutes), supernatant was discarded, and the pellet was resuspended in lysis buffer (40 mM sodium acetate pH 5.2, 1 mM EDTA pH 8). Test expression samples were boiled (96 °C, 20 minutes) and centrifuged (14000 rpm, 20

minutes). The pellet was resuspended and boiled again without centrifugation. The supernatant and pellet were run with the pre-induction sample on a SDS-PAGE gel.

2.5.4.3 SDS PAGE

SDS-PAGE was used to separate proteins by molecular weight. A 4X loading buffer (200 mM Tris-HCl pH 6.8, 400 mM DTT, 8 % (w/v) SDS, 6 mM bromophenol blue, 40 % (v/v) glycerol) was added to protein samples (10-20 µg) to a final concentration of 1x. Proteins were denatured by boiling (10 minutes, 95 °C) and loaded onto resolving gels in the range 12 - 20% (w/v) pH 8.8 with 5% pH 6.8 stacking gel. To estimate the molecular weight, samples were run alongside a protein ladder (Pageruler^{TM} pre-stained). The gels were run in running buffer (25 mM Tris-HCl pH 7.5, 192 mM glycine, 0.1 % (w/v) SDS) at 180 V for 1 h. Gels were stained (15 minutes) with InstantBlue (Abcam), followed by destaining in MilliQ.

2.5.4.4 Large-Scale Protein Expression

Inoculated cultures (25 mL) were grown (18 h) in LB with the appropriate antibiotics (cell line and plasmid specific). Overnight cultures (25 mL) were reseeded into LB (1 L, 12 flasks). When cultures reached OD_{600} 0.4, IPTG was added (200 µM) to induce protein expression and grown (25 °C, 16 h). Cultures were centrifuged (4,000 rpm, 25 minutes, 4 °C), the supernatant was removed, and pellets were stored (-80 °C) until required.

2.5.4.5 Inclusion Body Isolation and Size Exclusion Purification

Bacterial cell pellets (3 L) were thawed and resuspended in wash buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM Na-EDTA, 1 mM benzamidine). The bacterial cell pellet was sonicated six times with a 15 s pulse at 40% amplitude. Inclusion bodies were pelleted by centrifugation (21,000 rpm, 20 minutes, 4 °C) and resuspended with wash buffer containing 1% (v/v) Triton X100 (TW buffer). This solution was centrifuged (12,000 rpm, 10 minutes, 4 °C), and pellets were washed twice in TW buffer and two times with wash buffer. The resulting pellet was soaked in 1 mL DMSO (30 minutes, room temperature) and unfolding buffer (7 M guanidinium hydrochloride, 20 mM Tris-HCl pH 7.5 and

10 mM DTT, 40 mL) was added slowly and stirred (1 h, room temperature). The resulting mixture was centrifuged (21,000 rpm, 10 minutes, 20 °C).

2.5.4.6 Histone Purification by Size Exclusion

The supernatant was concentrated and loaded onto a S200 pre-equilibrated with size exclusion buffer (7 M urea, 20 mM sodium acetate pH 5.2, 1M NaCl and 5 mM β-mercaptoethanol (BME)). The injected loop was washed with size exclusion buffer at 5 times the loop volume. Proteins were separated over one column volume. Flowthrough, wash, and elution fractions every 0.5 ml were collected for analysis by SDS PAGE. Histone-containing fractions were pooled and dialysed against water (5 µM BME) in three steps (1 h, 14 h, 3 h). All dialyses discussed in section 2.5 were conducted in a bag with a molecular weight cut-off of 6-8 kDa. Following this, proteins were dialysed into SAU200, loaded onto an SP column, and processed as described in 2.5.4.8, without Q column purification.

2.5.4.7 Histone Denaturation and Extraction from Bacterial Cell Pellets

Cell pellets (3 L) were thawed and resuspended in 150 mL lysis buffer (40 mM sodium acetate pH 5.2, 1 mM EDTA pH 8, 1 mM lysine, 200 mM sodium chloride, 6 M urea and 5 mM β-mercaptoethanol) with protease inhibitor cocktail tablets. This lysis buffer will hereafter be referred to as SAU200, where the number following SAU refers to the concentration of sodium chloride in mM. Following resuspension, the bacterial cell pellet was sonicated at 40% for 20 minutes in intervals of 15 s. This solution was centrifuged (21,000 rpm, 50 minutes, 4 °C), and the supernatant filtered (0.45 µM) ready for ionic exchange.

2.5.4.8 Histone Purification by Ionic Exchange

A Hi Trap sulphopropyl (SP) sepharose column was equilibrated into SAU200 for five column volumes. The filtered supernatant was loaded, and the column was washed for five column volumes. Histone proteins were eluted using a gradient of SAU200 to SAU600 over twenty-column volumes. Flowthrough, wash, and elution fractions were collected for analysis by SDS PAGE. The SP column was rinsed in SAU1000 for five column volumes to remove contaminants. Before storage (4 °C), the column

was washed into MilliQ, followed by 20% ethanol. Histone protein-containing fractions determined by SDS PAGE were pooled and dialysed against water (5 mM BME, 4 °C, 16 h). The dialysate was centrifuged (4,000 rpm, 10 minutes) to remove precipitants and supplemented with 15 mM-Tris-HCl pH 8. The filtered dialysate was loaded onto a Hi Trap quaternary (Q) amino group sepharose column equilibrated in a loading buffer (15 mM Tris, 0 M NaCl). The column was washed in ten-column volumes, and a gradient to 2M NaCl was run over five-column volumes. Flowthrough, wash, and elution fractions were collected for analysis by SDS PAGE. Before storage (4 °C), the column was washed into MilliQ, followed by 20% ethanol. The flowthrough contained the histone protein, which was concentrated (5 kDa filter) and a final concentration of 10 % (v/v) glycerol was added for storage in useable aliquots (-80 °C). Proteins were quantified using the Bradford assay detailed in 2.2.1.

2.5.4.9 Octamer Refolding

Histone constructs were required at a molar ratio of 0.9:0.9:1:1.1 (H2A (5.5 mg): H2B (5.4 mg): H3 (6.65 mg): H4 (5.4 mg)) for octamer formation. Each construct was at a starting concentration of at least 4 mg/ml and diluted using unfolding buffer (7 M guanidine HCl, 20 mM Tris pH 7.5 and 10 mM DTT) to 2 mg/ml. Unfolding of each construct proceeded for 1 h (room temperature). The four individual proteins were mixed, and the concentration of each histone was reduced to 1 mg/ml with unfolding buffer. This solution was dialysed against refolding buffer (2 M NaCl, 10 mM Tris pH7.5, 1 mM Na-EDTA, 5 mM BME 4 °C) in three steps (16 h, 4 h, 3 h). The dialysate was filtered (0.45 µM) and concentrated (30 kDa filter) for injection on a sephacryl S200 pre-equilibrated in refolding buffer. The injected loop was washed with refolding buffer at 5 times the loop volume. Proteins were separated over one column volume. Flowthrough, wash, and elution fractions every 1 ml were collected for analysis by SDS PAGE. Octamer-containing fractions were pooled and concentrated (30 kDa filter). Proteins were quantified using the Bradford assay detailed in 2.2.1. A final concentration of 50 % (v/v) glycerol was added for storage in useable aliquots (-80 °C).

2.5.4.10 Analytical Sizing

Analytical sizing was used to assess the multimeric state of proteins. A Superose 6 column was calibrated with standards (Cytiva), including ferritin (440), aldolase (158), conalbumin (75) and ovalbumin (44), where the number in brackets is the molecular weight in kDa. The recommended concentrations specified by Cytiva for each standard were used. To determine the molecular weight of the protein, the volume at which the purified protein was eluted was compared with the standard elution volume and the protein's known weight. The buffer used for the standards calibration was the same as the storage buffer for the protein of interest (POI).

2.5.4.11 Miniprep of DNA

Mini preps were carried out using a miniprep spin kit (QIAGEN). Pellets from overnight culture (5 mL) of 1x widom DNA were resuspended in resuspension buffer (250 µL). Lysis solution (250 µL) was added, and the tube was inverted 3 - 4 times and incubated (2 minutes, room temperature). Neutralisation buffer (350 µL) was added and inverted 4 - 6 times. The mixture was centrifuged at 13,000 rpm for 5 minutes, and the supernatant was transferred to a spin column. This was centrifuged for 1 minutes, and the flow through was discarded. Wash solution (500 µL) was added, the column centrifuged, the flow through discarded, and then repeated. The column was transferred to a fresh collection tube, incubated with dH_2O (50 µL) at room temperature for 20 minutes, and centrifuged for 2 minutes to collect the DNA.

2.5.4.12 Gigaprep of DNA

Each gigaprep can host 2.5 L cell culture. Pellets obtained from overnight culture of 12 x widom DNA (12 L) were resuspended in resuspension buffer, lysed, and neutralised as specified in the miniprep protocol in 2.5.4.11. The resulting solution was vacuum filtered on the provided Giga Filter. A binding buffer was added to the cleared lysate and inverted 10 times. This mixture was added to the provided Zymo-Spin VI- P column and vacuum filtered. The column was washed twice with the two

wash buffers provided in the kit. The column was transferred to a fresh tube, incubated (5 minutes, room temperature) with elution buffer and centrifuged (5 minutes, 4,000 rpm) to harvest the DNA.

2.5.4.13 Palindromic DNA Preparation

Transformed *E. coli* max efficiency DH5α cells with 12x widom DNA (114359) were grown in LB medium (75 mL) supplemented with 100 µg/ml streptomycin at 37 °C with shaking (18 h, 140 rpm). Overnight bacterial cultures (10 mL) were reseeded into fresh LB (1 L). Cultures were grown (1 h, 140 rpm), and cells were harvested by centrifugation (4,000 rpm, 25 minutes, 4 °C). DNA was purified using the pure plasmid gigaprep kit (Zymo) and quantified using 260 nm absorbance on a nanodrop. Plasmid DNA (10 mg) was incubated with *PmlI* (1 U/ 1 µg of DNA) for widom DNA extraction. The reaction mixture was divided into 1 ml fractions (10) and incubated (37 °C, 4 h). Following this, the reaction was stopped by inactivation of the restriction enzyme (65 °C, 30 minutes). The reaction efficiency was confirmed by separating samples on an agarose gel using control widom samples. Control widom sequences were generated using minipreps from the 601 widom sequence and PCR.

2.5.4.14 PCR of 601 DNA

The Q5 high-fidelity DNA polymerase protocol was followed with a 50 µL reaction size, with final concentrations in Table 2-1. The PCR reaction commenced with denaturation at 98 °C for 30 s, followed by 30 rounds of a three-step temperature cycle including denaturation at 98 °C for 5 s, followed by a 30 s annealing step at a range of temperatures (47, 47.7, 49.2, 51.6, 54.6, 56.9, 58.4, 59.1 °C) and finally an elongation step at 72 °C for 20 s. The PCR reaction is concluded with two minutes at 72 °C, and DNA is stored at -20 °C.

Reagent	Final Concentration
5x Q5 Reaction Buffer	1x
dNTPs	1 mM
Forward Primer ctggagaatcccggtgccg	5 µM
Reverse Primer acaggatgtatatatctgacacg	5 µM
Template DNA	25 ng
Q5 polymerase	0.04 U/µl
Q5 High GC enhancer	1x
MQ (sterile)	Made up to reaction volume

Table 2-1 PCR reaction components for widom DNA amplification

2.5.4.15 MonoQ Purification of Palindromic DNA

The excised DNA was diluted in MilliQ at a ratio of 1:10 for subsequent purification on the Mono Q column. The Mono Q was pre-equilibrated in five-column volumes of binding buffer (100 mM NaCl). The sample was loaded onto the column, after which the column was washed in five-column volumes wash buffer (400 mM NaCl) and eluted using a gradient of thirty column volumes (400 - 700 mM NaCl) followed by a second gradient for ten-column volumes (700 mM - 1 M NaCl). Flowthrough, wash, and elution fractions (0.8 mL) were collected for analysis using an agarose gel.

2.5.4.16 Agarose Gel

Agarose powder (4 g) was mixed with TAE buffer (100 mL, 40 mM Tris-acetate and 1 mM EDTA). This solution was microwaved for 1 minute until dissolved. Ethidium bromide (0.1 mg) was added before the gel was poured and allowed to set. 6X gel loading dye (0.25% bromophenol, 0.25% xylene cyanol, 30% glycerol) was added to DNA samples to a final concentration of 1x. Samples were

run alongside a DNA ladder (1 kb and 50 bp) to estimate DNA size. The gel was run at 80 V in TAE, and resulting DNA bands were visualised using UV.

2.5.5 MTase Glo Methyltransferase Assay

The MTase Glo methyltransferase assay was used with a few modifications in the 384-well plate format. Firstly, the pH stability of the assay buffer (20 mM Tris pH 8.5, 5 mM $MgCl_2$, 250 mM total anion, 1mM DTT, 0.1% CHAPS (3-((3-cholamidopropyl) dimethylammonio)-propane sulfonate) was determined with a concentration range of $NaHCO_3$ (0 mM) with NaCl (250 mM) to $NaHCO_3$ (250 mM) with NaCl (0 mM). Dot1L dependence (BPS bioscience or purified in-house at AstraZeneca) and substrate K_M were determined for use in the assay. For DOT1L inhibition, SYC-522 was pre-dispensed (100 µM, Echo 650). S-adenosyl homocysteine (SAH) standards were prepared by serial dilution in the range of 0 - 1000 nM for testing varying inorganic carbon concentrations. Hela oligo nucleosomes (Reaction Biology Corp, 0.1 mg/ml) and S-adenosyl-L-methionine (SAM, 1 µM) were mixed in the reaction buffer containing the range of inorganic carbon concentrations previously stated and incubated for 10 minutes. The methyltransferase (MTase) reaction was started by adding Dot1L (10 nM) was added to start the reaction. Plates were centrifuged (2 minutes, 1,000 rpm) and placed on an orbital shaker (2 minutes). 10 X MTase Glo reagent was thawed on ice, mixed, and equilibrated to room temperature. The MTase reaction was stopped using TFA (0.5%) after the appropriate incubation time (10 - 60 minutes). Plates were centrifuged (2 minutes, 1,000 rpm), placed on an orbital shaker (3 minutes) and kept at this stage until all reaction incubations were complete. 6X MTase Glo reagent (2 µl) was added to wells, plates centrifuged, shook, and incubated (30 minutes, room temperature). MTase Glo detection solution (10 µl) was added to all wells, and plates were centrifuged, shaken, and incubated (30 minutes, room temperature). Luminescence was read on a plate-reading luminometer.

2.5.6 Dot1L Inhibition under Varying Carbon Dioxide Incubation

2.5.6.1 MTT Assay

HEK293 cells cultured as previously stated were treated with pinometostat (0, 0.5, 1 µM) for various incubation lengths (2,3,4,7,10 days) and assessed for viability using the MTT (3-[4,5-dimethylthiazol-2-yl]-2,5 diphenyl tetrazolium bromide) assay (Abcam). Control sample cells were treated with DMSO to reflect 100% viability. Cells were reseeded every 2 days, and fresh media with pinometostat or DMSO was used. After the appropriate incubation time, cells were centrifuged (1,000 rpm, 5 minutes) and counted. The treatment media was discarded, and cells were resuspended into phenol red and serum-free media for use in the assay. Cells were seeded into a 96-well plate and mixed with an equal volume of MTT solution. Cells were incubated (37 °C, 3 h, 5% CO_2). Following incubation, MTT solvent was added to the reaction incubated in the dark on an orbital shaker (15 minutes), and the absorbance at 590 nm was read.

2.5.6.2 Elisa Plate Protocol

Pre-coated Histone 3 (H3) Modification Kits, including pan methyl H3K79 and H3 multiplex kit, were used without modifications. In brief, native nucleosomes were extracted (described in 2.5.1) and quantified by Bradford assay (described in 2.2.1). Histone extracts were diluted to be within the range of the assay, this was normalised between samples using the protein concentration from the Bradford assay. A dilution series of the ELISA assay control protein was performed to produce a standard curve. Standards (1 - 100 ng/µl) and histone extracts (25 ng, 3 µl) were added to the plate in the provided antibody buffer and incubated (room temperature, 90 minutes). Wells were washed three times with the provided wash buffer. The secondary antibody was added, and the plate was incubated (1 ug/ml, room temperature, 60 minutes). Wells were washed six times, and the plate was incubated (8 minutes) with the colour developer solution. Finally, a stop solution buffer was added, and absorbance was read at 450 and 655 nm within 4 minutes.

2.5.6.3 HEK293 Cell Preparation

HEK293 cells were treated with pinometostat (1 µM, 10 days) as specified in section 2.5.6. On day 8, cells were seeded into 6 well plates (0.18 x10^6 ml, 2 ml). On day 9, the media for the 24 h CO_2 samples was exchanged for CO_2 experimental media with fresh inhibitor/DMSO and placed in the appropriate CO_2 incubator (5%, 10%, 20%). CO_2 experimental media for all the other CO_2 incubation time points (1, 3, 6, 9 h) was placed in the appropriate CO_2 incubator overnight (until day 10) when it was exchanged with the cell culture media on the remaining samples. HEK293 cells were incubated for the specified time points and harvested in an RNase-free environment with PBS, and cells were transferred to a microcentrifuge tube. Cells were pelleted (1,000 rpm, 10 minutes), the supernatant was decanted, and stored at -80 °C until required.

2.5.6.4 RNAse-free Environment

All materials and the workspace were cleaned before and after each RNA experiment. DNAZap (Thermo Fisher Scientific) was first applied, followed by RNAse free water, 70% ethanol, RNase Zap (Thermo Fisher Scientific), and finally 70% ethanol. PBS and water were treated with diethylpyrocarbonate (0.1%) for 20 minutes and autoclaved before use. All plastics used were certified as RNAse and DNAse free.

2.5.6.5 Total RNA Purification

The Total RNA purification kit (Norgen) was used without modifications. In brief, buffer RL was added to the frozen cell pellet and cells were lysed by vortexing (15 s). To the resulting lysate, ethanol was added with vortexing (10 s). This solution was added to the spin column and centrifuged (1 minute, 6,000 rpm). The flowthrough was discarded, and the on-column DNA removal kit (Noragen) was used. DNAse 1 was added to the enzyme incubation buffer (10 minutes). Following this, the column was washed with wash buffer 3 times using centrifugation (1 minute, 14,000 rpm). After the final wash, the column was centrifuged (2 minutes, 14,000 rpm) to ensure the resin was dry. The column was

transferred to an elution tube. Each RNA sample was quantified on the nanodrop twice, and the readings were averaged. RNA quality was assessed using the absorbance ratio at 260/280 and 230/260.

2.5.6.6 cDNA Synthesis

RNA extracted from HEK293 cells treated with DMSO were processed using the qScript cDNA Synthesis kit (QuantaBio). A master mix for the reaction contained 10 parts nuclease-free water, 4 parts 5x qScript reaction mix, and 1 part qScript reverse transcriptase (RT). RNA (100 ng, 5 µl) was added to the master mix (15 µl) vortexed (10 s) and centrifuged briefly. The reaction was placed in a thermocycler and ran for one cycle (22 °C, 5 minutes, followed by 42 °C, 30 minutes and finally 85 °C, 5 minutes). cDNA was stored (-20 °C) until required.

2.5.6.7 Quantitative Polymerase Chain Reaction (qPCR)

The components of the qPCR reaction (20 µl reactions for each biological sample with 3 technical replicates) are outlined in Table 2-2. To prepare these reagents SSo Advanced Universal SYBR Green Supermix (Bio-Rad) was thawed on ice protected from light, and cDNA and primers were diluted in water. The master mix was made on ice and scaled according to how many samples were being tested. Samples were placed in the CFX Connect Real-Time System. The procedure run is outlined in Table 2-3.

Component	Volume/ µL	Final concentration
2X SSo Advanced Universal SYBR Green Supermix	31.5	1x
Forward primer	1.9	250 nM
Reverse primer	1.9	250 nM
cDNA	1	2.5 ng/µl determined from RNA concentration
Nuclease free water	26.7	-

Table 2-2 Components for qPCR reaction

Temperature / °C	Length time/minutes
95	3
95	0:30
60	0:30

Plate Read	
72	0:30
Repeat steps 2-4, 40 times.	
95	0:30
55	0:30
50	0:30
Melt curve 55 to 95 in increment 0.5	0:10 - Plate Read

Table 2-3 qPCR parameters

2.5.6.8 RNA Sequencing and Data Analysis

RNA (≥ 50 ng/µl) was provided to Cambridge Genomics Services. RNA-seq libraries were prepared using a TruSeq-stranded total RNA sample preparation kit. RNA-seq libraries were sequenced on the NextSeq 2000 using the P2 -100 kit. The binary base call files (bcl) from the sequencer were converted into Fastq files using bcl2fastq. The quality of the sequencing data was analysed with FastQC v0.11.4, and reads were trimmed using TrimGalore v0.5.0. Reads were mapped to the reference human genome (Ensembl GRCh38 GTF file) using STAR v2.7.9 to generate BAM files. Reads that map to certain genomic locations were determined using Picard Tools v2.1.1. To count the number of reads mapped to each genomic feature, HTseq v0.6.1 was used. Normalised and log-transformed data was obtained using DESeq2 v1.24.1 and samples were clustered using Principal Component Analysis (PCA). Finally, EdgeR v3.26.5 was used to filter genes with low read counts, calculate normalisation factors, correct for GC content and gene length bias, and finally produce lists of genes with differential expression. Gene lists shared between specified pairwise comparisons were analysed in Enrichr.

3. A Proteomic Screen for Carbamylation in a HEK293 Lysate

3.1 Overview

As previously outlined in Chapter 1, carbamylation is the reversible post-translational modification that occurs on proteins at neutral amine sites. CO_2 has been implicated in a diverse number of biological processes, but surprisingly, the knowledge of proteins that are direct CO_2 targets remains limited. Furthermore, the carbamate PTMs published to date have been shown to mediate downstream cellular signalling and adapt to environmental changes. Three biologically relevant carbamate PTMs have been previously identified on mammalian proteins, including haemoglobin, connexin and ubiquitin. [83] This chapter aims to identify and validate carbamylation sites on mammalian proteins by screening the HEK293 lysate. Mass spectrometry (MS) has been instrumental in the identification and quantification of a vast number of eukaryotic PTMs.[90] The trapping methodology, which irreversibly modifies carbamates for MS application, developed by Linthwaite *et al.* [1] was applied to a HEK293 lysate to identify novel mammalian carbamate sites. The HEK293 proteomic screen discussed in this chapter was performed to address the hypothesis that the carbamate PTM is widespread in the mammalian proteome and varies according to carbon dioxide concentration.

A fractionation-based workflow was developed following initial test experiments to improve coverage across the proteome. Peptides were fractionated by hydrophobicity and recombined into 16 fractions to reduce the complexity of each sample injection into the mass spectrometer.

The carboxyethyl PTM is detected by a shift of 72.02 Da on modified lysine sites whereby the CO_2 modification has a mass of 44.01 Da and the trapping reagent group has a mass of 28.01 Da. A set of conditions was defined to classify carbamylation sites into low, medium, high and no confidence by assessing support from y and b ions and isotopes on the mass spectra. In addition, the bioinformatics pipeline was optimised to reduce the number of false positive identifications.

The lysate screening was repeated using Ci with the 13C isotope instead of 12C for further verification of hits. This additional screening allowed actual carboxyethyl modifications to be separated

from methylglyoxal-derived advanced glycation end products (AGEs), which are PTMs that also exhibit a 72.02 Da mass shift on modified lysine sites.[91]

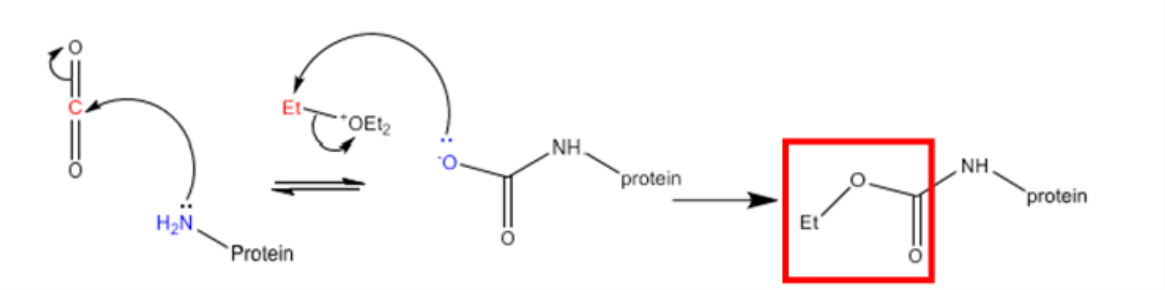
The background to this chapter covers the method for trapping carbamates, previously reported mammalian carbamates, the methylglyoxal-derived AGE, the detection of PTMs by mass spectrometry and the database search algorithms used in this investigation. The results discussed in this chapter include the optimisation of sample preparation and the analysis workflow to produce a list of carbamates identified from 12C and 13C lysate screens.

3.2 Proteome-Wide Carbamate Detection Strategies.

The two carbamate MS detection strategies discussed in sections 3.2.1 and 3.2.2 are distinct from the alternative techniques discussed earlier (sections 1.3 and 1.5) because they are experimental approaches which can be applied on a proteome scale to localise the carbamate PTM to a specific lysine or N-terminus residue of a protein.

3.2.1 Triethyloxonium Trapping

Previously, the direct, widespread detection of carbamylation on a specific lysine or N-terminal residue on a protein was regarded as challenging due to the transient nature of the modification. In 2018, Linthwaite *et al.* addressed this problem by developing a trapping technique for carbamate PTM identification.[1] Linthwaite et al.'s method utilized the Meerwins reagent, TEO which artificially modifies carbamylated neutral amines by covalent modification with an ethyl group, creating an irreversible PTM which is suitable for analysis by Liquid Chromatography-Tandem Mass Spectrometry (LCMSMS). Scheme 3-1 details the reaction mechanism for TEO trapping of a carbamate modification onto a protein.



Scheme 3-1 Irreversible trapping of the carbamylated PTM using triethyloxonium (TEO). The trapped CO_2 modification corresponds to a mass shift of 72.02 Da, as shown by the group highlighted in red.

3.2.2 Chemoproteomic Carbamate Identification

In 2022, King *et al*. proposed a new chemical biology-based method for discovering carbamate sites.[72] The researchers proposed and verified the use of a lysine-selective chemical probe to identify carbamate sites using a competitive activity-based profiling method, where CO_2 and the chemical probe compete to modify reactive lysine sites. The selected chemical probe was the CO_2 analogue, isocyanic acid (OCNH), which reacts irreversibly at lysine to form homocitrulline. Quantification of the OCNH-modified lysine adducts was measured by MSMS under varying PCO_2 to detect carbamates. This technique was verified using known CO_2 target proteins and then applied to the *Synechocystis* proteome. King *et al*.'s method offers an orthogonal approach to Linthwaite *et al*.'s carbamate identification and validation method.

3.2.3 Mammalian Carbamates

In the literature, the mammalian proteins identified as carbamylation targets are haemoglobin, connexin 26 and ubiquitin. The biological relevance of these carbamates and the detection method used for identifying these CO_2 targets are detailed in sections 3.2.3.1 - 3.

3.2.3.1 Haemoglobin

Haemoglobin (Hb) is a protein found in red blood cells vital for oxygen transport from the lungs to the tissues. The Bohr and Haldane effects define the oxygen and carbon dioxide transport properties of haemoglobin, respectively.[92] The Bohr effect describes the changes in haemoglobin's affinity for

oxygen in response to metabolic activity, whereby the affinity of Hb for oxygen is maximised in the lungs and lowered in the tissues with high oxygen demand. The Haldane effect describes that the affinity of Hb for CO_2 is dependent on haemoglobin's degree of oxygenation. Christiansen *et al*. proved that the deoxygenation of the blood in the tissues increases the affinity of Hb for CO_2. Conversely, the oxygenated blood in the lungs facilitates the excretion of CO_2 by exhalation.[93]

Specifically, there are four CO_2 binding sites in haemoglobin, the N-terminal valines of each Hb chain. In deoxygenated blood, CO_2 binds to Hb to produce carbaminohaemoglobin.[94] This observation highlights the physiological role of carbamate formation in the ventilatory response. The carbamylation sites on Hb were identified by mutation of the N terminal valine with cyanate, which inhibited CO_2 uptake[82,95] and 13C-NMR spectroscopy.[96]

3.2.3.2 Connexin

Connexins are hexameric gap junctions positioned in the intracellular membrane. When these hemichannels are open, neuronal depolarization occurs, and cellular conductance is altered.[97] Under increased CO_2, the CO_2-sensitive connexins Cx26, Cx30 and Cx32 [86] are opened, releasing ATP. The direct interaction of CO_2 forming a salt bridge between the subunits of Cx26 has been identified by mass spectrometry [98] and verified by mutagenesis, [99] X-ray crystallography, [100] and fluorescence imaging.[101] Connexins are ubiquitously expressed, and this carbamylation example suggests CO_2 could act as a signalling molecule in a diverse range of cellular processes.

3.2.3.3 Ubiquitin

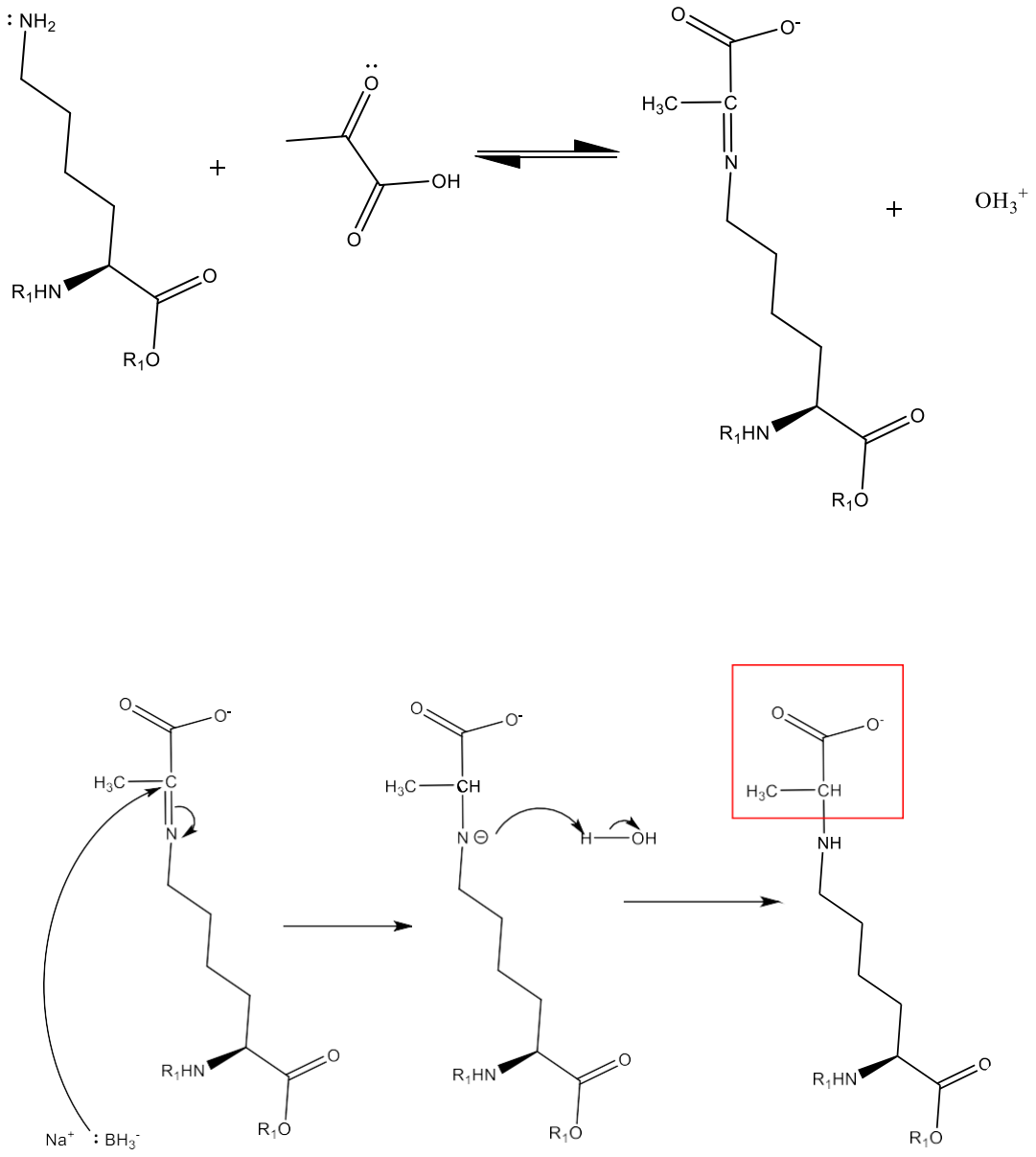
Ubiquitin is a small, eukaryotic protein that regulates cellular activity. A range of heterotypic and homotypic ubiquitin crosslinks exist, each with a different function, as depicted later in Figure 4-1. For example, Ub K48 crosslinking is a proteasome targeting signal that results in protein degradation, whilst the less well-characterized and prevalent K33 ubiquitin chains regulate enzyme activity.[102] Linthwaite *et al*. verified the presence of carbamylation at K33 and K48 on ubiquitin by mass spectrometry and ^{13}C-NMR. Furthermore, the direct downstream effects of K48 carbamylation were studied using mutagenesis controls in the context of the NF-κB pathway.[84]

3.3 Advanced Glycation end products (AGEs)

Protein glycation arises from a reducing sugar covalently bonding to a primary amine and is a PTM commonly associated with oxidative stress.[103] Glycation adducts are commonly derived from glucose and fructose metabolism. Proteomic-based methods have been used to identify and quantify these modification products. [91]

Carboxymethyl-lysine (CML) is a well-characterised AGE [104] identifiable by a 72.02 Da mass shift on the modified lysine. Initial identification of the CML group was aided by chemical synthesis, whereby pyruvate reacted with lysine to form a Schiff base. Then, the imine bond was reduced by sodium borohydride, as shown in Scheme 3-2.[105] The *in vivo* reaction differs slightly from the chemical synthesis reaction. It involves glucose reacting with lysine to form a Schiff base, a rearrangement to an intermediate Amadori product followed by irreversible oxidation, which results in the CML product.[106] A secondary *in vivo* CML formation mechanism involves lysine's reaction with the 1,2 dicarbonyl glyoxal via the Cannizzaro reaction.[107]

AGEs cover a wide research area due to their association with ageing and disease and a range of glycation adducts exist.[108] However, the AGEs are discussed in this investigation due to the similarities between the CML modification and the TEO-trapped carbamates. These modifications target the lysine amino acid, are the same mass (72.02 Da), and are formed spontaneously.



Scheme 3-2 The chemical synthesis of the carboxymethyl-lysine post-translational modification corresponds to a mass shift of 72.02 Da, as shown by the group highlighted in red.

3.4 Mass Spectrometry

Mass spectrometry has been applied to various research areas due to the diverse instrumentation developed for each pipeline step and the ability to obtain quantitative and qualitative information.[109] Importantly, MS has become an instrumental tool in the proteomics field and is the preferred technique for site-specific identification of PTMs.

A bottom-up MS regime is followed here, whereby a protease digests the proteins, and peptides are injected into the mass spectrometer, as described in section 2.3 and Figure 3-1. Before the MS analysis begins, the injected peptide mixture is cleaned by a trap column (C18), which concentrates the sample of interest by extracting non-analyte material, which can affect the run.[110] The peptides are then eluted and directly applied to the resolving column, coupled to the mass spectrometer. Reverse phase high-performance liquid chromatography (HPLC) on the resolving C18 column is carried out. Hydrophobic peptides will elute last because these peptides have the strongest Van der Waals interactions with the matrix.[111] This method is known as trap and elute, which removes contaminants and increases the throughput of sample loading.

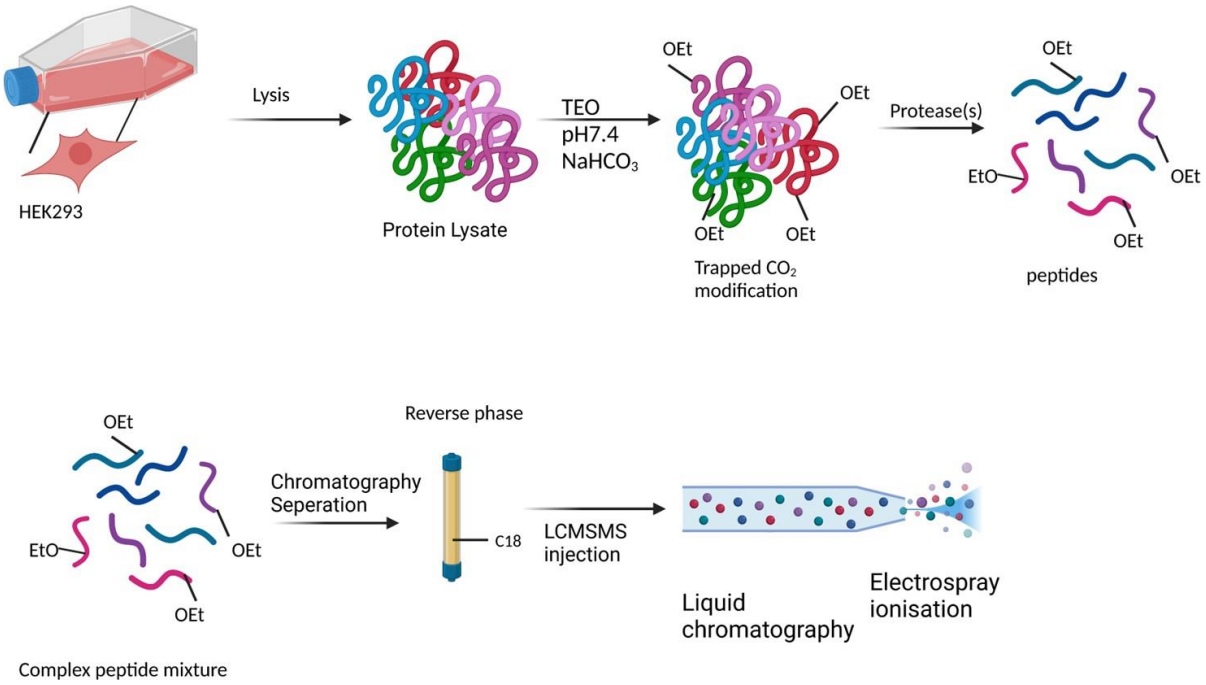


Figure 3-1 Sample preparation for proteomic screening of carbamylation in a HEK293 lysate created using BioRender.

The first step in the MS pipeline is to ionize molecules of interest and create precursor ions. The ionisation method used in this workflow was electrospray ionisation (ESI), illustrated in Figure 3-2. For ESI, peptides are dissolved in 0.1% formic acid and passed through a needle at high electric potential.[112] The applied electric field causes the dispersal of the liquid into a spray of small, highly charged droplets. To form ions in the gaseous phase, solvent evaporation is required and mediated by increasing the local temperature at the ESI source or by a stream of nitrogen gas. As the solvent

evaporates, the charged droplets reduce in size, reaching a critical point where ions at the surface of the droplets are ejected into the gaseous phase.[113]

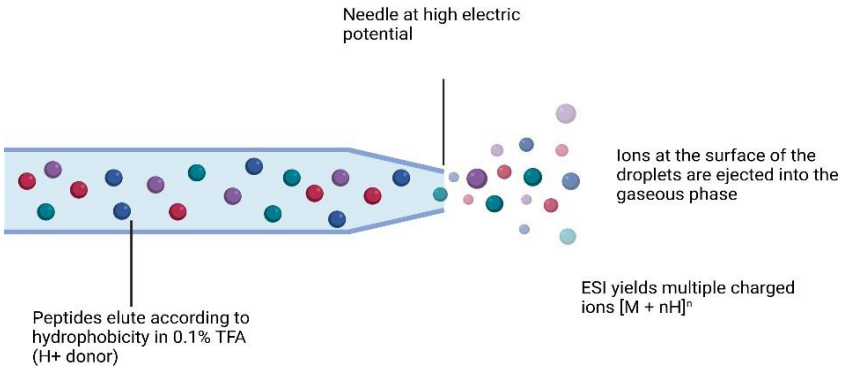


Figure 3-2 The Principle of Electrospray Ionisation

Biomolecule analysis requires using tandem mass spectrometry, commonly referred to as MSMS. This technique relies on two rounds of ion separation by two mass analysers, which separate ions based on their mass-to-charge ratio (m/z). In this investigation, a quadrupole-time of flight (Q-TOF) mass spectrometer was used, and the ion path through the analyser is shown in Figure 3-3. A Q-TOF mass spectrometer uses the Q mass analyser to separate the precursor ions produced by ESI and the TOF mass analyser to separate the product ions produced by fragmentation in the CID cell.[114] In the workflow used here, only the thirty highest precursor ions are selected for further fragmentation due to applying the data-dependent acquisition mode at a threshold of 30.[115] Product ions are grouped into b and y ions, where b ions extend from the peptide N-terminus and y ions from the C terminus.[116] The b and y ions travel through the TOF tube. They are detected on a microchannel plate detector by voltage pulses, which are converted to the ion intensity and the time taken for the ion to reach the detector represents the m/z value.[117] Mass spectra are plotted from the relative intensity of ions against the m/z. The pattern of b and y ions are used to sequence peptides, leading to protein group assignment. An extensive array of algorithms is available to apply database searching and novel PTM discovery, and those utilized in this investigation are discussed in section 3.5.

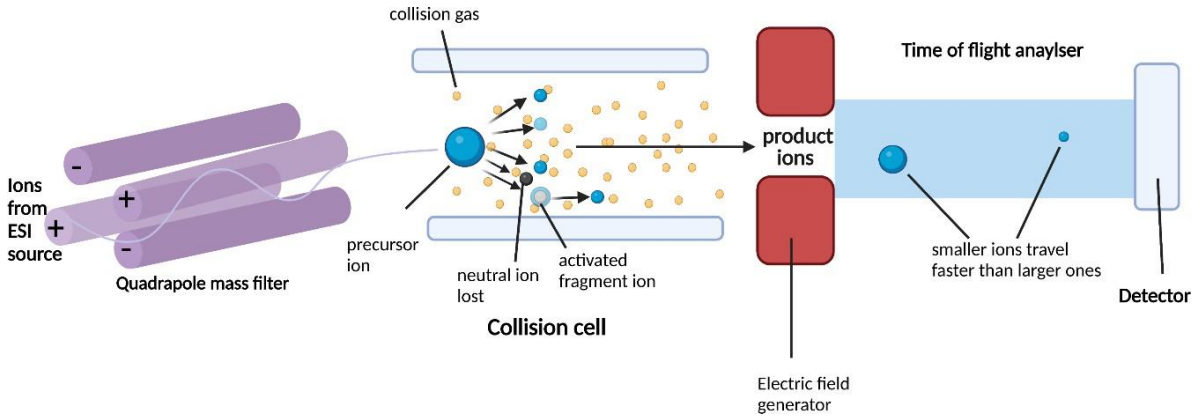


Figure 3-3 The path of ions in a quadrupole- time of flight (Q-TOF) mass spectrometer.

3.5 Database Search Algorithms

To screen HEK293 lysates for the presence of carbamylated proteins, two database search algorithms were used, namely, PEAKs [118] and Protein Pilot.[119] The first step in both analysis pipelines was to convert the raw mass spectrometry data into a readable format suitable for processing by the search algorithm. Raw MSMS data is complex because it consists of distributed m/z values for each ion detected. The peak-picking data conversion algorithm simplifies and filters the raw data into centroided data, which contains only the most intense m/z peak for each ion detected.[120] For PEAKs, this step was completed using proteowizard, where raw, wiff files were converted to mgf files.[121] Meanwhile, for Protein Pilot, this step was already built into the analysis software.

The centroided data is then analysed by a specific database search algorithm consisting of three stages. Firstly, the observed precursor and product ions are matched to peptide sequences. Secondly, the peptide-to-spectrum match (PSM) is given a confidence score, and thirdly, peptides with a high enough PSM confidence are mapped to proteins. Finally, the output data is filtered using a FDR threshold for both peptides and proteins.[122] To aid the FDR calculations, a decoy protein search algorithm is commonly used to assign a minimum threshold for PSM scores and improve the accuracy of statistical thresholding.[123] Each step is conducted by a unique algorithm. Here, the peptide

identification algorithms for PEAKs and Protein Pilot are discussed due to the importance of this step for PTM identification.

Traditionally, there are two main methods for peptide identification from tandem MS spectra: database search and de novo sequencing. In database search sequencing, the search space is limited by querying a database for the peptide that best explains the MSMS spectrum fragmentation pattern. Meanwhile, in de novo peptide sequencing, the best amino acid match is assigned for each product ion detected. [124] However, both methods have limitations, in particular, database searching can suffer from low identification rates [125] and de novo sequencing from imperfections in MSMS spectra, leading to incorrect amino acid assignment.[124] A third approach, which combines the de novo and database search strategies, known as the tag-based search algorithm, which has been developed to combat these issues. In this method, the software matches spectra to a partial peptide sequence known as peptide sequence tag using de novo sequencing, which is then queried against the sequence database to interpret the remainder of the sequence.[126] PEAKs and Protein Pilot both use a tag-based search algorithm for peptide identification to improve the sensitivity and accuracy of data. However, the particulars for each analysis tool differ and are discussed in sections 3.5.1 - 3.5.2.

3.5.1 PEAKs

Peptide identification in PEAKs is first conducted by a de novo sequencing algorithm [127] whereby each sequenced amino acid in the peptide sequence tag is associated with a confidence score. These tags are quality screened, amino acids that pass the confidence threshold are retained, and those that do not are replaced with mass segments that correspond to the spectra. PEAKs peptide sequence tags are comprised of amino acids and mass segments which are searched against the protein database. The number of amino acids in common (CAA) between the de novo sequence tag and the database peptide is used to score proteins, and the proteins with the highest CAA score are shortlisted for downstream analysis. These selected proteins are then digested to create a list of possible peptide hypotheses, which are modified in turn by each of the PTMs specified by the user in

the run. This process creates multiple possible peptide queries, which are compared and scored using the spectra data to identify significant PSMs.

3.5.2 Protein Pilot

Protein Pilot utilises the paragon algorithm for peptide identification.[128] This algorithm relies on two facets to acquire peptide hypotheses, including computation of sequence temperature values (STVs) and modelling peptide feature probabilities. Peptide features correspond to PTMs, amino acid substitutions and cleavages, each associated with a probability encoded in the software. In this investigation, carbamylation was given a probability of 0.001 on lysine compared with other pre-set modifications, such as acetylation at 0.01 on the protein N-terminus and carbamidomethylation at 0.994 on cysteine following iodoacetamide treatment.

The paragon algorithm conducts a taglet-based search, whereby de novo peptide sequences are split into two to three amino acid sections. These taglets are composed of modified and unmodified peptide features depending on the mass-to-charge shift identified in the MS spectrum and whether the probability score of the peptide feature passes a predetermined threshold. The taglets are quality screened and queried against the database which is split into sequences of seven amino acids in length. The number and confidence of assigned taglets to the database sequences are quantified by the metric known as the STV. For regions with high STVs, all possible peptide hypotheses are considered, including low-probability peptide features. In contrast, for regions with low STVs, the search space is narrowed, and only the most probable features are considered. Finally, a threshold is applied to the combined results of the STV and the feature probabilities to select the peptide hypotheses which are suitable for PSM scoring.

PEAKs and Protein Pilot were used in this investigation to address the false positive identification of carbamates and verify real carbamate hits, as discussed in section 3.8.

3.6 Protocol Optimisation Results

The optimisation steps to produce the HEK293 lysate carbamate screening protocol detailed in section 2.3 are discussed here (section 3.6). Section 3.6.1 details the biomolecule quantification performed at each step of the sample preparation process. Section 3.6.2 details the optimisation of the MS method to increase the coverage of the mammalian proteome. These optimisation steps were performed in collaboration with a postdoc in the Cann lab.

3.6.1 Biomolecule Quantification

Biomolecule quantification was performed at three stages throughout the sample preparation process. These three stages were before trapping, before digestion and after digestion. The Bradford assay was performed before trapping to determine the starting amount of protein in the HEK293 lysates. The BCA assay was performed before digestion to obtain a consistent protein-to-protease ratio across the sample set. Finally, the Pierce Peptide assay was performed after digestion to normalise the loading of peptide injections on the mass spectrometer.

HEK293 lysates with a protein amount of 3-4 mg were prepared for trapping as described in section 2.3.3. Protein amounts were estimated at this stage using the Bradford assay. The Bradford dye binds to amino acid side chains, resulting in deprotonation of the dye via hydrogen transfer, which shifts the dye's absorbance from 465 nm to 595 nm.[129] Section 2.2.1 describes the procedure for the Bradford assay. Figure 3-4 is a standard curve which plots the absorbance detected over a range of known BSA concentrations. The equation of the line for the curve is given in Equation 3-1, which is used to quantify unknown protein concentrations.

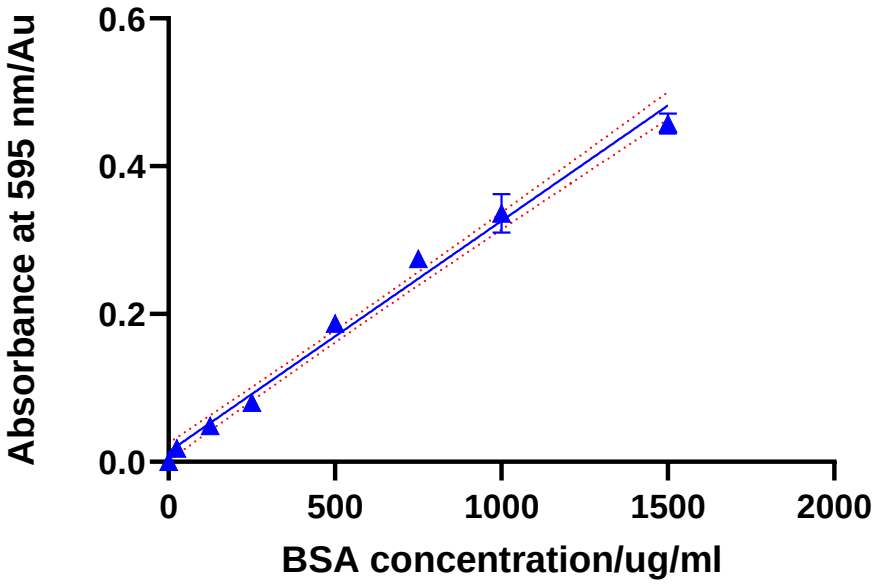


Figure 3-4 The absorbance detected at 595 nm versus the known concentration of bovine serum albumin (BSA) protein standards to obtain a standard curve for protein quantification. A linear regression is fitted to the data, 95% confidence limits are plotted as red dotted lines, and the R squared value for the fit is 0.9857.

$$y = 0.000312x + 0.01347$$

Equation 3-1 Standard curve equation from Figure 3-4.

Following trapping, samples were prepped for S-trap digestion as described in section 2.3.4. The samples were resuspended in the S-trap lysis buffer containing 5% SDS, a chemical incompatible with the Bradford reagent.[130] Therefore, samples were quantified at this stage using the BCA assay. Two reaction steps occur in the BCA assay. The first step involves reducing Cu^{2+} from copper (II) sulfate to Cu^{1+} by the amide backbone of proteins in the Biuret reaction process. Cu^{1+} ions then react with BCA to form a complex which absorbs at 562 nm.[131] Section 2.2.2 describes the procedure for the BCA assay. Figure 3-5 is a standard curve which plots the absorbance detected over a range of known BSA concentrations. The equation of the line for the curve is given in Equation 3-2, which is used to quantify unknown protein concentrations.

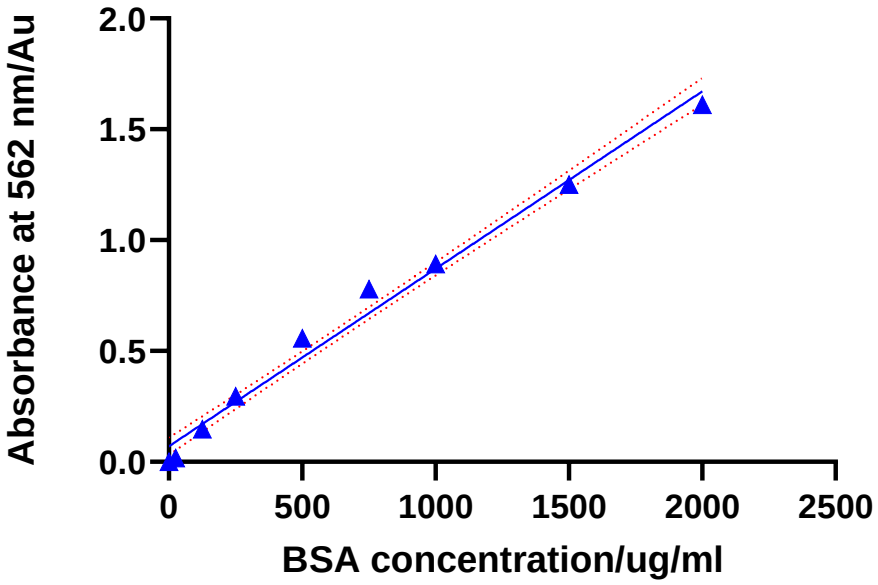


Figure 3-5 The absorbance detected at 562 nm versus the known concentration of bovine serum albumin (BSA) protein standards to obtain a standard curve for protein quantification. A linear regression is fitted to the data, 95% confidence limits are plotted as red dotted lines, and the R squared value for the fit is 0.9850.

$$y = 0.07002x + 0.0008004$$

Equation 3-2 Standard curve equation from Figure 3-5

Following digestion, peptides were quantified using the Pierce Peptide assay. The peptide assay is a modified version of the BCA assay, with the first step being the biuret reaction followed by the formation of a Cu^{1+} complex, which absorbs at 480 nm (Thermo Fisher Scientific, 23275). Section 2.2.3 describes the procedure for the Pierce Peptide assay. Figure 3-6 is a standard curve that plots the absorbance detected over a range of known peptide concentrations using the provided standard. The equation of the line for the curve is given in Equation 3-3, which is used to quantify unknown peptide concentrations.

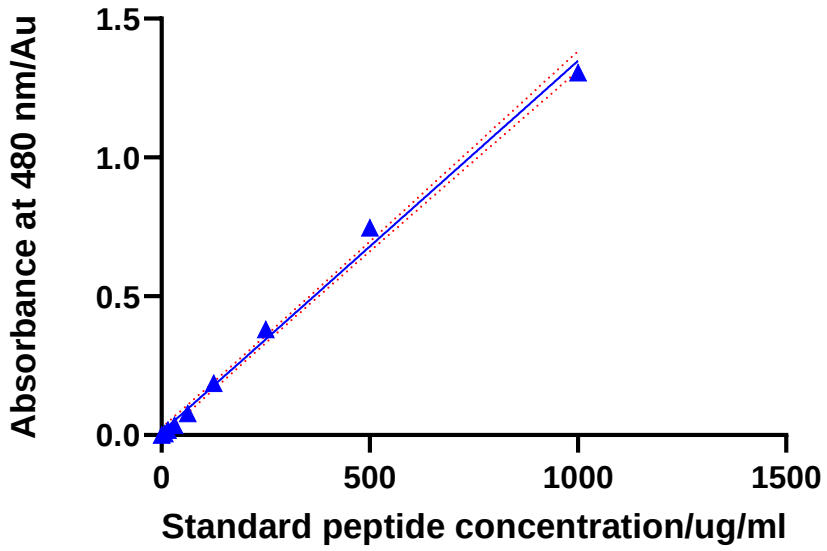


Figure 3-6 The absorbance detected at 480 nm versus the known concentration of a standard peptide to obtain a standard curve for peptide quantification. A linear regression is fitted to the data, 95% confidence limits are plotted as red dotted lines, and the R squared value for the fit is 0.9942.

$$y = 0.001337x + 0.01061$$

Equation 3-3 Standard curve equation from Figure 3-6

It is important to note that the standard curves shown in Figure 3-3 – 3-6 were re-measured between sample sets to obtain accurate biomolecule estimations. Figure 3-7 shows the quantity of biomolecules present at each stage from three replicate samples in the 12C HEK293 lysate screen. The amount of protein reported by the BCA assay after trapping is 11.4% less than that reported by the Bradford assay before trapping. The digestion stage also reports a loss in biomolecules where the peptide amount quantified is 35% less than the protein detected by the BCA assay. However, different assay procedures are used at each quantification step, so a direct comparison of the exact amounts is not suitable. A previous study reported that the precise amount of protein is not directly comparable between the Bradford and BCA assay because the Bradford assay underestimates protein amounts whilst the BCA assay overestimates.[132] Despite this, the biomolecule quantification steps were used to

determine the starting amount of protein required for the downstream preparation procedures of protein trapping, digestion, and peptide fractionation. The biomolecule amounts reported in Figure 3-7 were suitable for the MS sample preparation pipeline, and therefore, between 3 and 4 mg of starting protein were used for the HEK293 lysate screening datasets.

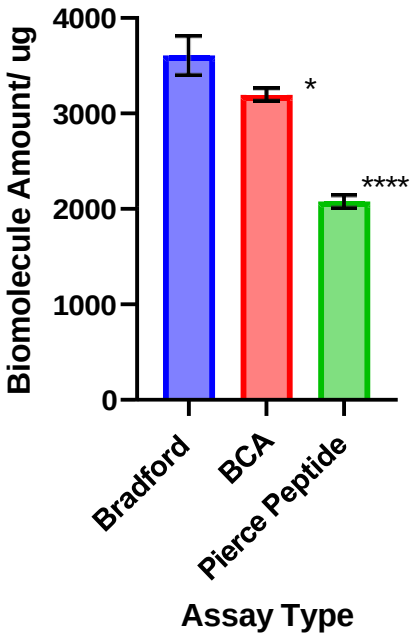


Figure 3-7 The quantity of biomolecule determined from each assay type at different stages of the HEK293 lysate carbamate screening preparation process. All data points are represented as the mean, and the error bars are the standard deviation from the mean. A one-way ANOVA assessment showed a statistically significant difference between the amount of biomolecule detected at each stage of the preparation process at a significance threshold of $p < 0.05$ where n=3. Asterisks indicate levels of significance (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$) from multiple comparison tests (MCTS). The MCTs compare the protein amounts quantified by the Bradford assay with those quantified by the BCA after trapping and those quantified by BCA before digestion with the peptide amount quantified by the peptide assay post-digestion.

3.6.2 Optimisation of the Mass Spectrometry Workflow

HEK293 coverage obtained from trapped samples by LCMSMS is presented in this section (3.6.2) using three coverage metrics. These metrics include the number of protein groups, unique peptides and the percentage of unique ethylated peptides identified from three replicate samples. All data discussed in this section was obtained using PEAKs.

3.6.2.1 Sample Clean Up

The preliminary proteomic experiments in this investigation were performed on complex peptide mixtures injected onto the mass spectrometer following digestion. Due to the complexity of the mixture, an additional clean-up step before injection was considered, known as HILLIC, as described in section 2.3.5. Figure 3-8 shows the HEK293 coverage when using the HILLIC column clean- up step before the C18 trap and elute (sections 2.3.6 and 3.4) versus only using the C18 clean-up method.

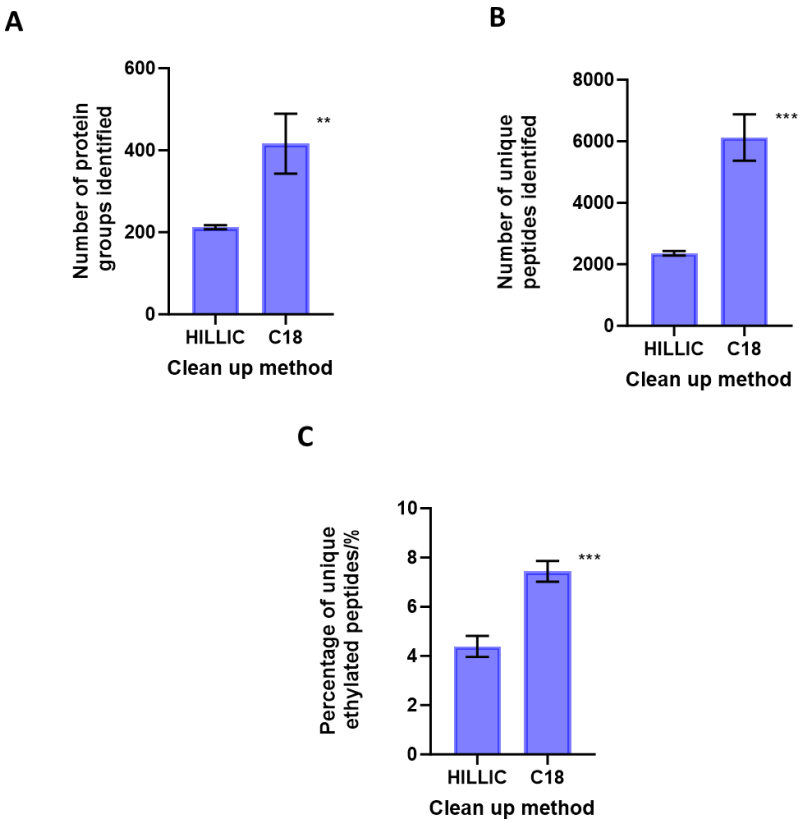


Figure 3-8 The number of (A) protein groups, (B) unique peptides and (C) the percentage of unique ethylated peptides identified in a HEK293 lysate trapped with 12C inorganic carbon versus the clean-up method. Samples were processed with HILLIC before mass spectrometry injection and C18 clean-up. In contrast, C18 samples were directly injected and cleaned up only by the C18 column. All data points are represented as the mean where n=3. The error bars are the standard deviation from the mean; in some cases, these errors are smaller than the individual data points. A one-sample t-test assessed the statistical significance of the differences in the coverage metrics between the sample clean-up methods. Asterisks indicate levels of significance (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$)

The coverage for all three metrics was reduced significantly using the HILLIC clean-up method, as shown in Figure 3-8. Of particular concern was the loss in the percentage of unique ethylated peptides by 3% when using the HILLIC column versus only using the C18 trap and elute. Ethylated peptides are those that TEO has modified, and therefore, if fewer are identified, the probability of identifying carboxyethyl modifications on proteins reduces. The HILLIC column separates peptides based on polarity, whereby the binding step utilizes a mobile phase, which is 85% organic. The polar compounds bind to the column, and subsequently, these compounds are eluted using a mobile phase that is 5% organic. The column does not retain the hydrophobic peptides, and these are eluted during the wash stage.[133] Resultant peptides are primarily hydrophilic, which explains the loss of ethylated peptides identified following HILLIC treatment. Consequently, this method was not pursued further in HEK293 lysate screening.

3.6.2.2 Peptide Fractionation and Recombination for LCMSMS Injection

A previous study identified ~8500 protein groups in a HEK293 lysate[134] more than twenty-fold greater than those identified in the preliminary data without HILLIC clean-up in Figure 3-8. A fractionation-based protocol was implemented to reduce the peptide mixture complexity and improve the coverage of the mammalian proteome. The method described in section 2.3.6 was adapted from Wang *et al*.'s protocol. Following fractionation, Wang *et al*. recombined fractionated peptides into 26

fractions to be injected into the MS and identified between 7300 - 8900 protein groups in the various cell lines that were tested.[135] However, in this investigation, fractionated peptides were recombined into 16 fractions for injection. Figure 3-9 displays the coverage summary data from an unfractionated and fractionated dataset.

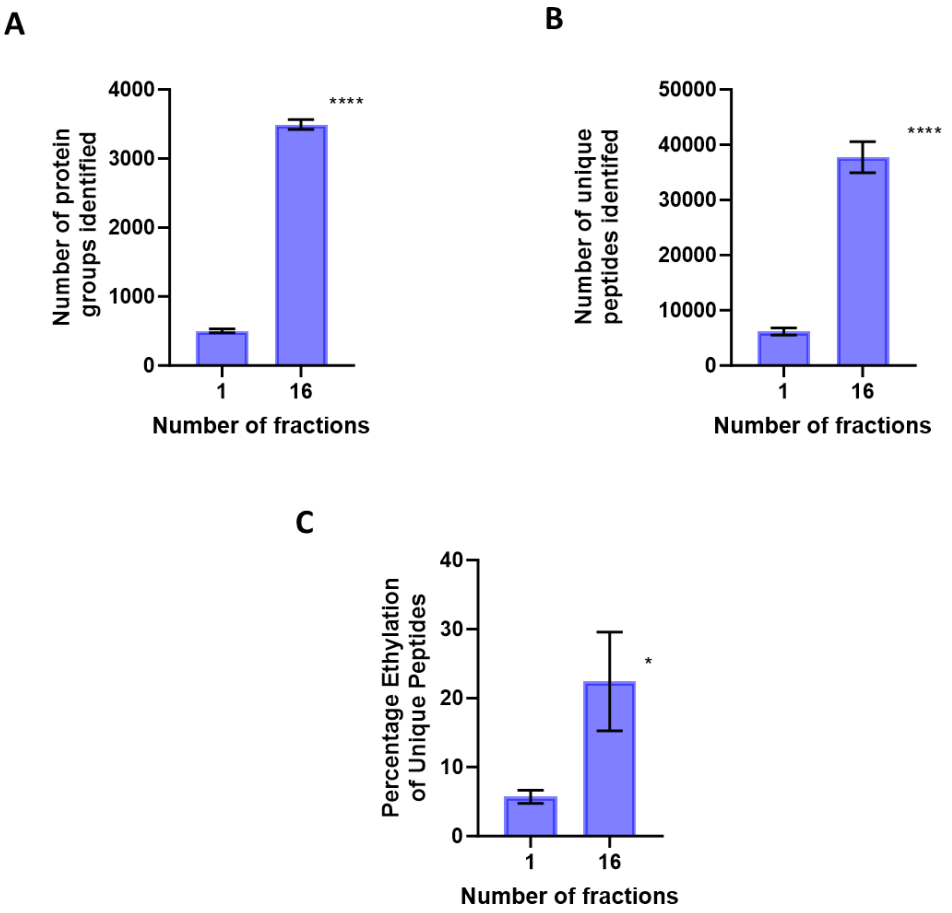


Figure 3-9 The number of (A) protein groups, (B) unique peptides and (C) the percentage of unique ethylated peptides identified in a HEK293 lysate trapped with 12C inorganic carbon versus the number of fractions injected per sample. All data points are represented as the mean where n=3. The error bars displayed are the standard deviation from the mean, and in some cases, these errors are smaller than the individual data points. A one-sample t-test assessed the statistical significance of differences in the coverage metrics between the number of injections per sample. Asterisks indicate levels of significance (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$).

Protein group coverage was increased 7-fold for a trapped sample when each peptide preparation was injected across 16 fractions compared to 1 fraction. The protein group identification rate for the 16 fractionated trapped samples in Figure 3-9 covers 40% of the protein groups identified by the previous HEK293 lysate data.[134] From this result, it was concluded that 16 fractions would be injected for each sample. The UV profile for the C18 reverse phase fractionation process is shown in Figure 3-10.

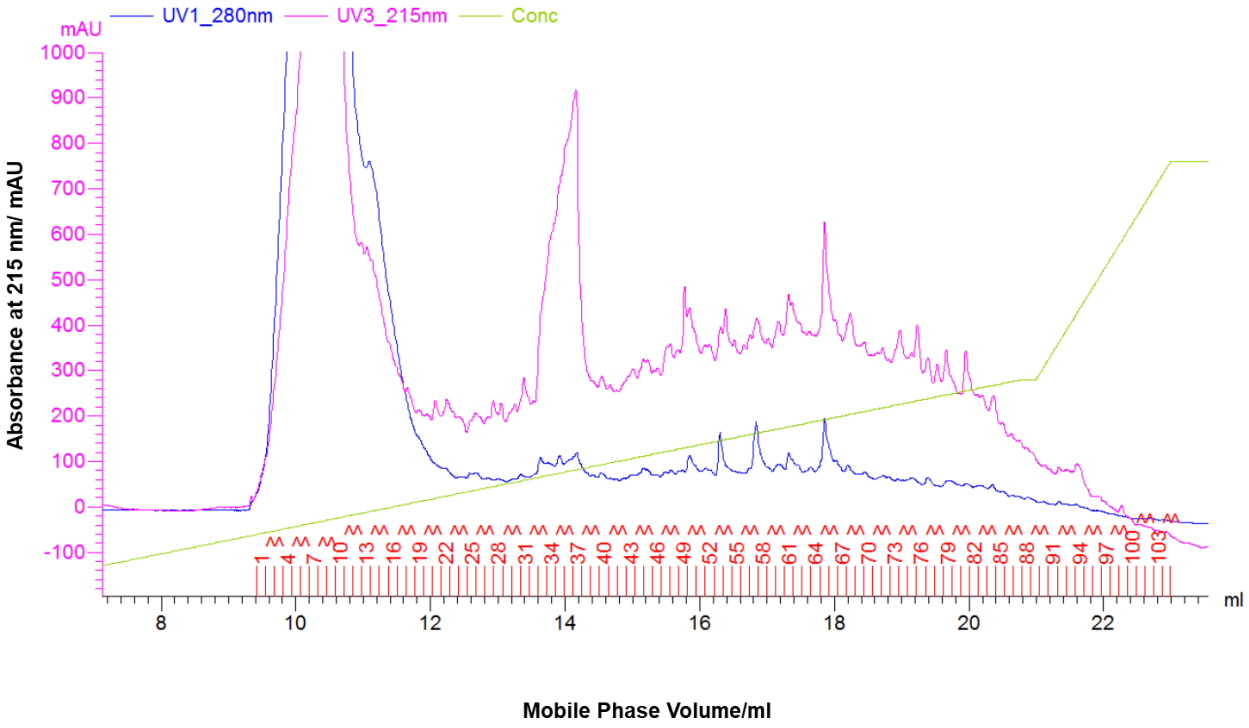


Figure 3-10 The UV chromatogram displaying UV in milli absorbance units (mAU) measured at 215 nm versus the volume of mobile phase run on the C18 column. The pink line is the absorbance at 215 nm, which represents peptides. The blue line is measured at 280 nm, representing proteins. The green line is the percentage of the organic phase in the solution being run over the column where the first step in the gradient is run between 0 - 40% acetonitrile, the second step is 40 - 80% acetonitrile, and the plateau is at 80%. The red numbers along the bottom represent the fractions collected every 130 µl of the mobile phase, which is the bandwidth of a peptide as represented by the pink peaks.

In Figure 3-10, most peptides elute between 0 - 40% acetonitrile in the first gradient step. Two chromatogram areas did not exhibit a typical peptide profile: the first 20 fractions and fractions 33 to 38. It was hypothesised that these contaminants could affect the performance of the mass spectrometer. Therefore, two further optimisation steps were performed. The first was to test the gradient acquisition length for each peptide injection at both 80 and 50 minutes across ten fractions. The second was to run the 16 fractions of reconstituted peptides with (C) and without (NC) the incorporation of the contaminant peaks to determine whether the exclusion of these fractions affected the coverage. The coverage data for these two tests are shown in Figure 3-11.

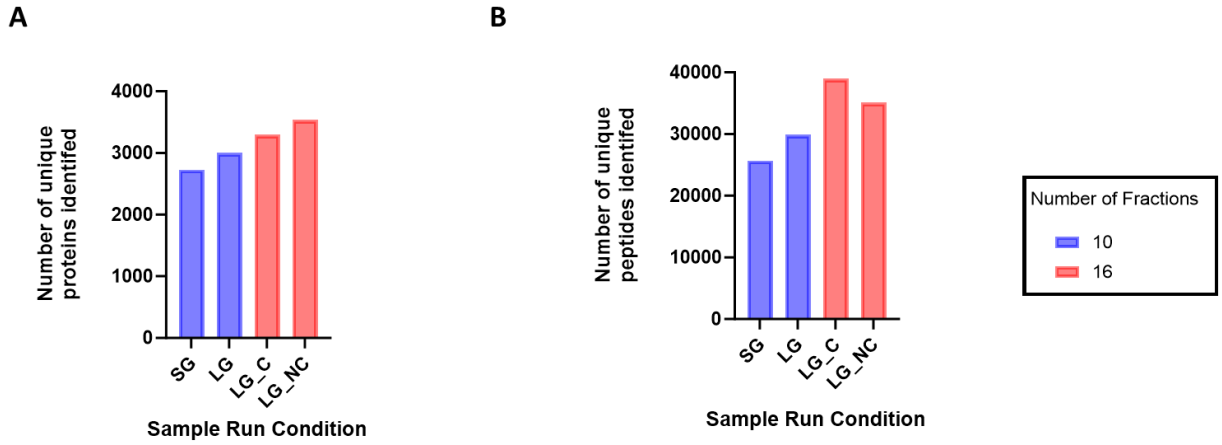


Figure 3-11 The number of (A) protein groups and (B) unique peptides identified in a HEK293 lysate trapped with 12C inorganic carbon versus the mass spectrometer run or sample condition used where n=1. SG and LG stand for a short gradient at 50 minutes and a long gradient at 80 minutes for data acquisition, respectively. The gradient comparison was performed for a sample injected over ten fractions (blue). C and NC denote which type of fractionation recombination was used. As shown in Figure 3-10, the suspected contaminant peak of fractions 33-38 were recombined in the C fractions but not in the NC fractions. The contaminant recombination was performed for a sample injected over 16 fractions (red).

Figure 3-11 was limited to one biological replicate due to the mass spectrometer run time required for this investigation. The run time for each fraction is 100 minutes, therefore, to run one sample the time required is 1600 minutes and the number of samples for the three screening datasets was 23, therefore the total run time for the LCMSMS was 613 hours and 20 minutes. These times exclude the recalibration of the TOF-MS and the blank samples that are run every four fractions. One biological replicate was sufficient to determine the run and sample conditions required for the HEK293 lysate screen. Figure 3-11 shows that a longer acquisition gradient increases the proteome coverage as predicted. Therefore, the data acquisition gradient was set at 80 minutes from this point onwards. Across the four samples in Figure 3-11, the highest number of protein groups identified was for the NC sample injected across 16 fractions. However, the highest number of unique peptides identified was for the C sample injected across 16 fractions. The conclusion drawn from this result was that removing the contaminant fractions from peptide reconstitutions had no drastic effect on the proteome coverage. To preserve the mass spectrometer performance between runs, the peptide fractions for each sample were recombined as NC from this point onwards.

3.7 Coverage for the HEK293 Lysate Screens

The optimised protocol of fractionating and reconstituting peptides into 16 fractions for injection without the contaminant peak (NC) using an acquisition gradient of 80 minutes for each injection was applied to HEK293 lysate screening. Firstly, a screen was completed using 12C Ci during the trapping stage, and later, to address false positive identification, a 13C Ci screen was run to verify the 12C carbamylated hits.

The samples tested for each dataset are detailed in Table 3-1. The 12C lysate screen was performed across two CO_2 partial pressure incubations, whereby the HEK293 cells were incubated at 5% and 10% PCO_2 for eighteen hours before harvesting, whilst the 13C screen was only performed at 5% PCO_2. Figures 3-12 and 3-13 show the coverage data for both proteomic screens. The MSMS data obtained from these screens were interrogated for carboxyethyl hits, and the PSMs modified with carboxyethyl reported by the software were validated, as detailed in section 3.8.

12C dataset, HEK293 lysates extracted from cells incubated at 5% CO_2	12C dataset, HEK293 lysates extracted from cells incubated at 10% CO_2	13C dataset, HEK293 lysates extracted from cells incubated at 5% CO_2
No_TEO_0 mM_1	-	No_TEO_0 mM_1
No_TEO_0 mM_2	-	No_TEO_0 mM_2
-	-	No_TEO_0 mM_3
TEO_20 mM_1	TEO_20 mM_1	TEO_20 mM_1
TEO_20 mM_2	TEO_20 mM_2	TEO_20 mM_2
TEO_20 mM_3	TEO_20 mM_3	TEO_20 mM_3
TEO_50 mM_1	TEO_50 mM_1	TEO_50 mM_1
TEO_50 mM_2	TEO_50 mM_2	TEO_50 mM_2
TEO_50 mM_3	TEO_50 mM_3	TEO_50 mM_3

Table 3-1 Sample descriptions for HEK293 lysate proteomic screening in 12C and 13C inorganic carbon. TEO is triethyloxonium, the trapping reagent and those labelled as no TEO are control samples. The number (0, 20 or 50) relates to the concentration of inorganic carbon in the trapping process. The final number (1, 2 or 3) is the replicate number for the sample. Blank spaces are present when there was no sample with equivalent conditions to the other screens.

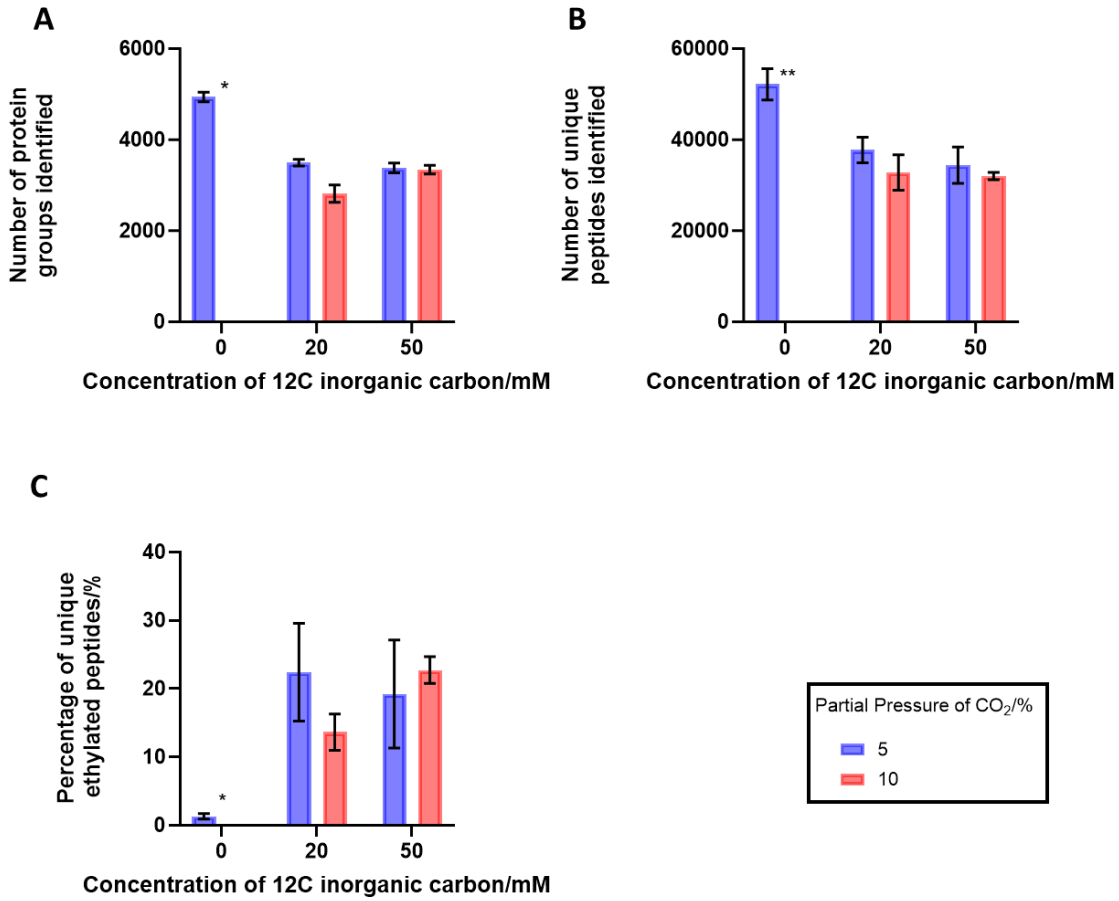


Figure 3-12 The number of (A) protein groups, (B) unique peptides and (C) the percentage of unique ethylated peptides identified in a HEK293 lysate versus concentration of 12C inorganic carbon used during trapping where the 0 mM condition does not contain the trapping reagent. The 12C lysate screen was performed across two CO_2 partial pressure incubations, whereby the HEK293 cells were incubated at 5% (blue) and 10% CO_2 (red) for eighteen hours before harvesting. The error bars displayed are the standard deviation from the mean, and in some cases, these errors are smaller than the individual data points. A one-sample t-test assessed the statistical significance of coverage metrics between trapped and untrapped samples where n=3 and n=2, respectively. Asterisks indicate levels of significance (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$).

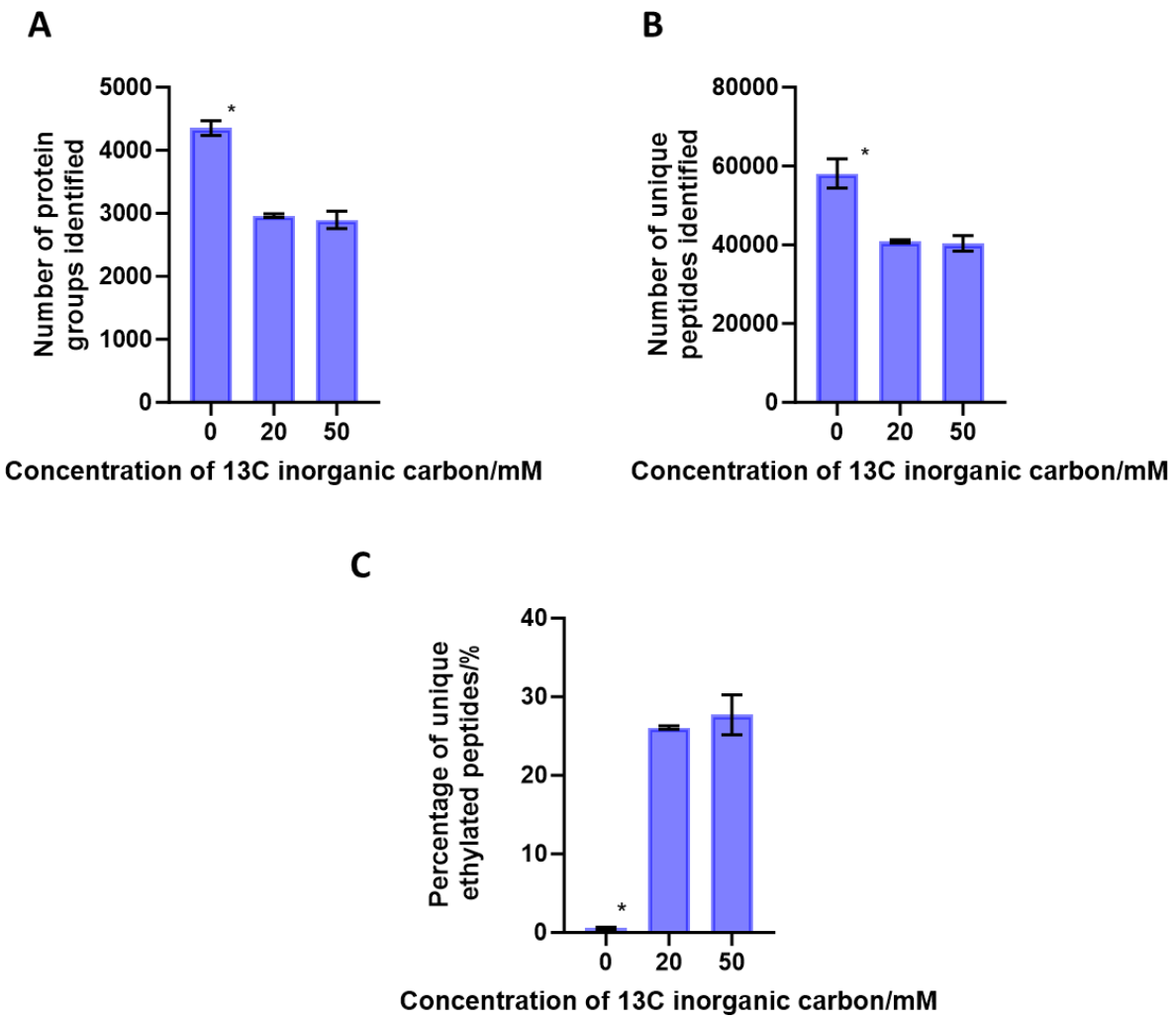


Figure 3-13 The number of (A) protein groups, (B) unique peptides and (C) the percentage of unique ethylated peptides identified in a HEK293 lysate versus the concentration of 13C inorganic carbon used during trapping and the 0 mM condition does not contain the trapping reagent. The 13C lysate samples were harvested from HEK293 cells incubated at a partial pressure of 5% CO_2. The error bars displayed are the standard deviation from the mean, and in some cases, these errors are smaller than the individual data points. A one-sample t-test assessed the statistical significance of coverage metrics between trapped and untrapped samples where n=3. Asterisks indicate levels of significance (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$).

Figures 3-12 and 3-13 show that the trapping process significantly reduces the coverage of the mammalian proteome. The number of unique protein groups and peptides identified in the untrapped samples in the 12C data are 43.6% and 44.6% higher than those identified for these metrics in the trapped samples when comparing the 5% CO_2 incubation datasets. Similarly, the number of unique protein groups and peptides identified in the untrapped samples in the 13C data is 48.5% and 43.0 % higher than those identified for these metrics in the trapped samples. The untrapped samples contain negligible amounts of unique ethylated peptides, while the trapped samples contain 15-25 % of unique ethylated peptides.

3.8 Carbamate Validation

Following data acquisition for the HEK293 lysate screens, the MS data was analysed by database search algorithms. The method used for HEK293 lysate carbamate screening in this investigation enabled the rapid systematic discovery of physiologically relevant carbamate sites despite the harsh conditions of MS. However, limitations of the chosen method included the necessity to manually verify each carboxyethyl-modified peptide spectrum and the identification of false positives. This section (3.8) discusses the verification criteria of hits using four confidence levels, the strategies used to reduce false positives, and the list of validated carboxyethyl hits from 12C and 13C.

3.8.1 Carboxyethyl Confidence Assignment

PTM identification by database search algorithms can be filtered using the assigned confidence level for the modification reported by the software. This metric in PEAKs was called the AScore, and for Protein Pilot, it was called conf. However, these metrics were unsuitable for filtering the carboxyethyl PTM because high confidence levels were reported for carboxyethyl groups on C terminal peptides, which would not be recognisable sites for trypsin digestion. Therefore, this metric was not used for assigning whether a carboxyethyl hit should be regarded as real.

Instead, the spectra that were reported to match carboxyethylated peptides had to be manually verified. A set of validation conditions split across four confidence levels was developed for screening hits, detailed in Table 3-2. To interpret the table, it is important to note the following mass spectra nomenclature. Mass spectra are composed of b and y ions numbered consecutively from the C and N terminus of the peptide, respectively.[136] Spectra, which correspond to peptides modified by PTMs, exhibit a mass shift on the target amino acid. The b and y ions in a modified (mod) peptide follow the nomenclature of y/b mod for the modified amino acid and y/b mod +1 or -1, depending on whether the mass relates to the amino acid after or before the modification. Neutral loss ions are obtained from alternative fragmentation events during CID [137] and the neutral loss ions searched for by PEAKs consist of b and y ions without water or ammonia. Doubly charged ions are generated during the ionisation

phase, whereby the ejected ion generated contains a secondary proton, resulting in a +2 charge.[138] Example spectra for each confidence level are displayed in the supplementary information in Figures 8-1, A to F.

Confidence Level	Mass Spectrum Features
High	- y or b modification (y or b mod) support - Good y ion coverage before and after with good intensity- designated a string of y ions. - At least one doubly charged or neutral loss supporting ion after the modification. - b ion support on peptide even if not in the modification vicinity.
Medium	- No doubly charged or neutral loss ion support but meets other criteria specified by high confidence. - No y or b mod support but a string of y ions and some doubly charged or neutral loss ion support, particularly after the modification. - y coverage is mainly before the modification and limited after, for example, only mod and mod+1 after, but meets other criteria specified by high confidence.
Low	- no y or b (mod) support - N terminal with good b ion coverage. - Support only before mod for y and b - Low-intensity spectra - no string of b or y ions
No	- The peptide sequence does not end at a trypsin cut site. For example, a C terminal carboxyethyl-modified lysine is

	not a possible at the C terminus, as trypsin would not recognise this site. - No supporting ions surround the modification.

Table 3-2 Carboxyethyl validation conditions

3.8.2 Challenges in Analysing Data

Initial data analysis of the 12C HEK293 lysate screen using the PEAKs search algorithm and the carboxyethyl validation conditions outlined in Table 3-2 identified a clear issue of false positive carboxyethyl hits. The false positive identification rate of carboxyethyl groups was ~30% in untrapped samples compared to trapped samples. Previous studies had been met with this challenge, for example, Jones *et al*. [139] and the future direction for data analysis was considered with this study in mind. Two strategies were implemented to reduce false positive identification. The first was using a different database search algorithm called Protein Pilot, which implemented distinct raw data conversion and peptide identification steps, as detailed in section 3.5, compared to PEAKS. The second was to interrogate the data for carboxyethyl identification on a decoy amino acid.

Protein Pilot was used to interrogate the 12C dataset for the presence of carboxyethylated residues using the target decoy-based approach.[139] Arginine was selected as the decoy amino acid modified by carboxyethyl because it is also a trypsin cut site. The software did not identify any carboxyethyl modifications on arginine, leading to the conclusion that the modification is a genuine carboxyethyl modification.

Following this result, the 12C dataset run in Protein Pilot was analysed for carboxyethyl lysine hits. Protein Pilot identified significantly fewer false positive hits when comparing untrapped samples to trapped samples compared to PEAKs, with only a false positive identification rate of 3.2%. Therefore, the settings in PEAKs were compared to those used in Protein Pilot. The precursor ion error tolerance defined by Protein Pilot for the TripleTof mass spectrometer was found to be set at 0.05 Daltons (Da). In PEAKs, the error tolerance was defined as parts per million (ppm) instead of being defined as a mass

in Protein Pilot. In contrast to the absolute mass error defined in Protein Pilot, the ppm value applies a consistent relative error across the dataset, which depends on each mass reading. The literature states that the maximum error tolerance should be set at 30 ppm for the TripleTof mass spectrometer.[140] Therefore, the 12C data was reprocessed using PEAKs and the error tolerance set for the precursor ion fragment was reduced from 50 to 20 parts ppm. This more stringent precursor error tolerance reduced the false positive identification rate using PEAKs to 3.5%, comparable to the Protein Pilot result.

A list of high and medium confidence carboxyethyl hits identified in the same sample in both PEAKs and Protein Pilot was produced, and any hits at these confidence levels that were placed in more than one sample were regarded here as confident hits in the 12C HEK293 lysate screen. Any reproducible false positives identified in the untrapped samples of the 12C dataset were regarded as the methylglyoxal-derived AGE modification, which is also associated with a mass shift of 72.02 Da on modified lysines, as detailed in section 3.3.

Finally, to further validate carboxyethyl identifications, a 13C HEK293 lysate screen was run to separate the identification of AGEs from carbamate sites on proteins. A key finding from this screen was that the false positive carboxyethyl hits seen with 12C were only seen in the 13C screening when searching with 72.02 Da but not with the 73.02 Da modification. This confirms that the false positive hits are the methylglyoxal-derived AGE modification. The validated carboxyethyl modifications and false positives identified in the 12C and 13C lysate screens are presented in sections 3.8.3 and 3.8.4.

3.8.3 Summary of Identified Carbamate Hits.

Tables 3-3 - 3-5 display the valid carboxyethyl hits identified from the 12C HEK293 lysate screen using PEAKs and Protein Pilot. Table 3-3 shows the hits identified multiple times with the same sample ID across the 12C lysate screen by both search algorithms. Table 3-4 shows the hits identified multiple times across the 12C lysate screen but were not found in the same sample or only found once in the same sample by both search algorithms. Table 3-5 shows the hits that were identified multiple times but only by one of the search algorithms. The spectra for the carbamates listed in Table 3-3 and 3-4 are plotted in Figure 3-14 and the supplementary Figure 8-2, respectively. The carbamates listed in Table 3-5 are not plotted due to not being identified by both software.

Protein Accession	Protein Name	Modification Site	Number of times the site was identified in PEAKs	Number of times the site was identified in Protein Pilot	Number of Samples in which site is ID in both
P15531/ P22392	Nucleoside diphosphate kinase A/B	12	9	10	7
P06748	Nucleophosmin	267	5	7	5
P16403/P16402/P10412	Histone H1.2/1.3/1.4	63	6	6	4
P68431/Q71DI3/ P84243	Histone H3.1/3.2/3.3	79	3	3	3
P06733	Alpha-enolase	80	2	2	2
P62805	Histone H4	32	3	3	3
P08865	Small ribosomal subunit protein	57	3	5	3

Table 3-3 The carboxyethyl hits identified in multiple samples with the same sample ID across the 12C lysate screen were analysed using both database search algorithms. The protein accession in column one is the unique identifier in the UniProt database for the sequenced protein which is named in column two. The modification site in column 3 relates to the carboxyethyl-modified lysine in the protein sequence. The number of times Peaks and Protein Pilot identified the site is reported. The final

column shows the number of times the carboxyethyl hit was seen in the same sample across both search algorithms.

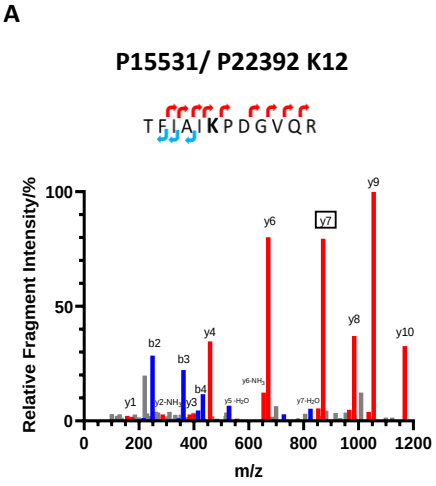
Protein Accession	Protein Name	Modification Site	Number of times the site was identified in PEAKs	Number of times the site was identified in Protein Pilot	Number of Samples in which site is ID in both
P04406	Glyceraldehyde-3-phosphate dehydrogenase - GAPDH	194	1	4	1
P16403/P16402/P10412	Histone H1.2/1.3/1.4	106	3	1	1
P06733	Alpha-enolase	233	3	1	1
P16403/P16402/P10412	Histone H1.2/1.3/1.4	85	2	2	1
P63261	Actin Cytoplasmic	61	1	3	1
P04406	Glyceraldehyde-3-phosphate dehydrogenase - GAPDH	263	1	1	1
Q9NYF8	Bcl-2-associated transcription factor 1	164	1	1	1
P68104	Elongation factor 1-alpha 1	457	1	4	0
P68431/Q71DI3/ P84243	Histone H3.1/3.2/3.3	123	2	1	0

Table 3-4 The carboxyethyl hits that were identified multiple times across the 12C lysate screen when analysed using both database search algorithms. The columns are described in Table 3-3. However, the carboxyethyl hits reported here were only identified once or not in the same sample across both search algorithms, as shown by the fifth column.

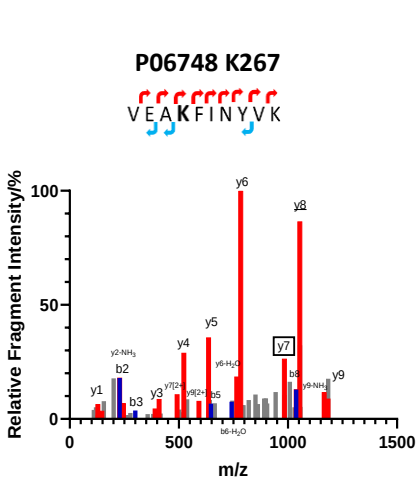
Protein Accession	Protein Name	Modification Site	Number of times identified	Analysis Type which detected the CE site
P62805	Histone H4	92	7	Protein Pilot
O94906	Pre-mRNA-processing factor 6.	299	4	Protein Pilot
Q32Q12	Nucleoside diphosphate kinase	37	3	PEAKs
O60814	Histone H2B type 1-K	109	2	Protein Pilot
Q92576	PHD finger protein 3.	323	2	Protein Pilot
Q09666	Neuroblast differentiation-associated protein AHNAK	976	2	Protein Pilot

Table 3-5 The carboxyethyl hits identified multiple times across the 12C lysate screen that were only identified by one of the database search algorithms. The first three columns are described in Table 3-3, and the database search algorithm which identified the hit is detailed in column.

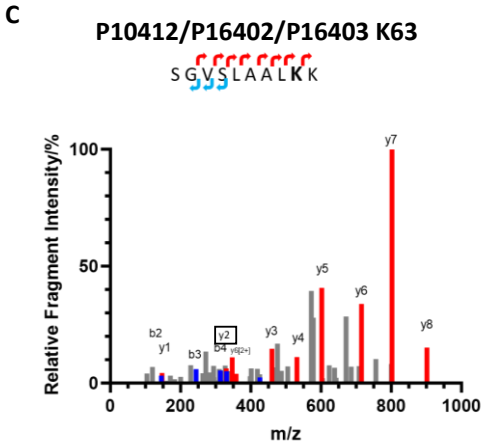
A


P15531/ P22392 K12
TFIAIKPDGVQR
Relative Fragment Intensity/%
m/z

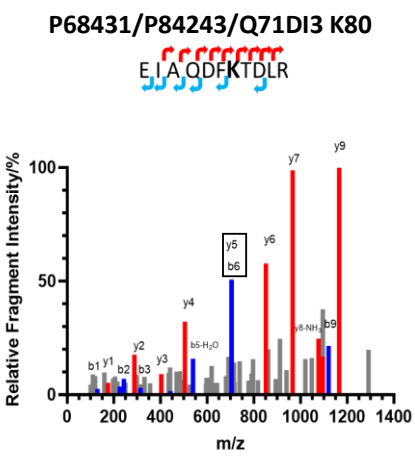
B


P06748 K267
VEAKFINYVK
Relative Fragment Intensity/%
m/z

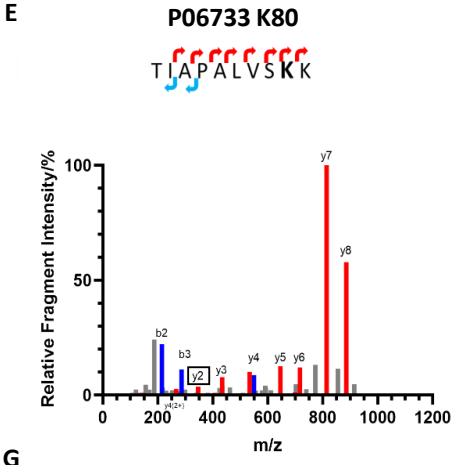
C


P10412/P16402/P16403 K63
SGVSLAALKK
Relative Fragment Intensity/%
m/z

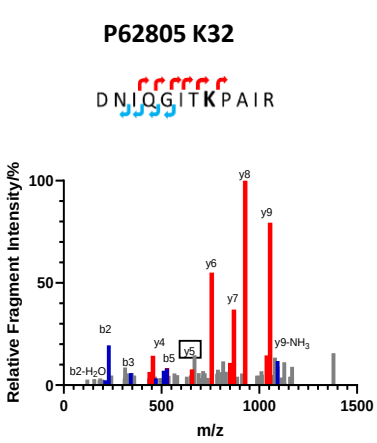
D


P68431/P84243/Q71DI3 K80
EIAQDFKTDLR
Relative Fragment Intensity/%
m/z

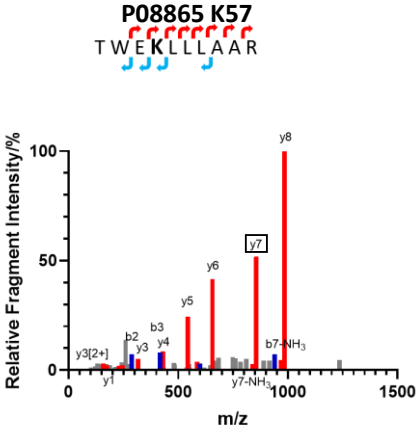
E


P06733 K80
TIAPALVSKK
Relative Fragment Intensity/%
m/z

F


P62805 K32
DNIQGITKPAIR
Relative Fragment Intensity/%
m/z

G


P08865 K57
TWEKLLLAAR
Relative Fragment Intensity/%
m/z

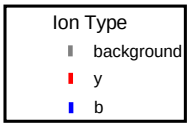

Ion Type
background
y
b

Figure 3-14 Identification of carbamate hits from the 12C HEK293 lysate screening that were identified multiple times by both database search algorithms in the same sample and are listed in Table 3-3. Plots of relative fragment intensity versus m/z from LCMSMS identifying trapped carbamates on (A) Nucleoside diphosphate kinase A/B, P15531/ P22392 K12, (B) Nucleophosmin, P06748 K267, (C) Histone H1.2/1.3/1.4, P10412/P16402/P16403 K63, (D) Histone H3.1/3.2/3.3, P68431/P84243/Q71DI3 K80, (E) Alpha-enolase, P06733 K80, (F) Histone H4, P62805 K32 and (G) Small ribosomal subunit protein, P08865 K57 in the presence of $12CO_2$. Each spectrum has a peptide sequence identifying predominant y (red) and b (blue) ions. The grey peaks represent background ions, and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

To support the 12C carboxyethyl validations, a 13C screen was completed to verify carboxyethyl hits further. Only PEAKs software was used to analyse the 13C screen due to time and resource constraints. Tables 3-6 – 3-8 display the valid carbamate hits identified from the 13C HEK293 lysate screen. Table 3-6 displays the hits identified in the 13C screen that were identified in both search algorithms with shared sample ID in the 12C screen. Table 3-7 shows the hits identified in the 13C screen that were identified by one of the search algorithms or did not share the same sample ID when found by both algorithms in the 12C screen. Table 3-8 displays the hits identified multiple times or that are of interest in this investigation in the 13C screen but not in the 12C screen. The spectra for the carbamates listed in Tables 3-6, 3-7 and 3-8 are plotted in the supplementary Figure 8-3 and Figures 3-15 and 3-16, respectively.

Protein Accession	Protein Name	Modification Site	Number of times ID in 12C PEAKs or Protein Pilot	Number of times ID in 13C
P62805	Histone H4	92	7	3
Q32Q12	Nucleoside diphosphate kinase (NDK)	37	3	5
P68431/Q71DI3 /P84243	Histone H3.1/3.2/3.3	123	2	2
P16403/P16402/P10412	Histone H1.2/1.3/1.4	46	1	1

Table 3-6 Carboxyethyl hits identified in the 12C dataset by both database search algorithms with shared sample IDs also identified in the 13C dataset. Column one lists the protein accession and the unique identifier given in the Uniprot database for the sequenced protein which is named in column two. The modification site in column three relates to the carboxyethyl-modified lysine in the protein sequence. The number of times the site was identified in the 12C screen by both search algorithms with a shared sample ID is listed in column four. The final column shows the number of times the carboxyethyl hit was seen in the 13C screen.

Protein Accession	Protein Name	Modification Site	Number of times identified in the same sample across both analysis methods in the 12C screen.	Number of times identified in the 13C screen
P06748	Nucleophosmin	267	5	4
P62805	Histone H4	32	3	5
P16403/P16402/P10412	Histone H1.2/1.3/1.4	85	1	3
P68431/Q71DI3 /P84243	Histone H3.1/3.2/3.3	79	3	1
P16403/P16402/P10412	Histone H1.2/1.3/1.4	106	1	2

Table 3-7 Carboxyethyl hits identified in the 12C dataset by only one of the database search algorithms or did not share sample ID when found by both database search algorithms also identified in the 13C dataset. Columns are as described in Table 3-6 except for column four which is the number of times the hit was seen in either search algorithm in the 12C data screen.

Protein Accession	Protein Name	Modification Site	Number of times ID in 13C
P62987	Ubiquitin-ribosomal protein eL40 fusion protein	48	3
P31948	Stress-induced-phosphoprotein 1	229	2
P68104/ Q05639	Elongation factor 1-alpha 1	255	2
F5H6Q2	Ubiquitin C	48	1
P68431/Q71DI3 /P84243	Histone H3.1/3.2/3.3	57	1

Table 3-8 Carboxyethyl hits identified multiple times or are of interest in this investigation in the 13C HEK293 lysate screen. Equivalent columns are described in Table 3-6.

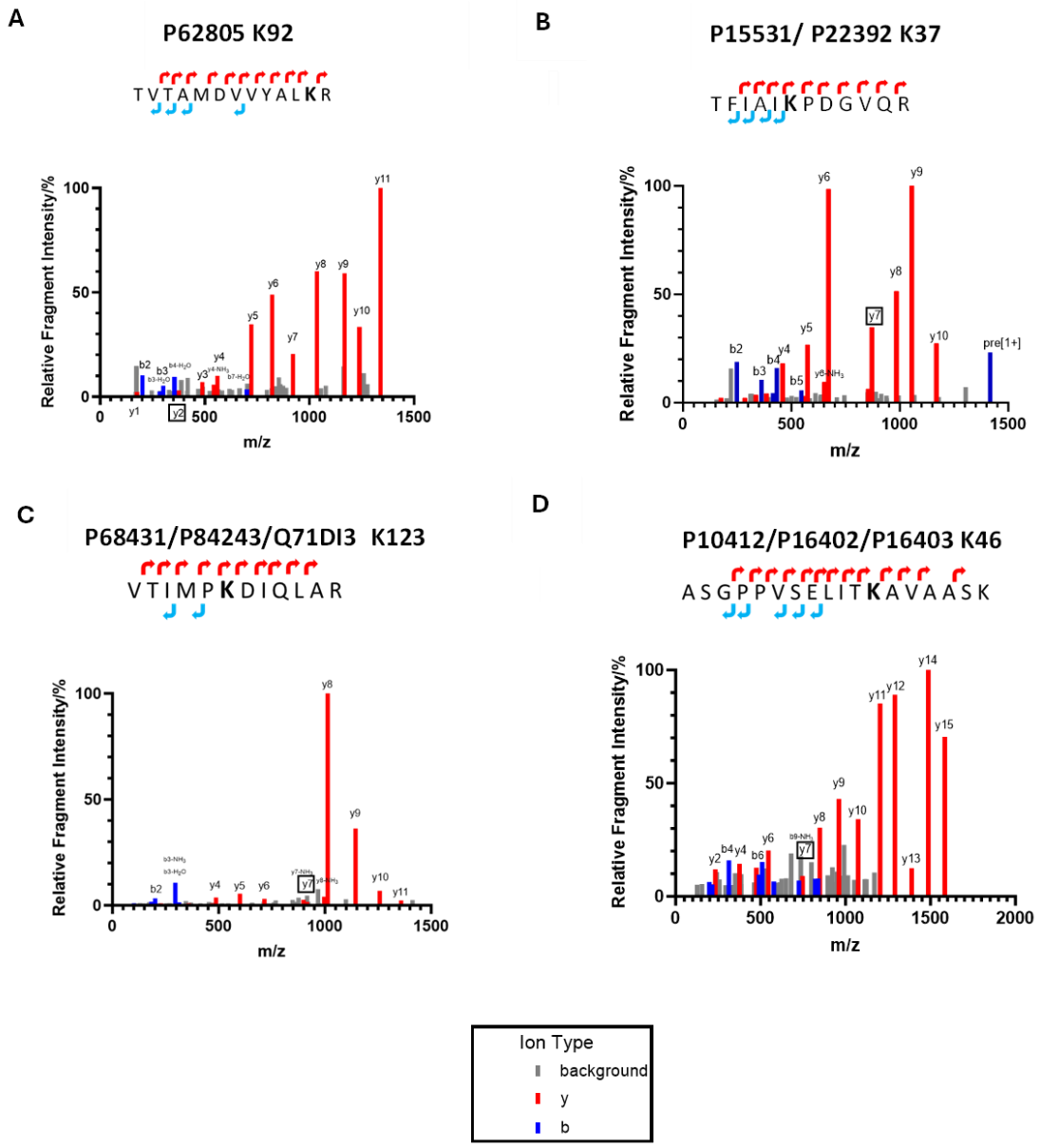


Figure 3-15 Identification of carbamate hits from the 13C HEK293 lysate screening identified in the 12C dataset by only one of the database search algorithms or did not share sample ID when found by both algorithms and are listed in Table 3-7. Plots of relative fragment intensity versus m/z from LCMSMS identifying trapped carbamates on (A) Histone H4 P62805 K92, (B) Nucleoside diphosphate kinase A/B P15531/ P22392 K37, (C) Histone H3.1/3.2/3, P68431/P84243/Q71DI3 K123, (D) Histone H1.2/1.3/1.4, P10412/P16402/P16403 K46 in the presence of $^{13}CO_2$. Each spectrum has a peptide sequence identifying predominant y (red) and b (blue) ions. The grey peaks represent background ions, and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

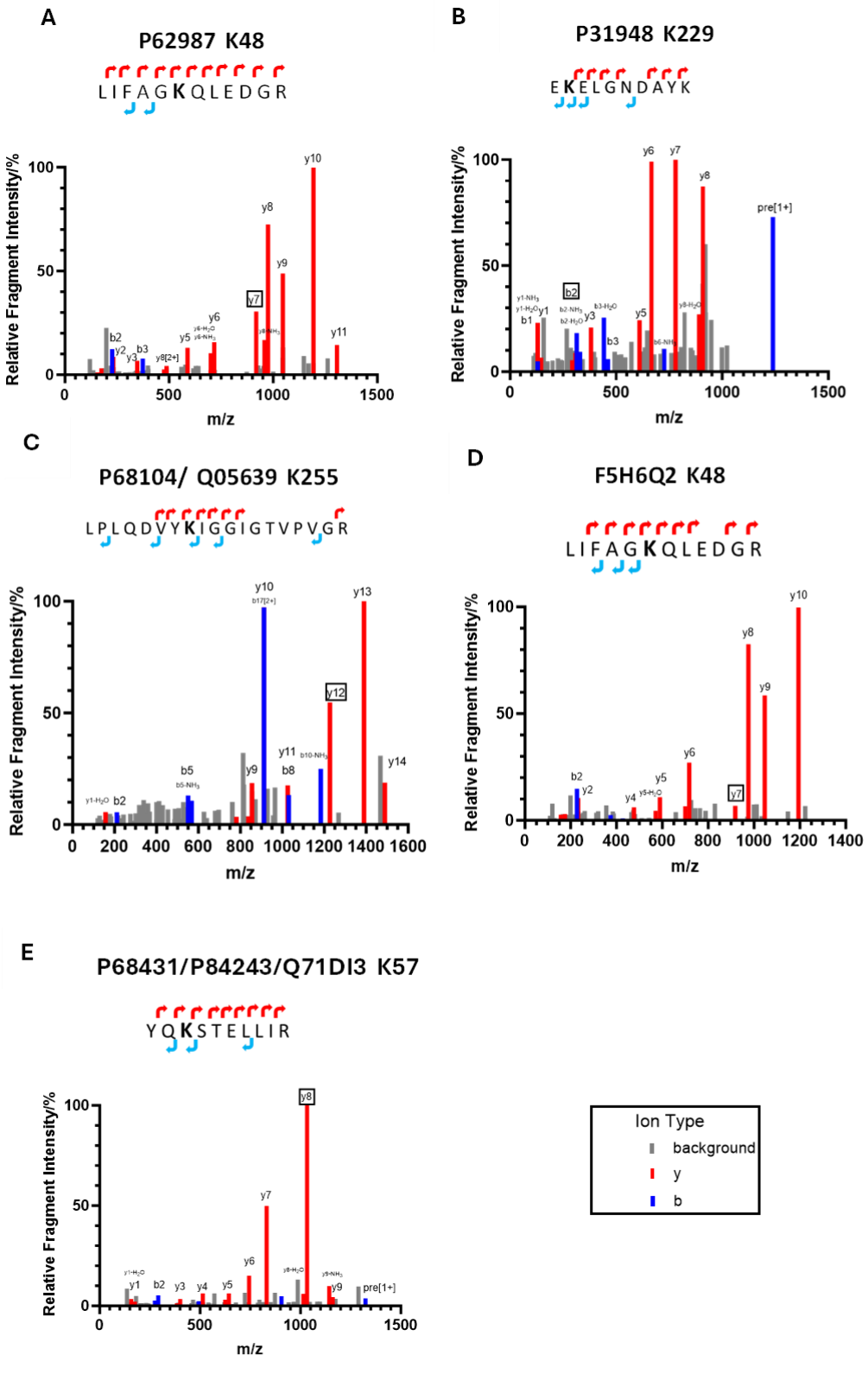


Figure 3-16 Identification of carbamate hits from the 13C HEK293 lysate screening that were identified multiple times or are of interest in this investigation but not found in the 12C dataset and are listed in Table 3-8. Plots of relative fragment intensity versus m/z from LCMSMS identifying trapped

carbamates on (A) Ubiquitin- ribosomal protein eL40 fusion protein P62987 K48, (B) Stress-induced-phosphoprotein 1, P31948 K229, (C) Elongation factor 1-alpha 1, P68104/ Q05639 K255, (D) Ubiquitin C, F5H6Q2 K48, (E) Histone H3.1/3.2/3.2, P68431/P84243/Q71DI3 K57 in the presence of $^{13}CO_2$. Each spectrum has a peptide sequence identifying predominant y (red) and b (blue) ions. The grey peaks represent background ions, and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

3.8.4 False Positives

The false positives identified in the 12C dataset are presented in Table 3-9. Table 3-9 shows that only four false positives had to be removed from the trapped datasets. The other false positives reported were not identified in trapped samples.

Protein Accession	Protein Name	Modification Site	Number of times the false positive was detected in trapped samples across both 12C analysis methods (sample is only counted once)	The number of times the false positive was detected in untrapped samples across both 12C analysis methods, where each sample is only counted once.	Analysis Type which detected the CE site
P16152	Carbonyl Reductase [NADPH] 1	239	10	2	Both
O60814/ P62807	Histone H2B type 1-K/C/E/F/G/I	6	3	1	Both
P05062	Fructose-bisphosphate aldolase B	147	0	2	Both
Q96EP5	Daz-associated protein 1.	57	0	1	PEAKs
Q02878	Large ribosomal subunit protein eL6	237	0	1	PEAKs

P23588	Eukaryotic translation initiation factor 4B	578	0	1	PEAKs
P24844	Myosin regulatory light polypeptide 9	1724	0	1	PEAKs
Q9H307	Pinin	676	8	2	Protein Pilot
P22102	Trifunctional purine biosynthetic protein adenosine-3	977	1	1	Protein Pilot
P62807	Histone H2B type 1-C/E/F/G/I	12	0	2	Protein Pilot
Q9NU22	Midasin	14	0	1	Protein Pilot
P10809	60 kDa heat shock protein, mitochondrial	87	0	1	Protein Pilot
P26038	Moesin	151	0	1	Protein Pilot
P31948	Stress-induced-phosphoprotein 1	337	0	1	Protein Pilot
Q13618	Cullin-3	646	0	1	Protein Pilot
Q9Y2H0	Disks large-associated protein 4	759	0	1	Protein Pilot
Q13472	DNA topoisomerase 3-alpha	910	0	1	Protein Pilot
Q86U86	Protein polybromo-1	1425	0	1	Protein Pilot

Table 3-9 False positives were identified across the two untrapped samples and the number of times these false positives were identified in the trapped samples from the 12C HEK293 lysate screen using two database search algorithms.

To assess the presence of a false positive in the 13C lysate screen, it is important to reiterate that no false positives were identified in the untrapped samples with a mass shift corresponding to 73.02 Da. The false positives identified in the trapped and untrapped samples from the 13C screen corresponded to a mass shift of 72.02 Da. Table 3-10 presents these false positives, and from this data, it was concluded that the P16152 and O60814/ P62807 were modified by CML on sites 239 and 6, respectively.

Protein Accession	Protein Name	Modification Site	The number of times the false positive was detected in untrapped or 12C searches for samples in the 13C dataset.	The number of times the false positive was detected in untrapped or 12C searches for samples in the 13C across both lysate screens.
P16152	Carbonyl reductase [NADPH] 1	239	4	6
O60814/ P62807	Histone H2B type 1-K/C/E/F/G/I	6	2	3
Q9Y5B9	FACT complex subunit SPT16	79	1	1
P13667	Protein disulfide-isomerase A4	218	1	1
P38646	Stress-70 protein, mitochondrial	300	1	1

Table 3-10 False positives were identified across the 13C samples, which correspond to a 72.02 Da mass shift and the number of times these false positives were identified as 12C carboxyethyl modifications in the trapped samples across both datasets. None of these hits were identified in the 13C carboxyethyl search.

The spectra for the two verified methylglyoxal-derived AGE hits in this screen are given in Figure 3-17. The other false positives were either only identified once or not identified across any of the trapped samples in the lysate screens.

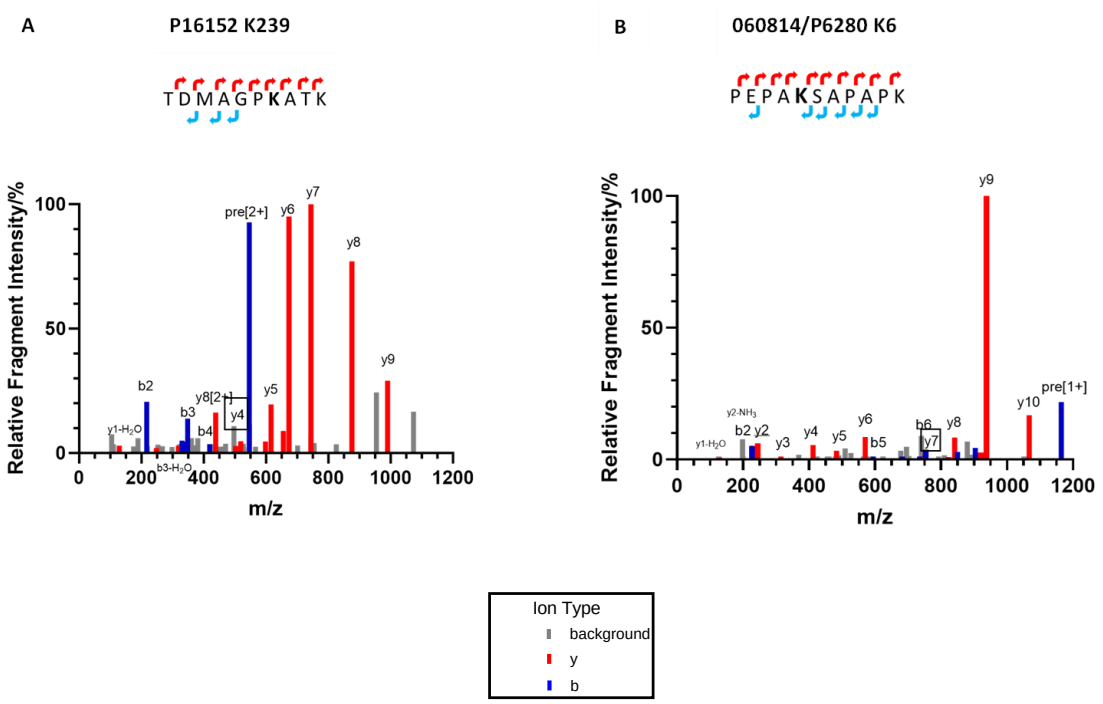


Figure 3-17 False positive identification of the 72.02 Da shift in samples from the 12C and 13C HEK293 lysate screens that were identified more than once in both datasets. Plots of relative fragment intensity versus m/z from LCMSMS identifying the AGE, carboxymethyl-lysine on (A) Carbonyl reductase [NADPH] 1, P16152 K239, and (B) Histone H2B type 1-K/C/E/F/G/I, O60814/P62807 K6 in the presence and absence of the trapping reagent and atmospheric levels of $^{12}CO_2$. Each spectrum has a peptide sequence identifying predominant y (red) and b (blue) ions. The grey peaks represent background ions, and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

3.9 Discussion

The optimised workflow used for sample preparation produced untrapped samples in the 12C and 13C datasets with a mean protein group coverage of 58% and 51%, respectively, compared to the reported protein group number identified in the published HEK293 mammalian proteome screen.[134] For the trapped samples, these values dropped to 38% and 34.5% for the 12C and 13C datasets, respectively. The reduction in coverage for trapped samples can be explained by the increased hydrophobicity of peptides modified by TEO. The detection of highly hydrophobic proteins, for example, membrane-bound proteins, has been recognised as a limitation of MS proteome screening.[141] The difference in coverage between the 12C and 13C datasets could reflect the mass spectrometer performance during data retrieval. In summary, the coverage achieved here was deemed sufficient due to the MS resource demand of the investigation.

The processing pipeline used in LCMSMS-based PTM discovery workflows is integral to the correct assignment of PTM hits, as discussed in the context of false positive identification in section 3.8.2. Initially, false positive identification of carbamates by PEAKs was ~30%, and it was hypothesised that the data file conversion step could be causing this issue. Therefore, a second database search algorithm was used for assigning carbamate hits. Protein Pilot was selected due to being provided by SCIEX and being suitable for processing raw wiff data produced from the Q-TOF. Protein Pilot directly converts the raw files before database searching. The MS data was also processed using a decoy amino acid-based strategy. In this strategy, the database search algorithm interrogated the MS data for the presence of the carboxyethyl modification on arginine. It was found that Protein Pilot could only identify carboxyethyl at lysine or N-terminal residues. The conclusion reached from these tests was that both PEAKs and Protein Pilot can identify real carboxyethyl hits when searched at a stringent precursor mass error tolerance that is reflective of the mass spectrometer.

After the 12C dataset had been properly processed using both of the database search algorithms, carboxyethyl hits were manually assigned a confidence level, as detailed in Table 3-2. The

confidence level assignments were developed by looking at numerous peptide spectrum matches to reduce subjectivity when assigning carbamate hits. It has been stated previously in the literature that y ions are more abundant than b ions due to the former being more stable.[142] This was evident in the HEK293 12C and 13C lysate screen analyses and explains why the confidence level assignments focus more on y ion detection than b. In addition, N terminal carbamates were not assigned to a confidence level higher than low confidence using the validation criteria. This condition contrasts with the fact that N-terminal carbamates exist, such as valine on haemoglobin. Still, the MS spectra produced here did not give enough evidence to support the validation of a carbamate modification at the protein N terminus.

Following correct processing and PTM validation, the 72.02 Da mass shift was identified in untrapped samples with a false positive identification rate in PEAKs and Protein Pilot of 3.5 and 3.2%, respectively. Literature searching of PTMs proved the presence of a rare CML modification formed due to protein glycation, which is also associated with a mass shift of 72.02 Da, as discussed in section 3.3. Following this realisation, a 13C screening was implemented to separate the identification of real carboxyethyl hits from the CML modification.

The HEK293 lysate carbamate screening identified 23 and 12 high-confidence reproducible carbamate-modified protein sites from the 12C and 13C datasets, respectively. From these sites, nine were identified in both the 12C and 13C screens.

The highest confidence hits in this screening are those identified in both the 12C and 13C screens displayed in Tables 3-6 – 3-7. Histones were identified multiple times as carbamylation targets across different lysine sites. A range of PTMs modify histone proteins and have been identified as integral to DNA accessibility and transcription. The role of carbamylation on DNA transcription was investigated further in Chapter 5. Nucleophosmin is an important cellular protein with a range of functionalities in metabolic pathways. These pathways include chromatin assembly, DNA repair and apoptosis. [143] Interestingly, nucleophosmin has been linked to the tumour suppressor activity of p53[143]

which is involved in crosstalk with the NFkB signalling pathway [144,145] which relies on CO_2-sensitive transcription factors.[33] NDK is an enzyme which catalyses the reversible transfer of phosphate onto nuclear diphosphates to synthesize nuclear triphosphates (NTPs). NDK activity is tightly regulated to maintain the homeostasis of cellular NTP pools.[146] In addition, NDK is a pleiotropic effector involved in a diverse range of biological processes, including gene transcription, DNA damage, protein phosphorylation and more.[147,148] Therefore carbamylation on NDK may be involved in the CO_2-dependent transcriptional response. The 12C dataset identified a higher number of hits compared to the 13C dataset. However, this is because two MS database search algorithms were used to analyse the 12C dataset. When the PEAKs and Protein Pilot data are considered individually, 12 and 17 reproducible carbamate hits are identified, respectively.

The false positives identified by the 12C dataset are listed in Table 3-9. The false positives were reported to ensure accuracy in reporting real carboxyethyl hits, even if they were only seen once. However, all reported trapped carboxyethyl hits in the 12C and 13C datasets had to be identified at least twice. The only exception to this rule was for the last two trapped carbamate hits reported in Table 3-8 due to being of broader interest in this investigation. The number of false positives reported by Protein Pilot was higher than that reported by PEAKs, as shown in Table 3-9. In total, four of the hits seen in the untrapped data were removed from the 12C trapped data. Two of these hits were seen in both search algorithms, while the others were only seen in Protein Pilot. This result contrasts with Protein Pilot's lower false positive identification rate, which was reported as 0.3% lower than PEAKs. The false positive identification rate was calculated using the number of unique hits in untrapped samples as a percentage of the unique carbamate hits in trapped samples. Out of the two 12C analysis pipelines, Protein Pilot identified more hits overall, and the carboxyethyl hits identified were more likely to appear only once per sample compared to PEAKs. Therefore, the exact number of false positive hits is a more useful metric to determine which software is more error-prone, which is Protein Pilot. Due to the time-consuming nature of manual spectra validation and the identification of fewer false positives using PEAKs than Protein Pilot, the 13C lysate screen was only analysed using PEAKs.

From the 13C dataset, no false positive hits were assigned a high or medium confidence level for the 13C-associated carboxyethyl modification, corresponding to a mass shift of 73.02 Da; false positives were only assigned for a mass shift of 72.02 Da. The presence of only 72.02 Da false positives specific to lysine led to the conclusion that any false positive hits were the CML AGE. From both data screens, only two CML sites were confidently identified, including carbonyl reductase K239, identified previously [105] and Histone H2B K6.

In summary, the approaches taken here to optimise the HEK293 lysate screening during the peptide preparation and database search stages were suitable for identifying and verifying the presence of carboxyethyl modifications. The reported carboxyethyl hits are reproducible and, in some cases, have been identified by the 12C and 13C lysate trapping screens.

3.10 Conclusion

In conclusion, an MS-based workflow for carbamate discovery has been successfully implemented on a mammalian lysate extracted from the CO_2-sensitive cell line HEK293. The screening results are split into three main sections, as detailed in sections 3.6-3.8. The preliminary experiments in 3.6 optimised the sample preparation workflow to improve proteome coverage. Following this, lysate screening was conducted firstly in a 12C dataset trapped with 12C Ci and later in a 13C dataset trapped with 13C Ci. Section 3.7 details the coverage obtained for trapped and untrapped samples in both datasets. The final stage, described in section 3.8, discusses the steps to optimise the bioinformatic pipeline, reduce false positive hit identification and report a list of reproducible carbamate hits.

The 12C and 13C proteome screening datasets identified nine novel high-confidence carbamates across the proteome. However, the three mammalian carbamates already stated in the literature were not identified except for ubiquitin K48, which was only found once in the 13C dataset. Out of the nine carbamates identified in both screens, seven of these hits are identified on histone proteins, which are of biological significance to DNA transcription.

The analysis of the 12C screen showed that PEAKs was less error-prone than Protein Pilot and should be designated as the database search algorithm for future carbamate discovery experiments. The data presented also gives a strong argument for verifying 12C carboxyethyl hits with 13C Ci to separate the presence of a carbamate from the AGE carboxymethyl modification.

Before this investigation, the trapping methodology had been utilised on purified proteins and applied to proteome screens using Arabidopsis.[1] The HEK293 lysate screening performed in this chapter has highlighted the limitations of trapping on the proteomic scale. In particular, TEO ethylates at aspartic acid, glutamic acid, and lysine and is highly reactive in solution with a half-life of 7 minutes. In a lysate screen, many possible TEO modification sites exist; therefore, carbamate sites will be missed due to not being irreversibly modified during the trapping stage. The bioinformatic analysis is resource-

heavy due to manual carbamate validation, and the detection sensitivity of identifying carbamate PTMs, particularly at the N terminus of proteins, is limited without using a higher-grade mass spectrometer or a PTM affinity-based workflow.[149]

3.11 Future Work

The carbamylation hits identified by this screening are useful avenues for future work. It is recommended that these carbamate hits reported here should be independently verified by trapping purified protein and designing experiments to assess the biological relevance of carbamylation on the identified proteins. In this study, histone carbamylation was concluded to be a worthwhile route for future investigation. The work pursued is detailed in Chapter 5 of this investigation and is particularly focused on the carbamylation of Histone H3 K79 due to the site's biological relevance to DNA transcription.

The identification of AGEs in this lysate screen has highlighted that future trapping experiments should also include using 13C Ci to verify that a reported hit is truly a trapped carbamate. Finally, a computational approach using the hits identified in this screen, alongside other known carbamates, and physiological factors like pKa, could be used to train a carbamate identification model to identify future hits.

4. Assessment of PROTAC's Activity at Normal and Hypercapnic Levels of CO_2.

4.1 Overview

Proteolysis-targeting chimaeras (PROTACs) are an emerging therapy designed to target the degradation of disease-related proteins by hijacking the ubiquitin-proteasome system (UPS).[150] Polyubiquitination at ubiquitin (Ub) lysine 48 (K48) is the primary signal for degradation by the proteasome. Intriguingly, UbK48 has been identified as a biologically relevant carbamylation site [84] and this post-translational modification could alter the degradation activity of PROTACs. In this chapter, several PROTAC compounds across two protein targets (BRD4 and SMARCA2) and cell lines (HEK293 and NCIH838) were tested under normal and hypercapnic levels of CO_2 to expand carbamate research into a pharmaceutical setting. It was hypothesised that the biologically relevant carbamate on UbK48 inhibits polyubiquitination on UbK48, decreasing the potency of PROTAC compounds.

PROTACs are dual-headed compounds with two targeting ligands, one targeting the disease protein and the other targeting the Really Interesting New Gene (RING) E3 ligase, the final enzyme in the ubiquitin enzyme cascade. The compounds used in this study target Von Hippel Landau (VHL) and Cereblon (CRBN) Ub RING E3 ligases, which are a focus for PROTAC development. At least 30 proteins have been targeted and degraded by CRBN PROTACs and at least 20 by VHL PROTACs.[151] The background to this chapter details the UPS, RING E3 ligase targets and the mechanism of action for PROTACs. The results in this chapter describe a method used to assess the dose response of PROTACs incubated at 5% and 10% (v/v) CO_2.

4.2 The Ubiquitin Enzyme Cascade for Protein Degradation

Ub is a small (8.5 kDa) eukaryotic polypeptide essential for protein regulation, amino acid recycling [152] and the DNA damage response.[153] Substrates are modified by Ub through isopeptide bonding at N-terminal or lysine residues in a process known as ubiquitination. Types of ubiquitination include mono, the attachment of a single Ub; multi-mono, the binding of Ub at various substrate sites; and Ub chains, which are linked by distinct lysine residues. Many chain types exist due to the seven lysine residues on Ub. In addition, the linkages between these lysines can either be unbranched, always to the same residue or branched to different residues.[154155]. Importantly, a protein tagged with a polyubiquitin chain linked by K48 is recognised by the 26S proteasome and degraded. The various types of ubiquitination and their associated biological function are described in Figure 4-1.

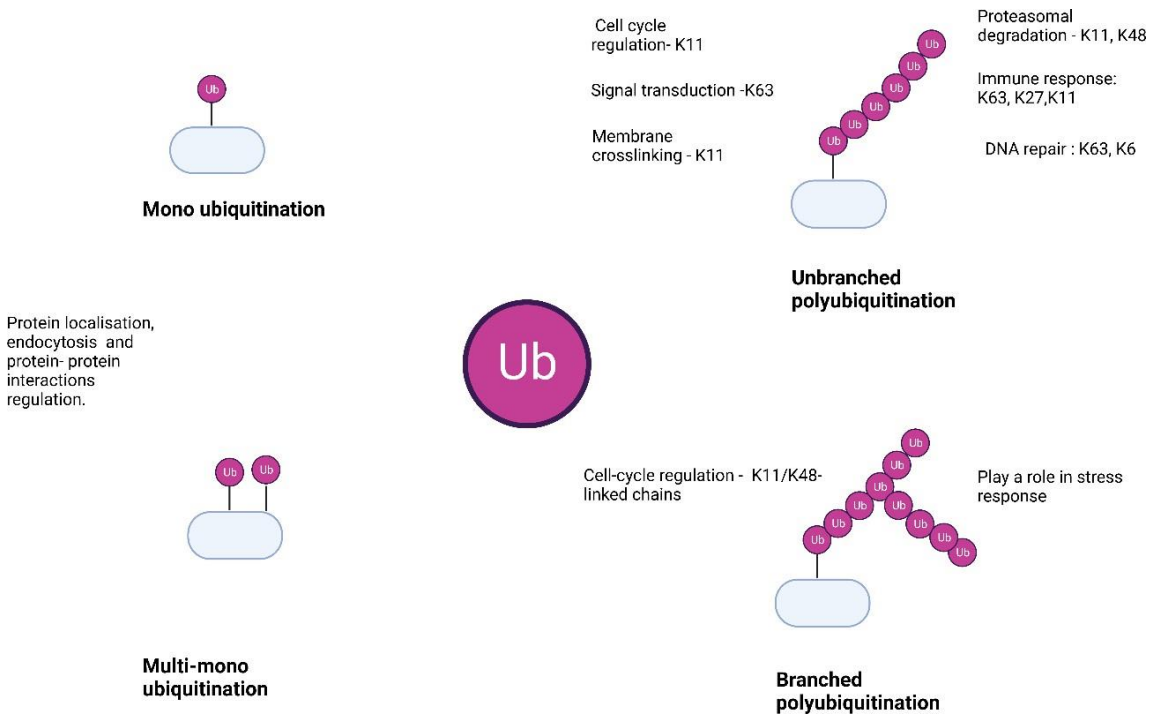


Figure 4-1 Types of ubiquitination: mono, multi mono and branched/ unbranched polyubiquitination. Modified from [156] and created using BioRender.

Ubiquitination involves an enzymatic cascade dependent on adenosine triphosphate (ATP) hydrolysis, as shown in Figure 4-2. After the first round of ATP hydrolysis, an adenosine monophosphate (AMP) modified C terminal ubiquitin binds to the active site cysteine residue on a ubiquitin-activating enzyme E1. Following a second round of ATP hydrolysis (Appendix Figure 8-4), the activated ubiquitin is shuttled from E1 to the ubiquitin-conjugating enzyme E2. The final transfer of ubiquitin is catalysed by the ubiquitin-protein ligase E3, which brings E2 and the target protein into proximity. [157]

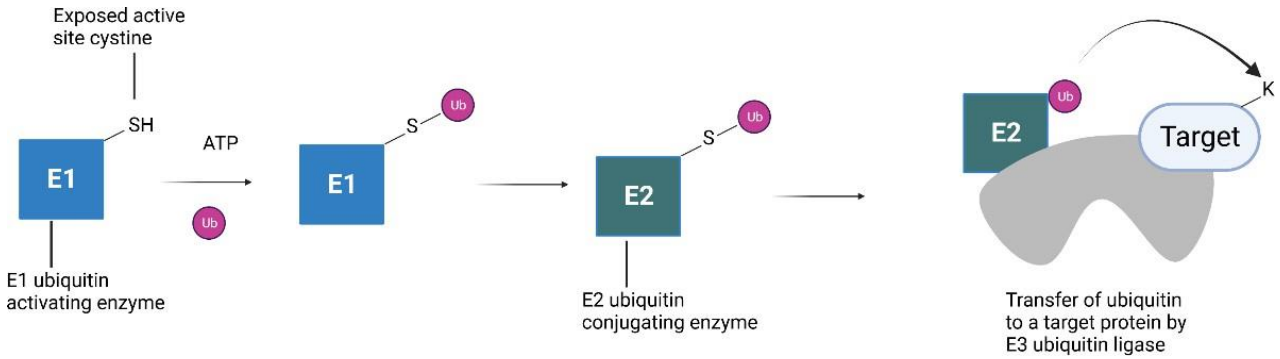


Figure 4-2 The ubiquitin enzyme cascade for protein degradation. Adapted from [158] and created using BioRender.

4.3 E3 Ligases

Over 800 E3 ligases have been identified in humans.[159] These are divided into various classes, which include homologous to E6AP C terminus (HECT), RING, U-box, and RING between RING (RBR) E3 ligases. The RING family E3 ligases are mechanistically distinct from the other classes because RING-mediated Ub transfer does not rely on an E3 ligase- Ub intermediate.[158] The research carried out in this chapter considers an important subclass of RING E3 ligases supported by the hydrophobic scaffolding of cullin proteins, which mediate 20% of ubiquitin-initiated proteolysis.[160] Cullins are typically complexed with a RING finger protein, either RING box protein 1 or 2 (Rbx1/ Rbx2) at the C terminus and an adaptor and receptor domain at the N terminus.[161162] Rbx1/Rbx2 recruits the E2 enzyme, whilst the adaptor protein is directly involved in substrate recruitment.[163] The structure of a Cullin RING E3 ligase (CRL) is outlined in Figure 4-3.

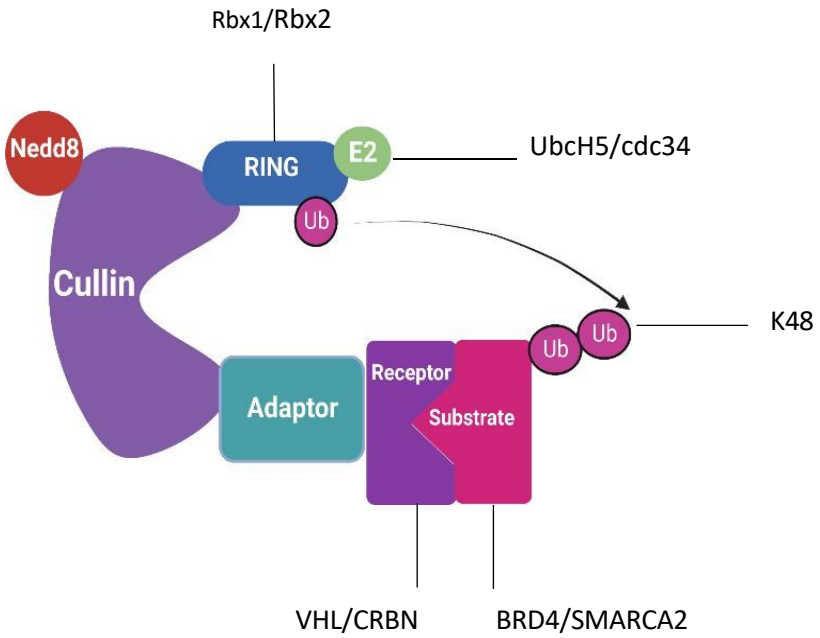


Figure 4-3 Subunit composition of Cullin RING E3 ligase, created using BioRender.

In clinical studies, CRLs composed of the receptor domains VHL and CRBN are of particular interest due to being ubiquitously expressed in humans.[164] The CRBN receptor is complexed with Rbx1, cullin4 (CRL4) and damaged DNA–binding protein 1 (DDB1) in the $CRL4^{CBRN}$ multi-subunit protein. The

VHL receptor is complexed with Rbx1, cullin2 (CRL2) and elongin BC in the $CRL2^{VHL}$ multi-subunit protein. The ubiquitin linkage activity for CRLs is determined by the E2 ligase recruited by the Rbx subunit.[165] Rbx1 recruits and activates the cell division cycle 34 (cdc34) or ubiquitin-conjugating enzyme E2 H5 (ubcH5) E2 enzymes.[166] Rbx1 preferentially recruits cdc34 in its ubiquitin-conjugated form compared to the unconjugated form by a factor of 50, determined by assessing the binding affinity (K_d) using titration.[167] UbcH5 attaches the initial ubiquitin to the substrate, whereas cdc34 is responsible for the successive addition of ubiquitin at K48 to the substrate during polyubiquitin chain formation.[168] This mechanism was first discovered in the Skp1-cullin 1-F-box (SCF) complex but has been shown to apply to other CRLs.[169]

Another important PTM for CRLs' activity is neddylation, which is closely related to ubiquitination. Neddylation induces a conformational change in the CRL E3 ligase complex and activates Ub transfer.[170] Fluorescence resonance energy transfer (FRET) studies have shown that neddylation of the cullin subunit enhances the interaction of CRL ligases with ubiquitin activated E2 enzymes. Additionally, the neddylation modification positions the E2 ubiquitinated site close to the Ub-accepting sites on substrates (e.g. K48) whilst simultaneously preventing the association of CRL with the Cullin-associated and neddylation-dissociated 1 (CAND1) complex which is a negative regulator of CRLs.[171] In Figure 4-3, the CRL complex has been modified by Neuronal precursor cell-expressed developmentally down-regulated protein 8 (Nedd8) to illustrate the activated form of the complex.

The E3 ligase family is diverse and targets many substrates in different cell locations. E3 ligases exhibit varying expression levels dependent on cell state. The biological functions of CRBN and VHL complexes are discussed in sections 4.3.1 and 4.3.2.

4.3.1 Biological Functions of the CRBN Complex

Several CRBN-mediated degradation substrates exist. In this section, these targets and their associated biological functions are highlighted. CRBN knockdown has shown decreased cell viability

whilst overexpression promotes cell proliferation, elucidating roles in cell metabolism and apoptosis.[172] In particular, CRBN binds to AMPK and glutamine synthetase, thus having a key role in metabolism. Moreover, voltage-gated chloride channel-2 (CIC2) and large conductance calcium- activated potassium channels (BKCa) are substrates for CRBN, making this E3 ligase complex important for ion balance in the cell.[173]

CRBN plays a key role in Wingless-related integration (Wnt) signalling because casein kinase 1α (CK1α) is a CRBN substrate. The canonical Wnt signalling pathway is mediated by β-catenin, where in the presence of Wnt, β-catenin is translocated to the nucleus to act as a transcriptional activator and in the absence of Wnt, β-catenin is degraded by the proteasome. When Wnt proteins bind extracellularly, the destruction complex composed of CK1α, axin, adenomatosis polyposis coli (APC), protein phosphatase 2A (PP2A) and glycogen synthase kinase 3 (GSK3) is inactivated. In particular, the Wnt ligand promotes CRBN-dependent degradation of CK1α and enhances transcriptional activation.[174]

4.3.2 Biological Functions of the VHL Complex

VHL-mediated degradation of Hypoxia Inducible Factor 1 Subunit Alpha (HIF1α) is the best-characterised function of this E3 ligase. HIF1α is an important transcriptional regulator for the cellular response to lack of oxygen.[175] HIF1α is degraded by VHL-dependent ubiquitination under normal levels of oxygen. VHL only recognises HIF1α in the hydroxylated form. Hydroxylation is carried out by prolyl hydroxylase domain (PHD) enzymes whose activity is tightly controlled by oxygen levels.[176] When the cell is in a state of hypoxia, with low oxygen levels, the PHD enzymes are inactive, and therefore HIF1α is not recognised by VHL. Other substrates of VHL include Sprouty2 (Spry2), which regulates cell proliferation and migration and the epidermal growth factor receptor (EGFR).

4.4 Proteolysis Targeting Chimera (PROTACS)

PROTACs have been developed to interact with the UPS and drive the rate of K48-ubiquitinated degradation. PROTACs consist of a ligand binding to a substrate joined to a ligand binding to an E3 ligase, as shown in Figure 4-4. PROTACs are a new modality of drug discovery that aims to use the cell's degradation machinery to remove proteins instead of the traditional approach of protein inhibition. [177] This approach has offered an alternative to targeting proteins previously classed as undruggable by classical inhibition.[178] The therapy has promising applications in cancer, [179,180] autoimmune, [181] metabolism disorders, [182] cell cycle regulation, [183] and neurodegenerative diseases.[184]

CBRN and VHL E3 ligases have been most frequently used to design PROTACS. The ligands developed for these E3 ligases have favourable binding properties, acceptable physiochemical profiles, and structural information on their binding modes.[185]

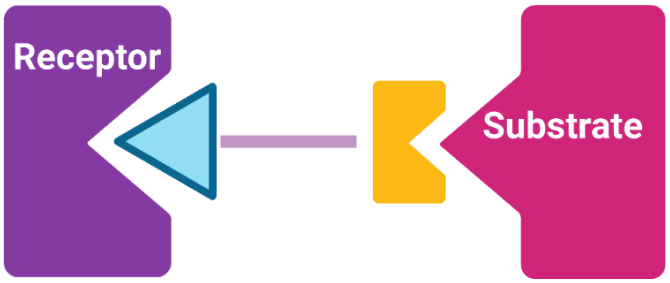


Figure 4-4 A dual-headed PROTAC compound targeting a protein of interest for degradation created using BioRender.

4.5 Substrate Targets for PROTAC Compounds.

This therapy can be applied to a range of disease-causing proteins. However, in this investigation, the chosen substrates were BRD4 and SMARCA2. These substrates and cell lines were primarily chosen due to the tools for the investigation being available and AZ's strategic research interest due to their link with cancer and the clinical interest in PROTACs as a cancer therapy.

4.5.1 Bromodomain-Containing Protein 4 (BRD4)

Bromodomain-containing protein 4 (BRD4) is a nuclear protein that contains two N-terminal bromodomains (BD1 and BD2), which bind acetylated lysines on histone proteins, altering gene transcription. This epigenetic modifier is essential in cellular regulation through chromatin remodelling, cell cycle progression and cellular differentiation.[186] Additionally, BRD4 has been identified as a cancer target as it affects the expression of oncogenes, for example, MYC.[187]

4.5.2 Switch/Sucrose Non-Fermentable (SWI/SNF) Related, Matrix Associated, Actin Dependent Regulator of Chromatin, Subfamily A, Member 2 (SMARCA2)

Switch/Sucrose Non-Fermentable (SWI/SNF) complexes are chromatin remodelling complexes that use ATP hydrolysis to alter nucleosome binding and transcription.[188] SWI/SNF complexes have two ATPase subunits, including SMARCA2 and SMARCA4. SWI/SNF complexes are a focus of cancer research because they have been repeatedly identified as mutated. Around 20% of all human tumour samples have shown abnormalities in SWI/SNF. The two ATPase subunits display antagonistic behaviour, as discussed by Martinez *et al*.[189] SMARCA2 is an attractive degradation target in SMARAC4 mutant cancer because these tumor cells require SMARCA2 protein to survive.[190]

4.6 Ubiquitin K48 as a Carbamylation Binding Site

Linthwaite *et al*. showed that ubiquitin conjugation at K48 was downregulated in higher concentrations of inorganic carbon using a ubiquitin conjugation assay. This result was rationalised by identifying a carbamate residing on Ub K48 by NMR and LCMSMS. Notably, the residency time for carbamylation at Ub K48 was long enough to block the transfer of free Ub.[84]

4.7 Nano-Glo HiBiT Lytic Detection System

To assess the effects of carbamylation at ubiquitin K48 in the context of PROTACs, the Nano-Glo HiBiT lysis assay was used, as depicted in Figure 4-5. The degradation activity of PROTACS targeted to BRD4 and SMARAC2 under 5% and 10% (v/v) CO_2 was measured in a dose-dependent manner.

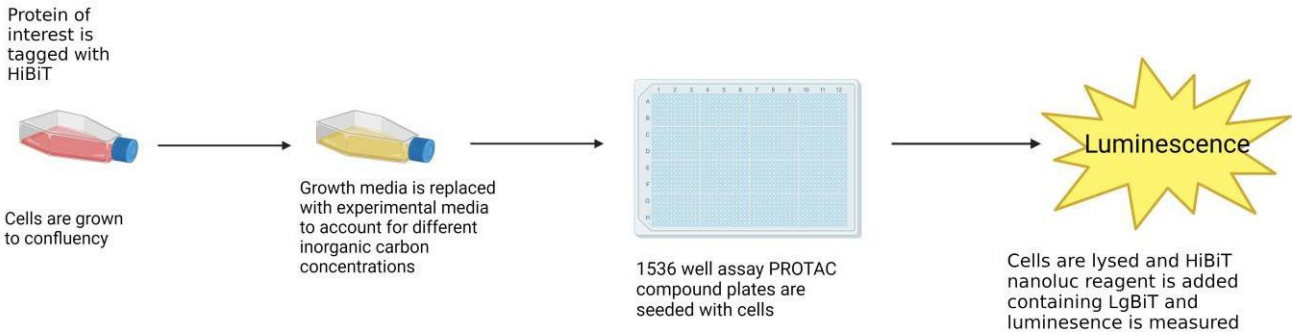


Figure 4-5 Experimental process for testing tool PROTAC compounds under 5% and 10 % (v/v) CO_2. Created using BioRender.

Target proteins, BRD4 and SMARCA2, were modified with a high bioluminescence tag (HiBiT) in HEK293 and NCI-H838 cell lines, respectively. CRISPR/Cas9 knock-in technology enables the addition of a small 11 amino acid HiBiT tag to any protein of interest at the N or C terminus.[191] The NanoGlo HiBit lysis assay works on the basis of quantitative protein complementation. Following lysis, the HiBiT-tagged protein can be quantified by adding the large bioluminescence tag (LgBiT) and furimazine to produce a luminescent signal. This approach works by reconstituting the two subunits of the split luciferase, HiBiT and LgBiT. The reaction mechanism is outlined in Scheme 4-1.[192]

$$\mathrm{LgBiT} + \mathrm{HiBiT} \rightarrow \mathrm{NanoBiT\ Luciferase}$$

$$Luciferin + ATP + O_2 \xrightarrow[Luciferase]{Mg^{2+}} Oxyluciferin + luminesence\ @\ 520\ nm$$

Scheme 4-1 Reaction of Nano Glo HiBiT lysis assay

This assay is attractive for measuring the PROTAC dose response (Figure 4-6) because the luminescence produced will be proportional to the HiBiT-tagged BRD4 or SMARCA2 protein quantity. PROTAC compounds will degrade the POI proportional to the concentration of PROTAC used, which can be plotted as a dose-response curve (DRC). Controls to this assay included DMSO treatment, where the maximum luminescence signal was obtained, and an inhibitor treatment at a high concentration known to cause 100% degradation, which gave the background luminescence. False positives of this approach include luciferase-inhibiting or toxic compounds.

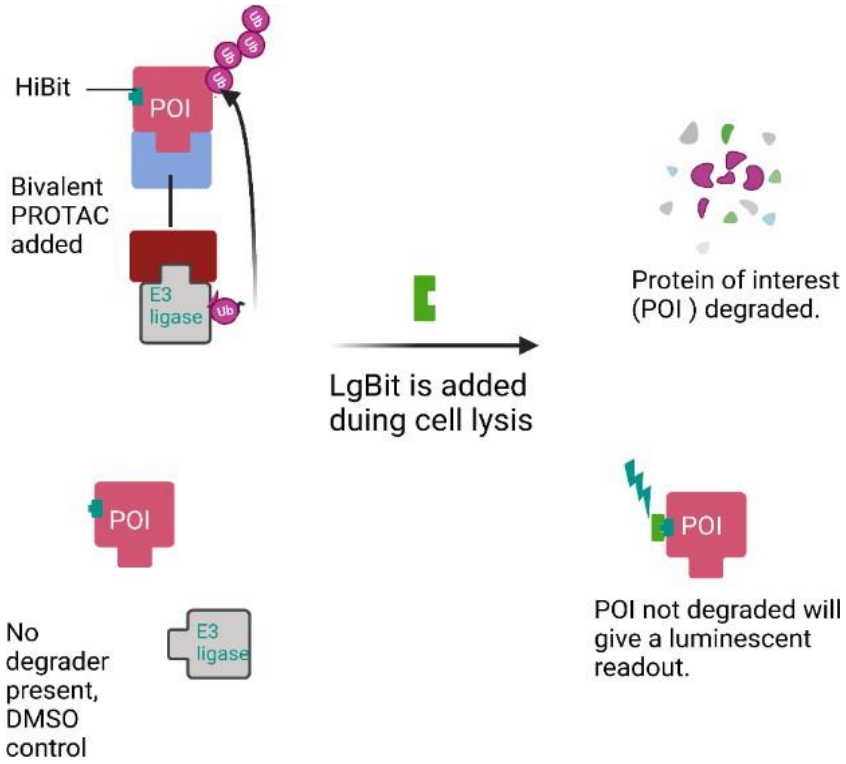


Figure 4-6 The effect on luminescence in the presence or absence of a PROTAC degrader compound created using BioRender.

4.8 HiBiT lysis Assay Quality

The maximum and minimum signal controls were passed through various quality metrics to threshold assay data as outlined in detail by Iversen *et al.*[193] The equations in this section contain Greek letters as defined by statistics.

4.8.1 Z-prime (Z')

The Z-prime (Z') statistic is a widely used measure of assay quality. The Z'-factor describes how well separated the positive and negative controls are to define a signal window and indicate the likelihood of false positives or negatives. Equation 4-1 outlines the calculation of this measurement where the plateau refers to the positive control (minimum compound), and the baseline (DMSO) refers to the negative control.

$$Z^{'} = 1 - \frac{3(\sigma plateau + \sigma baseline)}{|\mu plateau - \mu baseline|}$$

Equation 4-1 Z' calculation

4.8.2 Robust Z' (RZ')

Instead of Z' being the cut-off parameter for assays, the robust Z' (RZ') is used by AZ's HTS department. RZ' is less sensitive to outliers as it uses the median and median absolute deviation (MAD) instead of the mean and the standard deviation. The calculations for MAD and RZ' are shown in Equations 4-2 and 4-3, respectively. All assay plates passed the threshold of an RZ' of 0.5 or above.

$$MAD = median(|x - median(x)|)$$

Equation 4-2 Calculation of the median absolute deviation.

$$RZ' = 1 - \frac{3(MADplateau + MADbaseline)}{\left|\tilde{x}plateau - \tilde{x}baseline\right|}$$

Equation 4-3 Calculation of the robust Z' statistic.

4.8.3 Signal to Background Ratio

The signal-to-background ratio (S: B) is also an important parameter to determine, as defined in Equation 4-4. Figure 4-7 is an example of the S: B for a HiBiT assay carried out in this investigation. All assay plates presented had a S: B of greater than 3.

$$S{:}\,B = \frac{Raw\ luminesence_{Neutral\ control}}{Raw\ luminesence_{Inhibitor\ control}}$$

Equation 4-4 Signal-to-background ratio calculation

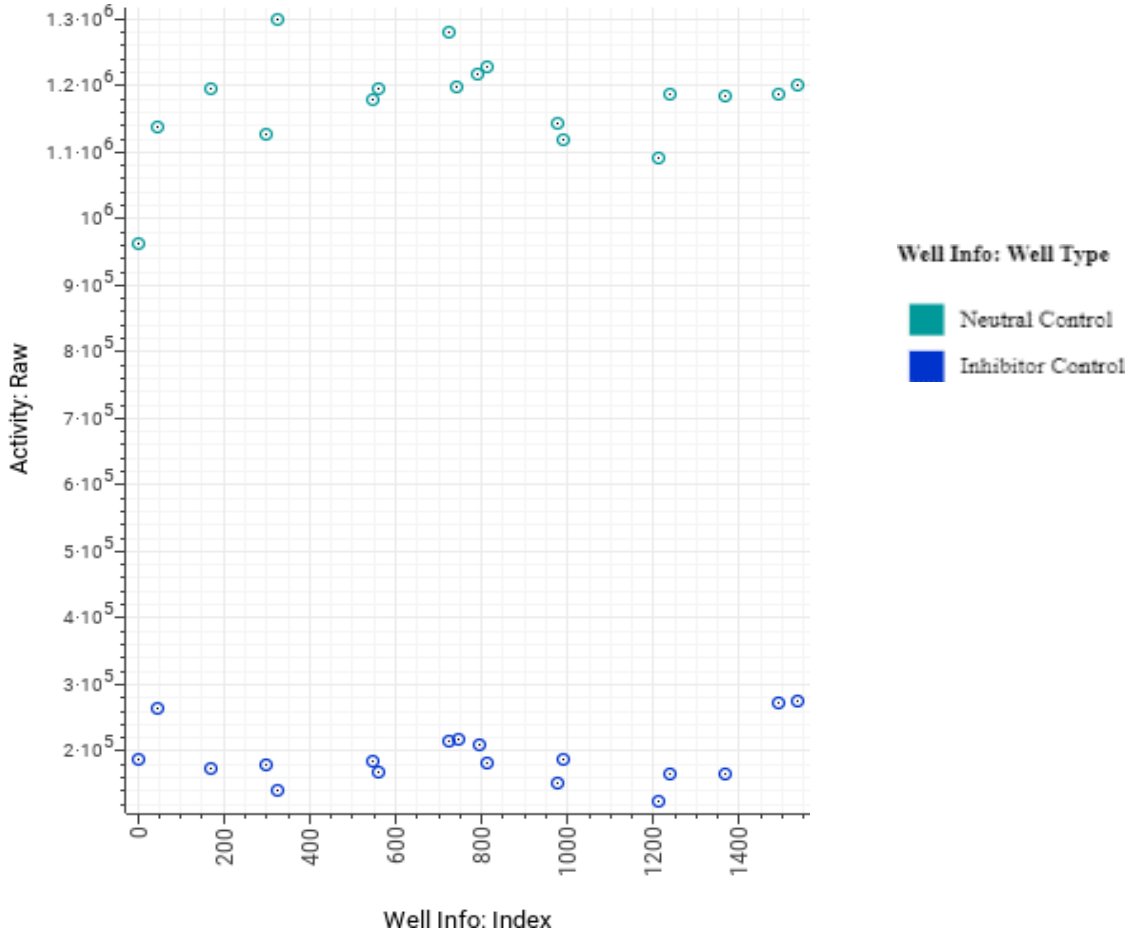


Figure 4-7 Raw luminescence values of controls versus the position of the control on a 1546 well plate. The neutral controls are HEK293-HiBiT-BRD4 cells treated with DMSO and the inhibitor controls are HEK293-HiBiT-BRD4 cells treated with a concentration of inhibitor known to give the maximum degradation response which is the background luminescence where n= 18. Individual values are plotted, and the replicates measured are spread evenly across the plate to reduce plate patterns. The signal-to-background ratio for the HiBiT lytic Nano Glo assay controls in this experiment was 6.

4.8.4 Coefficient of Variation

The coefficient of variation for the maximum controls was expressed as a percentage. This calculation is shown in Equation 4-5. A threshold of less than 10% was applied to assay plates.

$$\%\ CV = \frac{\mu}{\sigma} \times 100$$

Equation 4-5 Percentage coefficient of variation.

4.9 Dose-Response Curves (DRCs)

The dose-response of effective PROTACs displayed an inhibition sigmoidal-shaped curve, as shown in Figure 4-8. This is characterised into three main sections: a flat line at low concentrations until a minimum concentration of the drug displays a significant response, a steep descending linear line, and a plateau where the maximum response is reached and increasing the dose no longer has any effect. PROTAC concentrations were converted into the logarithmic scale to increase the linear range of the DRC. The more negative the log concentration, the lower the concentration of the compound. Genedata software derives the percentage activity from the raw luminescence value for each concentration by normalisation to the controls. From these values, a DRC is fitted based within the 95% confidence limits.

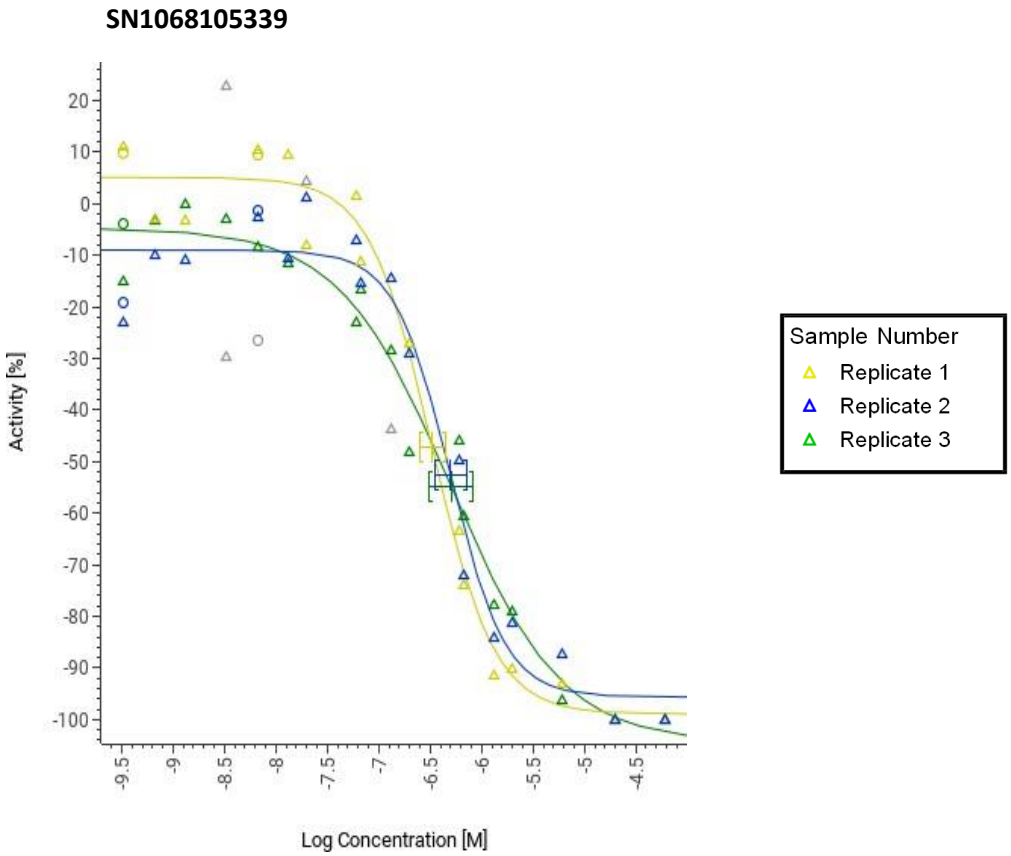


Figure 4-8 A typical dose-response curve for a BRD4 targeting PROTAC measured in biological triplicate. The normalized activity (%) derived from the luminescence readout signal is plotted against the specified log concentration represented here by each triangular data point. Replicate 1 is in yellow, replicate 2 is in blue and replicate 3 is in green. Genedata fits a dose-response curve to the data points and calculates a logIC_{50} value with an associated standard deviation for each curve fit displayed by the error bars shown here where n=3.

Parameters that can be analysed from DRCs include the drug efficacy, potency, and the Hill coefficient. The efficacy can be described as the minimal drug dose required for a maximal effect. Potency is the half-maximal inhibitory concentration (IC_{50}), the inhibitor concentration required to produce 50% of the maximal response. Potency and efficacy are analysed in pharmaceutical studies as they are measures of drug effectiveness and therapeutic potential. Whilst potency only assesses drug dosage, efficacy is a complex metric influenced by affinity, pharmacokinetics, and other essential

variables. The steepness of the DRC is defined by the Hill coefficient, which can give information about the binding mechanics.[194] A very steep curve (Figure 4-9A) is undesirable due to difficulty in achieving precise dose dependence. Figure 4-9B displays an inactive compound with no effect when increasing the compound dose. Figure 4-9C is a weakly active compound that does not display the expected DRC and has a significant standard deviation associated with the logIC_{50} which limits knowledge about the compound's effect.

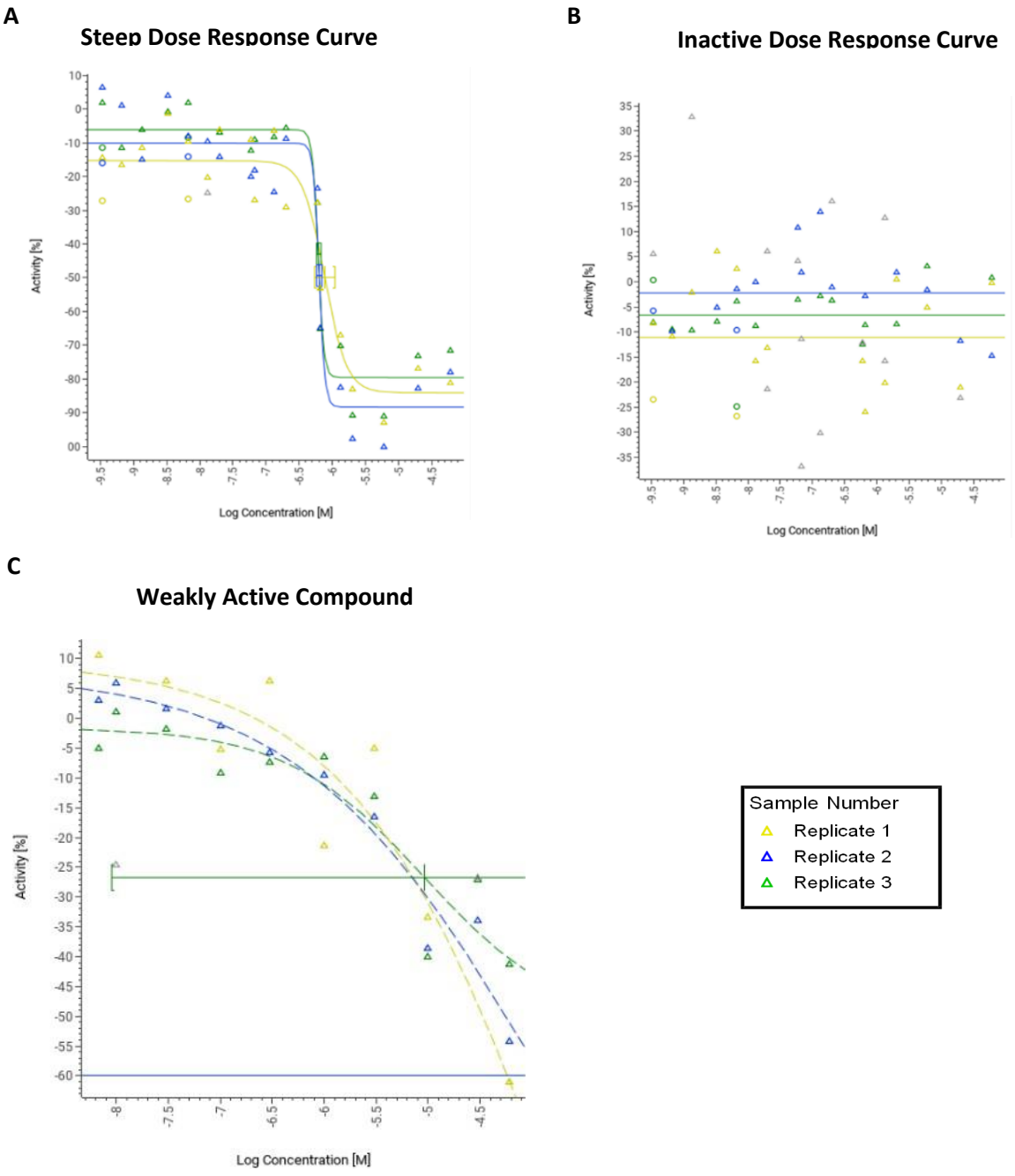


Figure 4-9 Dose-response curve examples with the same format as described in Figure 4-8. The normalized activity (%) derived from the luminescence readout signal is plotted against the specified log concentration represented here by each triangular data point. Replicate 1 is in yellow, replicate 2

is in blue and replicate 3 is in green. The DRCs are classed as (A) steep curve, (B) inactive and (C) a weakly active compound showing high variance for the logIC_{50} value where n=3.

4.10 PROTAC Degradation in the HiBiT Lysis Assay

It is important to note that each PROTAC is given an SN code corresponding to the labelling system for all compounds in the AZ compound library. The results in this section (4.10) all passed the quality screening metrics outlined in section 4.8. The log concentrations for the PROTAC compounds used in the BRD4 experiments were -4.2 to -9.5 M, and for the SMARCA2 experiments, -4.2 to -8.16 M.

4.10.1 PROTAC Selection

The degradation of BRD4 was tested using eighteen CRLs substrate recognition unit based PROTACs. From this preliminary data performed in biological triplicate, the compounds with the best DRC fit (5, CRBN) were chosen alongside an inactive (1, CRBN) and steep curve (1, VHL) compound to investigate if CO_2 incubation would impact PROTACs' potency. This selection process was also completed in biological triplicate for SMARCA2, where eighteen PROTACS were tested. Nine were selected, including those with the best curve fit (7, 3 CRBN, 4 VHL), a steep curve (1, CRBN), and a low activity (1, VHL) compound. The categorisation of dose-response curves developed by HTS, AZ for consistency across drug discovery screens was used to choose compounds with the best DRC fit.[195] The DRCs for the tool compounds and corresponding logIC_{50} values are shown in the appendix Figures 8-5 - 8-6, and Tables 8-1 - 8-2. The outlined graphs were chosen for the CO_2 experiment. During the compound selection process, if the DRC fit and potency were not closely correlated between biological replicates, these compounds were not selected for further study. In addition, compounds with the best fit reached full inhibition, as seen on DRC (Figure 4-8). When the linear portion of the curve was between -20% and -80% activity, this indicated the highest quality chemistry because this slope type indicates specific dosing effects. The exceptions to these screening parameters were the inactive, low inhibition and steep slope compounds that were selected.

Several compounds, particularly those that target SMARCA2, showed solubility issues or the hook effect; when the concentration of the PROTAC was too high, the compound began to inhibit itself. These outlier points were excluded by the Genedata software to not interfere with the dose-response curve fit, as depicted in Figure 4-10 by the grey triangles.

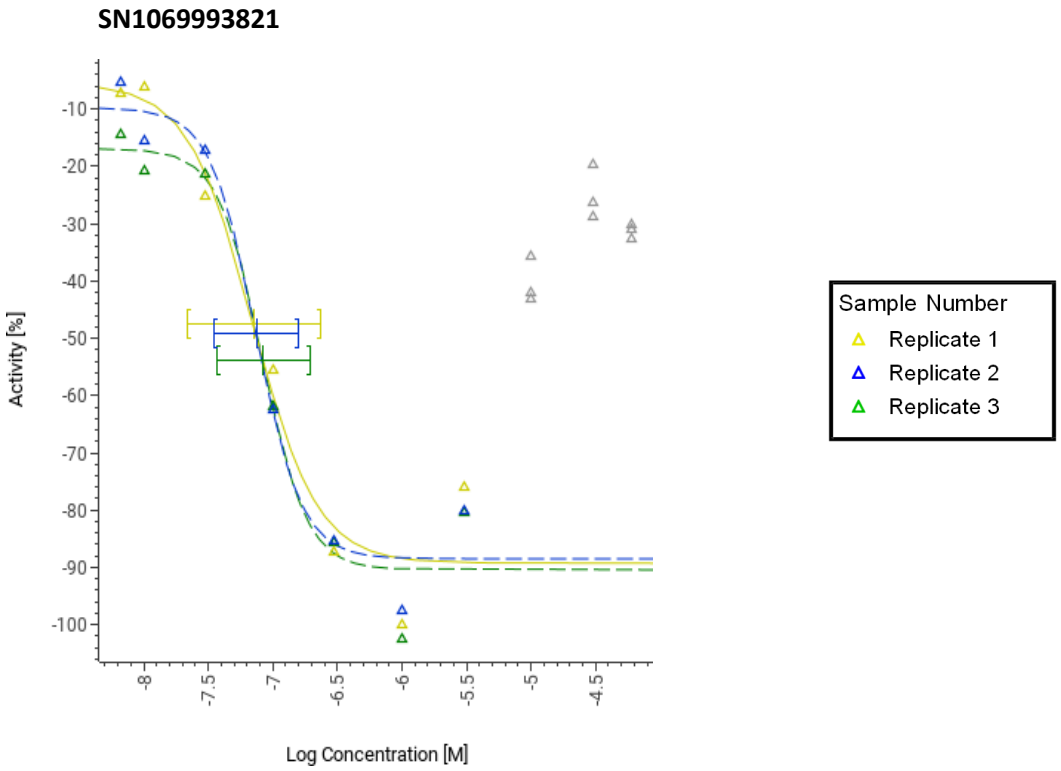


Figure 4-10 A dose-response curve (DRC) as described in Figure 4-8 for a SMARCA2 PROTAC where n=3. The normalized activity (%) derived from the luminescence readout signal is plotted against the specified log concentration represented here by each triangular data point. Replicate 1 is in yellow, replicate 2 is in blue and replicate 3 is in green. This compound exhibits the hook effect or solubility issues and when the concentration increases beyond a certain point, the data points show reduced inhibition and therefore are excluded (coloured in grey) as outliers when fitting the DRC.

The logIC_{50} was chosen as the parameter for comparing PROTAC activity under varying CO_2 as it is a clear comparison metric calculated from the DRC. A higher logIC_{50} (the highest concentration tested in this experiment was -4.2 M on the logarithmic scale) indicates that a higher dose is required to achieve the half-maximal effect. A lower logIC_{50} requires a lower compound concentration for the half-maximal response and reflects a more potent compound.

Figures 4-11 and 4-12 are the mean $logIC_{50}$ values calculated from three biological replicates for each tool compound tested for the degradation of BRD4 and SMARCA2, respectively. The error bars in these figures use the standard deviation values associated with the DRC fit for each replicate and propagate this variance into the calculation of the mean $logIC_{50}$, as discussed in section 4.10.3. The inactive compounds that were tested (SN1068695989**,** SN1151619913**,** SN1065656271) do not have a $logIC_{50}$ value because, at the highest concentration tested (100 μM), there is no change in the luminescence signal, as shown in Figure 4-9B. A few compounds had standard deviations larger than the range of concentrations tested in the experiment, plotted in Figures 4-11B and 4-12B and C. A number of these compounds had only one replicate value with a larger than expected standard deviation due to the DRC fit and, in these cases, the variable value was excluded as an outlier and replotted in Figures 4-11C and 4-12D. It is clear from these figures that SMARCA2 tool compounds are generally a lot more variable in their activity and this inherent variability will affect further experiments. The curve category, $logIC_{50}$ and standard deviation of the $logIC_{50}$ for all the tool compounds are given in Appendix Tables 8-1 and 8-2.

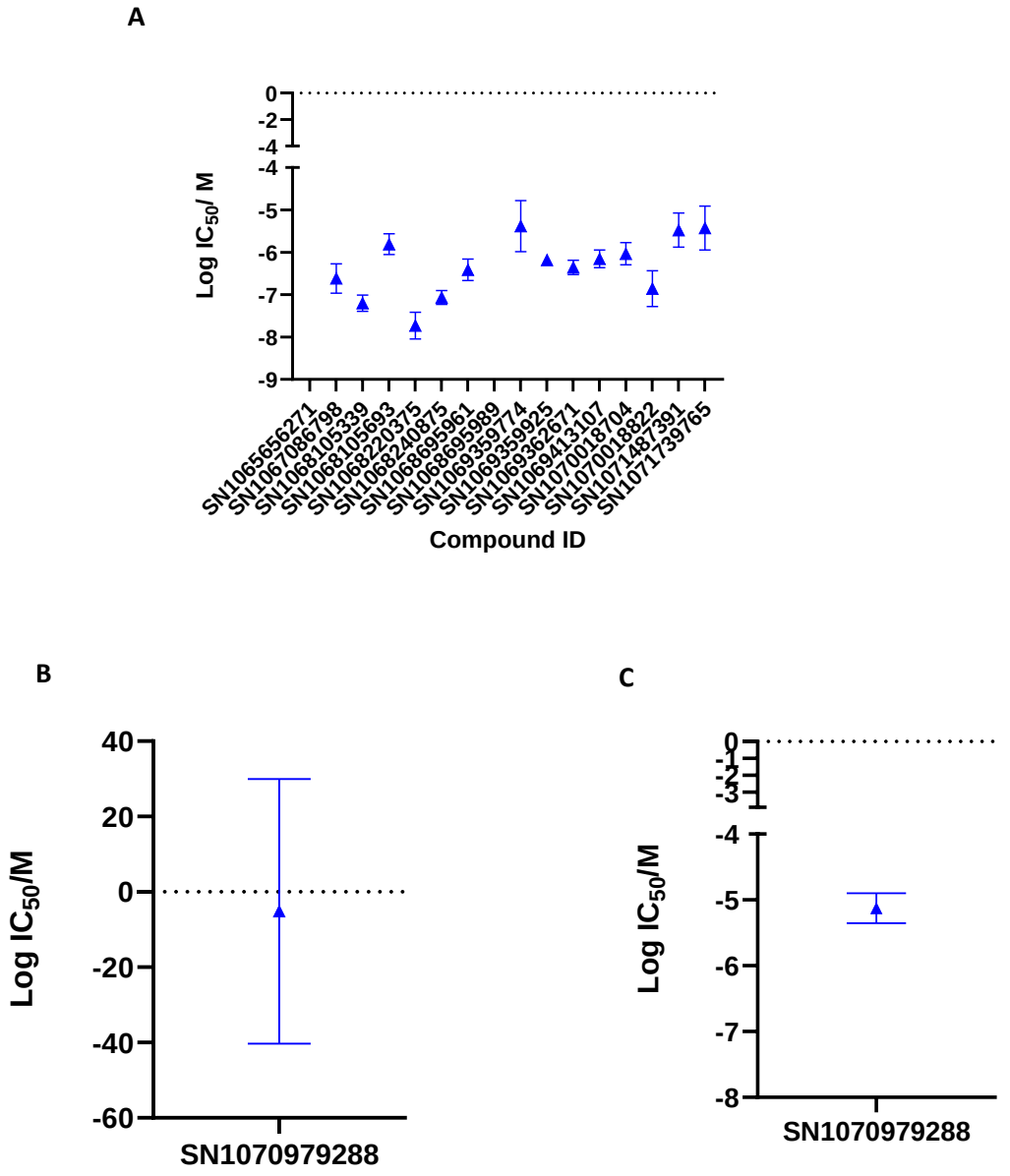
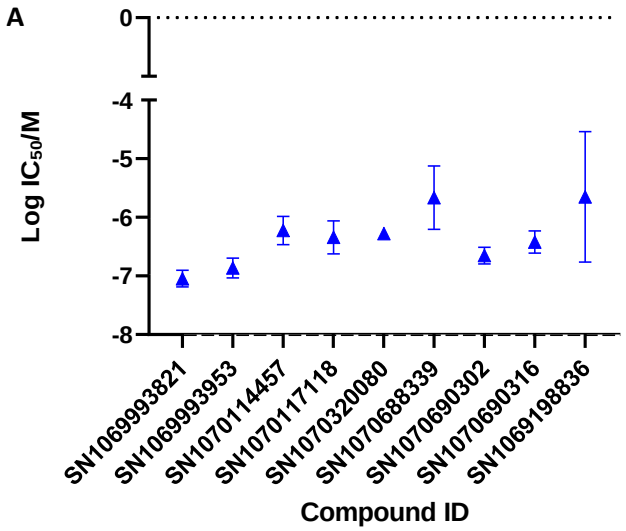
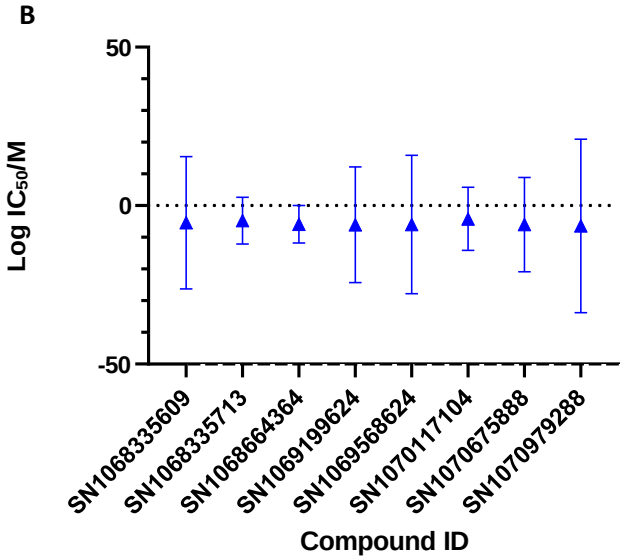
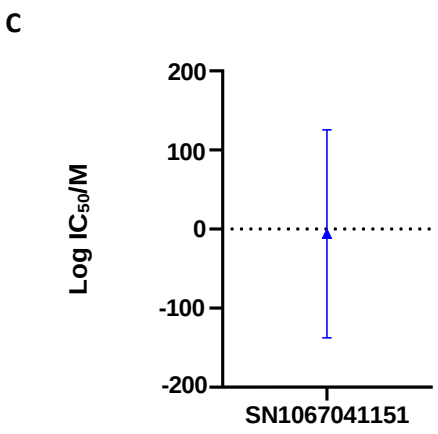


Figure 4-11 LogIC_{50} values determined from the dose-response curve plotted against the tool BRD4 compound tested where all values are represented as the mean with error bars calculated from the absolute standard deviation where n=3. (A) The error bars are within the concentration range tested and the compound, SN1068695989 has no logI_{C50} because this is an inactive compound. The only compound missing from Figure A is SN10726900031 which is plotted in Figures (B) as measured and (C) by excluding a singular data point.


A
0
-4
-5
-6
-7
-8
Log IC50/M
SN1069993821
SN1069993953
SN1070114457
SN1070117118
SN1070320080
SN1070688339
SN1070690302
SN1070690316
SN1069198836
Compound ID


B
50
0
-50
Log IC50/M
SN1068335609
SN1068335713
SN1068664364
SN1069199624
SN1069568624
SN1070117104
SN1070675888
SN1070979288
Compound ID


C
200
100
0
-100
-200
Log IC50/M
SN1067041151

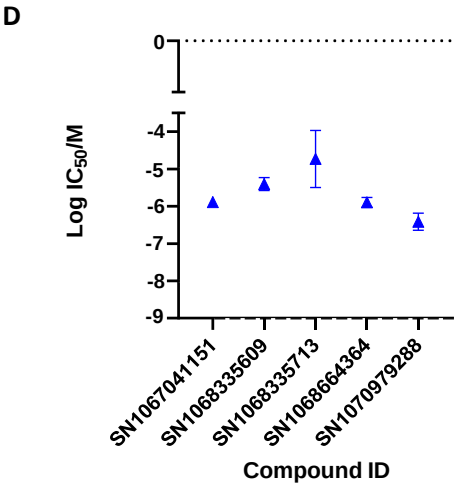

D
0
-4
-5
-6
-7
-8
-9
Log IC50/M
SN1067041151
SN1068335609
SN1068335713
SN1068664364
SN1070979288
Compound ID

Figure 4-12 LogIC_{50} values determined from the dose-response curve plotted against the tool SMARCA2 compound tested where all values are represented as the mean with error bars calculated from the absolute standard deviation where n=3. (A) The error bars are within the concentrations range tested (B) error bars are outside the concentrations tested. (C) the most variable logIC_{50} compound. (D) variable compounds replotted by excluding a singular data point from the triplicate measurements. If the compound was not replotted, more than one data point was attributed to this variance.

4.10.2 The Effect of CO_2 Treatment on PROTAC Activity

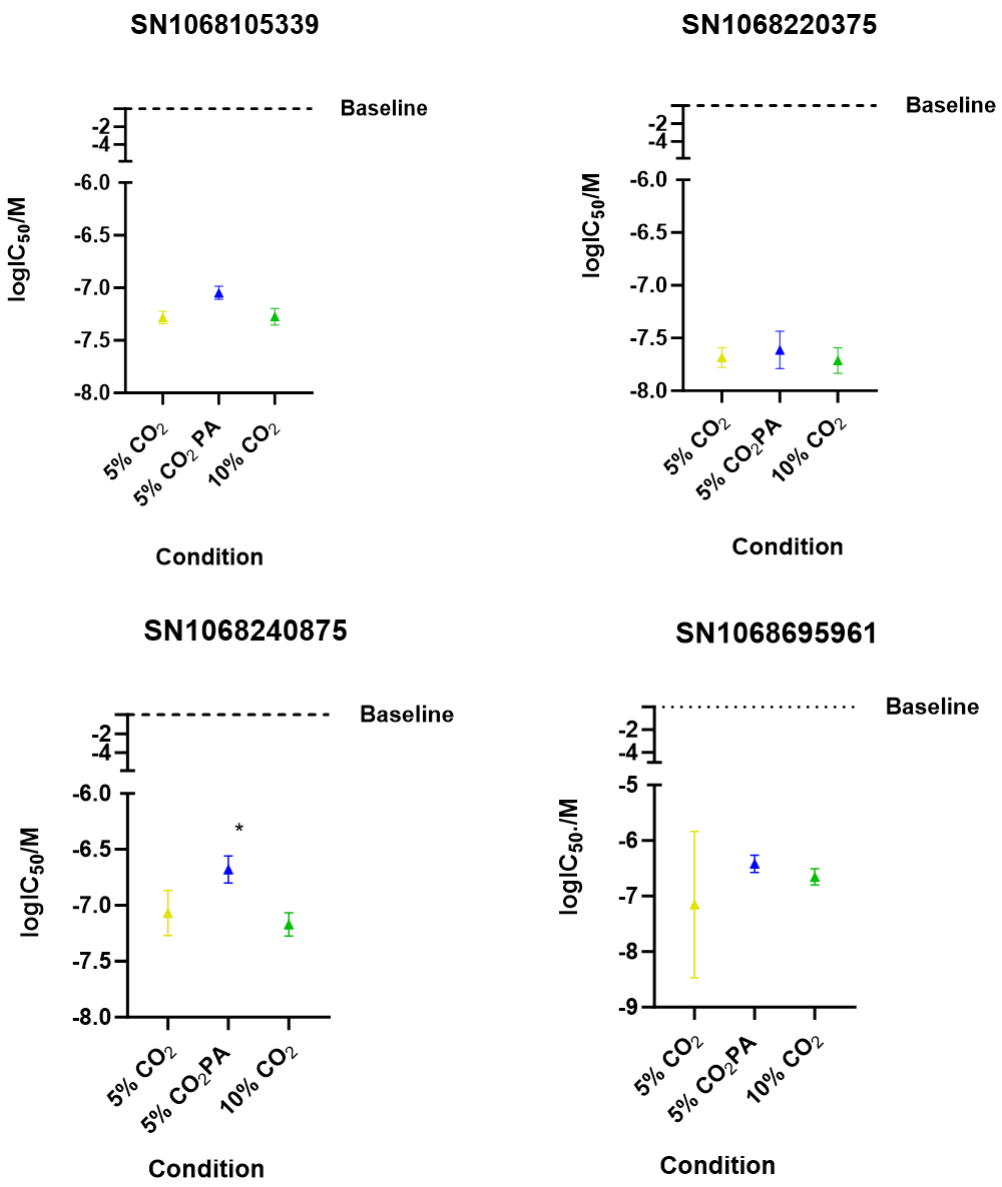
In the following experiments, the DRCs of the selected PROTACs were obtained under the CO_2 concentrations of 5% CO_2 with and without treatment of 2 mM propionic acid and 10% CO_2 in biological triplicate. The rationale for treatment with propionic acid as an intracellular acidity control is discussed in section 1.5.

Sections 4.10.3 and 4.10.4 plot the mean logIC_{50} for the selected BRD4 and SMARCA2 PROTACs from three biological repeats. Initially, the error bars plotted only considered the standard deviation from calculating the mean from the three logIC_{50} values (Figures 4-13 and 4-14). Outliers when considering this type of variance only include whether the DRC shape has changed or shifted along the x-axis to require a different inhibitory concentration to achieve 50% activity. However, this method does not consider the standard deviation of calculating each replicate's logIC_{50} value from the DRC fit. This error was incorporated in the error bars in Figures 4-15 and 4-16 and is the measure of true variance across the experiment. The DRC curves for the CO_2 experiments for BRD4 and SMARCA2 can be seen in Appendix Figures 8-7 – 8-13A-C and 8-14 – 8-22A-C, respectively.

4.1.1.1 The effect of CO_2 on PROTAC-mediated degradation of BRD4

Figure 4-13 shows logIC_{50} values ranging from - 6.5 to -8 M with strong concordance between the replicates reflected in the small error bars. SN106869589 is the inactive PROTAC, and no logIC_{50} is given. However, the software will estimate the value and assign a q log$AC_{50,}$ which is quoted as the highest concentration tested, -4.2 M and these are values plotted here. Aside from the inactive

compound, the compound exhibiting the highest logIC_{50} is SN106936271 at -6.25 M, the steep curve compound as shown in Figure 4-9A. The error bars are small for all compounds except for SN1068695961 in the 5% condition. However, there is a data point that is likely an outlier based on distance from the mean of the other data points.



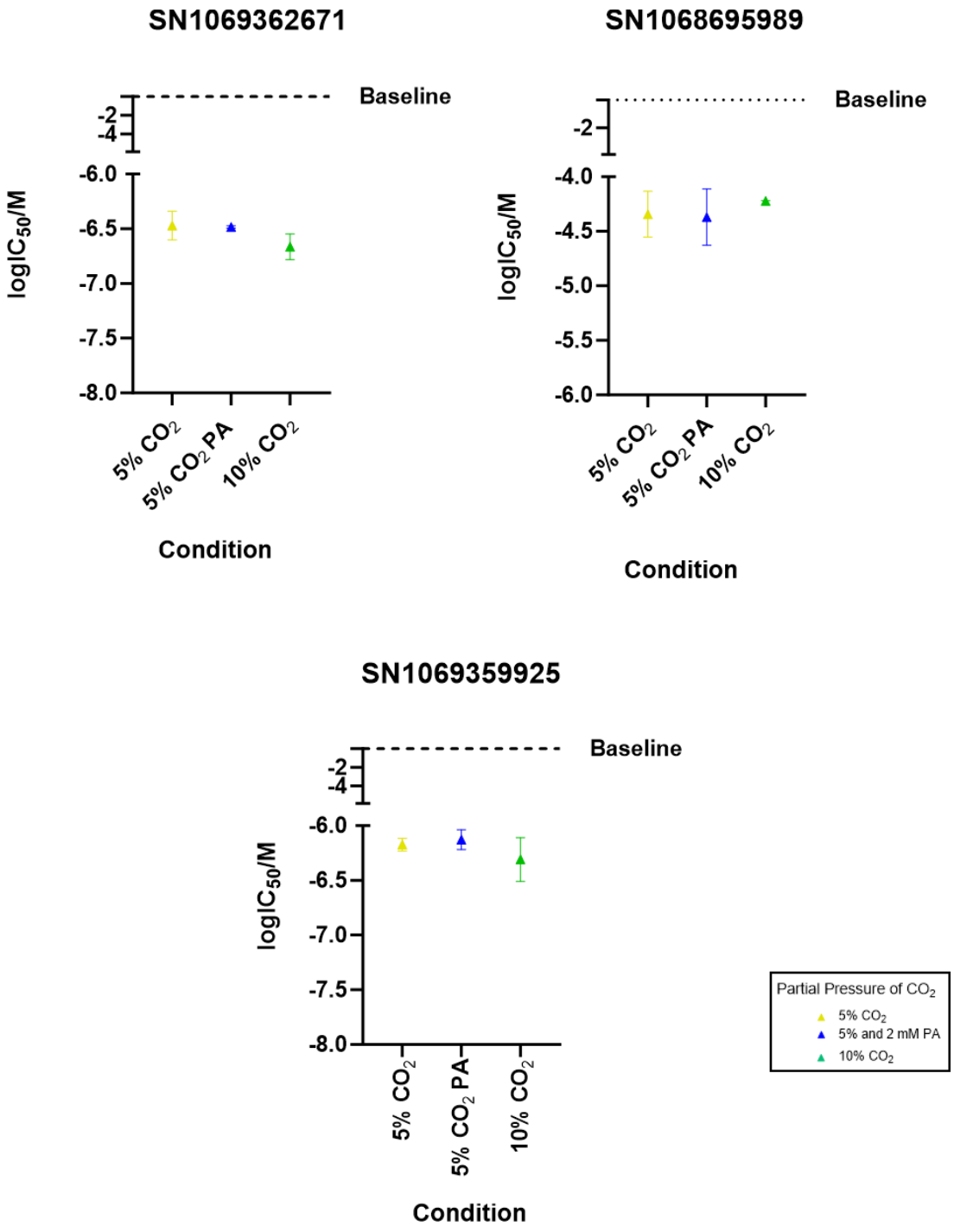
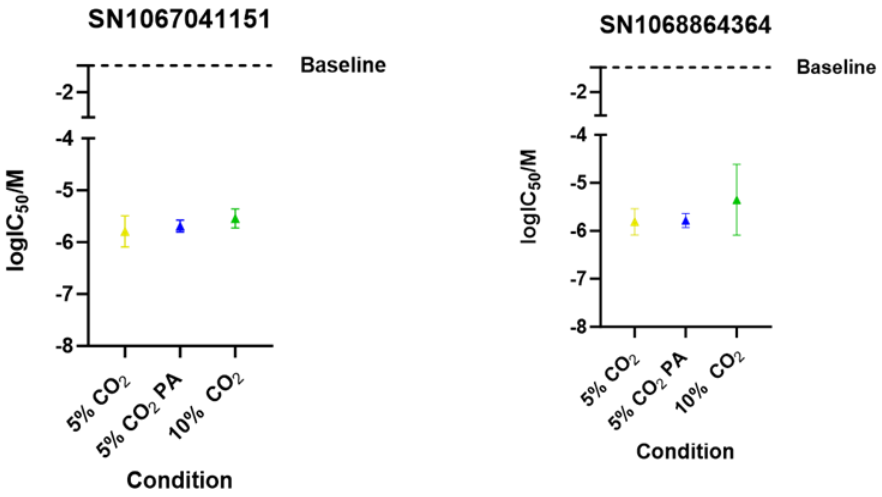


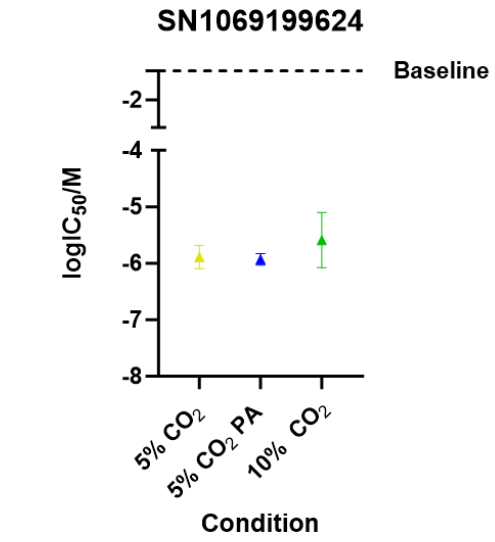
Figure 4-13 The $logIC_{50}$ plotted against the CO_2 experiment condition for each BRD4 targeting PROTAC labelled with an SN code as determined from the dose-response curve. The condition 5% CO_2 PA describes cells incubated at 5% CO_2 treated with 2 mM propanoic acid. All values are represented as

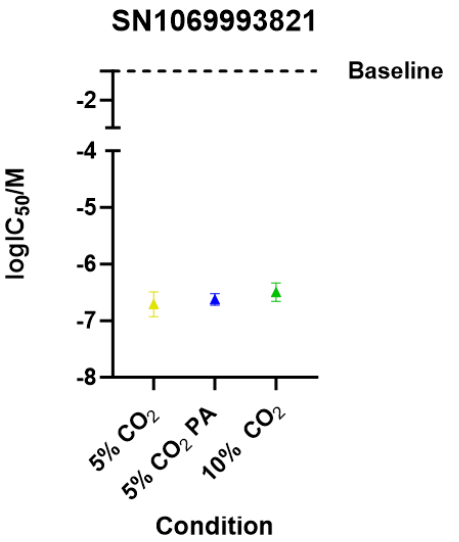
the mean with error bars calculated from the standard deviation for the mean calculation of the $logIC_{50}$ values where n=3. The significance of the data was determined by one-way ANOVA and multiple comparison tests showed the individual relationships which were significant where * is $p<0.05$.

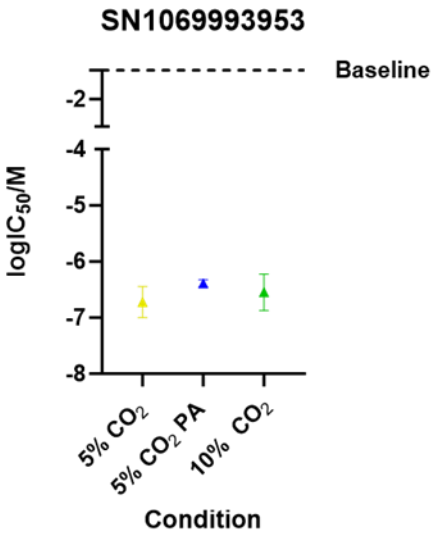
4.10.3 The Effect of CO_2 on PROTAC-mediated Degradation of SMARCA2

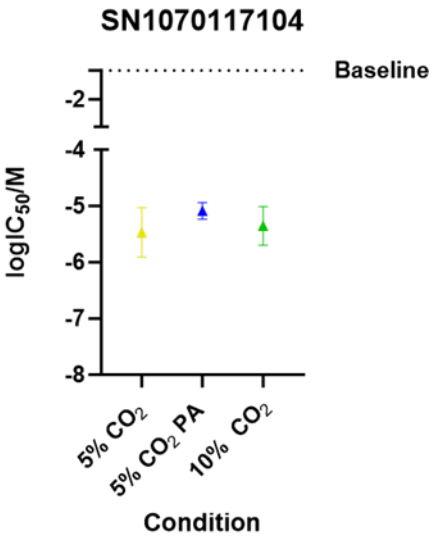
Figure 4-14 shows $logIC_{50}$ values for all the tested compounds. The compound that was selected as weakly active was SN1070117104. The data supports this as it has the highest $logIC_{50}$ at ~ -5.5 M. All the other tested compounds, including the steep curve compound SN1070320080, had a $logIC_{50}$ within a range of -6 to -7 M. Similar to the BRD4 dataset, the error bars for the mean $logIC_{50}$ calculation are very small, excluding one data point for SN1068864364 10% CO_2 which could again be considered as outlier.

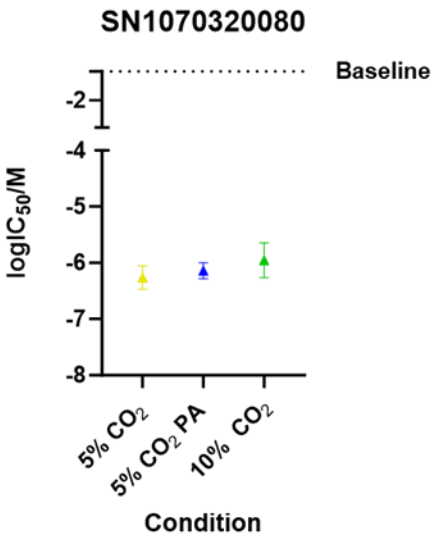


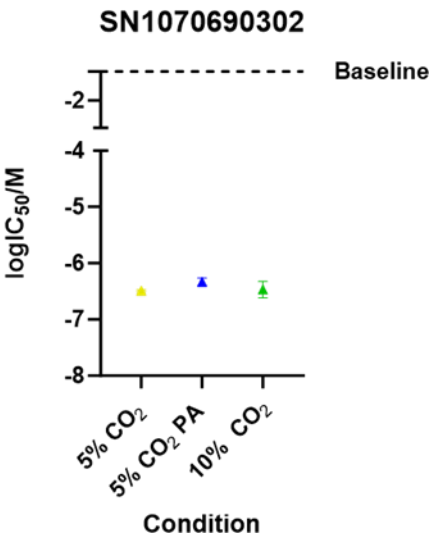

SN1069199624
Baseline
-2
-4
-5
-6
-7
-8
logIC50/M
5% CO2
5% CO2 PA
10% CO2
Condition


SN1069993821
Baseline
-2
-4
-5
-6
-7
-8
logIC50/M
5% CO2
5% CO2 PA
10% CO2
Condition


SN1069993953
Baseline
-2
-4
-5
-6
-7
-8
logIC50/M
5% CO2
5% CO2 PA
10% CO2
Condition


SN1070117104
Baseline
-2
-4
-5
-6
-7
-8
logIC50/M
5% CO2
5% CO2 PA
10% CO2
Condition


SN1070320080
Baseline
-2
-4
-5
-6
-7
-8
logIC50/M
5% CO2
5% CO2 PA
10% CO2
Condition


SN1070690302
Baseline
-2
-4
-5
-6
-7
-8
logIC50/M
5% CO2
5% CO2 PA
10% CO2
Condition

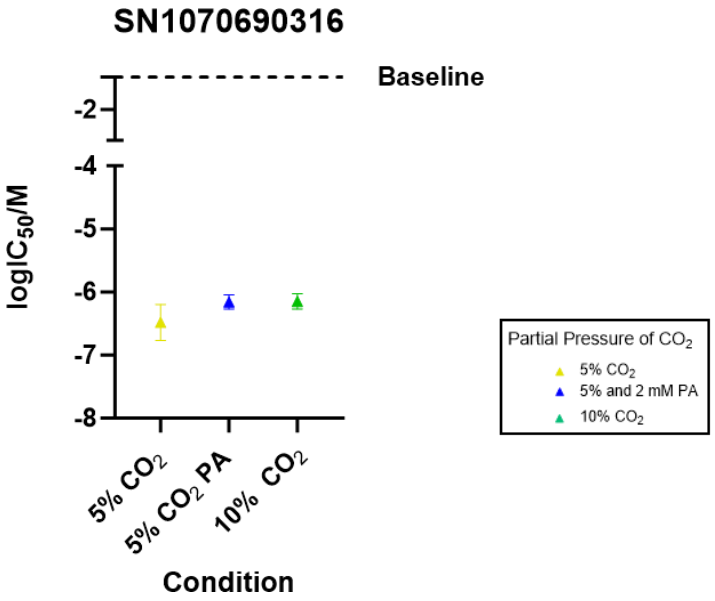


Figure 4-14 The $logIC_{50}$ plotted against the CO_2 experiment condition for each SMARCA2 targeting PROTAC labelled with an SN code as determined from the dose-response curve. The condition 5% CO_2 PA describes cells incubated at 5% CO_2 treated with 2 mM propanoic acid. All values are represented as the mean with error bars calculated from the standard deviation for the mean calculation of the $logIC_{50}$ values where n=3. The significance of the data was determined by one-way ANOVA and multiple comparison tests showed the individual relationships which were significant at a threshold of $p<0.05$.

4.10.4 Propagation of $logIC_{50}$ Error

As mentioned in sections 4.10.3 and 4.10.4, the standard deviation from calculating the mean $logIC_{50}$ value results in small error bars. However, the $logIC_{50}$ value assigned by Genedata is associated with a standard deviation, the statistic that considers the deviation of response values from the fitted DRC. The variance associated with the calculated $logIC_{50}$ is compound specific, as shown from different DRC fit examples in Figures 4-9 and 4-10. In Figure 4-10, the error bars associated with the $logIC_{50}$ are shown to be about one logarithmic unit for all the biological repeats. In contrast, those in Figure 4-9A are insignificant, and the error associated with the $logIC_{50}$ value in Figure 4-9C is outside the bounds of the measured concentrations. Therefore, this error must be propagated to find the variance within and between experiments as shown in Equation 4-6.

$$Propogated\ standard\ deviation\ = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}+\sum_{i=1}^{n}(\sigma_i)^2}$$

Equation 4-6 Calculation for the propagated error, where x is the $logIC_{50}$ for each data point, n is the number of biological replicates (3 in this experiment), μ is the mean of the $logIC_{50}$ values and σ_i is the standard deviation from the DRC fit of each curve.

The propagation of error calculation was applied to the BRD4 and SMARCA2 CO_2 datasets, and the data was replotted in Figures 4-15A and 4-15B, respectively. The data presented in Figure 4-15 is more variable than in Figures 4-13 and 4-14. This more accurately reflects the variability of the compound's DRC fit. BRD4 compounds have less inherent variability in their $logIC_{50}$ value than the SMARCA2 compounds.

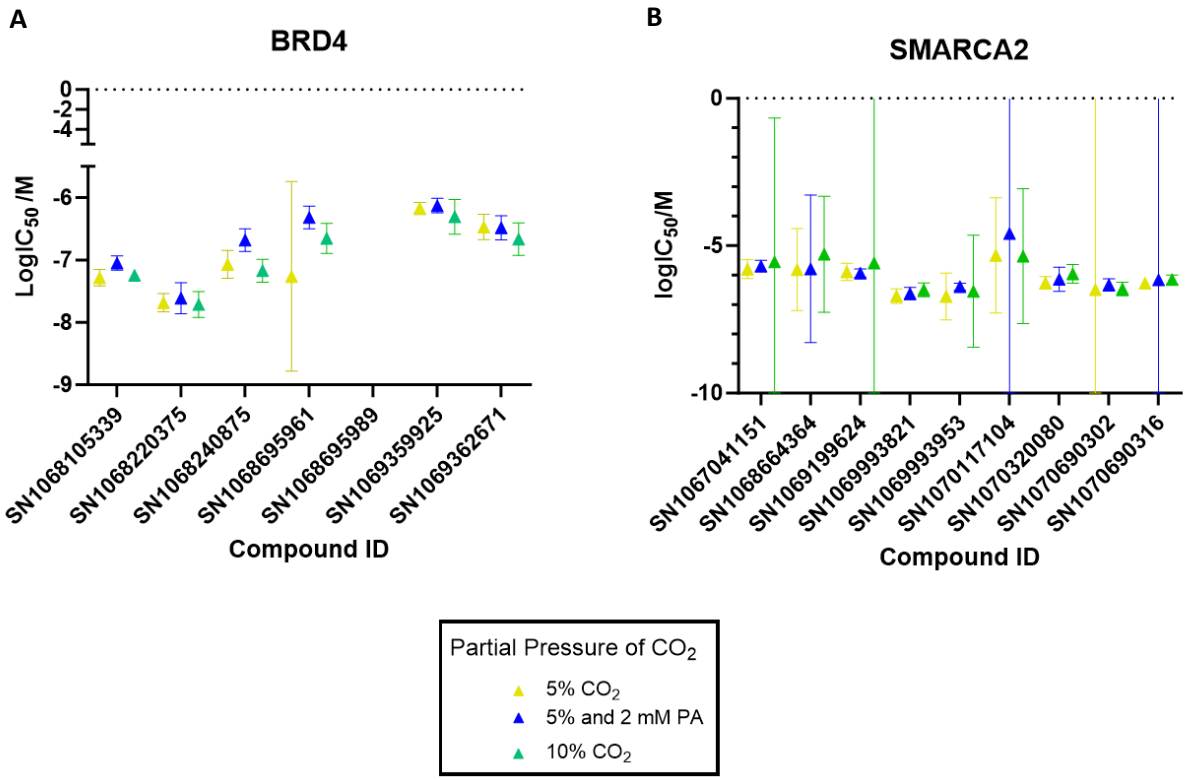


Figure 4-15 $LogIC_{50}$ values plotted against the compound ID for (A) BRD4 and (B) SMARCA2 tool compounds for the CO_2 experiment conditions summarised in the legend. All values are represented as the mean with error bars calculated from the absolute standard deviation from both the mean calculation and the DRC fit for n=3. The significance of the data was determined by one-way ANOVA and multiple comparison tests showed the individual relationships which were significant at a threshold of $p<0.05$.

Across the dataset, there was often one data point with a much higher standard deviation for the $logIC_{50}$ DRC fit than the others, and this could be excluded as an outlier. As Figure 4-13 showed, the BRD4 compound SN1069695961 had an outlier $logIC_{50}$ value arising from a shifted DRC; therefore, this data point was excluded in Figure 4-16A. In the SMARCA2 CO_2 experiment, the $logIC_{50}$ values associated with a high standard deviation from a singular replicate were regarded as outliers and removed to give Figure 4-16B. SN1070117104 was the weakly active compound, corresponding to having the highest variability, and there was more than one value with a high standard deviation

associated with DRC fit. The mean $logIC_{50}$ graphs for each compound, excluding inactive ones, are included in Appendix Figures 8-7 - 8-13D-E and 8-14 – 8-22D-E for BRD4 and SMARCA2, respectively.

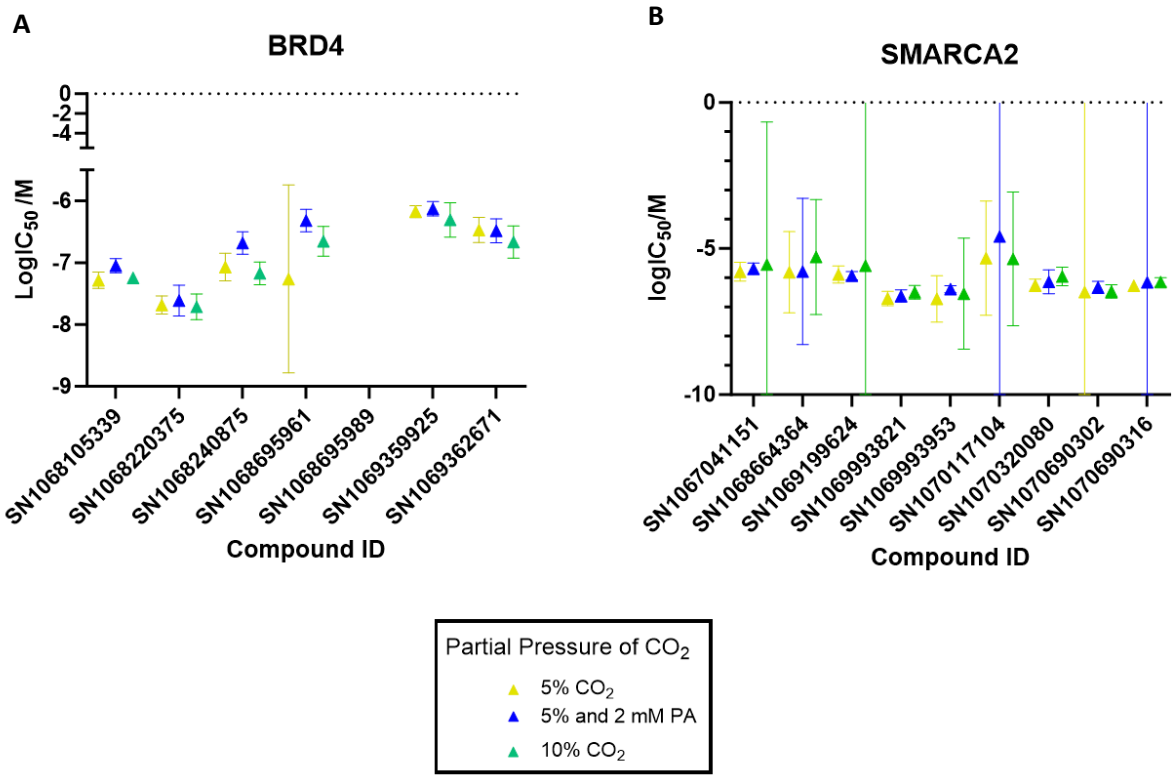


Figure 4-16 $LogIC_{50}$ values plotted against the compound ID for (A) BRD4 and (B) SMARCA2 tool compounds for the CO_2 experiment conditions summarised in the legend. All values are represented as the mean with error bars calculated from the absolute standard deviation from both the mean calculation and the DRC fit for n=3 with outliers removed. The significance of the data was determined by one-way ANOVA and multiple comparison tests showed the individual relationships which were significant at a threshold of $p<0.05$.

4.10.5 Statistical Analysis of the HiBit Assay Results

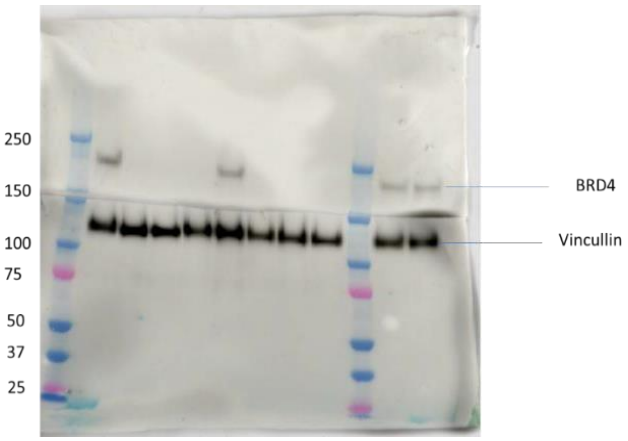
All the data passed the normality Shapiro-Wilk test, so it could be analysed using normal distribution. Statistical significance was assessed by a one-way Analysis of Variance (ANOVA), which compares the means of all three conditions to report whether they are statistically different. The variability used for the one-way ANOVA was the propagated standard deviation. Individual relationships between the conditions were assessed with a multiple comparisons test. Data concordance of the q logAC50 values between and within experiments were evaluated using Spotfire with y=x (solid line), and 95% confidence intervals (CI,dotted lines) were plotted. This data can be seen for the BRD4 compounds in Appendix Figures 8-23 – 8-28 and SMARCA2 compound in 8.29 – 8.34. The q $logAC_{50}$ value is used here to produce a value for the inactive compounds. However, compounds with a $logIC_{50}$ value will have the same $qlogAC_{50}$ value.

4.11 An Orthogonal Technique

A western blot was performed to create an orthogonal technique for the HiBiT lysis assay to assess the influence of CO_2 levels on PROTAC substrate degradation. The western blot in Figure 4-17 used four different concentrations of the PROTAC, SN1068220375, in HiBiT-BRD4-HEK293 cells under 5% and 10% CO_2. The PROTAC concentrations were chosen from each section of the DRC: 0 nM as a control, 7 nM (log -9.15 M) before the compound showed any response, 35 nM (log -7.45 M) in the linear portion of the DRC and finally 125 nM (log -6.9 M), for 100% inhibition.

The protocol was optimised by using different antibodies for the target BRD4 protein, altering the antibody concentration for BRD4 and vinculin as well as the number of washing steps and transfer time. The optimised western (Figure 4-17A) was clean with little background interference, suggesting an optimised process. Vinculin was probed to act as loading control for the cell lysate and is well separated from BRD4 by SDS-PAGE due to the size difference of the two proteins, as labelled in Figure 4-17A.

A



B

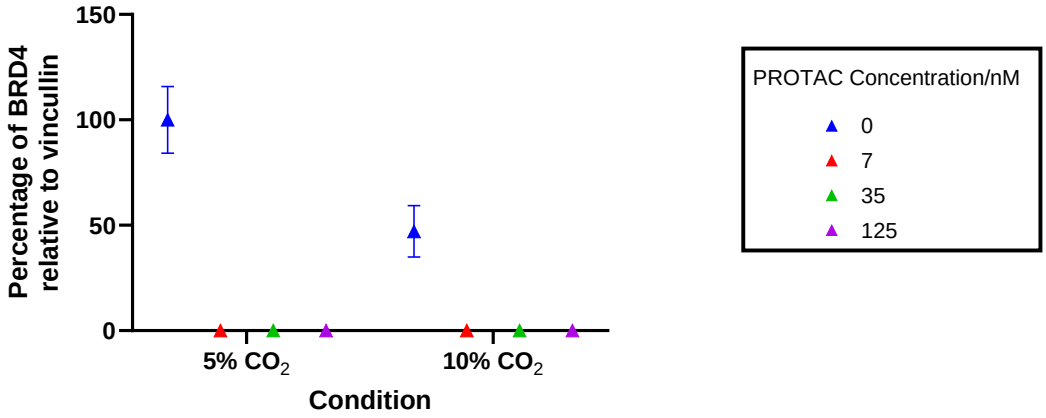


Figure 4-17 (A) Western Blot of SN1068220375 at 4 doses and two CO_2 concentrations. Lanes 1 and 10 are molecular weight ladders with weights specified in kDa. Lanes 2- 5 are treated at 5% CO_2 at varying concentrations of SN1068220375 from lane 2 at 0 nM (DMSO used instead), lane 3 at 7 nM, lane 4 at 35 nM and lane 5 at 125 nM. Lanes 6-9 are the same conditions but at 10% CO_2. Lanes 11 and 12 are technical repeats of 2 and 6, respectively. BRD4 is labelled with a molecular weight of 200 kDa, whilst vinculin is at 124 kDa. (B) The relative percentage of BRD4 normalised to vinculin plotted against the CO_2 treatment condition tested across a range of PROTAC concentrations as defined in the legend. All values are represented as the mean with error bars calculated from the standard deviation where n=2.

Figure 4-17B represents one biological sample in the same western blot; therefore, it is impossible to draw any significant conclusion from this result. However, a few main issues were highlighted in this analysis. Firstly, the relative density of the BRD4 bands showed a discrepancy

between the 5% and 10% CO_2 conditions when normalised to the vinculin band. Secondly, the standard deviation plotted indicates this method is variable within the same technical process. Therefore, between experiments, there could be large discrepancies. Finally, BRD4 bands were only seen in the DMSO control samples, which were less intense than the vinculin bands. This result was surprising because treatment with 7 nM of SN1068220375 resulted in no BRD4 detected on the blot. However, the dose-response curve from the HiBiT experiments consistently showed that this concentration was not causing degradation. If further experiments using this technique were required, perhaps the results could have been improved by using a higher concentration of the BRD4 antibody to detect the protein in the PROTAC-treated samples.

4.12 Discussion

The key finding from these experiments was that the incubation at hypercapnic levels of CO_2 had no significant effect on the activity of PROTACs compared with normal CO_2 levels. This is verified by the strong concordance in mean $logIC_{50}$ values for BRD4 and SMARCA2 (Figures 8-23 – 8-34). In addition, the technical error of the experiment is low, and instead, the majority of variance comes from dose-response curve fitting, which is compound dependent. Across both target datasets, when including and excluding the outliers, the only statistically significant finding was from the BRD4 dataset depicted with a * in Figure 4-16A. This was compound SN1068240875, which had a p-value of 0.049 for the one-way ANOVA, meaning it was only just below the threshold of 0.05 for being significant. A possible explanation for this result is that decreasing the intracellular pH is either due to PA treatment or associated acidosis of hypercapnia. This pH decrease could affect the protonation state of specific functional groups in the PROTAC compound, altering the association with the E3 ligase or degradation substrates. However, it is unlikely the pH change is drastic enough to change SN1068240875's mechanism of action because the value has only just passed the significance threshold.

Following the tool compound selection experiment, the HiBiT-tagged BRD4 substrate in HEK293 cells was chosen as the first target for the CO_2 experiment. The dose-response curves for the BRD4 compounds were reproducible with low variability (Figures 4-15 and 4-16 A), which gave concordant calculated $logIC_{50}$ values across replicates (Figures 8-23 - 8-25) and CO_2 treatment conditions (Figures 8-26 – 8-28). The BRD4 compounds identified outside the confidence interval lines of the concordance assessment were SN1068695961 and SN1068240875. The $logIC_{50}$ for SN1068695961 is far from the confidence intervals in Figures 8-23 A, B, 8-26 A and 8-27 B. This result relates to the wide error bar for this compound in the 5% CO_2 condition shown in Figures 4-13 and 4-15A and supports removing this data point as an outlier. SN1068240875 was just outside the CI for replicate 3, comparing the conditions of 10% and 5% PA (Figure 8-28A), but the $logIC_{50}$ was concordant

between replicates. This correlates with SN1068240875 being a statistically significant compound between treatment conditions in the CO_2 experiment**.**

Despite CO_2 having no influence in BRD4 targeting PROTACs, the investigation was continued with the HiBiT-tagged SMARCA2 substrate in the NCI-H838 cell line to ensure the effects were not cell line or substrate-specific. However, there were no statistically significant findings found in this dataset either. Despite the inherent variability of SMARCA2 compounds due to the DRC fits, the concordance assessment of the data showed a strong correlation between the $logIC_{50}$ values compared between the CO_2 conditions (Figures 8-32 – 8-34) and for biological replicates (Figures 8-29 – 8-31). In the SMARCA2 dataset, replicate three tends to have greater variability than the other repeats with more compounds outside the confidence intervals. One of the main reasons for the SMARCA2 compound's DRC fit variability was the hook effect or solubility issues at higher concentrations (Figure 4-10), leading to steep slopes on the DRC. The fit for these DRCs may have benefited from smaller incremental concentration increases to reduce plotting variability. In addition, the NCI-H838 cell line was more difficult to work with and grew slower than HEK293, which may have contributed to the variability in the SMARCA2 experiments.

The experimental aim of the western blot was to verify the HiBiT lysis assay results. However, no compound was identified in the HiBiT assay to have a marked difference in potency under the CO_2 conditions tested. The western blot was less sensitive, less precise, and had a higher variability than the HiBiT assay. When the HiBiT assay did not identify any significant findings, it was concluded that the western blot could not detect any change in activity that may arise from CO_2 incubation; therefore, this line of investigation was not pursued any further.

In summary, the SMARCA2-NCI-H838 and BRD4-HEK293 datasets prove that the HiBiT lytic detection system is a practical, reproducible method for measuring PROTAC DRCs. The DRCs were closely correlated between biological and technical repeats. Potency was chosen as the metric for comparison because it is a single readout value that is easy to interpret and describes the compounds'

structure-activity relationship (SAR). As explained previously, the drawback of using this metric is that the DRC fit can skew it.

4.13 Conclusion

This chapter discussed the application of the Nano-Glo HiBiT assay to PROTAC-mediated degradation of BRD4 and SMARCA2 in the context of CO_2. The key finding was that no significant difference was found for any of the compounds tested in either cell line. Overall, the data was robust, reproducible, and not cell line or substrate dependent. The effect of carbamylation on polyubiquitin chain formation at K48 was represented in a di-Ub conjugation assay where an approximate 12% decrease was found between 0 to 3.3 mM CO_2 (representing hypercapnia).[84] This infers that when PROTAC compounds are used to hijack the UPS, the effect of carbamate formation at K48 on polyubiquitin chain formation is too mild to influence these compounds. Furthermore, the VHL and CRBN E3 cullin ligases are expressed at sufficient levels not to inhibit PROTACs. In conclusion, this result indicates that PROTAC potency under physiologically relevant levels of normoxic and hypercapnic CO_2 is not significantly altered.

4.14 Future Work

As detailed in section 4.13, the DR of PROTACs is not dependent on CO_2. The diverse applications of PROTACs beyond VHL and CRBN E3 ligases are discussed in the review cited here.[196] However, PROTACs can potentially target many more E3 ligases and expand the number of disease targets. This is an important development in this field due to the varying expression of E3 ligases across cell and tumour types.[151] A clear conclusion was met in both disease targets for all the PROTACs tested under normal and hypercapnic levels of CO_2. Therefore, there is no future work for this investigation.

5. Histone Carbamylation and Transcription Regulation

5.1 Overview

Eukaryotic genomes are packaged by nucleosomes into chromatin. Chromatin is a complex structure which exists in two major forms called heterochromatin and euchromatin. Heterochromatin is highly condensed and contains genes which are rarely transcribed whereas euchromatin is involved in active gene transcription where DNA is converted to messenger RNA (mRNA). Nucleosomes consist of DNA wrapped around histone proteins and DNA transcription is regulated by a diverse range of histone PTMs as detailed by the histone code.[197] These PTMs work individually and together to alter the chromatin structure and induce specific gene expression. Altered gene expression is associated with elevated CO_2 levels and several mammalian signalling pathways are responsive to hypercapnia.[198] This chapter aims to identify the effects of carbamylation on DNA transcription and make a case for adding the carbamate PTM to the histone code.

In chapter three, seven nucleosome carbamylation sites were identified in both 12C and 13C trapped HEK293 lysates. From the seven sites identified, histone 3 lysine 79 (H3K79) was chosen as a model for studying nucleosome carbamylation due to the extensive research on mono, di and tri methylation at this site. H3K79 methylation is a transcriptionally active mark and is mediated solely by the enzyme disrupter of telomeric silencing like (DOT1L) in humans.[199] The specificity of this enzyme enabled the design of *in-vitro* and *in-cellulo* experiments targeted at understanding the effects of H3K79 carbamylation on DNA transcription.

The background to this chapter covers; the nucleosome structure, complexity of the histone code, the H3K79 methyltransferase DOT1L and the DOT1L inhibitor, pinometostat. In addition, CO_2-dependent transcription factors and signalling pathways are discussed. The results discussed in this chapter include native nucleosome trapping, recombinant histone octamer production and trapping, an *in-vitro* assessment of carbamylation on H3K79 using the methyltransferase Glo (MTase-Glo) assay

and the development of a RNA-sequencing screen for assessing transcriptional change related to DOT1L inhibition under normal and disease state CO_2 conditions.

5.2 Nucleosome Structure

Nucleosomes are the subunits of chromatin, which consist of 146 base pairs (bps) of DNA and two copies of each core histone protein, namely, Histone 2A (H2A), Histone (H2B), Histone 3 (H3), Histone 4 (H4) as depicted by Figure 5-1. One-fifth of the amino acids in histone proteins are arginine and lysine residues which are positively charged at physiological pH resulting in the basic property of the histone octamer. The backbone of DNA contains alternating deoxyribose sugar and negatively charged phosphate groups. These opposing charges lead to the association of these biomolecules by electrostatic interactions. Externally, there is a linker Histone 1 (H1) bound at the DNA entry and exit sites of each nucleosome which results in further compaction.[200]

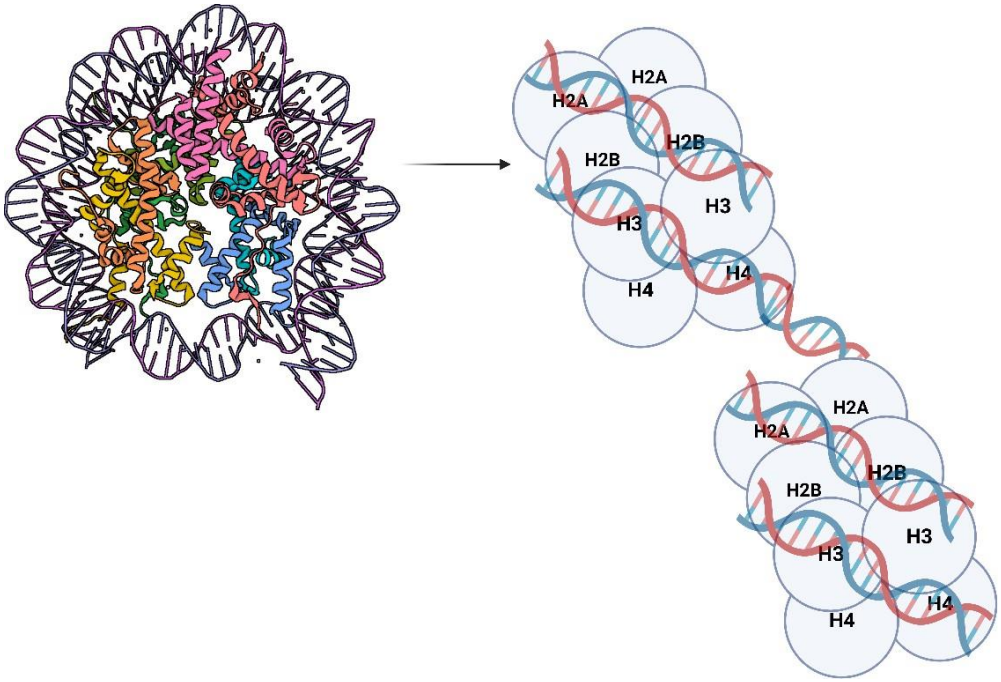


Figure 5-1 The crystal structure of a nucleosome illustrating 146-base pairs of DNA bound to the histone octamer (PDB, 2CV5 [201]), alongside a schematic to depict adjacent nucleosomes joined together by linker DNA created using BioRender.

Small changes in the amino acid composition of core histone proteins give rise to several histone variants. The expression of these variants ranges from uniform to tissue specific. Histone variants have evolved to exhibit distinct effects on the nucleosome structure and function in DNA-mediated processes. Furthermore, histone variants have been subdivided into canonical and non-canonical, characterised by their expression being replication dependent and independent, respectively. During replication-independent chromatin assembly, the epigenetic state can be changed by the exchange of histone variants which can alter or erase the pattern of PTMs.[202] A detailed overview of the mammalian histone variants and their functional characteristics is outlined elsewhere.[203–205]

5.3 The Histone Code

DNA transcription is tightly regulated by a wide range of PTMs covalently bound to histone proteins. Histone PTMs are mediated by enzymes which are grouped into writers, readers, and erasers. Writers catalyse the addition of a specific PTM, readers recognise PTM sites via a specific domain and erasers catalyse the removal of a PTM.[206] The histone code refers to the vast number of possible PTMs that can occur on histone proteins, arising from the multitude of possible types, modification sites, and histone variants.

Many histone PTMs have been identified for example, acetylation, methylation, ubiquitination, and phosphorylation. The most frequently modified areas of the nucleosome are the N and C termini of all four histone variants which lie outside the nucleosome core. These are highly flexible regions of the structure otherwise known as histone tails. Histone PTMs are epigenetic markers associated with effects on histone-DNA and histone-histone interactions which continue to be an evolving area of research. Detailed reviews [207–209] and machine learning approaches [210] which analyse the histone code and its significance in genomic function can be referred to for further information.

Histone PTM patterns associated with certain genomic regions have been identified with chromatin immunoprecipitation sequencing (CHIP-seq).[211] Acetylated histones correlate with

increased transcription because the acetyl group neutralises the positive charge on lysine residues and weakens the DNA-histone interaction leading to transcriptional activation.[212] The effects on transcription due to histone methylation are residue and degree (mono, di, tri) dependent as detailed elsewhere.[213] Methylated H3K9, H3K27 and H4K20 are heterochromatin markers whilst methylated H3K4, H3K36 and H3K79 are euchromatin markers. Histone phosphorylation is known to play roles in DNA repair, transcription, apoptosis, and cell cycle progression. [214] Mono-ubiquitination is most common on H2A and H2B and this covalent modification plays a key role in the DNA damage response. [215] Over the past decade, several biologically relevant histone PTMs for example, monoaminylation, glutarylation and glycation have been added to an expanding network of epigenetic modulators of chromatin, illustrating the dynamic nature of this research area.[216]

5.4 Histone Code Complexity and Crosstalk

Histone PTM crosstalk can be described as the coordination between two or more PTMs which work together to alter the transcriptional outcome. Histone PTM crosstalk can take many forms, three of which are discussed here.

Firstly, the modification of a histone residue stimulates a writer enzyme of another residue between and within histone proteins. Examples include the phosphorylation of H3S10 which stimulates the acetyltransferase, Gcn5 to acetylate H3K14 [217] and H2B mono-ubiquitination which stimulates H3K4 methylation by COMPASS [218] and H3K79 methylation by DOT1L.[219] The modification of a histone residue can also have the opposite effect, whereby the PTM can prevent the recognition of another residue by its modifying enzyme. For example, H3R2 methylation inhibits the methylation of H3K4 by COMPASS.[220]

Secondly, several epigenetic modifiers associate with complexes working in tandem to alter the chromatin structure. For example, the MLL3/4 Set1-H3K4 methyltransferase complex catalyses the removal of the repressive H3K27 methylation mark whilst methylating at H3K4 to activate transcription.[221] Further examples of these complexes can be found elsewhere.[222]

Lastly, the flexible tails of the histone octamer are subject to irreversible proteolytic cleavage due to the abundance of cleavage sites. Histone degradation particularly for H3, has been linked to the processes of infection, cell differentiation and aging through gene expression regulation.[223] The activity of the protease cathepsin L, which cleaves H3 is influenced by PTM crosstalk. It was found that pan-acetylation of lysine residues reduced cleavage whilst H3K27me2 increased cleavage.[224]

5.5 Histone H3

In *Homo Sapiens*, four H3 variants have been identified, namely, H3.1, H3.2, H3.3 and H3t. The canonical variants are H3.1 and H3.2, whilst H3.3 is defined as the centromere-specific variant and H3t is localised to the testes.[225] These variants have high sequence homology and the potential to exhibit similar PTMs. However, H3.1, H3.2, H3.3 and H3t, show distinct profiles along the genome alongside varying PTM enrichment indicating that each variant has a distinct biological function.[226]

H3 undergoes the highest number of modifications across the histone family, due to having the most lysine residues. Figure 5-2 shows a handful of H3 methylated residues and their transcriptional effects. In this study, the primary site of interest is H3K79 which can be modified by methylation either in the mono, di or tri state and this modification has been linked to active transcription. Mass spectrometry studies on various cell lines have shown that the majority of H3K79 is unmethylated (70-95%), mono-methylation is the most prevalent methylation state at 6-30%, followed by di-methylation at 0.1-10% and tri-methylation is undetectable or extremely low at levels less than 0.1%.[227]

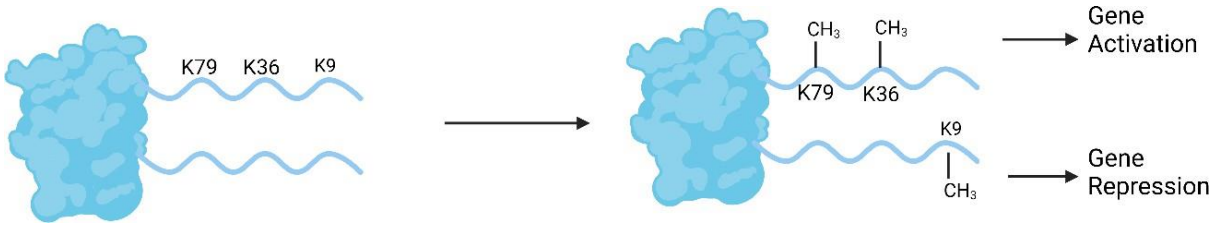


Figure 5-2 Transcriptional changes exhibited upon methylation at different H3 residues created using BioRender.

5.6 The DOT1L Complex

DOT1L exhibits distinct properties when compared to other histone methyltransferases (HMTs). Firstly, H3K79 is exclusively methylated by DOT1L, whilst other histone modification sites are subject to modification by multiple HMTs, for example, H3K4 can be methylated by 10 known enzymes.[228] Secondly, DOT1L is only active on nucleosomes and does not modify free H3 proteins or peptides which could be due to PTM crosstalk.[229] Lastly, DOT1L does not contain a SET domain which is conserved between other HMTs.[227]

The mechanism of DOT1L-mediated methylation requires the substrate, S-adenosyl methionine (SAM) as a methyl donor which is converted into S-adenosyl-L-homocysteine (SAH). DOT1L is a distributive methyltransferase, to establish the different degrees of methylation, the enzyme repeatedly binds and dissociates from H3, it has been shown that the transfer of SAM to SAH requires enzyme dissociation.[230,231]

Methylation at H3K79 depends on histone crosstalk. Mono-ubiquitination of H2BK123 stimulates DOT1L activity, by stabilising the enzyme's interaction with the nucleosome as determined by cryogenic electron microscopy (CryoEM).[232] H4K16 acetylation is also essential to the distribution of DOT1L-mediated H3K79 methylation.[233] DOT1L and the Silent Information Regulator complex (SIR3), which assembles heterochromatin, bind to the same site on molecular chromatin. CHIP analyses identified that H4K16 acetylation displaces SIR3 which enhances the binding of Dot1L and reduces SIR3 spreading.[234]

DOT1L can associate with several fusion partners, either as part of the DOT1L complex (DotCom) or through single protein-protein interactions with *c-Myc*, Bat3, and RNA polymerase II (RNA pol II) as depicted in Figure 5-3.[227] The fundamental subunits of DotCom are the ALL1-Fused gene from chromosome protein 10 (AF10), AF9 or Eleven Nineteen-leukaemia (ENL), and DOT1L. Additional non-essential subunits, include AF17 and proteins from the Wingless-related integration (Wnt) pathway. In the absence of AF10, only mono methylation at H3K79 is observed therefore it is likely that AF10

localises DOT1L to certain areas of the genome to increase the degree of methylation.[235] AF9/ENL contains a YEATS domain which is critical for chromatin association.[236] The oncogene*, c-Myc* has been associated with DOT1L to form an activating complex with the histone acetyltransferase CBP/p300 which can accelerate breast cancer progression. Bat3 appears to assist Dot1L's interaction with H3 as it interacts with both proteins. DOT1L interacts with the phosphorylated C terminal domain (CTD) repeats of the RNA polymerase II which may enhance DOT1L-mediated methyl transfer in transcriptionally active regions.[237]

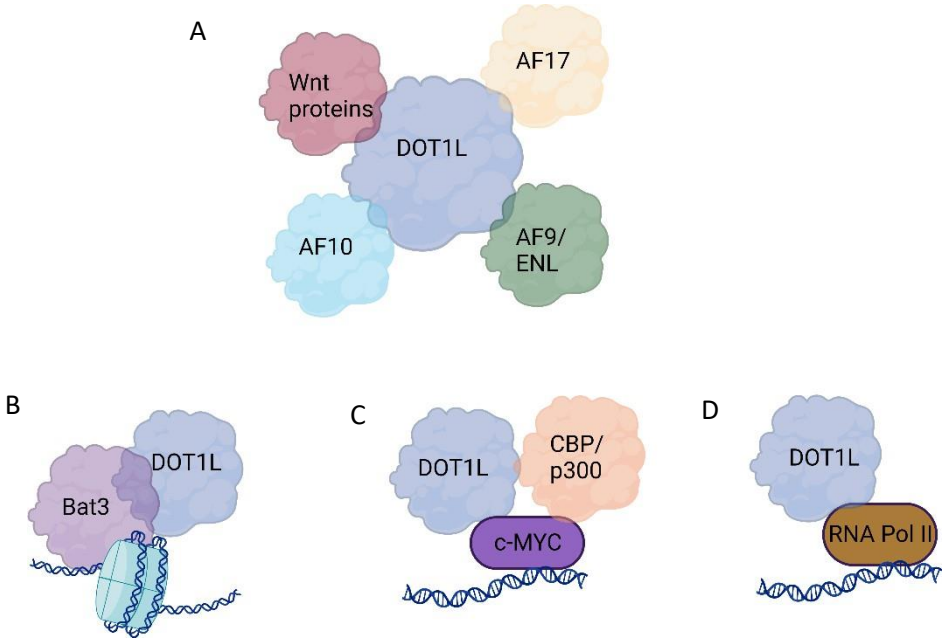


Figure 5-3 DOT1L fusion partners, figure adapted from[227] (A) DotCom composed of five interacting subunit proteins. (B) Bat3 mediates the interaction of DOT1L with the nucleosome. (C) Activating complex formed at the *c-MYC* gene. (D) The phosphorylated C terminus of RNA polymerase II interacting with DOT1L in actively transcribed areas of the genome created using BioRender.

5.6.1 DOT1L Downstream Targets

Methylation of H3K79 engages in a diverse range of cellular processes including the regulation of telomeric silencing, transcription, cellular development, cell-cycle checkpoint, DNA repair and thermogenesis.[238] An overview of the biological role of DOT1L and H3K79 methylation is discussed below.

DOT1 was first discovered in yeast by a genetic screen looking for genes causing defects in telomeric silencing.[239,240] DOT1 and the mammalian equivalent DOT1L play a regulatory role in telomeric chromatin structure. A deficiency of Dot1L has not only shown a global loss of H3K79 methylation but also associated loss of heterochromatin markers at telomeres and centromeres.[241] Interestingly, the overexpression of DOT1L also reduces telomeric silencing.[242] DOT1L's role in telomeric silencing has not been fully elucidated however these findings suggest that there is an interplay between Dot1L and the SIR. The SIR complex mediates the assembly of heterochromatin domains, and Dot1L-mediated methylation at H3K79 affects the assembly of these domains by restricting the SIR's access to chromatin at certain genomic regions.

There are four phases to mitotic cell division, growth is the first stage known as the G1 phase, followed by DNA replication in the S phase, the third stage is a secondary growth phase known as G2 and finally the cell splits into two daughter cells known as the M phase. Throughout the four stages of the cell cycle, the levels of each H3K79 methylation type (mono, di and tri) fluctuate between species and cell lines. [227,240] These findings suggest that the role these modifications play in cell cycle regulation is not conserved across species and could be cell-type specific.[238] The first study to link H3K79 methylation with DNA replication start sites was whole genome sequencing of human cancer cells. The researchers found that DOT1L depletion resulted in genomic over replication and it was proposed that H3K79 methylation acts as a marker to ensure replication only occurs once per cell cycle.[243] DOT1L mutant studies have been performed in various human cell lines. For example, a study using erythroid progenitor cells showed that DOT1L knockouts experience G1 cell cycle arrest [244] whereas a study in

embryonic stem cells (ES cells) showed that DOT1L mutants experience G2 cell cycle arrest. [245] These studies show that DOT1L has diverse roles in cell cycle regulation.

Chromatin assembly factor 1 (CAF1) is a histone chaperone important for DNA regulation and repair. CAF1's major role is to deposit histones onto DNA to form nucleosomes. Zhou *et al*. investigated the role of Dot1p in yeast in the context of CAF1 and DNA damage. Mass spectrometry identified methylated H3K79 as a binding site for CAF1. Double mutants lacking Cac1 a subunit of CAF1 and Dot1p were more sensitive to DNA damaging agents than single mutants suggesting these two proteins interact in response to DNA damage.[246]

Brown adipose tissue (BAT) regulates thermogenesis, which can be defined as the process which creates heat to regulate body temperature. BAT is rich in mitochondria which contain the uncoupling protein 1 (Upc1), which disrupts ATP synthesis to generate heat instead. Zc3h10 is a cold response transcription factor, enriched in BAT tissue which activates the Upc1 promoter. Yi *et al*. showed that Dot1L is an interacting partner of Zc3h10 which drives transcriptional activation of thermogenic genes. [247]

Finally, Dot1L is associated with Wnt signalling genes and Mohan *et a*l. previously linked the knockdown of Dot1L to a reduction in Wnt signalling.[248] However, Gibbons *et al.* investigated this hypothesis using the Wnt pathway inducible HEK293 cells and human colon adenocarcinoma cell lines by inhibiting DOT1L.[249] Wnt gene expression was not altered in this study when cells were treated with and without the Dot1L inhibitor (EPZ004777). This work concluded that H3K79 methylation is not essential for the canonical Wnt signalling pathway, either for maintenance or activation of the Wnt pathway targeting gene expression.

5.6.2 DOT1L Inhibition

Aberrant expression of DOT1L has been linked with a range of cancer types.[250] Mixed lineage leukaemia (MLL) has been linked to DOT1L through the interaction with MLL proteins including AF4, AF9, AF10 or ENL. When DOT1L is overexpressed, the MLL genes, Homeobox A9 (*Hoxa9*) and Myeloid ecotropic viral integration site 1 (*Meis1*) are actively transcribed.[251,252] The first H3K79 de-methyltransferase was not identified until 2018 [253] therefore the inhibition of DOT1L has been explored as a primary therapeutic option for cancer treatment. DOT1L inhibition is specific and has few off-target effects because it is the only methyltransferase that does not contain a SET domain. The first small molecule DOT1L inhibitor that successfully inhibited the proliferation of MLL cells was EPZ004777. EPZ00477 is a potent selective compound however it is limited by poor pharmacokinetics thus unsuitable for clinical use.[254] EPZ00477's mechanism of action was used in the development of a more clinically relevant inhibitor EPZ -5676 otherwise known as pinometostat. The efficacy, potency, bioavailability, and clearance properties of this compound are desirable [255] and led to the use of pinometostat in clinical trials to treat MLL leukaemia.[256] EPZ-5676 (pinometostat) induces a conformational change by competing for the specific SAM-binding pocket of DOT1L. In this work, pinometostat is used as a DOT1L inhibitor.

5.7 Transcriptional Changes under Hypercapnia

In chapter one, hypercapnic acidosis and the cellular response mediated by pH and CO_2 sensing were discussed. This section (5.7) covers, key CO_2-dependent transcription factors and their role in signal transduction. Several mammalian networks are responsive to elevated CO_2 including the NF-κB and Wnt signalling pathways.

NF-κB activation is comprised of two signalling pathways; canonical and non-canonical which have both been implicated in the cellular response to hypercapnia. Both pathways are dependent on NF-κB inducing kinase (NIK) and inhibitory-κB kinase (IKK) phosphorylation which lead to nuclear translocation of NF-κB dimers which interact with DNA to activate immune response genes. The inflammatory genes: *ICAM1* and *IL-8* are markedly decreased in hypercapnia due to dysregulation of these pathways.[47] The exact mechanism of action by CO_2 on these networks is covered in detail in the cited reviews.[5,33,257]

A multispecies transcriptomics study was conducted by Shigemura *et al*. to measure hypercapnia-induced genomic responses. A key finding from this study was that hypercapnia activates genes involved in the Wnt signalling pathway and that these effects are conserved across flies, nematodes, and humans.[258] This study showed that the length of time cells and tissues are exposed to CO_2 impacts the number of differentially expressed genes (DEGs). This thesis uses the Shigemura *et al.* study as a basis for selecting the CO_2 sensitive, Wnt signalling genes; Frizzled class receptor 9 (*Fzd9*) and Wnt family member 7A (*Wnt7a*) for qPCR reactions designed to detect transcriptional changes due to CO_2 exposure. These genes were selected because they had been robustly identified by Shigemura *et al.* to be upregulated under CO_2 exposure in both mouse tissues and cells.

Hypercapnic patients are susceptible to bacterial infection and compromised wound healing. Interestingly, elevated CO_2 levels have been linked to immune response gene expression changes via the NF-κB pathway which is described in further detail in Chapter 1. Despite this until recently, the

global pH-independent transcriptional change induced by hypercapnia in immune cells was poorly understood. In 2023, Phelan *et al*. researched the transcriptional change induced by elevated CO_2 in monocytes and macrophages.[259]

Basal and inflammation-exposed immune cell lines were assessed for CO_2-dependent transcriptional change. Hypercapnia elicits a transcriptional change which is significantly enhanced by inflammation exposure. Transcripts related to the mitochondria, lipid metabolism and glycolysis were identified by RNA-sequencing as deregulated due to elevated CO_2. Cellular assays were used to support RNA-sequencing data including qPCR and an MTT assay which indicated that mitochondrial function is dependent on CO_2 levels.[259]

5.8 Initial Carbamate Screening on Native Nucleosomes

The proteomic screen discussed in Chapter 3 showed multiple carbamylation modifications on histone proteins including H1K46, H1K85, H1K106, H3K79, H3K123, H4K32, H4K92 which were identified in both the 12C and 13C datasets. H1K63 and H3K57 were also identified in the 12C and 13C datasets, respectively. The spectra for these carbamylation hits are plotted in Figures 8-35 – 8-37. Following these results, a targeted approach using native nucleosomes extracted from HEK293 cells was implemented. Figure 5-4 shows the four histone proteins which make up the nucleosome alongside the contaminants which were also extracted using the method outlined in section 2.5.1. Table 5-1 displays the molecular weight (Mw) of each histone in kilodaltons (kDa) which was used to identify each variant on the SDS PAGE gel in Figure 5-4.

The proprietary kit contains three buffers which are used for native nucleosome extraction. Firstly, the pre-lysis buffer was at pH 6.5-7.5 to break down the cell membrane and release nuclei. Next, the lysis buffer used was strongly acidic at pH 2 to release the histone proteins, and finally, the balance buffer was used to neutralise the pH of the extracted histones.

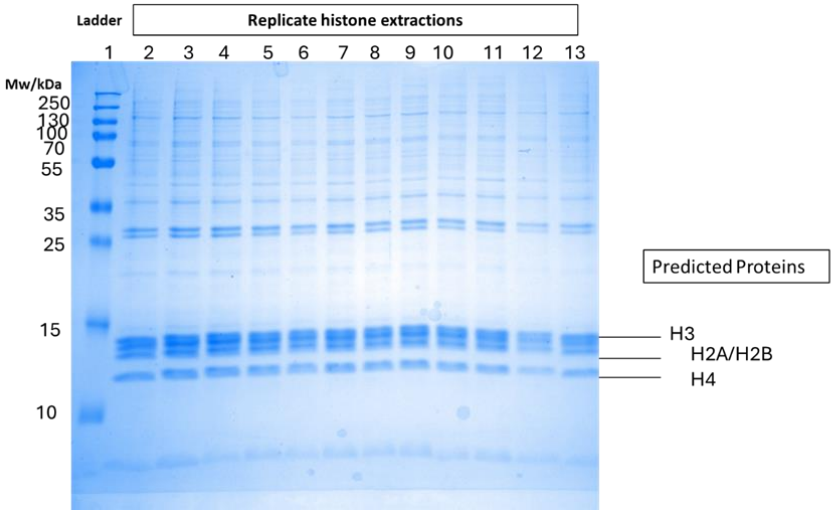


Figure 5-4 SDS-PAGE of predicted native nucleosomes extracted from HEK293 cells. Lane 1 is a molecular weight marker kDa with weights listed beside it. Lanes 2-13 were loaded with replicate histone extractions from HEK293 cells, passage 14.

Histone	Molecular Weight/ kDa	Isoelectric point (pI)
H2A	13.99	11.05
H2B	13.77	11.13
H3	15.27	10.31
H4	11.24	11.36
Histone Octamer	108.42	11.01

Table 5-1 The molecular weight and isoelectric point (pI) for histone proteins.

The native nucleosomes were run through the trapping process using ^{12}Ci concentrations of 20 and 50 mM with TEO and a negative control without TEO or inorganic carbon. Table 5-2 summarises the results obtained from this experiment including the coverage of each histone variant and the carbamate hits. The mass spectra for identified hits from this initial experiment included H1K85, H1K63 and H4K32 as shown in Figure 5-5 which supports results from the HEK293 proteome lysate screen.

Sample	Histone variant coverage /%	High-confidence carbamate histone hits identified
No_TEO_no_bicarbonate_1	H1.2 – 39 % H1.4 – 41% H1X- 11% H2A.X- 43% H2A.1 – 24% H2.A–8% H2A.Z- 31% H4 – 58%	None
No_TEO_no_bicarbonate_2	H1.2- 39% H1.4- 41% H1X-23% H2A.X- 44% H2A.1 -31% H2A.Z-31% H3.3 -51% H4- 68%	None
No_TEO_no_bicarbonate_3	None	None
TEO_20mM_bicarbonate_1	H4-68%	H4K32
TEO_20mM_bicarbonate_2	H1.2- 46% H1.4- 41%	H1.2/1.4K63 H1.2/1.4K 85
TEO_20mM_bicarbonate_3	None	None
TEO_50mM_bicarbonate_1	H1.2- 39% H1.3-37% H1.4- 47% H4-63%	H1.2/H1.4K85, H4K32
TEO_50mM_bicarbonate_2	H1.2- 45% H1.3 – 38% H1.4- 47% H1.5 -23 % H4 – 69%	H1.2/H1.4K85, H1.5K88
TEO_50mM_bicarbonate_3	None	None

Table 5-2 An initial native nucleosome carbamate screen, displaying coverage for specific variants and the hits identified. The histone is listed, when it passed FDR of 1% thresholding and at least two unique peptides were assigned to the protein.

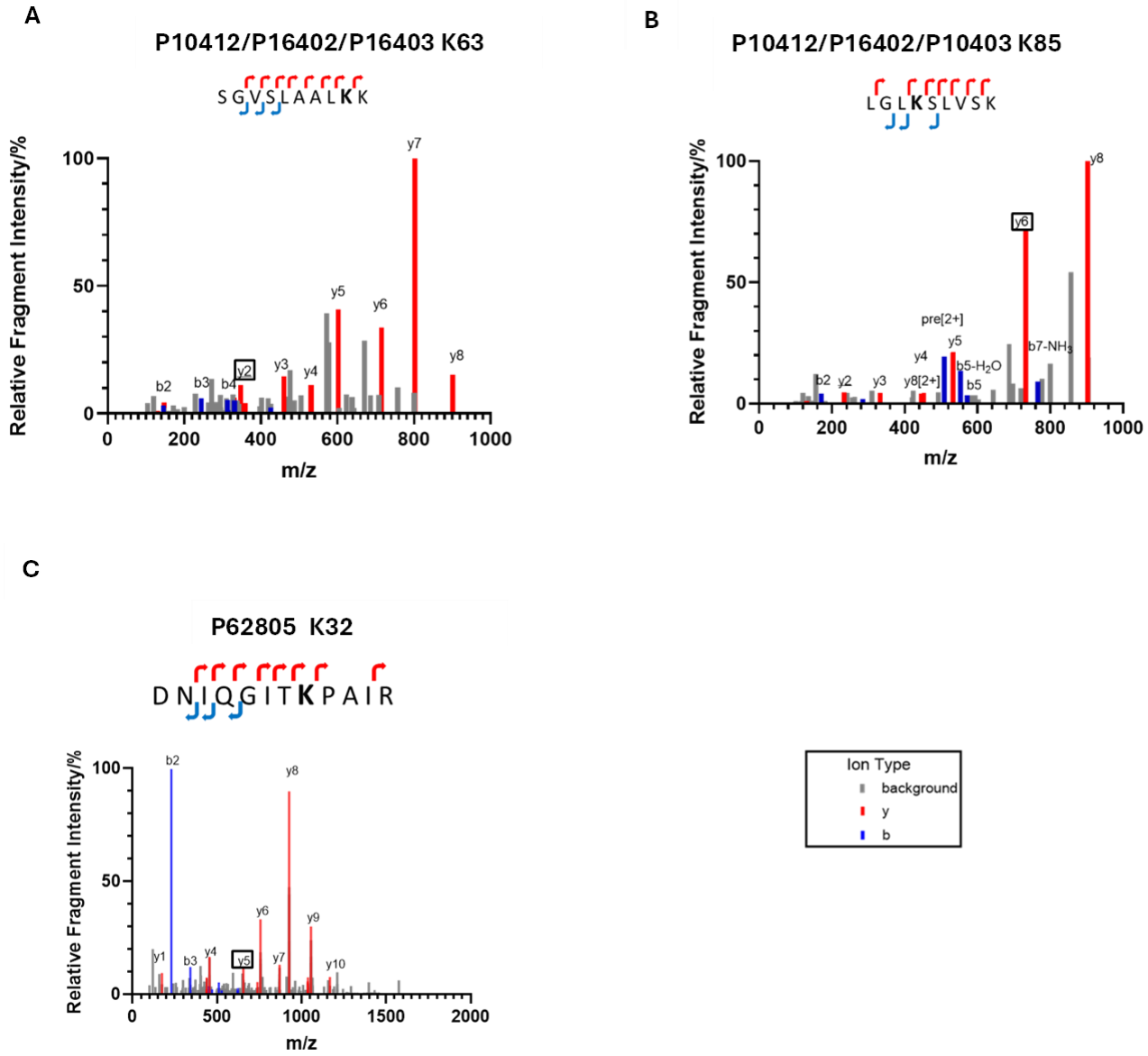


Figure 5-5 Identification of carbamate histone hits from the native nucleosome screening. Plots of relative fragment intensity versus m/z from LCMSMS identifying trapped carbamates on histones, (A) H1.2/1.3/1.4K63 (P10412/P16402/P16403), (B) H1.2/1.3/1.4K85 (P10412/P16402/P16403) and (C) H4K32 (P62805 K32) in the presence of $^{12}CO_2$. Each spectrum is associated with a peptide sequence illustrating the identification of predominant y (red) and b (blue) ions. The grey peaks represent background ions, and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

Sample preparation was consistent across the dataset in Table 5-2 however the histone variants identified were inconsistent. The discrepancies in histone coverage across replicates are due to sample contamination arising from both the protein extraction phase and sample handling. Contaminants and histone proteins exist in different proportions in each sample therefore the ratio of trypsin to a specific histone variant may not be optimal leading to the histone variant not being identified.

5.9 Improving Coverage of Nucleosomes

Certain histone proteins are represented better than others in the initial screening dataset, for example, H1 and H4 which have coverages of ~40% and ~60%, respectively. The observed coverage of histone proteins is linked to the sample preparation step and the amino acid sequence of each histone variant. As discussed, histones are highly basic due to a high composition of lysine and arginine residues for DNA binding. These residues are also trypsin cut sites leading to short peptide fragments that do not separate as efficiently by LC when cut with trypsin. To improve overall histone coverage, two approaches were trialled; changing the protease type to Arg-C or Lys-C to extend peptide fragment lengths to improve coverage on untrapped samples and reducing the number of trypsin cut sites by lysine propionylation.

5.9.1 Protease Choice

Table 5-3 details the identified histone proteins using three different proteases in the peptide preparation step. As the names suggest, Arg-C only cuts at the C terminus of arginine, whilst Lys-C only cuts at the C terminus of lysine and trypsin cuts at the C terminus of both arginine and lysine sites.

Protease	Histone Coverage
Arg-C	None
Lys-C	H1.2- 18% H2A- 27% H3.1/2/3- 11% H4 – 24%
Trypsin	H2A-38% H3- 15% H4 -36%

Table 5-3 Coverage of identified histones when using different proteases, n=1.

Arg-C did not identify any histone variants, whilst Lys-C identified H1, H2A, H3 and H4 however the coverage was lower than when digesting with trypsin. In addition, if Lys-C was applied to a trapped sample the number of possible sites for Lys-C activity would reduce due to lysines being modified by carbamylation or ethylation. This result showed that Arg-C and Lys-C were inefficient at cleaving the Arg-C/Lys-C cut sites to improve coverage therefore trypsin remained as the protease of choice for histone peptide preparation.

5.9.2 Propionylation

Propionylation of histones at lysines was used in a previous study as a successful technique to improve histone coverage.[260] The cited study makes the case for two rounds of propionylation before and following digestion for optimal results. Figure 5-6A outlines the mechanism for propionylation, the reagent propionic anhydride reacts with primary amino groups at the protein's N terminus and lysine sites. Propionic anhydride (PA) treatment is followed by hydroxylamine (HA) treatment to remove any unspecific propionyl groups at serine and threonine, Figure 5-6B displays the acyl group removal. In this investigation, it is important to consider the reaction conditions in the context of carbamates. Carbamate-modified lysine or N-terminal functional groups are not nucleophilic therefore propionic

anhydride will not react there. It was hypothesised that HA treatment would not cause carbamate dissociation, because the carbamate is more stable than the ester formed from PA reacting at a hydroxyl (OH) group and HA is too weak a base to dissociate a carbamylated group. The peptide synthesis literature was consulted because carbamate groups are used in the field for primary amine protection. Carbamate protective groups used in peptide synthesis are commonly attached to bulkier R groups than the ethyl group from TEO. Orthogonal deprotection strategies are detailed in the literature to deprotect specific sites during multistep synthesis and base deprotection occurs using an acidic hydrogen on the peptide with stronger bases such as piperidine (pK_b ~4) which supports the hypothesis that the carbamate modification would be stable to HA (pK_b ~8) treatment.[261]

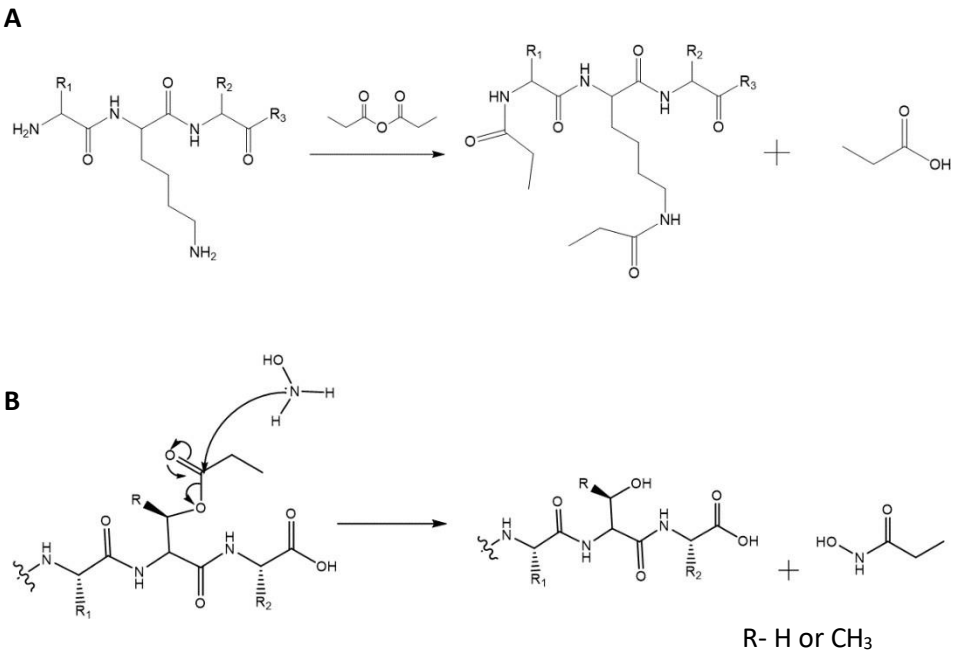


Figure 5-6 The two-step process for the preparation of propionylated histone peptides (A) propionic anhydride reacting with primary amino groups and (B) hydroxylamine treatment to remove unspecific propionyl groups from serine (R= H) and threonine (R=CH_3) amino acids.

A dataset comprised of non-propionylated (unmodified) and propionylated (modified) samples, tested in triplicate at 12 and 13C Ci concentrations of 0, 20, and 50 mM were trapped with TEO. Control samples were prepared with 0 mM Ci without TEO. Coverage tables for the propionylated

and non-propionylated samples are included in the appendix for each histone variant. (Table 8-3 – 8-6). This data is summarised by Table 5-4 which shows the average coverage percentage for each histone variant across the dataset; the numbers in brackets in the table accounts for how many times the histone variant was found across the total number of samples. For example, in the 12C unmodified sample set, H2B has a high average coverage at 70% but was only seen twice across twelve samples.

	Average coverage for each histone variant when ID /%			
Histone Variant	12C unmodified	13C unmodified	12C modified	13C modified
H1	30 (12/12)	28 (6/6)	8 (1/12)	10 (2/6)
H2A	25 (12/12)	31 (6/6)	21 (10/12)	29 (4/6)
H2B	70 (2/12)	26 (1/6)	0 (0/12)	0 (0/6)
H3	35 (10/12)	24 (5/6)	34(6/12)	40 (6/6)
H4	59 (12/12)	29 (5/6)	47(12/12)	56 (6/6)

Table 5-4 The average coverage of each histone variant across the sample set. Unmodified and modified relate to propionylated and non-propionylated samples. The first number is the coverage found as a percentage whilst the number in brackets is the number of times the variant was identified across the total number of samples in the dataset. For simplicity, the average coverage for each histone was derived from all the variants identified e.g., H1.2, H1.3, H1.4 and H1x and matched to the larger subgroup e.g. H1.

H4 followed by H3 are the best-represented histone variants, both in terms of being identified in the highest number of samples and the percentage coverage. Histone H2B was the most challenging histone to identify across this dataset. Importantly, this data showed that propionylation did not improve coverage as suspected. The reason for this may be that trapping modifies the number of potentially modifiable lysines influencing the effectiveness of propionylation. In addition to this, the native nucleosome preparation contains a high proportion of impurities and contaminant peptides which interfere with the identification of histone variant peptides.

Following coverage assessment, the carbamate identification in samples was assessed and data is shown in Table 5-5. From this table, most hits are seen in the unmodified samples and the 12C dataset. It can be said with high confidence, that H4K32 and H4K92 are real sites as they were identified in two distinct datasets, namely the HEK293 lysate screening and the native nucleosome propionylation/ non-propionylation dataset with both 12 Ci and 13 Ci. Carbamates on H1K46, H1K85, and H3K57 were also identified across both of these datasets however in the native nucleosome propionylation/non-propionylation dataset these hits were only seen with 12 Ci. H1K63 was seen in all three datasets, namely the HEK293 lysate screening, the initial native nucleosome screen and the propionylation/ non-propionylation dataset however in all three datasets this site was only seen in 12 Ci. The H3 carbamate hits of H3K79 and H3K123 identified by the HEK293 screening were not seen in either of the native nucleosome screens, however, coverage of this region was limited in the native nucleosome trapping. Figure 5-7 displays the spectra for the H1K90 carbamate site that was only identified in the propionylation versus non-propionylation LCMSMS dataset.

Carbamate Site	Number of Samples ID	Carbon concentration/ mM	Propionylation	Carbon isotopes
H4K32	6	Both	Both	Both
H4K92	3	Both	Unmodified	Both
H1.2/1.3/1.4K63	4	Both	Unmodified	12C
H3.1/3.2/3.3K57	1	20	Unmodified	12C
H1.2/3/4K90	1	50	Unmodified	12C
H1.2/3/4K85	1	50	Unmodified	12C
H1,2/1.3K46	1	50	Unmodified	12C

Table 5-5 Carbamate sites identified in the native nucleosome dataset under the conditions evaluated. Unmodified and modified relate to propionylated and non-propionylated samples.

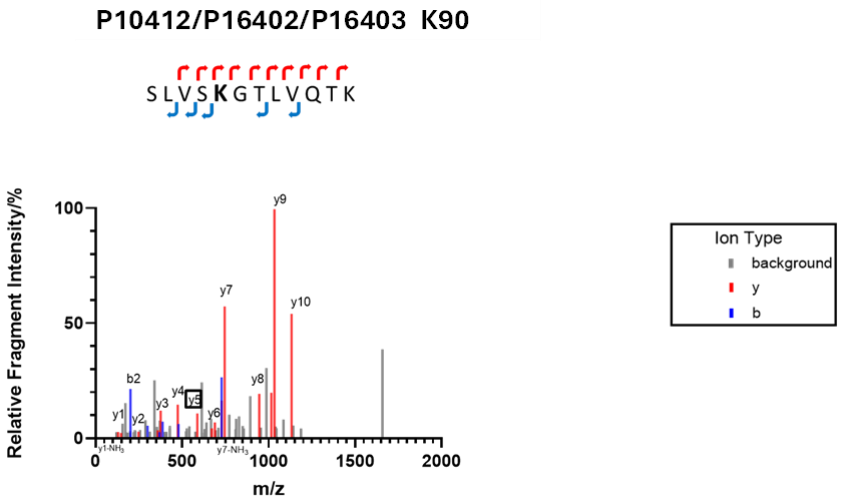


Figure 5-7 Identification of carbamate histone hits identified only in the propionylation/ non-propionylation dataset. Plots of relative fragment intensity versus m/z from LCMSMS identifying trapped carbamates on H1.2/1.3/1.4K90 (P10412/P16402/P16403) in the presence of $^{12}CO_2$. Each spectrum is associated with a peptide sequence illustrating the identification of predominant y (red) and b (blue) ions. The grey peaks represent background ions, and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

Further investigation into carbamylation on nucleosomes was pursued by purification of recombinant nucleosomes to reduce impurities as described in sections 5.10 and 5.11. This avenue was pursued due to the core histone variants and their PTMs being a more attractive research target than the linker Histone H1.

5.10 Recombinant Protein Expression Factors

5.10.1 Choice of Expression System and Cell Strain

Escherichia coli (*E. coli)* is commonly selected as a host organism for recombinant protein production due to its fast doubling time, high cell density cultures which result in high yields of the POI and the ease of genetic manipulation via transformation of exogenous DNA.[262] Due to its popularity, many laboratory strains of *E. coli* have been engineered. This work utilized the DH5α strain and BL21 (DE3) strain for plasmid and protein production, respectively. DH5α includes two key mutations including, the bacterial recombination (*recA*) mutation, which increases plasmid insert stability and the endonuclease (*endA1*) mutation to eliminate non-specific digestion of plasmid.[263] The BL21 (DE3) strain is protease deficient (Lon and OmpT) therefore recombinant proteins are not degraded when expressed. In addition, BL21 (DE3) produces the T7 RNA polymerase required for gene transcription as detailed below in section 5.10.2.[264]

5.10.2 Choice of Expression Vector

Expression vectors for laboratory use, consist of three key components including an origin of replication (ori), antibiotic resistance marker and multiple restriction enzyme sites. The plasmid expression vector under T7 control (pET) 28a vector was used initially for the expression of all four recombinant histones. However, after initial expression tests, the pET24a vector was used for H2B and H4 expression. pET28a and pET24a plasmids have many features in common including the selection marker, which is kanamycin, the tag encoded by the vector which is a histidine (His) tag and the presence of the T7 promoter. Figure 5-8 is a simplified depiction of pET28a highlighting the important components for plasmid propagation. Kanamycin acts as a marker to indicate that the recombinant DNA has been effectively transformed into the cell line because only cells with the plasmid will survive kanamycin selection. Kanamycin is an antibiotic which inhibits protein synthesis by binding to the 30S ribosomal subunit resulting in mistranslation.[265] The main difference between pET28a and pET24a is the composition of restriction enzyme (RE) sites which are also referred to as multiple cloning sites.

The RE sites chosen for recombinant DNA insertion are unique along the plasmid backbone and are highlighted in Figure 5-8 as *NcoI* and *NotI* for the pET28a vector. These sites are *NdeI* and *NotI* for the pET24a vector. In addition, the RE sites chosen for recombinant DNA insertion are located downstream from the T7 promoter so that expression is T7 dependent.

pET vectors use the T7-specific inducible system to yield a high expression of recombinant protein. For protein production, the T7 system requires T7 polymerase to bind to the T7 promoter which initiates the mRNA transcription of the target gene that is subsequently translated into the POI. IPTG is used as a non-hydrolysable artificial inducer for T7 protein expression due to its interaction with the lac operon.[266] During the lag phase of bacterial growth, the recombinant protein is not expressed because the lac repressor (LacI) is bound to the lac operator. When bacteria reach the exponential phase, protein expression is induced by IPTG which binds to LacI and relieves repression.

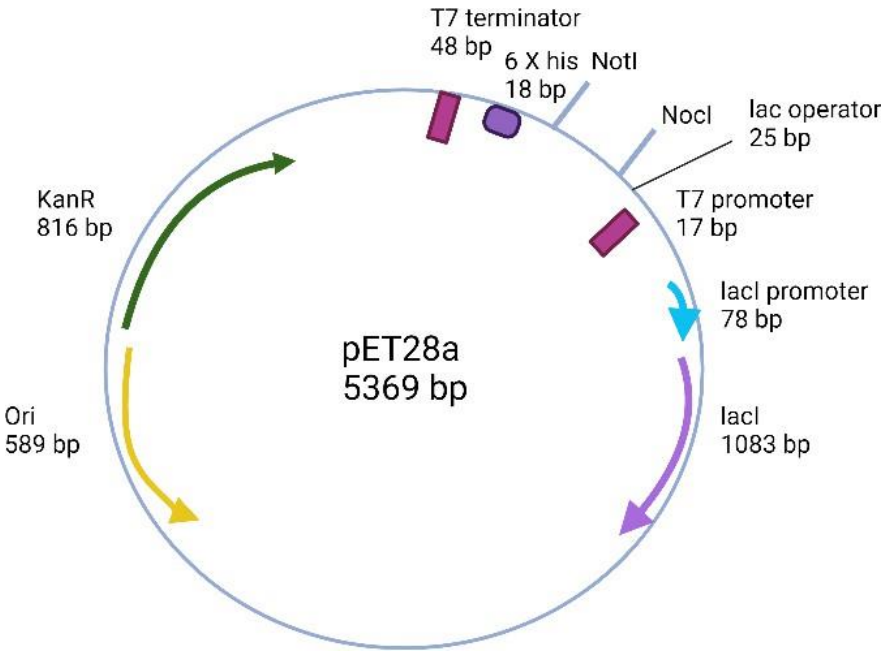


Figure 5-8 Plasmid sequencing map of Pet28a with key features for propagation.

5.10.3 Purification Tags.

The use of purification tags in protein production is a commonly used purification approach called affinity chromatography.[267] As previously mentioned, the pET28a and pET24 vectors contain a histidine tag which can be used for this purpose. However, in this work, the POIs were untagged, the elimination or incorporation of this tag is determined by the open reading frames (ORFs) found within the plasmid sequence. ORFs consist of start and stop codons which are recognised by the ribosome to translate a specific sequence of DNA into the recombinant protein. In this case, the His tag lies outside the ORF and therefore the POIs were untagged.

Despite the purification advantages of using a tag, these recombinant proteins did not have this functionality. Luger developed the first method that successfully reconstituted the nucleosome core particle (NCP).[268] An untagged approach was adopted because histones are small proteins, and the researchers hypothesised that the tag could affect the function or structure of histones. However, using tags in histone purification is possible because five years later Dyer updated Luger's method, one of these modifications included the use of non-tagged and histidine-tagged histones, however, the purification method was consistent between these proteins and did not include affinity chromatography.[269] Due to the interest in chromatin research and the laborious processes outlined by Luger and Dyer, many researchers have developed protocols to improve NCP preparation. One of these changes includes the use of a histidine tag and a nickel affinity-based purification step however this approach has been adopted for co-expression histone protein strategies.[270,271]

5.11 Recombinant Nucleosome Production

5.11.1 Histone Expression and Cell Growth

The DNA sequences which encode each histone variant (Figure 5-9) were cloned into the expression plasmids outlined in 5.10.2. One of the limitations of the *E. coli* expression system is that *E. coli* has differential codon usage when compared to the host system of the protein which is particularly an issue for eukaryotic protein expression. The Rosetta strain has been engineered to overcome this issue, the incorporation of the pRARE plasmid encodes rare *E. coli* tRNAs and enhances the expression of eukaryotic proteins which contain rare codons.[272] The Tuner strain has a lacZY mutation which allows uniform entry of IPTG to enable homogenous induction.[273] These strains were combined to give competent Rosetta Tuner cells.

Histone H2A type 1.

MSGRGKQGGKARAKAKTRSSRAGLQFPVGRVHRLLRKGNYAERVGAGAPVYLAAVLEYLTAEILELAGNA
ARDNKKTRIIPRHLQLAIRNDEELNKLLGKVTIAQGGVLPNIQAVLLPKKTESHHKAKGK

Histone H2B type 1-C/E/F/G/I

MPEPAKSAPAPKKGSKKAVTKAQKKDGKKRKRSRKESYSVYVYKVLKQVHPDTGISSKAMGIMNSFVNDIFERI
AGEASRLAHYNKRSTITSREIQTAVRLLLPGELAKHAVSEGTKAVTKYTSSK

Histone H3.1

MARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQR
LVREIAQDFKTDLRFQSSAVMALQEACEAYLVGLFEDTNLCAIHAKRVTIMPKDIQLARRIRGERA

Histone H4

MSGRGKGGKGLGKGGAKRHRKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKVFLENVIRDA
VTYTEHAKRKTVTAMDVVYALKRQGRTLYGFGG

Figure 5-9 The primary sequence for the specific histone variants used in this study.

Expression levels were tested on a small scale for histones cloned into the pET28a vector cultured in the Rosetta Tuner strain as detailed in 2.5.4.2. It is important to note that the expression conditions described here relate to the exponential phase of growth when cells reached an optical density at 600 nm (OD_{600}) of 0.4 - 0.6. Samples were obtained from cultures grown at 25 °C with and without IPTG induction for 0, 3, 4, 16 h. Both the soluble and insoluble fractions from the protein lysate were separated on an SDS-PAGE gel to test for recombinant protein production. The molecular weights for the histone proteins are given in Table 5-1. SDS-PAGE gels (Figures 5-10 and 5-11) indicated that H3 and H2A were successfully expressed using these conditions, as shown by the white box in Figures 5-10 and 5-11 surrounding the POI. Both H3 and H2A exhibit leaky expression and are present in both the insoluble and soluble fractions. Both the Tuner strain and the T7 expression system aim to reduce leaky expression because it can cause cell death or affect cell growth. Despite using the Tuner strain and the T7 expression system, leaky expression was still seen. Expression was scaled up to 12 L using the 16 h induced and non-induced condition to see if the proteins were overexpressed. Overexpression of predicted H2A (Figure 5-12) and predicted H3 was seen in large-scale cultures, therefore the 16 h induced condition was the selected expression condition. However, the expression of H2B and H4 was poor using this plasmid and strain as shown in Figures 5-13 and 5-14.

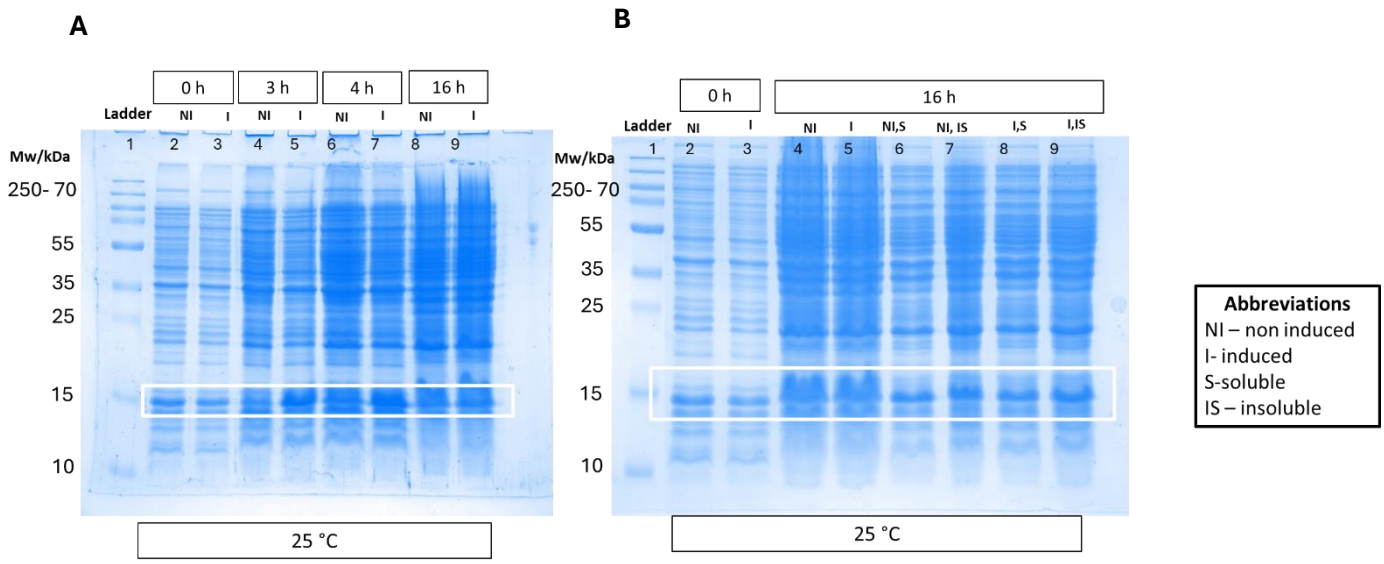


Figure 5-10 SDS-PAGE of test expression of predicted histone H3 at 25 °C with (A) four different time points and (B) assessing the solubility of H3. Lane 1 in both gels contains a MW marker (kDa) with weights specified. (A) Four time points were assessed: 0, 3, 4 and 16 h. Lanes 2 and 3 are at 0 h, lanes 4 and 5 at 3 h, lanes 6 and 7 at 4 h and lanes 8 and 9 at 16 h. Lanes 2,4,6 and 8 are non-induced whilst 3,5,7,9 are induced with 0.2 mM IPTG. (B) Lanes 2 and 3 are samples taken at 0 h and lanes 4 and 5 are at 16 h. The samples in lanes 6-9 are fractionated into soluble and insoluble whereby lanes 6 and 8 are soluble and lanes 7 and 9 are insoluble. Lanes 2,4,6 and 7 are non-induced, whilst lanes 3,5,8 and 9 are induced with 0.2 mM IPTG. The strain used was Rosetta Tuner and the white box surrounds the predicted POI.

.

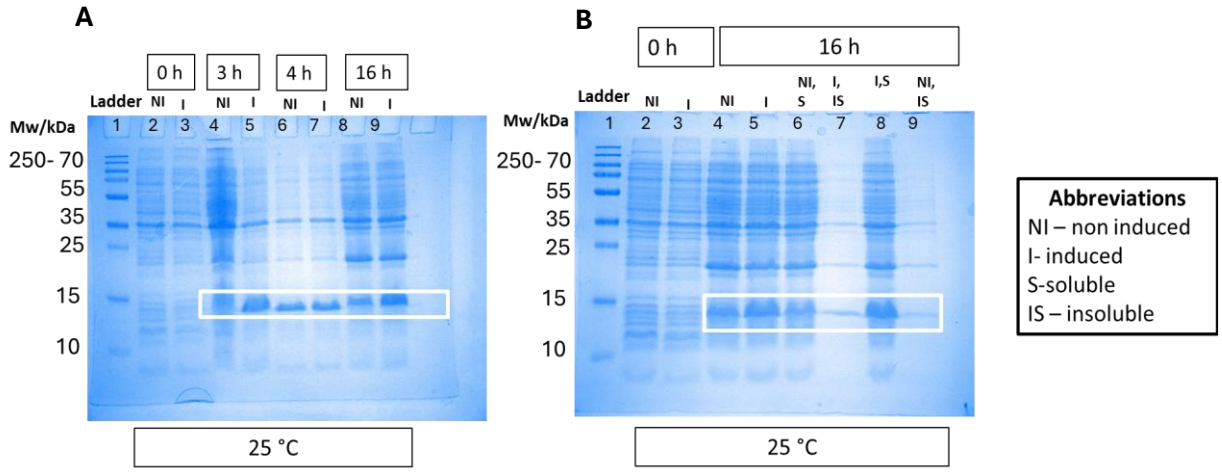


Figure 5-11 SDS-PAGE of test expression predicted histone H2A at 25 °C with conditions tested in each lane outlined in Figure 5-10. However, B has a different lane order for induced and non-induced samples. (B) Lanes 2,4,6 and 9 are non-induced, whilst lanes 3,5,7 and 8 are induced with 0.2 mM IPTG.

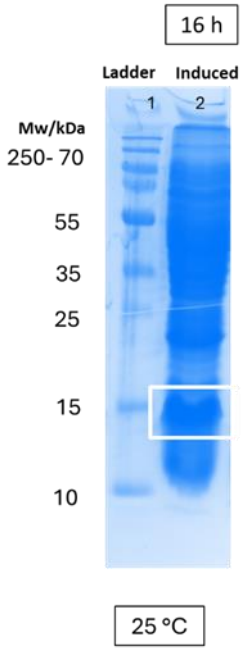
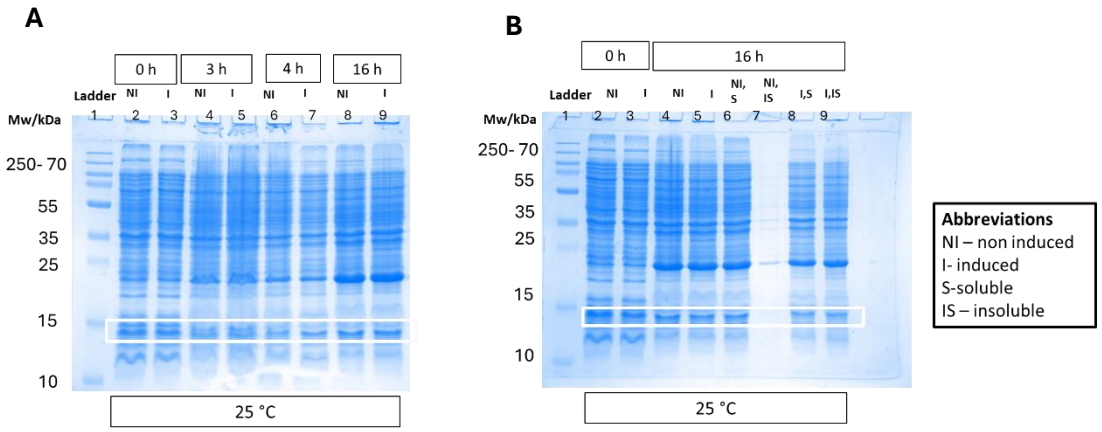


Figure 5-12 SDS- PAGE of large-scale predicted H2A expression grown for 16 hrs at 25 °C with IPTG induction in lane 2. Lane 1 in both gels contains a MW marker (kDa) with weights specified.



(5-13 SDS-PAGE of test expression of predicted histone H2B at 25 °C with conditions tested in each lane outlined in Figure 5-10.

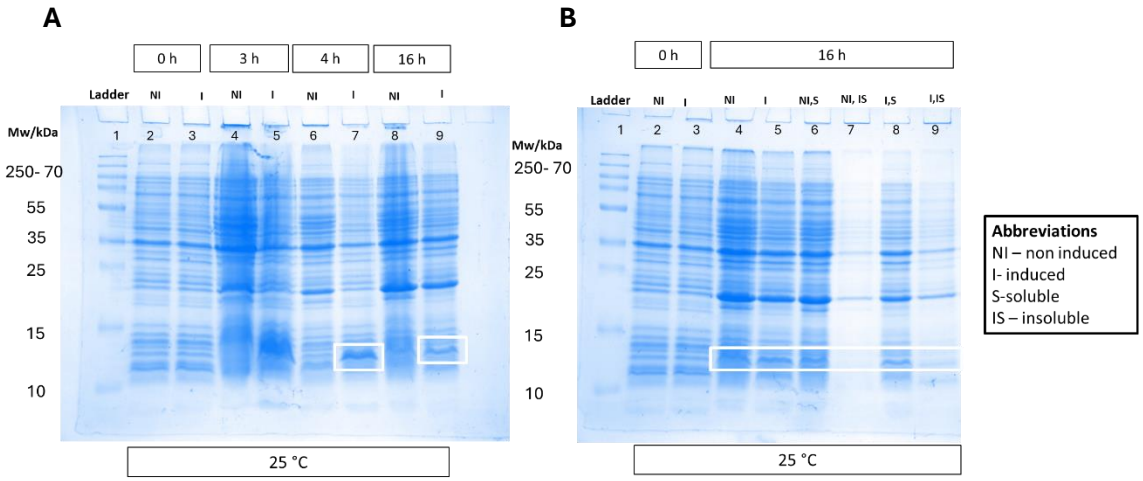


Figure 5-14 SDS-PAGE of test expression of predicted histone H4 at 25 °C with conditions tested in each lane outlined in Figure 5-10.

To improve the expression of H2B and H4, it was noted that Dyer *et al*. stated that optimal expression for some histone variants was observed with BL21-CodonPlus (DE3)-RIL cells. [269] The BL21-CodonPlus (DE3)-RIL cells encode for extra copies for four tRNAs which recognise rare codons which differs from the Rosetta strain which encodes for seven rare tRNAs. This strain provides an increased supply of tRNA-specific codons which code for arginine (R), isoleucine (I) and leucine (L) amino acids. These three amino acids constitute 16% and 24% of H2B and H4's primary sequence, respectively. In this study, H2B and H4 cloned into the pET24a plasmid were transformed into the BL21-CodonPlus

(DE3)-RIL strain. The results of overexpression of H2B and H4 in this strain are shown in Figures 5-15 and 5-16, respectively. The post-induction temperature of 25 °C gave a higher yield than 18 °C for both proteins. The BL21-CodonPlus (DE3)-RIL strain may have successfully overexpressed H2B and H4 because of the extra supply of tRNAs for the RIL amino acids whereas the Rosetta strain is not optimised specifically for these amino acids.

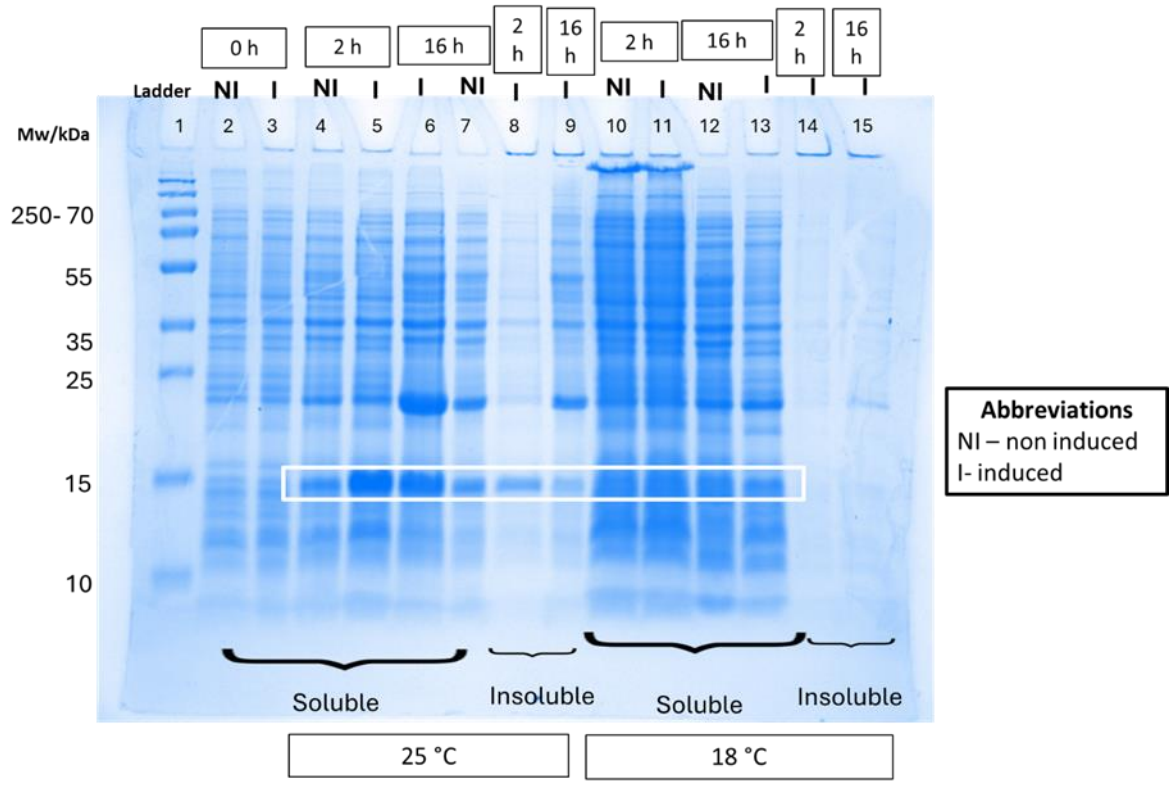


Figure 5-15 SDS-PAGE expression of predicted histone H2B in BL21 - Codon Plus (DE3) – RIL at two different temperatures SDS and three time points. Lane 1 contains a MW marker (kDa) with weights specified. Lane 2 -9 is grown at 25 °C and lanes 10-15 are at 18 °C. Lanes 2 and 3 are at 0 hours (h), 4, 5, 8, 10, 11, 14 are at 3 h and 6, 7, 9, 12, 13 and 15 at 16 h. Lanes 2,4,7,10 and 12 are not induced whilst all other lanes are induced with 0.2 µM IPTG. All lanes include the soluble fraction except for 8,9, 14 and 15 which contain the insoluble fraction.

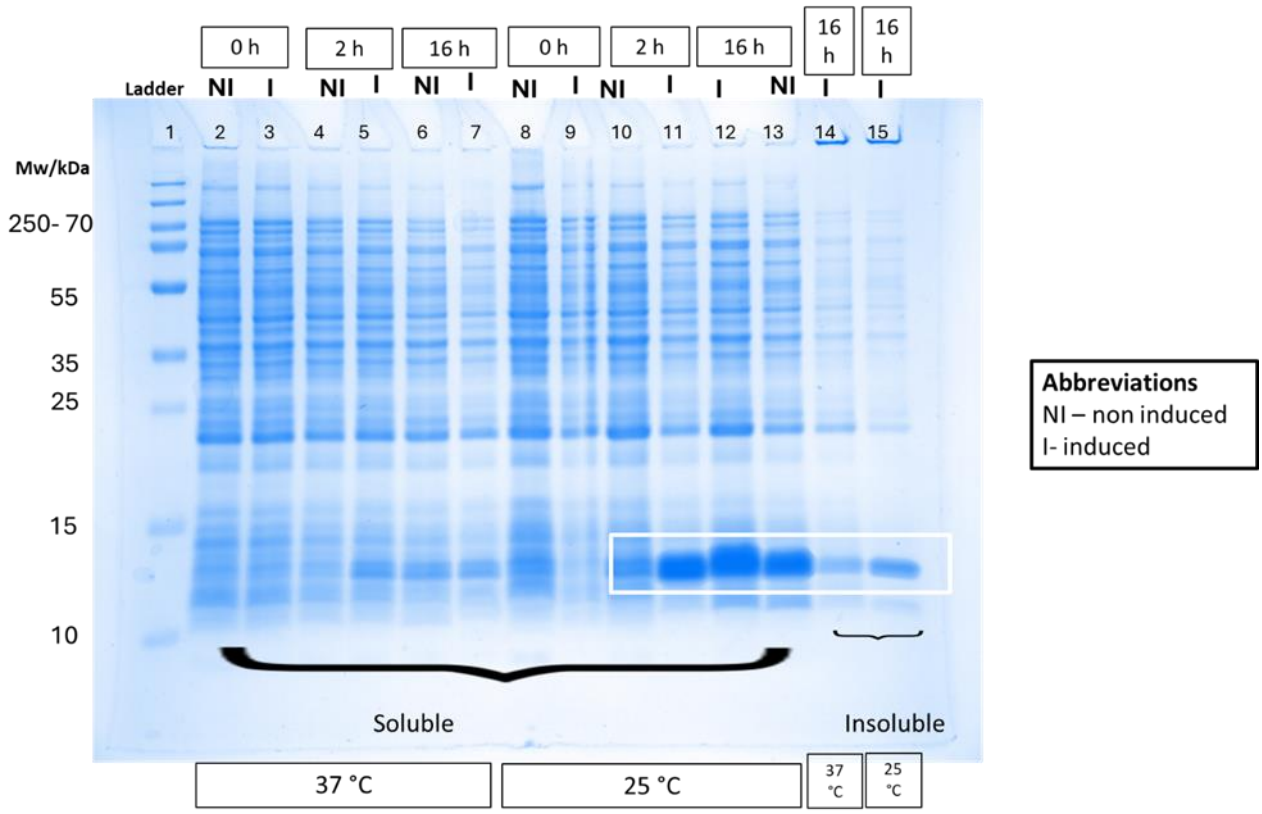


Figure 5-16 SDS PAGE expression of predicted H4 as outlined in Figure 5-15 however the conditions tested were slightly different. Lanes 2 -7, and 14 are grown at 37 °C and lanes 8-13 and 15 are at 25 °C. Lanes 2,3, 8 and 9 are at 0 hours (h), 4, 5, 10 and 11 are at 3 h and 6, 7, 12, 13, 14 and 15 at 16 h. Lanes 2,4,6,8, 10 and 13 are not induced whilst all other lanes are induced with 0.2 µM IPTG. All lanes include the soluble fraction except for 14 and 15 which contain the insoluble fraction.

For overexpression of all four histone proteins, the *E. coli* cells were grown to an OD_{600} of 0.4-0.6 at this point they were induced and grown at 25 °C for 16 h to overexpress the predicted POI. Leaky expression was an issue in these expression experiments, however further purification and validation steps confirmed that the desired histone proteins had been successfully expressed.

5.11.2 Histone Purification

Due to the insoluble nature of histone proteins, the overexpressed misfolded proteins aggregate into inclusion bodies. Luger *et al.* developed a method for extracting histone proteins [268] which involves isolating the inclusion bodies from *E. coli* cells using the detergent triton. Following this, histones were solubilised with a denaturant such as urea or guanidine hydrochloride. The Luger

purification of histones involves two steps, size exclusion purification under denaturing conditions, followed by cation exchange as detailed in the methods chapter in sections 2.5.4.5 and 2.5.4.6. Size exclusion is a purification technique which separates proteins based on their size, it will take small proteins longer to elute from the column than large ones because small proteins travel through the pores of the column resin. Figure 5-17B is an SDS PAGE analysis of H2A protein purity after using this method. From this result, it was decided that the Luger protocol was not creating a pure product, contaminant proteins were co-eluting with H2A following size exclusion (Figure 5-17A) and the final gel shows a very faint band for the POI (Figure 5-17B), indicating insufficient protein and degradation throughout the purification process.

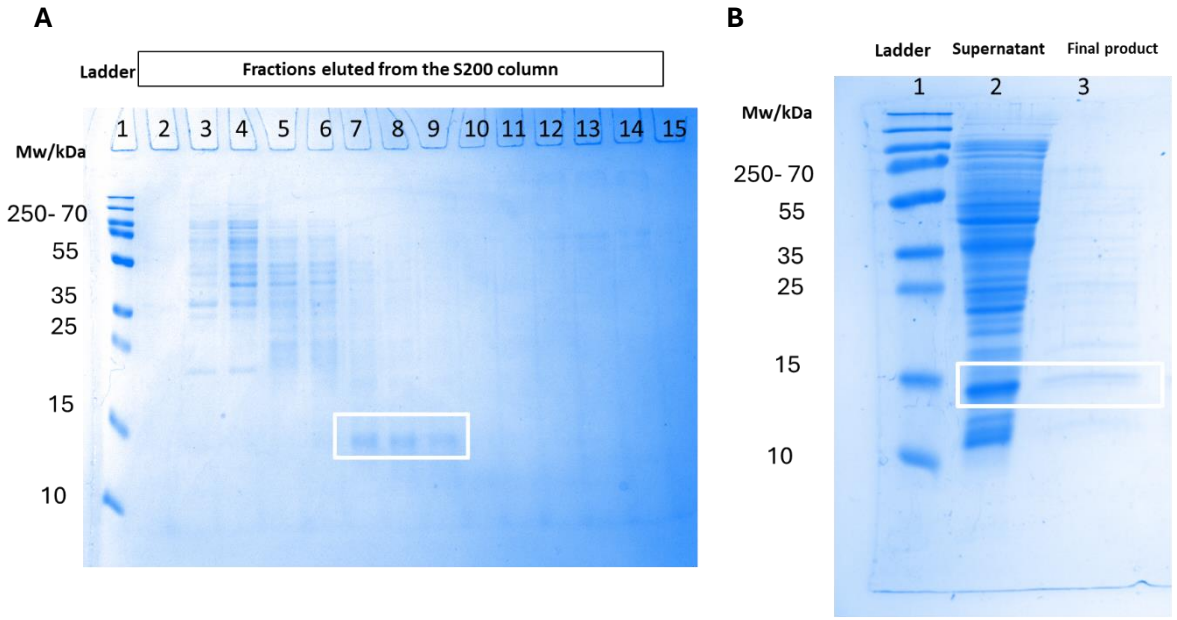


Figure 5-17 SDS PAGE of predicted histone H2A purified using the Luger method.[268] (A) size exclusion result where lanes 2-15 contain 0.5 ml fractions eluted from the S200 column, lanes 7-9 contained the POI as shown by the white box. (B) final product after size exclusion and ion exchange. Lane 2 is the supernatant of H2A that was obtained after the expression, lane 3 shows the final purification product. Lane 1 in A and B contains a MW marker (kDa) with weights specified.

There were two main limitations to this method. Firstly, the histones were found in both the insoluble and soluble fractions during expression tests, therefore the inclusion body formation reduced

the possible yield. Secondly, the Superdex 200 size exclusion column is not an optimal size for this purification. Histones are 11-15 kDa proteins, and the S200 resin did not effectively resolve the POI from the other contaminant proteins.

An alternative approach was considered whereby, insoluble histone proteins were directly solubilised using the denaturant urea, to purify both the soluble and insoluble fractions. This method was developed by Klinker to reduce the laborious nature of NCP preparation [274] and is described in sections 2.5.4.7 and 2.5.4.8. Following histone solubilisation, the purification involves two types of ion exchange, firstly denaturing cationic followed by anionic exchange. The most important parameters for ion exchange are the protein's pI and the buffer's pH. The pI of a protein can be defined as an average of all the amino acids' pKa values. Proteins are in an uncharged state when the buffer pH is the same as the protein's pI. Table 5-1 details the histone proteins' pI values; histones are highly basic proteins with high pI values. The cationic SP column is negatively charged and therefore retains positively charged analytes until elution with high salt. The histone proteins were buffered at pH 5.2 and therefore were positively charged, as the buffer pH is lower than the pI value. Conversely, the anionic Q column is positively charged and therefore retains negatively charged analytes until elution with high salt. The pH of the buffer used for this step was pH 8 and therefore the POI was positively charged and therefore would not interact with the column but the contaminants with negative charges would. Figure 5-18 shows the 280 nm UV chromatograms for ion exchange purification. Between the two purification steps, the histone protein is refolded overnight by dialysing out the urea under reducing conditions to prevent precipitation.

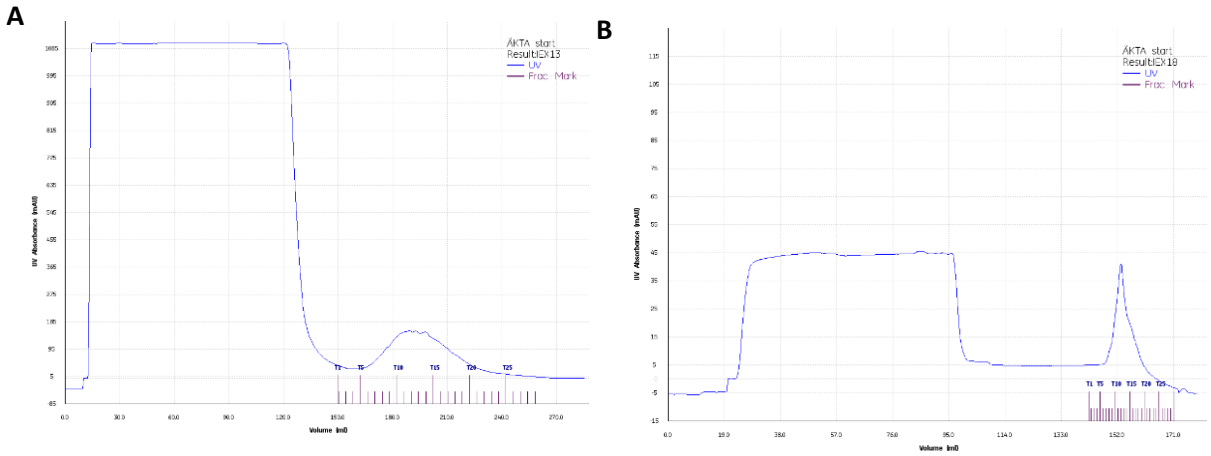


Figure 5-18 AKTA UV 280 nm Chromatograms where the UV absorbance at 280 nm in milli-absorbance units (mAU) is measured against the volume of buffer run over the purification column in millilitres (ml). The first flat peak indicates the flowthrough, and the second peak represents the fractionated elutions where (A) is the SP column and (B) is the Q column purification steps.

Figure 5-19 shows the purification products of H2A from 3L of induced culture grown at 25 °C. Figure 5-19C has a contaminant protein at 30 kDa, this was initially thought to be an H2A dimer that was not denatured under SDS PAGE conditions. However, it was confirmed by ESI-MSMS that this was the 50S ribosomal *E. coli* protein. The anionic purification step was able to remove the *E. coli* contaminant from the final purified H2A protein when purifying 1 L, (Figure 5-20) instead of 3 L of culture because the column was not at binding capacity. This purification technique was applied to all 4 histones with the final products shown in Figure 5-21, however, H3 still contained the ribosomal *E. coli* contaminant band. Each histone purification product was confirmed to be the specified histone variant by trypsin digest of the SDS-PAGE gel bands and ESI-MSMS.

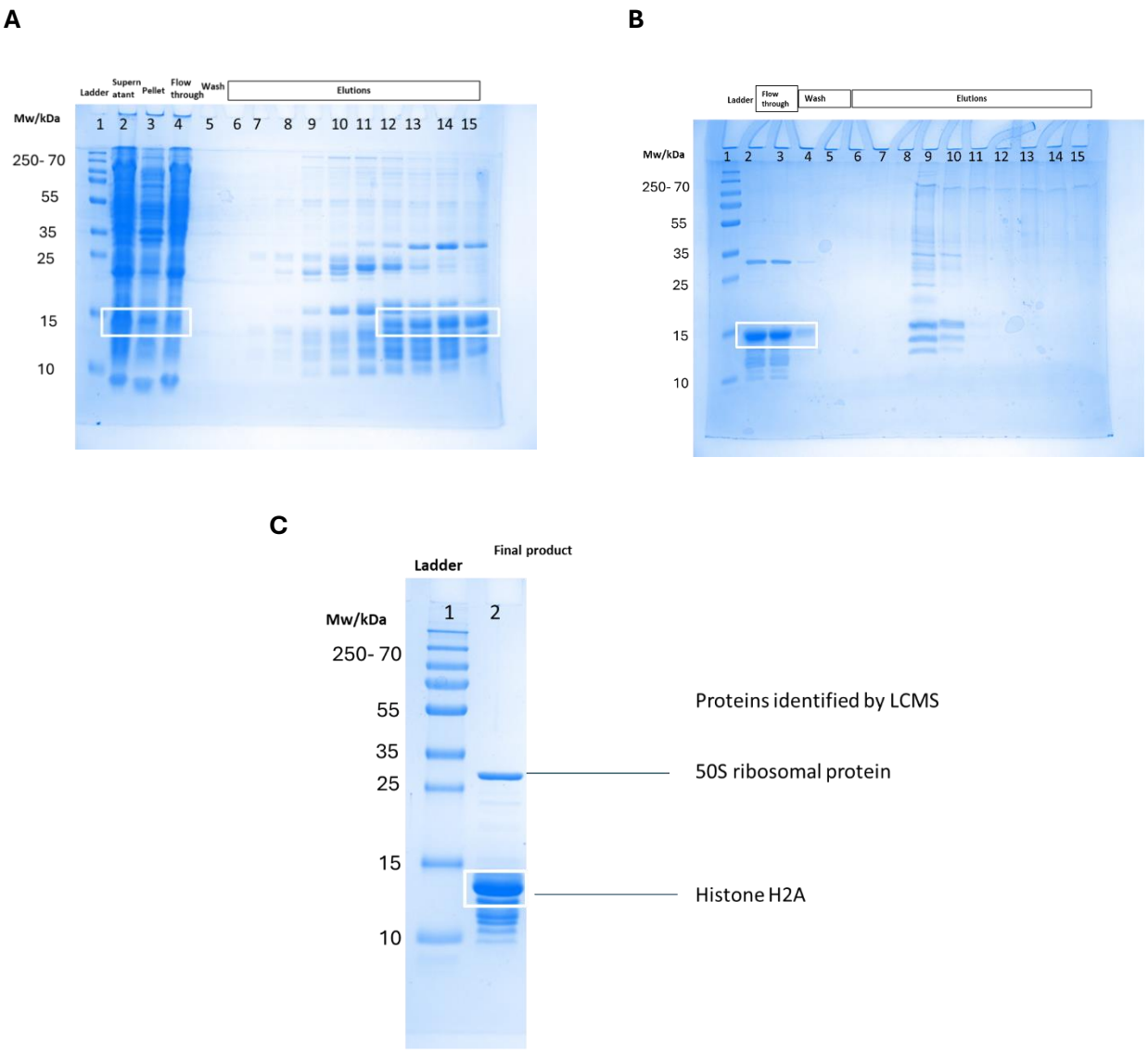


Figure 5-19 Purification of 3L of predicted overexpressed histone H2A using the rapid purification method by Klinker *et al*. [274] Lane 1 in all these gels contains a MW marker (kDa) with weights specified. The white box highlights the predicted H2A on the gel. (A) SP cationic exchange, lane 2 contains the supernatant, lane 3 contains the pellet after the urea treatment and centrifugation, lane 4 is the flow- through of the supernatant running over the column, lane 5 is a buffer wash of the column, lanes 7-15 are elutions over increasing salt concentrations. (B) Q anionic exchange where lanes 2 and 3 are the flowthrough, lane 4 is the wash and 6-15 are increasing salt concentration elutions. (C) The final purified product was confirmed by ESI-MSMS, it was thought the protein at ~30 kDa was a dimer of H2A however it was confirmed to be a contaminant *E. coli* protein with a weight of 29.9 kDa.

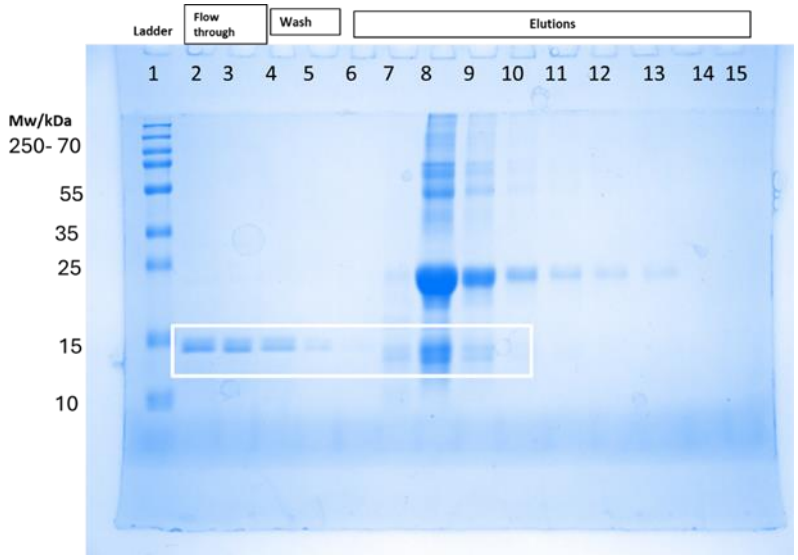


Figure 5-20 Q column purification of 1L of predicted overexpressed histone H2A. Lane 1 contains a MW marker (kDa) with weights specified. The white box highlights the predicted H2A on the gel. Q anionic exchange where lanes 2 and 3 are the flowthrough, lanes 4 and 5 are the wash and 6-15 are increasing salt concentration elutions.

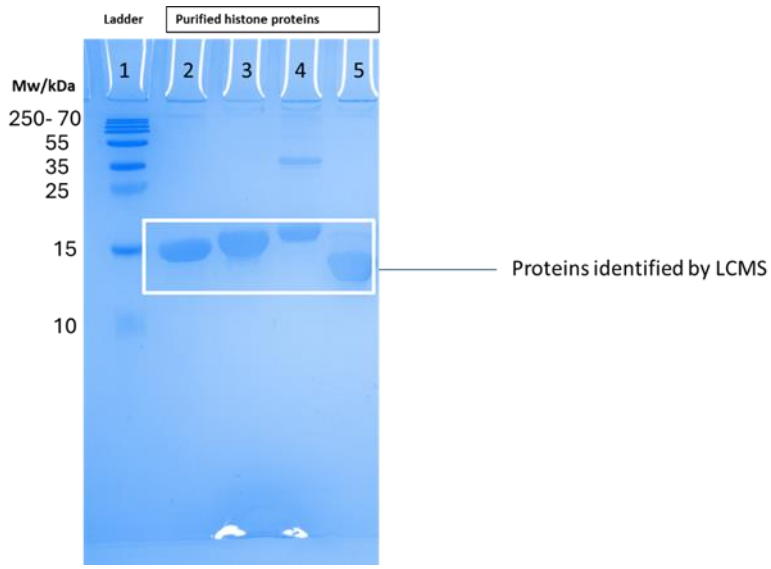


Figure 5-21 Purification products of each histone confirmed by ESI-MSMS. Lane 1 contains a MW marker (kDa) with weights specified. The white box highlights the presence of the POI. Lane 2 is H2A, lane 3 is H2B, lane 4 is H3 and lane 5 is H4.

5.11.3 Histone Octamer Reconstitution

All four histone products were purified successfully in high enough quantities for histone octamer reconstitution as described in section 2.5.4.9. It was important to ensure the correct ratios of each variant were used for the reconstitution. Any contaminants from the individual purification steps could be removed during the complex purification. Firstly, the proteins were unfolded individually, mixed, and dialysed into a refolding buffer. Following refolding, the histone octamer was purified by size exclusion (Figure 5-22) and fractions relating to each peak of the chromatogram were pooled after analysis by SDS-PAGE. The purification buffer required 2 M NaCl to prevent histone aggregation and stabilise the complex by ionic interactions.

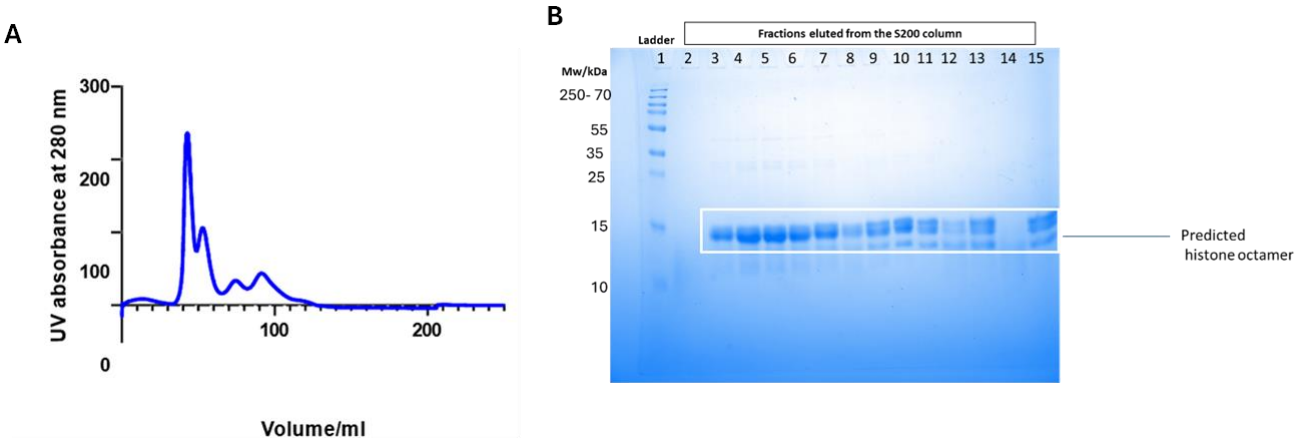


Figure 5-22 S200 purification of the histone octamer. (A) UV 280 nm chromatogram where the UV absorbance at 280 nm in milli-absorbance units (mAU) is measured against the volume of buffer in millilitres (ml) run over the purification column. (B) SDS-PAGE of predicted histone octamer following size exclusion, lane 1 is the molecular weight ladder with weights specified in kDa. Lane 2 is the flowthrough, whilst lanes 3-15 are the elution fractions which correspond to the first two peaks on the chromatogram.

An AKTA UV chromatogram is a good indicator of a protein's purity. A symmetrical sharp single peak indicates good resolution and purity of the POI. In Figure 5-22A there are four peaks and the first two were confirmed by SDS PAGE to contain all 4 histones (Figure 5-22B). The two later peaks were non-specific dimers of the histone variants. To assess the multimeric state of the histone complex, analytical sizing as described in section 2.5.4.10 was carried out.

Firstly, the Superose 6 was calibrated using 4 standards which eluted at specific volumes dependent on their Mw as shown in Figure 5-23A. A few parameters were needed to determine the partition coefficient (Kav) including, the geometric column volume (Vc) which is calculated from the length and radius of the column and the void volume (Vo) which is the volume of the mobile phase in the column shown as the first peak on the UV chromatogram, for the column used the Vo was at 7.4 ml. These values along with the elution volume (Ve) for each standard (Figure 5-23A) were used in Equation 5-1 to calculate the Kav. A linear regression of Kav against the logMw was plotted (Figure 5-23B) which can be used to determine the observed molecular weight of the unknown protein which in this case is the histone octamer. The elution profile of the purified octamer is shown in Figure 5-23C. The Kav can be calculated from the Ve which is converted into the observed molecular weight. The main elution peak for the histone octamer is at 16.11 ml which is converted to 125 kDa using the line of best fit plotted from the standards (Figure 5-23B). Table 5-1 states the Mw of the histone octamer at 108 kDa. The observed Mw in comparison to the theoretical Mw are slightly different because the method works on the assumption that proteins are perfectly spherical. The histone octamer is described as globular however does have flexible tail regions which could account for the difference in observed and theoretical Mw values. A smaller secondary peak at 20.5 ml was also observed, meaning that the sample was not completely homogeneous and likely there were H2A-H2B dimers in the final product as the observed Mw was at 31 kDa close to the theoretical Mw of 27.8kDa. The later peaks displayed on the chromatogram in Figure 5-22 were also run on the Superose column (Figure 5-23D) and these elute later due to being smaller dimers and singular histone proteins which were not included in the final product of the histone octamer which is shown in Figure 5-24.

$$Vc = \pi r^2 l$$

$$Kav = \frac{(Ve - Vo)}{(Vc - Vo)}$$

Equation 5-1 Calculation of the geometric column volume and the partition coefficient for analytical sizing.

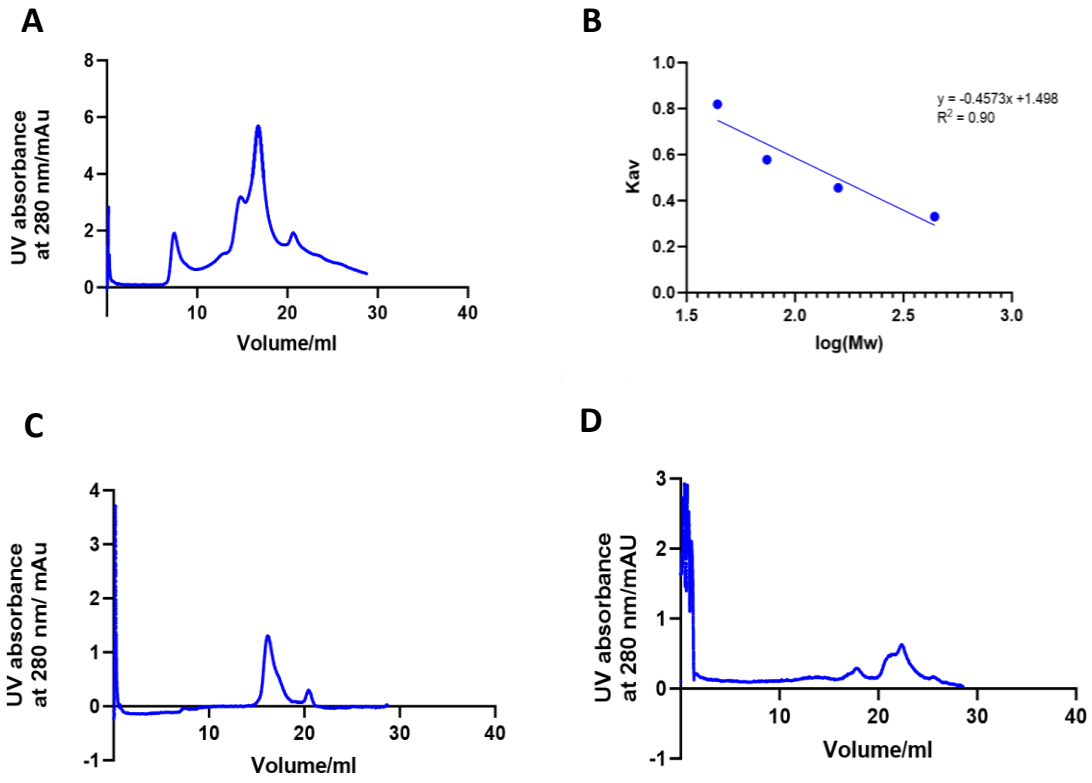


Figure 5-23 Analytical sizing of the histone octamer on the superose 6 UV chromatograms in A, C and D display the UV absorbance at 280 nm in milli-absorbance units (mAU) against the volume of buffer in millilitres (ml) run over the purification column. (A) Column calibration using ferritin, aldolase, conalbumin and ovalbumin. (B) Linear regression of the partition coefficient (Kav) versus log molecular weight (Mw) with the y = mx+c equation shown with R^2 displaying the goodness of fit where n=1. (C) Run profile of the histone octamer complex (D) Run profile of the non-specific complexes separated from the histone octamer including dimers and monomers.

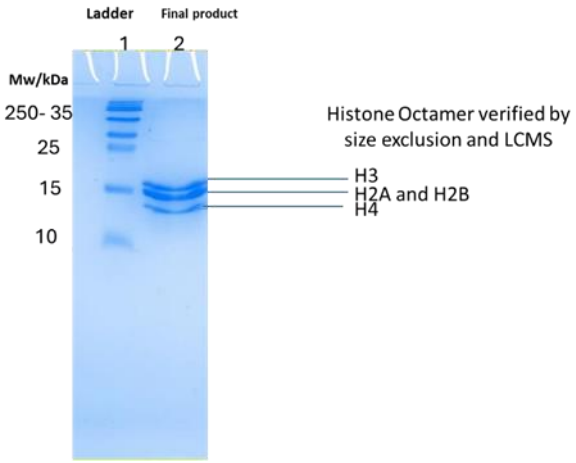


Figure 5-24 SDS-PAGE of the ESI-MSMS histone octamer (2) against the molecular weight ladder in kDa (1).

5.11.4 Widom DNA Large-Scale Purification

Two milligrams (mg) of DNA were needed to produce enough recombinant NCP for a trapping experiment. To produce enough DNA, a 12-copy sequence of 177 bp DNA was cloned into the phagemid (pBS) vector and cultured in DH5α as described in section 2.5.4.13. The widom 147 bp sequence is a synthetic DNA sequence with strong nucleosome positioning preferences. Here a 177 bp sequence is used, composed of the 601 sequence and a 30 bp linker. The plasmid is described as having 12 copies of the DNA, however, there are only 12 PmlI restriction enzyme sites, meaning that 11 copies of the 177 bp sequence can be excised from the plasmid (Figure 5-25). PmlI is a blunt end restriction enzyme which means all DNA produced from digestion is double stranded with no overhangs which is a key requirement for the NCP DNA.

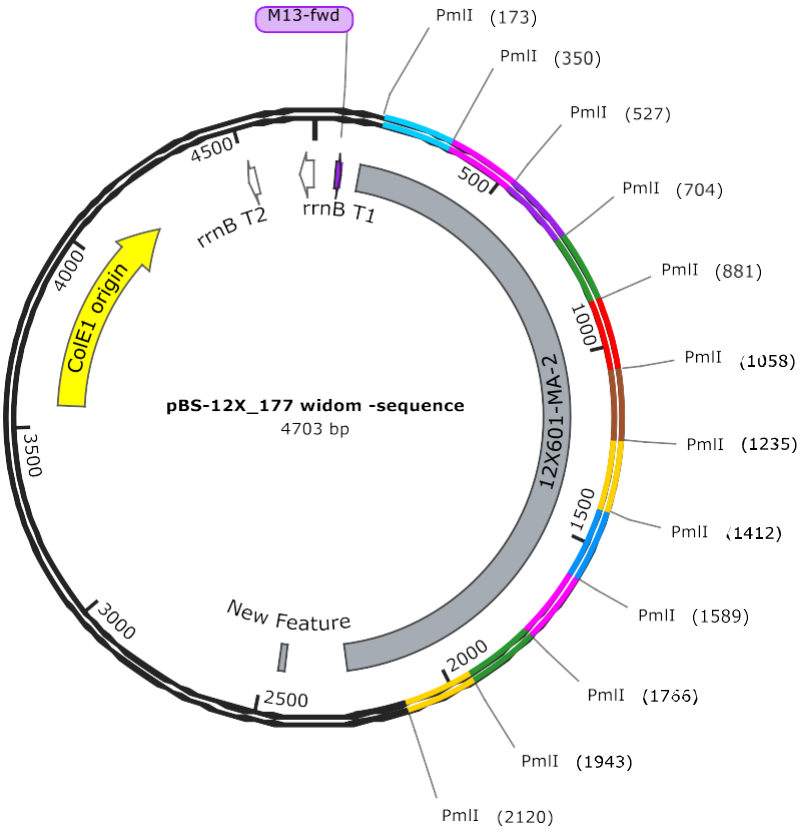


Figure 5-25 The pBS plasmid cloned with 177 bp widom sequence with PmlI restriction enzyme sites labelled. Created using Snapgene.

To assess whether the digestion of this plasmid was producing the correct length DNA fragment, agarose gels with a ladder of DNA fragments with a known number of base pairs were run. In addition to the ladder, a control sequence of 146 bps obtained via PCR was used. A pGEM-3z/601 plasmid from DH5α culture was miniprepped and the 146 bp sequence was amplified by PCR with specific primers as detailed in sections 2.5.4.11 and 2.5.4.14. Figure 5-26 shows the PCR products of the reaction with a gradient of annealing temperatures compared with a PmlI digest of the 12x copy plasmid. Figure 5-27 shows a few different enzyme concentrations used to complete the digest; the result of this experiment was that 1 U enzyme per 1 µg of DNA was needed for a complete digest.

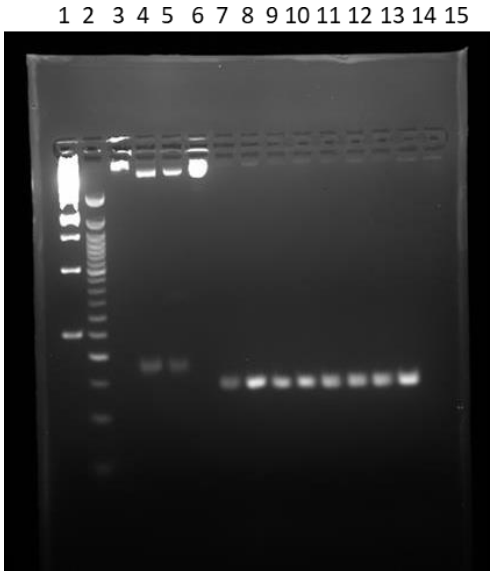


Figure 5-26 Agarose gel of digest (177 bp, lanes 5 and 6) and control (146 bp, 8-15) PCR products. Lanes 1 and 2, are base pair markers of 1kb and 50 bp, respectively. Lane 1 contains DNA fragments at 10,000, 8,000, 6,000, 5,000, 4,000, 3,500, 3,000, 2,500, 2,000, 1,500, 1,000, 750, 500, 250 bp. Lane 2 contains DNA fragments at 1,350, 916, 766, 700, 650, 600, 550,500, 450, 400, 350, 300, 250, 200, 150, 100 and 50 bp. Lane 3 is the intact 12x copy plasmid, and lane 6 is the intact 1x widom DNA plasmid, these do not run far in the gel as they are above 3,000 bp in size. Lanes 4 and 5 show an excision digest of the 12x copy DNA with two main bands one at the desired 177 bp mark and the other for the rest of the plasmid. The digest time for lanes 4 and 5 was 15 minutes and 4 hours, respectively. Lanes 7-15 are the expected PCR products from 1x widom DNA amplification each well had a different annealing temperature ranging from 47- 59 °C, with temperature increasing from left to right.

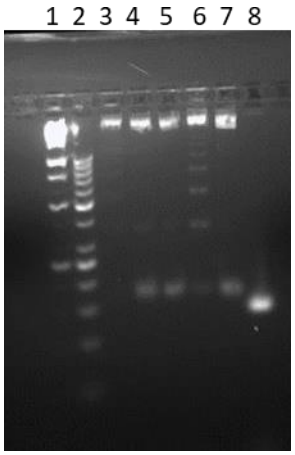


Figure 5-27 Lane 1 and 2, are the base pair markers detailed in Figure 5-26. Lanes 3-5 is a miniprep isolation of the pBS 12x 177 bp DNA whilst 6 and 7 are from a bigger giga-prep scale isolation. Lanes 3-7 are the digested expected DNA products produced with different concentrations of enzyme, lanes 3 and 6 were digested with 0.5 U/ 1 µg, lanes 4 and 7 with 1 U/ 1 µg 1 and lanes 5 with 10 U/ 1 µg. Lane 8 is the 146 bp PCR product control. The digest time for all lanes was 4 h.

Following these initial tests, the digest using 1 U enzyme per 1 µg of DNA was scaled up and completed on 15 mg of the pBS- 12x-177bp plasmid. The DNA products of this digest are shown in Figure 5-28. The gel shows banding of different length DNA fragments indicating that the digest does not go to completion on this large scale. A few of the PmlI cut sites across the plasmid copies have been targeted but not all of them, leading to a range of DNA fragments. After pooling the digested DNA and dialysing it, the DNA present was quantified to be 10 mg.

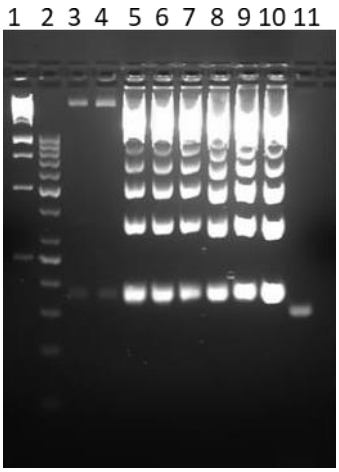


Figure 5-28 Agarose gel of large-scale digest, lanes 1 and 2 are the base pair markers of 1kb and 50 bp, respectively as described in Figure 5-26. Lanes 3 and 4 a miniprep and gigaprep plasmid digested previously with 1 U/1 µg. Lanes 5-10 are the expected DNA products from the 15 mg of DNA digested with 1 U/ 1 µg DNA split into 1 ml fractions for the digest and each lane is from a different fraction. Lane 11 is the 146 bp PCR product control.

Despite the inefficiency of the digest process, there was a high enough quantity of DNA, to proceed to the next step. Before purification, dialysis of the digested DNA into water was essential to remove salts present in the digestion buffer. A Mono Q purification was used due to having a higher resolution than a HiTrap Q column, this step aimed to separate the 177 bp fragment from the others. The result of this purification process is shown in Figure 5-29. The process of anion exchange was described earlier in 5.11.2, however in this instance instead of the protein's pI value, DNA is a negatively charged biomolecule and will interact with the Q resin. The column volumes and salt concentrations used are specified in 2.5.4.15. The result in Figure 5-29 showed that the 177 bp fragment elutes both before and with the other DNA fragments. This result led to the decision to extend the purification gradient between 40-70% of 1 M NaCl from 30 column volumes (CV) to 40 CV. Figure 5-30 is the result of this extended gradient and shows a better separation of the target sequence from the unspecific digest products. However, from these purifications, only 0.1 mg of the 177 bp DNA was produced from a starting amount of 1 mg. This loss can be accounted for by the number of unspecific fragments produced during the digest.

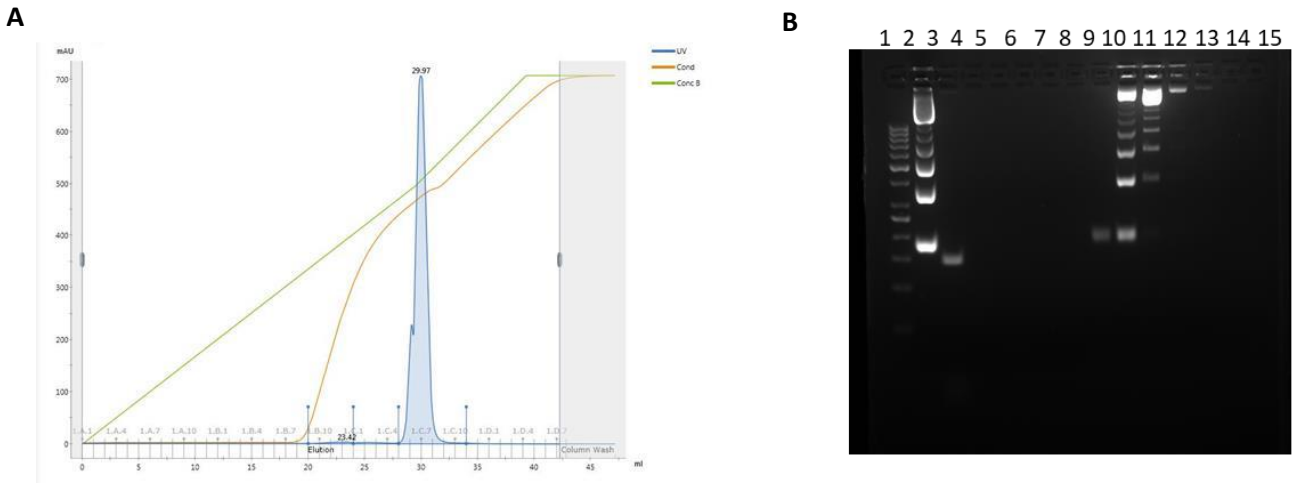


Figure 5-29 Purification of the 12-177 bp DNA using a monoQ column. (A) AKTA chromatogram where the UV absorbance at 280 nm in milli-absorbance units (mAU) is measured against the volume of buffer in millilitres (ml) run over the purification column. (B) Agarose gel of monoQ purification products, lane 1 is the 50 bp ladder with weights specified in Figure 5-26. Lane 2 is the digested PmlI DNA from Figure 5-28. Lane 3 is the 1x widom PCR control. Lanes 4 and 5 are flowthrough and wash, respectively. The remaining lanes are consecutive fractions collected. Lane 9 shows the expected 177 bp fragment separated from the other digest products, however the next fraction also contains the desired fragment with other contaminants.

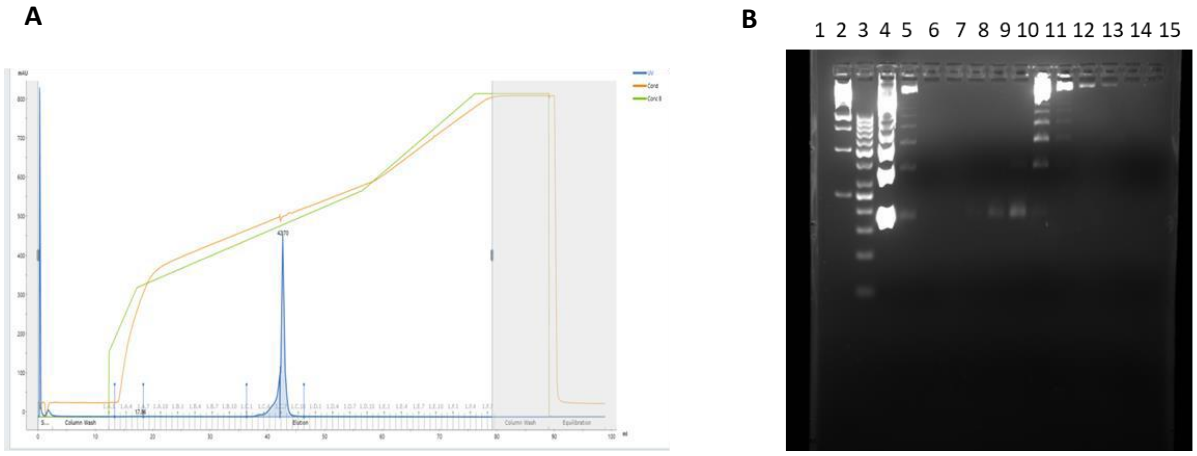


Figure 5-30 Purification of the 12-177 bp DNA using a monoQ column with an extended separation gradient. (A) AKTA chromatogram where the UV absorbance at 280 nm in milli-absorbance units (mAU) is measured against the volume of buffer in millilitres (ml) run over the purification column (B) Agarose

gel of monoQ purification products, lanes 1 and 2 are the base pair markers of 1kb and 50 bp respectively as described in Figure 5-26. Lane 3 is the digested PmlI DNA from Figure 5-28. Lanes 4 and 5 are flowthrough and wash, respectively. The remaining lanes are consecutive fractions collected. Lanes 7- 9 show the expected 177 bp fragment separated from the other digest products. Lane 10 has a small amount of the expected 177 bp DNA eluting with the contaminants.

It was surprising that the digest worked effectively with the same conditions at a smaller scale in Figure 5-27 compared with the large scale in Figure 5-28. Due to these inconsistencies, it was thought that PCR could be used as an effective method of producing the required widom DNA. This was trialled with the single copy plasmid as detailed previously, the yield from a single PCR reaction indicated that 2,000 PCR reactions would be needed to make 2 mg of DNA. It was decided that this would be too expensive and laborious, and due to time constraints, the histone octamer was used in the trapping experiment instead of the NCP.

5.11.5 Histone Octamer Trapping and Identification by LCMSMS

A similar experiment to the one described in section 5.9.2 was implemented using the recombinant histone octamer instead of the native nucleosomes. The dataset was composed of unmodified and modified samples, where n=1 using 12C and 13C Ci concentrations of 0, 20, and 50 mM trapped with TEO. Control samples were prepared with 0 mM Ci without TEO.

The histone octamer trapping experiment was complementary to the analytical size exclusion process to verify the presence of the histone variants. The purity of the trapped histone octamer is shown in Figure 5-24, when compared with Figure 5-4, the recombinant histone octamer is much cleaner than the native nucleosome. It was hypothesised coverage may be improved particularly in the propionylation samples. Tables 8-7 and 8-8 give the coverage for each histone variant identified in the unmodified and modified histone octamer trapping experiments, respectively. This data is summarised by Table 5-6 which shows the average coverage percentage for each histone variant across the dataset;

the numbers in brackets in the table account for how many times the histone variant was found across the total number of samples.

H2A and H2B coverage was reduced by propionylation which can be explained by the primary amino acid sequence of these variants. Coverage was improved for H3 and H4 in the propionylated recombinant histone octamer samples. The sequence homology between the histone variants leads to H1 being identified despite not being present in the reconstituted histone octamer sample. The H1 variant was identified at low coverage in the unmodified samples but absent in the modified samples due to there being fewer arginine cleavage sites in the amino acid sequence so when modified by propionylation the amino acid lengths are too long. This data indicates that propionylation on pure recombinant nucleosomes would be a suitable avenue for improving coverage of H3 and H4.

Histone Variant	Average coverage when ID /% in unmodified samples	Average coverage when ID /% in modified samples
H1.1/1.2/1.3/1.4/1t	12 (4/6)	0 (0/6)
H2A type 1-H/J/type 2-C	44 (5/6)	23 (6/6)
H2B type 1 - C/E/F/G/I/D/N/M/H/K/ F-S/TYPE 2 F	38 (3/6)	26 (6/6)
H3.1/2/3.3/It	26 (6/6)	39 (6/6)
H4	49 (6/6)	54 (5/6)

Table 5-6 Coverage across the histone octamer dataset where unmodified relates to no propionylation treatment, whereas modified relates to propionylated samples.

The carbamylation sites identified from the recombinant octamer included H4K32, H3K79 and H4K92 which gave further verification that these sites were real. The results from the histone octamer trapping experiment are shown in Table 5-7. As these sites have been identified previously the peptide sequence and mass spectrum for the site identification are not plotted. The results obtained here are limited because the octamer has been used instead of the nucleosome due to the reasons stated in 5.11.4. The histone octamer is less biologically relevant than the nucleosome for the identification of these sites due to structural differences and the amount of PTM crosstalk mediated by nucleosomes. However, the carbamate sites that were identified across these experiments were reproducible, found in nucleosome/HEK293 lysate samples and worth further investigation into whether they are biologically relevant to nucleosome-mediated processes such as transcription.

Carbamate Site	Number of samples ID	Carbon concentrations/mM	Propionylation	Carbon Isotope
H4K32	7	both	Both	Both
H3K79	1	20	Propionylated	12C
H4K92	2	50	Both	Both

Table 5-7 Carbamate Sites identified across the recombinant histone octamer dataset under the conditions tested.

The carbamate site, H3K79 was chosen for future investigation into carbamylation effects on transcription. H3K79 is a well-characterised methylation site mediated by a singular methyltransferase known as DOT1L and participates in PTM crosstalk with H2BK123 ubiquitination and H4K16 acetylation to alter transcriptional outcomes. The features of H3K79 which make the site interesting for further study in the context of CO_2 are further detailed in section 5.6. Alongside this, inhibitors of DOT1L have been developed and an *in-vitro* proprietary assay for assessing methylation can be applied to a carbamylation context.

5.12 Methyltransferase (MTase) Glo Assay

In this section (5.12), the MTase-Glo assay is discussed, including the principle, validation steps and the effect of normal compared to hypercapnic levels of inorganic carbon on the DOT1L methylation rate. This assay aimed to determine whether H3K79 carbamylation affects the methylation activity of DOT1L.

5.12.1 Assay Principle

Histone methylation requires two substrates and the enzyme specific to the modification site which in this case is Dot1L. The substrates include the methyl donor, SAM which is synthesized from ATP and methionine as shown in Figure 5-31 and the histone modification site which for *in-vitro* assays can take the form of a peptide, protein or in some cases the nucleosome. H3K79 methylation is only feasible when the nucleosome is used due to PTM crosstalk with H2BK123 and H4K16. Therefore, in this study, native nucleosomes (Reaction Biology) were used because the extraction procedures do not modify the PTM landscape whereas recombinant nucleosomes would have been unsuitable due to the lack of post-translational machinery in *E. coli.* The MTase-Glo assay monitors the formation of SAH which is the reaction product remaining after methyl transfer. Figure 5-32 is a depiction of this assay, where SAH is converted to ADP and then ATP which is detected as luminescence by a luciferase reaction.

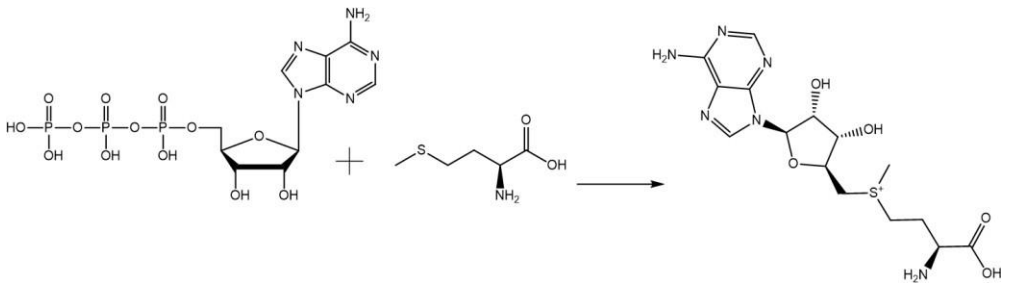


Figure 5-31 S-adenosyl methionine (SAM) produced from adenosine triphosphate and methionine.

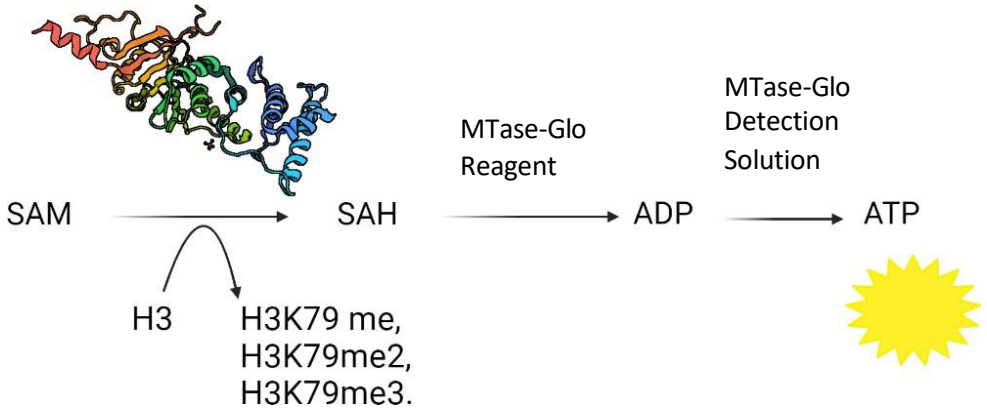


Figure 5-32 Schematic of the MTase-Glo Assay where DOT1L methylates H3K79 and the methylation rate is quantifiable by luminescence. Created using BioRender.

5.12.2 Validation and Suitability of Assay

The MTase-Glo assay was used previously to assess inhibitors of DOT1L activity at HTS, AZ. The assay validation step at HTS showed that pH 8.5 was optimal for DOT1L activity and at physiological pH 7.4 there was no DOT1L activity. (G. Davies, personal communication, 18th May 2022) Therefore, the assay buffer used here was buffered at pH 8.5 and to reflect hypercapnic conditions, a higher starting concentration of inorganic carbon was required at pH 8.5 compared to at pH 7.4. Figure 1-1 and Table 5-8 outline the proportion of dissociated inorganic carbon species at different pH values. The rate equations for deriving these values are described in the cited reference.[275] The hypercapnic Ci concentration used for this experiment was 250 mM which dissociates into a lower concentration of CO_2 than an assay at 7.4 under 50 mM. However, it is hypothesised that at pH 8.5, CO_2 degassing is minimal and the CO_2 will stay in solution for longer than at pH 7.5.

pH	Ci/ mM	CO_2/ mM	HCO_3^-/mM	CO_3^-/mM
7.4	20	1.5	18.5	0
8.5	20	0.1	19.6	0.3
7.4	50	2.9	47.0	0.1
8.5	250	1.5	244.6	3.9

Table 5-8 The dissociation of inorganic carbon (Ci) into ionic species and CO_2 at different pHs.

The pH stability of the assay buffer was tested over the inorganic carbon range of the assay between 0 and 250 mM inorganic carbon (CO_2/HCO_3^-) to isolate CO_2 effects from CO_2- associated acidosis. The results of this are shown in Figure 5-33, it is important to note that the total anion concentration was kept consistent across all conditions tested, by supplementation with NaCl. The change in pH is within 0.1 units or less between the Ci concentrations tested. Therefore, it was concluded that the buffering capacity was sufficient for the range of inorganic carbon concentrations across the assay timescale.

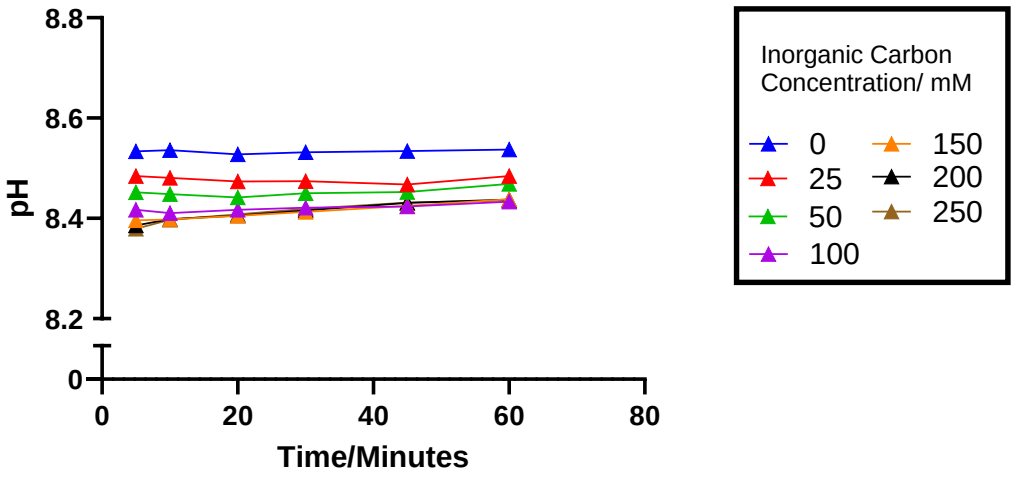


Figure 5-33 The measured pH of the MTase-Glo assay buffer versus the time in minutes to assess pH buffer stability under 0-250 mM CO_2/HCO_3^- supplemented with the Cl^- anion. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points.

For accurate quantification of H3K79 methylation by DOT1L it was essential to operate within the linear range of the kinetic reaction. Michaelis-Menten kinetic assays were applied, to determine the enzyme kinetic constants: the maximum velocity (V_{max}) and Michaelmas constant (K_M) concentrations for DOT1L under set concentrations of nucleosome and SAM. The V_{max} describes the maximum reaction rate where the enzyme is completely saturated by the substrate and increasing the

substrate concentration beyond this point without increasing the enzyme concentration will have no effect. The K_M is defined as the concentration of the substrate which permits the enzyme to achieve half of V_{max}.

The substrate, SAM was supplied in excess, to saturate the DOT1L enzyme. This was particularly important in the context of carbamylation because the N-terminal amine site of SAM has a pKa of 9 which under a structurally privileged environment via interaction with other biomolecules could be carbamylated and deplete the presence of SAM by altering the structure. However, when SAM concentration is above V_{max} the carbamylation reaction on SAM would not outcompete methyl group dissociation on SAM due to the reactivity of the positively charged sulfur atom. The excess concentration of SAM used for the assay was defined by Promega as 10 µM.

Previous HTS validation had determined that a nucleosome concentration of 0.05 mg/ml with a Dot1L concentration of 3.5 nM was within the linear range of the assay. Initially, 0.05 mg/ml of nucleosome across a 0-20 nM range of DOT1L was tested, however, the production of SAH was outside the linear range of the assay. After, troubleshooting, the appropriate conditions were identified as shown in Figure 5-34. In Figure 5-34, luminescence is doubled between 0.1 mg/ml compared with 0.05 mg/ml of nucleosome under 10 nM enzyme incubation. The concentrations of 0.1 mg/ml nucleosome and 10 nM Dot1L were selected for future use in the bicarbonate assay.

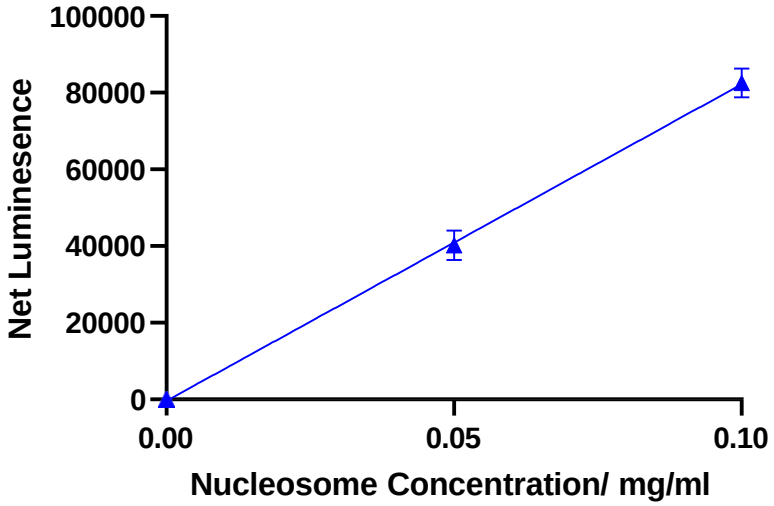


Figure 5-34 Net luminescence produced from the DOT1L methyltransferase reaction versus the nucleosome concentration where enzyme concentration is constant. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. A line of best fit is drawn displaying the linear relationship between these conditions.

An important consideration for the experiment design was whether trifluoracetic acid (TFA) could effectively quench the first step of the reaction. Figure 5-35 shows the results of stopping the methyl transfer at different reaction lengths using TFA. The net luminescence produced is proportional to the reaction length with the lowest signal at 20 minutes and the highest signal at 60 minutes. The results indicate that 0.5% TFA is acidic enough to inactivate Dot1L and stop the production of SAH. Effective TFA treatment is essential for staggering the first reaction step across an experiment. The MTase-Glo reagent followed by the detection solution can be added uniformly across the experiment after all methyltransferase incubations are complete. The result in Figure 5-35 also shows that the luminescence signal is stable for at least three hours however plates within an experiment were read at the same time to maintain consistency across the experiment.

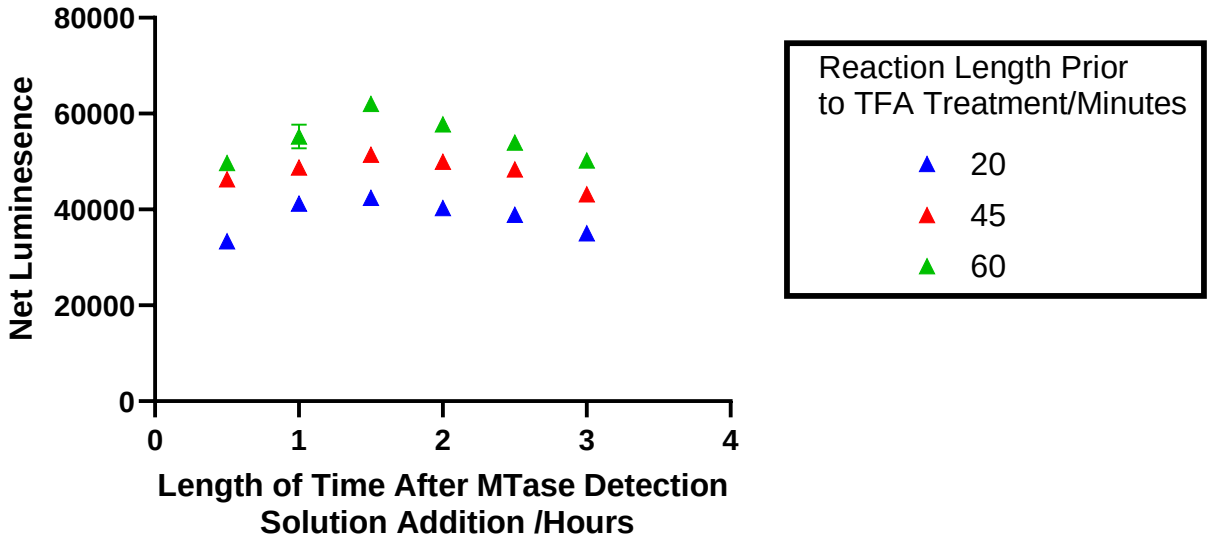


Figure 5-35 The stability of the net luminescence signal detected versus the time after the MTase detection solution was added for three different methyltransferase reaction lengths stopped by 0.5% TFA using 10 nM Dot1L and 0.05 mg/ml nucleosome. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points.

The selected inorganic carbon concentration range for this assay was 0-250 mM. As discussed previously the total anionic concentration must be consistent across the experiment and is accounted for by changing the proportion of NaCl. The Promega protocol for the MTase-Glo assay, suggests a total anionic concentration of 50 mM. Therefore, the consistency in luminescence signal between higher anionic concentrations was assessed. Figure 5-36 shows the net luminescence signal across three repeats for the methyltransferase reaction under two different anionic concentrations. An unpaired t-test showed there was no significant difference between the luminescence values for 250 mM anionic and 50 mM anionic buffering using NaCl. This result shows that it is possible to perform the MTase-Glo assay at an anionic concentration of 250 mM.

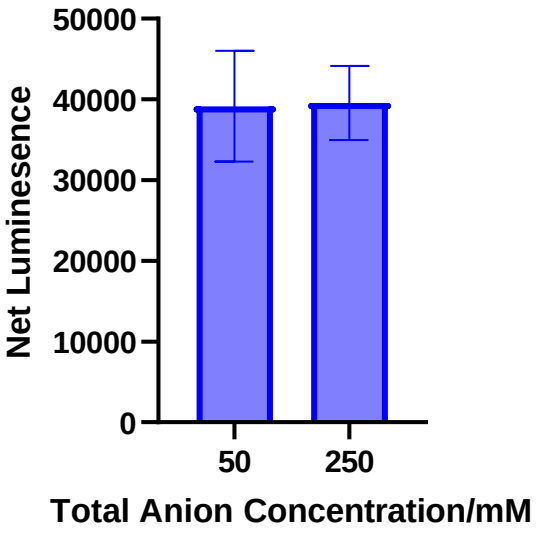


Figure 5-36 Net Luminescence versus the total anion concentrations (NaCl) for 0.05 mg/ml nucleosome and 10 nM Dot1L produced by the MTase-Glo assay normalised to the background luminescence produced without the nucleosome substrate. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. Data passed the Shapiro-Wilks test for normality. There was no statistically significant difference in the luminescence values produced by the two anion concentrations as shown by an unpaired t-test with a significance threshold of $p < 0.05$.

In summary, the validation steps have shown this is a suitable approach for assessing the effect of Ci on DOT1L methyltransferase activity. The buffering capacity of the MTase reaction buffer is high enough for the inorganic carbon concentrations selected. TFA treatment stops the methyltransferase reaction and therefore a time course of methyltransferase activity can be studied. Finally, the concentrations of substrate and enzyme were determined within the reaction's linear range and can be performed at an anionic concentration of 250 mM which is suitable to reflect hypercapnic conditions at pH 8.5.

5.12.3 Results Under Varying Inorganic Carbon Concentrations

The MTase-Glo assay was performed using a Ci concentration range of 0-250 mM across three methyltransferase incubation time points in triplicate. In this experiment, the S-adenosyl homocysteine standard curve (Figure 8-38) was used to calculate the concentration of SAH from the luminescence value obtained (Figure 8-39). Figure 5-37 is a plot of the amount of SAH produced from methyl transfer under varying Ci concentrations. The data shows that under elevated physiological levels of inorganic carbon, the methyltransferase activity of Dot1L on H3K79 is stimulated. All the 25 - 250 mM Ci data points when compared to 0 mM Ci showed a statistically significant increase in SAH concentration. However, this effect becomes saturated because there is no statistical difference between the consecutive Ci concentrations tested after 25 mM Ci across all three incubation time points as determined by multiple comparison tests (MCTs).

Following the result in Figure 5-37, a further experiment using smaller incremental increases of Ci was conducted. The concentration range selected was 0 - 150 mM Ci with a total of eight Ci concentrations. Alongside the concentration range modification, another incubation time point was added for the methyltransferase reaction, otherwise, all other parameters remained constant. The amount of SAH produced was calculated from the luminescence values and the SAH standard curve (Figures 8-40 and 8-41) as shown in Figure 5-38. The Dot1L stimulation was statistically significant across all the Ci concentrations compared with 0 mM. The effect was saturated in this experiment at around 75 mM Ci. This was determined by MCTs whereby there is no significant increase in SAH concentration between 75 mM and the higher Ci concentrations but there are significant differences between consecutive Ci concentrations below a Ci of 75 mM.

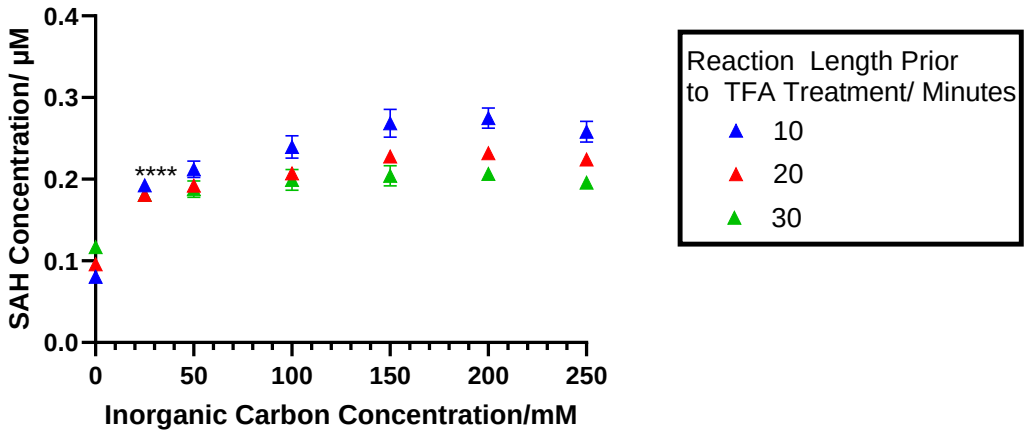


Figure 5-37 The concentration of SAH produced by methyltransferase reactions incubated for 10,20 and 30 minutes versus the inorganic carbon concentration. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. A one-way ANOVA assessment showed there was a statistically significant difference between samples incubated with and without inorganic carbon across all three time points at a significance threshold of $p < 0.05$. Multiple comparison tests showed that the production of SAH was statistically significant between 0 and 25 mM but not between other consecutive Ci treatments across all three time points. Asterisks indicate levels of significance (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$).

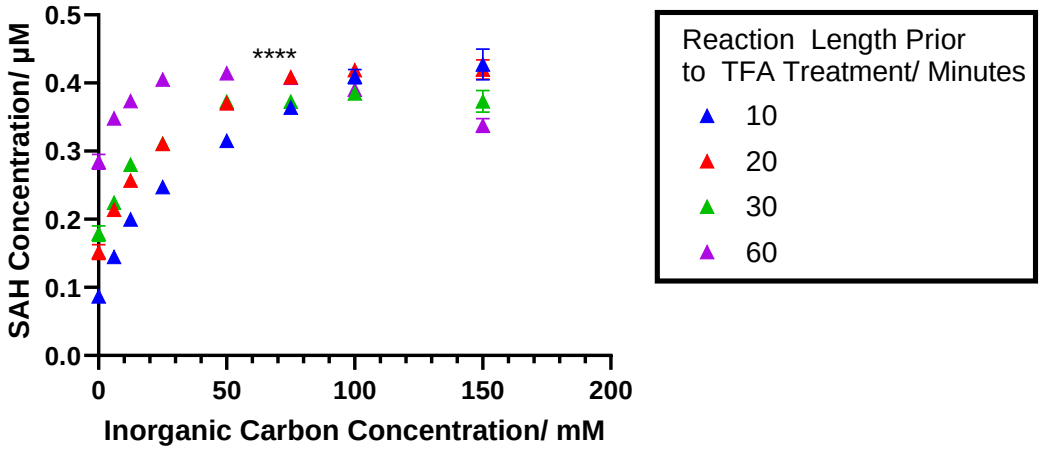


Figure 5-38 The concentration of SAH produced by methyltransferase reactions incubated for 10, 20, 30 and 60 minutes versus the inorganic carbon concentration. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. A one-way ANOVA assessment showed there was a statistically significant difference between samples incubated with and without inorganic carbon across all three time points at a significance threshold of $p < 0.05$. Multiple comparison tests showed that the production of SAH was statistically significant between up to 75 mM but not between higher consecutive Ci treatments across all three time points. Asterisks indicate levels of significance (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$).

5.12.4 MTase-Glo Assay Discussion

The data obtained from this *in-vitro* assay indicates that Ci concentration alters Dot1L methylation activity. Control experiments were completed to ensure this was not an artefact of the assay and was linked to the methyltransferase reaction these can be seen in Appendix Figures 8-42 – 8-46. It is important to note that the control assays indicated a nucleosome-only dependent increase in the luminescence signal recorded between a low and high concentration of Ci (Figure 8-45 - 8-46). Therefore, there is an unspecific ATP reaction also driving the luminescence signal alongside the methyltransferase activity. To understand the impact of the unspecific ATP reaction on luminescence, the proportional increase in luminescence across the Ci concentration range at the 20 minute incubation time point of the control (Figure 8-46A) and test assay (Figure 5-38) were compared. When all three components for the assay are present the signal increase between 0 and 150 mM Ci was 3-fold higher compared to the control with both substrates present but without Dot1L the signal was only 1.3-fold higher. This 1.3-fold difference would have remained at the same magnitude when all three components were present for the methyltransferase reaction if only the unspecific ATP production reaction was driving the signal. However, the 3-fold difference in luminescence between 0 - 150 mM Ci when all three components were present indicated that the methyltransferase reaction was stimulated at higher Ci concentrations. Figure 8-45 supports this finding because when Dot1L was inhibited the increase in the luminescence was 1.6-fold between 0 and 150 mM Ci. Under Dot1L inhibition the signal was remarkably close to background levels and the production of SAH was 10-fold lower than when all three components are present indicating that this unspecific reaction is not solely responsible for the luminescence increase at higher Ci concentrations when all components are present.

The hypothesis for this experiment was that carbamylation at H3K79 would inhibit Dot1L activity via substrate depletion however the opposite result was shown by the data. Dot1L stimulation under increased Ci concentration plateaued within the detection range of luminescence as shown by the SAH standard curve. There are several explanations for Dot1L stimulation under increased Ci

concentration. Firstly, the spontaneous carbamate modification is liable, and the carbamylated H3K79 site could be involved in the recruitment of Dot1L, however before Dot1L binding the carbamate group dissociates to enable H3K79 methylation. Alternatively, carbamylation could be happening at another nucleosome site and stabilising DOT1L binding. This hypothesis is supported by the fact that H3K79 methylation is dependent on nucleosome PTM crosstalk and there may be an unidentified carbamylated modification site responsible for these effects or a carbamylation-mediated change in proportion of H2BK123 ubiquitination or H4K16 acetylation. Nucleosome mutant testing in the MTase-Glo assay could provide insight into the mechanism of Dot1L stimulation under increased Ci. Lastly, it is important to consider the distributive nature of the DOT1L enzyme when interpreting the results because the luminescence recorded is from the composite of methyl forms (mono, di and tri). Therefore, a targeted antibody approach would need to be used to identify the effect of Ci concentration on the methylation type.

5.13 In-cellulo Validation of H3K79 Carbamylation Effects

The aim of this section (5.13) was to demonstrate the *in-cellulo* relevance of the *in-vitro* methyltransferase assay result. An experiment was devised using HEK293 cells treated with pinometostat for a defined incubation time and concentration known to cause Dot1L methyltransferase inhibition. Inhibited and non-inhibited cells were incubated at normal and hypercapnic levels of PCO_2 to assess the transcriptional change between samples. The first step in this process was to determine a non-toxic incubation length and concentration of pinometostat. Secondly, qPCR reactions to assess transcriptional change due to CO_2 incubation was trialled. Finally, RNA was extracted from cells treated with and without the Dot1L inhibitor exposed to normal and hypercapnic levels of CO_2 for three and six hours and analysed by RNA sequencing.

5.13.1 MTT Assay

The MTT assay is a technique used to determine cell viability. The MTT assay quantifies the metabolic activity of cells using absorbance. The MTT assay quantitates the total metabolism which is directly proportional to the viable cell number. Whereby, viable cells convert tetrazolium to formazan which results in a detectable colour change from yellow to purple.[276] Elevated CO_2 levels have been linked to an altered metabolic activity [53,259] therefore this assay was performed only at normoxic CO_2.

This method was applied to HEK293 cells treated with the DOT1L inhibitor, pinometostat for 2, 3, 4, 7, and 10 days at a concentration of 0.5 µM or 1 µM. The range of conditions was determined from previous *in-cellulo* studies using pinometostat as a DOT1L inhibitor in various cell lines as shown in Table 5-9.

Cell line	Concentration of DOT1L inhibitor/µM
231	1 for 10 days [277]
468	0.5 for 10 days. [277]
MV4-11	0.1 for 4 days[278]
HL-60	1 for 4 days [279]

Table 5-9 Incubation lengths and concentration of pinometostat used in various cell lines with relevant studies cited in brackets.

Figure 5-39 is the percentage of cell viability across the conditions tested normalised to a daily DMSO control. This experiment showed that cell viability decreases significantly between DMSO and pinometostat treatment on day 4 however by day 7 the cells become used to the inhibitor treatment and viability increases again and by day 10, there is no significant difference in cells viability between DMSO and pinometostat treatment.

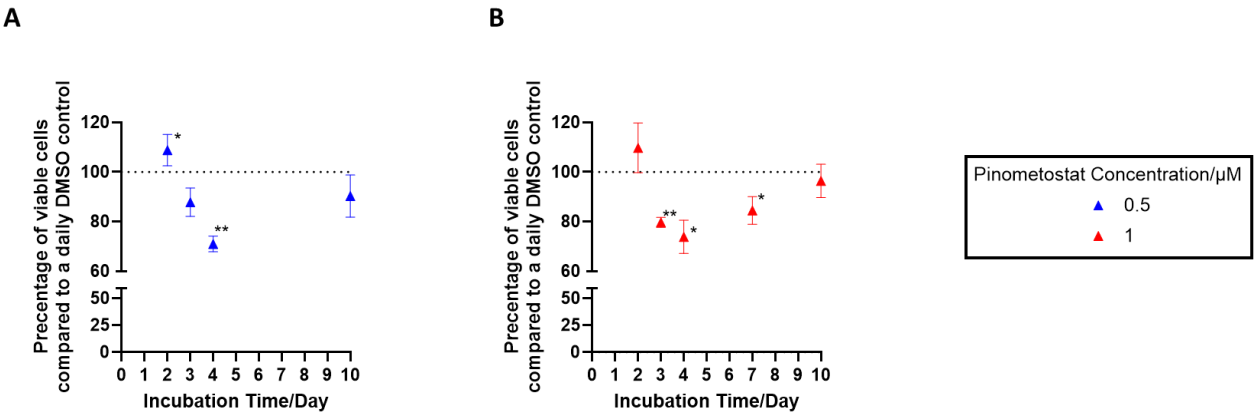


Figure 5-39 The percentage viability of HEK2933 cells compared to a daily DMSO control versus the pinometostat incubation length for 2,3,4,7, and 10 days at (A) 0.5 µM (blue) and (B) 1 µM (red). All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. Data passed the Shapiro-Wilks test for normality and a one-sample t-test comparing the data points to a hypothetical mean of 100 (represented by the dashed line) was applied at a significance threshold of p>0.05 and n=3. Asterisks indicate levels of significance (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$).

The supplementary Figure 8-47 shows the pre-normalisation absorbance data which was assessed using one-way ANOVA analysis and MCTs. There is a slight difference in the significance results for day 2 in the raw data (Figure 8-47) compared with the percentage viability data (Figure 5-39). This can be explained by the fact that the data undergo different statistical testing. The raw data is tested by considering the change in absorbance between samples whereas the percentage viability data compares data points to the theoretical mean of 100% viability. The normalisation method does not consider the variability in the DMSO reading for the statistical testing because the average of the readings is calculated and assigned to be 100%. Importantly, the statistical tests on both the raw and normalised data show that by day 10 there is no significant difference in the number of metabolically viable cells between the conditions tested.

5.13.2 Enzyme-linked Immunosorbent Assay (ELISA) for Testing H3K79 Methylation State of Histone Extracts from Pinometostat-treated HEK293 Cells.

Methylation changes due to pinometostat treatment were explored using an enzyme-linked immunosorbent assay (ELISA). An ELISA is an antibody-based technique, where assay plates are precoated with an antibody specific to a protein or a protein PTM of interest. The analyte binds to the plate and the captured protein is detected by a detection antibody, which catalyses an enzymatic reaction and produces a quantifiable absorbance value. In the experiment discussed in this section (5.13.3), the intensity of the absorbance measured at a wavelength of 450 nm is proportional to the amount of H3K79 mono, di or tri methylation. Figure 8-48 displays the raw absorbance values obtained from DMSO and pinometostat-treated HEK293 lysate analytes at various incubation time points.

The raw absorbance data needed to be normalised however the commercial ELISA plate did not contain wells which detected the total H3 content. The ELISA plate was loaded with an identical concentration of HEK293 lysate for each sample and therefore the H3 content was assumed to be consistent across samples as advised by the supplier. The relative percentage of each methylation form of H3K79 under the range of treatment conditions was calculated by normalisation to the relevant daily DMSO control treatment which was assigned a value of 100%. Figure 5-40 displays the change in methylation when using the inhibitor treatment at the concentrations 0.5 µM and 1 µM for 2, 3, 4, 7, and 10 days across the three forms of H3K79 methylation. Due to the fixed number of antibody wells on the pre-coated plate, not all degrees of methylation could be quantified for each condition in the experiment. The treatment condition with the greatest marked decrease in methylation across all methylation degrees was the 10-day incubation with 1 µM pinometostat when compared with DMSO treatment. This result together with the MTT assay determined that this was a suitable treatment condition for downstream experiments.

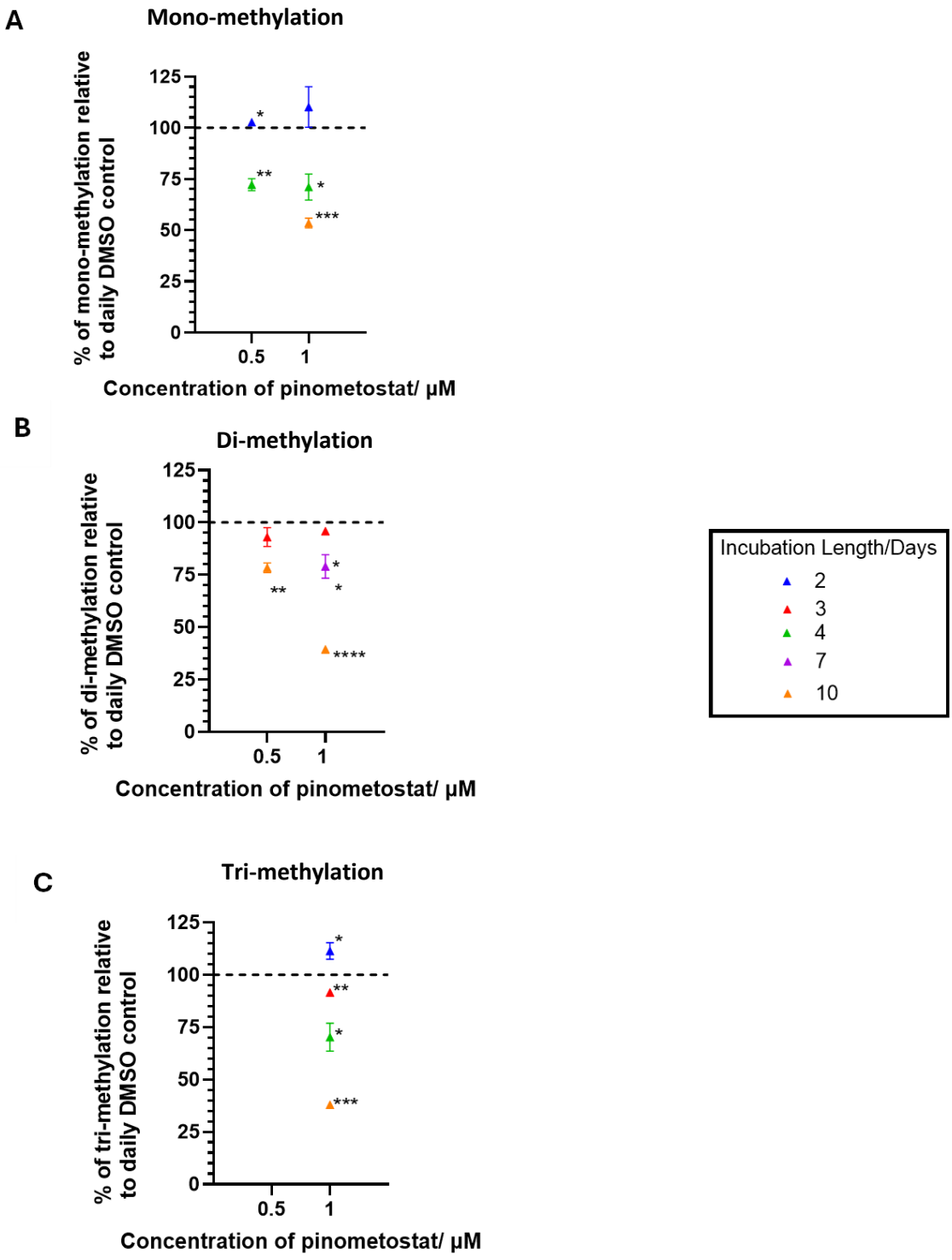


Figure 5-40 The percentage of H3K79 methylation where the relevant daily DMSO control represents 100% methylation versus the pinometostat concentration at various incubation lengths where (A) represents mono-methylation (B) di-methylation and (C) tri-methylation. Due to the fixed number of antibody wells on the pre-coated plate, not all degrees of methylation could be quantified for each

condition in this experiment. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. All data passed the Shapiro-Wilks test, and one sample t-tests were performed against a hypothetical mean of 100% representing the DMSO control for each treatment condition. Asterisks indicate levels of significance (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$).

5.13.3 Quantitative Polymerase Chain Reaction (qPCR)

The aim of this section (5.13.4) was to identify a suitable CO_2 exposure time point that leads to a hypercapnic-induced transcriptional change in HEK293 cells using qPCR. The Wnt signalling genes, *Fzd9* and *Wnt7a* were chosen for this experiment as previously mentioned in section 5.7. *Fzd9* and *Wnt7a* were identified as CO_2-sensitive genes by Shigemura *et al*.[258] in a multi-tissue microarray and validated by qPCR in MLE-12 and ASM cell lines. Shigemura *et al*. reported the largest fold change in *Fzd9* and *Wnt7a* expression at the six-hour incubation time point between 5% and 20% PCO_2 (Figure 8-49). *Fzd9* and *Wnt7a* were upregulated at 20% PCO_2 when compared with 5% PCO_2.

The principle of qPCR relies on a fluorescence reporter to detect the amount of amplicon after each PCR cycle. In this investigation, SYBR-green was used, which is a DNA-binding dye that fluoresces when bound to double-stranded DNA (dsDNA). The PCR amplicon product will exponentially increase for each reaction cycle and then plateau as the reaction components are used up. The data is plotted as an amplification curve where the PCR cycle number is plotted against the SYBR green fluorescence. A key parameter in qPCR analysis is the quantitation cycle number (Cq). The Cq can be defined as the cycle number at which the fluorescence signal detected is above the threshold background fluorescence set by the reaction cycler. Cq values are a direct measure of the amount of target cDNA in the sample, where high expression genes are associated with low Cq values because they are detected at an earlier cycle number compared to low expression genes which are associated with high Cq values and are detected at a later cycle number. Cq values can be compared between samples when normalised to a housekeeping gene to identify fold changes in expression.

5.13.1.1 qPCR Optimisation and Validation

The cDNA used for qPCR reactions was synthesised from RNA extracted from HEK293 cells incubated at 5, 10 or 20% PCO_2 across five different time points and treated with DMSO as described in sections 2.5.6.5- 2.5.6.6. The concentration of cDNA used in qPCR reactions was determined at the

RNA stage. During RNA preparation, an extra step was implemented to remove all interfering genomic DNA (gDNA) as shown in the agarose gel in the supplementary information Figure 8-50.

The initial qPCR reactions used primer sequences and reaction conditions obtained from Shigemura *et al.*'s protocol, and the primer sequences are given in the supplementary information in Table 8-9 (Set 1 for each gene). The primer concentration was 500 nM, and the cDNA concentration was 2.5 ng/µl and the other reaction conditions are detailed in section 2.5.6.7.

The amplification curves for the two Wnt signalling genes (*Fzd9* and *Wnt7a*) and the housekeeping gene, ribosomal protein L19 (*RPL19*) alongside the control curves for each primer set without template cDNA are shown in Figure 5-41. The Cq values associated with each amplification curve are given in Table 5-10 and the disparity between them can be explained by RNA expression levels for each gene in HEK293 cells. The RNA expression levels for these three genes were obtained from the human protein atlas [280] and given in Table 5-10 as the normalised number of transcripts of the gene of interest per one million full-length RNA transcripts (nTPM). Table 5-11 details the Cq values for the control amplification plots, when no cDNA is present in the reaction.

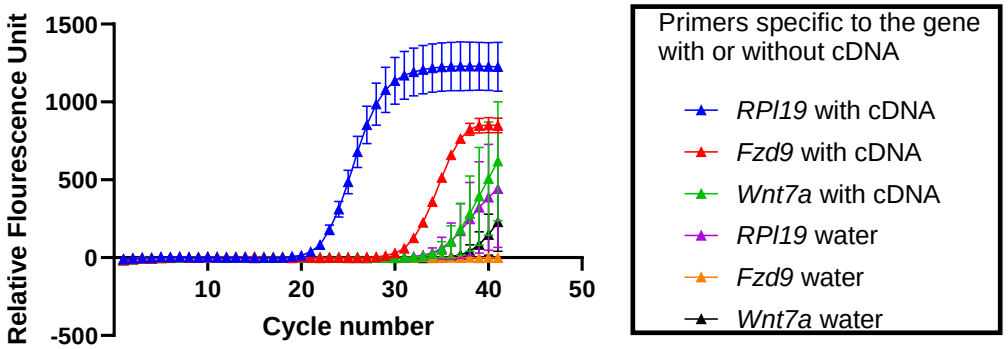


Figure 5-41 qPCR amplification curves plotted from the relative fluorescence are plotted against the PCR cycle number for three genes of interest in the presence and absence of cDNA. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points.

Gene	Cq	Normalised target transcripts per million (nTPM) in a HEK293 RNA sample	Expression Level
RPL19	20.6	4625	High
Fzd9	29.9	4.2	Low
Wnt7a	35.1	0	Low

Table 5-10 Genes of interest with associated Cq values and expression levels in HEK293 cells.

Control experiments with primers but excluding DNA	Cq
RPL19	33.3
Fzd9	0 – not detectable
Wnt7a	38.0

Table 5-11 Control Cq values from amplification plots where the primer targeting the gene is present, but cDNA is absent and instead, water is used.

The next step for optimising the qPCR reaction was to reduce the signal from the no cDNA template control. The *RPL19* primers without cDNA have a lower Cq value (33) than the Cq value for *Wnt7a* primers with the cDNA template (35). Therefore, the qPCR reaction products from Figure 5-41 were run on an agarose gel to test for template cDNA contamination in the controls and this gel is shown in Figure 5-42. No PCR products are detected by the agarose gel in the control samples but for the reactions with cDNA present, PCR products are detected between 100-150 bp for all three genes. This result indicates that DNA detection using an agarose gel is not as sensitive a technique as the amplification plot produced from the reaction cycler. The quantitative detection of *Wnt7a* from HEK293 cDNA is a challenge due to having a reported nTPM value of 0 in HEK293 cells and therefore was not used in further qPCR experiments.

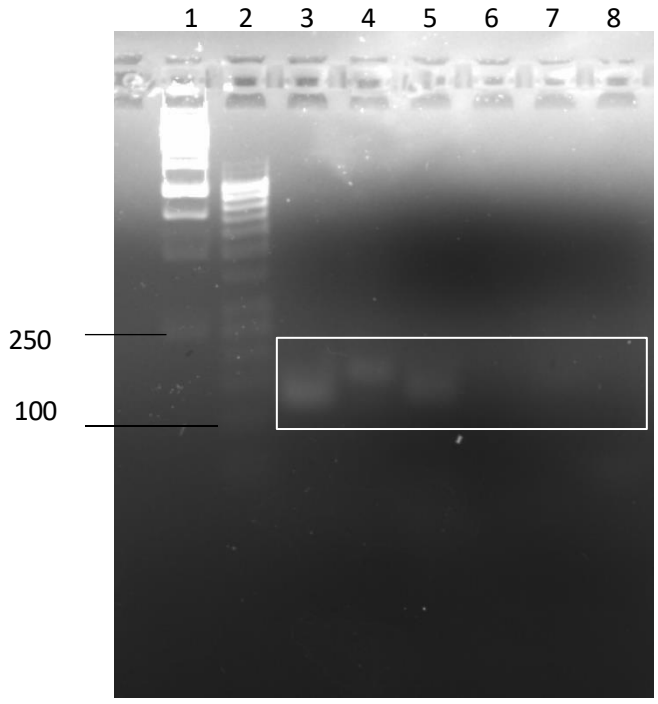


Figure 5-42 Agarose gel of PCR amplicons from the initial qPCR experiment. Lanes 1 and 2, are base pair markers of 1kb and 50 bp, respectively. Lane 1 contains DNA fragments at 10,000, 8,000, 6,000, 5,000, 4,000, 3,500, 3,000, 2,500, 2,000, 1,500, 1,000, 750, 500 and 250 bp where 250 bp is labelled. Lane 2 contains DNA fragments at 1,350, 916, 766, 700, 650, 600, 550, 500, 450, 400, 350, 300, 250, 200, 150, 100 and 50 bp where 100 bp is labelled. Lanes 3-8 are the qPCR products from samples run with (3-5) and without DNA (6-8) using *RPL19* (3,6), *Fzd9* (4,7) and *Wnt7a* (5,8) primers. The white box highlights the area at which the DNA sequence of interest would run.

Primer hybridisation is a contaminating signal source that can arise from primer sequence or concentration. To check for this type of signal contamination, a step was added to the protocol after the qPCR amplification reaction whereby the temperature of the sample is increased incrementally, and the fluorescence signal is measured at increments to produce a melt curve. A single peak represents a specific amplicon product, whilst multiple peaks indicate that other side reactions are occurring, and that the primer sequence should be reconsidered.

Three primer sequences given in Table 8-9 for *RPL19* and *Fzd9* were tested to identify whether any of these primer variants produced less background fluorescence in the no cDNA template control.

Figures 5-43 and 5-44 show the amplification and melt curves for the *Fzd9* and *RPL19* primers tested, respectively. Analysis of this dataset indicated that further qPCR reactions should be run with primer set 2 for both amplicon targets. From the *RPL19* primer sets tested, set 2 had the best separation between Cq values between the test and control samples. The *RPL19* amplification plot without cDNA for primer set 2 reached the Cq at cycle number 39 however with cDNA the Cq was 21.7. For the *Fzd9* gene, the primer sets all showed similar profiles for the amplification plot and no Cq was assigned for the without cDNA samples. However, the melt curves for the *Fzd9* primer set 1 did not exhibit a symmetrical peak synonymous with one PCR product. The melt curve profile for the *Fzd9* primer sets 2 and 3 had one single peak and either primer could have been used in further reactions. The agarose gel in Figure 5-45 supports the melt curve analyses that one specific amplicon product is produced with each primer set tested for *RPL19* and *Fzd9*.

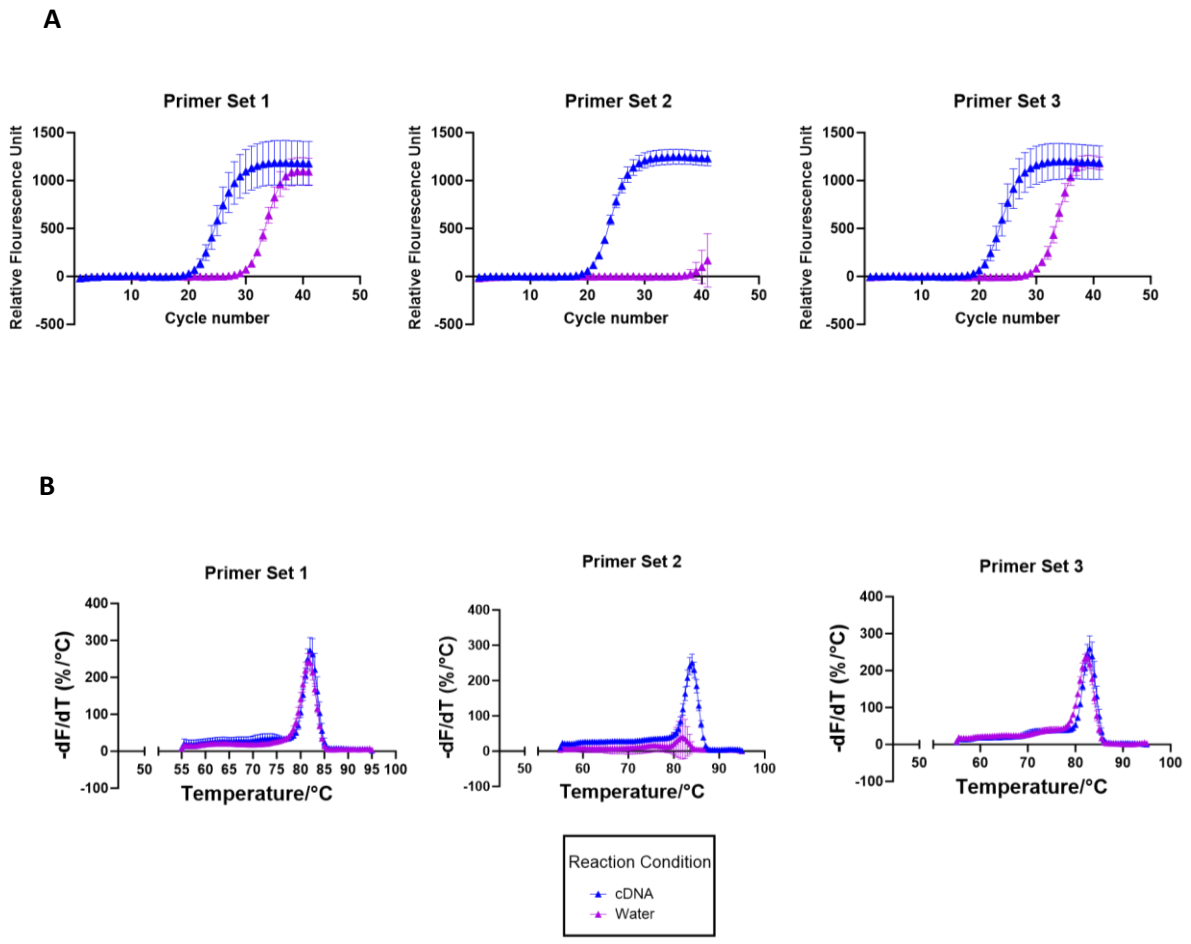


Figure 5-43 qPCR primers tested for targeting the *RPL19* amplicon in a reaction with and without cDNA.

Blue curves are those with cDNA in the reaction mixture and the purple curves are those without cDNA where cDNA was substituted with an equal volume of water. A) are amplification curves where the relative fluorescence is plotted against the cycle number and (B) are melt curves where the negative derivative of fluorescence (F) with respect to temperature (T) otherwise known as the rate of change in fluorescence (-dF/dT) is plotted against the temperature (°C). All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points.

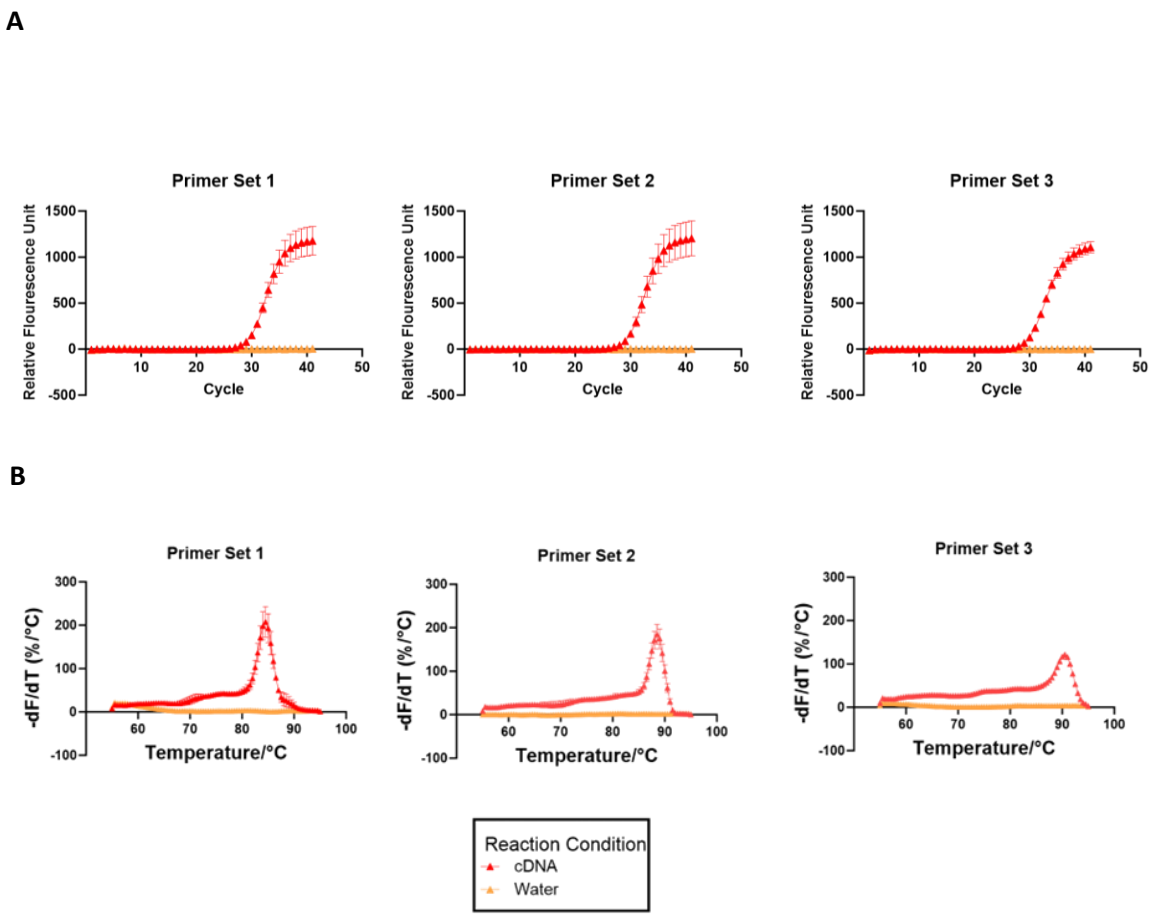


Figure 5-44 qPCR primers tested for targeting the *Fzd9* amplicon in a reaction with and without cDNA. Red curves are those with cDNA in the reaction mixture and the orange curves are those without cDNA where cDNA was substituted with an equal volume of water. (A) are amplification curves where the relative fluorescence is plotted against the cycle number and (B) are melt curves where the negative derivative of fluorescence (F) with respect to temperature (T) otherwise known as the rate of change in fluorescence (-dF/dT) is plotted against the temperature (°C). All values are represented as mean

with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points.

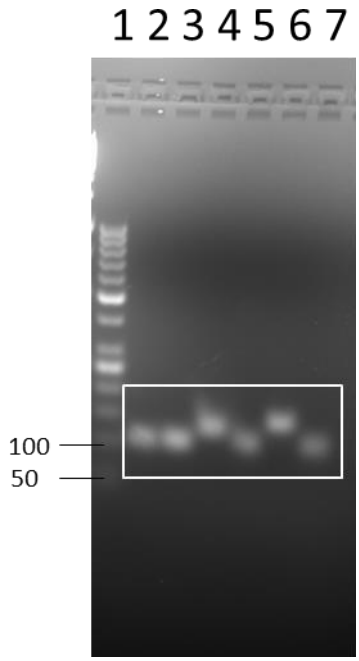


Figure 5-45 Agarose gel of PCR amplicons from the primer testing qPCR experiment. Lane 1 is a base pair marker 50 bp and contains DNA fragments at 1,350, 916, 766, 700, 650, 600, 550, 500, 450, 400, 350, 300, 250, 200, 150, 100 and 50 bp with 100 and 50 bp labelled. Lanes 2-7 are the qPCR products from samples run with cDNA using *RPL19* primer set 1 (2), primer set 2 (3), primer set 3 (4) and *Fzd9* primer set 1 (5), primer set 2 (6) and primer set 3 (7). The white box highlights the area at which the DNA sequence of interest would run.

Despite the success in limiting the background luminescence in the no cDNA template controls using primer set two, the problem of *RPL19* control experiment Cq values being close to the *Fzd9* Cq values reoccurred. Instead of altering primer sequences the next optimisation step was to identify the optimal concentration of primer and cDNA to use. Figures 5-46 and 5-47 show the melt curves (A) and threshold Cq values (B) with (2.5 ng/ µl) and without cDNA for a range of primer concentrations and for a range of cDNA concentrations with 250 nM of primer, respectively.

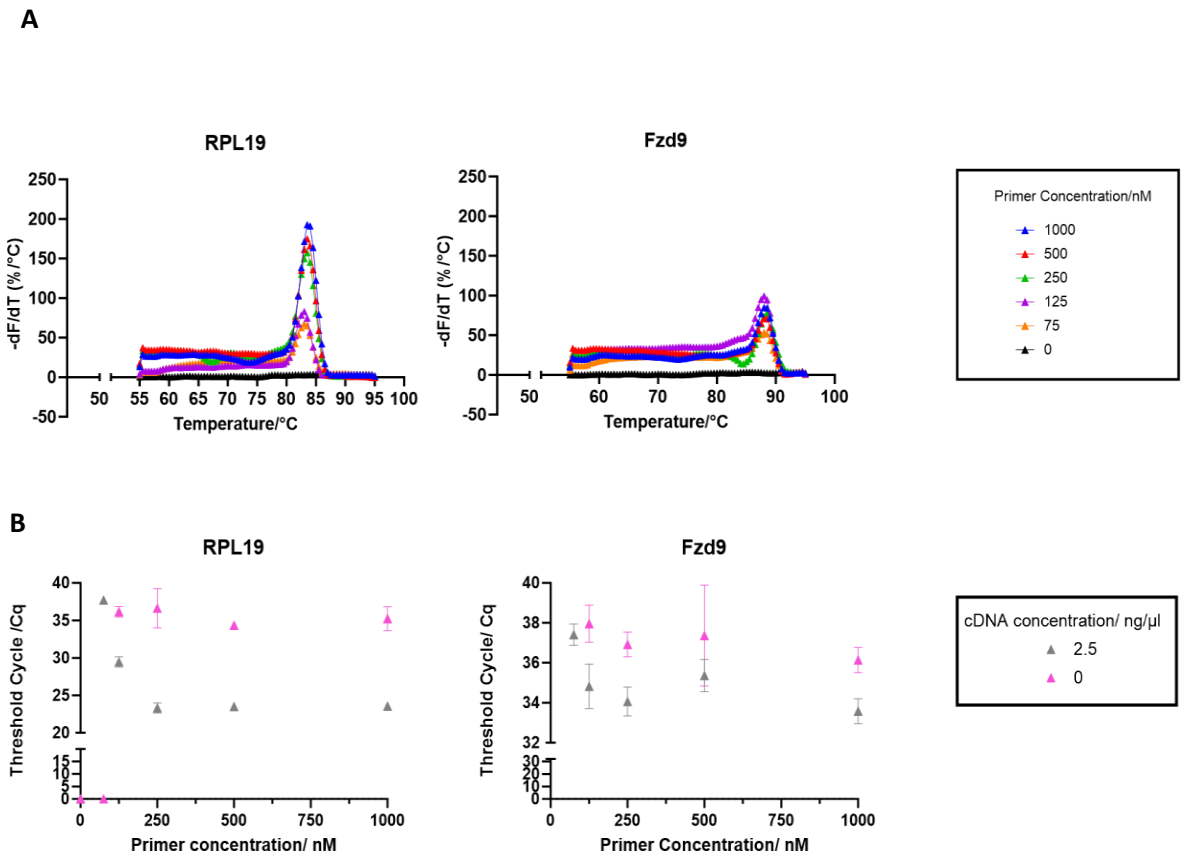


Figure 5-46 Relative amount of target amplicons *RPL19* and *Fzd9* produced under varying primer concentrations. (A) melt curves for *RPL19* and *Fzd9* targets are plotted where the negative derivative of fluorescence (F) with respect to temperature (T) otherwise known as the rate of change in fluorescence (-dF/dT) is plotted against the temperature (°C). The legend displays the primer concentration used. (B) a scatter plot of threshold cycle numbers against primer concentration with (grey) and without (pink) cDNA. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points.

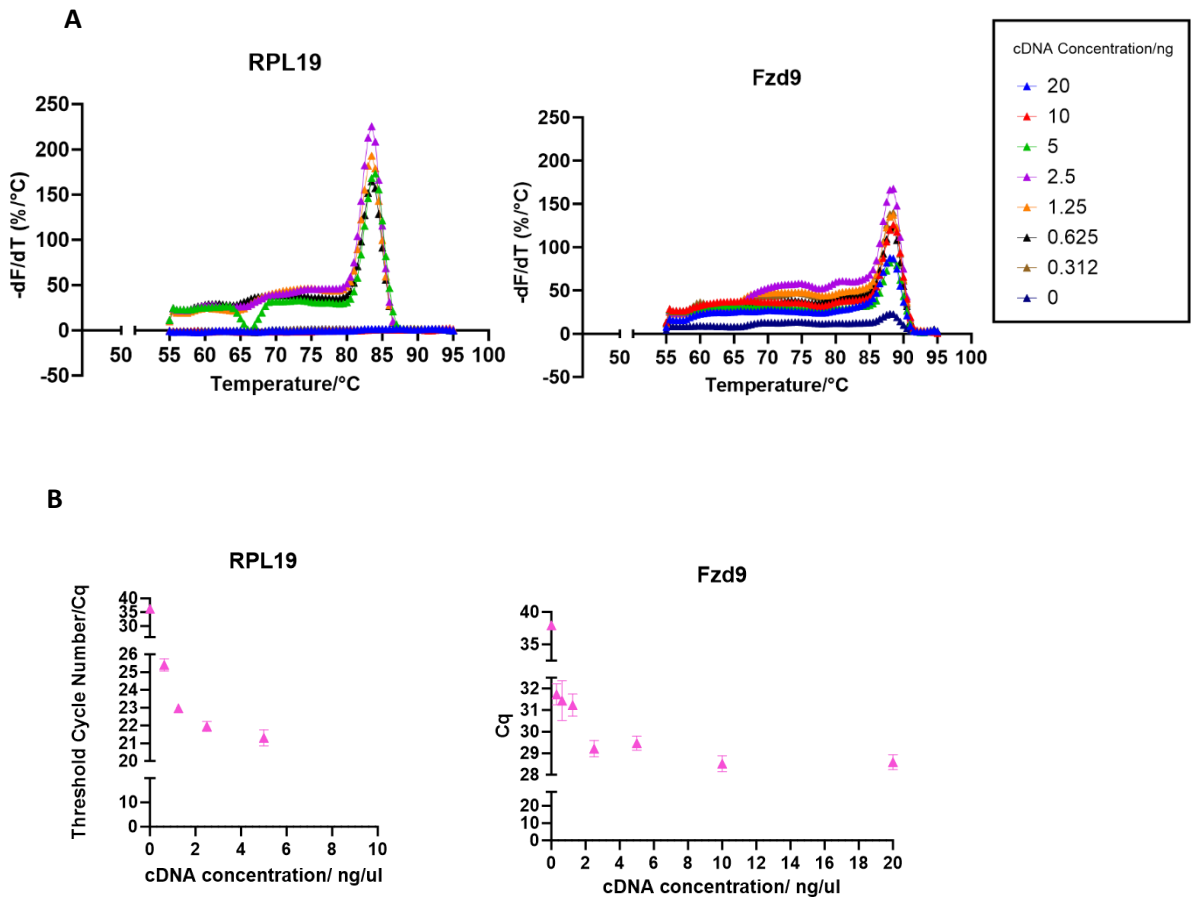


Figure 5-47 Relative amount of target amplicon produced under varying cDNA concentrations. (A) melt curves for *RPL19* and *Fzd9* targets are plotted where the negative derivative of fluorescence (F) with respect to temperature (T) otherwise known as the rate of change in fluorescence (-dF/dT) is plotted against the temperature (°C). The legend displays the concentration of cDNA used. (B) a scatter plot of threshold cycle numbers against cDNA concentration using 250 nM of primer. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points.

To identify an optimal primer concentration, it was important to consider the extremes. Background contamination from gDNA traces in the control samples will be detected faster on the amplification curve if the primer concentration is too high. When the primer concentration is too low it will limit the qPCR reaction and cDNA containing test samples will amplify at later cycles than expected. In this work, the primer concentration selected for further qPCR reactions was 250 nM for

both amplicons. For *RPL19*, 23.5 was the average Cq value for a reaction with 2.5 ng/µl cDNA and 250 nM primer and the same reaction without cDNA had an average Cq value of 36.6. For *Fzd9* the average Cq value for a reaction with 2.5 ng/µl cDNA and 250 nM primer was 34 and for no cDNA, it was 37. For both amplicons, the signal response at 250 nM of primer with cDNA was similar in magnitude to the response for higher primer concentration samples at 500 and 1000 nM. In addition, at 250 nM primer concentration, the best separation between Cq values for the test and control samples was seen.

The principles applied to the selection of a primer concentration were also applied to identify the optimal cDNA concentration. In theory, the Cq value should increase by one when the concentration of cDNA is halved. The *RPL19* cDNA dilution series exhibited this linear relationship between three of the data points collected. Table 5-12 shows that at 5ng/µl cDNA the Cq value was 21, for 2.5 ng/µl it was 22 and for 1.25 ng/µl it was 23, 0.625ng/µl lay just outside this linear range and was at a Cq of 25. For the cDNA concentrations tested at 10 and 20 ng/ µl no Cq value was detected, indicating that the ratio between primer and cDNA for the highly expressing *RPL19* was unsuitable. For *Fzd9* the cDNA dilution series did not give the linear relationship expected indicating that the lower expression of the *Fzd9* as shown in Table 5-10 is a limiting factor in this experiment. For consistency across future qPCR reactions, the 2.5 ng/µl cDNA concentration was selected for both amplicons.

cDNA concentration/ng/µl	*RPL19* Cq	*Fzd9* Cq
20	-	28.6
10	-	28.5
5	21.3	29.5
2.5	21.9	29.2
1.25	23.0	31.2
0.625	25.4	31.4
0.310	-	31.7
0	36.2	38

Table 5-12 Average Cq values for the genes of interest under varying cDNA concentrations.

Finally, the housekeeping ability of *RPL19* under two PCO_2 was tested. Figure 5-48 shows the amplification plots for *RPL19* and *Fzd9* under 5 and 10% PCO_2. Only the 24 h time point was tested and the Cq value for *RPL19* at 5% PCO_2 was 22.2 and for 10% PCO_2 was 21.8, these Cq values indicate that

RPL19 is a suitable control gene. The *Fzd9* gene Cq values were 30 and 30.7 for 5% and 10% PCO_2 conditions, respectively. These initial results indicate that other CO_2 incubation time points should be tested to identify the biggest differential in the expression of *Fzd9*.

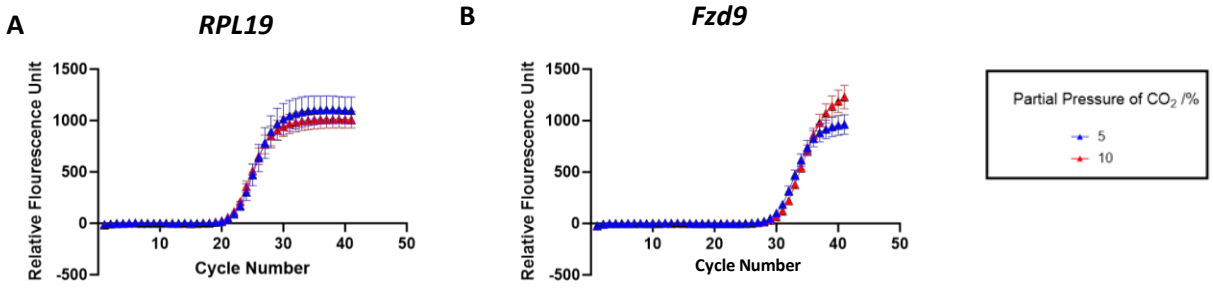


Figure 5-48 qPCR amplification plots where the relative fluorescence is plotted against the cycle number and a curve of best fit is plotted to the data points for (A) *RPL19* and (B) *Fzd9* for cDNA synthesised from RNA extracted from HEK293 cells incubated for 24 h under 5% (blue) and 10% (red) partial pressure of CO_2. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points.

5.13.1.2 Identification of a CO_2 Incubation Time Point

The next step was to use the optimised qPCR approach to assess the relative expression of *Fzd9* between HEK293 cells incubated at 5, 10 and 20 % PCO_2 for various time points. The first time point to analyse was the six-hour incubation, due to having the biggest transcriptional change in *Fzd9* expression across the five different time points in the Shigemura transcriptomic study. Figure 5-49 displays the qPCR results for *RPL19* and *Fzd9* amplicons from cDNA extracted from HEK293 cells incubated for six hours under varying PCO_2.

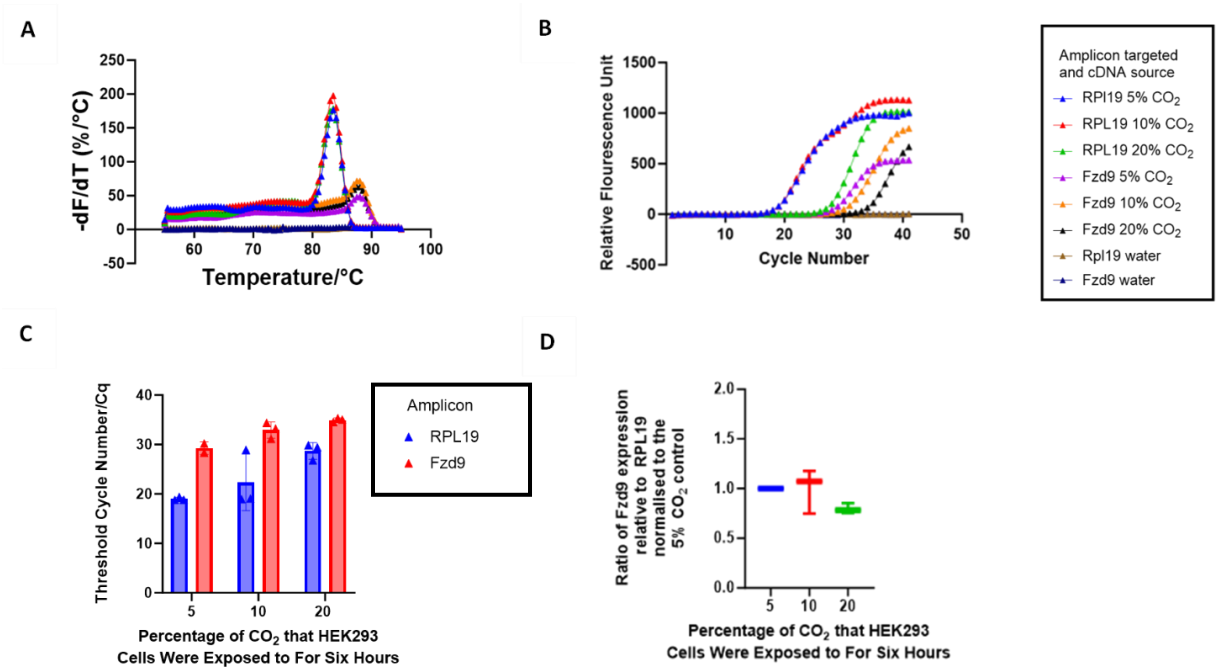


Figure 5-49 qPCR results for *RPL19* and *Fzd9* amplicons with cDNA extracted from HEK293 cells incubated for six hours at differential CO_2 concentrations. (A) Melt for *RPL19* and *Fzd9* targets are plotted where the negative derivative of fluorescence (F) with respect to temperature (T) otherwise known as the rate of change in fluorescence (-dF/dT) is plotted against the temperature (°C) and (B) amplification curves where the relative fluorescence is plotted against the cycle number and a curve of best fit is plotted to the data points for *RPL19* and *Fzd9*. The legend is applicable for both A and B where the experimental conditions are defined. (C) Threshold cycle (Cq) values for both amplicons at the various conditions tested against the percentage of CO_2 that HEK293 cells were exposed to for six hours and (D) Box plot of *Fzd9* expression relative to *RPL19* at 10% and 20% partial pressure of CO_2 (PCO_2) incubation normalised to the 5% PCO_2 control against the percentage of CO_2 that HEK293 cells were exposed to for six hours. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. Data passed the Shapiro-Wilks test for normality and significance was assessed by one-way ANOVA, multiple comparison tests and a one-sample t-test at a threshold of $p<0.05$. There was no significant difference in *Fzd9* expression between 5% and 10% PCO_2.

Figures 5-49A and B show no amplification in the control samples without cDNA for either amplicon indicating no contamination was present in samples. Figure 5-49C and D summarise the results of the qPCR reaction. The main issue with this dataset was that the 20% PCO_2 cDNA samples had a much higher than expected Cq value for the *RPL19* amplicon. The amplification curve for *RPL19* quantification from the 20% HEK293 cDNA is shown in green in Figure 5-49A and is consistent between the replicates therefore it is likely there was an issue during the cDNA synthesis from the RNA, storage of samples or that *RPL19* is not a suitable housekeeping gene for this experiment. This led to the decision that the 20% PCO_2 data should be disregarded when analysing this data. The 10% PCO_2 data was compared with the 5% PCO_2 data using a one-sample t-test where the hypothetical mean should be 1 and there was no significant difference in *Fzd9* expression for these two samples.

Further experiments were trialled, but it proved difficult to maintain consistency between qPCR runs which might have been due to the disparity in expression of these two genes or due to the cDNA synthesis sample quality. Strategies considered for improving results were to use a different housekeeping gene with lower expression in HEK293 cells for example *TUBA1A* (130 TPM), resynthesise all cDNA from newly extracted RNA or identify a different CO_2 sensitive gene with higher expression in HEK293 cells. All three strategies had limitations including, the literature on CO_2-sensitive genes identified previously with qPCR was limited and restarting from RNA extraction or reoptimizing the experiment for new genes would have been time consuming. Therefore, due to time constraints, it was decided that the time points of three hours and six hours would be used for RNA sequencing based on the previous findings by Shigemura *et al.*

5.13.4 RNA Sequencing

RNA sequencing (RNA-seq) analyses differential gene expression across control and treated samples. RNA-seq involves three main steps, including, library preparation, sequencing, and data analysis. The first step is library preparation where extracted RNA is broken into small fragments and converted into dsDNA. Sequencing adapters are attached to the DNA sequences by random priming. The DNA is loaded onto a flow cell and the adapter-modified clustered DNA fragments are enriched by PCR amplification, which is also known as cluster generation, creating millions of copies of single-stranded DNA.

Following library concentration and fragment length validation, the next step is library sequencing. Illumina sequencing uses the sequencing by synthesis (SBS) method where fluorescent nucleotide probes bind to the template nucleotide chain dependent on the base identity in a sequential fashion, one base at a time. An SBS kit specifies the number of cycles performed to determine the read length and, in this investigation, the read length was set at 100 bps. Each base in the oligonucleotide read sequence has an associated quality score. The quality score reflects the confidence level that the correct base has been called and is dependent on the strength of the signal detected from the fluorescent probe and the diversity of the surrounding sequence.

The last step is data analysis which screens reads for quality, maps high-quality reads to the genome and read count data is normalised to analyse the relative transcription of active genes between samples. The bioinformatics pipeline for the data analysis performed in this study by Cambridge Genomic Services is covered in section 2.5.6.8. and the results of this are detailed below in sections 5.13.4.1 – 6. The RNA expression levels for all genes in HEK293 cells were obtained from the human protein atlas.[280]

Twenty-four samples were submitted to RNA sequencing, one for each condition across the two-time points in triplicate. The details of these samples are given in Table 5-13 and grouped into replicates in Table 5-14. Samples at 10% CO_2 were buffered accordingly to account for acidic changes due to hypercapnia.

Sample ID	PCO_2/%	Pinometostat treatment	Time/hours
207_Ctrl_NT_3h	5	No	3
208_Ctrl_NT_3h	5	No	3
209_Ctrl_NT_3h	5	No	3
210_Ctrl_T_3h	5	Yes	3
211_Ctrl_T_3h	5	Yes	3
212_Ctrl_T_3h	5	Yes	3
267_C_NT_3h	10	No	3
268_C_NT_3h	10	No	3
269_C_NT_3h	10	No	3
270_C_T_3h	10	Yes	3
271_C_T_3h	10	Yes	3
272_C_T_3h	10	Yes	3
213_Ctrl_NT_6h	5	No	6
214_Ctrl_NT_6h	5	No	6
215_Ctrl_NT_6h	5	No	6
216_Ctrl_T_6h	5	Yes	6
217_Ctrl_T_6h	5	Yes	6
218_Ctrl_T_6h	5	Yes	6
273_C_NT_6h	10	No	6

274_C_NT_6h	10	No	6
275_C_NT_6h	10	No	6
276_C_T_6h	10	Yes	6
277_C_T_6h	10	Yes	6
278_C_T_6h	10	Yes	6

Table 5-13 Sample List for RNA sequencing experiment with a unique ID for samples. Ctrl refers to control PCO_2 which is 5% and C refers to elevated PCO_2 which is 10%. T refers to treatment with pinometostat and NT is without treatment. The number is related to incubation length in hours (h) at the defined PCO_2.

Group ID	Samples
5pct_D3	207,208,209
5pct_P3	210,211,212
10pct_D3	267,268,269
10pct_P3	270,271,272
5pct_D6	213,214,215
5pct_P6	216,217,218
10pct_D6	273,274,275
10pct_P6	276,277,278

Table 5-14 The conditions tested in this RNA seq experiment and the sample numbers which belong to each group. The 5pct and 10pct represent 5% PCO_2 and 10% PCO_2 incubation respectively. D and P represent DMSO and pinometostat, respectively whilst 3 and 6 are the length of incubation in hours at the defined PCO_2.

5.13.4.1 RNA-seq Quality

FastQC [281] analyses a range of metrics for assessing the data quality of each sample in a dataset both at the per base and per sequence level as discussed in the following sections 5.13.4.1.1-6. The read quality is influenced by library preparation and sequencing. Bias resulting in low-quality data can arise from a range of sources including, poor RNA sample quality, adapter contamination, PCR amplification errors, cluster density and phasing.[282]

5.13.4.1.1 Per Base Sequence Quality

The per base sequence quality was performed for each sample, and a box plot for an example sample is shown in Figure 5-50. Every base call across every sequence is associated with a quality score between 1 - 40 which relates to the error rate of calling the incorrect base. The quality score is also known as the Phred score where a score of 10 is an error rate of 1 in 10 and a score of 30 is an error rate of 1 in 1000.

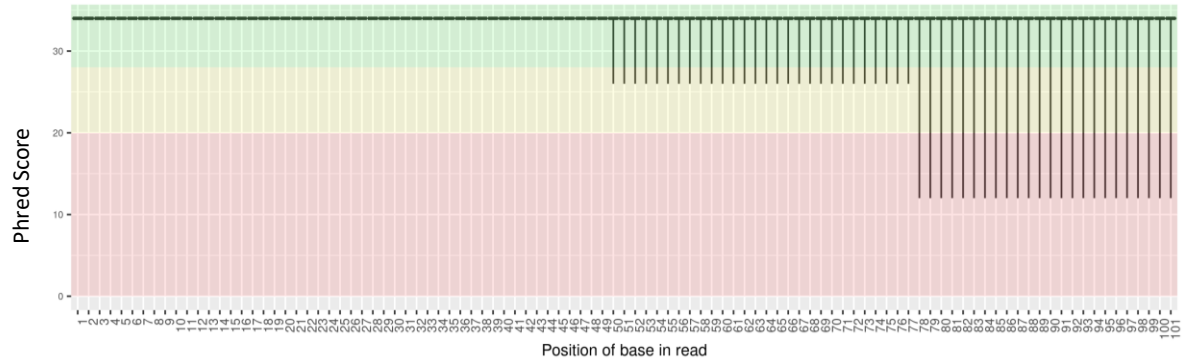


Figure 5-50 A box and whisker plot of phred/quality scores at each base position across the sequences identified. The black line represents the median quality score for the base at each position across all the reads and the whiskers are the 10th and 90th percentiles. The colour gradient displays acceptable quality scores in green, with decreasing base call confidence from yellow to red.

Figure 5-50 shows that the median quality score for all base positions across the reads is always in the green zone, meaning the median quality score is consistently above 30. The whiskers increase in size as the length of read increases due to the quality scores being lower at higher base positions because the different fragments in DNA clusters are more likely to be out of phase in later cycles of the SBS. All twenty-four samples used for RNA-sequencing had similar profiles to Figure 5-50 for the per base quality metric. The profile described here for a per-base sequencing box plot is reflective of high-quality data.

5.13.4.1.2 Per Sequence Quality Score

The distribution of the mean quality score across the reads is plotted in Figure 5-51. The expectation is that there should be a peak at a quality score of 30 or above to ensure that most sequences have a low error rate.

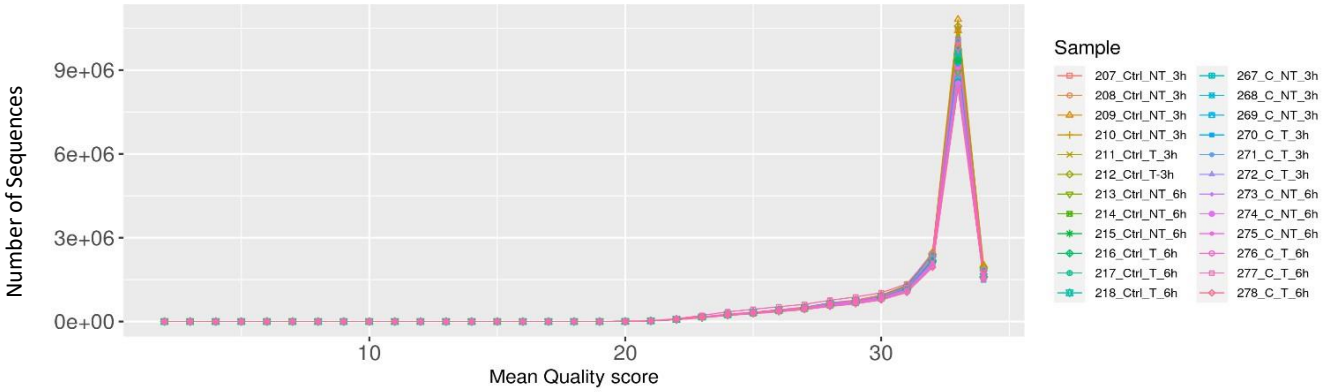


Figure 5-51 The number of sequences against the mean quality score for all RNA sequencing samples.

Figure 5-51 shows the overall quality scores for the reads across all the samples. This data shows that the quality is above 30 and consistent across all samples.

5.13.4.1.3 Per Base Sequence Content

The per base sequence content quality metric assesses the balance of the four nucleotides across the reads. In an unbiased genome sequencing experiment, the expected result would be that 25% of each base would be present across the reads giving parallel lines across these plots. However, it is important to note that this experiment sequences the transcriptome, not the genome. Assessing the proportion of each nucleobase can detect problems such as over-represented sequences, biased fragmentation, and biased composition libraries.

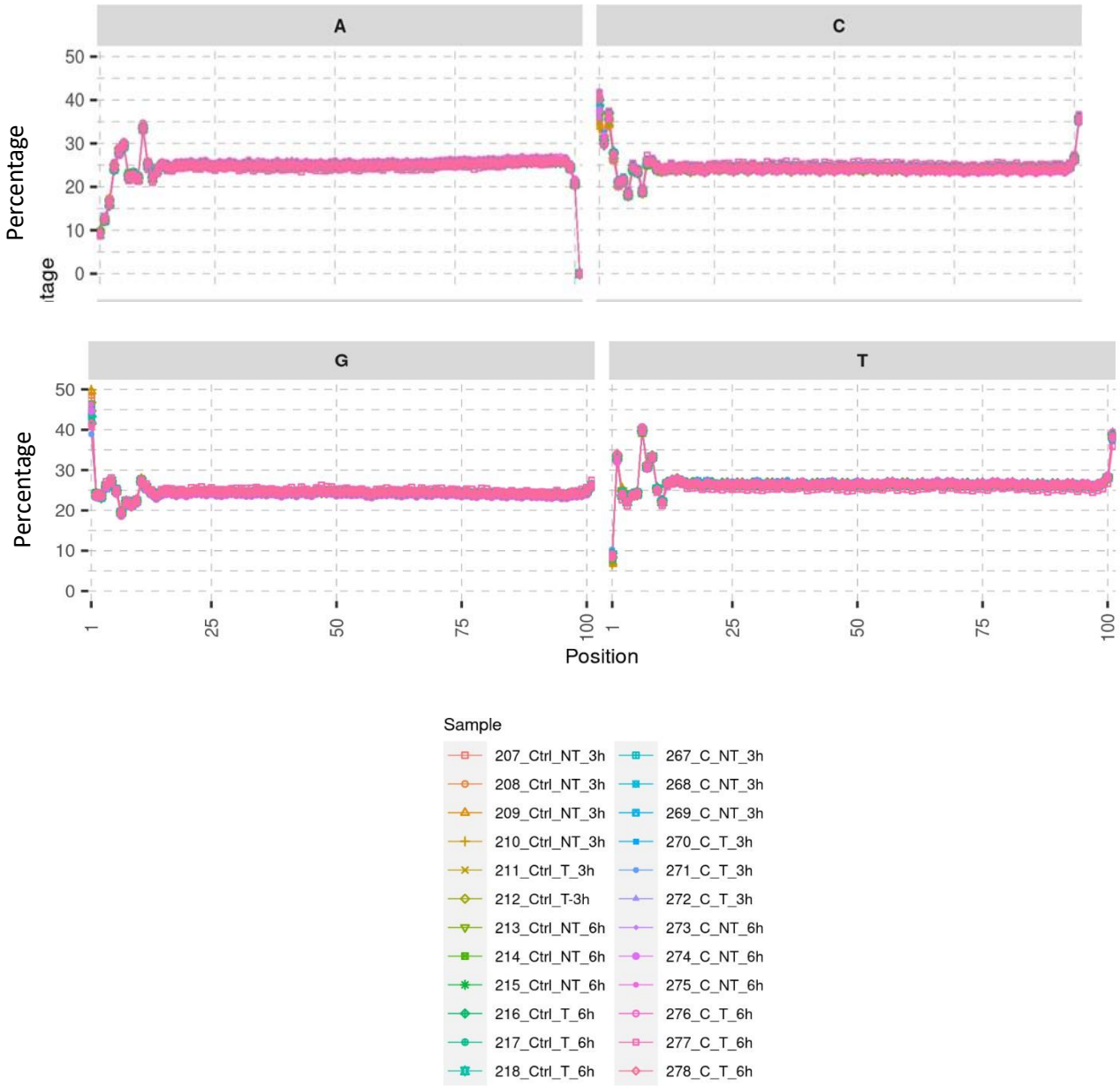


Figure 5-52 The percentage of each base per sequence versus the read position across all RNA sequencing samples.

Across the dataset, the proportion of each nucleobase at each read position is around 25% resulting in consistent parallel lines in Figure 5-52 except for at the start and end read positions. The variation at the start can be explained by the library preparation process. The addition of adapters to dsDNA by random priming is not completely random therefore bias at the beginning of the reads is expected.[283] The proportion of A drops at the end of the read due to trimming of adapters as discussed later in section 5.13.5.2.1.

5.13.4.1.4 Per Sequence GC content

The distribution of the C and G nucleotides can detect bias in samples. For an unbiased sample, a normal distribution centred at the overall GC content of the genome is expected.

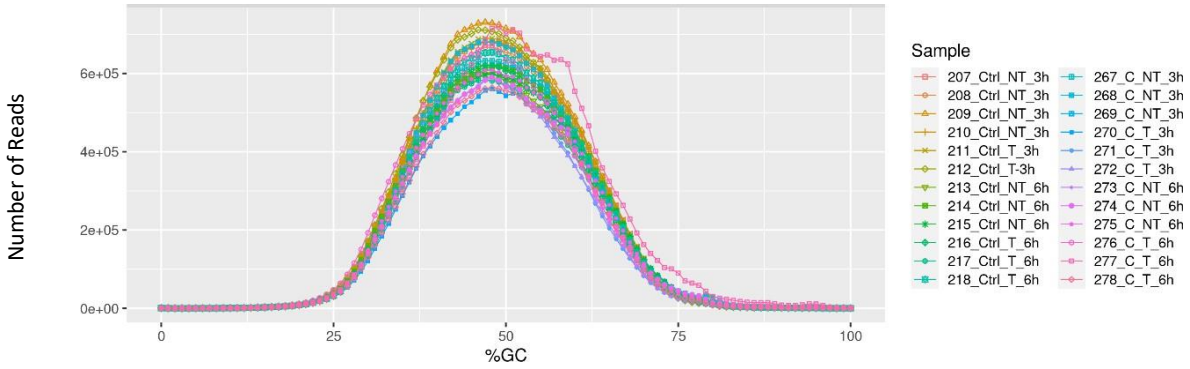


Figure 5-53 The number of reads versus the per sequence GC content distribution expressed as a percentage across the RNA sequencing dataset.

Figure 5-53 shows that the distribution peaks are consistent across the dataset and centred around 50% GC which is representative of the total RNA samples supplied. The only sample with a slight deviation from the bell curve is 277_C_T_6h, however, it is not significant enough to impact the data.

5.13.4.1.5 Sequence Length Distribution

The length of read is determined by the SBS cycle kit and in this investigation was set at 100. Figure 5-54 shows that the expected length of 100 is seen for the reads in all samples.

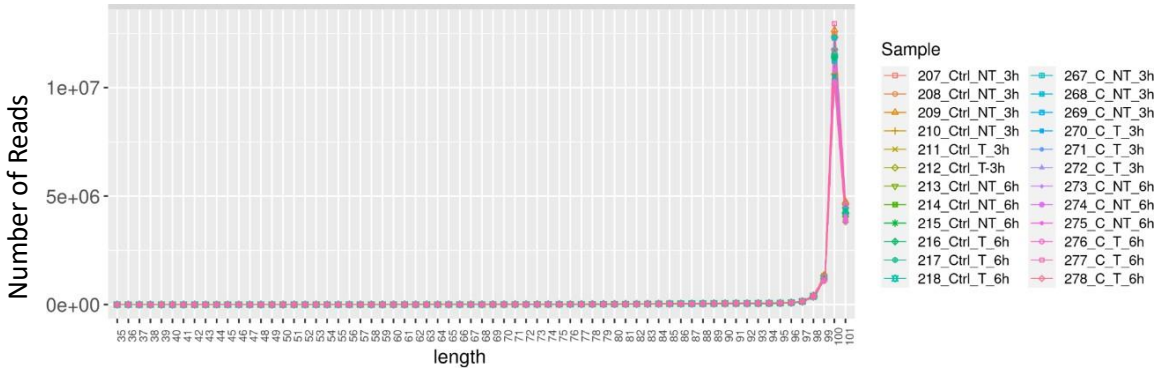


Figure 5-54 The number of reads versus the sequence length distribution across the RNA sequencing dataset.

5.13.4.1.6 Sequence Duplication Levels

The sequence duplication metric is a quality control for assessing bias arising from the PCR enrichment step during library preparation. The percentage of the library represented by identical reads grouped into duplication bins is plotted.

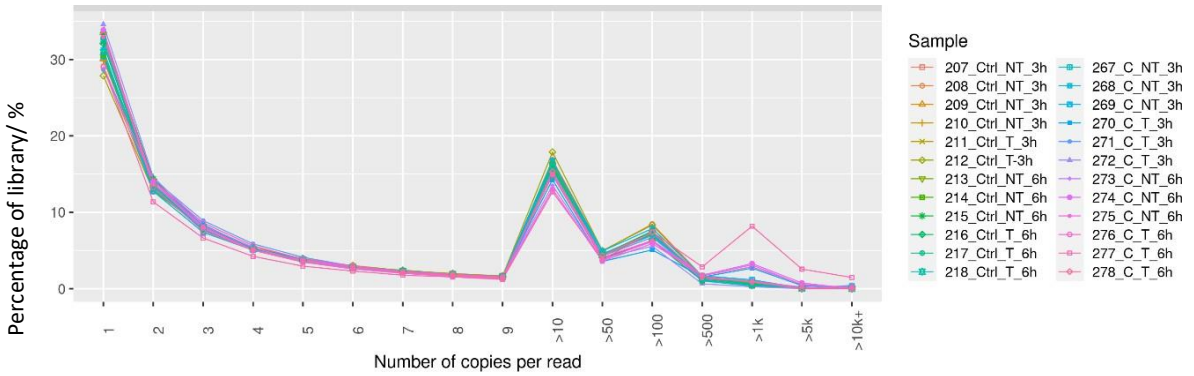


Figure 5-55 The percentage of the RNA sequencing library for each sample represented by the number of duplicated reads grouped into duplication bins.

Figure 5-55 shows that the highest percentage proportion of the library is composed of single copy reads. The next highest represented group is for the over 10 copies per read (>10) which is common in RNA sequencing analysis because certain transcripts are highly expressed and will be amplified and sequenced in the same region multiple times. This metric also shows the data is highly consistent across the dataset as all samples have a similar profile, with only 277_C_T_6h having a slightly different peak at >1000 copies.

The quality assessment metrics described in this section (5.13.5.1) show that the data is consistent across the sample set and that high-quality data has been sequenced.

5.13.4.2 RNA Sequencing Data Processing

The RNA sequencing bioinformatic pipeline involves many steps before obtaining a list of differentially expressed genes. These steps are outlined in sections 5.13.4.2.1- 6.

5.13.4.2.1 Trimming Low Quality Reads

Trim Galore [284] is an algorithm which uses the base quality score to determine which bases to trim from the 3' end of reads. Following the removal of low-quality bases, the algorithm finds and removes the adapter sequences. The cut-off read length after trimming was set at 20 bases and any reads at or below this length were discarded from the dataset. The average percentage of reads removed from samples across the dataset was 0.293 ± 0.195%. This value is extremely low and is to be expected from high-quality reads.

5.13.4.2.2 Mapping

Reads were mapped using STAR [3] v2.7.9 [285] to the Ensembl Homo_sapiens. GRCh38 (release 109) reference genome.[286] Reads from mapping experiments are assigned into three categories including, uniquely mapped, mapped to multiple loci and unmapped reads, expressed as a percentage as shown for samples in Figure 5-56. It is expected that over 75% of reads map uniquely to the genome and samples in the dataset met this expectation except for 277_C_T_6h which had 70% of uniquely mapped reads. The multiple loci mapped reads ranged from 6 - 13% for all samples except for 277_C_T_6h which was at 22%. The multiple loci reads were discarded to reduce false positives in the differential expression analysis.

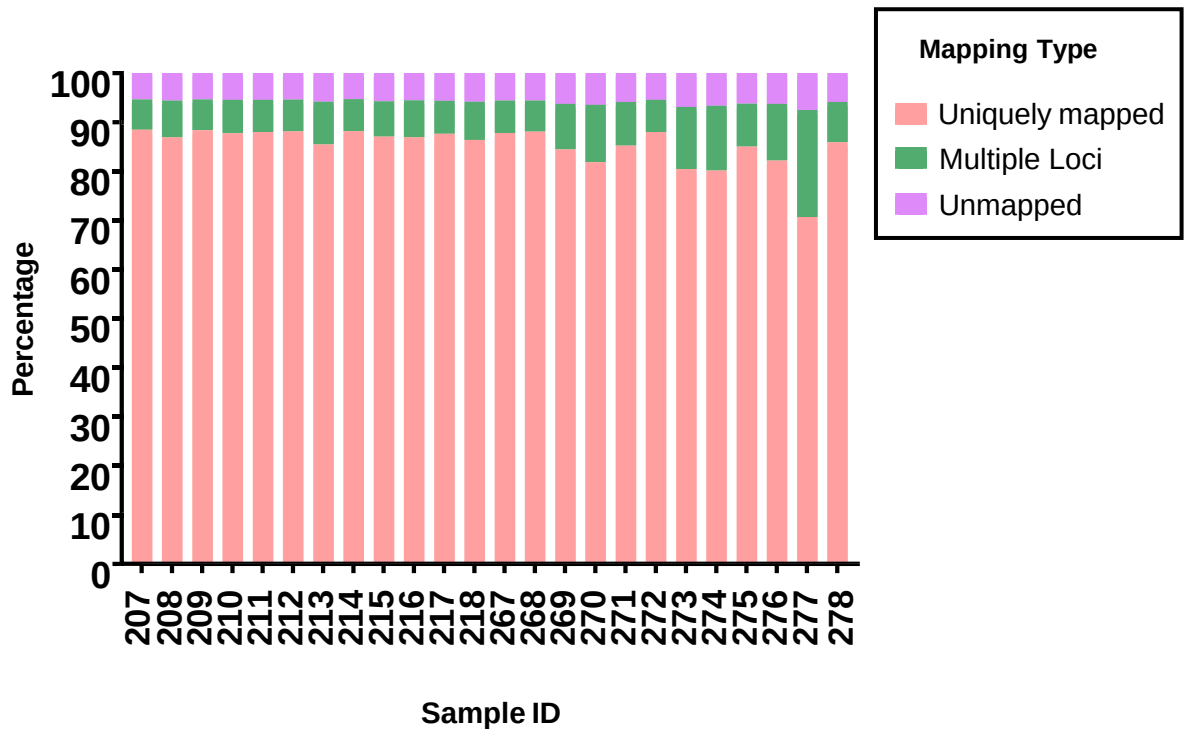


Figure 5-56 A 100% stacked bar chart depicting the proportion of reads uniquely mapped, mapped to multiple loci and unmapped to the genome versus the sample ID.

The percentage of reads from each sample which map to certain genomic locations was analysed as shown in Figure 5-57. The highest proportion of reads are expected to map to coding regions which was seen in this dataset. Coding regions represent the genes that code for proteins, whilst non-coding regions are composed of three types including, intergenic which represents non-coding sequences between genes, intronic which is a region residing in a gene which does not translate into protein and untranslated which is a regulatory domain not translated from the mRNA into the protein. In addition, the ribosomal RNA is encoded by the genome which leads to ribosome production.

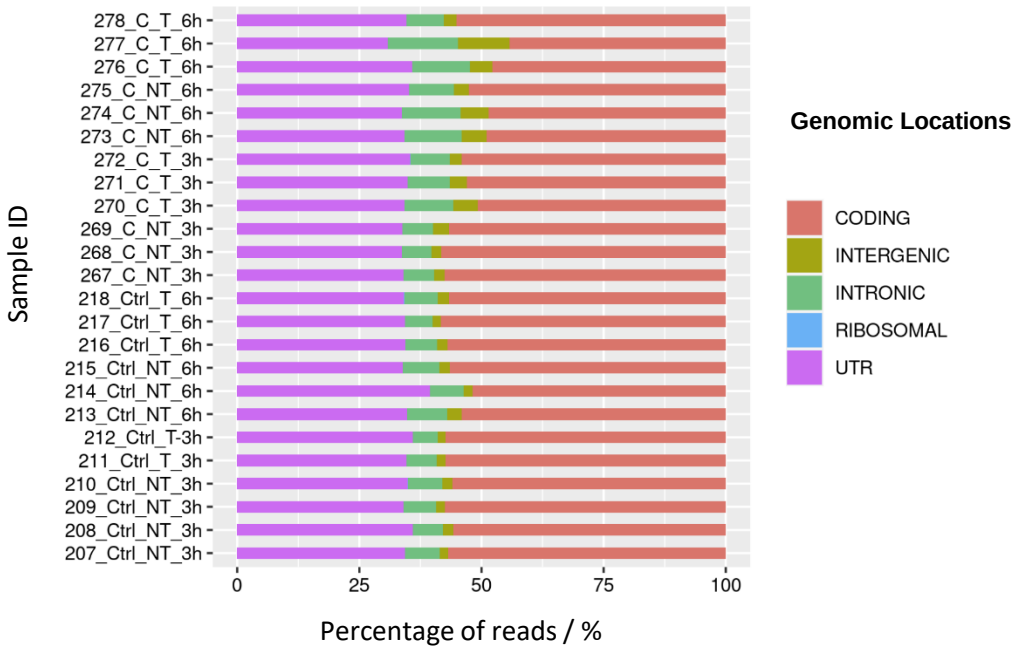


Figure 5-57 A 100% stacked bar chart which displays for each sample ID, the proportion of reads mapped to different genomic locations, where Utr represents the untranslated region.

5.13.4.2.3 Read Counting

Counts are the number of reads that map to a genomic location. In this investigation, the HTSeq algorithm [287] was used to calculate read counts. Low-quality (at a quality score threshold of 10), overlapping and duplicated mapped reads are discarded at this stage to reduce false positives. Figure 5-58 shows the results of read mapping to genes across the RNA-seq dataset. A feature is considered as the union of all exons belonging to a gene whose genomic coordinates are determined from the gtf database file.[288] This terminology means that in-features relates to, reads mapped to genes, not in-features relates to other genomic regions and the third category refers to ambiguous locations. As expected, the vast majority of reads map within gene features, for all samples.

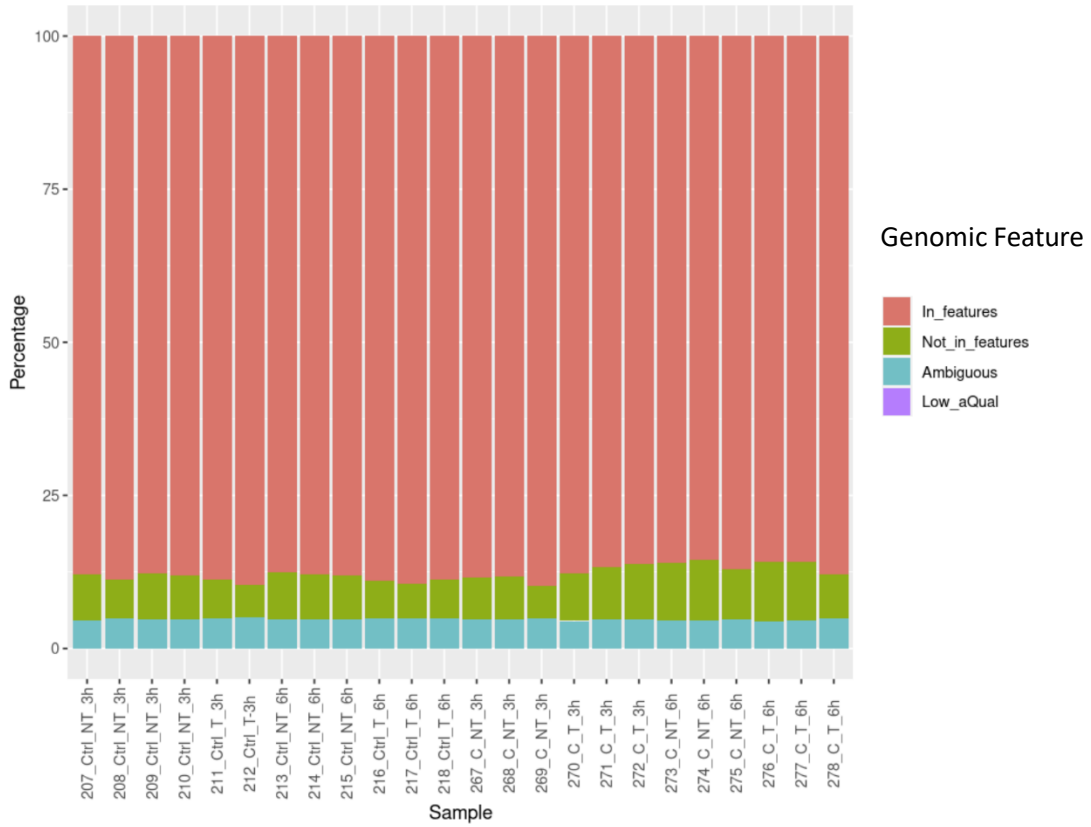


Figure 5-58 A 100% stacked bar chart depicting the proportion of mapped reads to genomic features versus the sample ID.

Following mapping, the number of genes detected in each sample is plotted in Figure 5-59. A count value threshold of greater than or equal to one was applied to all the genes identified across the dataset. Over 20,000 genes were identified in each sample.

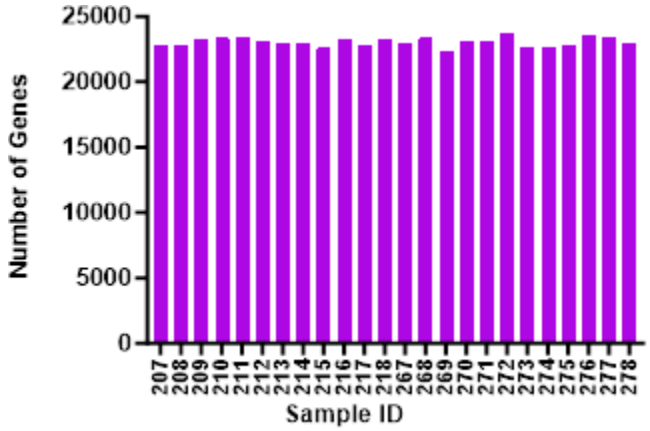


Figure 5-59 A bar graph of the number of genes above the count threshold versus the sample ID in the RNA-seq dataset.

5.13.4.2.4 Data Normalisation and Log Transformation

Reads map to genes in relation to three variables including, the gene length, gene expression, and library depth. Following gene mapping and counting, reads must be normalised to isolate gene expression from the other variables. The normalisation process used the trimmed median of means (TMM) as described in Scheme 5-1 performed by the DESeq2 algorithm.[289]

1. The raw count data is organised into a matrix where the rows are genes, and the columns are samples.

2. The per row geometric mean of the counts was calculated across the matrix.

$$Geometric\ mean\ for\ each\ row = \sqrt[n]{Multiply\ the\ raw\ counts\ for\ a\ gene\ across\ all\ samples}$$

$$n\ =\ number\ of\ samples$$

3. The per row ratio of the raw count divided by the geometric mean was calculated across the matrix. Whereby the raw count is the observation for that gene in one sample and the geometric mean is representative of all the observations of that gene across the samples in the experiment.

$$Per\ row\ Ratio = \frac{Raw\ count\ for\ a\ gene}{Per\ row\ geometric\ mean}$$

4. The per column median was calculated from the per row ratios across the matrix.

$$Size\ Factor\ =\ Median\ of\ the\ ratios\ in\ the\ columns$$

5. Finally, every raw count observation was divided by the size factor across the matrix to give the normalised data.

$$Library\ Depth\ Normalised\ Data\ = \frac{Raw\ count}{Size\ Factor}$$

Scheme 5-1 Calculation of Size Factors for Trimmed Median of Means Normalisation Calculations to account for library depth.

The geometric mean calculation is a robust metric resistant to outliers that removes rows containing 0 count data, which would affect the normalisation process. The size factor calculated in step 4 of Scheme 5-1 is the normalisation factor which reduces the variation in raw count data that arises from the library depth as shown in step 5. In addition to the size factor, every observation is

moderated by the gene length for each row in the matrix resulting in data normalised to gene length and library depth.

Further data manipulation is required because RNA-sequencing data is not normally distributed, however log-transformed data is normally distributed. Crucially, RNA-seq count data is heteroskedastic which can be defined as the data variance not being stable across the data range. Low-count data is highly variable and high-count data is associated with low variance. DESeq performs a regularized log (rlog) transformation to the normalised count data to account for the changing variance across the data.

5.13.4.2.5 Sample Clustering

Principal Component Analysis (PCA) plots are a useful visualisation technique used to analyse the variance across the samples in a dataset. Two principal components (PC) usually those with the highest variance expressed as a percentage are named PC1 and PC2 are the x and y axes for the plot. Further PCs are labelled consecutively, for example, PC4 represents the factor ranked as the fourth most variable in the data. Figure 5-60 shows the clustering of samples in this investigation where PC1 vs PC3 clusters the replicate samples into independent groups. Data points on a PCA plot that are close together are more similar and those further away are more different. PC1 variance arose from the time point of either 3 or 6-hour incubation of cells at 5 and 10% PCO_2 where all four of the three-hour times are clustered on the left-hand side and all six-hour time points are on the right. PC3 appears to separate data based on the treatment with DMSO or pinometostat. From the PC1 vs PC2 plot (Figure 8-50), PC2 appears to separate the data based on the percentage of PCO_2 incubation (5 or 10%) however this data does not cluster as distinctly as PC1 vs PC3.

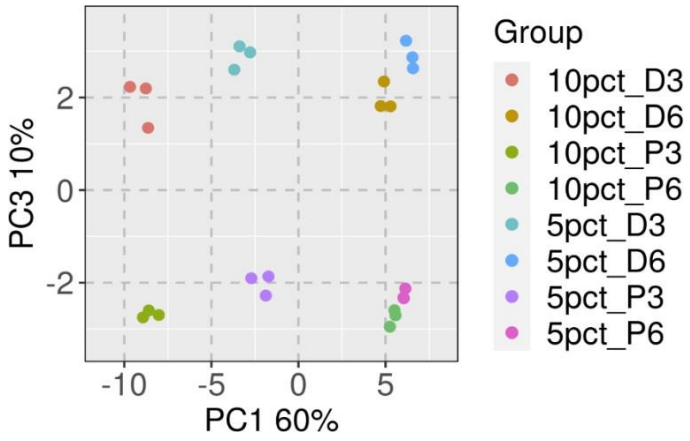


Figure 5-60 A Principal Component Analysis (PCA) plot for PC1 vs PC3 for all samples grouped into the different conditions tested in the RNA-seq dataset where each group contains three replicates.

5.13.4.2.6 EdgeR Data Processing

EdgeR software [290] was used for differential gene expression analyses.[291] For software compatibility reasons, the normalisation factor was recalculated by EdgeR in the same manner described for the DESeq2 data which was needed for PCA plots. Following this GC content and gene length bias were corrected for by the GQN package c1.30.0 [292] to reduce bias in read coverage. The dispersion of each gene was estimated by EdgeR which is a measure of the uncertainty of the read count. The dispersion is a sum of the biological and technical variance. For low-expression genes, the technical variance dominates whereas for high-expression genes, the biological variance dominates. The method used for estimating the dispersion for each data point is called Bayesian shrinkage. In an RNA sequence experiment, there are many variables, i.e., expression values for many genes but a limited number of samples. Therefore, the variance across the whole experiment is considered to give an estimate of variance for each individual gene to increase the data robustness. From the negative binomial distributed data, EdgeR will finally determine differential gene expression between two sample groups at a time using the statistical method called the Exact test. The differentially expressed genes can be analysed by the log in base 2 fold change values where the first group in the comparison

is the baseline and each gene is associated with a p-value and FDR value which is given as an adjusted p-value accounting for multiple testing.

After read trimming, mapping, counting, data normalisation and transformation, samples were clustered and DEGs between pairwise comparisons as detailed in Table 5-15 were identified.

5.13.4.3 Analysis of Differentially Expressed Genes (DEGs)

The fold change in gene expression between sample groups is expressed on the logarithmic scale, to increase the expressivity of numbers so that the scale for downregulated genes (0 to - ∞) is the same magnitude as upregulated genes (0 to + ∞). Whereas on the linear scale downregulated genes scale would be on a scale of 1 to 0 and upregulated genes would be on scale of 0 to + ∞. Differentially expressed genes are filtered by the adjusted p-value to assess statistically significant changes. Table 5-15 shows the number of statistically significant DEGs at a threshold value of FDR < 0.05 identified across the pairwise comparison groups in this study.

Comparison	Statistically Significant DEGs FDR <0.05	Total DEGs
5pct_D3__VS_ 10pct_D3	1213	10840
5pct_D3__VS_ 5pct_P3	752	10977
5pct_D3__VS_ 10pct_P3	2816	11091
10pct_D3__VS_ 10pct_P3	2013	11039
5pct_P3__VS _ 10pct_P3	2440	11173
5pct_D6__VS_ 10pct_D6	3053	11129
5pct_D6__VS_ 5pct_P6	2308	11219
5pct_D6__VS_ 10pct_P6	2727	11238
10pct_D6__VS_ 10pct_P6	1162	11190
5pct_P6__VS _ 10pct_P6	2829	11241

Table 5-15 The total and statistically significant number of DEGs identified in the pairwise comparisons performed in this RNA-seq dataset. The sample notation used here is detailed in Table 5-13.

It is important to note that the FDR is used when assessing significant change to reduce false positives due to reducing the type I error, but the p-value is used when assessing non-significant change because using the FDR increases the type II error.

5.13.4.3.1 CO_2-sensitive Genes That Are Not Altered under Pinometostat Treatment.

The principal aim of the study was to identify genes which were significantly changed between 5% and 10% PCO_2 treated with DMSO in both conditions but not altered under 5% PCO_2 DMSO and 10% PCO_2 pinometostat treatment. All DEGs pass the statistically significant threshold of FDR ≤ 0.05, genes that do not significantly change have a $p > 0.05$ and the baseline condition used for comparison is 5% PCO_2 DMSO. Gene tables detailing the top ten deregulated genes for each of these comparisons are shown in Tables 5-16 to 5-27.

5.13.4.3.2 Upregulated Genes

For the three-hour CO_2 incubation time point, 561 genes were significantly upregulated in 10% PCO_2 DMSO compared with 5% PCO_2 DMSO (Figure 5-16), of which 132 genes were not significantly changed in 10% PCO_2 pinometostat compared with 5% PCO_2 DMSO ((Figure 5-17).

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000015520	NPC1L1	2.20	$3.14X10^{-22}$	$1.13X10^{-18}$
ENSG00000261341	SMIM47	1.73	$3.43X10^{-11}$	$6.52X10^{-9}$
ENSG00000196378	ZNF34	1.69	$4.33X10^{-19}$	$5.87X10^{-16}$
ENSG00000105219	CCNP	1.69	$7.69X10^{-12}$	$1.81X10^{-9}$
ENSG00000156427	FGF18	1.68	$3.97X10^{-11}$	$7.42X10^{-9}$
ENSG00000110876	SELPLG	1.55	$4.39X10^{-13}$	$1.70X10^{-10}$
ENSG00000014914	MTMR11	1.43	$4.49X10^{-12}$	$1.16X10^{-9}$
ENSG00000136881	BAAT	1.42	$1.91X10^{-12}$	$5.60X10^{-10}$
ENSG00000187796	CARD9	1.41	$1.79X10^{-10}$	$2.93X10^{-8}$
ENSG00000130203	APOE	1.31	$5.60X10^{-20}$	$8.67X10^{-17}$

Table 5-16 Top ten genes upregulated in response to buffered hypercapnia at 3-hour differential CO2 incubation listed in order of magnitude of fold change with associated P and FDR values.

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000232434	AJM1	0.93	2.87X10-6	1.24X10-4
ENSG00000100027	YPEL1	0.91	3.81X10-7	2.27X10-5
ENSG00000206535	LNP1	0.90	1.02X10-4	2.37X10-3
ENSG00000042286	AIFM2	0.66	2.93X10-3	3.07X10-2
ENSG00000131188	PRR7	0.64	5.61X10-4	8.80X10-3
ENSG00000102984	ZNF821	0.64	7.48X10-4	1.09X10-2
ENSG00000148926	ADM	0.63	5.51X10-5	1.42X10-3
ENSG00000162772	ATF3	0.62	2.52X10-12	6.65X10-10
ENSG00000111266	DUSP16	0.61	3.35X10-13	1.40X10-10
ENSG00000148677	ANKRD1	0.61	2.44X10-3	2.69X10-2

Table 5-17 Top ten genes upregulated in response to buffered hypercapnia at the 3-hour differential CO_2 incubation that were not significantly deregulated under pinometostat treatment listed in order of magnitude of fold change with associated P and FDR values.

For the six-hour CO_2 incubation time point, 1507 genes are significantly upregulated in 10% PCO_2 DMSO compared with 5% PCO_2 DMSO, (Table 5-18) of which 502 genes are not significantly changed in 10% PCO_2 pinometostat compared with 5% PCO_2 DMSO (Table 5-19).

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000144821	MYH15	1.74	$2.20X10^{-13}$	$7.21X10^{-11}$
ENSG00000089820	ARHGAP4	1.63	$4.45X10^{-13}$	$1.34X10^{-10}$
ENSG00000183091	NEB	1.60	$7.83X10^{-5}$	$8.78X10^{-4}$
ENSG00000116254	CHD5	1.48	$1.05X10^{-13}$	$3.66X10^{-11}$
ENSG00000182326	C1S	1.46	$5.27X10^{-10}$	$5.48X10^{-8}$
ENSG00000182179	UBA7	1.40	$5.85X10^{-15}$	$2.96X10^{-12}$
ENSG00000164309	CMYA5	1.35	$2.21X10^{-10}$	$2.65X10^{-8}$
ENSG00000182759	MAFA	1.25	$5.18X10^{-15}$	$2.88X10^{-12}$
ENSG00000125740	FOSB	1.21	$1.14X10^{-7}$	$4.86X10^{-6}$
ENSG00000187837	H1-2	1.20	$3.67X10^{-8}$	$1.84X10^{-6}$

Table 5-18 Top ten genes upregulated in response to buffered hypercapnia at 6-hour differential CO_2 incubation listed in order of magnitude of fold change with associated P and FDR values.

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000156711	MAPK13	1.01	$3.00X10^{-7}$	$1.03X10^{-5}$
ENSG00000186594	MIR22HG	0.94	$4.21X10^{-8}$	$2.05X10^{-6}$
ENSG00000255198	SNHG9	0.87	$3.79X10^{-6}$	$8.05X10^{-5}$
ENSG00000165949	IFI27	0.82	$4.20X10^{-5}$	$5.47X10^{-4}$
ENSG00000129355	CDKN2D	0.78	$9.25X10^{-6}$	$1.62X10^{-4}$
ENSG00000204248	COL11A2	0.78	$9.92X10^{-5}$	$1.06X10^{-3}$
ENSG00000180479	ZNF571	0.76	$1.72X10^{-5}$	$2.63X10^{-4}$
ENSG00000143387	CTSK	0.75	$7.51X10^{-5}$	$8.51X10^{-4}$
ENSG00000131471	AOC3	0.75	$5.48X10^{-5}$	$6.67X10^{-4}$
ENSG00000272145	NFYC-AS1	0.67	$2.85X10^{-4}$	$2.48X10^{-3}$

Table 5-19 Top ten genes upregulated in response to buffered hypercapnia at the 6-hour differential CO_2 incubation that were not significantly deregulated under pinometostat treatment listed in order of magnitude of fold change with associated P and FDR values.

In response to buffered hypercapnia, 195 genes were upregulated at 10% CO_2 compared to 5% CO_2 at both time points (Table 5-20) There are 13 genes identified at both time points between the lists of genes upregulated between 5% and 10% PCO_2 with DMSO treatment but not changed in 5% DMSO vs 10% PCO_2 pinometostat treatment (Table 5-21). Statistical testing with Fisher's exact test gives a p value of 0.004657 meaning there is a significant overlap between these gene lists.

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000130203	APOE	1.31	$5.60X10^{-20}$	$8.67X10^{-17}$
ENSG00000131471	AOC3	1.24	$7.13X10^{-23}$	$3.86X10^{-19}$
ENSG00000149591	TAGLN	1.17	$1.54X10^{-12}$	$4.91X10^{-10}$
ENSG00000156711	MAPK13	1.09	$6.31X10^{-6}$	$2.40X10^{-4}$
ENSG00000182326	C1S	1.08	$2.37X10^{-12}$	$6.65X10^{-10}$
ENSG00000167733	HSD11B1L	1.05	$6.62X10^{-6}$	$2.46X10^{-4}$
ENSG00000085465	OVGP1	1.05	$1.11X10^{-15}$	$9.23X10^{-13}$
ENSG00000158715	SLC45A3	0.94	$7.42X10^{-8}$	$5.62X10^{-6}$
ENSG00000144821	MYH15	0.94	$9.39X10^{-12}$	$2.12X10^{-9}$
ENSG00000223573	TINCR	0.90	$4.73X10^{-18}$	$5.13X10^{-15}$

Table 5-20 Top ten genes upregulated in response to buffered hypercapnia at both 3- and 6-hour differential CO_2 incubations listed in order of magnitude of fold change with associated P and FDR values.

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000148677	ANKRD1	0.61	$2.44X10^{-3}$	$2.69X10^{-2}$
ENSG00000178381	ZFAND2A	0.44	$2.15X10^{-3}$	$2.44X10^{-2}$
ENSG00000081320	STK17B	0.43	$6.17X10^{-6}$	$2.36X10^{-4}$
ENSG00000104447	TRPS1	0.37	$1.37X10^{-5}$	$4.42X10^{-4}$
ENSG00000025156	HSF2	0.35	$1.00X10^{-4}$	$2.34X10^{-3}$
ENSG00000120690	ELF1	0.33	$1.96X10^{-3}$	$2.28X10^{-2}$
ENSG00000185551	NR2F2	0.31	$1.31X10^{-3}$	$1.67X10^{-2}$
ENSG00000163602	RYBP	0.31	$2.05X10^{-4}$	$4.08X10^{-3}$
ENSG00000188997	KCTD21	0.30	$2.90X10^{-3}$	$3.05X10^{-2}$
ENSG00000158158	CNNM4	0.29	$8.24X10^{-4}$	$1.19X10^{-2}$
ENSG00000070423	RNF126	0.29	$4.34X10^{-4}$	$7.14X10^{-3}$
ENSG00000112290	WASF1	0.24	$7.15X10^{-4}$	$1.06X10^{-2}$
ENSG00000237649	KIFC1	0.21	$4.25X10^{-3}$	$4.02X10^{-2}$

Table 5-21 The genes that were upregulated in response to buffered hypercapnia at 3- and 6-hour differential CO_2 incubations that were not significantly deregulated under pinometostat treatment listed in order of magnitude of fold change with associated P and FDR values.

5.13.4.3.3 Downregulated Genes

For the three-hour CO_2 incubation time point, 652 genes were significantly downregulated in 10% PCO_2 DMSO compared with 5% PCO_2 DMSO (Table 5-22), of which 271 genes were not significantly changed in 10% PCO_2 pinometostat compared with 5% PCO_2 DMSO (Table 5-23).

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000291168	ANKRD19P	-1.46	$1.77X10^{-9}$	$2.15X10^{-7}$
ENSG00000177352	CCDC71	-1.20	$1.55X10^{-20}$	$2.79X10^{-17}$
ENSG00000223749	MIR503HG	-1.12	$1.68X10^{-13}$	$8.13X10^{-11}$
ENSG00000214021	TTLL3	-1.08	$2.51X10^{-15}$	$1.94X10^{-12}$
ENSG00000114626	ABTB1	-1.06	$1.66X10^{-6}$	$8.01X10^{-5}$
ENSG00000197128	ZNF772	-1.06	$7.27X10^{-6}$	$2.64X10^{-4}$
ENSG00000181004	BBS12	-1.02	$1.93X10^{-7}$	$1.26X10^{-5}$
ENSG00000002016	RAD52	-0.98	$6.92X10^{-17}$	$6.82X10^{-14}$
ENSG00000215012	RTL10	-0.98	$7.32X10^{-16}$	$6.61X10^{-13}$
ENSG00000162066	AMDH	-0.97	$5.05X10^{-13}$	$1.89X10^{-10}$

Table 5-22 Top ten genes downregulated in response to buffered hypercapnia at 3-hour differential CO_2 incubation listed in order of magnitude of fold change with associated P and FDR values.

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000223749	MIR503HG	-1.12	$1.68X10^{-13}$	$8.13X10^{-11}$
ENSG00000159958	TNFRSF13C	-0.93	$6.15X10^{-7}$	$3.42X10^{-5}$
ENSG00000163463	KRTCAP2	-0.84	$1.09X10^{-6}$	$5.64X10^{-5}$
ENSG00000186272	ZNF17	-0.84	$1.97X10^{-6}$	$9.12X10^{-5}$
ENSG00000270069	MIR222HG	-0.83	$1.47X10^{-5}$	$4.71X10^{-4}$
ENSG00000163703	CRELD1	-0.82	$6.91X10^{-4}$	$1.04X10^{-2}$
ENSG00000117480	FAAH	-0.82	$1.03X10^{-4}$	$2.39X10^{-3}$
ENSG00000112218	GPR63	-0.80	$3.77X10^{-6}$	$1.56X10^{-4}$
ENSG00000134253	TRIM45	-0.78	$5.10X10^{-5}$	$1.33X10^{-3}$
ENSG00000291118	ZNF767P	-0.76	$1.11X10^{-9}$	$1.45X10^{-7}$

Table 5-23 Top ten genes downregulated in response to buffered hypercapnia at the 3-hour differential CO_2 incubation that were not significantly deregulated under pinometostat treatment listed in order of magnitude of fold change with associated P and FDR values.

For the six-hour CO_2 incubation time point, 1546 genes are significantly downregulated in 10% PCO_2 DMSO compared with 5% PCO_2 DMSO (Table 5-24), of which 510 genes are not significantly changed in 10% PCO_2 pinometostat compared with 5% PCO_2 DMSO (Table 5-25).

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000204388	HSPA1B	-1.83	$2.16X10^{-115}$	$2.40X10^{-111}$
ENSG00000291049	FAM86B3P	-1.32	$2.19X10^{-10}$	$2.64X10^{-8}$
ENSG00000287110	ACAD9-DT	-1.32	$1.75X10^{-9}$	$1.48X10^{-7}$
ENSG00000126562	WNK4	-1.28	$5.22X10^{-9}$	$3.57X10^{-7}$
ENSG00000247796	MOCS2-DT	-1.27	$5.54X10^{-9}$	$3.73X10^{-7}$
ENSG00000267493	CIRBP-AS1	-1.24	$1.61X10^{-8}$	$8.95X10^{-7}$
ENSG00000183798	EMILIN3	-1.14	$1.31X10^{-12}$	$3.30X10^{-10}$
ENSG00000183317	EPHA10	-1.08	$5.77X10^{-6}$	$1.11X10^{-4}$
ENSG00000189362	NEMP2	-1.07	$2.87X10^{-5}$	$3.98X10^{-4}$
ENSG00000162999	DUSP19	-1.05	$5.39X10^{-8}$	$2.56X10^{-6}$

Table 5-24 Top ten genes downregulated in response to buffered hypercapnia at 6-hour differential CO_2 incubation listed in order of magnitude of fold change with associated P and FDR values.

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000157214	STEAP2	-0.94	$7.45X10^{-5}$	$8.49X10^{-4}$
ENSG00000181458	TMEM45A	-0.93	$9.32X10^{-6}$	$1.62X10^{-4}$
ENSG00000178026	LRRC75B	-0.91	$5.04X10^{-5}$	$6.23X10^{-4}$
ENSG00000047597	XK	-0.86	$4.79X10^{-6}$	$9.65X10^{-5}$
ENSG00000186998	EMID1	-0.85	$2.62X10^{-4}$	$2.33X10^{-3}$
ENSG00000149927	DOC2A	-0.81	$6.69X10^{-9}$	$4.32X10^{-7}$
ENSG00000183873	SCN5A	-0.81	$4.86X10^{-4}$	$3.72X10^{-3}$
ENSG00000107020	PLGRKT	-0.78	$8.42X10^{-4}$	$5.73X10^{-3}$
ENSG00000163393	SLC22A15	-0.76	$5.26X10^{-4}$	$3.96X10^{-3}$
ENSG00000119946	CNNM1	-0.75	$8.76X10^{-6}$	$1.56X10^{-4}$

Table 5-25 Top ten genes downregulated in response to buffered hypercapnia at the 6-hour differential CO_2 incubation that were not significantly deregulated under pinometostat treatment listed in order of magnitude of fold change with associated P and FDR values.

In response to buffered hypercapnia, 89 genes were downregulated at 10% CO2 compared to 5% CO2 at both time points (Table 5-26). There are 3 genes identified at both time points between the lists of genes downregulated between 5% and 10% PCO_2 with DMSO treatment but not changed in 5% DMSO vs 10% PCO_2 pinometostat treatment (Table 5-27). Statistical testing with Fisher's exact test gives a p-value of 0.9997 meaning there is no significant overlap between these gene lists.

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000114626	ABTB1	-1.064	$1.66X10^{-6}$	$8.01X10^{-5}$
ENSG00000218510	LINC00339	-0.956	$1.93X10^{-6}$	$9.03X10^{-5}$
ENSG00000223764	LINC02593	-0.939	$2.53X10^{-11}$	$5.07X10^{-9}$
ENSG00000250571	GLI4	-0.859	$6.46X10^{-8}$	$5.04X10^{-6}$
ENSG00000134253	TRIM45	-0.784	$5.10X10^{-5}$	$1.33X10^{-3}$
ENSG00000108469	RECQL5	-0.753	$1.77X10^{-11}$	$3.68X10^{-9}$
ENSG00000271270	TMCC1-DT	-0.749	$4.10X10^{-7}$	$2.42X10^{-5}$
ENSG00000158106	RHPN1	-0.720	$3.92X10^{-15}$	$2.65X10^{-12}$
ENSG00000175283	DOLK	-0.702	$2.12X10^{-8}$	$1.98X10^{-6}$
ENSG00000168010	ATG16L2	-0.645	$3.70X10^{-5}$	$1.02X10^{-3}$

Table 5-26 Top ten genes downregulated in response to buffered hypercapnia at both the 3- and 6-hour differential CO_2 incubations listed in order of magnitude of fold change with associated P and FDR values.

Ensembl Gene ID	GeneName	logFC	PValue	FDR
ENSG00000088451	TGDS	-0.36	$9.81X10^{-4}$	$1.34X10^{-2}$
ENSG00000132405	TBC1D14	-0.30	$7.33X10^{-4}$	$1.08X10^{-2}$
ENSG00000048828	FAM120A	-0.23	$1.49X10^{-3}$	$1.85X10^{-2}$

Table 5-27 The genes that were downregulated in response to buffered hypercapnia at 3- and 6-hour differential CO_2 incubations that were not significantly deregulated under pinometostat treatment listed in order of magnitude of fold change with associated P and FDR values.

5.13.4.3.4 Gene Ontology Enrichment Analysis

Gene Ontology (GO) enrichment is a method of classifying gene lists into predefined categories to describe the functional characteristics of genes.[293] GO terms consist of three main classes: molecular function, cellular component, and biological process. EnrichR was developed to analyse the overlap of GO terms associated with the experimental gene list compared with the in-built library which contains functional terms from gene sets across diverse contexts.[294] The significance of the overlap of enriched terms is assessed by the Fishers exact test and terms are filtered by an adjusted $p \leq 0.05$.

5.13.4.3.4.1 GO Terms for Genes Upregulated at 10% PCO_2 DMSO but not at 10% PCO_2 Pinometostat Compared to 5% PCO_2 DMSO

Various input gene lists as detailed in sections 5.13.4.3.2 - 3 were analysed with EnrichR to assess the enrichment of GO terms that were altered in 10% PCO_2 when treated with DMSO but not changed when treated with pinometostat compared with the baseline condition of 5% PCO_2 DMSO. The combined score is another useful metric in the EnrichR output which multiplies together the natural log of the p-value multiplied by the z-score, where the z-score is the deviation from the expected rank which is calculated from randomized gene sets. Here the biological process enriched GO terms are stated, the number of genes which are associated with the GO term in the gene list compared to the gene library is shown in the overlap column, the statistical significance (adjusted p value) of the term and the combined score are listed.

The gene lists which contained the 132 and 502 genes which are upregulated at 10% PCO_2 DMSO but not at 10% PCO_2 pinometostat compared to 5% PCO_2 DMSO at 3 and 6 hours, respectively were analysed by EnrichR. The GO term results are given in Tables 5-28 and 5-29 for the three- and six-hour incubation time points, respectively.

GO terms	GO identifier	Overlap	Adjusted P-value	Combined Score
Positive Regulation of Cardiac Muscle Cell Differentiation	GO:2000727	2/5	0.04257	790.46
Mitotic Spindle Organisation	GO:0007052	6/85	0.01549	128.46
Positive Regulation of Nucleic Acid-Templated Transcription	GO:1903508	12/557	0.04257	28.47
Regulation of DNA-templated Transcription	GO:0006355	28/1922	0.01752	25.43
Positive Regulation of DNA-templated Transcription	GO:0045893	20/1243	0.04257	23.36
Positive Regulation of Transcription by RNA polymerase II	GO:0045944	16/938	0.04257	21.76
Negative Regulation of DNA-templated Transcription	GO:0045892	17/1025	0.04257	21.50
Regulation of Transcription by RNA Polymerase II	GO:0006357	27/2028	0.04257	18.64

Table 5-28 The biological process GO enrichment terms for genes upregulated at 10% PCO_2 DMSO but not at 10% PCO_2 pinometostat compared to 5% PCO_2 DMSO at a three-hour incubation of higher PCO_2, alongside the gene list and library overlap and the statistical significance of the term at a cut-off of an adjusted $p \leq 0.05$ ordered by the combined score.

GO terms	GO identifier	Overlap	Adjusted P-value	Combined Score
Plasma Membrane Bounded Cell Projection Morphogenesis	GO:0120039	7/50	0.03588	56.57
Negative Regulation of Transcription by RNA polymerase II	GO:0000122	43/763	0.00031	38.10
Regulation of RNA Splicing	GO:0043484	10/102	0.03588	38.02
Negative Regulation of DNA-templated Transcription	GO:0045892	52/1025	0.00031	33.45
Regulation of Transcription by RNA Polymerase II	GO:0006357	86/2028	0.00031	30.10
DNA Damage Response	GO:0006974	24/384	0.00892	29.21
Regulation of DNA-templated Transcription	GO:0006355	81/1922	0.00031	27.51
Chromatin Remodelling	GO:0006338	16/228	0.03509	27.22
Positive Regulation of DNA-templated Transcription	GO:0045893	56/1243	0.00260	24.15
Positive Regulation of Nucleic Acid-Templated Transcription	GO:1903508	30/557	0.01167	23.74
Regulation of Gene Expression	GO:0010468	50/1127	0.00896	20.47

Table 5-29 The biological processes GO enrichment terms for genes upregulated at 10% PCO_2 DMSO but not at 10% PCO_2 pinometostat compared to 5% PCO_2 DMSO at a six-hour incubation of higher PCO_2, alongside the gene list and library overlap and the statistical significance of the term at a cut-off of an adjusted $p \leq 0.05$ ordered by the combined score.

The thirteen genes that overlap between the two PCO_2 incubation time points which are upregulated at higher PCO_2 when treated with DMSO but do not change under pinometostat treatment were also analysed using EnrichR, the data for this is given in Table 5-30.

GO terms	GO identifier	Overlap	Adjusted P-value	Combined Score
Positive Regulation of Programmed Cell Death	GO:0043068	3/245	0.03203	187.34
Positive Regulation of Apoptotic Process	GO:0043065	3/270	0.03203	163.31
Ubiquitin-Dependent Protein Catabolic Process	GO:0006511	3/367	0.04968	104.85
Regulation Of DNA-templated Transcription	GO: 0006355	6/1922	0.03203	58.29

Table 5-30 The biological processes GO enrichment terms for genes that are upregulated at 10% PCO_2 DMSO but not at 10% PCO_2 pinometostat compared to 5% PCO_2 DMSO that were found at both the three and the six-hour incubation of higher PCO_2, alongside the gene list and library overlap and the statistical significance of the term at a cut-off of adjusted $p \leq 0.05$ ordered by the combined score.

5.13.4.3.4.2 GO terms for Genes Downregulated 10% PCO_2 DMSO but not under 10% PCO_2 Pinometostat Compared to 5% PCO_2 DMSO.

The gene lists which contained the 271 and 510 genes which are downregulated at 10% PCO_2 DMSO but not at 10% PCO_2 pinometostat compared to 5% PCO_2 DMSO at 3 and 6 hours respectively were analysed by EnrichR. The three-hour incubation gave no statistically significant enriched biological GO terms. The GO term results are given in Table 5-31 for the six-hour incubation.

GO terms	GO identifier	Overlap	Adjusted P-value	Combined Score
Protein Insertion into Mitochondrial Membrane	GO:0051204	6/28	0.045535	101.68
Proton Motive Force-Driven ATP Synthesis	GO:0015986	9/60	0.02164	73.91
Mitochondrial Translation	GO:0032543	12/98	0.01616	64.01

Table 5-31 The biological processes GO enrichment terms for genes that are downregulated at 10% PCO_2 DMSO but not at 10% PCO_2 pinometostat compared to 5% PCO_2 DMSO that were found at the six-hour incubation of higher PCO_2, alongside the gene list and library overlap and the statistical significance of the term at a cut-off of adjusted $p \leq 0.05$ ordered by the combined score.

Only three genes overlap between the two PCO_2 incubation time points that are downregulated at higher PCO_2 when treated with DMSO but do not change under pinometostat treatment. Any identified overlap between the GO terms for the gene list and library only relates to one gene in Table 5-32, and the combined score is skewed as the sample list of genes is so small, making it hard to draw conclusions from this data.

GO terms	GO identifier	Overlap	Adjusted P-value	Combined Score
Regulation Of Vacuole Organization	GO:0044088	1/17	0.026994433	2366.46
Regulation Of Autophagosome Assembly	GO:2000785	1/40	0.026994433	824.08
Negative Regulation of Autophagy	GO:0010507	1/61	0.026994433	488.52
Negative Regulation of Cellular Catabolic Process	GO:0031330	1/67	0.026994433	434.58
Regulation Of Cilium Assembly	GO:1902017	1/71	0.026994433	404.19
Regulation Of Plasma Membrane Bounded Cell Projection Assembly	GO:0120032	1/75	0.026994433	377.38
Regulation Of Organelle Assembly	GO:1902115	1/80	0.026994433	348.01
Retrograde Transport, Endosome to Golgi	GO:0042147	1/94	0.026994433	283.98
Activation Of GTPase Activity	GO:0090630	1/102	0.026994433	256.06
Golgi Vesicle Transport	GO:0048193	1/197	0.046589205	109.40
Positive Regulation of GTPase Activity	GO:0043547	1/234	0.04733902	87.06
Regulation Of Autophagy	GO:0010506	1/241	0.04733902	83.70

Table 5-32 The biological processes GO enrichment terms for genes that are downregulated at 10% PCO_2 DMSO but not at 10% PCO_2 pinometostat compared to 5% PCO_2 DMSO that was found at both the three and the six-hour incubation of higher PCO_2, alongside the gene list and library overlap and the statistical significance of the term at a cut-off of adjusted $p \leq 0.05$ ordered by the combined score.

5.13.4.4 Pinometostat Sensitive Genes.

Section 5.6.2 discusses the upregulation of *Hoxa9* and *Meis1* genes in MLL leukaemia. Studies have shown that pinometostat reduces *Hoxa9* gene expression which is central to its anticancer activity.[295] Table 5-33 shows the *Hoxa9* gene expression between the pairwise comparisons made between the samples, where the first sample listed is the baseline condition. There is no change in

expression of *Hoxa9* between the two incubation time points for DMSO treatment (5pct_D vs 10pct_D, 3 and 6 h) however *Hoxa9* is downregulated when treated with pinometostat compared with a baseline DMSO condition (5pct_D or 10pct_D) across all sample comparisons which is consistent with the literature. The log fold change (logFC) represents the difference in expression between the pairwise comparisons.

Meis1 was not detected as a housekeeping or as a differential expressed gene across the RNA-seq dataset therefore the expression levels of *Meis1* in HEK293 cells at 6.5 nTPM were likely too low to be detected.

Sample Comparison.	Log_2FC *Hoxa9* (116.9 nTPM)	Log_2FC *RUBICON* (12.9 nTPM)
5pct_D3_vs_5pct_P3	-0.6125	Not changed
5pct_D3_vs_10pct_D3	Not changed	Not changed
5pct_D3_vs_10pct_P3	-0.432	Not Changed
10pct_D3_vs_10_P3	-0.462	Not Identified
5pct_D6_vs_5pct_P6	-0.812	Not Changed
5pct_D6_vs_10pct_D6	Not changed	+0.448
5pct_D6_vs_10pct_P6	-0.757	+0.371
10pct_D6_vs_10_P6	-0.783	Not changed

Table 5-33 *Hoxa9* and *RUBICON* gene expression across sample comparisons where the differential expression stated as the Log_2FC meets the threshold cut-off of FDR ≤0.05. The sample notation used here is detailed in Table 5-13.

A study in human lung epithelial cells showed that Rubicon Autophagy Regulator (*RUBICON*) and Tripartite Motif Containing 25 (*TRIM25*) genes were upregulated under pinometostat treatment versus the DMSO control and that *BCL3* was downregulated under pinometostat treatment versus the DMSO control.[296] *RUBICON* is only upregulated in two of the pairwise comparisons including 10% PCO_2 with DMSO compared to the 5% PCO_2 DMSO baseline indicating that this may be a lung epithelial cell-specific response to pinometostat treatment. *TRIM25* (32.6 nTPM) was a housekeeping gene across all pairwise comparisons in this dataset and *BCL3* was at too low an expression level (0.5 nTPM) to be detected.

5.13.4.5 CO_2 Sensitive Genes

RPL19 is an appropriate housekeeping gene for assessing transcriptional change between 5 and 10% PCO_2 incubations because across both incubation time points there is no change in expression of RPL19.

Section 5.7 covers the transcriptional changes associated with hypercapnia that are of interest in the literature. The inflammatory response genes, *ICAM1* and *IL-8* were not identified in this RNA-seq experiment due to low expression levels in HEK293 cells at 0.1 nTPM and 0.2 nTPM, respectively. The differential expression of the Wnt signalling genes; *Fzd9* and *Wnt7a* between the two PCO_2 levels of 5% and 10% were analysed. Similarly, to *ICAM1* and *IL-8*, *Wnt7a* was at too low an expression level to be detected in HEK293 cells. Across all the pairwise comparisons, *Fzd9* expression was unchanged by higher CO_2 incubation at both 3 and 6 hours. The expres-ssion of *TMEM267* human equivalent of the murine gene *GM7120* was also assessed because the murine gene had been identified as transcriptionally upregulated under hypercapnia by Shigemura *et al*.[258] However, there was no change in *TMEM267* expression between CO_2 incubations at both time points in this study.

CO_2-sensitive genes established in the literature were not identified in this RNA-seq HEK293 dataset. A possible explanation for this is that the CO_2-sensitive genes identified by Shigemura *et al.* were significantly upregulated at 20% PCO_2 compared with 5% PCO_2 whereas this study uses 10% PCO_2 as the hypercapnic condition.

Transcriptional change linked to the inflammatory response was assessed by the Phelan et al,[259] study as detailed in Section 5.7. The deregulated gene lists from Tables 5-16, 5-18, 5-20, 5-22, 5-24, 5-26 and Phelan's study were both performed under buffered hypercapnia therefore could be compared to assess whether any common genes are identified. However, it is important to note that differences may arise due to the studies being performed in different cell lines and exposure to different stimuli.

Section 5.13.4.3.2 states that the number of genes significantly upregulated at 10% PCO_2 when

comparing 5% and 10% PCO_2 both treated with DMSO was 561 and 1507 for three and six- hour incubation time points, respectively. The number of genes significantly downregulated at 10% PCO_2 under 5% PCO_2 versus 10% PCO_2 was 652 and 1546 for the three and six- hour incubation time points, respectively. These deregulated genes across three replicates show that transcriptional change was evident between the CO_2 incubations at both incubation time frames and that the greater change was experienced following the six-hour incubation. GO terms that are consistent between the gene lists at three and six hours are regulatory terms associated with both RNA polymerase II and DNA-templated transcription. These terms indicate that differential CO_2 incubation impacts gene expression.

5.13.4.6 RNA-sequencing Dataset Discussion

This is the first study to assess the effects of pinometostat treatment on hypercapnic-induced transcriptional change in the HEK293 cell line. When designing this experiment, the use of pinometostat in HEK293 cells had not been performed yet, due to the interest of using the inhibitor in the cancer research field. However, a recent paper did use pinometostat in an RNA sequencing study with the HEK293T cell line to assess a labelling strategy for studying protein-protein interactions.[297] Publicly available RNA-sequencing analyses using pinometostat primarily use ALL and Acute Myeloid Leukaemia (AML) cell lines. *Hoxa9* had previously been identified as a downregulated gene due to pinometostat treatment.[295] In this study, *Hoxa9* was consistently downregulated in pinometostat treated cells compared with DMSO treated cells supporting the ELISA data that pinometostat treatment was effectively inhibiting H3K79 methylation.

Hoxa9 downregulation between 5% DMSO and 5% pinometostat with 5% DMSO and 10% pinometostat at both time points was looked at in more detail because the magnitude of downregulation was lower when the PCO_2 was changed between treatment and no treatment. For the three-hour incubation, *Hoxa9* is downregulated at a magnitude of -0.6 between 5% PCO_2 DMSO and 5% PCO_2 pinometostat whereas for 5% PCO_2 DMSO and 10% PCO_2 pinometostat, there is a downregulation of -0.4. Similarly, for the six-hour incubation, *Hoxa9* is downregulated at a magnitude

of -0.8 between 5% PCO_2 DMSO and 5% PCO_2 pinometostat whereas, for 5% PCO_2 DMSO and 10% PCO_2 pinometostat, there is a downregulation of -0.75. This finding supports the MTase-Glo assay data where DOT1L mediated methylation is stimulated under higher Ci under the assumption that the pinometostat treatment does not completely inhibit Dot1L. This assumption is consistent with the ELISA data in Figure 5-40 where mono-methylation was at ~50%, di-methylation at ~40% and trimethylation at ~30% compared with DMSO. However, when samples incubated at 5% PCO_2 treated with pinometostat are compared with samples incubated at 10% PCO_2 treated with pinometostat the fold change in *Hoxa9* expression is not statistically significant with a logFC of 0.18 at a P value of 0.08, FDR of 0.2 and a logFC of 0.06 at a P value of 0.46, FDR 0.62 for 3 hours and six hours, respectively. The logFC in *Hoxa9* expression at the three-hour incubation of differential CO_2 is greater and closer to the significance threshold than the equivalent logFC for the six-hour incubation. This indicates that a repeat experiment with the same pairwise comparisons using a different time incubation time point could give a significant FC in *Hoxa9* expression between samples incubated at 5% PCO_2 treated with pinometostat compared to samples incubated at 10% PCO_2 treated with pinometostat to support the MTase-Glo assay. However, this proposed experiment may be limited by the fact that CO_2 stimulation of DOT1L could be masked by pinometostat treatment.

Large-scale CO_2 transcriptomic studies in the mammalian context are divided into the deregulation of immunity/ inflammatory signalling [5,257] and the Wnt Signalling pathway.[258] Genes upregulated in 10% PCO_2 DMSO but not under 10% PCO_2 pinometostat when compared with 5% PCO_2 DMSO for both incubation time points (three and six hours) are linked to three enriched GO terms including, programmed cell death, apoptosis, and ubiquitin-dependent protein catabolic processing. Abnormal regulation of these terms is associated with immunological disorders which could be linked to hypercapnic induced NF-κB signalling suppression.[298] However, the deregulation of Wnt signalling was not identified in this dataset across any of the pairwise comparisons. Hypercapnic-induced transcriptional change is cell-line specific and current literature has a limited scope because at the time of writing, there are no CO_2 transcriptional change studies which use the HEK293 cell line.

The GO terms listed in section 5.13.4.3.4 are ordered by the largest to smallest combined score and in RNA-seq data a combined score threshold is commonly applied where the higher the number, the more stringent the cut-off. The terms which have a combined score of equal to or above 30 are listed here for genes that change under hypercapnia but not under hypercapnia treated with pinometostat at three and six-hour incubation time points. In addition to this, the overlap of genes between both the three and six-hour incubation time points that are deregulated in 10% PCO_2 DMSO but not in 10% PCO_2 pinometostat treatment are only statistically significant for the upregulated genes therefore the GO terms for downregulation at both time points are not discussed.

At the three-hour time point, the biological processes of positive regulation of cardiac muscle cell differentiation and mitotic spindle organisation were upregulated. At the six-hour time point, the biological process of positive plasma membrane-bounded cell projection morphogenesis was upregulated. The biological processes of protein insertion into the mitochondrial membrane, proton motive force-driven ATP synthesis and mitochondrial translation were downregulated. At both time points, in addition to the immunologically linked biological processes discussed above, DNA-templated transcription was upregulated.

The key takeaway from this experiment was that the transcriptional change that normally occurs between 5% and 10% PCO_2 is altered in the presence of pinometostat. Therefore, DOT1L and subsequently H3K79 methylation does influence CO_2-dependent transcriptional change. The GO terms listed could provide useful leads into future experiment designs to assess CO_2 sensitivity with and without pinometostat treatment.

5.14 Discussion

Histone carbamylation sites that can be stated with high confidence identified in this study are H1K46, H1K85, H3K79, H4K32, and H4K92 as these sites were identified more than once across different experiments in both 12C and 13C Ci experiments. The identification of histone carbamylation sites in native nucleosomes and the use of lysine propionylation to improve nucleosome coverage was discussed in sections 5.8-5.10. Propionylation of native nucleosomes did not improve coverage which may have been due to contamination associated with preparation from whole HEK293 cell lysates. A purification strategy for recombinant nucleosomes was implemented as discussed in sections 5.10 and 5.11. Analytical sizing indicated the successful purification of the recombinant histone octamer. Difficulties were experienced when trying to scale up widom DNA production for nucleosome preparation as discussed in section 5.11.4. Therefore, despite the histone octamer being less biologically relevant it was used in a trapping screen as detailed in section 5.11.5. Propionylation was identified as a suitable technique for improving H3 and H4 coverage in purified nucleosomes. H4K32, H3K79 and H4K92 were identified as carbamate sites in the histone octamer screening.

Subsequent experiments performed in this chapter were focused on the histone site H3K79. This site was an attractive target for several reasons including that H3K79 methylation is only performed by the enzyme, Dot1L, methylation at the site is linked to increased transcriptional activity and the site is actively researched due to the role of DOT1L in a range of cellular processes meaning effective Dot1L inhibitors are readily available.

An *in-vitro* assessment of Dot1L-mediated methylation under varying Ci showed that DOT1L activity was stimulated at higher increased Ci concentrations which then plateaued within the detection range of the MTase-Glo assay. An ELISA plate to assess the degree of methylation under different Ci conditions was considered. However, the data in the ELISA assay in 5.13.2 indicates that the percentage of H3K79 methylation is much higher than has been detected previously by MS as described in section 5.5. Therefore, this route was deemed too unspecific for detecting Ci-related

changes and not pursued. The exact mechanism to explain the MTase-Glo assay result is too complex to identify from this preliminary study and would require nucleosome mutation variants and further trapping experiments to explain as detailed further in section 5.12.4. The conclusion drawn from this *in-vitro* experiment was that the stimulation of DOT1L activity by increasing Ci leads to a higher proportion of H3K79 being methylated and an increase in DNA transcription.

To complement the results of the MTase-Glo assay, an RNA sequencing screen which could assess *in-cellulo* changes due to carbamylation on H3K79 was developed. The first step of this work was to identify a pinometostat concentration and incubation length which led to Dot1L inhibition. This was identified as 1 µM of pinometostat for 10 days using an ELISA and MTT assay. qPCR experiments were run to identify an incubation time point for CO_2 exposure. However, due to low expression of CO_2 sensitive genes in HEK293 cells, this proved difficult therefore incubation lengths of three- hours and six- hours were chosen for RNA sequencing. The RNA-seq data in this study was of high quality as shown by all FastQ metrics outlined in section 5.13.4.1. Data processing showed independent clustering of replicates from each sample group meaning data was suitable for differential gene expression analysis. EnrichR identified enriched GO terms for 10% PCO_2 treated with DMSO but not when treated with pinometostat compared to the 5% PCO_2 DMSO baseline as discussed in greater detail in section 5.13.4.3.4. Importantly RNA sequencing identified a subset of genes that are dependent on DOT1L activity under hypercapnic CO_2 levels, a result which is consistent with the *in-vitro* assay.

In conclusion, LCMSMS experiments have identified carbamates on multiple histone lysines with the potential of verifying further sites in the future. The *in-vitro* and *in-cellulo* studies indicate that carbamylation on H3K79 is biologically relevant and influences gene transcription. Further studies could be performed on the other identified histone sites to determine their biological relevance and build evidence for adding the carbamylation PTM to the histone code.

5.15 Future Work

The work discussed in this chapter has opened several potential avenues for future work. After successful DNA and histone octamer reconstitution, sucrose gradient ultracentrifugation is required for effective nucleosome purification.[299] Further screening using recombinant nucleosomes trapped with 12C and 13C inorganic carbon would be useful for further validation of histone carbamate hits. Furthermore, trapping with varying 12C and 13C Ci with a range of Ci concentrations between 0 and 50 mM could be performed to quantify whether the percentage of carbamylated histone peptides compared with total histone content is Ci concentration dependent. The H4 sites of HK32 and H4K92 were identified as carbamylation hits multiple times and therefore their biological relevance could be assessed.

To build a profile of carbamylation effects across the different histone sites, a global assessment of chromatin packing could be implemented using chromosome conformation capture and DNA sequencing of interacting fragments between DNA from cells incubated at hypercapnic and normoxic levels of CO_2.[300] An iTraq study to quantify well-known histone modifications for example acetylation and methylation under hypercapnic and normoxic CO_2 conditions could provide insight into the importance of histone carbamylation sites and crosstalk mechanisms that are affected by CO_2.

In terms of further study regarding H3K79, trapping on Dot1L is necessary to ensure carbamylation is nucleosome specific. The MTase-Glo assay could be performed with specific mutants to delineate the stimulation effects of increasing Ci concentration on Dot1L activity. A qPCR study to assess *Hoxa9* expression under 5% PCO_2 treated with pinometostat compared with 10% PCO_2 treated with pinometostat normalised to 5% PCO_2 treated with DMSO using varying incubation lengths at differential PCO_2 could be performed to validate DOT1L stimulation at higher Ci concentrations. The GO terms identified from the RNA-sequencing study indicate biological processes to target for further delineating H3K79 carbamylation effects. Finally, further gene comparison studies on the RNA sequencing data could be conducted using publicly available RNA-seq datasets from the GEO website.

6. Synopsis

6.1 Introduction

CO_2 is an important research target with known roles in respiration, [301] metabolism, [32] and cell signalling.[33] In mammalian systems, CO_2 homeostasis is crucial for preventing hypercapnia-linked diseases, which may arise due to elevated PCO_2-induced transcriptomic change.[141] This thesis aimed to identify and biologically validate mammalian CO_2 target proteins.

CO_2 interacts with proteins via hydrogen bonding [66] and the carbamylation of neutral amines.[302] In silico [64–66] and mass spectrometry approaches [1,72] have been applied to identify CO_2 binding sites in proteins. Interestingly, a carbamate prediction model indicated that carbamylation could be used by at least 1.3% of large proteins.[71] This model was trained using previously identified stable carbamates buried within the protein structure or stabilised by a metal cation. Therefore, it is likely that this model only represents a subset of stable carbamates because the modification can spontaneously occur on solvent-exposed structurally privileged amines which are deprotonated at physiological pH.[303] This thesis aims to build upon the knowledge of the relatively understudied and liable carbamylation PTM by using a proteomic tool to identify carbamates across the mammalian proteome systematically.

Following carbamate screening, two CO_2-target proteins were selected to illustrate the biological relevance of the carbamate PTM. The first target was ubiquitin K48, first identified as a carbamate site by Linthwaite *et al.*[84] This provided the foundation for assessing the effects of carbamylation on PROTAC-mediated proteasomal degradation. The second carbamylation target studied was Histone H3 K79, which was identified by the MS proteome screen. The effects of carbamylation on DNA transcription were investigated in the context of H3K79. This chapter summarises the results outlined in this investigation and the future direction for this study.

6.2 A Mammalian Carbamylation Proteome Screen.

The work conducted in Chapter 3 builds on Linthwaite *et al.'s* trapping methodology for carbamate detection.[1,87] The trapping method was adapted to a HEK293 lysate proteome screen, and these adaptations improved coverage and facilitated the validation of carbamate sites. A fractionation-based workflow [135] was used and increased the mammalian proteome coverage by 7-fold, and two database search algorithms, namely, PEAKs and Protein Pilot were used for carbamate identification. In addition, two isotopes of Ci were used in two trapping datasets to separate carbamates from the AGE-derived CML modification.[104] Twenty-seven reproducible novel carbamate sites were identified in the mammalian lysate screen across search algorithms and isotope datasets. Out of the twenty-seven hits, nine carbamate sites were identified by both the 12C and 13C datasets using PEAKs.

The results showed that PEAKs was less error-prone than Protein Pilot, and 13C Ci should be used for carboxyethyl site identification. Although the screening method was successfully applied to HEK293 lysates, it was limited by identifying false positives, available resources, and the proteomic scale. In conclusion, several carbamate sites were identified, which are worthwhile avenues for future study, and this screening directed the work completed in Chapter 5 on nucleosome carbamylation.

6.3 PROTACs and CO_2

The aim of Chapter 4 was to apply carbamylation to a pharmaceutical setting. PROTAC compounds have been developed to hijack the UPS to increase the degradation rate of disease-related proteins.[150] K48-linked polyubiquitin chains are recognisable by the 26S proteasome, which breaks down proteins into amino acids for protein turnover.[102] Linthwaite *et al*. reported that carbamylation at ubiquitin K48 is sensitive to Ci concentration, particularly under elevated Ci, K48-linked ubiquitin conjugation decreases.[84]

A HiBiT- nano Glo luminescence assay was used to investigate the activity of PROTACS under elevated CO_2.[304] A sensitive dose-response assay was applied in two cell lines, each with a different chromatin remodelling target protein. The assay produced reproducible data across both targets and was substrate and cell-line-independent. However, there was no significant effect on the activity of any of the PROTACs tested between normoxic and hypercapnic CO_2 levels. This is a positive outcome for the clinical application of PROTACs because hypercapnia does not affect the potency of these compounds. It was concluded that the effects of CO_2 on polyubiquitin chain formation are too mild to influence the degradation efficiency of PROTACs. Due to this result, nucleosomal carbamylation sites identified by the mammalian proteomic screen (section 6.2) were considered for future study instead.

6.4 Histone Carbamylation

Chromatin is a dynamic DNA packaging structure composed of nucleosome subunits that play an important role in regulating DNA accessibility and transcription.[305] The nucleosome contains four histone proteins, namely H2A, H2B, H3 and H4, which can be post-translationally modified to alter the interaction of the histone proteins with DNA, and these PTMs are summarised by the histone code.[306] A considerable number of histone PTMs have been reported to date, and this is an active research area due to the linked importance of histone PTMs in altering gene expression.[307]

Out of the nine reproducible carbamate hits identified by both the 12C and 13C HEK293 lysate screens, seven of these hits were identified on histone proteins. Three further histone carbamate hits were identified in HEK293 screening but were only seen in the 12C or 13C datasets.

Before choosing a specific histone target site for future study, a screen for carbamates on native nucleosomes was completed. Following nucleosome trapping, the sample preparation procedure was optimised by trailing three proteases and introducing a chemical modification called propionylation to increase the length of histone peptides to improve coverage.[260] The native nucleosome datasets identified seven histone hits, six of which were identified in the HEK293 lysate screen. However, only three of these histone hits were reproducible across the native nucleosome datasets. These optimisation datasets indicated that the native nucleosome extraction process was impure and contaminant proteins could interfere with histone protein coverage. To address this hypothesis, the preparation of purified recombinant nucleosomes was pursued.

The recombinant histone octamer was purified in this investigation using small modifications to Klinker *et al.'s* protocol.[274] First, the four individual histone variants were overexpressed and then extracted using urea to solubilise the target histone. A denaturing cationic exchange was performed after solubilisation, followed by anionic exchange. When all four variants were produced in high enough quantities, each purified histone was unfolded individually. Then, the four variants were recombined together in a specific ratio to enable refolding into the histone octamer. Purification of

the DNA sequence for nucleosome formation was attempted; however, it proved difficult to produce DNA quantities at the required scale. Therefore, the histone octamer was used for trapping experiments to profile histone carbamylation sites. The histone octamer samples were either propionylated or not propionylated. Three carbamate hits were identified, two reproducible, and all three had been identified previously in the HEK293 lysate screen. In these datasets, propionylation was better for H3 and H4 coverage but not for H2A and H2B coverage, corresponding with the expected result due to protein sequence for these variants. This dataset was limited by the amount of histone octamers available and was not as biologically relevant as the trapping on recombinant nucleosomes. The histone hits and the number of times they were identified across the various screening stages (HEK293 lysates, native nucleosome, and recombinant histone octamer) are detailed in Table 6-1.

Histone Variant	Modified Lysine	12C PEAKS HEK293	12C PP HEK293	13C HEK293	Native Nucleosome 12C/13C, modified/ unmodified	Histone Octamer 12C/13C, modified/ unmodified
P62805	32	3	3	5	6	7
P10412/P16402/ P16403	63	6	6	0	4	0
P62805	92	0	7	3	3	0
P68431/P84243/ Q71DI3	80	3	3	1	0	1
P10412/P16402/ P16403	85	2	2	3	1	0
P10412/P16402/ P16403	106	3	1	2	0	0
P68431/P84243/ Q71DI3	123	2	1	2	0	0
P10412/P16402/ P16403	46	1	1	1	1	0
P10412/P16402/ P16403	90	0	0	0	1	2
P68431/P84243/ Q71DI3	57	0	0	1	1	0
O60814	109	0	2	0	0	0

Table 6-1 Histones identified as carbamate-modified proteins across the different stages of hit identification listed by the total number of times identified from highest to lowest. The 12C HEK293

dataset is split into the two database search algorithms used, and PP stands for Protein Pilot. All further searches were completed by PEAKs only. For simplicity, the native nucleosome and histone octamer datasets are grouped.

From the carbamate sites identified in Table 6-1, H3K79 was selected as the target site for future work. The reasoning was that H3K79 plays a diverse biologically relevant regulation role in various processes via the post-translational modification of a mono, di or tri methylation group mediated solely by the methyltransferase DOT1L.[238] The availability of the propriety MTase Glo assay, DOT1L inhibitors and the literature on DOT1L-mediated changes in gene expression due to H3K79 methylation [251,252,295] provided a foundation for conducting carbamylation research on this target.

The MTase Glo assay was run to assess whether carbamylation at H3K79 altered the methylation rate by DOT1L. The luminescence-based assay measured the conversion of the methyl donor SAM to SAH, which is the methyl transfer product. The assay optimisation steps were used to identify the reaction's linear range, the buffer's stability under varying Ci and the capability of 0.5% TFA to stop the reaction. The final result of the assay indicated that under high levels of Ci, the methylation rate of H3K79 is stimulated. This result contrasted with the original hypothesis that H3K79 methylation would be reduced due to a carbamate residing on the DOT1L target site. Due to the complexities of the histone code and the phenomenon of PTM crosstalk, determining the exact reason for the DOT1L stimulation under elevated Ci requires further supporting experiments.

To support the *in-vitro* MTase Glo assay result, an *in-cellulo* approach was used to assess the transcriptional changes of elevated CO_2 in the context of H3K79 methylation. Firstly, the incubation time frame and concentration of pinometostat required to inhibit DOT1L methylation successfully were determined using an ELISA plate with antibodies for mono, di, and tri-methylated H3K79. Secondly, the time point to expose HEK293 cells to high CO_2 was investigated by qPCR. The expression levels of the CO_2-sensitive *Fzd9* gene [258] were tested at various time points. However, it was concluded that the expression levels of *Fzd9* were too low to identify a robust CO_2 incubation length. RNA samples

extracted from HEK293 cells incubated at three or six hours at normoxic or hypercapnic CO_2 levels treated with or without pinometostat were submitted in triplicate for RNA sequencing.

The data obtained passed all RNA sequencing quality metrics, and the PCA showed a clear separation between the conditions tested. The pinometostat-sensitive gene, *Hoxa9*, was shown to be downregulated between pinometostat and DMSO-treated samples, confirming that the inhibitor treatment length and concentration were suitable for DOT1L inhibition. The RNA sequencing data identified several genes with deregulated expression when HEK293 cells were incubated at 10% CO_2 and treated with pinometostat compared to cells incubated at 10% CO_2 and treated with DMSO when using the 5% CO_2 cells treated with DMSO as a baseline. This result highlights that the transcriptional change associated with normoxic and hypercapnic PCO_2 is altered in the presence of pinometostat. This result is of wider importance because it supports the hypothesis that DOT1L and subsequently, H3K79 methylation influences CO_2-dependent transcriptional change.

In conclusion, the data outlined in this section indicates that histone carbamylation influences DNA transcription. The exact mechanism underpinning this change will be a challenge to identify due to the complexity of the histone code and the challenges associated with global histone carbamate identification.

6.5 Conclusions

The results described in this thesis present the trapping methodology on the proteome scale and investigates the biological relevance of two carbamate hits. Carbamate hit validation across the HEK293 proteome revealed PEAKs is a suitable database search algorithm for carbamate identification. Using 13C Ci during the trapping stages further validates the presence of a carbamate. In total, 27 reproducible novel carbamate sites were identified in the mammalian lysate screen. The histone carbamylation sites were the most prevalent from these hits, and H3K79 was selected as a target site for future study. The other carbamate site selected was Ub K48, a key biological target due to its importance in protein turnover.

Carbamylation on Ubiquitin K48 had previously been characterised as biologically relevant, and in this study, the effect was examined using PROTACs. A firm conclusion was reached in this work that carbamylation at Ub K48 has no significant impact on PROTAC degradation efficiencies.

Carbamylation of H3K79 was biologically assessed by *in-vivo* and *in-cellulo* methodologies. The results of these experiments indicate that increased Ci or PCO_2 levels alter the activity of DOT1L and result in DNA transcriptional change. In conclusion, the specific regulation mechanism is still to be uncovered, but this thesis presents evidence to suggest that carbamylation is part of the histone code.

6.6 Future Work

The principal focus for future work should be on the regulation of histone proteins by carbamylation. The identified histone carbamylation sites should be verified in recombinant nucleosomes and assessed for biological relevance where appropriate. A useful avenue to build a profile of possible biologically relevant histone carbamylation sites would be to use quantitative mass spectrometry to assess acetylation and methylation levels of nucleosomes under high and low CO_2 levels. The data obtained could be verified by consulting previous literature on known histone crosstalk mechanisms, and mutational analysis studies to delineate biologically relevant histone carbamylation sites.

Further investigation into H3K79 methylation levels under hypercapnia could be investigated using the RNA sequencing dataset as a foundation. The GO terms identified from the RNA-sequencing study indicate biological processes to target for further delineating H3K79 methylation-associated carbamylation effects. Effects of carbamylation in other cell lines could be investigated to verify these transcriptional changes are not isolated to HEK293 cells. Trapping on DOT1L would be a useful experimental approach to confirm the methylation effects are nucleosome dependent.

The PROTAC work requires no future study because a clear conclusion was reached. Similarly, the proteomic screening of carbamylation sites using the trapping method has a limited scope. This study performed the proteomic screen at a standard that meets the trapping methodology's capabilities. A pre-trapping protein purification method, such as ammonium sulfate precipitation, could be performed to increase coverage further. However, it is hypothesised that the limitations of the results arise from applying this analysis on a proteome scale and the type of mass spectrometer used.

In the future, it is desirable to use the trapping methodology on purified proteins of interest with 13C Ci. The carbamates detected in this study should be verified using trapping on the recombinant protein or protein complex to verify these sites further before biological relevance

studies. Computational approaches using previously identified carbamates and those in this study to train a carbamate prediction model could provide an alternative approach for global carbamate identification.

7. Bibliography

1. Linthwaite, V. L. *et al.* The identification of carbon dioxide mediated protein post-translational modifications. *Nat Commun* (2018) doi:10.1038/s41467-018-05475-z.

2. Martin, W. F., Bryant, D. A. & Beatty, J. T. A physiological perspective on the origin and evolution of photosynthesis. *FEMS Microbiology Reviews* vol. 42 205–231 Preprint at https://doi.org/10.1093/FEMSRE/FUX056 (2018).

3. Jensen, F. B. *Red Blood Cell PH, the Bohr Effect, and Other Oxygenation-Linked Phenomena in Blood O 2 and CO 2 Transport*. *Acta Physiol Scand* vol. 182 (2004).

4. Cummins, E. P., Selfridge, A. C., Sporn, P. H., Sznajder, J. I. & Taylor, C. T. Carbon dioxide-sensing in organisms and its implications for human disease. *Cellular and Molecular Life Sciences* vol. 71 831–845 Preprint at https://doi.org/10.1007/s00018-013-1470-6 (2014).

5. Taylor, C. T. & Cummins, E. P. Regulation of gene expression by carbon dioxide. *J Physiol* **589**, 797–803 (2011).

6. Beard, D. A., Wu, F., Cabrera, M. E. & Dash, R. K. Modeling of cellular metabolism and microcirculatory transport. *Microcirculation* **15**, 777–793 (2008).

7. Lee Hamm, L., Nakhoul, N. & Hering-Smith, K. S. Acid-base homeostasis. *Clinical Journal of the American Society of Nephrology* **10**, (2015).

8. Rosenberg, R. M. & Peticolas, W. L. Henry's law: A retrospective. *J Chem Educ* **81**, (2004).

9. Green, O. *et al.* Activity-Based Approach for Selective Molecular CO2Sensing. *J Am Chem Soc* (2022) doi:10.1021/jacs.2c02361.

10. Pedersen, O., Colmer, T. D. & Sand-Jensen, K. Underwater photosynthesis of submerged plants - Recent advances and methods. *Front Plant Sci* **4**, (2013).

11. KREBS, H. A. Carbonic anhydrase as a tool in studying the mechanism of enzymic reactions involving H2CO3, CO2 or HCO3. *Biochem J* (1948) doi:10.1042/bj0430550.

12. Supuran, C. T. Carbonic anhydrases: Novel therapeutic applications for inhibitors and activators. *Nature Reviews Drug Discovery* vol. 7 168–181 Preprint at https://doi.org/10.1038/nrd2467 (2008).

13. Sjöblom, B., Polentarutti, M. & Djinović-Carugo, K. *Structural Study of X-Ray Induced Activation of Carbonic Anhydrase*. *Proceedings of the National Academy of Sciences of the United States of America* vol. 106 www.pnas.org/cgi/content/full/ (2009).

14. Itada, N. & Forster, R. E. *Carbonic Anhydrase Activity in Intact Red Blood Cells Measured with 18O Exchange**. *THE JOURNAL OF BIOLOGICAL CHEMISTRY* vol. 252 (1977).

15. Endeward, V. *et al.* Evidence that aquaporin 1 is a major pathway for CO 2 transport across the human erythrocyte membrane. *The FASEB Journal* (2006) doi:10.1096/fj.04-3300com.

16. Zhang, X., Barraza, K. M. & Beauchamp, J. L. Cholesterol provides nonsacrificial protection of membrane lipids from chemical damage at air–water interface. *Proc Natl Acad Sci U S A* **115**, 3255–3260 (2018).

17. Waisbren, S., Geibel, J., Modlin, I. & Boron, W. Unusual permeability properties of gastric gland cells. *Nature* 332–335 (1994).

18. Gruswitz, F. *et al.* Function of human Rh based on structure of RhCG at 2.1 Å. *PNAS* **107**, 9638–9643 (2010).

19. Forster, R. E., Gros, G., Lin, L., Ono, Y. & Wunder, M. *The Effect of 4,4-Diisothiocyanato-Stilbene-2,2-Disulfonate on CO 2 Permeability of the Red Blood Cell Membrane*. *Physiology* vol. 95 www.pnas.org. (1998).

20. Geyer, R. R., Musa-Aziz, R., Qin, X. & Boron, W. F. Relative CO2/NH3 selectivities of mammalian aquaporins 0-9. *Am J Physiol Cell Physiol* (2013) doi:10.1152/ajpcell.00033.2013.

21. Michenkova, M. *et al.* Carbon dioxide transport across membranes. *Interface Focus* vol. 11 Preprint at https://doi.org/10.1098/rsfs.2020.0090 (2021).

22. Parker, M. D. & Boron Walter F. The Divergence, Actions, Roles, and Relatives of Sodium-Coupled Bicarbonate Transporters. *Physiol Rev* **93**, 803–959 (2013).

23. Kustu, S. & Inwood, W. Biological gas channels for NH3 and CO2: evidence that Rh (Rhesus) proteins are CO2 channels. *Transfusion Clinique et Biologique* **13**, 103–110 (2006).

24. Niciu, M. J., Kelmendi, B. & Sanacora, G. Overview of glutamatergic neurotransmission in the nervous system. *Pharmacology Biochemistry and Behavior* vol. 100 656–664 Preprint at https://doi.org/10.1016/j.pbb.2011.08.008 (2012).

25. Bonham, A. C. *Frontiers Review Neurotransmitters in the CNS Control of Breathing*. *Respiration Physiology* vol. 101 (1995).

26. Vaughan-jones, R. D. & Spitzer, K. W. Role of bicarbonate in the regulation of intracellular pH in the mammalian ventricular myocyte. *BioChemistry and Cell Biology* **80**, 579–596 (2002).

27. Alka, K. & Casey, J. R. Bicarbonate transport in health and disease. *IUBMB Life* vol. 66 596–615 Preprint at https://doi.org/10.1002/iub.1315 (2014).

28. Lee, M. G., Ohana, E., Park, H. W., Yang, D. & Muallem, S. Molecular mechanism of pancreatic and salivary gland fluid and HCO 3- secretion. *Physiological Reviews* vol. 92 39–74 Preprint at https://doi.org/10.1152/physrev.00011.2011 (2012).

29. Lee, D. & Hong, J. H. The fundamental role of bicarbonate transporters and associated carbonic anhydrase enzymes in maintaining ion and pH homeostasis in non-secretory organs. *International Journal of Molecular Sciences* vol. 21 Preprint at https://doi.org/10.3390/ijms21010339 (2020).

30. Guyenet, P. G. & Bayliss, D. A. Neural Control of Breathing and CO2 Homeostasis. *Neuron* vol. 87 946–961 Preprint at https://doi.org/10.1016/j.neuron.2015.08.001 (2015).

31. Nattie, E. & Li, A. Central chemoreceptors: Locations and functions. *Compr Physiol* **2**, 221–254 (2012).

32. Cummins, E. P., Strowitzki, M. J. & Taylor, C. T. Mechanisms and consequences of oxygen and carbon dioxide sensing in mammals. *Physiol Rev* **100**, 463–488 (2020).

33. Phelan, D. E., Mota, C., Lai, C., Kierans, S. J. & Cummins, E. P. Carbon dioxide-dependent signal transduction in mammalian systems. *Interface Focus* **11**, 20200033 (2021).

34. Bayliss, D. A., Barhanin, J., Gestreau, C. & Guyenet, P. G. The role of pH-sensitive TASK channels in central respiratory chemoreception. *Pflugers Archiv European Journal of Physiology* vol. 467 917–929 Preprint at https://doi.org/10.1007/s00424-014-1633-9 (2015).

35. Kumar, N. *et al.* Regulation of breathing by CO2 requires the proton-activated receptor GPR4 in retrotrapezoid nucleus neurons. *Science (1979)* **384**, 1255–1260 (2015).

36. Chandrashekar, J. *et al.* The taste of carbonation. *Science (1979)* **326**, 443–445 (2009).

37. Darst, S. A. *et al. Soluble Adenylyl Cyclase as an Evolutionarily Conserved Bicarbonate Sensor*. *Cold Spring Harbor Symp. Quant. Biol* vol. 3 https://www.science.org (1999).

38. Sun, L. *et al. Guanylyl Cyclase-D in the Olfactory CO 2 Neurons Is Activated by Bicarbonate*. www.pnas.org/cgi/content/full/ (2008).

39. Zhou, Y. *et al.* Role of Receptor Protein Tyrosine Phosphatase γ in Sensing Extracellular CO2 and HCO3. *J Am Soc Nephrol* **27**, 2616–2621 (2016).

40. Stojanovska, V., Miller, S. L., Hooper, S. B. & Polglase, G. R. The consequences of preterm birth and chorioamnionitis on brainstem respiratory centers: Implications for neurochemical development and altered functions by inflammation and prostaglandins. *Frontiers in Cellular Neuroscience* vol. 12 Preprint at https://doi.org/10.3389/fncel.2018.00026 (2018).

41. Stams, T. *et al. Crystal Structure of the Secretory Form of Membrane-Associated Human Carbonic Anhydrase IV at 2.8-Å Resolution (Protein Crystallographyzinc Enzymecarbon Dioxide Hydration)*. *Biochemistry* vol. 93 (1996).

42. Sheriff, S. *et al.* Small molecule receptor protein tyrosine phosphatase γ (RPTPγ) ligands that inhibit phosphatase activity via perturbation of the tryptophan-proline-aspartate (WPD) loop. *J Med Chem* **54**, 6548–6562 (2011).

43. Phelan, D. E., Mota, C., Lai, C., Kierans, S. J. & Cummins, E. P. Carbon dioxide-dependent signal transduction in mammalian systems. *Interface Focus* **11**, 20200033 (2021).

44. Bonizzi, G. & Karin, M. The two NF-κB activation pathways and their role in innate and adaptive immunity. *Trends in Immunology* vol. 25 Preprint at https://doi.org/10.1016/j.it.2004.03.008 (2004).

45. Yu, H., Lin, L., Zhang, Z., Zhang, H. & Hu, H. Targeting NF-κB pathway for the therapy of diseases: mechanism and clinical study. *Signal Transduction and Targeted Therapy* vol. 5 Preprint at https://doi.org/10.1038/s41392-020-00312-6 (2020).

46. Cummins, E. P. & Keogh, C. E. Respiratory gases and the regulation of transcription. *Experimental Physiology* Preprint at https://doi.org/10.1113/EP085715 (2016).

47. Takeshita, K. *et al.* Hypercapnic Acidosis Attenuates Endotoxin-Induced Nuclear Factor-κB Activation. *Am J Respir Cell Mol Biol* **29**, 124–132 (2003).

48. Kawai, T. & Akira, S. Signaling to NF-κB by Toll-like receptors. *Trends in Molecular Medicine* vol. 13 Preprint at https://doi.org/10.1016/j.molmed.2007.09.002 (2007).

49. Tang, S. E. *et al.* Pre-treatment with ten-minute carbon dioxide inhalation prevents lipopolysaccharide-induced lung injury in mice via down-regulation of toll-like receptor 4 expression. *Int J Mol Sci* **20**, (2019).

50. Jaitovich, A. *et al.* High CO2 levels cause skeletal muscle atrophy via AMP-activated kinase (AMPK), FoxO3a protein, and muscle-specific ring finger protein 1 (MuRF1). *Journal of Biological Chemistry* (2015) doi:10.1074/jbc.M114.625715.

51. Jaitovich, A. *et al.* High CO2 Levels Cause Skeletal Muscle Atrophy via AMP-activated Kinase (AMPK), FoxO3a Protein, and Muscle-specific Ring Finger Protein 1 (MuRF1). *Journal of Biological Chemistry* **290**, 9183–9194 (2015).

52. Vohwinkel, C. U. *et al.* Elevated CO2 Levels Cause Mitochondrial Dysfunction and Impair Cell Proliferation. *Journal of Biological Chemistry* **286**, 37067–37076 (2011).

53. Casalino-Matsuda, S. M. *et al.* Hypercapnia Alters Expression of Immune Response, Nucleosome Assembly and Lipid Metabolism Genes in Differentiated Human Bronchial Epithelial Cells. *Sci Rep* **8**, 13508 (2018).

54. Bautista, A. F. & Akca, O. Hypercapnia: is it protective in lung injury? *Med Gas Res* **3**, (2013).

55. De Vito, E. L., Roncoroni, A. J., Berizzo, E. E. A. & Pessolano, F. Effects of spontaneous and hypercapnic hyperventilation on inspiratory effort sensation in normal subjects. *Am J Respir Crit Care Med* **158**, (1998).

56. May, A. *et al.* Non-invasive carbon dioxide monitoring in patients with cystic fibrosis during general anesthesia: end-tidal versus transcutaneous techniques. *J Anesth* (2020) doi:10.1007/s00540-019-02706-5.

57. B.Y., S. Obesity Hypoventilation: Pathophysiology, Diagnosis, and Treatment. *Curr Pulmonol Rep* (2019) doi:10.1007/s13665-019-0223-x LK - http://sfx.library.uu.nl/utrecht?sid=EMBASE&issn=21992428&id=doi:10.1007%2Fs13665-019-0223-x&atitle=Obesity+Hypoventilation%3A+Pathophysiology%2C+Diagnosis%2C+and+Treatment&stitle=Curr.+Pulm.+Reports&title=Current+Pulmonology+Reports&volume=8&issue=2&spage=31&epage=39&aulast=Sunwoo&aufirst=Bernie+Young&auinit=B.Y.&aufull=Sunwoo+B.Y.&coden=&isbn=&pages=31-39&date=2019&auinit1=B&auinitm=Y.

58. Nin, N., Angulo, M. & Briva, A. Effects of hypercapnia in acute respiratory distress syndrome. *Ann Transl Med* (2018) doi:10.21037/atm.2018.01.09.

59. Bustamante-Fermosel, A., De Miguel-Yanes, J. M., Duffort-Falcó, M. & Muñoz, J. Mortality-related factors after hospitalization for acute exacerbation of chronic obstructive pulmonary disease: the burden of clinical features. *American Journal of Emergency Medicine* (2007) doi:10.1016/j.ajem.2006.09.014.

60. Wang, L., Yang, L., Yang, J. & Shan, S. Effects of Permissive Hypercapnia on Laparoscopic Surgery for Rectal Carcinoma. *Gastroenterol Res Pract* (2019) doi:10.1155/2019/3903451.

61. Morales-Quinteros, L. *et al.* The role of hypercapnia in acute respiratory failure. *Intensive Care Med Exp* (2019) doi:10.1186/s40635-019-0239-0.

62. Ambalavanan, N. & Carlo, W. A. Hypocapnia and hypercapnia in respiratory management of newborn infants. *Clin Perinatol* (2001) doi:10.1016/S0095-5108(05)70104-4.

63. Uribarri, J. & Oh, M. S. Acid-Base Balance in Dialysis Patients: A Reassessment. *Seminars in Dialysis* Preprint at https://doi.org/10.1111/j.1525-139X.1995.tb00339.x (1995).

64. Drummond, M. L., Wilson, A. K. & Cundari, T. R. The importance of secondary structure in determining CO2-protein binding patterns. *J Mol Model* **18**, 2527–2541 (2012).

65. Drummond, M. L., Wilson, A. K. & Cundari, T. R. Nature of protein-CO2 interactions as elucidated via molecular dynamics. *Journal of Physical Chemistry B* **116**, 11578–11593 (2012).

66. Cundari, T. R. *et al.* CO2-formatics: How do proteins bind carbon dioxide? *J Chem Inf Model* **49**, 2111–2115 (2009).

67. Lorimer, G. H. Carbon dioxide and carbamate formation: the makings of a biochemical control system. *Trends Biochem Sci* (1983) doi:10.1016/0968-0004(83)90393-6.

68. Lide, D. R. *et al.* Properties of Amino Acids Table in CRC Handbook of Chemistry and Physics. *CRC Press* (2005).

69. Andersen, C. B. Understanding carbonate equilibria by measuring alkalinity in experimental and natural systems. *Journal of Geoscience Education* **50**, 389–403 (2002).

70. Nolting, D. *et al.* pH-induced protonation of lysine in aqueous solution causes chemical shifts in X-ray photoelectron spectroscopy. *J Am Chem Soc* **129**, 14068–14073 (2007).

71. Jimenez-Morales, D., Adamian, L., Shi, D. & Liang, J. Lysine carboxylation: Unveiling a spontaneous post-translational modification. *Acta Crystallogr D Biol Crystallogr* **70**, 48–57 (2014).

72. King, D. T. *et al.* Chemoproteomic identification of CO2-dependent lysine carboxylation in proteins. *Nat Chem Biol* **18**, (2022).

73. Kollipara, L. & Zahedi, R. P. Protein carbamylation: In vivo modification or in vitro artefact? *Proteomics* **13**, (2013).

74. Sun, S., Zhou, J. Y., Yang, W. & Zhang, H. Inhibition of protein carbamylation in urea solution using ammonium-containing buffers. *Anal Biochem* **446**, (2014).

75. Matthew, J. B., Morrow, J. S., Wittebort, R. J., And, $ & Gurd, F. R. N. *Quantitative Determination of Carbamino Adducts of a and p Chains in Human Adult Hemoglobin in Presence and Absence of Carbon Monoxide and 2,3-Diphosphoglycerate*. THE JOURNAL OF B~LOGICAL CHEMISTRY* vol. 252 (1977).

76. Knight, S., Andersson, I. & Brändén, C. I. Crystallographic analysis of ribulose 1,5-bisphosphate carboxylase from spinach at 2·4 Å resolution. Subunit interactions and active site. *J Mol Biol* (1990) doi:10.1016/S0022-2836(05)80100-7.

77. Cleland, W. W., Andrews, T. J., Gutteridge, S., Hartman, F. C. & Lorimer, G. H. *Mechanism of Rubisco: The Carbamate as General Base X*. https://pubs.acs.org/sharingguidelines (1998).

78. Morrow, J. S., Keim, P. & Gurd, F. R. N. CO2 adducts of certain amino acids, peptides, and sperm whale myoglobin studied by carbon 13 and proton nuclear magnetic resonance. *Journal of Biological Chemistry* **249**, 7484–7494 (1974).

79. Morollo, A. A., Petsko, G. A. & Ringe, D. Structure of a Michaelis complex analogue: Propionate binds in the substrate carboxylate site of alanine racemase. *Biochemistry* **38**, 3293–3301 (1999).

80. Park, I.-S. & Hausinger, R. P. Requirement of Carbon Dioxide for in Vitro Assembly of the Urease Nickel Metallocenter. *Science (1979)* **267**, 1156–1158 (1995).

81. Golemi, D. *et al.* Critical involvement of a carbamylated lysine in catalytic function of class D-lactamases. *PNAS* 14280–14285 (2001).

82. Kilmartin, J. V. & Rossi-Bernardi, L. Inhibition of CO2 combination and reduction of the bohr effect in haemoglobin chemically modified at its α-amino groups. *Nature* **222**, (1969).

83. Blake, L. I. & Cann, M. J. Carbon Dioxide and the Carbamate Post-Translational Modification. *Front Mol Biosci* **9**, (2022).

84. Linthwaite, V. L. *et al. Ubiquitin Is a Carbon Dioxide-Binding Protein*. *Sci. Adv* vol. 7 https://www.science.org (2021).

85. Isom, D. G., Castañeda, C. A., Cannon, B. R. & García-Moreno, B. Large shifts in pK a values of lysine residues buried inside a protein. doi:10.1073/pnas.1010750108/-/DCSupplemental.

86. Huckstepp, R. T. R., Eason, R., Sachdev, A. & Dale, N. CO2-dependent opening of connexin 26 and related β connexins. *Journal of Physiology* **588**, 3921–3931 (2010).

87. Linthwaite, V. L., Cummins, E. & Cann, M. J. Carbon dioxide detection in biological systems. *Interface Focus* vol. 11 Preprint at https://doi.org/10.1098/rsfs.2021.0001 (2021).

88. Cook, Z. C., Gray, M. A. & Cann, M. J. Elevated carbon dioxide blunts mammalian cAMP signaling dependent on inositol 1,4,5-triphosphate receptor-mediated Ca2+ release. *Journal of Biological Chemistry* **287**, 26291–26301 (2012).

89. Girish, V. & Vijayalakshmi, A. Affordable image analysis using NIH Image/ImageJ. *Indian J Cancer* **41**, 47 (2004).

90. Doll, S. & Burlingame, A. L. Mass Spectrometry-Based Detection and Assignment of Protein Posttranslational Modifications. *ACS Chem Biol* **10**, 63–71 (2015).

91. Rabbani, N., Ashour, A. & Thornalley, P. J. Mass spectrometric determination of early and advanced glycation in biology. *Glycoconj J* **33**, (2016).

92. Malte, H. & Lykkeboe, G. The Bohr/Haldane effect: a model-based uncovering of the full extent of its impact on O 2 delivery to and CO 2 removal from tissues. *J Appl Physiol* **125**, 916–922 (2018).

93. Christiansen, J., Douglas, C. G. & Haldane, J. S. The absorption and dissociation of carbon dioxide by human blood. *J Physiol* **48**, (1914).

94. Bartels, H. & Baumann, R. Respiratory function of hemoglobin. *International review of physiology* vol. 14 Preprint at https://doi.org/10.1056/nejm199801223380407 (1977).

95. Kilmartin, J. V. & Rossi-Bernardi, L. The binding of carbon dioxide by horse haemoglobin. *Biochem J* **124**, (1971).

96. Morrow, J. S., Matthew, J. B., Wittebort, R. J. & Gurd, F. R. N. Carbon 13 resonances of 13CO2 carbamino adducts of α and β chains in human adult hemoglobin. *Journal of Biological Chemistry* **251**, (1976).

97. Hill, E., Dale, N. & Wall, M. J. Co2-sensitive connexin hemichannels in neurons and glia: Three different modes of signalling? *International Journal of Molecular Sciences* vol. 22 Preprint at https://doi.org/10.3390/ijms22147254 (2021).

98. Brotherton, D. H. *et al.* Conformational changes and channel gating induced by CO2 binding to Connexin26. doi:10.1101/2020.08.11.243964.

99. Meigh, L. *et al.* CO2 directly modulates connexin 26 by formation of carbamate bridges between subunits. *Elife* (2013) doi:10.7554/elife.01213.

100. Maeda, S. *et al.* Structure of the connexin 26 gap junction channel at 3.5 Å resolution. *Nature* **458**, 597–602 (2009).

101. Nijjar, S. *et al.* Opposing modulation of Cx26 gap junctions and hemichannels by CO2. *bioRxiv* (2019) doi:10.1101/584722.

102. Komander, D. The emerging complexity of protein ubiquitination. *Biochemical Society Transactions* Preprint at https://doi.org/10.1042/BST0370937 (2009).

103. Giacco, F. & Brownlee, M. Oxidative stress and diabetic complications. *Circulation Research* vol. 107 Preprint at https://doi.org/10.1161/CIRCRESAHA.110.223545 (2010).

104. Ahmed, M. U., Thorpe, S. R. & Baynes, J. W. Identification of N(ε)-carboxymethyllysine as a degradation product of fructoselysine in glycated protein. *Journal of Biological Chemistry* **261**, (1986).

105. Krook, M., Ghosh, D., Strömberg, R., Carlquist, M. & Jörnvall, H. Carboxyethyllysine in a protein: Native carbonyl reductase/NADP+-dependent prostaglandin dehydrogenase. *Proc Natl Acad Sci U S A* **90**, (1993).

106. Baldensperger, T., Jost, T., Zipprich, A. & Glomb, M. A. Novel α-Oxoamide Advanced-Glycation Endproducts within the N6-Carboxymethyl Lysine and N6-Carboxyethyl Lysine Reaction Cascades. *J Agric Food Chem* **66**, (2018).

107. Glomb, M. A. & Monnier, V. M. Mechanism of protein modification by glyoxal and glycolaldehyde, reactive intermediates of the Maillard reaction. *Journal of Biological Chemistry* **270**, (1995).

108. Prasad, C., Davis, K. E., Imrhan, V., Juma, S. & Vijayagopal, P. Advanced Glycation End Products and Risks for Chronic Diseases: Intervening Through Lifestyle Modification. *American Journal of Lifestyle Medicine* vol. 13 Preprint at https://doi.org/10.1177/1559827617708991 (2019).

109. Gleave, M. The strengths of mass spectrometry are not just sensitivity and selectivity. *Bioanalysis* vol. 3 Preprint at https://doi.org/10.4155/bio.10.195 (2011).

110. Fitz, V., El Abiead, Y., Berger, D. & Koellensperger, G. Systematic Investigation of LC Miniaturization to Increase Sensitivity in Wide-Target LC-MS-Based Trace Bioanalysis of Small Molecules. *Front Mol Biosci* **9**, (2022).

111. Meek, J. L. & Rossetti, Z. L. Factors affecting retention and resolution of peptides in high-performance liquid chromatography. *J Chromatogr A* **211**, (1981).

112. Walther, T. C. & Mann, M. Mass spectrometry-based proteomics in cell biology. *Journal of Cell Biology* vol. 190 Preprint at https://doi.org/10.1083/jcb.201004052 (2010).

113. Ho, C. S. *et al.* Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin Biochem Rev* **24**, (2003).

114. Allen, D. R. & McWhinney, B. C. Quadrupole Time-of-Flight Mass Spectrometry: A Paradigm Shift in Toxicology Screening Applications. *Clinical Biochemist Reviews* **40**, (2019).

115. Davies, V. *et al.* Rapid Development of Improved Data-Dependent Acquisition Strategies. *Anal Chem* **93**, (2021).

116. Chu, I. K. *et al.* Proposed nomenclature for peptide ion fragmentation. *Int J Mass Spectrom* **390**, (2015).

117. Liu, R., Li, Q. & Smith, L. M. Detection of large ions in time-of-flight mass spectrometry: Effects of ion mass and acceleration voltage on microchannel plate detector response. *J Am Soc Mass Spectrom* **25**, (2014).

118. Xin, L. *et al.* A streamlined platform for analyzing tera-scale DDA and DIA mass spectrometry data enables highly sensitive immunopeptidomics. *Nat Commun* **13**, (2022).

119. Seymour, S. L. & Hunter, C. L. ProteinPilot ™ Software Overview. *Biomarkers and Omics* (2017).

120. Deutsch, E. W. File formats commonly used in mass spectrometry proteomics. *Molecular and Cellular Proteomics* vol. 11 Preprint at https://doi.org/10.1074/mcp.R112.019695 (2012).

121. Holman, J. D., Tabb, D. L. & Mallick, P. Employing ProteoWizard to convert raw mass spectrometry data. *Curr Protoc Bioinformatics* (2014) doi:10.1002/0471250953.bi1324s46.

122. Dams, M., Dores-Sousa, J. L., Lamers, R. J., Treumann, A. & Eeltink, S. High-Resolution Nano-Liquid Chromatography with Tandem Mass Spectrometric Detection for the Bottom-Up Analysis of Complex Proteomic Samples. *Chromatographia* vol. 82 Preprint at https://doi.org/10.1007/s10337-018-3647-5 (2019).

123. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol Biol* **604**, (2010).

124. Wang, P. & Wilson, S. R. Mass spectrometry-based protein identification by integrating de novo sequencing with database searching. *BMC Bioinformatics* **14 Suppl 2**, (2013).

125. Kapp, E. A. *et al.* An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *Proteomics* **5**, (2005).

126. Ma, B. & Johnson, R. De novo sequencing and homology searching. *Molecular and Cellular Proteomics* vol. 11 Preprint at https://doi.org/10.1074/mcp.O111.014902 (2012).

127. Zhang, J. *et al.* PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular and Cellular Proteomics* (2012) doi:10.1074/mcp.M111.010587.

128. Shilov, I. V. *et al.* The paragon algorithm, a next generation search engine that uses sequence temperature values sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Molecular and Cellular Proteomics* **6**, (2007).

129. Sapan, C. V, Lundblad, R. L. & Price, N. C. Colorimetric protein assay techniques. *Biotechnol Appl Biochem* **29 (Pt 2)**, (1999).

130. ThermoFisher Scientific. Protein quantitation assay compatibility table. *Assets.Thermofisher.Com* (2021).

131. Huang, T. Competitive Binding to Cuprous Ions of Protein and BCA in the Bicinchoninic Acid Protein Assay. *Open Biomed Eng J* **4**, (2010).

132. Fountoulakis, M., Juranville, J. F. & Manneberg, M. Comparison of the Coomassie brilliant blue, bicinchoninic acid and Lowry quantitation assays, using non-glycosylated and glycosylated proteins. *J Biochem Biophys Methods* **24**, (1992).

133. Lämmerhofer, M. & Lindner, W. Separation Methods in Drug Synthesis and Purification. *Handbook of Analytical Separations* **1**, (2000).

134. Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular and Cellular Proteomics* **11**, (2012).

135. Wang, H. *et al.* An off-line high pH reversed-phase fractionation and nano-liquid chromatography-mass spectrometry method for global proteomic profiling of cell lines. *J Chromatogr B Analyt Technol Biomed Life Sci* **974**, 90–95 (2015).

136. Gallia, J., Lavrich, K., Tan-Wilson, A. & Madden, P. H. Filtering of MS/MS data for peptide identification. *BMC Genomics* (2013) doi:10.1186/1471-2164-14-S7-S2.

137. Martin, D. B., Eng, J. K., Nesvizhskii, A. I., Gemmill, A. & Aebersold, R. Investigation of neutral loss during collision-induced dissociation of peptide ions. *Anal Chem* **77**, (2005).

138. Jörg, J., Houriet, R. & Spiteller, G. Massenspektren von Pflanzenschutzmitteln. *Monatsh Chem* **97**, (1966).

139. Jones, A. R., Siepen, J. A., Hubbard, S. J. & Paton, N. W. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics* **9**, 1220–1229 (2009).

140. Zhu, Z. J. *et al.* Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database. *Nat Protoc* **8**, (2013).

141. Carroll, J., Altman, M. C., Fearnley, I. M. & Walker, J. E. Identification of membrane proteins by tandem mass spectrometry of protein ions. *Proc Natl Acad Sci U S A* **104**, (2007).

142. Ren, J., Tian, Y., Hossain, E. & Connolly, M. D. Fragmentation Patterns and Mechanisms of Singly and Doubly Protonated Peptoids Studied by Collision Induced Dissociation. *J Am Soc Mass Spectrom* **27**, (2016).

143. Box, J. K. *et al.* Nucleophosmin: From structure and function to disease development. *BMC Mol Biol* **17**, (2016).

144. Cooks, T. *et al.* Mutant p53 Prolongs NF-κB Activation and Promotes Chronic Inflammation and Inflammation-Associated Colorectal Cancer. *Cancer Cell* **23**, (2013).

145. Zhu, B. S. *et al.* Blocking NF-κB nuclear translocation leads to p53-related autophagy activation and cell apoptosis. *World J Gastroenterol* **17**, (2011).

146. Yu, H., Rao, X. & Zhang, K. Nucleoside diphosphate kinase (Ndk): A pleiotropic effector manipulating bacterial virulence and adaptive responses. *Microbiological Research* vol. 205 Preprint at https://doi.org/10.1016/j.micres.2017.09.001 (2017).

147. Yu, H., Rao, X. & Zhang, K. Nucleoside diphosphate kinase (Ndk): A pleiotropic effector manipulating bacterial virulence and adaptive responses. *Microbiol Res* **205**, 125–134 (2017).

148. Kapoor, I., Emam, E. A. F., Shaw, A. & Varshney, U. Nucleoside Diphosphate Kinase Escalates A-to-C Mutations in MutT-Deficient Strains of Escherichia coli. *J Bacteriol* **202**, (2019).

149. Zhao, Y. & Jensen, O. N. Modification-specific proteomics: Strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* vol. 9 4632–4641 Preprint at https://doi.org/10.1002/pmic.200900398 (2009).

150. Sun, X. *et al.* PROTACs: great opportunities for academia and industry. *Signal Transduct Target Ther* **4**, 64 (2019).

151. Kramer, L. T. & Zhang, X. Expanding the landscape of E3 ligases for targeted protein degradation. *Current Research in Chemical Biology* **2**, 100020–100024 (2022).

152. Haas, A. L. & Siepmann, T. J. Pathways of ubiquitin conjugation. *The FASEB Journal* **11**, 1257–1268 (1997).

153. Ulrich, H. D. & Walden, H. Ubiquitin signalling in DNA replication and repair. *Nat Rev Mol Cell Biol* **11**, 479–489 (2010).

154. Haakonsen, D. L. & Rape, M. Branching Out: Improved Signaling by Heterotypic Ubiquitin Chains. *Trends Cell Biol* **29**, 704–716 (2019).

155. Tracz, M. & Bialek, W. Beyond K48 and K63: non-canonical protein ubiquitination. *Cell Mol Biol Lett* **26**, (2021).

156. Yau, R. G. *et al.* Assembly and Function of Heterotypic Ubiquitin Chains in Cell-Cycle and Protein Quality Control. *Cell* **171**, 918-933.e20 (2017).

157. Pickart, C. M. & Eddins, M. J. Ubiquitin: structures, functions, mechanisms. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1695**, 55–72 (2004).

158. Sluimer, J. & Distel, B. Regulating the human HECT E3 ligases. *Cellular and Molecular Life Sciences* vol. 75 3121–3141 Preprint at https://doi.org/10.1007/s00018-018-2848-2 (2018).

159. Lescouzères, L. & Bomont, P. E3 Ubiquitin Ligases in Neurological Diseases: Focus on Gigaxonin and Autophagy. *Front Physiol* **11**, (2020).

160. Dubiel, W., Dubiel, D., Wolf, D. A. & Naumann, M. Cullin 3-Based Ubiquitin Ligases as Master Regulators of Mammalian Cell Differentiation. *Trends in Biochemical Sciences* vol. 43 95–107 Preprint at https://doi.org/10.1016/j.tibs.2017.11.010 (2018).

161. Cai, W. & Yang, H. The structure and regulation of Cullin 2 based E3 ubiquitin ligases and their biological functions. *Cell Division* vol. 11 Preprint at https://doi.org/10.1186/s13008-016-0020-7 (2016).

162. Fuchs, S. Y. *et al. The SCF HOS/-TRCP-ROC1 E3 Ubiquitin Ligase Utilizes Two Distinct Domains within CUL1 for Substrate Targeting and Ubiquitin Ligation*. *MOLECULAR AND CELLULAR BIOLOGY* vol. 20 (2000).

163. Xie, J., Jin, Y. & Wang, G. The role of SCF ubiquitin-ligase complex at the beginning of life. *Reproductive Biology and Endocrinology* vol. 17 Preprint at https://doi.org/10.1186/s12958-019-0547-y (2019).

164. Girardini, M., Maniaci, C., Hughes, S. J., Testa, A. & Ciulli, A. Cereblon versus VHL: Hijacking E3 ligases against each other using PROTACs. *Bioorg Med Chem* **27**, 2466–2479 (2019).

165. Buetow, L. & Huang, D. T. Structural insights into the catalysis and regulation of E3 ubiquitin ligases. *Nat Rev Mol Cell Biol* **17**, 626–642 (2016).

166. Williams, K. M. *et al.* Structural insights into E1 recognition and the ubiquitin-conjugating activity of the E2 enzyme Cdc34. *Nat Commun* **10**, (2019).

167. Spratt, D. E., Wu, K., Kovacev, J., Pan, Z.-Q. & Shaw, G. S. Selective Recruitment of an E2∼Ubiquitin Complex by an E3 Ubiquitin Ligase. *Journal of Biological Chemistry* **287**, 17374–17385 (2012).

168. Wu, K., Kovacev, J. & Pan, Z.-Q. Priming and Extending: A UbcH5/Cdc34 E2 Handoff Mechanism for Polyubiquitination on a SCF Substrate. *Mol Cell* **37**, 784–796 (2010).

169. Miller, F., Kentsis, A., Osman, R. & Pan, Z. Q. Inactivation of VHL by tumorigenic mutations that disrupt dynamic coupling of the pVHL·hypoxia-inducible transcription factor-1α complex. *Journal of Biological Chemistry* **280**, 7985–7996 (2005).

170. Duda, D. M. *et al. Structural Insights into NEDD8 Activation of Cullin-RING Ligases: Conformational Control of Conjugation*. (2008).

171. Li, J. M. & Jin, J. CRL ubiquitin ligases and DNA damage response. *Frontiers in Oncology* vol. 2 Preprint at https://doi.org/10.3389/fonc.2012.00029 (2012).

172. Shi, Q. & Chen, L. Cereblon: A Protein Crucial to the Multiple Functions of Immunomodulatory Drugs as well as Cell Metabolism and Disease Generation. *J Immunol Res* **2017**, 1–8 (2017).

173. Liu, J. *et al.* CRL4ACRBN E3 ubiquitin ligase restricts BK channel activity and prevents epileptogenesis. *Nat Commun* **5**, 3924 (2014).

174. Shen, C. *et al.* The E3 ubiquitin ligase component, Cereblon, is an evolutionarily conserved regulator of Wnt signaling. *Nat Commun* **12**, (2021).

175. Weidemann, A. & Johnson, R. S. Biology of HIF-1α. *Cell Death Differ* **15**, 621–627 (2008).

176. Cardote, T. A. F., Gadd, M. S. & Ciulli, A. Crystal Structure of the Cul2-Rbx1-EloBC-VHL Ubiquitin Ligase Complex. *Structure* **25**, 901-911.e3 (2017).

177. Konstantinidou, M. *et al.* PROTACs– a game-changing technology. *Expert Opin Drug Discov* **14**, 1255–1268 (2019).

178. Cecchini, C., Pannilunghi, S., Tardy, S. & Scapozza, L. From Conception to Development: Investigating PROTACs Features for Improved Cell Permeability and Successful Protein Degradation. *Front Chem* **9**, (2021).

179. Neklesa, T. K. *et al.* An oral androgen receptor PROTAC degrader for prostate cancer. *Journal of Clinical Oncology* **35**, 273–273 (2017).

180. Fouad, S., Wells, O. S., Hill, M. A. & D'Angiolella, V. Cullin Ring Ubiquitin Ligases (CRLs) in Cancer: Responses to Ionizing Radiation (IR) Treatment. *Frontiers in Physiology* vol. 10 Preprint at https://doi.org/10.3389/fphys.2019.01144 (2019).

181. Kargbo, R. B. PROTAC Molecules for the Treatment of Autoimmune Disorders. *ACS Med Chem Lett* **10**, 276–277 (2019).

182. Zhou, Q.-Q. *et al.* Advancing targeted protein degradation for metabolic diseases therapy. *Pharmacol Res* **188**, 106627 (2023).

183. Steinebach, C. *et al.* Systematic exploration of different E3 ubiquitin ligases: an approach towards potent and selective CDK6 degraders. *Chem Sci* **11**, 3474–3486 (2020).

184. Inuzuka, H., Liu, J., Wei, W. & Rezaeian, A.-H. PROTAC technology for the treatment of Alzheimer's disease: advances and perspectives. *Acta Materia Medica* **1**, (2022).

185. Ishida, T. & Ciulli, A. E3 Ligase Ligands for PROTACs: How They Were Found and How to Discover New Ones. *SLAS Discovery* **26**, 484–502 (2021).

186. Devaiah, B. N., Gegonne, A. & Singer, D. S. Bromodomain 4: a cellular Swiss army knife. *J Leukoc Biol* **100**, 679–686 (2016).

187. Devaiah, B. N. *et al.* MYC protein stability is negatively regulated by BRD4. *Proceedings of the National Academy of Sciences* **117**, 13457–13467 (2020).

188. Andrades, A. *et al.* SWI/SNF complexes in hematological malignancies: biological implications and therapeutic opportunities. *Mol Cancer* **22**, 39 (2023).

189. Guerrero-Martínez, J. A. & Reyes, J. C. High expression of SMARCA4 or SMARCA2 is frequently associated with an opposite prognosis in cancer. *Sci Rep* **8**, 2043 (2018).

190. Hoffman, G. R. *et al.* Functional epigenetics approach identifies BRM/SMARCA2 as a critical synthetic lethal target in BRG1-deficient cancers. *Proceedings of the National Academy of Sciences* **111**, 3128–3133 (2014).

191. Schwinn, M. K. *et al.* CRISPR-Mediated Tagging of Endogenous Proteins with a Luminescent Peptide. *ACS Chem Biol* **13**, 467–474 (2018).

192. Dixon, A. S. *et al.* NanoLuc Complementation Reporter Optimized for Accurate Measurement of Protein Interactions in Cells. *ACS Chem Biol* **11**, 400–408 (2016).

193. Iversen, P. W., Eastwood, B. J., Sittampalam, G. S. & Cox, K. L. A comparison of assay performance measures in screening assays: Signal window, Z' factor, and assay variability ratio. *J Biomol Screen* **11**, 247–252 (2006).

194. Prinz, H. Hill coefficients, dose–response curves and allosteric mechanisms. *J Chem Biol* **3**, 37–44 (2010).

195. Davies, G., Vincent, J., Packer, M. J. & Murray, D. Grouping concentration response curves by features of their shape to aid rapid and consistent analysis of large data sets in high throughput screens. *SLAS Discovery* **27**, 272–277 (2022).

196. Liu, Z. *et al.* An overview of PROTACs: a promising drug discovery paradigm. *Molecular Biomedicine* **3**, 46 (2022).

197. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).

198. Phelan, D. E., Mota, C., Lai, C., Kierans, S. J. & Cummins, E. P. Carbon dioxide-dependent signal transduction in mammalian systems. *Interface Focus* **11**, 20200033 (2021).

199. Cao, K. *et al.* DOT1L-controlled cell-fate determination and transcription elongation are independent of H3K79 methylation. *Proceedings of the National Academy of Sciences* **117**, 27365–27373 (2020).

200. Fyodorov, D. V., Zhou, B.-R., Skoultchi, A. I. & Bai, Y. Emerging roles of linker histones in regulating chromatin structure and function. *Nat Rev Mol Cell Biol* **19**, 192–206 (2018).

201. Tsunaka, Y. Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle. *Nucleic Acids Res* **33**, 3424–3434 (2005).

202. Henikoff, S. & Smith, M. M. Histone Variants and Epigenetics. *Cold Spring Harb Perspect Biol* **7**, a019364 (2015).

203. Martire, S. & Banaszynski, L. A. The roles of histone variants in fine-tuning chromatin organization and function. *Nat Rev Mol Cell Biol* **21**, 522–541 (2020).

204. Kurumizaka, H., Kujirai, T. & Takizawa, Y. Contributions of Histone Variants in Nucleosome Structure and Function. *J Mol Biol* **433**, 166678 (2021).

205. Talbert, P. B. & Henikoff, S. Histone variants at a glance. *J Cell Sci* **134**, (2021).

206. Gillette, T. G. & Hill, J. A. Readers, Writers, and Erasers: Chromatin as the Whiteboard of Heart Disease. *Circ Res* **116**, 1245–1253 (2015).

207. Rothbart, S. B. & Strahl, B. D. Interpreting the language of histone and DNA modifications. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1839**, 627–643 (2014).

208. Millán-Zambrano, G., Burton, A., Bannister, A. J. & Schneider, R. Histone post-translational modifications — cause and consequence of genome function. *Nat Rev Genet* **23**, 563–580 (2022).

209. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res* **21**, 381–395 (2011).

210. Lee, D., Yang, J. & Kim, S. Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer. *Nat Commun* **13**, 6678 (2022).

211. Millán-Zambrano, G., Burton, A., Bannister, A. J. & Schneider, R. Histone post-translational modifications — cause and consequence of genome function. *Nat Rev Genet* **23**, 563–580 (2022).

212. Xia, C., Tao, Y., Li, M., Che, T. & Qu, J. Protein acetylation and deacetylation: An important regulatory modification in gene transcription (Review). *Exp Ther Med* (2020) doi:10.3892/etm.2020.9073.

213. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet* **13**, 343–357 (2012).

214. Rossetto, D., Avvakumov, N. & Côté, J. Histone phosphorylation. *Epigenetics* **7**, 1098–1108 (2012).

215. Meas, R. & Mao, P. Histone ubiquitylation and its roles in transcription and DNA damage response. *DNA Repair (Amst)* **36**, 36–42 (2015).

216. Chan, J. C. & Maze, I. Nothing Is Yet Set in (Hi)stone: Novel Post-Translational Modifications Regulating Chromatin Function. *Trends Biochem Sci* **45**, 829–844 (2020).

217. Lo, W.-S. *et al.* Phosphorylation of Serine 10 in Histone H3 Is Functionally Linked In Vitro and In Vivo to Gcn5-Mediated Acetylation at Lysine 14. *Mol Cell* **5**, 917–926 (2000).

218. Worden, E. J., Zhang, X. & Wolberger, C. Structural basis for COMPASS recognition of an H2B-ubiquitinated nucleosome. *Elife* **9**, (2020).

219. Worden, E. J., Hoffmann, N. A., Hicks, C. W. & Wolberger, C. Mechanism of Cross-talk between H2B Ubiquitination and H3 Methylation by Dot1L. *Cell* **176**, 1490-1501.e12 (2019).

220. Lee, J.-S., Smith, E. & Shilatifard, A. The Language of Histone Crosstalk. *Cell* **142**, 682–685 (2010).

221. Cho, Y.-W. *et al.* PTIP Associates with MLL3- and MLL4-containing Histone H3 Lysine 4 Methyltransferase Complex. *Journal of Biological Chemistry* **282**, 20395–20406 (2007).

222. Suganuma, T. & Workman, J. L. Crosstalk among Histone Modifications. *Cell* **135**, 604–607 (2008).

223. Zhou, P., Wu, E., Alam, H. B. & Li, Y. Histone Cleavage as a Mechanism for Epigenetic Regulation: Current Insights and Perspectives. *Curr Mol Med* **14**, 1164–1172 (2014).

224. Duncan, E. M. *et al.* Cathepsin L Proteolytically Processes Histone H3 During Mouse Embryonic Stem Cell Differentiation. *Cell* **135**, 284–294 (2008).

225. Xu, Y.-M., Du, J.-Y. & Lau, A. T. Y. Posttranslational modifications of human histone H3: An update. *Proteomics* **14**, 2047–2060 (2014).

226. Clément, C. *et al.* High-resolution visualization of H3 variants during replication reveals their controlled recycling. *Nat Commun* **9**, 3181 (2018).

227. Vlaming, H. & van Leeuwen, F. The upstreams and downstreams of H3K79 methylation by DOT1L. *Chromosoma* **125**, 593–605 (2016).

228. Ruthenburg, A. J., Allis, C. D. & Wysocka, J. Methylation of Lysine 4 on Histone H3: Intricacy of Writing and Reading a Single Epigenetic Mark. *Mol Cell* **25**, 15–30 (2007).

229. Duan, Z. *et al.* Role of Dot1L and H3K79 methylation in regulating somatic hypermutation of immunoglobulin genes. *Proceedings of the National Academy of Sciences* **118**, (2021).

230. Frederiks, F. *et al.* Nonprocessive methylation by Dot1 leads to functional redundancy of histone H3K79 methylation states. *Nat Struct Mol Biol* **15**, 550–557 (2008).

231. Basavapathruni, A. *et al.* Conformational Adaptation Drives Potent, Selective and Durable Inhibition of the Human Protein Methyltransferase DOT1L. *Chem Biol Drug Des* **80**, 971–980 (2012).

232. Valencia-Sánchez, M. I. *et al.* Structural Basis of Dot1L Stimulation by Histone H2B Lysine 120 Ubiquitination. *Mol Cell* **74**, 1010-1019.e6 (2019).

233. Lee, S. *et al.* Dot1 regulates nucleosome dynamics by its inherent histone chaperone activity in yeast. *Nat Commun* **9**, 240 (2018).

234. Altaf, M. *et al.* Interplay of Chromatin Modifiers on a Short Basic Patch of Histone H4 Tail Defines the Boundary of Telomeric Heterochromatin. *Mol Cell* **28**, 1002–1014 (2007).

235. Deshpande, A. J. *et al.* AF10 Regulates Progressive H3K79 Methylation and HOX Gene Expression in Diverse AML Subtypes. *Cancer Cell* **26**, 896–908 (2014).

236. He, N. *et al.* Human Polymerase-Associated Factor complex (PAFc) connects the Super Elongation Complex (SEC) to RNA polymerase II on chromatin. *Proceedings of the National Academy of Sciences* **108**, (2011).

237. Wakeman, T. P., Wang, Q., Feng, J. & Wang, X.-F. Bat3 facilitates H3K79 dimethylation by DOT1L and promotes DNA damage-induced 53BP1 foci at G1/G2 cell-cycle phases. *EMBO J* **31**, 2169–2181 (2012).

238. Nguyen, A. T. & Zhang, Y. The diverse functions of Dot1 and H3K79 methylation. *Genes Dev* **25**, 1345–1358 (2011).

239. Singer, M. S. *et al.* Identification of High-Copy Disruptors of Telomeric Silencing in Saccharomyces cerevisiae. *Genetics* **150**, 613–632 (1998).

240. Farooq, Z., Banday, S., Pandita, T. K. & Altaf, M. The many faces of histone H3K79 methylation. *Mutation Research/Reviews in Mutation Research* **768**, 46–52 (2016).

241. Jones, B. *et al.* The Histone H3K79 Methyltransferase Dot1L Is Essential for Mammalian Development and Heterochromatin Structure. *PLoS Genet* **4**, e1000190 (2008).

242. van Leeuwen, F., Gafken, P. R. & Gottschling, D. E. Dot1p Modulates Silencing in Yeast by Methylation of the Nucleosome Core. *Cell* **109**, 745–756 (2002).

243. Fu, H. *et al.* Methylation of Histone H3 on Lysine 79 Associates with a Group of Replication Origins and Helps Limit DNA Replication Once per Cell Cycle. *PLoS Genet* **9**, e1003542 (2013).

244. Feng, Y. *et al.* Early mammalian erythropoiesis requires the Dot1L methyltransferase. *Blood* **116**, 4483–4491 (2010).

245. Barry, E. R. *et al.* ES Cell Cycle Progression and Differentiation Require the Action of the Histone Methyltransferase Dot1L. *Stem Cells* **27**, 1538–1547 (2009).

246. Zhou, H., Madden, B. J., Muddiman, D. C. & Zhang, Z. Chromatin Assembly Factor 1 Interacts with Histone H3 Methylated at Lysine 79 in the Processes of Epigenetic Silencing and DNA Repair. *Biochemistry* **45**, 2852–2861 (2006).

247. Yi, D. *et al.* Dot1l interacts with Zc3h10 to activate Ucp1 and other thermogenic genes. *Elife* **9**, (2020).

248. Mohan, M. *et al.* Linking H3K79 trimethylation to Wnt signaling through a novel Dot1-containing complex (DotCom). *Genes Dev* **24**, (2010).

249. Gibbons, G. S., Owens, S. R., Fearon, E. R. & Nikolovska-Coleska, Z. Regulation of Wnt signaling target gene expression by the histone methyltransferase DOT1L. *ACS Chem Biol* **10**, (2015).

250. Alexandrova, E. *et al.* Histone Methyltransferase DOT1L as a Promising Epigenetic Target for Treatment of Solid Tumors. *Front Genet* **13**, (2022).

251. Bernt, K. M. *et al.* MLL-Rearranged Leukemia Is Dependent on Aberrant H3K79 Methylation by DOT1L. *Cancer Cell* **20**, 66–78 (2011).

252. Milne, T. A., Martin, M. E., Brock, H. W., Slany, R. K. & Hess, J. L. Leukemogenic MLL Fusion Proteins Bind across a Broad Region of the Hoxa9 Locus, Promoting Transcription and Multiple Histone Modifications. *Cancer Res* **65**, 11367–11374 (2005).

253. Kang, J. *et al.* KDM2B is a histone H3K79 demethylase and induces transcriptional repression *via* sirtuin-1-mediated chromatin silencing. *The FASEB Journal* **32**, 5737–5750 (2018).

254. Daigle, S. R. *et al.* Selective Killing of Mixed Lineage Leukemia Cells by a Potent Small-Molecule DOT1L Inhibitor. *Cancer Cell* **20**, 53–65 (2011).

255. Daigle, S. R. *et al.* Potent inhibition of DOT1L as treatment of MLL-fusion leukemia. *Blood* **122**, 1017–1025 (2013).

256. Shukla, N. *et al.* Final Report of Phase 1 Study of the DOT1L Inhibitor, Pinometostat (EPZ-5676), in Children with Relapsed or Refractory MLL-r Acute Leukemia. *Blood* **128**, 2780–2780 (2016).

257. Keogh, C. E. *et al.* Carbon dioxide-dependent regulation of NF-κB family members RelB and p100 gives molecular insight into CO2-dependent immune regulation. *Journal of Biological Chemistry* **292**, 11561–11571 (2017).

258. Shigemura, M. *et al.* Elevated CO2 regulates the Wnt signaling pathway in mammals, Drosophila melanogaster and Caenorhabditis elegans. *Sci Rep* **9**, 18251 (2019).

259. Phelan, D. E. *et al.* Hypercapnia alters mitochondrial gene expression and acylcarnitine production in monocytes. *Immunol Cell Biol* **101**, (2023).

260. Meert, P., Govaert, E., Scheerlinck, E., Dhaenens, M. & Deforce, D. Pitfalls in histone propionylation during bottom-up mass spectrometry analysis. *Proteomics* **15**, 2966–2971 (2015).

261. Fields, G. B. Methods for Removing the Fmoc Group. in *Peptide Synthesis Protocols* 17–28 (Humana Press, New Jersey). doi:10.1385/0-89603-273-6:17.

262. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Front Microbiol* **5**, (2014).

263. Carnes, A. E., Hodgson, C. P. & Williams, J. A. Inducible Escherichia coli fermentation for increased plasmid DNA production. *Biotechnol Appl Biochem* **45**, 155 (2006).

264. Jeong, H., Kim, H. J. & Lee, S. J. Complete Genome Sequence of Escherichia coli Strain BL21. *Genome Announc* **3**, (2015).

265. Carter, A. P. *et al.* Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature* **407**, 340–348 (2000).

266. Daber, R., Stayrook, S., Rosenberg, A. & Lewis, M. Structural Analysis of Lac Repressor Bound to Allosteric Effectors. *J Mol Biol* **370**, 609–619 (2007).

267. Arora, S., Saxena, V. & Ayyar, B. V. Affinity chromatography: A versatile technique for antibody purification. *Methods* **116**, 84–94 (2017).

268. Luger, K., Rechsteiner, T. J. & Richmond, T. J. Preparation of nucleosome core particle from recombinant histones. in 3–19 (1999). doi:10.1016/S0076-6879(99)04003-3.

269. Dyer, P. N. *et al.* Reconstitution of Nucleosome Core Particles from Recombinant Histones and DNA. in 23–44 (2003). doi:10.1016/S0076-6879(03)75002-2.

270. Anderson, M. *et al.* Co-expression as a convenient method for the production and purification of core histones in bacteria. *Protein Expr Purif* **72**, 194–204 (2010).

271. Shim, Y., Duan, M.-R., Chen, X., Smerdon, M. J. & Min, J.-H. Polycistronic coexpression and nondenaturing purification of histone octamers. *Anal Biochem* **427**, 190–192 (2012).

272. Tegel, H., Tourle, S., Ottosson, J. & Persson, A. Increased levels of recombinant human proteins with the Escherichia coli strain Rosetta(DE3). *Protein Expr Purif* **69**, 159–167 (2010).

273. Chu, I.-T., Speer, S. L. & Pielak, G. J. Rheostatic Control of Protein Expression Using Tuner Cells. *Biochemistry* **59**, 733–735 (2020).

274. Klinker, H., Haas, C., Harrer, N., Becker, P. B. & Mueller-Planitz, F. Rapid Purification of Recombinant Histones. *PLoS One* **9**, e104029 (2014).

275. Mitchell, M. J., Jensen, O. E., Cliffe, K. A. & Maroto-Valer, M. M. A model of carbon dioxide dissolution and mineral carbonation kinetics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **466**, 1265–1290 (2010).

276. Mosmann, T. Rapid colorimetric assay for cellular growth and survival: Application to proliferation and cytotoxicity assays. *J Immunol Methods* **65**, 55–63 (1983).

277. Kurani, H. *et al.* DOT1L Is a Novel Cancer Stem Cell Target for Triple-Negative Breast Cancer. *Clinical Cancer Research* **28**, 1948–1965 (2022).

278. Richter, W. F., Shah, R. N. & Ruthenburg, A. J. Non-canonical H3K79me2-dependent pathways promote the survival of MLL-rearranged leukemia. *Elife* **10**, (2021).

279. Lonetti, A. *et al.* Inhibition of Methyltransferase DOT1L Sensitizes to Sorafenib Treatment AML Cells Irrespective of MLL-Rearrangements: A Novel Therapeutic Strategy for Pediatric AML. *Cancers (Basel)* **12**, 1972 (2020).

280. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (1979)* **347**, (2015).

281. Andrews, S. *et al.* FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. *Babraham Institute* (2012).

282. Fuller, C. W. *et al.* The challenges of sequencing by synthesis. *Nat Biotechnol* **27**, 1013–1023 (2009).

283. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**, e131–e131 (2010).

284. Krueger, F. Trim Galore: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type. *Babraham Bioinformatics* (2012).

285. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, (2013).

286. Martin, F. J. *et al.* Ensembl 2023. *Nucleic Acids Res* **51**, (2023).

287. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, (2015).

288. Zhao, S., Xi, L. & Zhang, B. Union exon based approach for RNA-seq gene quantification: To be or not to be? *PLoS One* **10**, (2015).

289. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, (2014).

290. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, (2009).

291. Chen, Y. *et al. EdgeR: Differential Analysis of Sequence Read Count Data User's Guide*. (2008).

292. Finotello, F. *et al.* A strategy to reduce technical variability and bias in RNA sequencing data. *EMBnet J* **18**, (2012).

293. Gene, T. & Consortium, O. Gene Ontology : tool for the unification of biology. *Gene Expr* **25**, (2000).

294. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

295. Sarno, F., Nebbioso, A. & Altucci, L. DOT1L: a key target in normal chromatin remodelling and in mixed-lineage leukaemia treatment. *Epigenetics* **15**, 439–453 (2020).

296. Marcos-Villar, L. & Nieto, A. The DOT1L inhibitor Pinometostat decreases the host-response against infections: Considerations about its use in human therapy. *Sci Rep* **9**, 16862 (2019).

297. Seath, C. P. *et al.* Tracking chromatin state changes using nanoscale photo-proximity labelling. *Nature* **616**, 574–580 (2023).

298. Yuan, L., Li, P., Zheng, Q., Wang, H. & Xiao, H. The Ubiquitin-Proteasome System in Apoptosis and Apoptotic Cell Clearance. *Front Cell Dev Biol* **10**, (2022).

299. Rogge, R. A. *et al.* Assembly of Nucleosomal Arrays from Recombinant Core Histones and Nucleosome Positioning DNA. *Journal of Visualized Experiments* (2013) doi:10.3791/50354.

300. Downes, D. J. *et al.* High-resolution targeted 3C interrogation of cis-regulatory element organization at genome-wide scale. *Nat Commun* **12**, 531 (2021).

301. Hetherington, A. M. & Raven, J. A. *The Biology of Carbon Dioxide*. *Current Biology* vol. 15 (2005).

302. Hampe, E. M. & Rudkevich, D. M. Exploring reversible reactions between CO2 and amines. *Tetrahedron* **59**, (2003).

303. Linthwaite, V. L. & Cann, M. J. A methodology for carbamate post-translational modification discovery and its application in Escherichia coli. *Interface Focus* **11**, (2021).

304. Riching, K. M., Mahan, S. D., Urh, M. & Daniels, D. L. High-throughput cellular profiling of targeted protein degradation compounds using hibit crispr cell lines. *Journal of Visualized Experiments* **2020**, (2020).

305. Swygert, S. G. & Peterson, C. L. Chromatin dynamics: Interplay between remodeling enzymes and histone modifications. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* vol. 1839 Preprint at https://doi.org/10.1016/j.bbagrm.2014.02.013 (2014).

306. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* vol. 293 Preprint at https://doi.org/10.1126/science.1063127 (2001).

307. Shahid, Z., Simpson, B., Miao, K. H. & Singh, G. Genetics, Histone Code. *StatPearls* (2023).

308. Cappadocia, L. & Lima, C. D. Ubiquitin-like Protein Conjugation: Structures, Chemistry, and Mechanism. *Chem Rev* **118**, 889–918 (2018).

8. Supplementary Information

8.1 Supplementary Data for Chapter 3

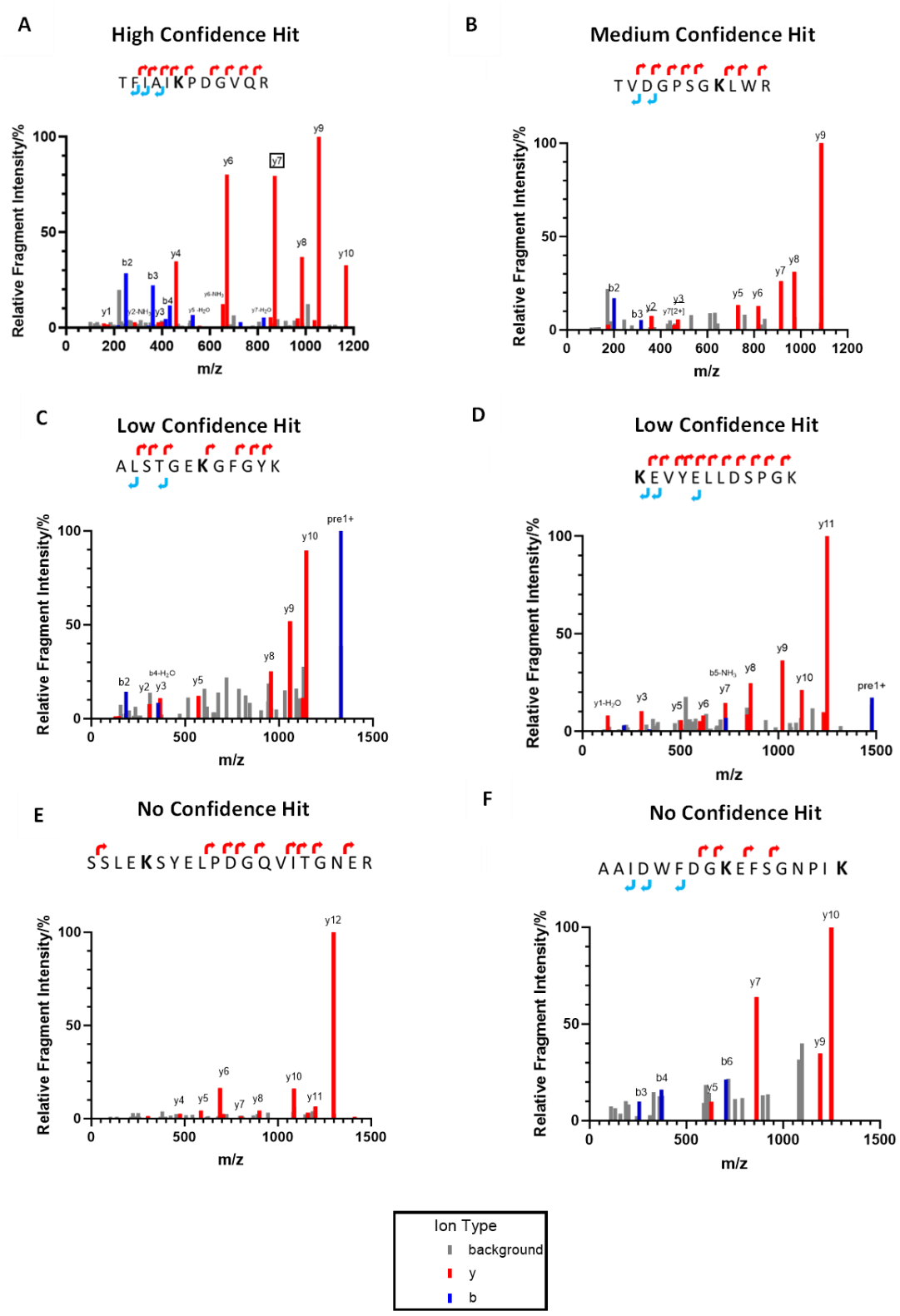
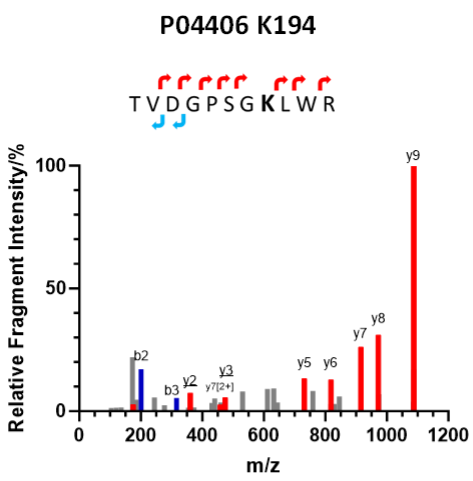
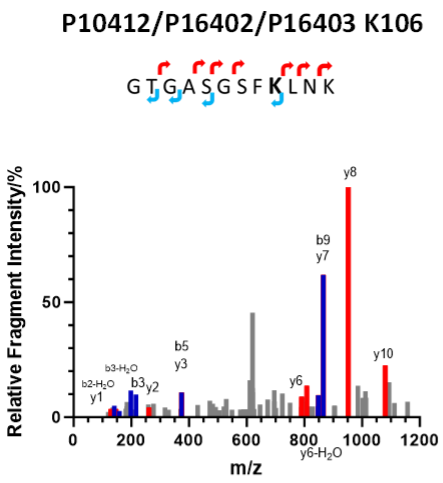


Figure 8-1 Example spectra of validating the carboxyethyl modification using the four confidence levels given in Table 3-2. Plots of relative fragment intensity versus m/z from LCMSMS identifying potential trapped carbamates in the presence of $^{12}CO_2$. Where (A) is a high confidence hit, (B) is a medium confidence hit due to no y mod but fulfilling other high criteria, (C) is a low confidence hit due to exhibiting sporadic y ion coverage and no y or b ion mod, (D) is a low confidence hit because the modified lysine is at the N terminus and there is not enough supporting information after the modification, (E) is a no-confidence hit because there are no surrounding supporting ions, and (F) is a no-confidence hit because the peptide does not terminate with a trypsin cut site. Each spectrum is associated with a peptide sequence illustrating the identification of predominant y (red) and b (blue) ions. The grey peaks represent background ions, and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

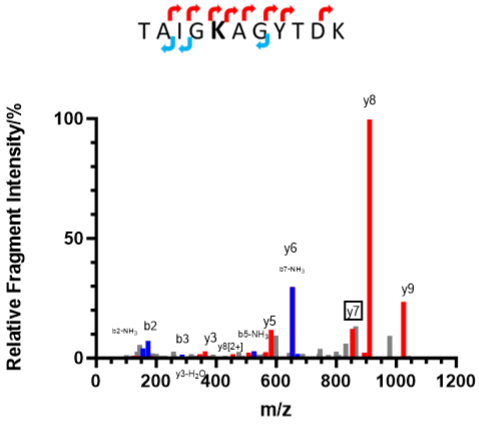
A

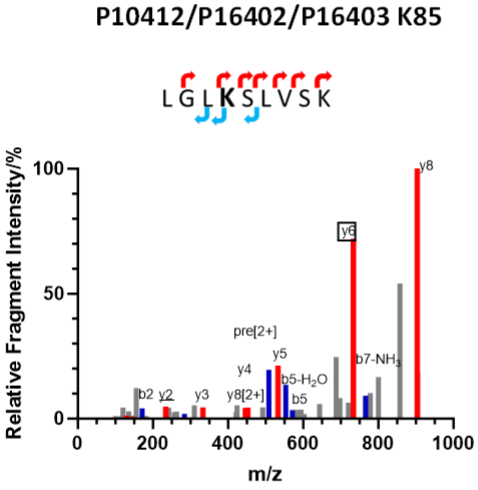

P04406 K194
TVDGPSGKLWR
Relative Fragment Intensity/%
m/z
b2
b3
y2
y3
y7[2+]
y5
y6
y7
y8
y9

B

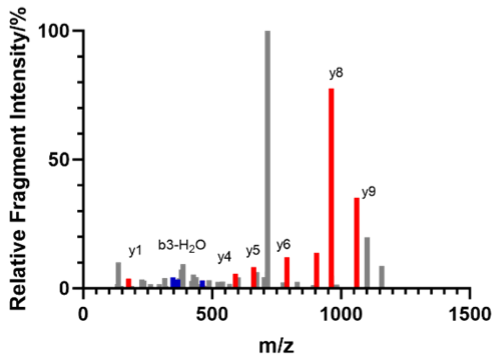

P10412/P16402/P16403 K106
GTGASGSFKLNK
Relative Fragment Intensity/%
m/z
b2-H2O
b3-H2O
y1
b3
y2
b5
y3
y6
b9
y7
y8
y10
y6-H2O

C

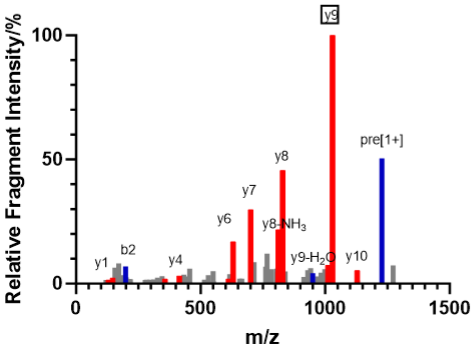

P06733 K233
TAIGKAGYTDK
Relative Fragment Intensity/%
m/z
b2-NH3
b2
b3
y3
y8[2+]
b5-NH3
y5
y6
b7-NH3
y7
y8
y9


P10412/P16402/P16403 K85
LGLKSLVSK
Relative Fragment Intensity/%
m/z
y6
pre[2+]
y4
y5
b5-H2O
b7-NH3
b2
y2
y3
y8[2+]
b5
y8

E


P63261 K61
DSYVGDEAQSKR
Relative Fragment Intensity/%
m/z
y1
b3-H2O
y4
y5
y6
y8
y9

F


P04406 K263
VVKQASEGPLK
Relative Fragment Intensity/%
m/z
y3
pre[1+]
y8
y7
y6
y8-NH3
y1
b2
y4
y9-H2O
y10

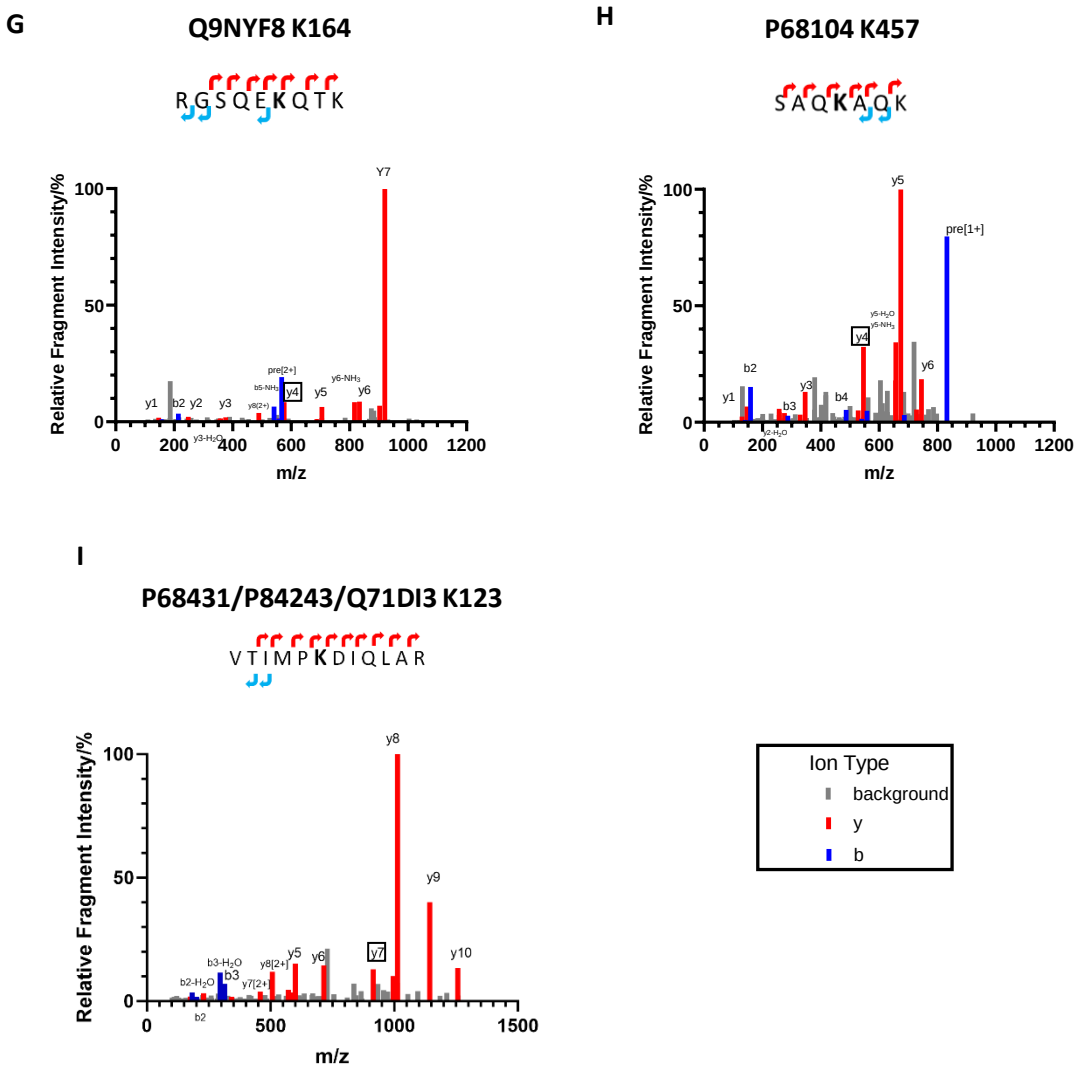


Figure 8-2 Identification of carbamate hits from the 12C HEK293 lysate screening that were identified multiple times by one of the database search algorithms but were only found once or were not identified in the same sample across both search algorithms and are listed in Table 3-4. Plots of relative fragment intensity versus m/z from LCMSMS identifying trapped carbamates on (A) P04406 K194, (B) P10412/P16402/P16403 K106, (C) P06733 K233, (D) P10412/P16402/P16403 K85, (E) P63261 K61, (F) P04406 K263, (G) Q9NYF8 K164, (H) P68104 K457, and (I)P68431/P84243/Q71DI3 K123 in the presence of $^{12}CO_2$. Each spectrum is associated with a peptide sequence illustrating the identification of predominant y (red) and b (blue) ions. The grey peaks represent background ions, and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

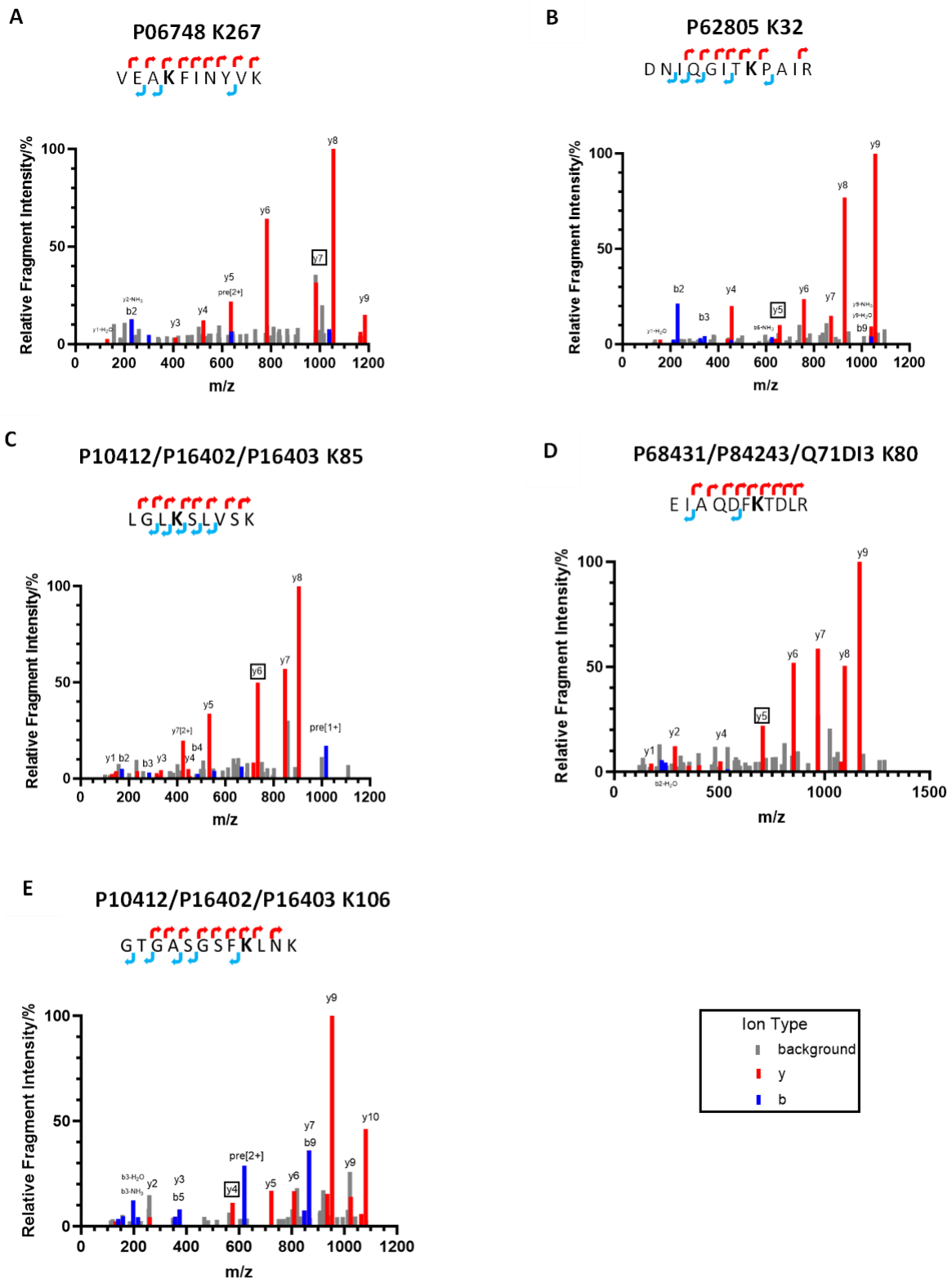


Figure 8-3 Identification of carbamate hits from the 13C HEK293 lysate screening that were identified in the 12C dataset by both database search algorithms with shared sample ID and are listed in Table 3-6. Plots of relative fragment intensity versus m/z from LCMSMS identifying trapped carbamates on (A)

P06748 K267, (B) P62805 K32 (C) P10412/P16402/P16403 K85, (D) P68431/P84243/Q71DI3 K80, (E) P10412/P16402/P16403 K106 in the presence of 13CO2. Each spectrum is associated with a peptide sequence illustrating the identification of predominant y (red) and b (blue) ions. The grey peaks represent background ions, and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

8.2 Supplementary Data for Chapter 4

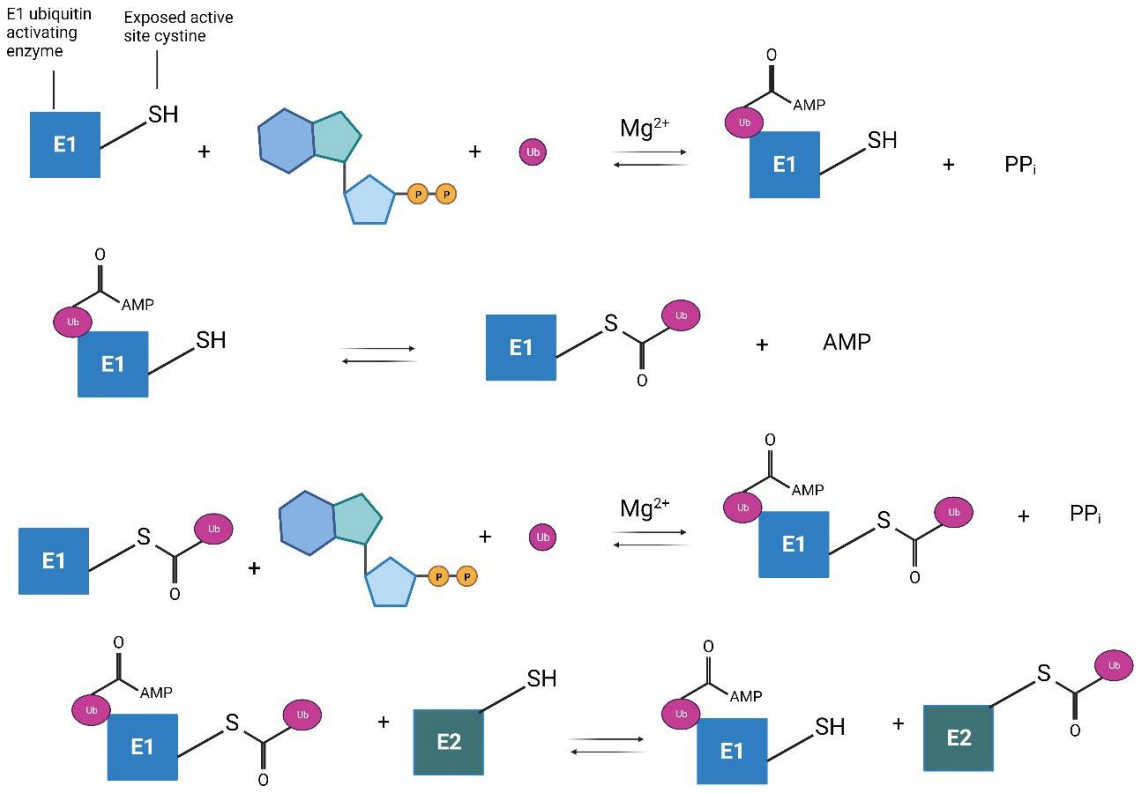


Figure 8-4 ATP hydrolysis mediated transfer of ubiquitin onto E1 and subsequently E2 adapted from reference. [308]

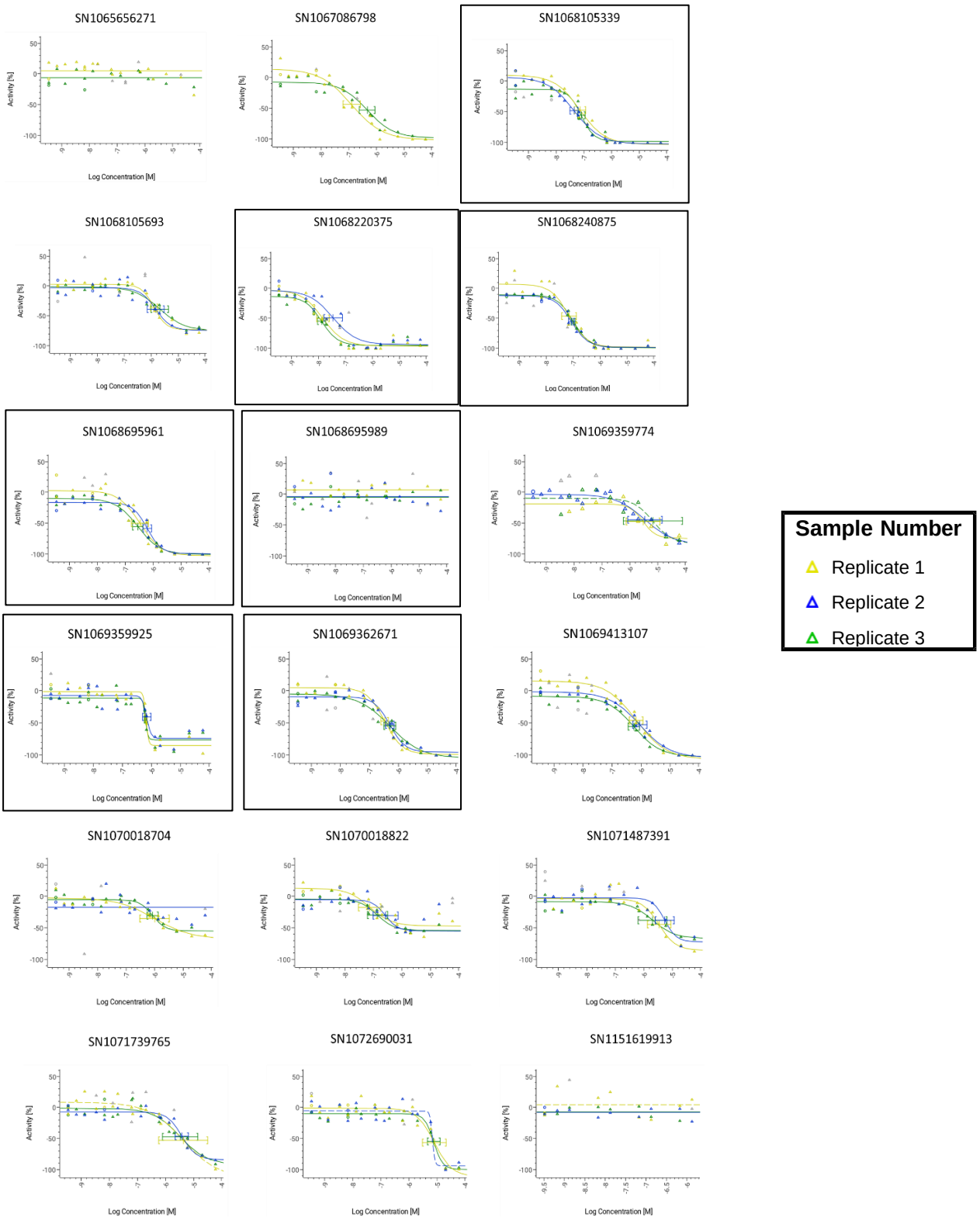


Figure 8-5 Dose-response curves (DRC) of BRD4 targeting tool PROTAC compounds where n=3. The normalized activity (%) derived from the luminescence readout signal is plotted against the specified log concentration. All data points are represented as the mean where n=3 and error bars are plotted as the standard deviation from the DRC fit. Percentage activity is normalised to the baseline

luminescence using 0 µM compound and 100% activity represents total degradation. Compounds which were selected for the CO_2 assay are marked by a border.

Compound ID	Curve Category replicate 1	$LogIC_{50}$ replicate 1	SE $LogIC_{50}$ replicate 1	Curve Category replicate 2	$LogIC_{50}$ replicate 2	SE $LogIC_{50}$ replicate 2	Curve Category replicate 3	$LogIC_{50}$ replicate 3	SE $LogIC_{50}$ replicate 3
SN1065656271	N-Inactive	-	-	N-Inactive	-	-	Unavailable	NA	NA
SN1067086798	Cat1 curve	-6.333	0.134	CI > 0.3	-6.901	0.150	Unavailable	NA	NA
SN1068105339	Cat2 curve	-7.098	0.065	Cat1 curve	-7.143	0.093	Cat1 curve	-7.354	0.070
SN1068105693	CI > 0.3	-5.688	0.150	Cat2 curve	-5.900	0.066	CI > 0.3	-5.836	0.147
SN1068220375	Cat1 curve	-7.889	0.069	Cat1 curve	-7.846	0.113	CI > 0.3	-7.463	0.163
SN1068240875	Cat1 curve	-6.991	0.037	Cat1 curve	-7.148	0.127	Cat1 curve	-7.061	0.050
SN1068695961	Cat1 curve	-6.551	0.087	CI > 0.3	-6.474	0.143	Cat2 curve	-6.215	0.076
SN1068695989	N-Inactive	-	-	N-Inactive	-	-	N-Inactive	-	-
SN1069359774	CI > 0.3	-5.171	0.436	CI > 0.3	-5.529	0.235	CI > 0.3	-5.455	0.285
SN1069359925	dS < 75-Partial	-6.202	0.024	nHill>4.0 Extremely steep	-6.183	0.018	dS < 75-Partial	-6.162	0.068
SN1069362671	Cat1 curve	-6.296	0.097	Cat1 curve	-6.475	0.060	Cat2 curve	-6.300	0.073
SN1069413107	Cat1 curve	-6.249	0.090	Cat1 curve	-6.166	0.112	Cat1 curve	-6.046	0.111
SN1070018704	dS < 75-Partial	-6.073	0.089	CI > 0.3	-5.996	0.243	N-Inactive		
SN1070018822	CI > 0.3	-6.865	0.166	CI > 0.3	-7.129	0.184	CI > 0.3	-6.583	0.204
SN1071487391	CI > 0.3	-5.694	0.238	CI > 0.3	-5.467	0.189	CI > 0.3	-5.268	0.154
SN1071739765	CI > 0.3	-5.480	0.293	CI > 0.3	-5.375	0.413	Cat2 curve	-5.426	0.106
SN1072690031	nHill>2.5< 4.0 Very steep	-5.121	0.102	CI > 0.3	-5.098	0.200	CI > 0.3	-5.160	35.125
SN1151619913	N-Inactive	-	-	N-Inactive	-	-	N-Inactive	-	-

Table 8-1 Key data used for the selection of PROTACs from the dose-response curves for each BRD4 tool compound tested. Compounds in bold were selected for the CO_2 assay.

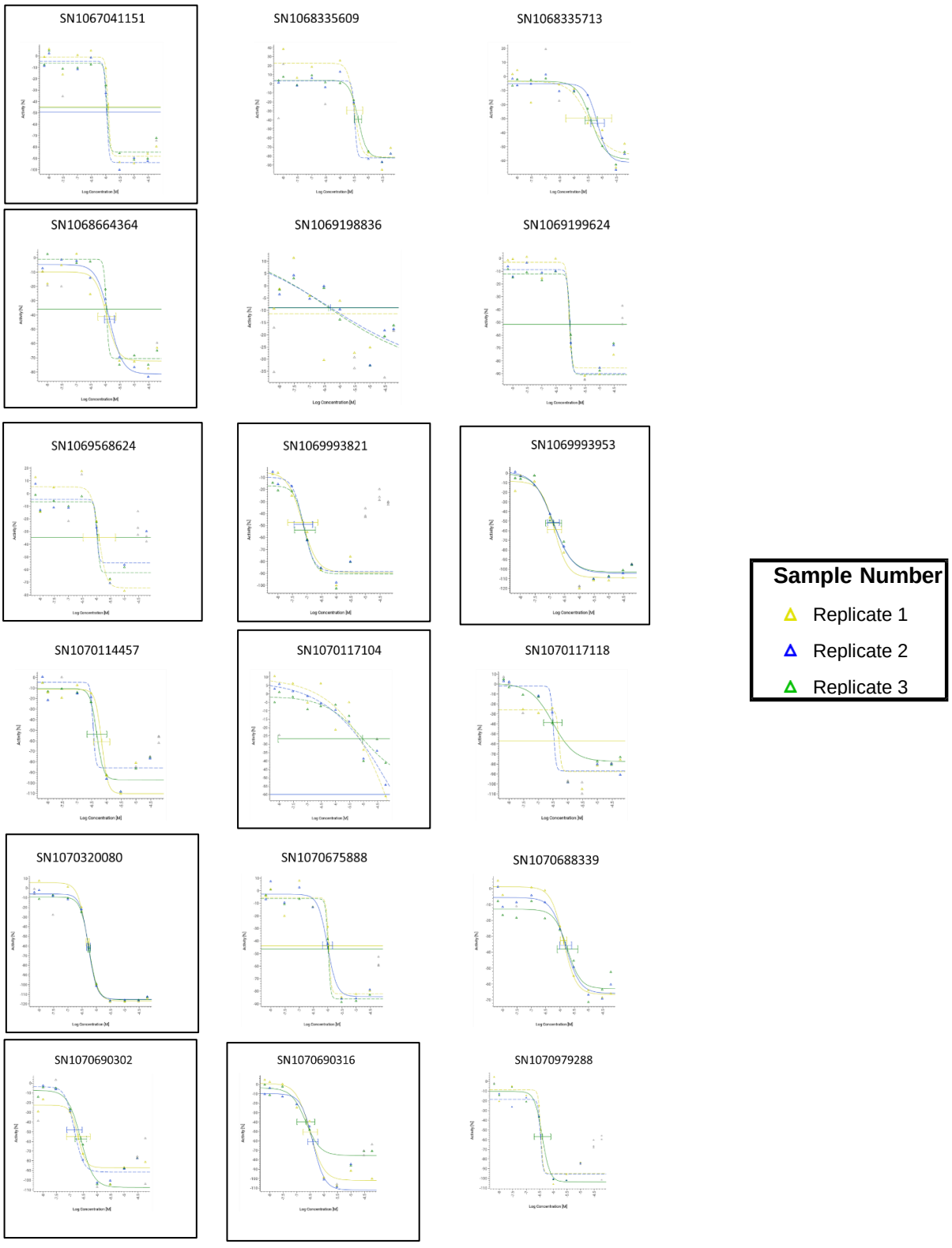


Figure 8-6 Dose-response curves (DRC) of SMARCA2 targeting tool PROTAC compounds where n=3. The normalized activity (%) derived from the luminescence readout signal is plotted against the specified log concentration. All data points are represented as the mean where n=3 and error bars are plotted as the standard deviation from the DRC fit. Percentage activity is normalised to the baseline

luminescence using 0 µM compound and 100% activity represents total degradation. Compounds which were selected for the CO_2 assay are marked by a border.

Compound ID	Curve Category replicate1	$LogIC_{50}$ replicate 1	SE $LogIC_{50}$ replicate 1	Curve Category replicate 2	$LogIC_{50}$ replicate 2	SE $LogIC_{50}$ replicate 2	Curve Category replicate 3	$LogIC_{50}$ replicate 3	SE $LogIC_{50}$ replicate 3
SN1067041151	nHill>2.5< 4.0 Very steep	-5.838	0.021	CI > 0.3	-5.867	0.108	CI > 0.3	-5.962	131.500
SN1068335609	Cat2 curve	-5.317	0.050	CI > 0.3	-5.409	0.144	CI > 0.3	-5.480	20.869
SN1068335713	dS < 75-Partial	-5.270	0.082	Weakly Active	-3.878	7.355	CI > 0.3	-5.044	0.140
SN1068664364	CI > 0.3	-5.961	5.932	CI > 0.3	-5.867	0.100	Cat2 curve	-5.855	0.062
SN1069198836	CI > 0.3	-5.796	0.579	N-Inactive			CI > 0.3	-5.506	0.939
SN1069199624	CI > 0.3	-6.018	8.503	CI > 0.3	-6.049	3.160	CI > 0.3	-6.037	15.829
SN1069568624	CI > 0.3	-5.959	13.954	CI > 0.3	-5.928	0.076	CI > 0.3	-5.962	16.788
SN1069993821	CI > 0.3	-6.999	0.091	Cat1 curve	-7.047	0.050	CI > 0.3	-7.089	0.082
SN1069993953	Cat1 curve	-6.869	0.108	Cat2 curve	-6.843	0.101	Cat1 curve	-6.881	0.078
SN1070114457	nHill>2.5< 4.0 Very steep	-6.241	0.044	nHill>4.0 Extremely steep	-6.163	0.089	CI > 0.3	-6.271	0.095
SN1070117104	CI > 0.3	-5.036	1.229	Weakly Active	-3.583	8.925	Weakly Active	-3.907	4.229
SN1070117118	Cat2 curve	-6.434	0.078	CI > 0.3	-6.254	0.245	CI > 0.3	-6.336	0.073
SN1070320080	nHill>2.5< 4.0 Very steep	-6.254	0.019	nHill>2.5< 4.0 Very steep	-6.301	0.018	nHill>2.5< 4.0 Very steep	-6.272	0.027
SN1070675888	CI > 0.3	-5.983	5.029	CI > 0.3	-5.963	14.002	Cat2 curve	-5.991	0.085
SN1070688339	CI > 0.3	-5.374	0.466	dS < 75-Partial	-5.871	0.045	dS < 75-Partial	-5.747	0.086
SN1070690302	Cat2 curve	-6.600	0.068	nHill>2.5< 4.0 Very steep	-6.587	0.063	Cat2 curve	-6.772	0.020
SN1070690316	Cat2 curve	-6.461	0.082	CI > 0.3	-6.407	0.155	Cat2 curve	-6.401	0.063
SN1070979288	CI > 0.3	-6.402	0.098	CI > 0.3	-6.442	27.351	CI > 0.3	-6.401	0.205

Table 8-2 Key data used for the selection of PROTACs from the dose-response curves for each SMARCA2 tool compound tested. Compounds in bold were selected for the CO_2 assay.

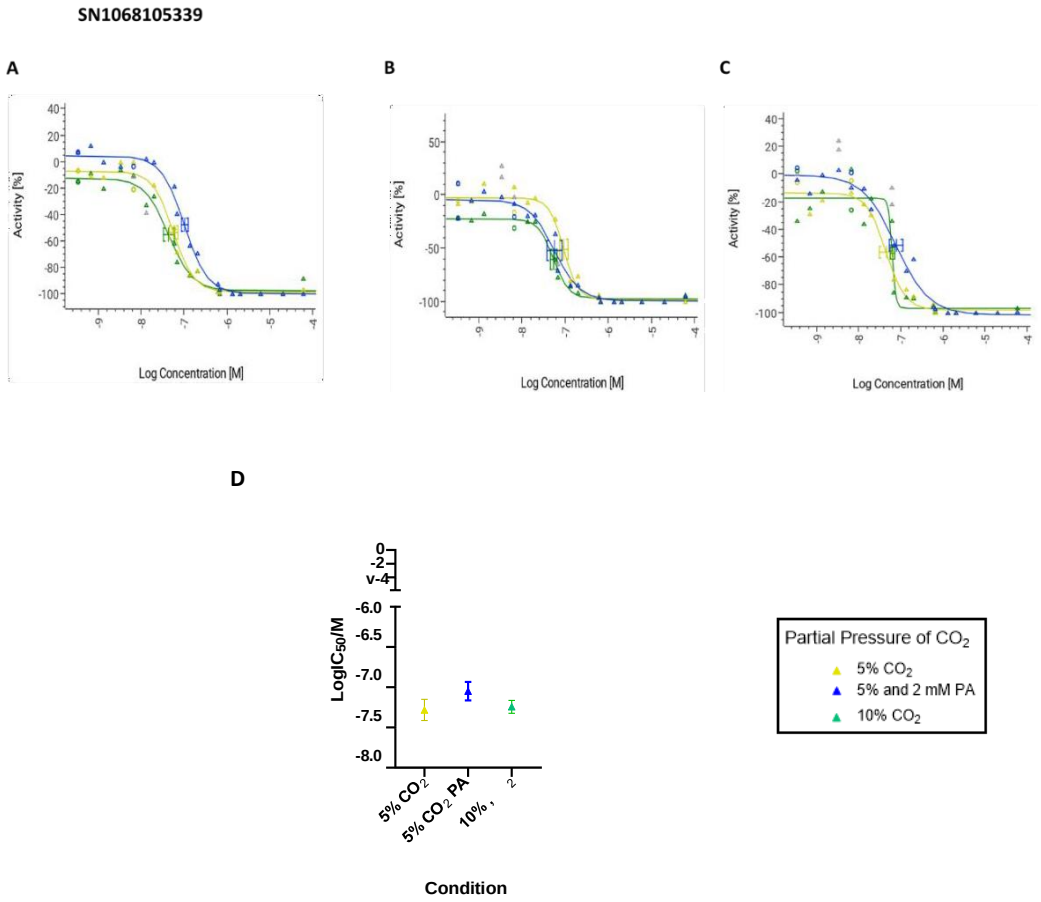


Figure 8-7 Dose-response data for SN1068105339 (A-C) Dose Response Curve (DRCs) for each condition where the normalized activity (%) derived from the luminescence readout signal is plotted against the specified log concentration. (D) logIC_{50} calculated from the dose-response curve fit plotted against the treatment condition. The colour for each condition across A-D is shown in the graph legend. All data points are represented as the mean where n=3 and error bars are plotted as the standard deviation from the DRC fit for A-C and as the absolute standard deviation calculated from the DRC fit and replicate IC_{50} values for D. One-Way ANOVA and multiple comparison test assessed significance at the threshold $p<0.05$.

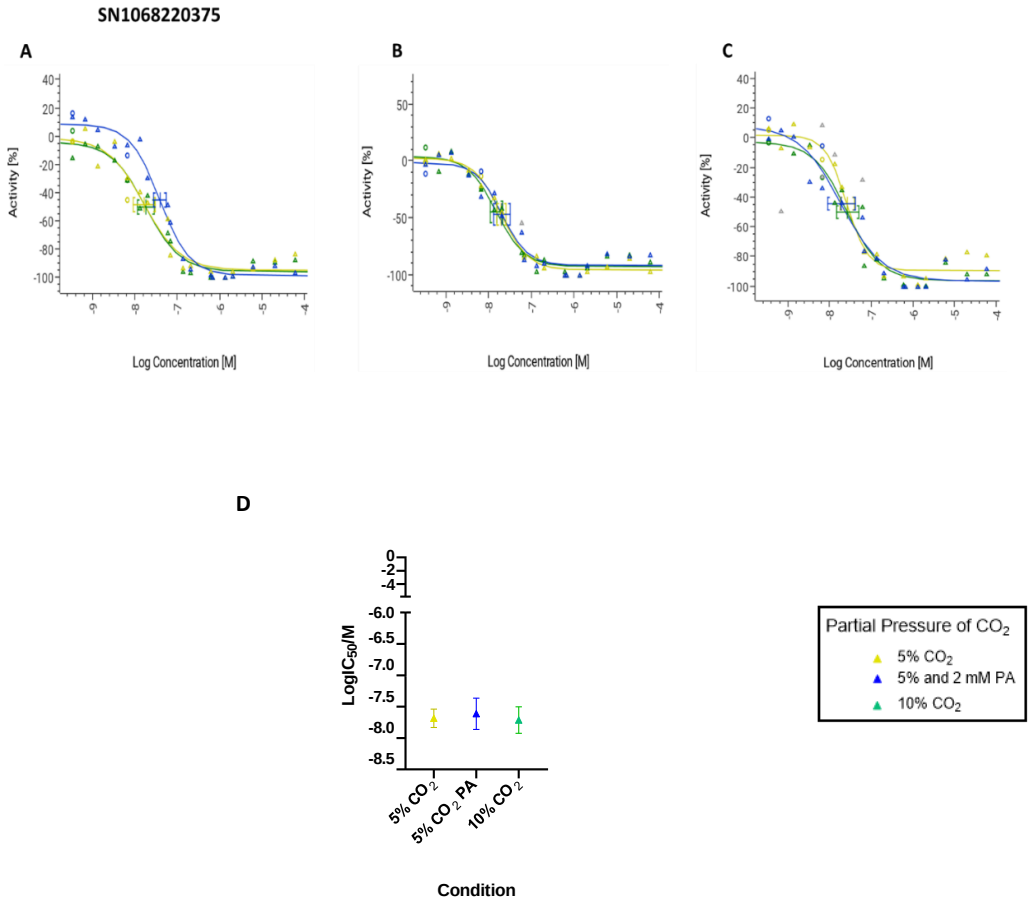


Figure 8-8 Dose-response data for SN1068220375, where A-D are detailed in Figure 8-7.

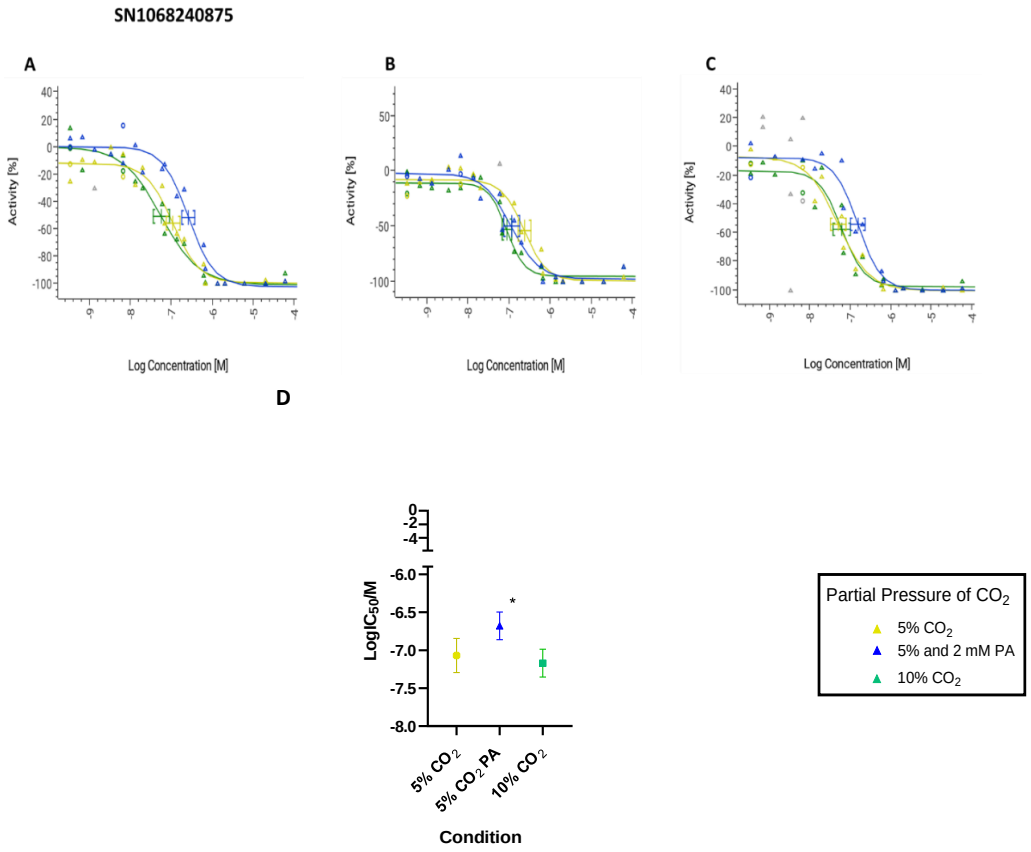


Figure 8-9 Dose-response data for SN1068240875, where A-D are detailed in Figure 8-7. Multiple comparison tests showed the $logIC_{50}$ values between 5% and 5% PA were statistically significant where * is $p<0.05$.

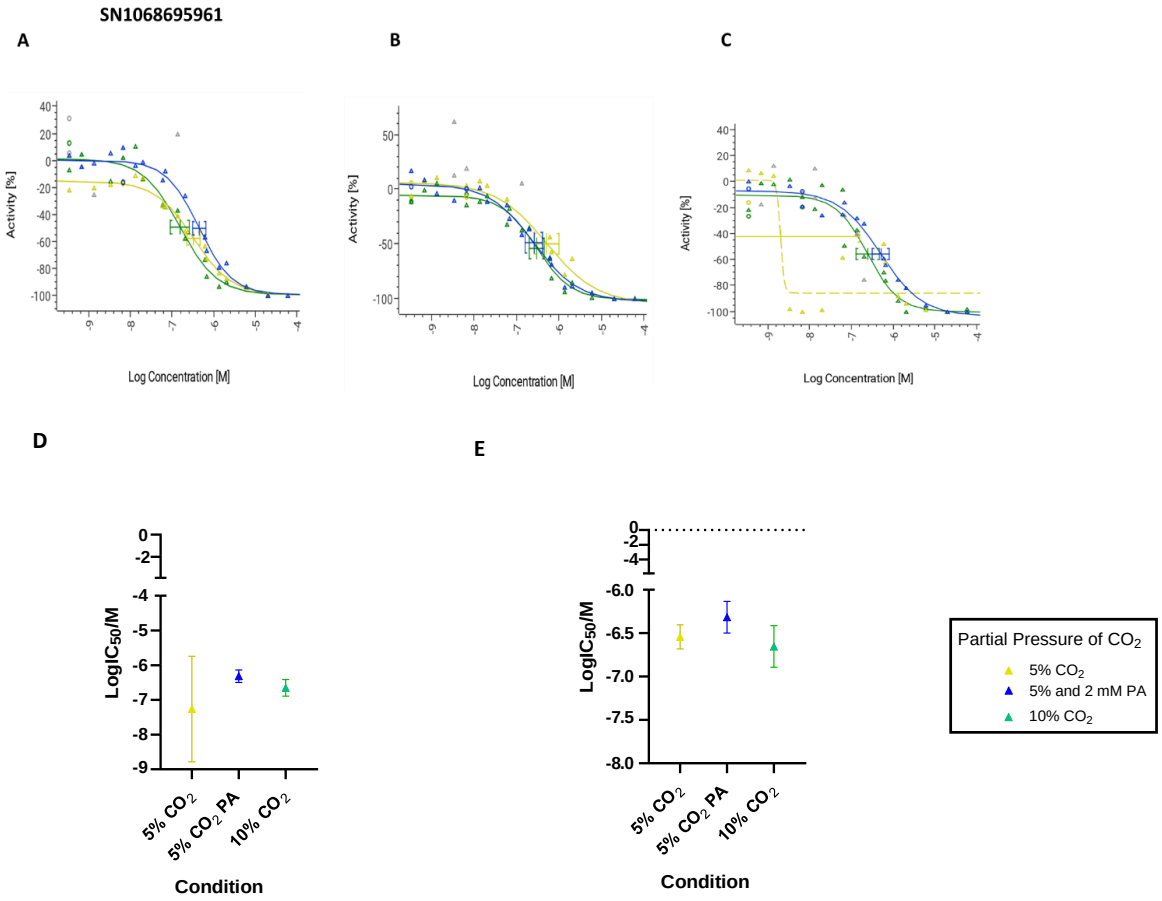


Figure 8-10 Dose-response data for SN1068695961, where A-D are detailed in Figure 8-7. (E) exclusion of outlier values where the corrected logIC_{50} is plotted against the treatment condition.

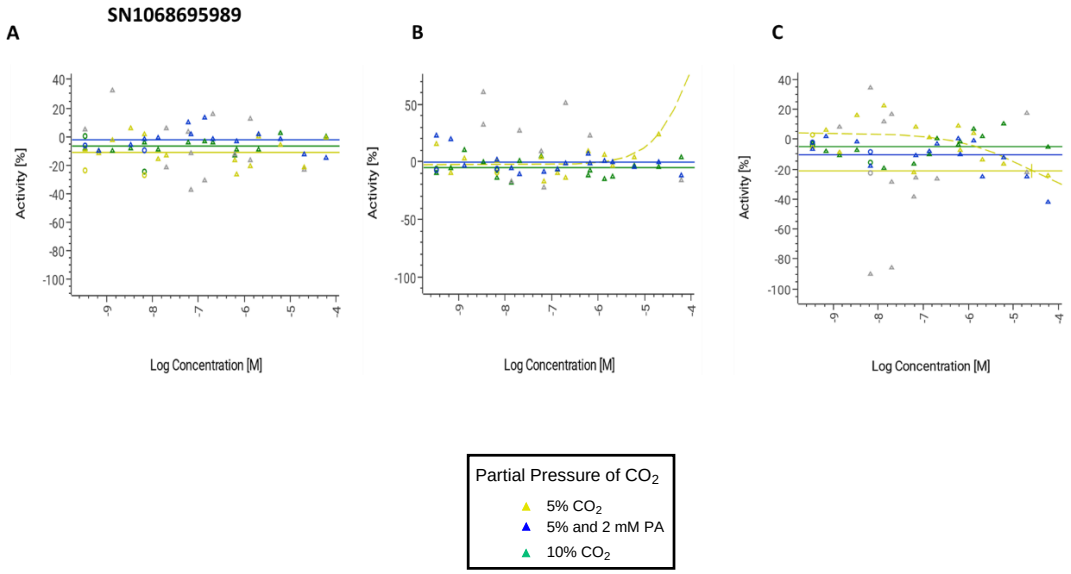


Figure 8-11 Dose-response data for SN1068695989, where A-C are detailed in Figure 8-7. There is no Figure D as the logIC_{50} is not calculated due to inactivity.

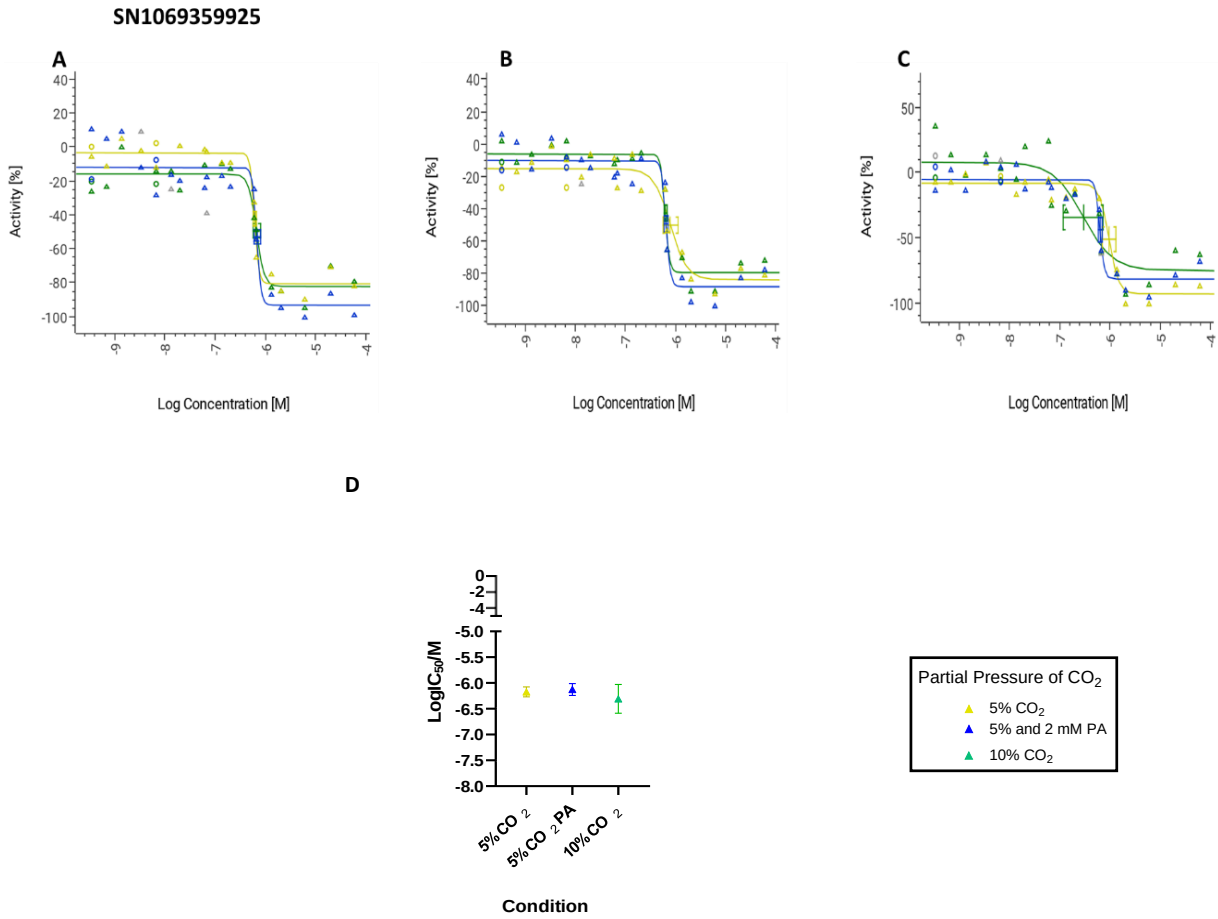


Figure 8-12 Dose-response data for SN1069359925, where A-D are detailed in Figure 8-7.

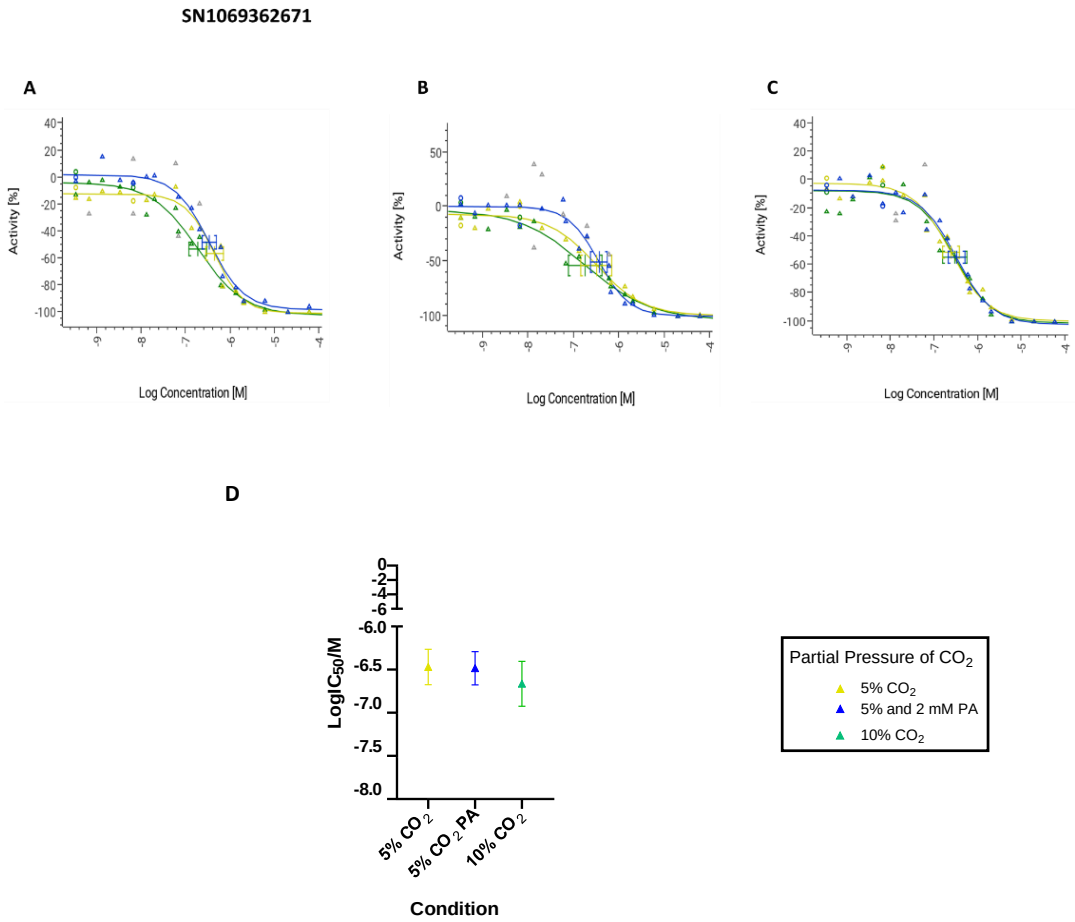


Figure 8-13 Dose-response data for SN1069362671, where A-D are detailed in Figure 8.7.

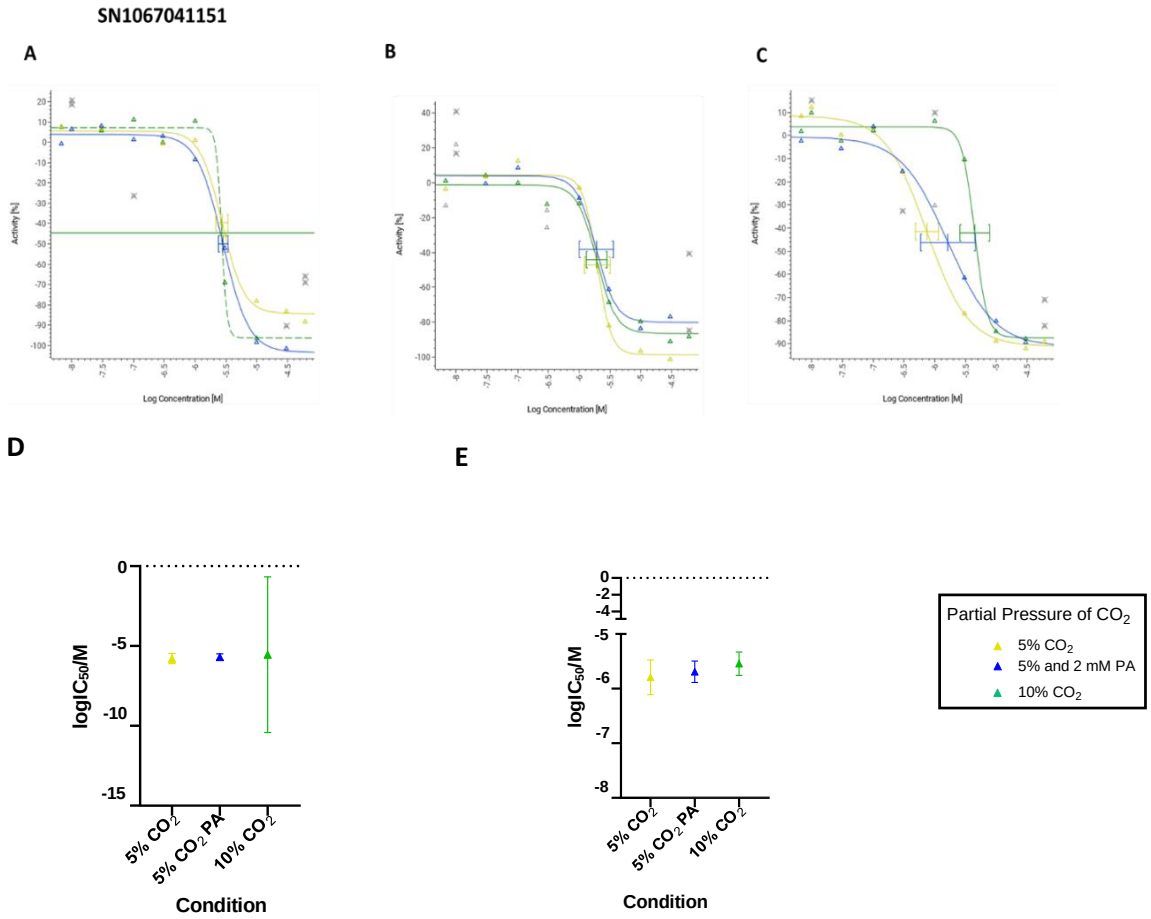


Figure 8-14 Dose-response data for SN10671151 (A-C) Dose Response Curve (DRCs) for each condition where the normalized activity (%) derived from the luminescence readout signal is plotted against the specified log concentration. (D) $logIC_{50}$ calculated from the dose-response curve fit against the treatment condition. (E) Exclusion of outlier standard deviation values where the corrected $logIC_{50}$ is plotted against the treatment condition. The colour for each condition across A-D is shown in the graph legend. All data points are represented as the mean where n=3 and error bars are plotted as the standard deviation from the DRC fit for A-C and as the absolute standard deviation from the DRC fit and replicate IC_{50} values for D and E. One-Way ANOVA and multiple comparison test assessed significance at the threshold $p<0.05$.

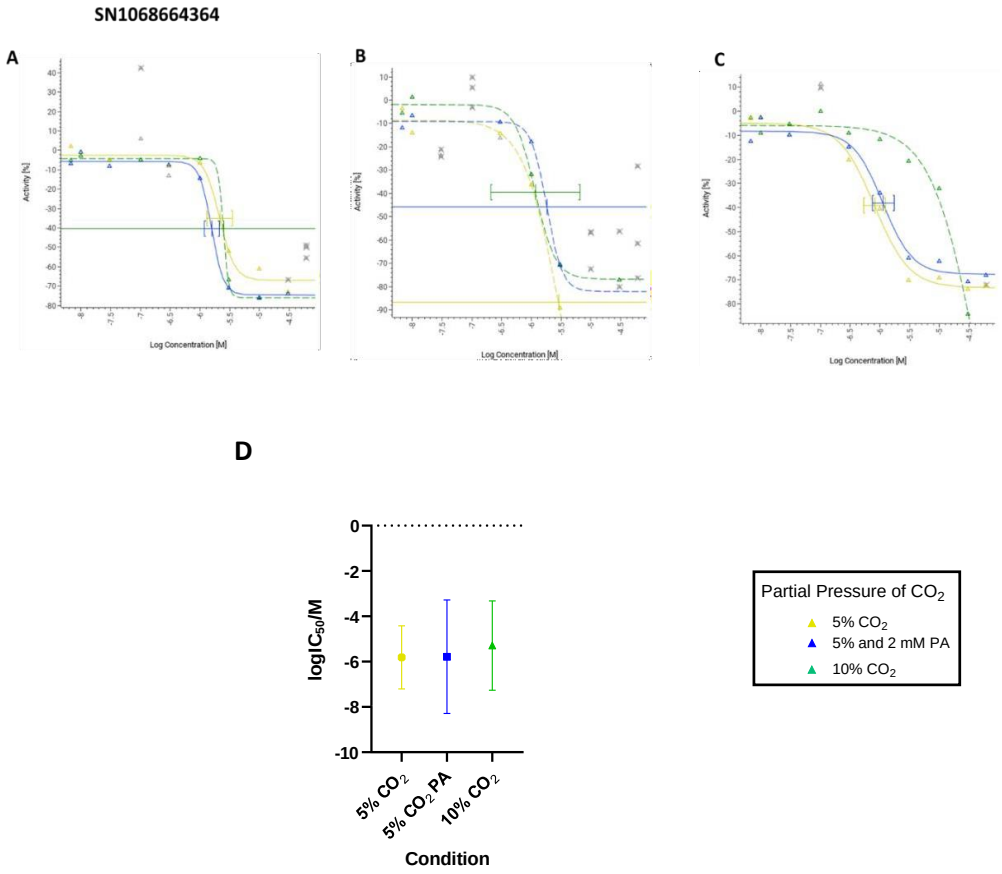


Figure 8-15 Dose-response data for SN1068664364 where A-D are detailed in Figure 8-14.

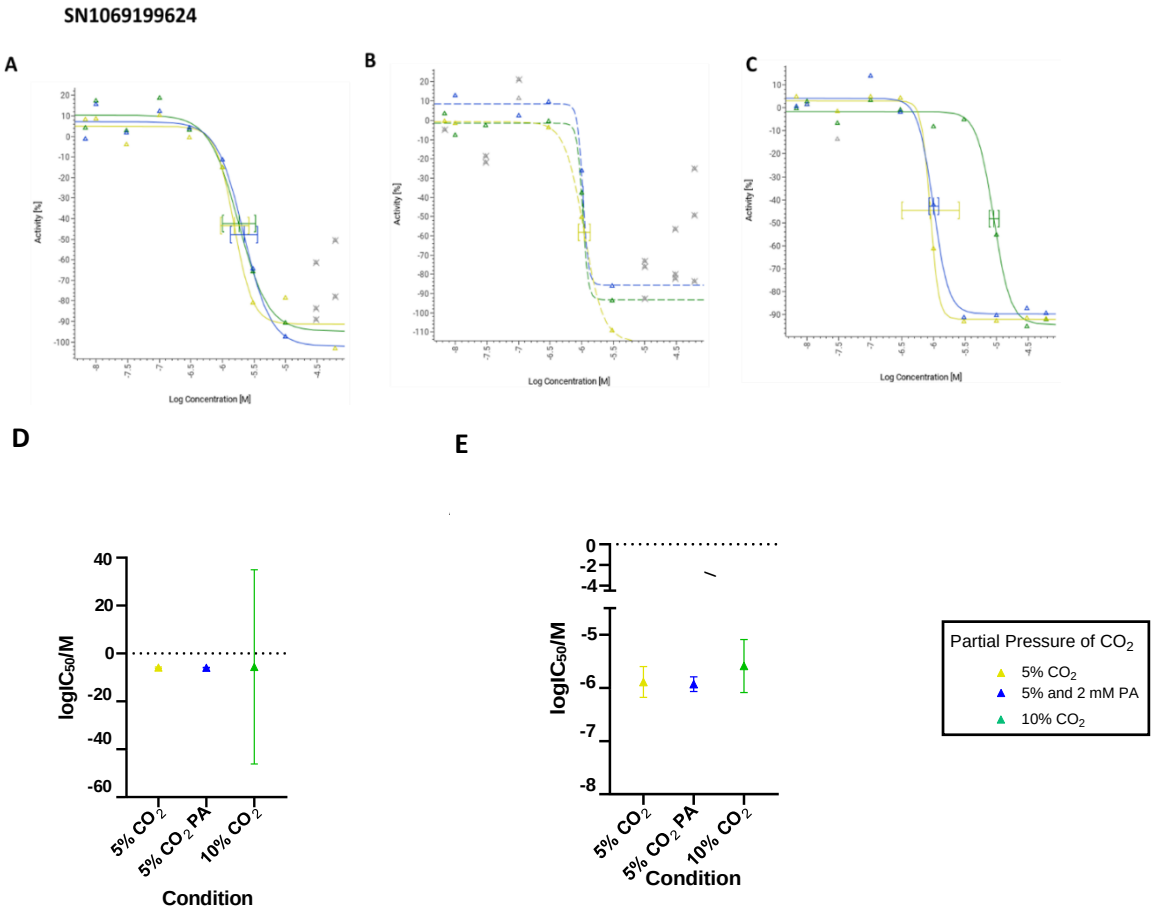


Figure 8-16 Dose-response data for SN1069199624 where A-E are detailed in Figure 8-14.

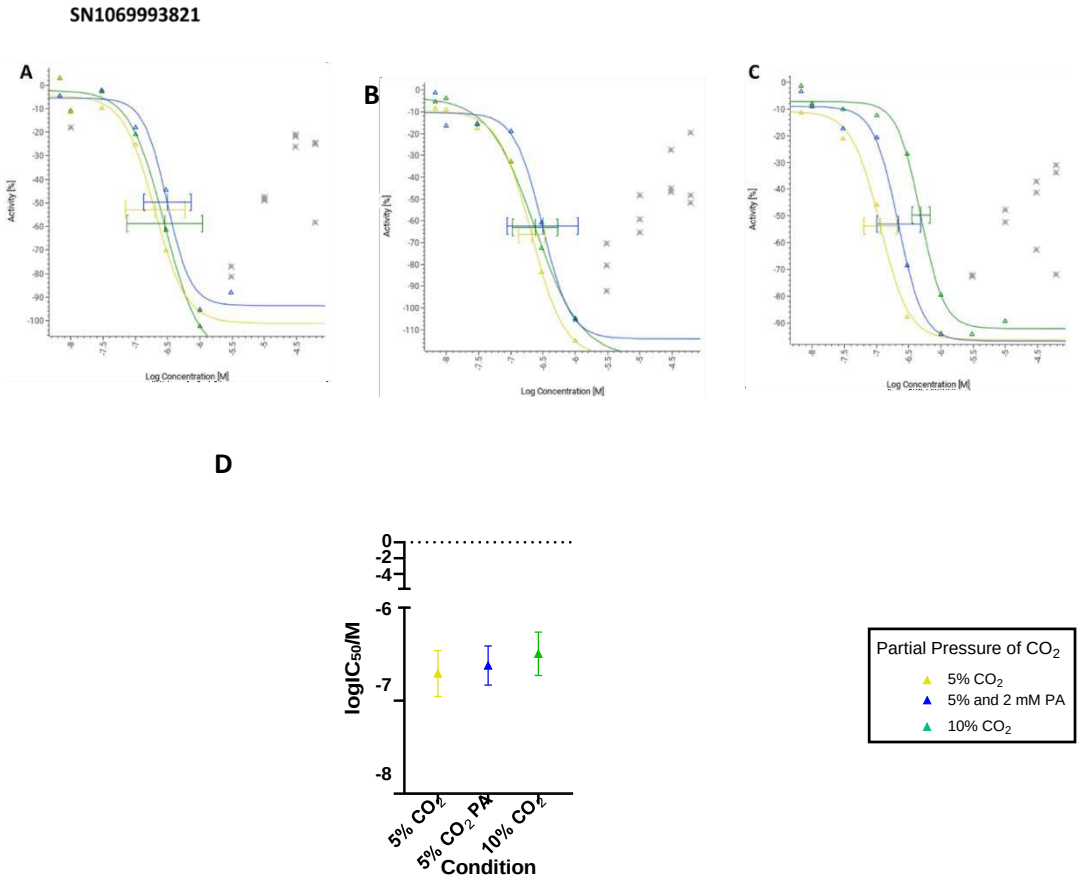


Figure 8-17 Dose-response data for SN1069993821 where A-D are detailed in Figure 8-14.

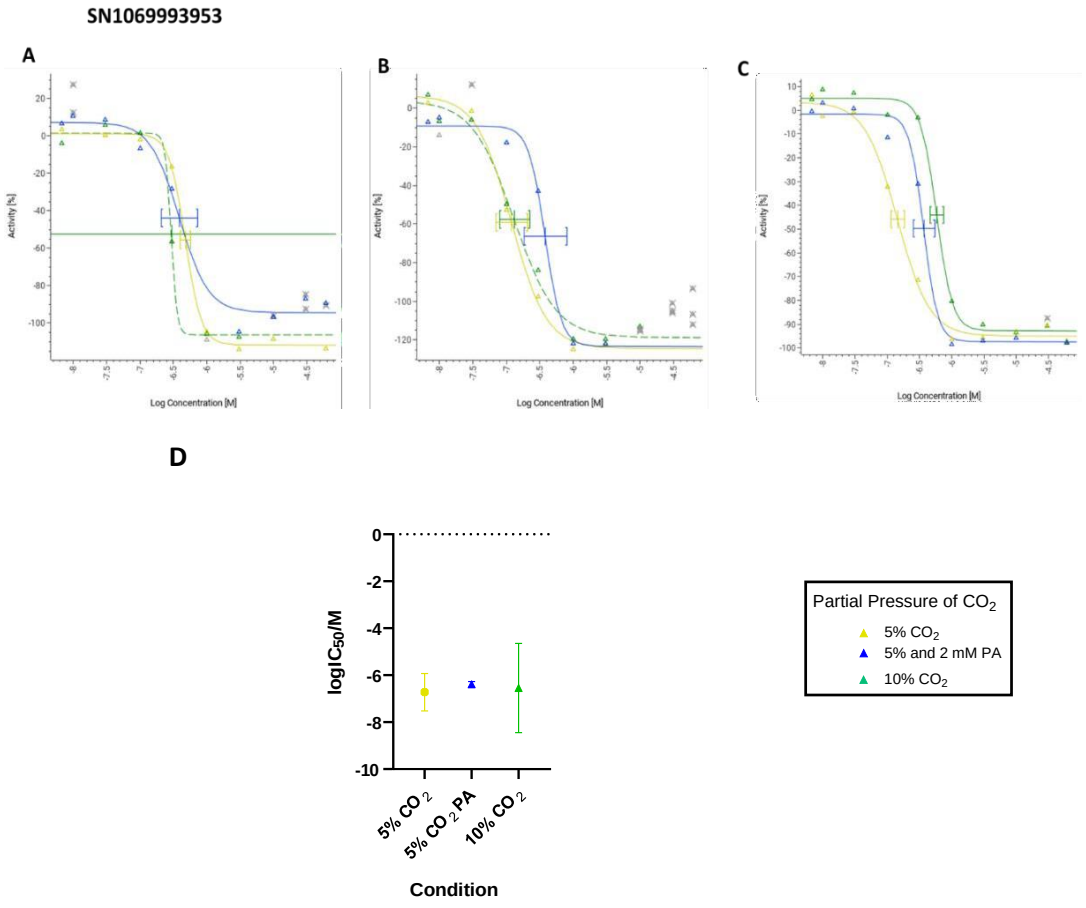


Figure 8-18 Dose-response data for SN1069993953 where A-D are detailed in Figure 8-14.

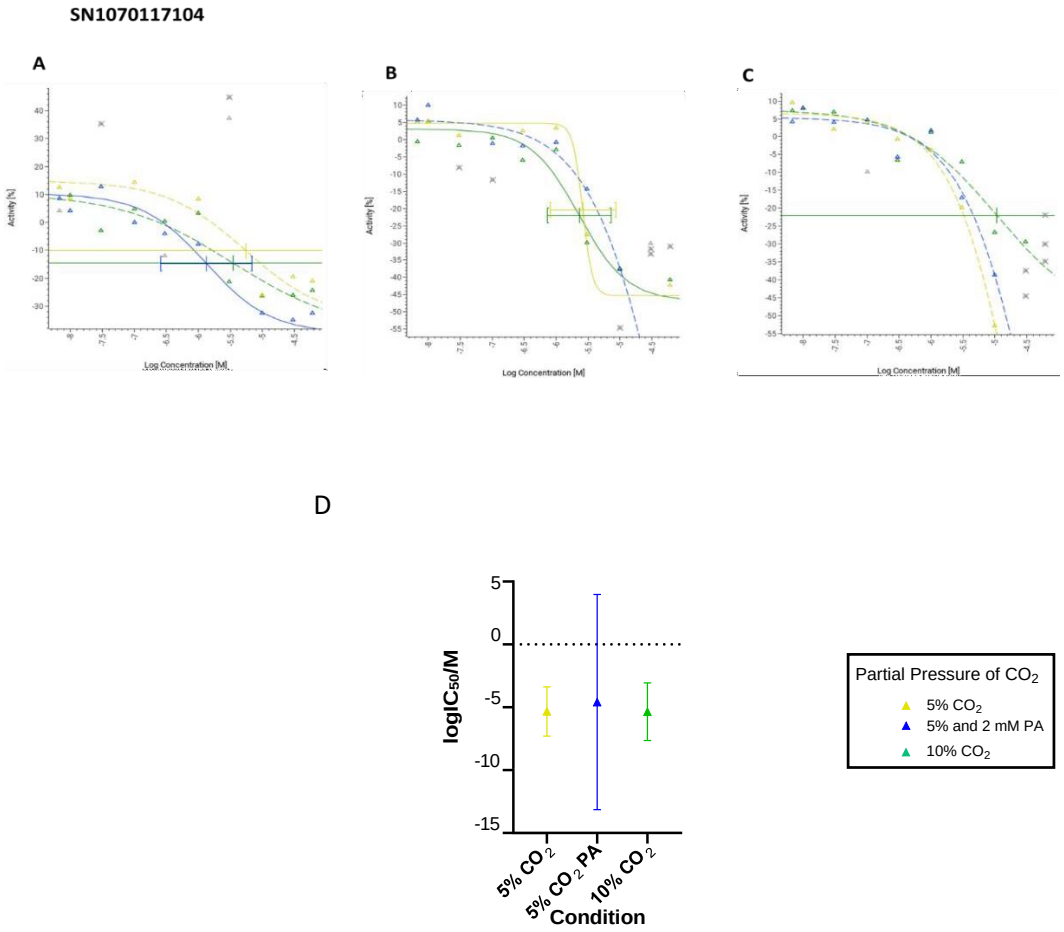


Figure 8-19 Dose-response data for SN1070117104 where A-D are detailed in Figure 8-14.

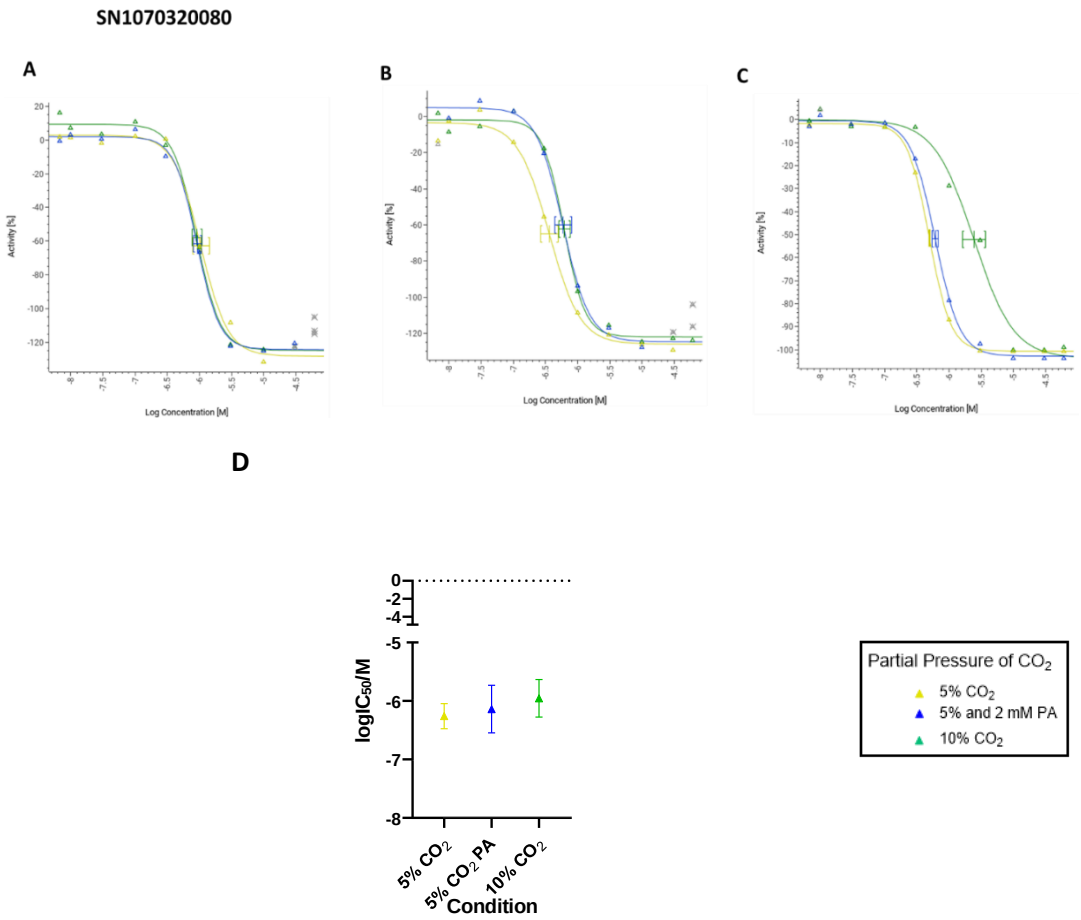


Figure 8-20 Dose-response data for SN1070320080 where A-D are detailed in Figure 8-14.

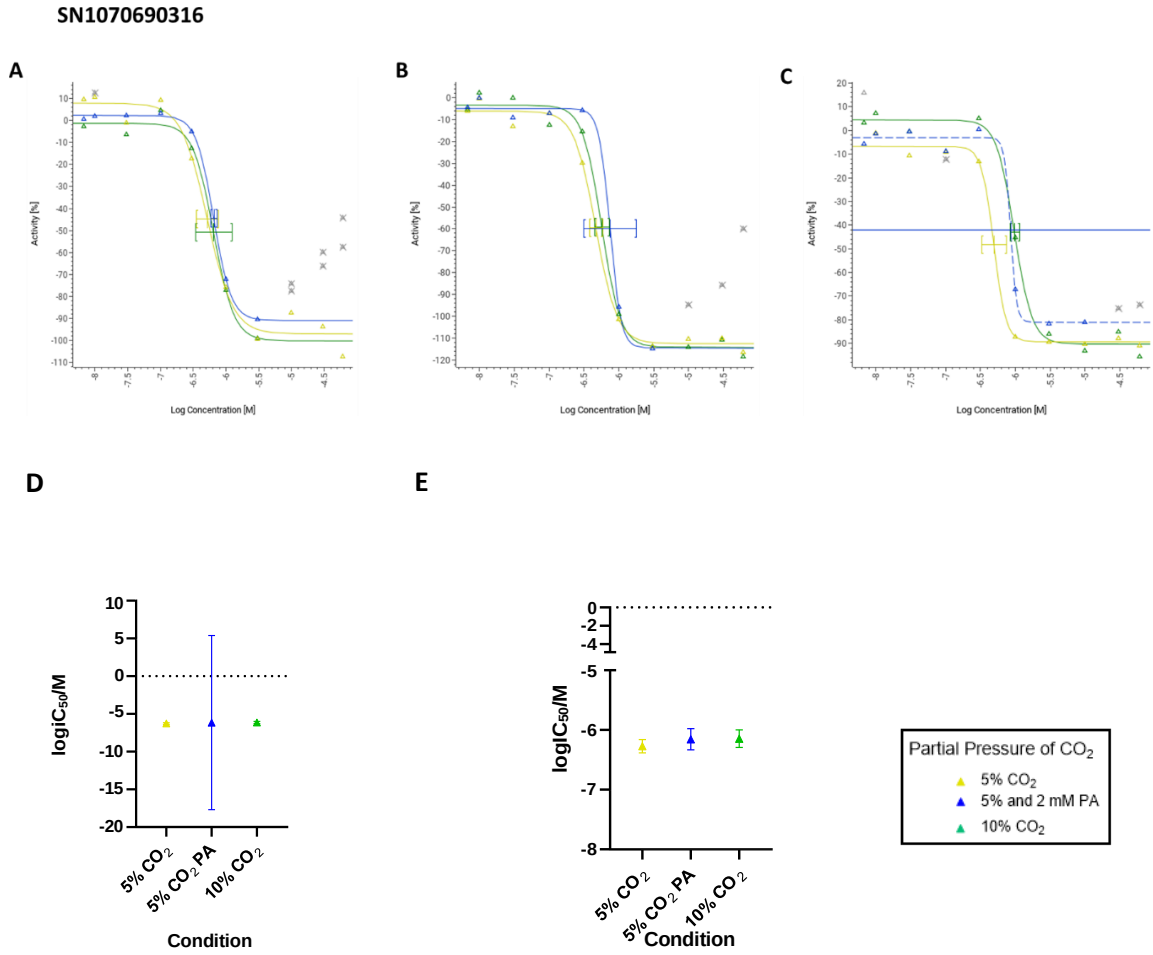


Figure 8-21 Dose-response data for SN1070690316 where A-E are detailed in Figure 8-14.

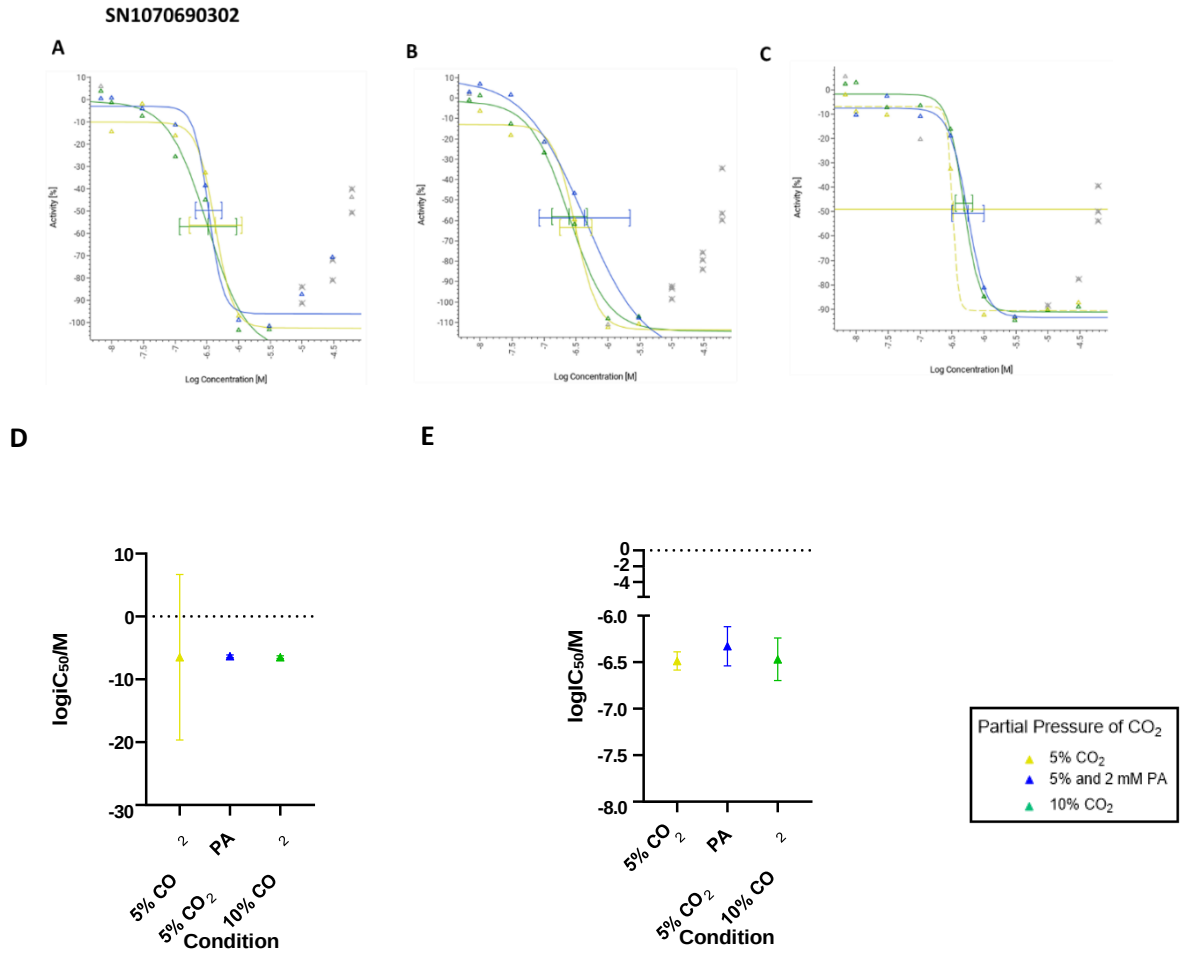
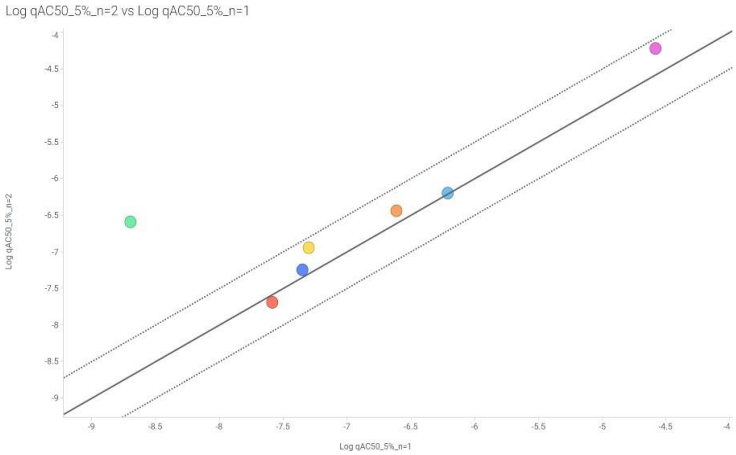
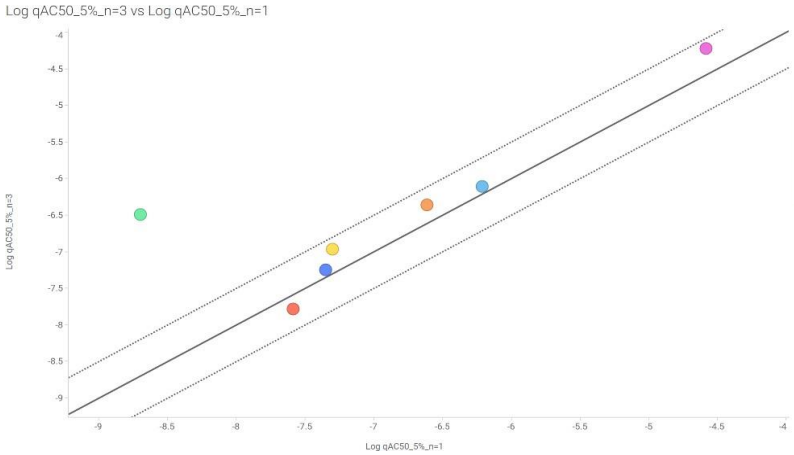


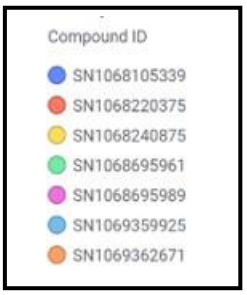
Figure 8-22 Dose-response data for SN1070690302 where A-E are detailed in Figure 8-14.

A



B





C

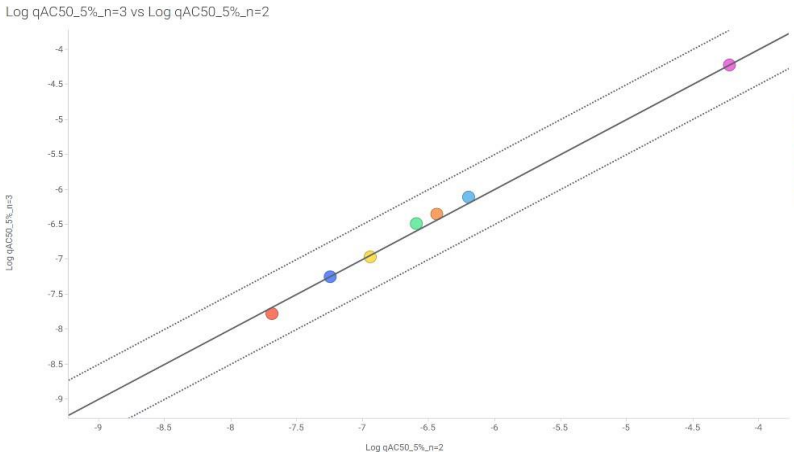


Figure 8-23 Concordance of mean logqIC_{50} values for the BRD4 compounds across the three 5% CO_2 replicates (A-C). The solid line is y=x and represents perfect concordance and the dotted lines are the 95% confidence interval which if points line with are regarded as strongly correlated.

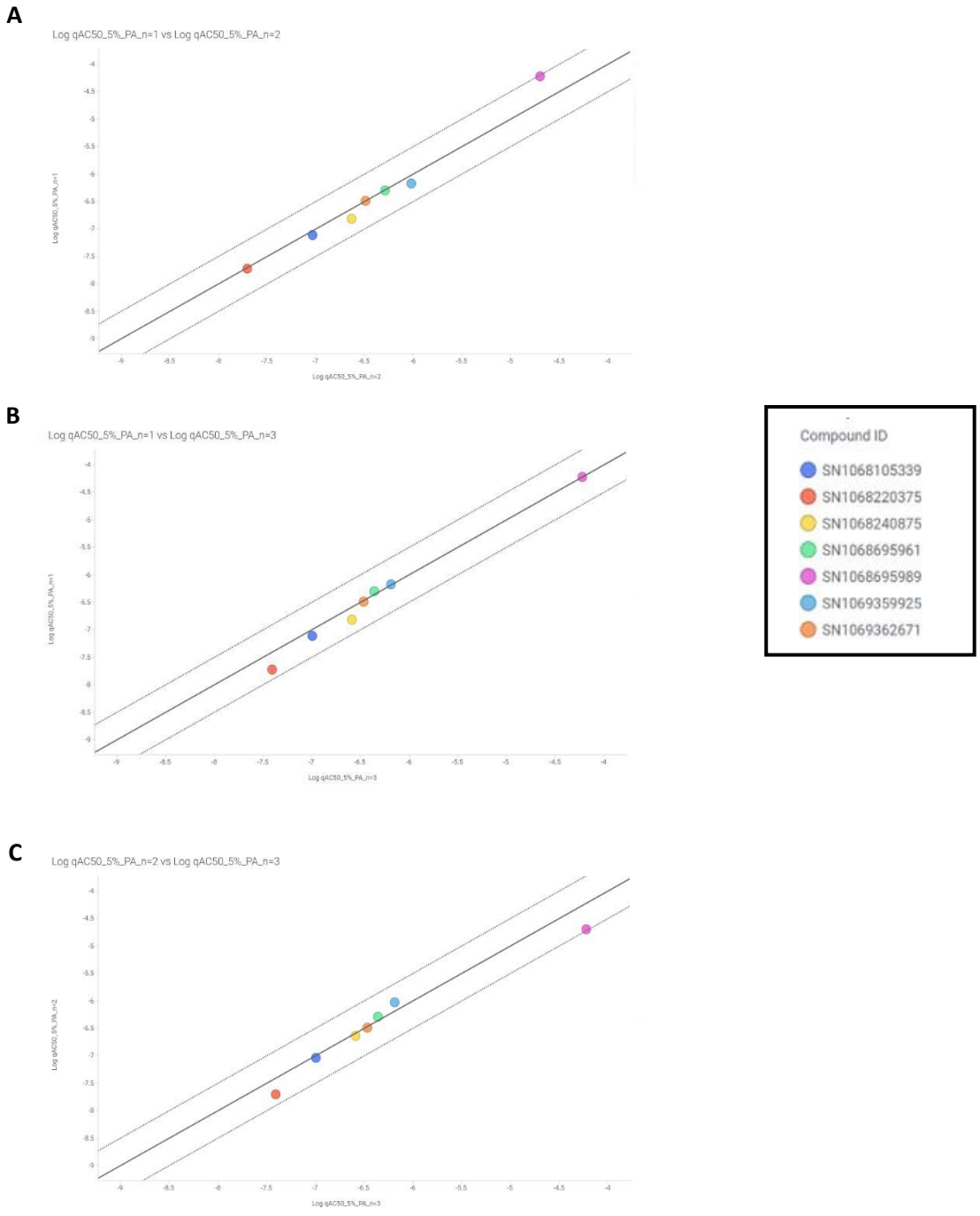


Figure 8-24 Concordance of mean logqIC_{50} values for the 5% CO_2 PA replicates (A-C) as detailed in Figure 8-23.

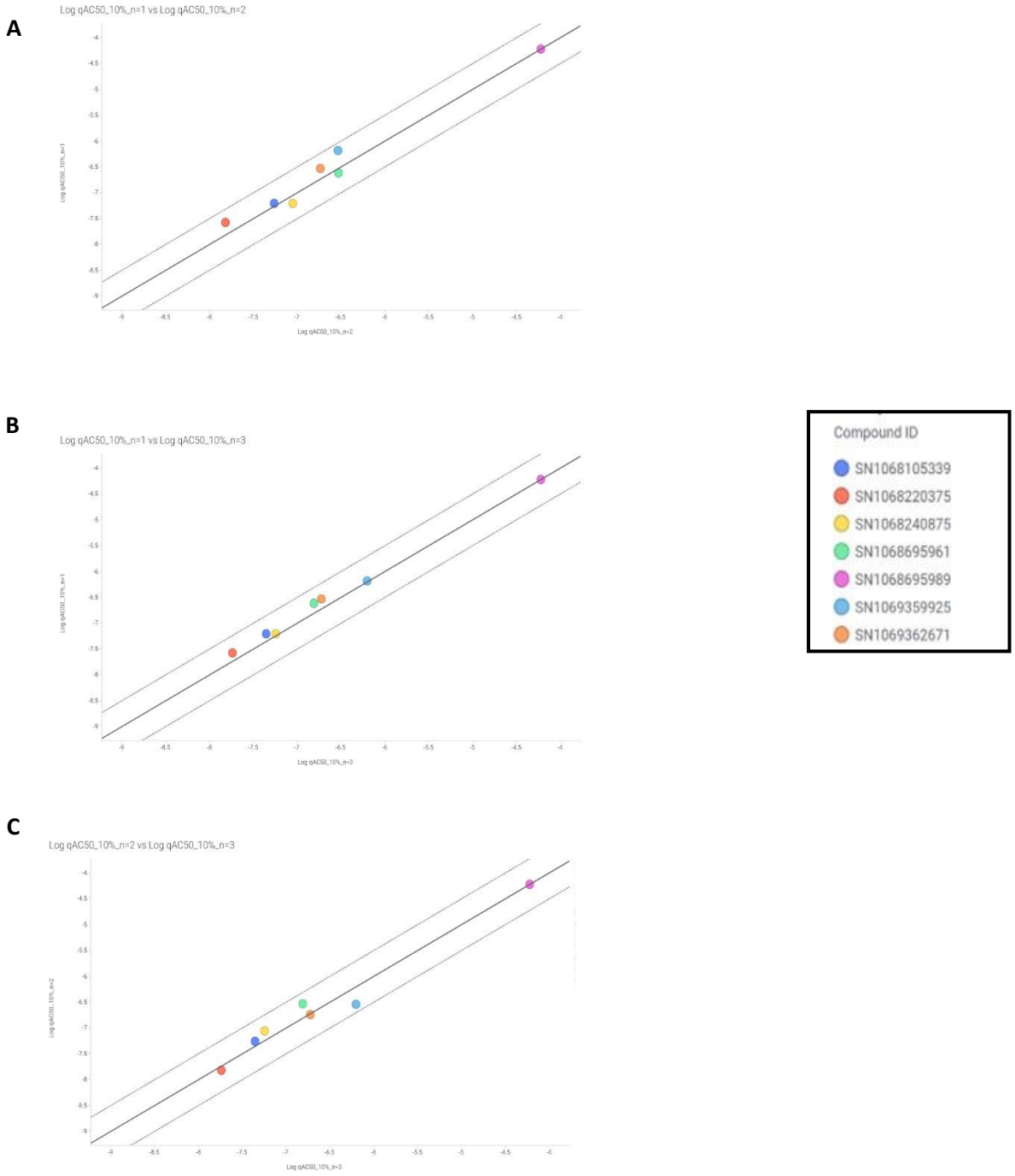
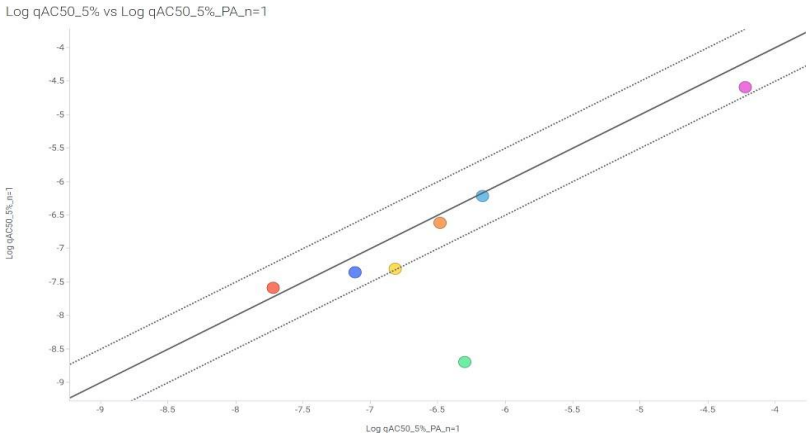
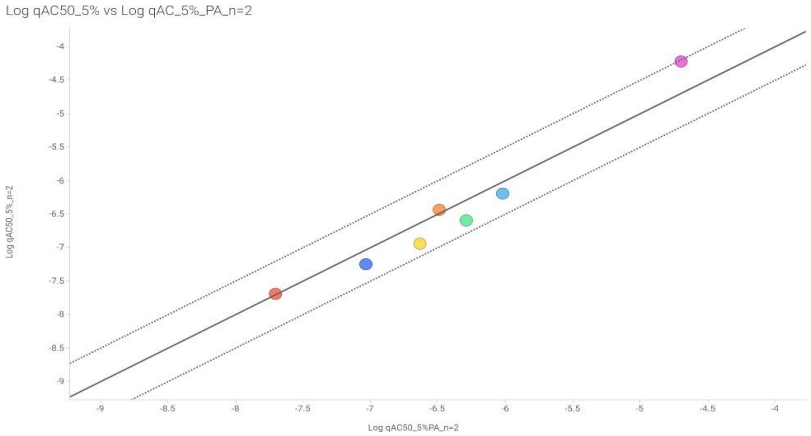


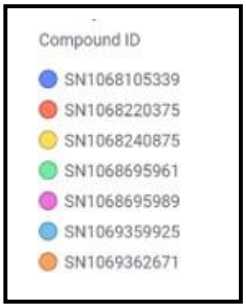
Figure 8-25 Concordance of mean logqIC_{50} values for the 10% CO_2 replicates (A-C) as detailed in Figure 8-23.

A



B





C

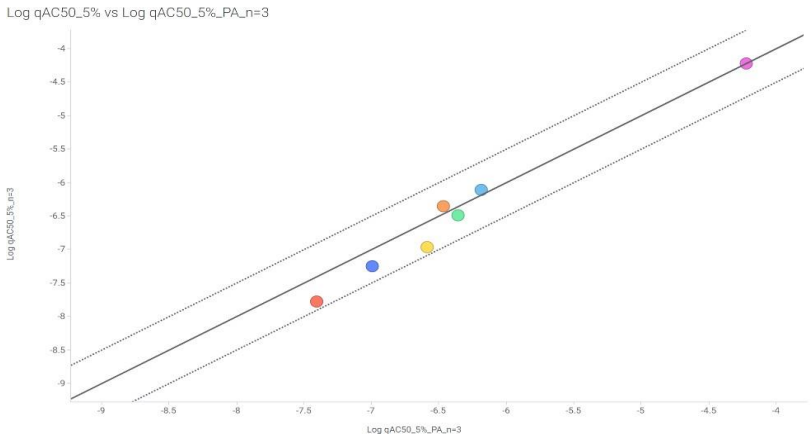


Figure 8-26 Concordance of mean logqIC_{50} values for the 5% CO_2 vs 5% CO_2 PA across 3 replicates (A-C) as detailed in Figure 8-23.

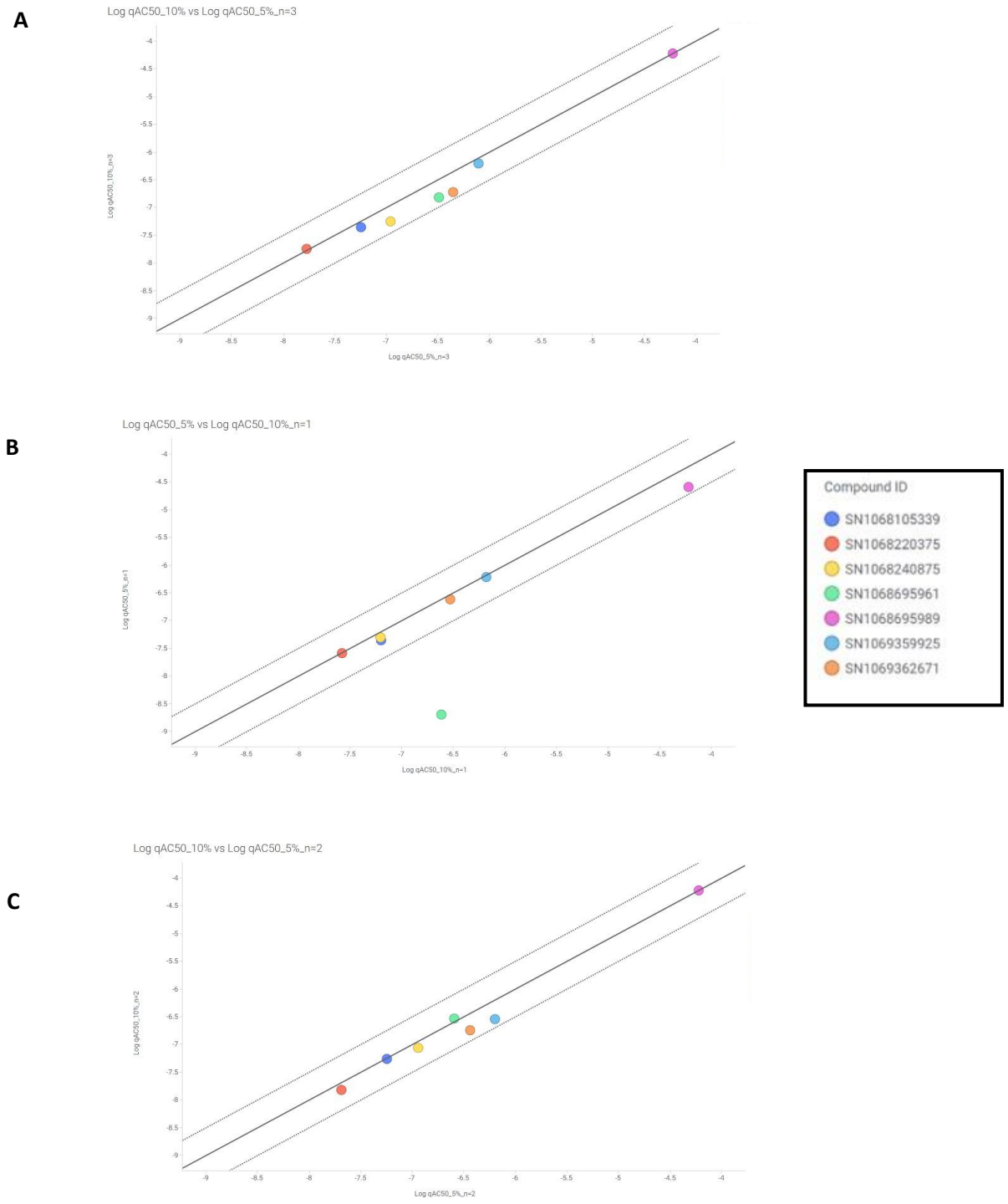
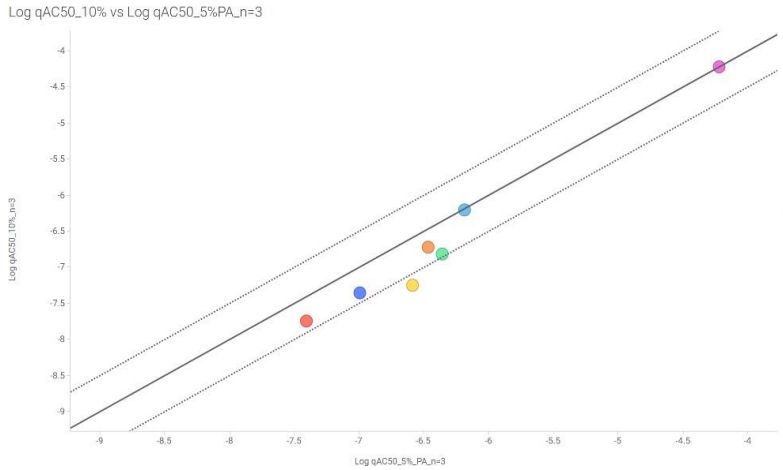
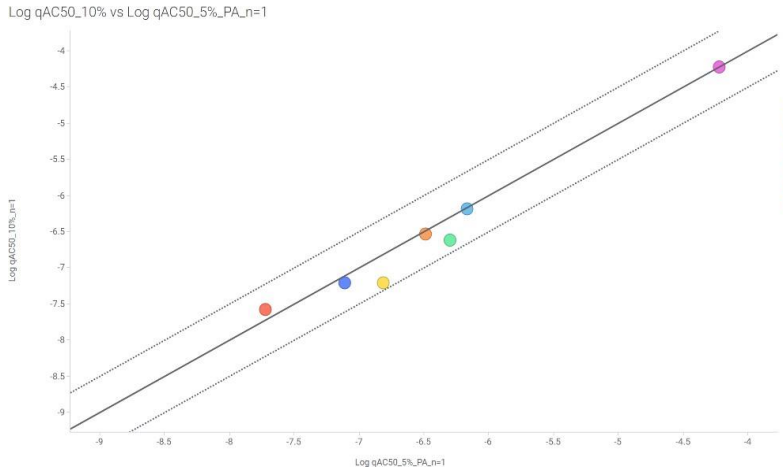


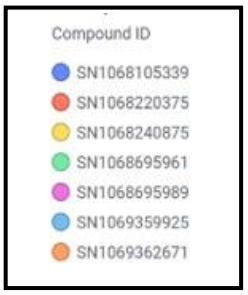
Figure 8-27 Concordance of mean logqIC_{50} values for the 5% CO_2 vs 10% CO_2 across 3 replicates (A-C) as detailed in Figure 8-23.

A



B





C

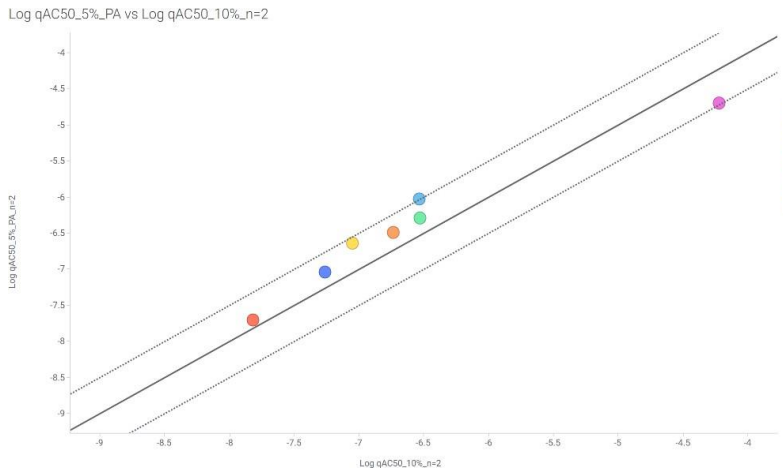


Figure 8-28 Concordance of mean logqIC_{50} values for the 5% CO_2 PA vs 10% CO_2 across 3 replicates (A-C) as detailed in Figure 8-23.

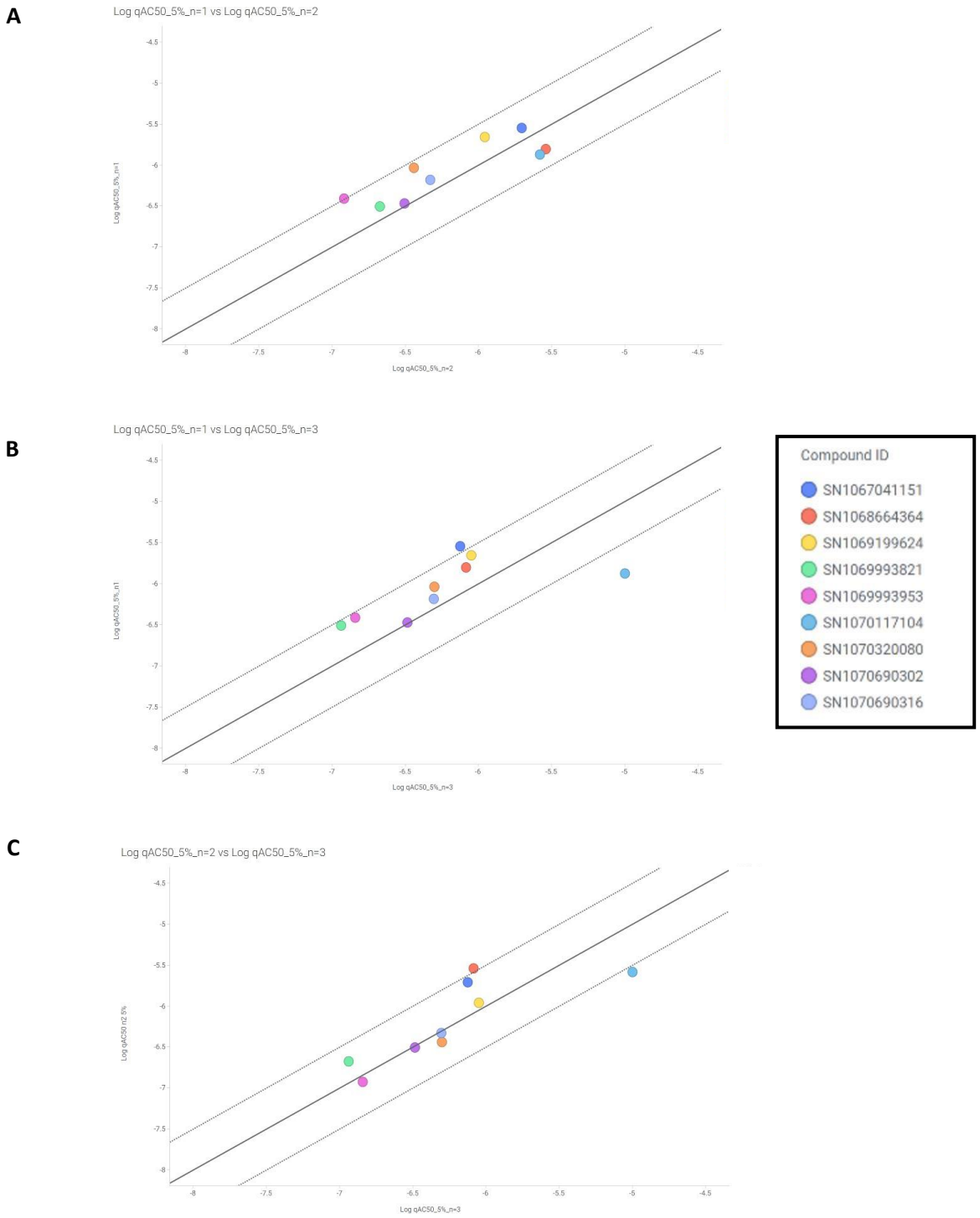
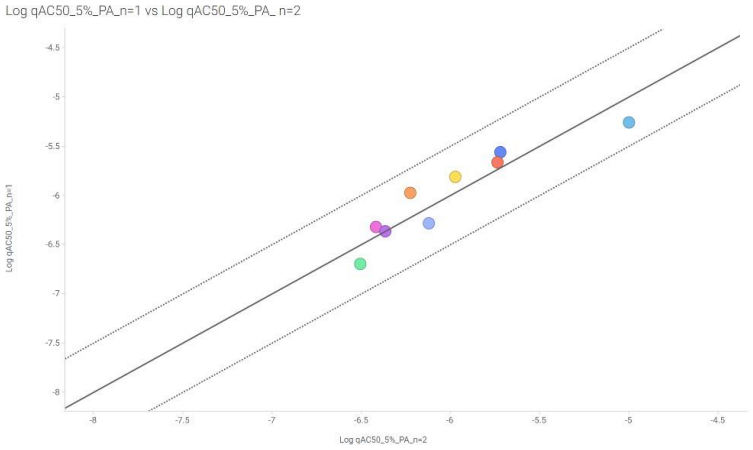
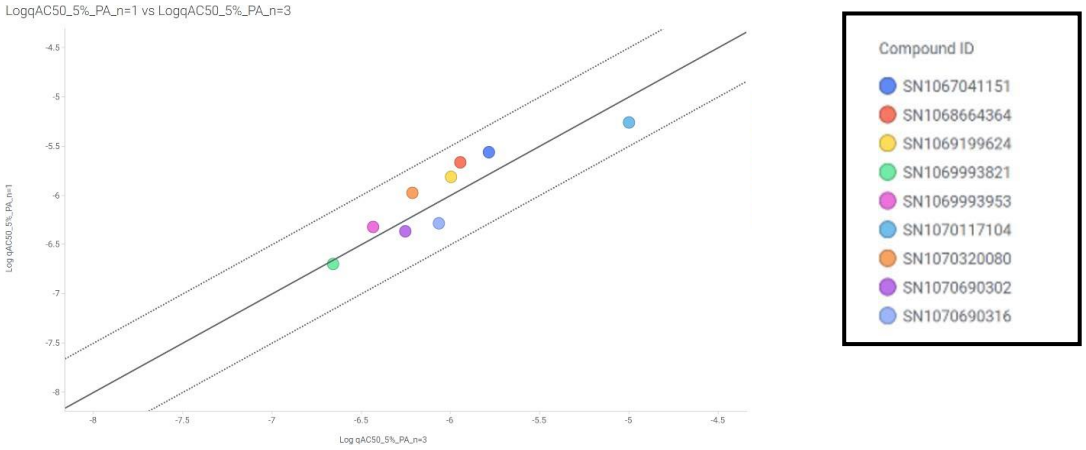


Figure 8-29 Concordance of mean logqIC$_{50}$ values for the SMARCA2 compounds across the three 5% CO_2 replicates (A-C). The solid line is y=x and represents perfect concordance and the dotted lines are the 95% confidence interval which if points line with are regarded as strongly correlated.

A



B



C

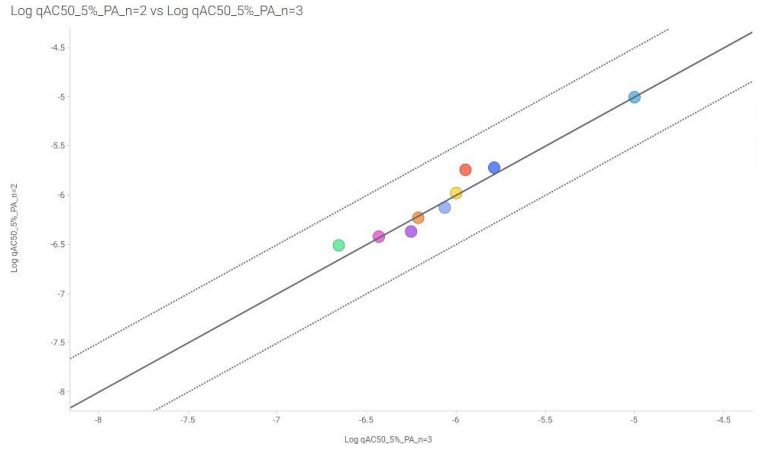


Figure 8-30 Concordance of mean logqIC_{50} values for the 5% CO_2 PA replicates (A-C) as detailed in Figure 8-29.

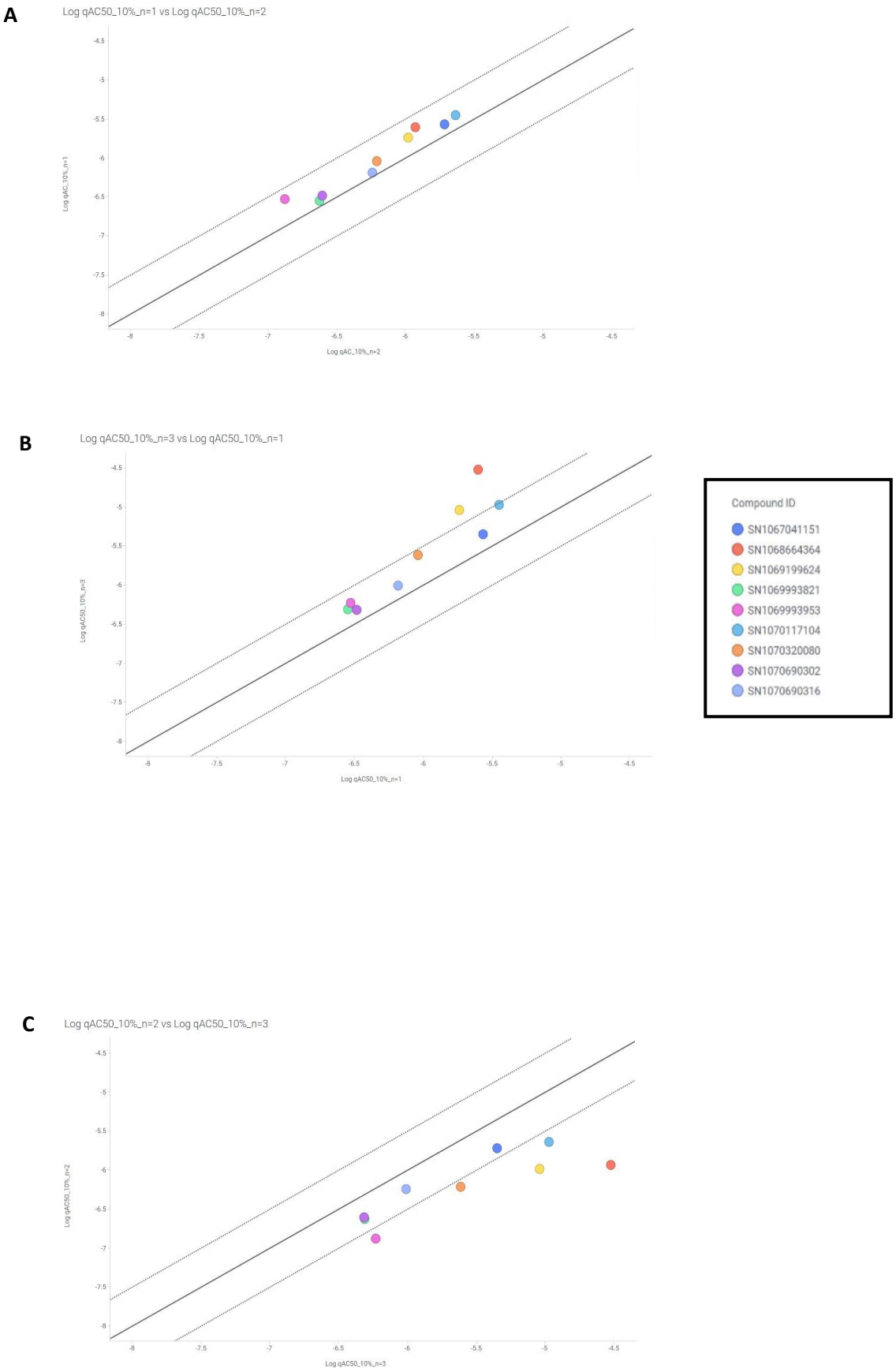
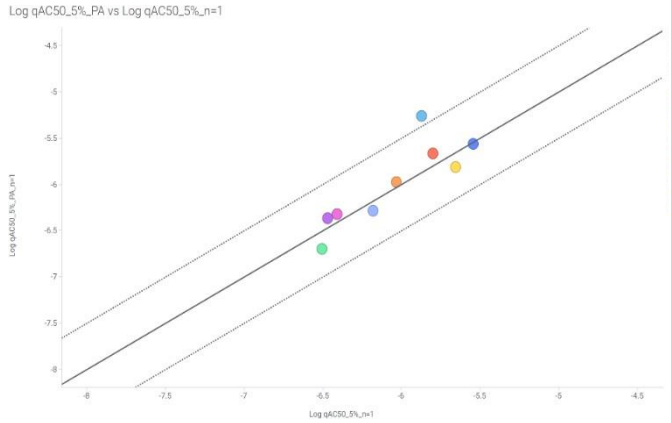
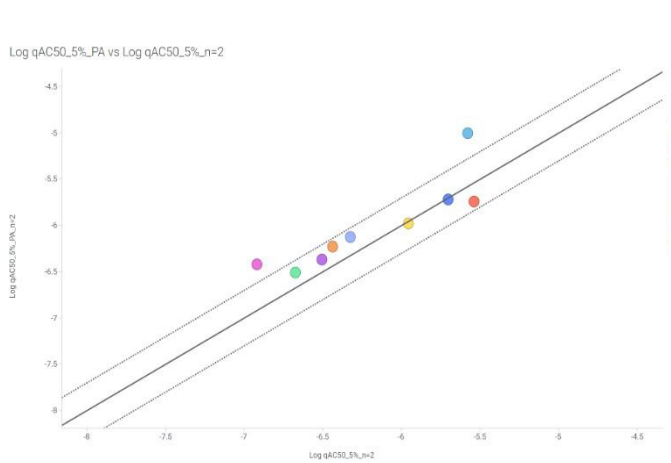


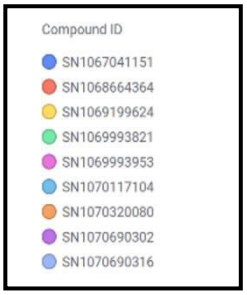
Figure 8-31 Concordance of mean logqIC_{50} values for the 10% CO_2 replicates (A-C) as detailed in Figure 8-29.

A



B





C

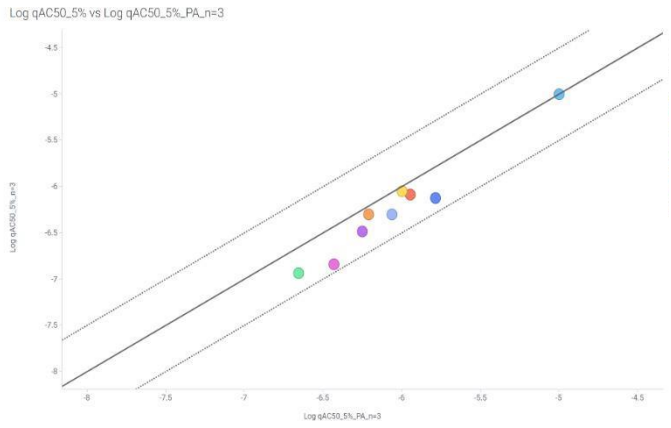
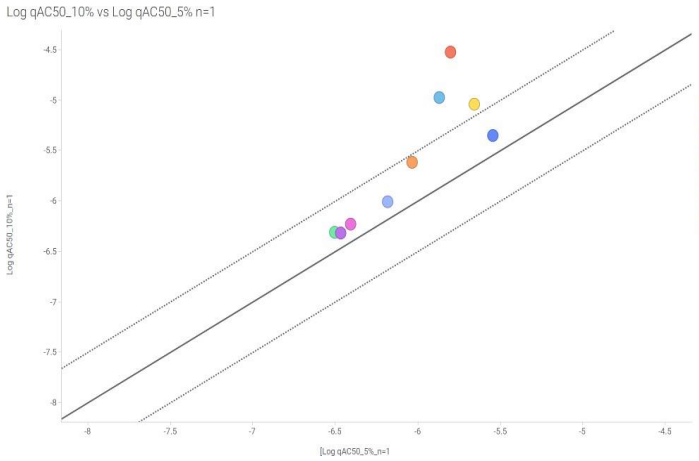
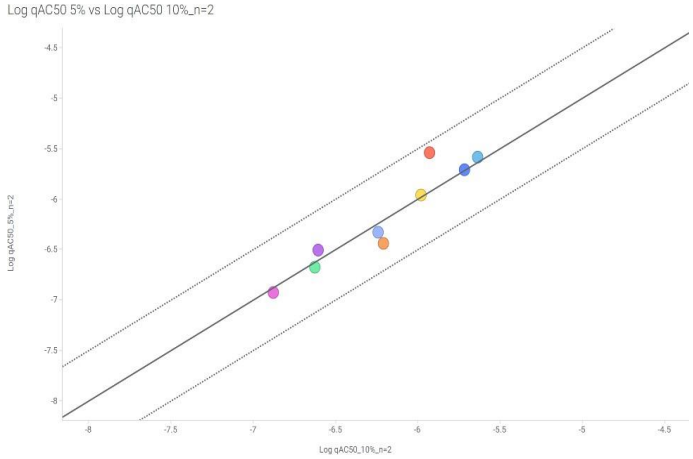


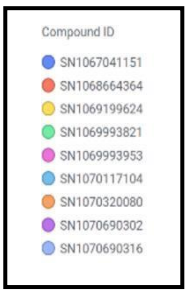
Figure 8-32 Concordance of mean logqIC_{50} values for the 5% CO_2 vs 5% CO_2 PA across 3 replicates (A-C) as detailed in Figure 8-29.

A



B





C

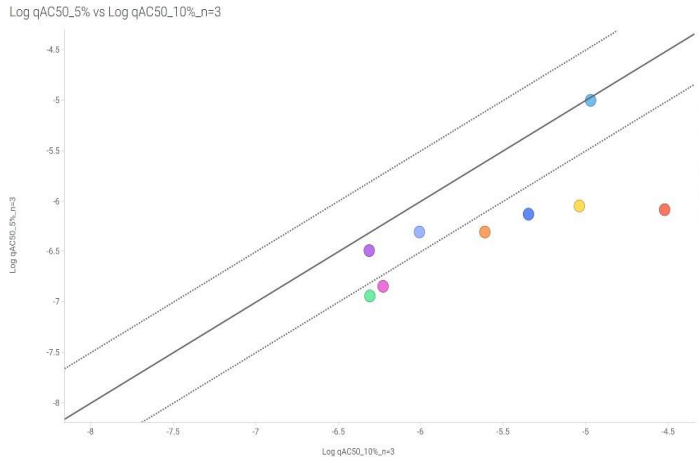


Figure 8-33 Concordance of mean logqIC_{50} values for the 5% CO_2 vs 10% CO_2 across 3 replicates (A-C) as detailed in Figure 8-29.

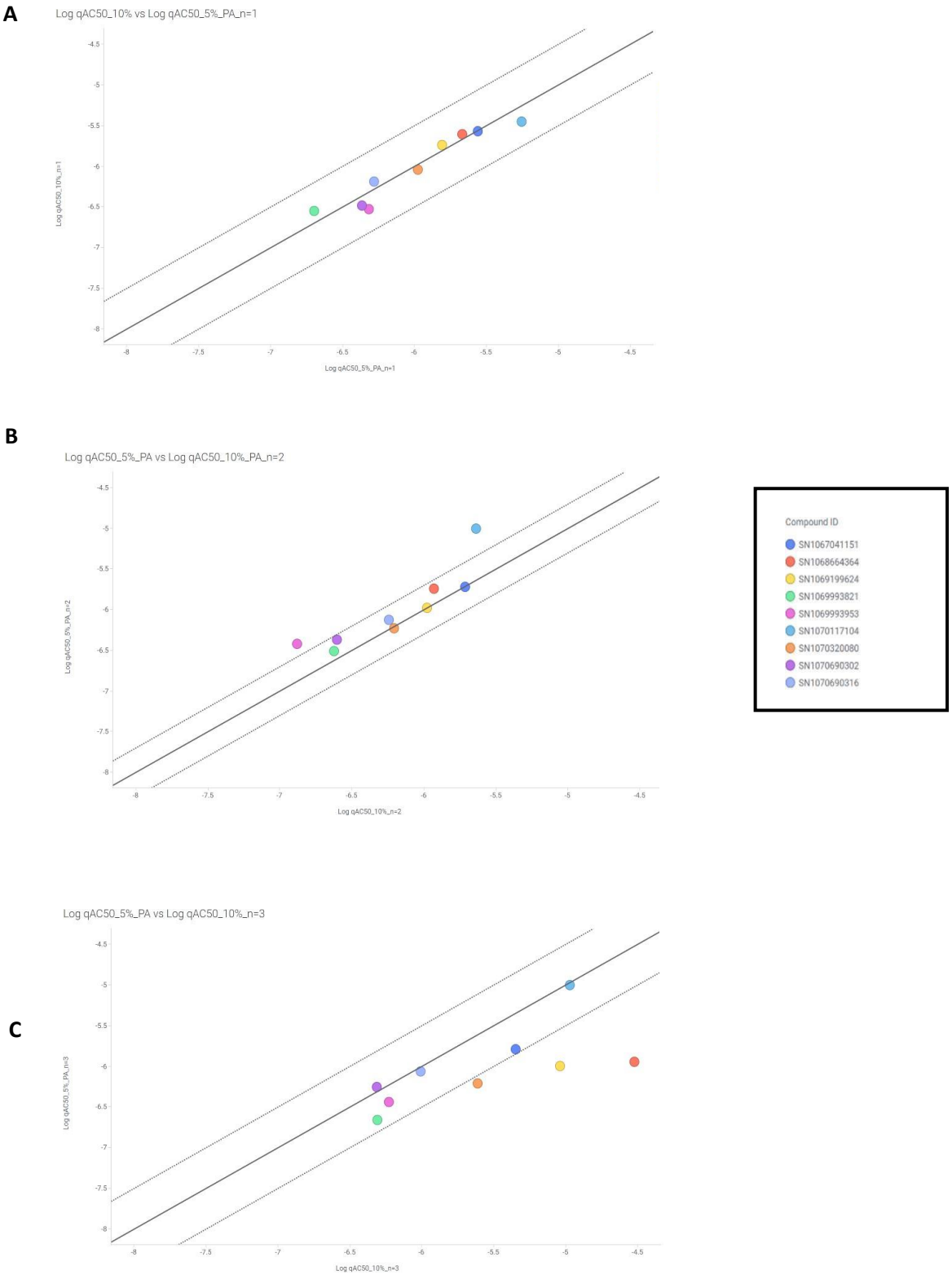


Figure 8-34 Concordance of mean logqIC_{50} values for the 5% CO_2 PA vs 10% CO_2 across 3 replicates (A-C) as detailed in Figure 8-29.

8.3 Supplementary Information for Chapter 5

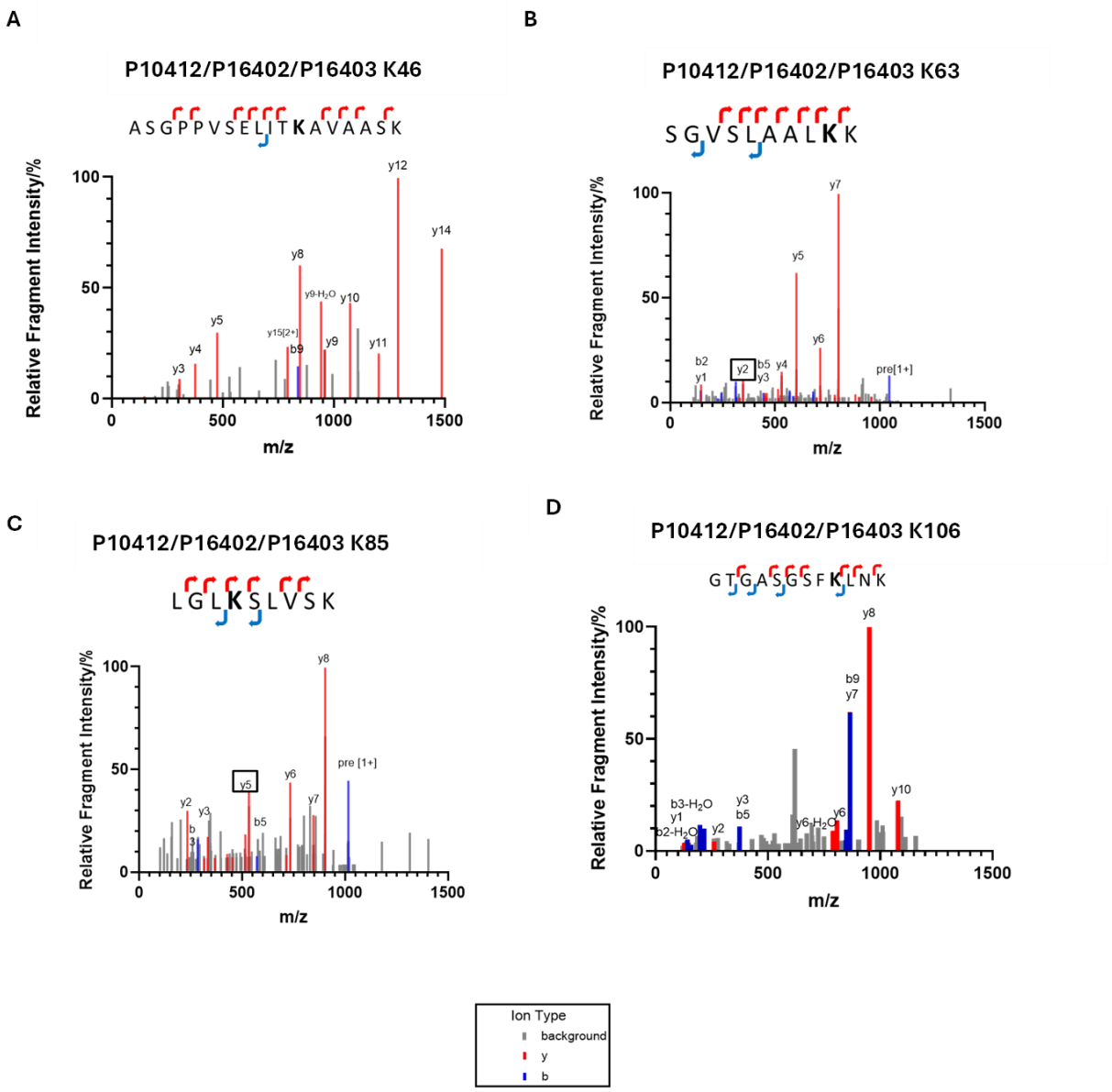


Figure 8-35 Identification of carbamate histone H1 hits from HEK293 lysate screening. Plots of relative fragment intensity versus m/z from LCMSMS identifying trapped carbamates on (A) H1K46 (B) H1K63, (C) H1K85 and (D) H1K106 in the presence of CO_2. Each spectrum is associated with a peptide sequence illustrating the identification of predominant y (red) and b (blue) ions. The grey peaks represent background ions, and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

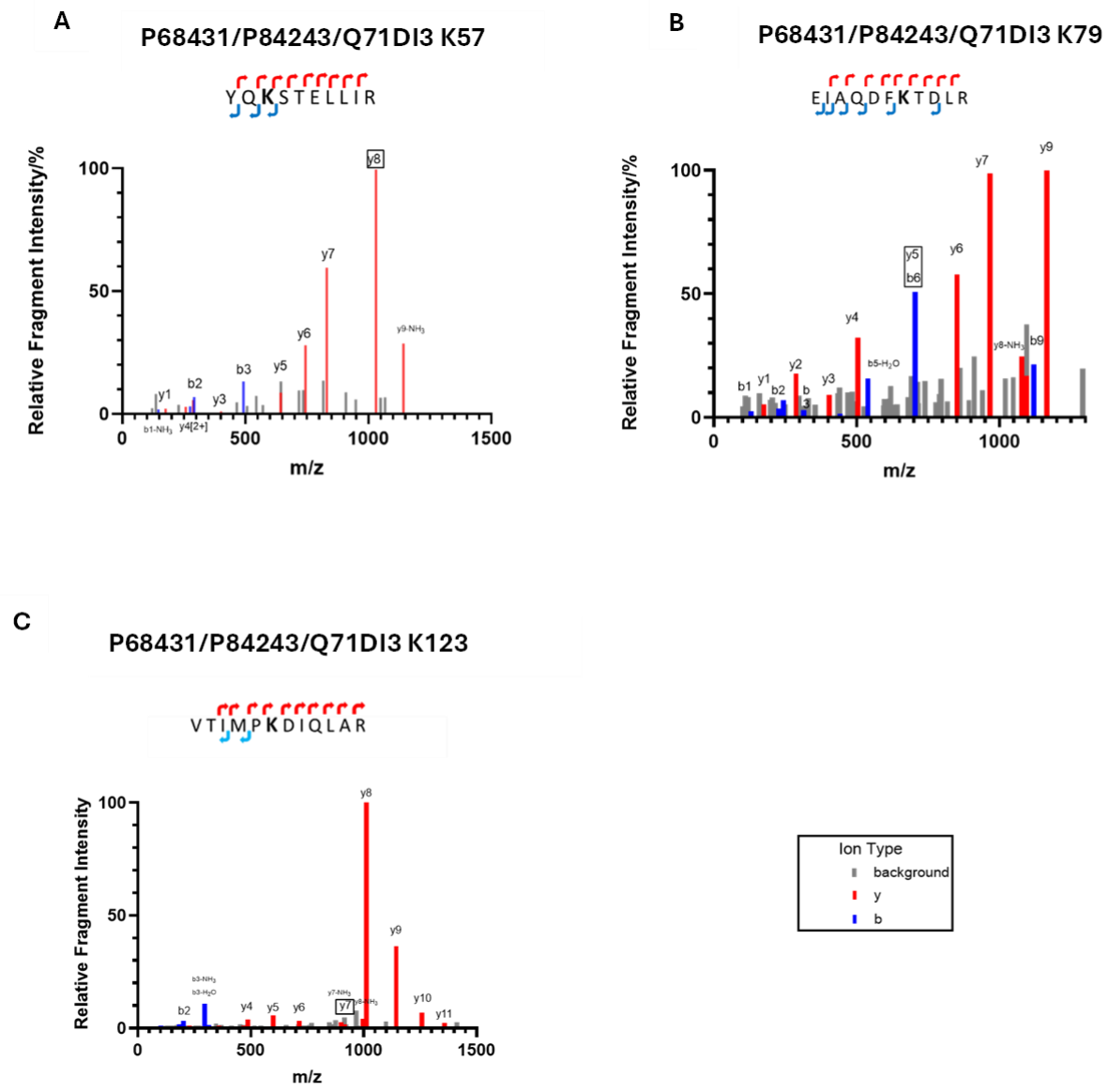


Figure 8-36 Identification of carbamate histone H3 hits from HEK293 lysate screening where (A) H3K57, (B) H3K79 and (C) H3K123. Plots of relative fragment intensity versus m/z from LCMSMS identifying the trapped carbamate in the presence of CO_2. Each spectrum is associated with a peptide sequence illustrating the identification of predominant y (red) and b (blue) ions. The grey peaks represent background ions and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

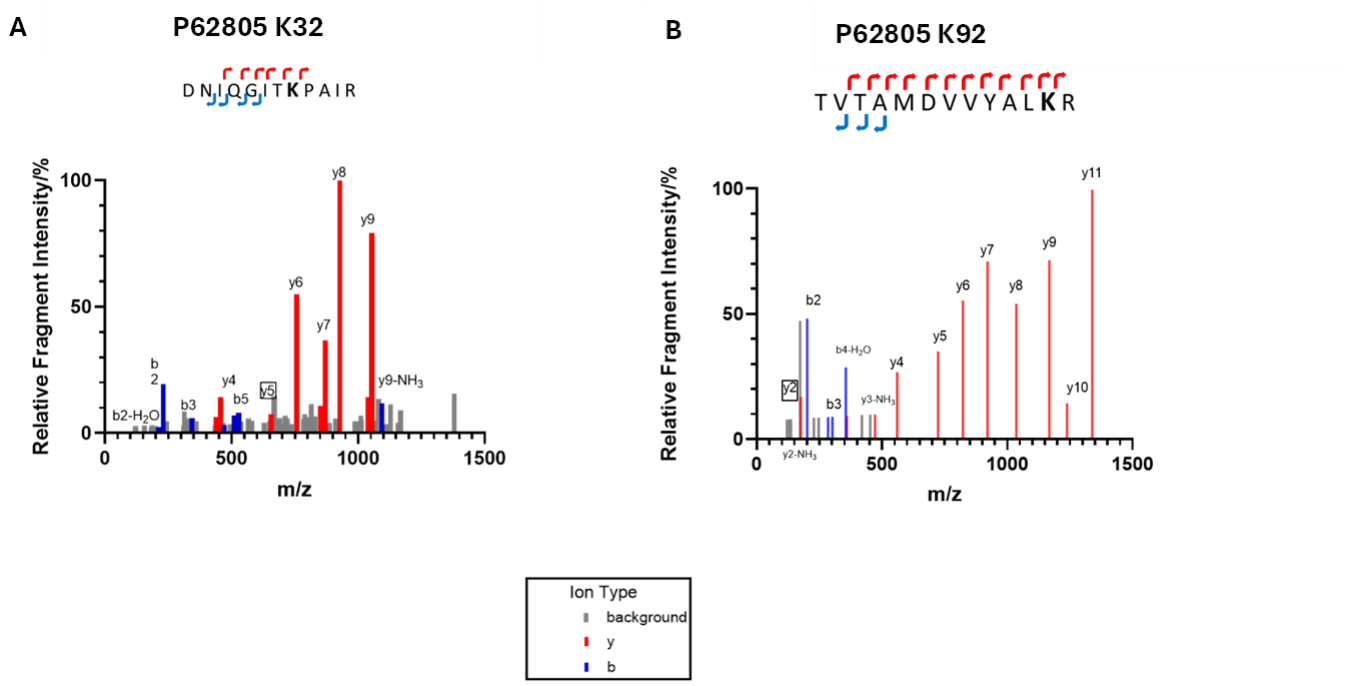


Figure 8-37 Identification of carbamate histone H4 hits from HEK293 lysate screening where (A) H4K32 and (B) H4K92. Plots of relative fragment intensity versus m/z from LCMSMS identifying the trapped carbamate in the presence of CO_2. Each spectrum is associated with a peptide sequence illustrating the identification of predominant y (red) and b (blue) ions. The grey peaks represent background ions, and the carbamate-modified residue is displayed in bold. The y ion corresponding to the carbamylated residue is highlighted.

Sample-12C Not Propionylated Sample	Histone Identified	Coverage/%
TEO_20 mM_bicarbonate_1	H1.2 H1.3 H1.4 H2A H3.1/2/3.3/It H4	46 38 44 31 38 64
TEO_20 mM_bicarbonate_2	H1.2 H1.3 H1.4 H2A H2B H3.1/2/3/1t H4	40 38 44 29 48 37 63
TEO_20 mM_bicarbonate_3	H1.0 H1.2 H1.4 H1X H2A.Z H2AX H2B.1 H2B.2 H3 H3.2 H4	11 42 43 17 31 45 82 79 32 41 59
TEO_50 mM_bicarbonate_1	H1.0 H1.2 H1.3 H1.4 H2A.1 H2A.V/Z H4	23 45 42 48 14 31 63
TEO_50 mM_bicarbonate_2	H1.2 H1.4 H1X H2AX H2A.Z/V H3.1/2/1t H4	38 39 23 42 29 36 63

TEO_50 mM_bicarbonate_3	H1.2 H1.3/4 H1X H2A.1 H2A.Z/V H3 H3.1/2 H4	42 40 31 41 31 36 45 61
TEO_0 mM_bicarbonate_1	H1.2 H1.3 H2A H2A.1 H2A.V/Z H2AX H3.1/1t/2 H4	37 36 9 5 31 38 26 52
TEO_0 mM_bicarbonate_2	H1.2 H1x H2A.1 H3.1/1t/2 H4	39 22 13 24 57
TEO_0 mM_bicarbonate_3	H1.2 H1.3 H1X H2A H2A.1 H2A.V/Z H4	32 31 15 9 5 29 57
No_TEO_0mM_bicarbonate_1	H1.2 H1.3 H1.4 H1X H2A.1 H2A.Z/V H2AX H4	36 35 39 22 27 31 57 57
No_TEO_0 mM_bicarbonate_2	H1.0 H1.2 H1.3 /4 H1X H2A.1 H2A.2 H2A.Z/V H3.1/3.2/3.3 H4	16 32 31 17 33 8 40 38 58
No_TEO_0 mM_bicarbonate_3	H1.0 H1.2 H1.3 H1.4 H1X	10 35 34 31 13

	H2AX H2A.2 H2A.Z/V H4	69 24 31 58

Table 8-3 Coverage of native nucleosome samples not modified by propionylation across the experiment conditions specified, with or without triethyloxonium (TEO) trapping, concentration of carbon 12 (12C) inorganic carbon and replicate number.

Sample-13C Not Propionylated Sample	Histone Identified	Coverage/%
TEO_20 mM_bicarbonate_1	H1.2 H1.3 H1.4 H1X H2A.1 H2A.Z/V H3.1/3.1T/3.2 H4	43 38 48 17 17 31 24 60
TEO_20 mM_bicarbonate_2	H1.0 H1.2 H1.3 H1.4 H1X H2A.1 H2A.Z/V H3.1/3.1T/3.2 H4	9 40 37 43 23 10 29 24 58
TEO_20 mM_bicarbonate_3	H1.2 H1.3 H1.4 H1X H2A.1 H2A type 1-C. H2AX H2A.Z/V H3.1/3.1T/3.2 H4	39 38 44 11 16 41 38 31 27 58
TEO_50 mM_bicarbonate_1	H1.2 H1.4 H2A type 1/type 3/type 1-J/type 1-H. H2B type 1-L. H3	37 36 38 26 19
TEO_50 mM_bicarbonate_2	iH1.2 H1.3 H1.4 H1X	38 35 42 22

	H2A type 1-C. H2A.1 H3.1/3.1T/3.2 H4	46 7 27 58
TEO_50 mM_bicarbonate_3	H1.0 H1.2 H1.3 H1.4 H1X H2A.1 H2A.Z/V H4	11 43 39 37 17 6 13 60

Table 8-4 Coverage of native nucleosome samples not modified by propionylation across the experiment conditions specified, with triethyloxonium (TEO) trapping, concentration of carbon 13 (13C) inorganic carbon and replicate number.

Sample 12C Propionylated Sample	Histone Identified	Coverage/%
TEO_20 mM_bicarbonate_1	H2A H2A.3 H4	20 26 43
TEO_20 mM_bicarbonate_2	H1.2 H1X H2A H4	7 9 28 62
TEO_20 mM_bicarbonate_3	H4	56
TEO_50 mM_bicarbonate_1	H2A H2A Type 1 -B/E/ 1-D H2A Type 3 H4	22 28 28 62
TEO_50 mM_bicarbonate_2	H2A Type 1 -B/E/ 1-D H2A H4	5 10 62
TEO_50 mM_bicarbonate_3	H2A Type 1 -C/Type 2-A H2A Type 1-J/1-H/H2A.J/H2A.2 H2A H2A.1 H3 H3.1/3.2/3.3 H4	28 29 10 5 33 22 50
TEO_0 mM_bicarbonate_1	H2A type 1-J/H/H2A.J/ type 2/type 3/ type 1-B/E/ type 1 C/ type D/type 2-A/type 1. H2A H2A.1 H3.1/1T/2 H4	21 16 5 31 48
TEO_0 mM_bicarbonate_2	H2A type 1-J/1-H/ H2A.J/type 2-C/type 1/type 2-A/type 1 -C H2A.2 core histone H3 H3.1/2/3 H4	27 26 6 38 37 33
TEO_0 mM_bicarbonate_3	H2A type 1-A/J/H/H2A.J/type 2-C/type 1/type 2-A/type 1-C. H3 H3.1/2/3 H4	18 38 37 48
No_TEO_0 mM_bicarbonate_1	H2A.2 H4	26 36
No_TEO_0 mM_bicarbonate_2	H3 H3.1/3.1T/3.3 H4	38 37 64
No_TEO_0 mM_bicarbonate_3	H2A.2 H3.1/3.2/3.3 H4	54 31 51

Table 8-5 Coverage of native nucleosome samples modified by propionylation across the experiment conditions specified, with or without triethyloxonium (TEO) trapping, concentration of carbon 12 (12C) inorganic carbon and replicate number.

Sample-13C Propionylated	Histone Identified	Coverage/%
TEO_20 mM_bicarbonate_1	H2A.Z/V H3.1/2/3 H4	21 22 43
TEO_20 mM_bicarbonate_2	H2A type 1-C. H3 H4	31 33 62
TEO_20 mM_bicarbonate_3	H1.4 H3 H4	10 33 60
TEO_50 mM_bicarbonate_1	H2A type 1-C. H3 H4	42 33 57
TEO_50 mM_bicarbonate_2	H1X H2A./V H3 H4	9 21 32 62
TEO_50 mM_bicarbonate_3	H3 H4	32 52

Table 8-6 Coverage of native nucleosome samples modified by propionylation across the experiment conditions specified, with triethyloxonium (TEO) trapping, concentration of carbon 13 (13C) inorganic carbon and replicate number.

Histone Oct Sample- Not Propionylated 12C unless otherwise stated	Histone Identified	Coverage/%
TEO_20 mM_bicarbonate_1	H1.2/4 H2A type 1-H/J/type 2-C. H2A type 1-A/C H3.1/2/3/1t H4	19 33 32 18 43
TEO_50 mM_bicarbonate_1	H1.2/1.3/1.4 H2A type 1-J/H H2B type 1 - C/E/F/G/I H3.1/3.1T H4	8 41 46 25 50
TEO_0mM_bicarbonate_1	H2A type 1-H/J H2A type 1/type 2A. H3.1/3.1T/3.2 H4	59 58 27 57
No_TEO_0 mM_bicarbonate_1	H1.1/1.2/1.3/1.4 H1t H2B type 1 - C/E/F/G/I/D/N/M/H/K/ F-S/TYPE 2 F H3.1/3.2/3.3 H4	9 10 54 29 58
TEO_20 mM_13C_bicarbonate_1	H1.2/1.3/1.4 Histone H2A type 2-C. Histone H2B type 1-C/E/F/G/I H3.1/3.2 H4	14 41 63 35 57
TEO_50 mM_13C _bicarbonate_1	H2A H2B type 1-C/E/F/G/I H3.1/3.1T/3.2 H4	41 53 23 26

Table 8-7 Coverage of recombinant histone octamer samples not modified by propionylation across the experiment conditions specified, with or without triethyloxonium (TEO) trapping, concentration of carbon 12 or 13 (12C or 13C) inorganic carbon and replicate number. If the carbon isotope is unspecified, it is 12C.

Histone Oct Sample- Propionylated 12C unless otherwise stated	Histone Identified	Coverage/%
TEO_20 mM_bicarbonate_1	H2A type 1-C. H2B type 1. H3.1/2/3 H4	29 23 28 50
TEO_50 mM_bicarbonate_1	H2A type 1-B/E/C/D/ type 2-A/C/ H2A.J/ type3. H2A type 1-H/J H2B type 1 - C/E/F/G/I H3.1/3.2/3.3/3.3C H4	22 23 30 35 52
TEO_0mM_bicarbonate_1	H2A type 1-H/J H2B type 1 - C/E/F/G/I H3.1/3.2/3.3 H4	27 31 43 61
No_TEO_0 mM_bicarbonate_1	Histone H2A type 1-J/H H2A type 1 -C /D /H2A.J/ type 2A/C. H2B type 1-C/E/F/G/I H3.1/3.2/3.3 H4	23 22 18 37 51
TEO_20 mM_13C_bicarbonate_1	H2A H2B type 1-C/E/F/G/I H3.1/3.2 H4	9 31 46 55
TEO_50 mM_13C _bicarbonate_1	H2A H2B H3.1/3.2/3.1t H4	31 24 46 52

Table 8-8 Coverage of recombinant histone octamer samples modified by propionylation across the experiment conditions specified, with or without triethyloxonium (TEO) trapping, concentration of carbon 12 or 13 (12C or 13C) inorganic carbon and replicate number. If the carbon isotope is unspecified, it is 12C.

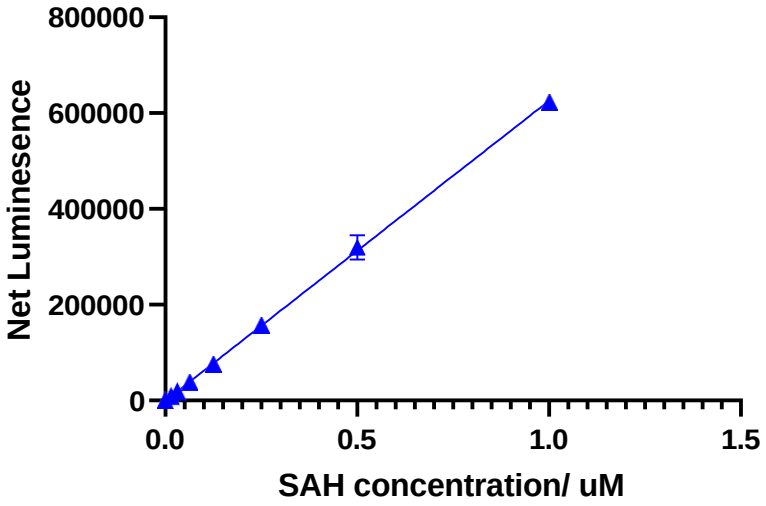


Figure 8-38 Net luminescence versus the concentration of SAH. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. Net luminescence refers to the luminescence readout minus the background luminescence at 0 µM SAH.

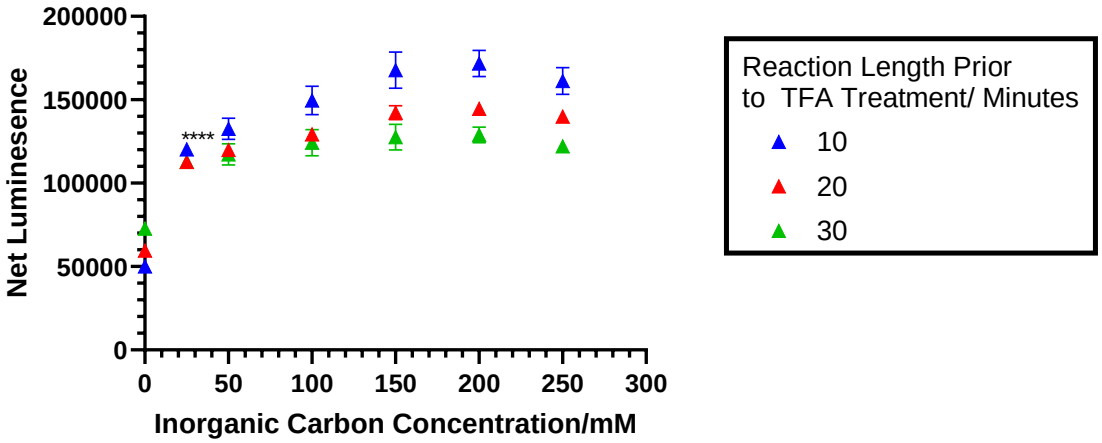


Figure 8-39 Raw luminescence produced from the methyltransferase reaction at three incubation times against varying inorganic carbon concentration. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points.

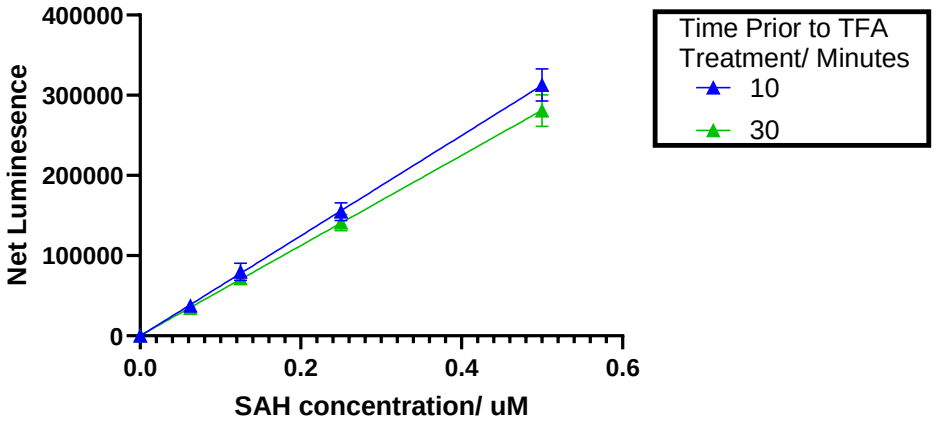


Figure 8-40 Net luminescence versus the concentration of SAH at two different incubation lengths to test SAH stability. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. Net luminescence refers to the luminescence readout minus the background luminescence at 0 µM SAH.

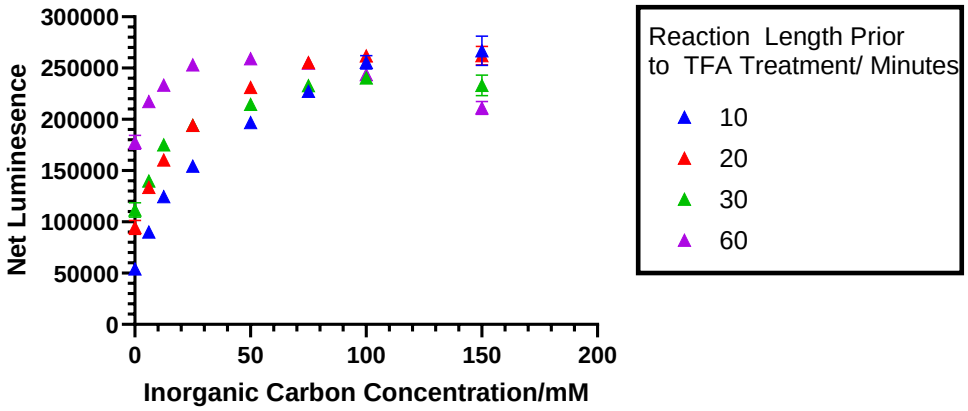


Figure 8-41 Net luminescence produced from the methyltransferase reaction at four incubation times against varying inorganic carbon concentration. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points.

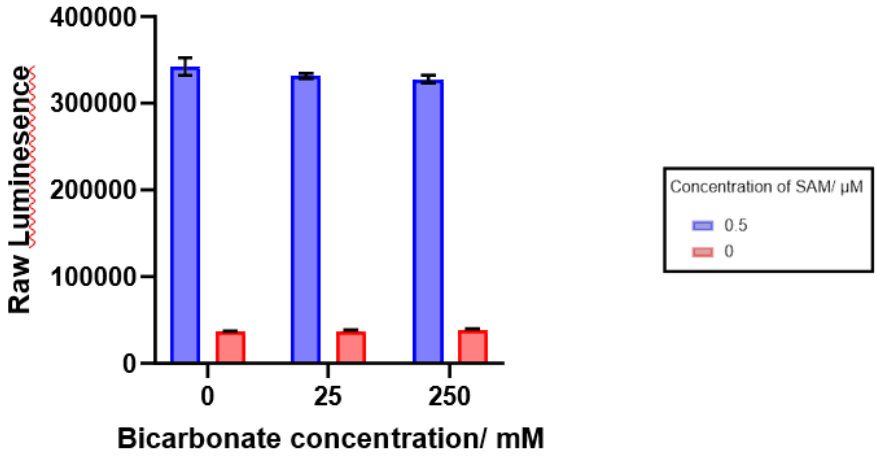


Figure 8-42 Luminescence readout versus the range of inorganic carbon concentrations (Ci) used in the assay. The red bars are the background luminescence values at 0-, 50- and 250-mM Ci. The blue bars are the readout luminescence values for 0.5 µM SAM at 0-, 50- and 250-mM Ci. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. There was no statistically significant difference in luminescence produced across the Ci concentrations for both SAM concentrations as determined by one-way ANOVA analyses.

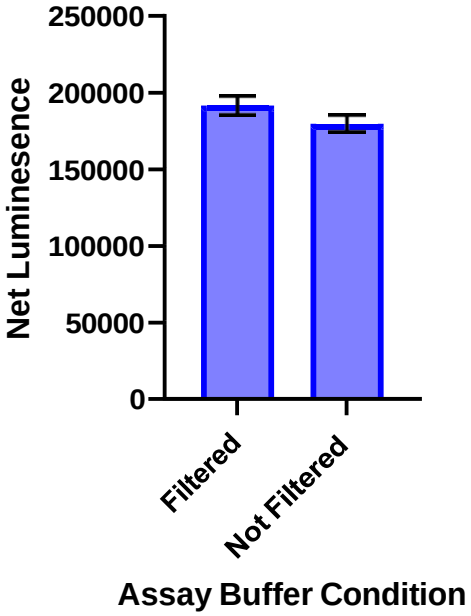


Figure 8-43 The luminescence produced versus the assay buffer condition. All values are represented as mean with error bars shown as the standard deviation where n=3. No significant difference was identified between these conditions as determined by an unpaired t-test. Where net luminescence is the luminescence readout minus the background luminescence at 0 µM SAH.

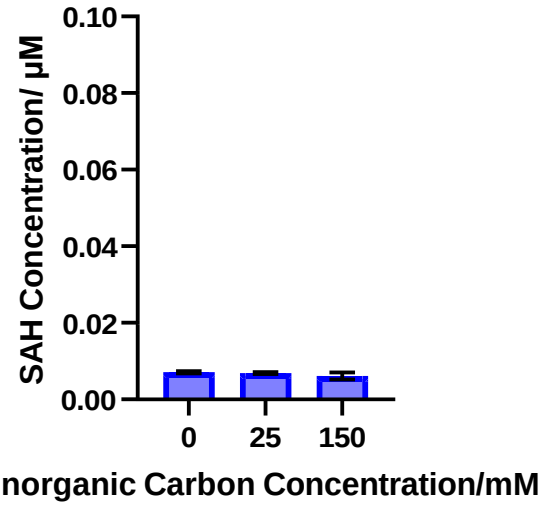


Figure 8-44 Concentration of S-adenosyl homocysteine (SAH) calculated from the luminescence readout of DOT1L incubated with S-adenosyl methionine (SAM) in the absence of the nucleosome

substrate against the inorganic carbon concentration following a 20-minute methyltransferase incubation time. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. The SAH concentration recorded is at background levels indicating there is no artificial increase in luminescence due to an unspecific ATP reaction from SAM or DOT1L under varying Ci. Bicarbonate is not causing a statistically significant artificial increase in luminescence across the inorganic carbon concentration range as determined by a one-way ANOVA analysis. Ci is not driving the conversion of SAM to SAH.

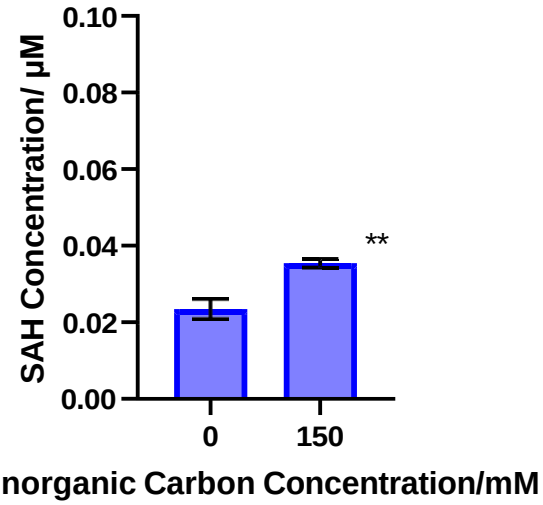


Figure 8-45 Concentration of S-adenosyl homocysteine (SAH) calculated from the luminescence readout of a methyltransferase reaction treated with SYC-522 DOT1L inhibitor versus the inorganic carbon concentration following a 20-minute methyltransferase incubation. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. The SAH concentration recorded is 10-fold lower than when all three components are active for methyl transfer activity. Bicarbonate is causing a statistically significant artificial increase in luminescence across the inorganic carbon concentration range as determined by a one-way ANOVA analysis where $P<0.002$.

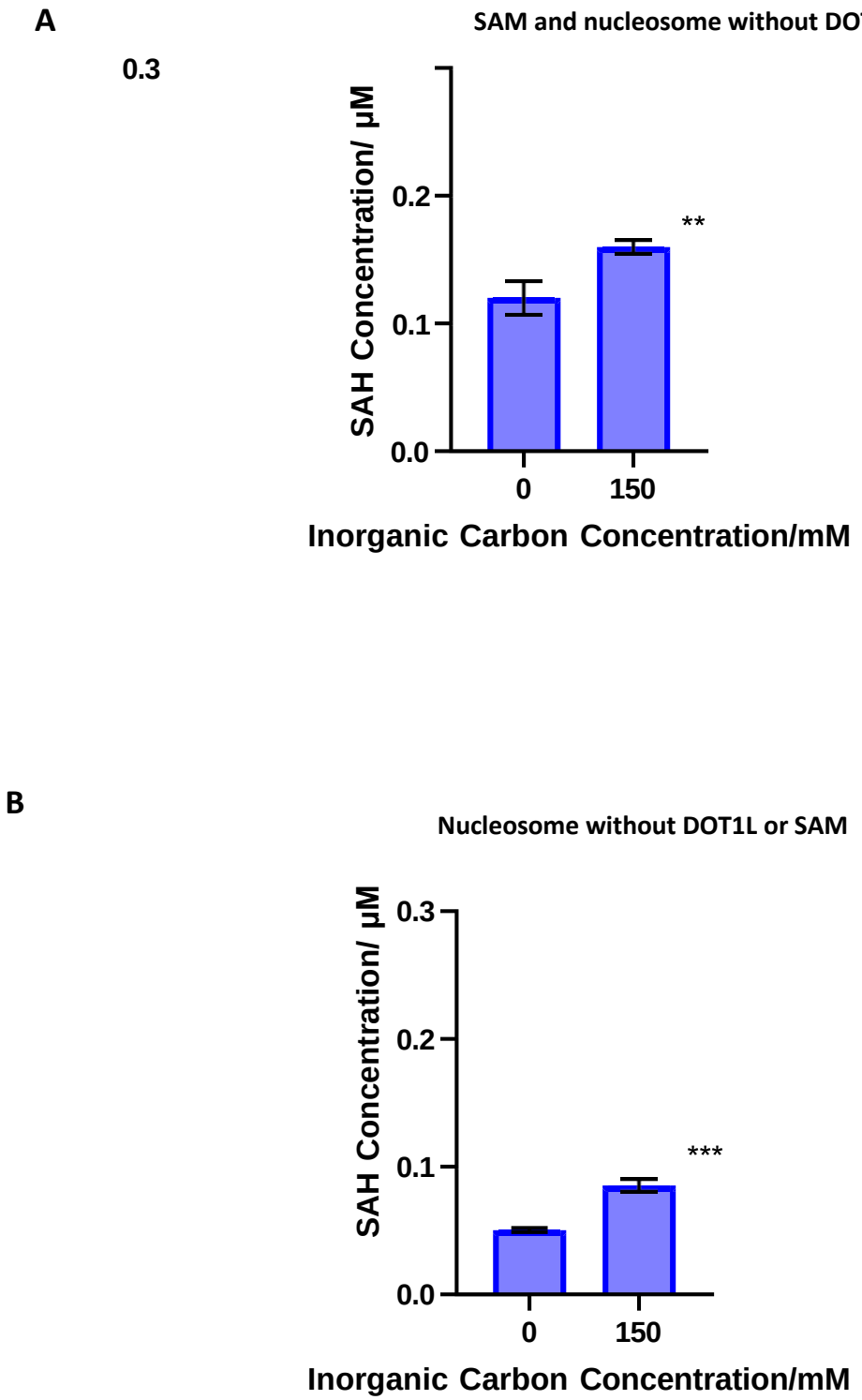


Figure 8-46 Nucleosome-dependent increase in the concentration of S-adenosyl homocysteine (SAH) calculated from the luminescence readout versus the inorganic carbon concentration following a 20-minute incubation. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. (A) SAM and nucleosome incubated together without DOT1L where there is a statistically significant

difference between SAH produced across the Ci concentration range as determined by a one-way ANOVA analysis and $p<0.002$ (B) Incubation of the nucleosome without DOT1lL or SAM where n-3 and the error bars displayed are the standard deviation from the mean and in some cases these errors are smaller than the individual data points and $p<0.0002$.

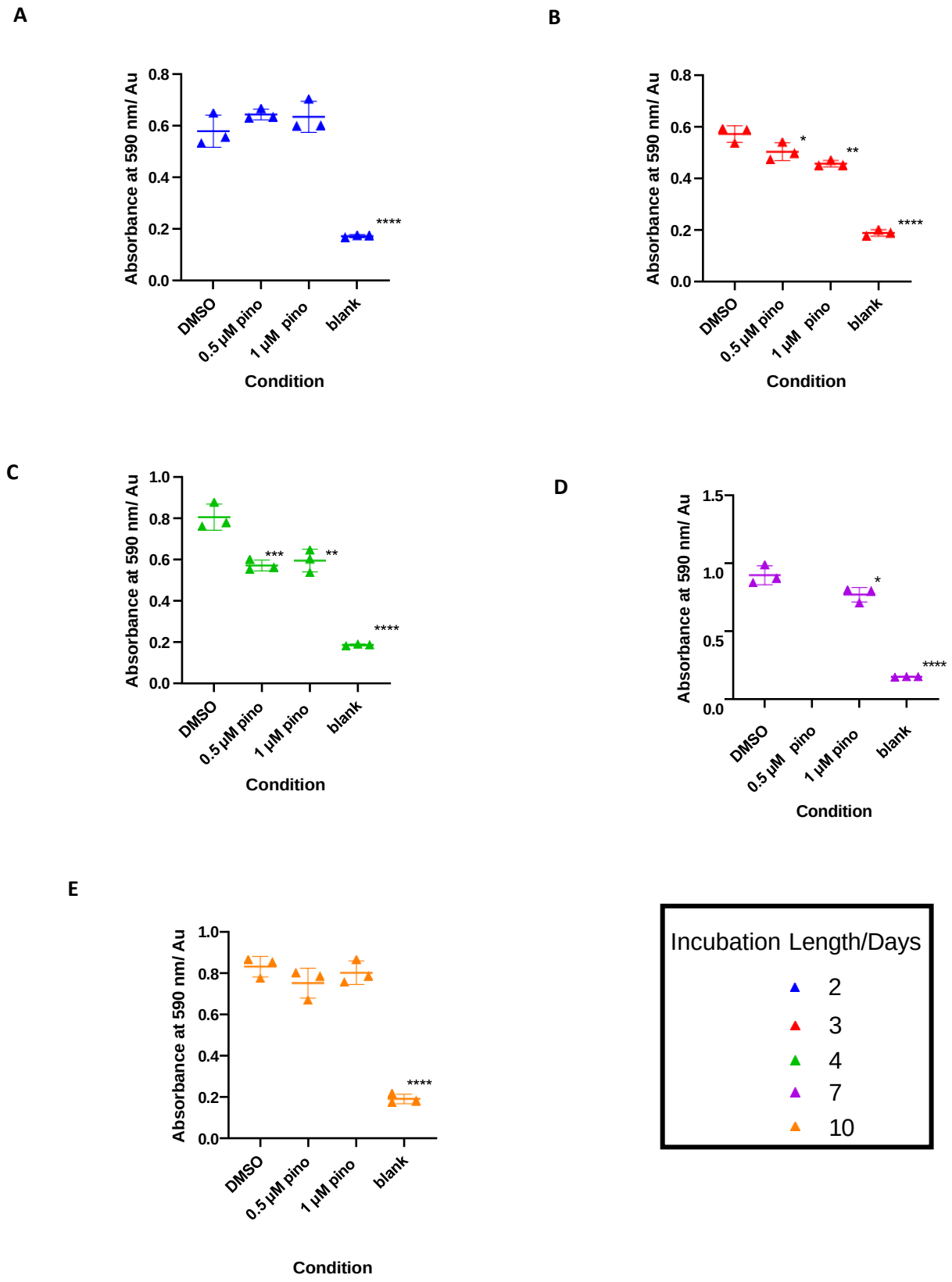


Figure 8-47 Absorbance Values obtained from an MTT assay measured at a wavelength of 590 nanometres versus the experiment treatment condition, including DMSO, and two concentrations of

pinometostat alongside the background absorbance (blank). Individual values are plotted, and the error bars represent the standard deviation from the mean where n=3. Data passed the Shapiro-Wilk normality test and was analysed by one-way ANOVA and multiple comparison tests. The significance threshold was $p<0.05$ and n=3.

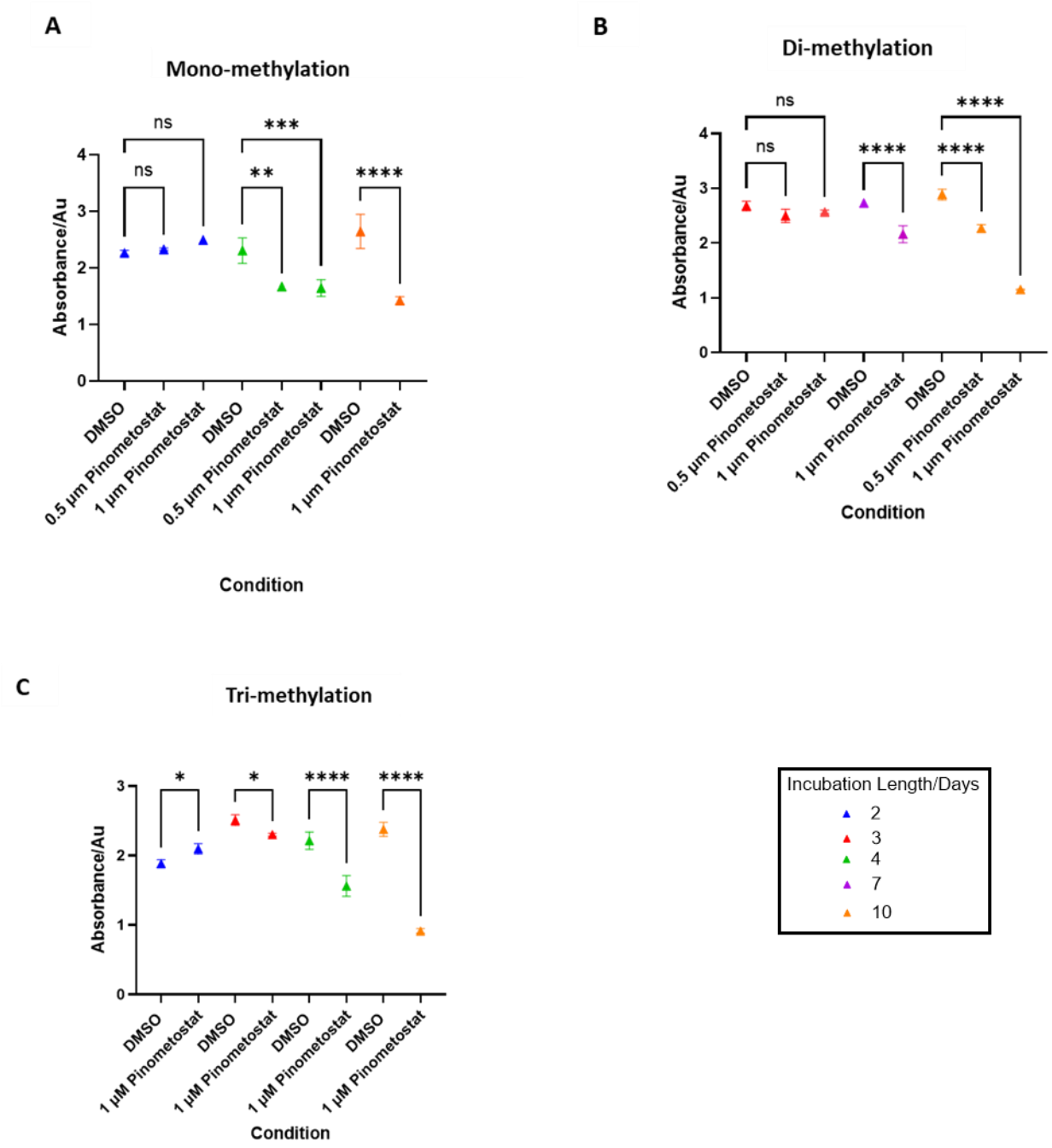


Figure 8-48 Raw absorbance measured at wavelength 450 nm for H3K79 methylation state detected versus the treatment condition of DMSO, 0.5 or 1 µM pinometostat across various incubation time

frames where (A) is mono (B) di and (C) tri methylation states. All values are represented as mean with error bars shown as the standard deviation where n=3 and in some cases these errors are smaller than the individual data points. One-way ANOVA and MCTs of n=3 were run to assess significant changes, the results of the MCTS are shown in each figure where the threshold for significance was p<0.05. Asterisks indicate levels of significance (* p ≤ 0.05, ** p ≤ 0.01, *** p ≤ 0.001).

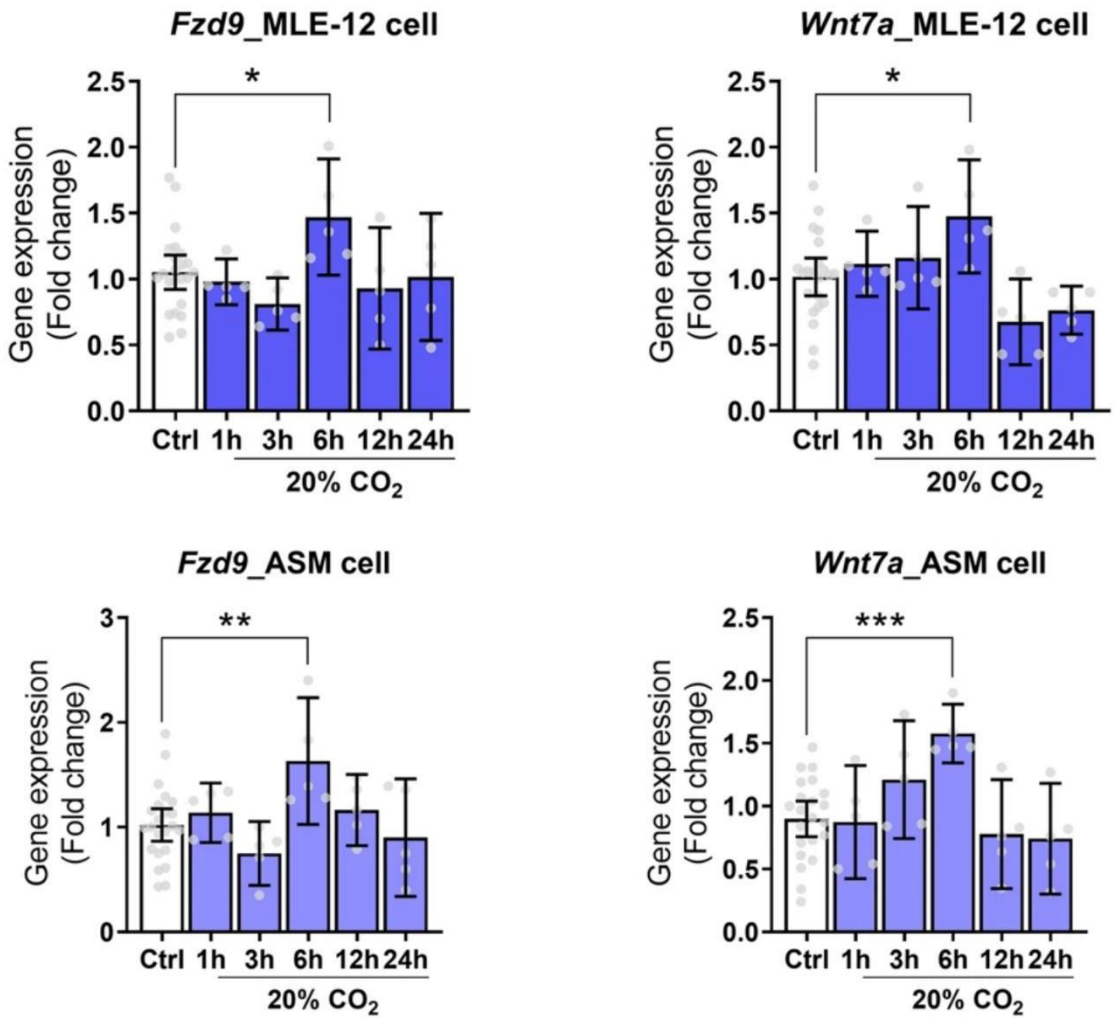


Figure 8-49 qPCR gene expression changes for *Fzd9* and *Wnt7a* in MLE-12 and ASM cell lines incubated at 20% CO_2 for various time points compared with incubation at 5% CO_2. All values are represented as mean with error bars shown as the 95% confidence interval. Statistical testing using the unpaired two-tailed Student's t-test or one-way ANOVA with Dunnett's post hoc test was performed. Asterisks indicate levels of significance (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$). Figure obtained from reference.[258]

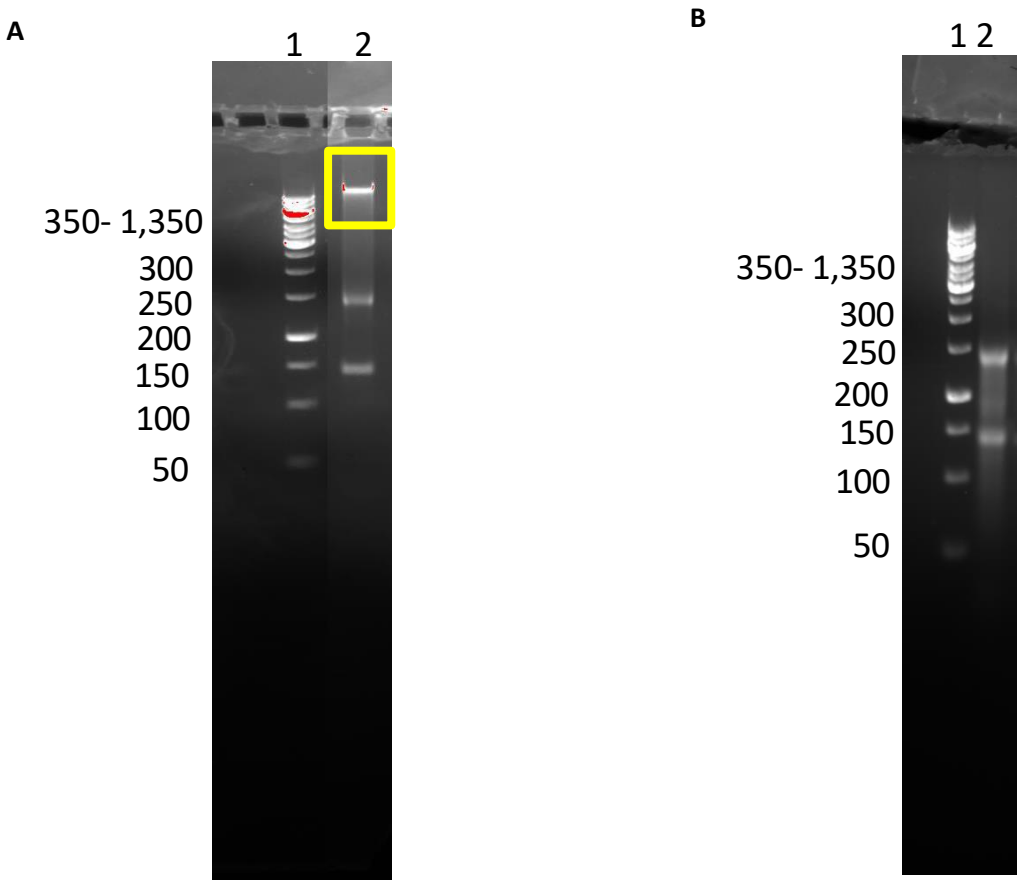


Figure 8-50 Agarose gels of extracted RNA from HEK293 cells using the Total RNA purification kit from Norgen where lane 1 is a DNA fragment reference ladder with number of base pairs specified and lane 2 is the RNA product, (A) without the on-column DNA removal kit and the yellow box indicates the presence of genomic DNA which is unsuitable for downstream analysis and (B) with the on-column DNA removal kit where only the two clear bands for 28S rRNA (~250 bp) and 18S rRNA (~150 bp) are seen.

Oligo Name	Sequence 5'- 3'
h*RPL19*_Forward_Primer_Set_1	ATGCCAGAGAAGGTCACATG
h*RPL19*_Reverse_Primer_Set_1	ACACATTCCCCTTCACCTTC
h*RPL19*_Forward_Primer_Set_2	GTATGCTCAGGCTTCAGAAGAG
h*RPL19*_Reverse_Primer_Set_2	GAGTTGGCATTGGCGATTTC
h*RPL19*_Forward_Primer_Set_3	GGTCACATGGATGAGGAGAATG
h*RPL19*_Reverse_Primer_Set_3	CTTCAGGTACAGGCTGTGATAC
h*Fzd9*_Forward_Primer_Set_1	GTTCCAGTACGTGGAGAAGAGC
h*Fzd9*_Reverse_Primer_Set_1	CAGCAAGAAGGTGAGCACAGTG
h*Fzd9*_Forward_Primer_Set_2	CTGGTCTTCCTACTGCTCTACT
h*Fzd9*_Reverse_Primer_Set_2	AGGCAGCCATGTGGAAATAG
h*Fzd9*_Forward_Primer_Set_3	CAACACAGAGAAGCTGGAGAAG
h*Fzd9*_Reverse_Primer_Set_3	GTTGAGGCGTTCGTAGACATAG
h*Wnt7a*_Forward_Primer_Set_1	GGGACTATGAACCGGAAAGC
h*WNt7a*_Reverse_Primer_Set_1	GGCCTGGGATCTTGTTACAG

Table 8-9 Primer sequences for targeting qPCR amplicons including *RPL19*, *Fzd9* and *Wnt7a*, the h prefix stands for human.

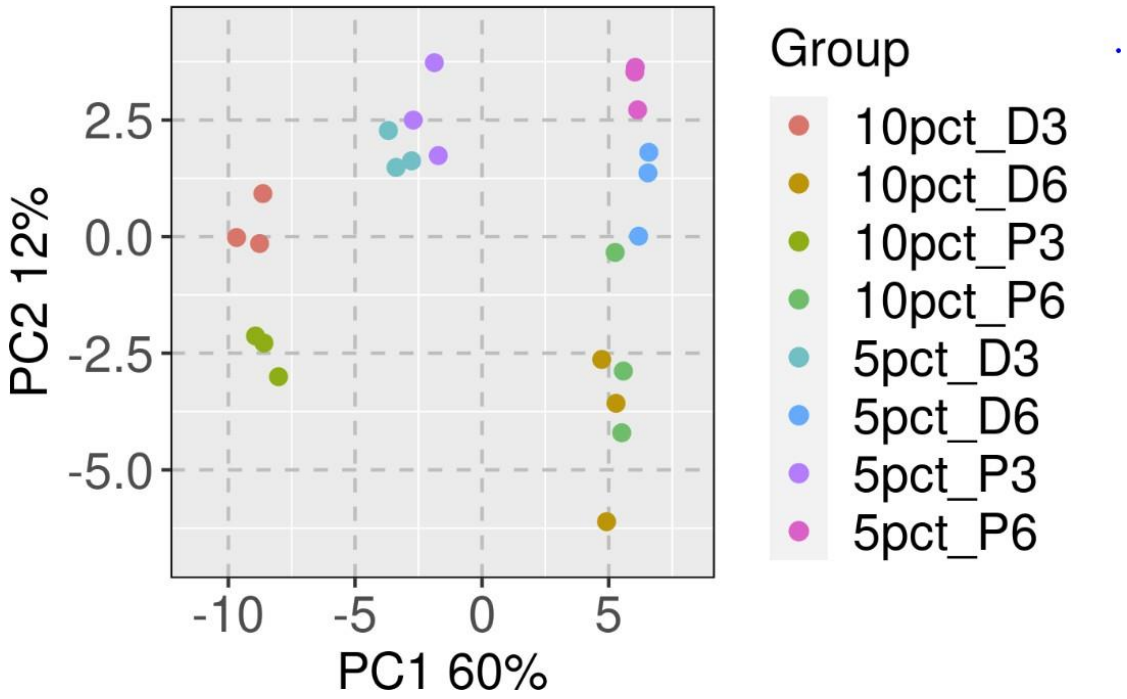


Figure 8-51 A Principal Component Analysis (PCA) plot for PC1 vs PC2 for all samples grouped into the different conditions tested in the RNA-seq dataset where each group contains three replicates.