write
read
erase
repeat

Investigating the fundamental chemical-biological interactions behind the readout process of the epigenetic histone trimethyllysine mark

Bas  J. G. E. Pieters
Write, Read, Erase, Repeat

Investigating the fundamental chemical-biological interactions behind the readout process of the epigenetic histone trimethyllysine mark

Bas J.G.E. Pieters
Write, Read, Erase, Repeat

Investigating the fundamental chemical-biological interactions behind the readout process of the epigenetic histone trimethyllysine mark

Proefschrift

ter verkrijging van de graad van doctor aan de Radboud Universiteit Nijmegen op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken, volgens besluit van het college van decanen in het openbaar te verdedigen op

vrijdag 16 juni 2017
om 10.30 uur precies

doort

Bas Johannes Gerardus Elizabeth Pieters
geboren op 7 juni 1988
te Meerssen
Promotor
Prof. dr. G.J.M. Pruijn

Copromotor
Dr. J. Mecinović

Manuscriptcommissie
Prof. dr. ir. J.C.M. van Hest (Technische Universiteit Eindhoven)
Prof. dr. H.T.M. Timmers (Universitair Medisch Centrum Utrecht)
Prof. dr. M. Vermeulen
Contents

List of abbreviations 6

Chapter 1 Introduction 11

Chapter 2 The effect of the length of histone H3K4me3 on the recognition by reader proteins 49

Chapter 3 The role of electrostatic interactions in binding of histone H3K4me2/3 to the SGF29 tandem tudor domain 71

Chapter 4 The chemical origin of the recognition of trimethyllysine by epigenetic reader domains 95

Chapter 5 Novel methyllysine analogs for investigating the chemical-biological role of lysine methylation 137

Chapter 6 General discussion and future perspectives 163

Summary in English 181

Nederlandse samenvatting 183

Dankwoord 189

CV & List of publications 195
List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-KG</td>
<td>α-ketoglutarate (2-OG)</td>
</tr>
<tr>
<td>μm</td>
<td>Micro Meter</td>
</tr>
<tr>
<td>19F-NMR</td>
<td>Fluorine-19 Nuclear Magnetic Resonance</td>
</tr>
<tr>
<td>2-OG</td>
<td>2-Oxoglutarate (α-KG)</td>
</tr>
<tr>
<td>5,6-dIF</td>
<td>Fluorinated position 5 and 6</td>
</tr>
<tr>
<td>5F</td>
<td>Fluorinated position 5</td>
</tr>
<tr>
<td>5mC</td>
<td>5-MethylCytosine</td>
</tr>
<tr>
<td>6F</td>
<td>Fluorinated position 6</td>
</tr>
<tr>
<td>a.u.</td>
<td>Absorbance Units</td>
</tr>
<tr>
<td>AA</td>
<td>Amino Acid</td>
</tr>
<tr>
<td>ACh</td>
<td>AcetylCholine</td>
</tr>
<tr>
<td>ADD</td>
<td>ATRX-DNMT3-DNMT3L</td>
</tr>
<tr>
<td>AdoMet</td>
<td>S-Adenosyl-L-Methionine</td>
</tr>
<tr>
<td>AP</td>
<td>APyrimidinic</td>
</tr>
<tr>
<td>ATRX</td>
<td>Alpha Thalasemia/Mental Retardation Syndrome X-Linked</td>
</tr>
<tr>
<td>AUX</td>
<td>Auxotrophic strain</td>
</tr>
<tr>
<td>BER</td>
<td>Base Excision Repair</td>
</tr>
<tr>
<td>BET</td>
<td>Bromodomain and Extra-Terminal motif</td>
</tr>
<tr>
<td>BL21</td>
<td>E. coli Rosetta BL21 (DE3) pLysS</td>
</tr>
<tr>
<td>BPTF</td>
<td>Bromodomain PHD finger Transcription Factor</td>
</tr>
<tr>
<td>CD</td>
<td>Circular Dichroism</td>
</tr>
<tr>
<td>CHD</td>
<td>Chromo-ATPase/Helicase-DNA-Binding</td>
</tr>
<tr>
<td>COMPASS</td>
<td>COMplex of Proteins ASsociated with Set1</td>
</tr>
<tr>
<td>CSD</td>
<td>Chromo SHadow Domain</td>
</tr>
<tr>
<td>D.I.T.</td>
<td>Data Integration Time</td>
</tr>
<tr>
<td>Da</td>
<td>Dalton</td>
</tr>
<tr>
<td>ddH₂O</td>
<td>Double Distilled water</td>
</tr>
<tr>
<td>DMSO</td>
<td>DimethylSulfOxide</td>
</tr>
<tr>
<td>DNA</td>
<td>DeoxyriboNucleic Acid</td>
</tr>
<tr>
<td>DNMT</td>
<td>DNA MethylTransferase</td>
</tr>
<tr>
<td>DOT1</td>
<td>Disruptor Of Telomeric silencing 1 (KMT4)</td>
</tr>
<tr>
<td>DOT1L</td>
<td>Disruptor Of Telomeric silencing 1 Like</td>
</tr>
<tr>
<td>DPF3b</td>
<td>D4, zinc and double PHD Fingers, family 3b</td>
</tr>
<tr>
<td>DSF</td>
<td>Differential Scanning Fluorimetry</td>
</tr>
<tr>
<td>DTT</td>
<td>DiThioTetrol</td>
</tr>
<tr>
<td>E. coli</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>EDT</td>
<td>Ethane-1,2-DiThiol</td>
</tr>
<tr>
<td>EDTA</td>
<td>EthyleneDiamineTetraacetic acid</td>
</tr>
<tr>
<td>ESC</td>
<td>Embryonic Stem Cell</td>
</tr>
<tr>
<td>ESI-MS</td>
<td>ElectroSpray Ionization Mass Spectrometry</td>
</tr>
<tr>
<td>FAD</td>
<td>Flavin Adenine Dinucleotide</td>
</tr>
<tr>
<td>FMOC</td>
<td>FluoromethoxyCarbonylChloride</td>
</tr>
<tr>
<td>FPLC</td>
<td>Fast Protein Liquid Chromatography</td>
</tr>
</tbody>
</table>
GST  Glutathione S-Transferase
H1  Histone 1
H2a  Histone 2a
H2b  Histone 2b
H2B-K123ub  Histone 2B Lysine 123 Ubiquitination
H3  Histone 3
H3K4me3  Histone 3, Lysine 4, trimethyl
H4  Histone 4
HAT  Histone AcetylTransferase
HDAC  Histone DeACetylace
HMT  Histone Methyltransferase
HP1  Heterochromatin protein 1
HPLC  High Performance Liquid Chromatography
i.e.  id est
IPTG  IsoPropyl β-D-1-ThioGalactopyranoside
ITC  Isothermal Titration Calorimetry
Jarid1A  Jumonji/ARID Domain-Containing Protein 1A (KDM5A)
JmJD  Jumonji C
Jumond2  Jumonji C domain containing protein 2 (KDM4)
Kc  Alkylated cysteine Lysine analog
kcal  Kilo Calories
KcΔme3  Alkylated cysteine TriMethylLysine Arsonium analog
KcButaneme3  Alkylated cysteine TriMethylLysine Butane analog
KcPentaneme3  Alkylated cysteine TriMethylLysine Pentane analog
KcPyridinem3  Alkylated cysteine TriMethylLysine Pyridine analog
kD  KiloDalton
Kd  Dissociation Constant
KDM  Lysine DeMethylase
KUM1  Lysine DeMethylase 1 (LSD1)
KUM4  Lysine specific DeMethylase 4 (JMJD2)
KUM4A  Lysine specific DeMethylase 4 A (JMJD2A)
KDM5A  Lysine specific DeMethylase 5A (JARID1A)
Kme0  non-Methylated Lysine
Kme1  MonoMethylLysine
Kme2  DiMethylLysine
Kme3  TriMethylLysine
Km  Lysine Methyl Transferase
Km14  Lysine Methyl Transferase 4 (DOT1)
L(3)MBTL  Lethal (3) Malignant Brain Tumor-Like protein
LB  Luria-Bertani Broth
LC-MS  Liquid Chromatography – Mass Spectrometry
LSD1  Lysine Specific Demethylase 1 (KDM1)
M  Molar
MALDI-TOF  Matrix Assisted Laser Desorption Ionisation – Time Of Flight
MAO  MonoAmine Oxidase
MBP  Maltose Binding Protein
MBT  Malignant Brain Tumor
MD  Molecular Dynamics
MLA  MethylLysine Analog
MLL  Mixed-Lineage Leukemia
MLL1  Mixed-Lineage Leukemia gene 1
MLL5  Mixed-Lineage Leukemia gene 1
MRE  Mean Residual Ellipticity
MRG  Morf-Related Gene
MWCO  Molecular Weight Cut-Off
NACHR  Nicotinic AcetylCholine Receptor
NCAA  Non-Canonical Amino Acid
NCP  Nucleosome Core Particle
Neil 1/2  Nei-like-DNA-glucosilase 1/2
Ni-NTA  Nickel - NitriloTriacetic Acid
nm  Nano Meter
NMM  New Minimal Medium
NMR  Nuclear Magnetic Resonance
NSD2  Nuclear receptor-binding SET Domain protein 2
NURF  NUCleosome Remodelling Factor
ON  Over-Night
P53  tumor Protein P53
P53BP1  P53 Binding Protein 1
PAGL  PolyAcrylamide GelElectrophoresis
Pc  Polycomb
PCH  Polymerase Chain Reaction
PDB ID  Protein DataBase IDentity
PEG  PolyEthylene Glycol
PHD  Plant HomeoDomain
PHF8  PHD Finger protein 8
pK a  Acid dissociation constant
PKMT  Protein Lysine MethylTransferase
PMT  Protein MethylTransferase
ppm  Parts Per Million
PRMT  Protein Arginine MethylTransferase
PTM  PostTranslational Modification
PWWP  Proline-Tryptophan-Tryptophan-Proline
PYGO  PYGOopus family PHD finger
RAG2  Recombination Activating Gene 2
RBBP4/7  RetinoBlastoma Binding Protein 4/7
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RING</td>
<td>Really Interesting New Gene</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root Mean Squared Deviation</td>
</tr>
<tr>
<td>RT</td>
<td>Room Temperature</td>
</tr>
<tr>
<td>SAGA</td>
<td>Spt-Ada-Gcn5 acetyltransferase</td>
</tr>
<tr>
<td>SAH</td>
<td>S-Adenosyl-L-Homocysteine</td>
</tr>
<tr>
<td>SAM</td>
<td>S-Adenosyl-L-Methionine</td>
</tr>
<tr>
<td>SASA</td>
<td>Solvent Accessible Surface Area</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium Dodecyl Sulfate</td>
</tr>
<tr>
<td>SET</td>
<td>Suppressor of Variegation–Enhancer of Zeste –Trithorax</td>
</tr>
<tr>
<td>SGF29</td>
<td>SAGA complex associated Factor 29</td>
</tr>
<tr>
<td>SRA</td>
<td>SET- and RING-Associated</td>
</tr>
<tr>
<td>SWIRM</td>
<td>Swi3p, Rsc8p, Moira</td>
</tr>
<tr>
<td>TAF3</td>
<td>TATA box binding factor Associated Factor 3</td>
</tr>
<tr>
<td>TB</td>
<td>Terrific Broth</td>
</tr>
<tr>
<td>TBP</td>
<td>TATA box Binding Protein</td>
</tr>
<tr>
<td>TDG</td>
<td>Thymine DNA-Glycosilase</td>
</tr>
<tr>
<td>TET</td>
<td>Ten-Eleven Translocation</td>
</tr>
<tr>
<td>TEV</td>
<td>Tobacco Etch Virus</td>
</tr>
<tr>
<td>TFA</td>
<td>TriFluoroacetic acid</td>
</tr>
<tr>
<td>TFIIID</td>
<td>Transcription Factor II D</td>
</tr>
<tr>
<td>TIPS</td>
<td>TriIsoPropylSylane</td>
</tr>
<tr>
<td>Tm</td>
<td>Melting Temperature</td>
</tr>
<tr>
<td>tRNA</td>
<td>Transfer Ribonucleic Acid</td>
</tr>
<tr>
<td>TRP2</td>
<td>2 Tryptophan residue</td>
</tr>
<tr>
<td>UHRF1</td>
<td>Ubiquitin-Like with PHD and RING finger domains 1</td>
</tr>
<tr>
<td>UNMM</td>
<td>Unnatural amino acid New Minimal Medium</td>
</tr>
<tr>
<td>WT</td>
<td>Wild Type</td>
</tr>
</tbody>
</table>
1

Introduction

Epigenetics:

“The study of changes in organisms caused by modification of gene expression rather than alteration of the genetic code itself.”
1.1. A general introduction to epigenetics

Even though the field of epigenetics has gained a lot of interest in the past decades, the term epigenetics itself had already been coined as early as 1942 by Conrad Waddington.[1] As is explained by the definition presented in the Oxford Dictionary: epigenetics is the study of changes in organisms caused by modification of gene expression rather than alteration of the genetic code itself. In the field of epigenetics, we try to understand exactly how the complex mechanisms that control and regulate our DNA exert their function. We know that the huge amount of information that is necessary to create and sustain a living being is locked in the genetic code, and to a large extent we have gotten very apt at reading and interpreting this genetic code.[2] We know the alphabet of genetics and how to translate this language into the functional building blocks, proteins, which make up a large part of living organisms. We also know the functions of a large portion of these proteins, why they are so important and how they do their amazing work. But what we do not know is equally (or perhaps even more?) interesting: how is the genetic code accessed and read by the cell? How does a cell know what part of the genome is important? How does it switch on and off the genes? We know that some genes that do not have an immediate function in some living cells are neatly folded and stored in the cell’s nucleus as a large archive of interesting and useful, but perhaps temporarily unnecessary information. On the other hand, the genetic information that is required for a cell’s immediate survival and function is freely accessible to the machineries that can read and use this information to keep the complex processes in a living cell running. What causes this difference?

The answer, we believe, lies in epigenetics: in the mechanisms surrounding our DNA that determine which genes are active, and when they should be active. These mechanisms are able to determine a cell’s phenotype and function, and are involved in complex but essential processes such as embryonic development. During development a totipotent cell divides into multiple cells that, eventually, are designated very specific functions such as the formation of the ectoderm, endoderm or mesoderm. This difference in cell function and phenotype is caused by the very precise regulation of gene activity within those cells. Genes that are necessary for the development of a specific cell type are switched on and are closely regulated, whereas the genes that are not involved for the development of this cell type are condensed, made inaccessible for genetic readout mechanisms.

Because epigenetic mechanisms are involved in gene regulation, it is not surprising that these mechanisms have also been found to be involved in various diseases ranging from hearing loss to cardiovascular disease, neurodegenerative disorders
and cancer.\[3-5\] This immediately implies that epigenetic processes involved in disease may be targets for drug development and in the past decades several drugs targeting epigenetic pathways have already been developed.\[6-8\] The main classes of epigenetic drugs are involved in the inhibition of DNA methylation by DNA methyl transferases (DNMTs), inhibition of histone modifications by phosphorylation enzymes (kinases), protein methyl-transferases (PMTs), histone demethylases (HDMs), histone acetyl-transferases (HATs), and histone deacetylases (HDACs), or the readout of these modifications by a variety of “reader proteins”. The list of epigenetic modifications and the proteins that regulate them in living cells is, however, much more extensive than the handful of examples listed here. Due to the myriad of different epigenetic modifications, the amount of potential druggable epigenetic targets is much larger than described above and it is expected that many more drugs targeting specific epigenetic mechanisms will be developed in the future.

Since the potential for new targets is still big and the scientific community continues their search for new druggable targets and the development of new medication, one way of developing new therapeutics could be based on the rational design of new drugs that target specific epigenetic mechanisms. A prerequisite for the rational design of novel drugs is a thorough understanding of the underlying biochemical and chemical-biological processes, however. And even though some impressive work has been done on a variety of targets, a lot of in-depth knowledge is still lacking for most proteins involved in epigenetic regulation.\[9\]

This brings us to the focus and purpose of this thesis: To gain a deeper understanding of the chemical-biological nature of epigenetics. More specifically: to understand how certain so-called “reader” proteins, specifically recognize their epigenetic targets. In the next section, epigenetic regulation is introduced in more detail, with a special focus on trimethylated lysine 4 on histone 3 (H3K4me3).

1.2. Epigenetics: acting on posttranslational modifications

Epigenetic regulation can occur on multiple levels.\[10, 11\] Here two major levels of epigenetic control will be highlighted, namely DNA methylation and histone post-translational modifications (PTM). Other levels of epigenetic regulation, such as non-coding RNA and non-canonical histones, for example, will not be addressed.\[12\] Both the DNA and the histones surrounding the DNA can be modified with a wide variety of modifications. DNA itself is known to be methylated and subsequently oxidized by the ten eleven translocation (TET) family of enzymes, and histones can
be posttranslationally modified in many different ways. Below, the DNA methylation process and the modification of histones are described in more detail.

**DNA methylation**

From a functional point of view, methylated DNA is generally associated with transcriptionally inactive genomic regions. Most of the methylation occurs in CpG islands, a region in the DNA which is rich in CG sequences, although on rare occasions it can also occur in CpH sequences where H can be either A, T or C.[13] DNA methylation is the addition of a methyl group to the C5 position of a cytosine base resulting in the formation of a 5-methylcytosine (5mC) residue (Figure 1A). These methyl groups are transferred from the methyl donor S-adenosyl-L-methionine (SAM) by enzymes belonging to the DNA methyltransferase (DNMT) family.[14]

There are two types of DNA methylation: maintenance of methylation during DNA replication, and de novo DNA methylation. DNMT1 acts primarily as a methylation state maintenance enzyme during DNA replication, where it methylates hemimethylated DNA.[15] This DNMT1 activity ensures the faithful inheritance of a DNA methylation pattern by a cell’s daughter cell after cell division. DNMT3A and DNMT3B, on the other hand, function as de novo methyltransferases.[16] DNMT2 mainly functions as an RNA methyltransferase and displays relatively weak DNA methylation activity. DNMT3L does not have any catalytic activity, but it increases DNMT3As and DNMT3Bs ability to bind to methyl groups and therefore increases DNA methylation. Additionally, DNMT3L can recognize unmethylated H3K4, and is able to recruit DNMT3A and DNMT3B to this position.[17]

In addition to the methylation of DNA, there are several ways to remove the 5mC residues from the genome (Figure 1A-B).[13, 18, 19] During certain developmental stages, for example, the DNA methylation pattern needs to be altered, which is important in generating pluripotent stem cells during embryogenesis. Demethylation can happen both passively and via active pathways. The passive loss of DNA methylation is simple passive dilution as DNA is replicated in the absence of active DNMT1. After several successive rounds of replication, the methylation status of the DNA will become more dilute and will gradually disappear upon subsequent replication.

The mammalian active demethylation pathway was not known until 2009 when the TET enzymes were found to be responsible for the conversion of 5mC to 5hmC (5-hydroxymethylcytosine, Figure 1A).[20] Research has shown that 5hmC can be further oxidized to 5fC (5-formylcytosine) and 5caC (5-carboxylycytosine), catalyzed by TET enzymes, in an Fe(II)/α-KG (alpha-ketoglutarate) dependent manner using
molecular oxygen as a substrate for oxidation.[19] The formed 5fC and 5caC can then be excised by thymine-DNA glycosylase (TDG, Figure 1B), which cleaves the N-glycosidic bond between the sugar and base yielding an apyrimidinic site (AP site). Upon removal of the modified cytosine base, TDG is displaced from the DNA strand by Nei-like-DNA-glycosylase 1/2 (Neil 1/2), which possibly reduces the risk of genomic instability upon base excision, and increases 5fC/5caC processing by TDG.[21] The AP site is then further processed by the base excision repair (BER) mechanism, leading to the incorporation of an unmodified cytosine into the DNA strand.
Histone modification

The second type of epigenetic modification discussed here is histone modification. [22] In order to pack the very long chromosomal DNA molecules into the nucleus of a single eukaryotic cell without losing the ability to transcribe the DNA, several layers of DNA condensation have evolved.[23] The first layer of DNA compaction is achieved by winding the DNA approximately 1.65 times around an octameric histone protein complex, which results in a nucleosome core particle (NCP, Figure 2A-B).

A nucleosome contains 146 base pairs of DNA and occurs approximately every 200 ± 40 bp in the eukaryotic genome.[24] The chain of NCPs assembles into a higher order structure that reduces the total length of the DNA by a factor of 30-40. The interactions of the histone octamer with the DNA strand are maintained by an extensive network of hydrogen bonds, ionic interactions, non-polar contacts and the positioning of the histone α-helices’ dipoles towards the DNA’s phosphate backbone. The histone octamer itself consists of a histone 3 - histone 4 (H3-H4) tetramer and two histone 2A - histone 2B (H2A-H2B) dimers (Figure 2C-D). The histone fold regions of the octameric structure can bind the central ~121 base pairs of DNA whereas the remaining terminal ~13 base pairs are bound by the histone tail extensions. At the nucleosomes dyad a single base pair is located which designates the nucleosomes symmetry axis.[25]

The histones that compose the nucleosome (H2A, H2B, H3 and H4) are not fully structured, however. They consist of a globular domain, making up the nucleosomes core region, but also of a linear, non-structured domain referred to as a histone tail. H2B, H3 and H4 all contribute a single N-terminal histone tail to the nucleosome, whereas H2A has both an N- and C-terminal tail, resulting in a total of 10 histone tails per NCP.[25] These histone tails protrude from the nucleosome complex and, due to their intrinsically unstructured nature, they are readily accessible for numerous post-translational modifications (PTMs) such as methylation, acetylation, phosphorylation, citrullination, ubiquitination and crotonylation.[11] These PTMs can have a wide variety of functional effects: direct effects, as is the case for lysine acetylation which causes DNA relaxation due to the removal of positive charges, and indirect effects involving the recruitment of multiple proteins and factors to a specific PTM on a specific location in the histone sequence. For example, the recruitment of TAF3 to H3K4me3 which is then capable of recruiting transcription factors and initiating transcription.[26] This notion of PTM-mediated complex downstream functions led to the postulation of the Histone Code stating that “multiple histone modifications, acting in a combinatorial or sequential fashion on one or multiple histone tails, specify unique downstream function”.[27, 28]
Three types of proteins are associated with histone PTM: writers (which add PTMs), erasers (which remove PTMs) and readers (also known as effectors, which recognize PTMs and can initiate complex downstream functions).[29] These readers, writers and erasers are often part of much larger proteins or protein complexes that generate complex downstream effects. A wide variety of specific readers, writers and erasers has been associated with the various histone PTMs that have been discovered thus far (Table 1). Acetylation is amongst the most extensively studied PTMs and this effort has already led to the development of several chemical probes and inhibitors.[6, 9]

Other well studied PTMs include methylation and ubiquitination. Of these PTMs, histone methylation will be discussed in more detail below.

**Histone lysine methylation**

Methylation is a relatively stable but dynamic PTM which can occur on lysine and arginine residues (Figure 3A-B).[33, 34] Lysine methylation can occur in three states,
INTRODUCTION

Table 1 | Readers, writers and erasers of histone methyllysine. A selection of histone methyllysine reader, writer and eraser proteins. [8, 30-32]

<table>
<thead>
<tr>
<th>Class</th>
<th>Protein</th>
<th>Domain</th>
<th>Modification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reader</td>
<td>53BP1</td>
<td>Tudor</td>
<td>H4K20me2</td>
</tr>
<tr>
<td></td>
<td>BPTF</td>
<td>PHD</td>
<td>H3K4me3</td>
</tr>
<tr>
<td></td>
<td>DNM13A</td>
<td>PWWP</td>
<td>H3K36me3</td>
</tr>
<tr>
<td></td>
<td>G9a/GLP</td>
<td>Ankyrin repeat</td>
<td>H3K9me1-2</td>
</tr>
<tr>
<td></td>
<td>HP1</td>
<td>Chromo</td>
<td>H3K9me3</td>
</tr>
<tr>
<td></td>
<td>JARID1A</td>
<td>PHD</td>
<td>H3K4me3</td>
</tr>
<tr>
<td></td>
<td>JMJD2A</td>
<td>Tudor</td>
<td>H3K4me3, H4K20me2-3</td>
</tr>
<tr>
<td></td>
<td>PHF20L1</td>
<td>MB1</td>
<td>H3K4me1, H4K20me1</td>
</tr>
<tr>
<td></td>
<td>SGF29</td>
<td>Tudor</td>
<td>H3K4me3</td>
</tr>
<tr>
<td></td>
<td>TAF3</td>
<td>PHD</td>
<td>H3K4me3</td>
</tr>
<tr>
<td>Writer</td>
<td>Dot1L</td>
<td>Non-SET</td>
<td>H3K79me1-3</td>
</tr>
<tr>
<td></td>
<td>G9a</td>
<td>SET</td>
<td>H3K9me1, H3K9me2, H3K27me1, H1K186me1, H1K25me1</td>
</tr>
<tr>
<td></td>
<td>GLP</td>
<td>SET</td>
<td>H3K9me2, H3K27me1, H1K25me1, H1K186me1</td>
</tr>
<tr>
<td></td>
<td>MLL</td>
<td>SET</td>
<td>H3K4me3</td>
</tr>
<tr>
<td></td>
<td>NSD1</td>
<td>SET</td>
<td>H3K36me2, H4K20me2</td>
</tr>
<tr>
<td></td>
<td>SETD7</td>
<td>SET</td>
<td>H3K4me1</td>
</tr>
<tr>
<td>Eraser</td>
<td>JARID1A (KDM5A)</td>
<td>JmjC</td>
<td>H3K4me2, H3K4me3</td>
</tr>
<tr>
<td></td>
<td>JMJD2A (KDM4A)</td>
<td>JmjC</td>
<td>H3K9me3, H3K36me3</td>
</tr>
<tr>
<td></td>
<td>LSD1A (KDM1A)</td>
<td>Non- JmjC</td>
<td>H3K4me1, H3K4me2, H3K9me1, H3K9me2</td>
</tr>
<tr>
<td></td>
<td>LSD1B (KDM1B)</td>
<td>Non- JmjC</td>
<td>H3K9me2</td>
</tr>
<tr>
<td></td>
<td>PHF8</td>
<td>JmjC</td>
<td>H3K9me2, H4K20me1, H3K27me2</td>
</tr>
</tbody>
</table>

namely mono-methylation (me1), di-methylation (me2) and tri-methylation (me3) and histone methylation is associated with both active and silenced genes, depending on the location of the methylated residue. For example, H3K4me3 is associated with transcriptionally active genes, whereas H3K9me3 and H3K27me3 are associated with transcriptionally inactive genes. On histone 3, the focus of this work, lysines 4, 9, 27, 36 and 79 are known to be potential methylation sites.
Figure 3] Methyllysine and methylarginine. A) Lysine (K), mono-methyllysine (Kme1), di-methyllysine (Kme2) and tri-methyllysine (Kme3). B) Arginine (R), mono-methylarginine (Rme1), asymmetric di-methylarginine (Rme2a) and symmetric di-methylarginine (Rme2s). C) Lysine methylation using S-adenosyl-L-methionine (SAM) as a methyl donor converts the substrates into monomethyllysine and S-adenosyl-L-homocysteine (SAH).
Upon methylation of a lysine residue, the methylammonium’s positive charge is spread across a larger area, increasing its cation-radius. Simultaneously, the ability of lysine to form hydrogen bonds becomes reduced and its hydrophobicity increases. Interestingly, the specific chemical properties of the methyllysine moiety change as the methylation state increases from Kme1 to Kme3. This allows for discrimination between the various methylation states by proteins that can explicitly recognize a specific methylation state.[35, 36]

Aside from lysine methylation, arginine methylation (Figure 3B) can also occur in three different states: monomethylation (Rme1), symmetric dimethylation (Rme2s) and asymmetric dimethylation (Rme2a). Similar to lysine methylation, arginine methylation also preserves the cationic charge, which is dispersed over a larger area, while hydrogen bonding capabilities are reduced and the hydrophobicity is increased.

Recently, the existence of bivalent histone modifications was proposed.[37] These are modifications that are suggested to allow a gene to be in a so-called poised state, preparing it for rapid activation and/or repression. For example, the H3K4me3 mark associated with transcriptional activation can be combined with a H3K27me3 mark, which is a known transcriptional repressor. The combination of such modifications allows for the rapid activation or inactivation of genes by removing either the activating or inactivation marks. Of these bivalent modifications the H3K4me3-H3K27me3 modifications have been most extensively studied. However, other PTMs such as H3K9/H3K14 acetylation may also have a bivalent character when paired with H3K27me3. [38-40]

**Writers of histone methyllysine**

Lysine methylation occurs sequentially by specialized writers which are able to specifically methylate lysine residues of a particular methylation state, at a defined position.[41] Enzymes capable of methylating lysine residues are grouped in the family of protein lysine methyltransferases (PKMTs), which can be subdivided into Suppressor of Variegation–Enhancer of Zeste–Trithorax (SET) domain containing PKMTs and the H3K79 methyltransferase Dot1L (disruptor of telomeric silencing 1-like).[35, 42, 43] Both SET domain containing writers and DOT1L are methyltransferases that transfer a methyl group from a SAM donor to lysine (Figure 3C).

PKMTTs have two binding sites, a cofactor binding site for the SAM donor molecule and a histone substrate binding site. Both sites are connected via a narrow hydrophobic channel where the lysine’s ε-amino group and SAM’s methylsulphonium ion come into close proximity. The methyl group is then transferred from SAM to lysine’s ε-amino group, via an S$_\text{N}$,2 reaction, to yield methyllysine and SAH.[44] The number of
rounds of catalysis on a single lysine residue varies between enzymes. SETD7 for example catalyzes only a single round of H3K4 methylation, yielding monomethyl-lysine.

G9a, on the other hand, is a HKMT responsible for mono- and di-methylation of H3K9.[45] G9a is a SET domain methyltransferase which has a conserved tyrosine residue (Y1154) that is part of the hydrophobic channel, but it also participates in the enzymes methyltransferase activity via its hydroxyl group. A second tyrosine residue (Y1067) hydrogen bonds with the methyl accepting \( \epsilon \)-amino group, bestowing the catalytic site of G9a with mono- and di-methylation specificity. By mutating the F1152 residue that is adjacent to the catalytic site to tyrosine, mono-methylation specificity can be imposed upon the G9a F1152Y mutant. Due to hydrogen bonding of the \( \epsilon \)-amino group by both Y1067 and Y1152, the nitrogen’s lone electron pair cannot be coordinated towards SAM once it has become mono-methylated, thus preventing di-methylation. Conversely by generating a Y1067F mutant, G9a is imposed with trimethylation catalytic activity as the di-methylation ammonium group is no longer hydrogen bound and the nitrogen’s lone pair can be coordinated towards SAM for catalysis.

Another well-known H3K4 methyltransferase is MLL1 (mixed-lineage leukemia), also known as KMT2A, which is a well-studied protein involved in hematopoiesis and genomic rearrangements of the MLL1 gene can result in potent chimeric oncoproteins leading to leukemia.[46] Its methyltransferase activity is antagonized by demethylases such as KDM5A-D (JARID1A-D) and LSD1, which are also involved in certain types of MLL. By aberrantly fusing the non-catalytic N-terminus of MLL1 with proteins which are part of the super elongation complex (SEC), MLL1 may recruit DOT1L to the chromatin via SEC which, in turn, may induce aberrant histone methylation.[47]

DOT1L is the only known non-SET domain PKMT, and is responsible for H3K79 methylation, which is associated with transcriptionally active genes.[33, 42] DOT1L is capable of catalyzing mono-, di-, and tri-methylation of H3K79 in a nucleosomal context. However, it is unable to methylate free H3 as binding of the DOT1L core to the nucleosome is required for its enzymatic activity. H2BK123 ubiquitination (H2BK123ub) has been shown to be required for H3K79 methylation, mediated possibly by direct DOT1L-H2BK123ub contacts or via a bridging protein and chromatin structural changes due to H2B123ub. Interestingly, DOT1L is more similar to protein arginine methyltransferases (PRMT) than to PKMT, but extensive investigations failed to demonstrate that it can catalyze arginine methylation. Recently, potent inhibitors have been developed that selectively block DOT1L methylation activity and these inhibitors are now under clinical investigation as potential drugs for the treatment of MLL-rearranged leukemia.[48]
Erasers of histone methyllysine

The first lysine demethylase (KDM) to be discovered was Lysine Specific Demethylase 1 (LSD1), now known as KDM1, which is capable of selectively demethylating H3K4me1/2. Later a series of Jumonji C (JmjC) domain containing Fe(II) 2-oxoglutarate dependent KDMs were discovered. This led to the recognition of various KDM subfamilies that have the ability to specifically demethylate mono-, di-, or tri-methylated lysine residues.[49]

KDM1 is a FAD-dependent monoamine oxidase (MAO) homolog capable of demethylating mono- and di-methylated H3K4 (Figure 4) and contains a SWIRM and amine oxidase domain, the latter being split into a substrate binding and a FAD-binding part. The catalytic site is located in between the substrate and FAD binding areas of the amine oxidase domain. At this catalytic site the methyl groups are removed resulting in the formation of formaldehyde and H2O2 via an imine intermediate. Due to the formation of an iminium ion intermediate, LSD1 is unable to catalyze the demethylation of trimethylated lysine residues, despite the fact that it is able to bind H3K4me3.[50] Its H3K4me1/2 substrate specificity is achieved by a combination of H3 sequence specificity and PTMs surrounding H3K4. K9 (methylation, acetylation), S10 (phosphorylation), K14 (acetylation), K18 (acetylation), R2 (methylation), R8 (methylation), and R17 (methylation) have been investigated in context with H3K4me1 demethylation and all of the investigated PTMs except K9 methylation resulted in a reduction in enzymatic activity.[51, 52] Further, it has been shown that H3K4me1 peptides shorter than 21 residues are less efficiently demethylated, and its amine oxidase domain is large enough to accommodate the bulky H3 tail. Additionally, the structural constraints of the catalytic site allow only 3 additional N-terminal residues to be present next to the K4me1/2 residue, further increasing N-terminal sequence specificity.

The JmjC domain KDMs are Fe(II), 2-oxoglutarate (2-OG) dependent demethylases capable of oxidatively demethylating Kme1-3 residues under the formation of formaldehyde via an hydroxymethylammonium intermediate (Figure 4).[53, 54] The first JmjC domain KDM to be discovered was JMJD2 (Jumonji C domain containing protein 2) also referred to as KDM4a. Structural studies of the KDM4a protein have shown that its catalytic site contains an asparagine or glutamine and two histidine residues which chelate the Fe(II) ion required for its catalytic activity.[54] This Asp/Glu-His-His triad has been shown to be conserved amongst other JmjC domain proteins. Variations in substrate binding sites were found in the JmjC subfamilies, which is not surprising considering the broad substrate specificity of the total JmjC domain KDMs (KDM2: H3K4me3 and H3K36me1/2; KDM3: H3K9me1/2; KDM4: H1K26me2/3 and H3K9/36me2/3; KDM5: H3K4me2/3; KDM6: H3K27me2/3).[55]
CHAPTER 1

Figure 4 | Lysine demethylation. Lysine demethylation by mono-/di-methyl KDM1 and di-/tri-methyl KDM2-7. Figure taken from Thinnes et al. BBA, 2014. DOI: 10.1016/j.bbagrm.2014.05.009
Readers of methyllysine
The posttranslational modifications placed on histone proteins, or any other protein for that matter, can function as recognition and docking sites for reader domains of other proteins or protein complexes. [56-58] Upon binding to a specific PTM, such protein complexes may facilitate complex functions such as gene transcription (e.g. as for the TFIID complex). Reader domains can recognize specific PTMs within sequence context with high degrees of affinity and specificity. The recognition domain of the reader is largely involved in generating ligand specificity: location and size of the binding pocket may induce steric hindrance whereas specific amino acid residues (charged, hydrophilic or hydrophobic) may attract or repulse certain PTMs. Due to the variety of PTMs available, a wide variety of reader domains have evolved. The amount of readers that recognize histone modifications is already quite extensive, and new readers are still being discovered and investigated. Due to the vastness of this topic, only readers of methyllysine will be described here, with a special focus on readers of H3K4me3.

Two major classes of methyllysine recognition domains exist, namely the Royal family which involves the chromodomain, tudor domain, PWWP (Pro-TRP-TRP-Pro) domain and MBT (malignant brain tumor) domain, and the second class, PHD (Plant HomeoDomain) tingers.[57] However, before the individual classes of readers are described, let us first consider an important feature which is present in readers of higher methylation states (Kme2/3): the aromatic cage.

1.3. The aromatic cage, cation-π interactions and reader domains

The aromatic cage is a structure, present in methyllysine reader domains, which is composed of several aromatic residues which flank the methylammonium moiety of the reader domain’s H3K4me3 ligand (Figure 5).[59] A cage with 4 aromatic residues is referred to as a full cage, whereas a cage flanking methylammonium on 2 or 3 sides is referred to as a half cage structure. In the case of half-cages, the aromatic cage is usually complemented by a negatively charged residue that can form electrostatic interactions with the trimethylammonium moiety, or in the case of Kme1/2 readers it can form a hydrogen bond with the methylammonium moiety. Especially in the case of readers that recognize the lower methylation states me1 and me2, hydrogen bonding, ionic interactions and steric exclusion of Kme3 play important roles. In the scientific literature it has been suggested that the interactions that are involved in the recognition of the trimethyllysine are hydrophobic and electrostatic (cation-π) in nature. To what extent these interactions contribute to the total readout,
has never been thoroughly investigated, however. Therefore, the functional properties of these aromatic cages and the contribution of cation-π to the overall biomolecular recognition are the main topic of this thesis.

In order to understand the nature of cation-π interactions in biological systems, scientists have attempted to translate fundamental studies to more biologically relevant settings. Some of the early work on translating the theoretical and gas-phase studies on cation-π interactions to a more aqueous environment has been conducted and described by the group of Dennis Dougherty.[60, 61] In short, cyclophane structures were used to study the cation-π interactions between these compounds and a variety of positively charged species in an aqueous environment (Figure 6). These experiments have shown that the cyclophanes’ hydrophobic binding site, composed of aromatic residues, can compete with the complete aqueous solvation of positively charged organic species, despite the substantial desolvation penalty that is incurred. The reason for this favorable binding can be rationalized by the fact that the used aromatic residues possess negatively charged quadrupoles that favor interactions with cationic species. The combination of this quadrupole effect with the aromaticity of the used cyclophanes allowed the binding of metal cations such as Li⁺, but also for the binding of trimethylanilinium via its trimethyl head group, at physiological pH in water.

Figure 5| SGF29 aromatic cage. SGF29’s half cage structure composed of 2 tyrosine residues, a phenylalanine residue and an aspartic acid residue. The protein is indicated in cyan, whereas the trimethyllysine residue is indicated in yellow. PDB ID: 3MEW.
In a biologically relevant context, cation-π interactions have been shown to exist between the aromatic amino acids Phe, Tyr and Trp and the basic amino acids Arg, Lys and charged His. Here cation-π can stabilize protein structures or influence side chain pKₐ, but it has also been shown to be involved in binding specific substrates. In the case of acetylcholine esterase, for example, the quaternary ammonium group of acetylcholine is located next to a conserved Trp residue which is present at the enzymes active site.[62]

Before more complex systems such as proteins are discussed, first the work done by Marcey Waters and co-workers should be considered, as they made a second leap in transferring the fundamental knowledge of cation-π to a more protein-like environment by making clever use of so called β-hairpin peptides (Figure 7).[63-66] Synthetic peptides capable of forming a hairpin structure were employed to investigate amino acid based cation-π interactions by selectively modifying paired cationic (such as lysine) and π-rich (such as tryptophan) residues. Using cyclohexylalanine as a phenylalanine analog, the effects of π-electrons of phenylalanine could be investigated in the Phe-Lys interaction. In a complementary approach norleucine was used to study the effects of the cationic state of lysine on the Phe-Lys interaction. These studies have shown that the Phe-Lys and Trp-Lys interaction in the hairpin model had both a hydrophobic and cation-π character.[63] Additional investigations expanded these studies to trimethyllysine-π systems showing that the methylated state of the lysine residue caused exceptional stability of the hairpin model when compared to the unmethylated state.[64] In order to explain these results, Waters and co-workers suggested the observed increase in interaction strength between Kme3-Trp when compared to Lys-Trp to be due to the formation of a tighter hydrophobic cluster.
between Kme3 and Trp, resulting in a stronger hydrophobic interaction. Additionally, the cation-π nature of the interaction was suggested to be slightly decreased due to an increase in lysine’s charge radius upon methylation of its ε-amine. These findings were subsequently translated into a biological context by comparing findings from the hairpin model to thermodynamic binding parameters obtained from the H3K9me3-HP1α (heterochromatin protein 1α, which contains an aromatic cage binding pocket) binding event. Here it was shown that the association was not solely hydrophobic in nature, but that the cation-π interaction was a critical component as well.[65] Despite this finding, it was at this time still unclear to what extent the cation-π interaction contributed to the Kme3 binding event of other Kme3 binding proteins and whether there are other contributions responsible for the specific recognition of Kme3. This knowledge gap largely inspired the work presented in this thesis.[67-69]

Another elegant example of cation-π interactions has been shown by Dougherty et al. who also extensively studied the nicotinic acetylcholine receptors (NACHR) gated transmembrane cation channels, involved in neuronal communication.[70] There are two distinct types of ACh receptors, namely the neuronal and muscular type. Both

**Figure 7** | β-hairpin model. This model was used to probe the charge effect of trimethyllysine on the formation of cation-π interactions between Tryptophan and trimethyllysine.
respond differently to nicotine: nicotine binds to the neuronal NACHr with much higher affinity than to its muscular counterpart. For a time, it was not known what caused this difference in affinity since both types of receptors share high degrees of homology at their agonist binding sites. A study where unnatural amino acids were incorporated into the NACHr “aromatic box” showed a marked reduction in binding affinity. Due to fluorination of a critical tryptophan residue, W149, the electronegativity of the indole ring was reduced. This resulted in weaker cation-π interactions between nicotine and the W149 residue of the neuronal NACHr, whereas the muscular receptor binding affinity remained largely unaffected. Interestingly, the aromatic box is identical for both muscular and neuronal receptors, which suggested that the reason for the difference in selectivity must lie outside the aromatic residues composing this binding site. As it turned out, a residue only 4 amino acids C-terminal of W149 was the likely cause of this difference in affinity. In high affinity receptors, residue 153 is a lysine whereas in low affinity receptors this residue is substituted by glycine. Mutating muscular G153 to K resulted in a substantial increase in nicotine binding affinity. It was suggested that the K153 residue helped shape the binding pocket in such a way that nicotine can make closer contact with W149, resulting in cation-π interactions that are naturally absent in low affinity receptors such as the muscular NACHr.[71]

**Variety in methyllysine reader domains**

As described above, the two major classes of methyllysine recognition domains are the Royal Family (Chromo, Tudor, MBT and PWWP) and the PHD fingers (Figure 8). [56, 57] These domains, and examples of proteins containing these domains, are addressed in more detail in the following sections.

**1) Chromodomains**

CHRoatin organization MOdifier domain proteins (Figure 8A-C) contain a conserved 30-70 residue domain which has first been described in heterochromatin 1 (HP1) and polycomb (Pc) proteins.[72] The majority of chromodomains is associated with H3 and H4 methyllysines. Proteins that share a high degree of sequence homology to the HP1 chromodomain are classified as canonical chromodomains, whereas chromodomain-like proteins are referred to as noncanonical chromodomains. The canonical chromodomains interact with a conserved ARKS motif usually found at H3K9/27me3 residues, although a wide variety of chromodomain targets have been described. The common mechanism of methyllysine recognition by chromodomains is achieved by a conserved Tyr-Trp-Tyr motif which constitutes a half-cage structure capable of binding Kme1-3. The residues surrounding this binding groove are involved in generating substrate specificity, as they are able to form interactions with the binding peptide. By subtly varying the amino acids which are able to interact with the peptide, sequence specificity can be generated for methyllysine marks such as
K3K9 and H3K27. Another histone PTM worth noting in the context of canonical chromodomains is the phosphorylation of Ser or Thr in the ARKS/T sequence. For chromodomains it is known that, upon S/T phosphorylation of its ligand, interactions with the chromodomain are markedly reduced, suggesting a bivalent regulatory role of S/Tphos histone modifications on chromodomain methyllysine binding.[72, 73]

Examples of non-canonical chromodomains are Chromo-ATPase/helicase-DNA-Binding (CHD) proteins which are involved in chromodomain remodeling. The Morf-related gene (MRG) family on the other hand is associated with histone acetyltransferases (HAT) and histone deacetylases (HDAC). The Chromo shadow domain (CSD) is a chromodomain that can be found C-terminal in HP1 proteins, although it has also been found to be present outside chromodomain-containing proteins. The CSD is unable to bind histones in its monomeric state, even though its structure is similar to HP1. CSD does no longer contain an aromatic cage and is only able to bind target peptides upon homo-dimerization or dimerization with a HP1 domain.

Figure 8| Examples of a Royal Family chromo domain and a PHD finger domain complexed with a H3K4me3 peptide. A-C) HP1 Chromodomain. D-F) BPTF PHD domain. PDBID: 1KNE and 2F6J for HP1 and BPTF respectively.
II) The Tudor domain

The second member of the Royal Family of readers is the Tudor domain, named after the Drosophila Tudor protein in which it was first discovered.[31] It has been shown that proteins containing Tudor domains can vary in the PTM they read as some interact with methylated lysines, whereas other proteins recognize methylated arginines. The domain itself is comprised of approximately 60 amino acid residues that form 4-5 antiparallel β-strands that form a barrel-like structure.

JMJD2A (Jumonji domain containing protein 2A, also known as KDM4A, Figure 9) was one of the first tudor proteins to be discovered and it has been found to be a H3K9me3/2 and H3K36me2/3 demethylase with a tandem tudor domain which is able to recognize H3K4me3 and H4K20me3. Remarkably, JMJD2A can recognize both methylation states via the same binding pocket and it is due to a difference in conformation of the H3 tail upon binding, which is induced by two aspartic acid residues, that a high specificity for either PTM can be achieved.[74]

SGF29 (SAGA complex associated factor 29) is another member of the Tudor family of reader domain proteins.[75, 76] This H3K4me3/2 binding, C-terminal tandem-Tudor containing protein has two Tudor domains which are located closely to one another in a distinct face-to-face orientation. Like JMJD2A, only one of SGF29s Tudor domains contains an aromatic cage structure, which is composed of four residues: Y238, Y245, F264 and a negatively charged D266 residue.

Figure 9 | JMJD2A tudor domain with H3K4me3 peptide. PDBID: 2GFA.
III) The PWWP domain

The PWWP domain is named after its semi-conserved Pro-Trp-Trp-Pro motif (Figure 10). [77] It is semi-conserved as only the fourth residue, proline, is fully conserved. Structural analysis has shown that the PWWP aromatic cage is comprised of three aromatic residues: a F/Y/W residue immediately preceding the PWWP motif, a W/Y residue which stems from the third residue of the PWWP motif and a F/Y/W residue which stems from the third β-strand of the PWWP domain. Upon binding, the histone peptides bind the PWWP domain in a structurally conserved binding groove. PWWP domains have been shown to specifically recognize H3K36me3 and H4K20me3 sequences, which have been associated with active and inactive genes respectively. Despite the general conservation of the aromatic cage, several PWWP domain proteins have been identified that do not have a complete cage structure.

An interesting feature of PWWP reader domains is that apart from being able to bind methyllysine residues, they are able to bind DNA as well.[77] Unfortunately, no crystal structures of PWWP readers complexed with DNA have been produced yet. Despite this lack of information, experimental and computational data suggests that the non-sequence specific DNA binding ability of the PWWP domain stems from a large positively charged surface on the opposite side of the methyllysine binding site in PWWP proteins.[78] Individually, binding affinities of PWWP readers for methyllysine and DNA is quite low (17 mM for H3K36me3 peptides and 150 μM for DNA in the case

Figure 10| PWWP domain protein DNMT3A. DNMT3A with “PWWP” motif (cyan) and additional aromatic cage residues (green). Note that the “PWWP” motif is actually a SWWP sequence and that second tryptophan residue of this sequence, W32, is one of the two tryptophan residues in the FWW aromatic cage structure. PDB ID: 3LLR.
of the PSIP1-PWWP). The combination, however, results in a striking increase in binding affinity (1.5 μM for H3K36me3 containing nucleosomes with PSIP1-PWWP).

Recently a H3K36me3-DNMT3B PWWP domain crystal structure has been resolved, which allowed for the proposal of a model for DNMT3A-nucleosome binding.[79] This suggested that the DNMT3A-PWWP domain binds nucleosomal H3K36me3 after which H3K4me0 is bound with its ADD (ATRX-DNMT3-DNMT3L) domain. This, in turn, leads to the methylation of a nearby cytosine. It is proposed that due to the dimerization of the catalytic domain of DNMT3A, a DNMT3A-DNMT3A link mediated nucleosomal dimer is formed that can lead to de novo DNA methylation on the linker DNA between both nucleosomes.

IV) MBT domains
Malignant brain tumor (MBT) domain proteins are reader domain proteins which recognize mono- and di-methylated lysine residues on H3 and H4 tails. [80] The MBT domain is found in Polycomb (Pc) group proteins and the L(3)MBT tumor suppressor family, from which the MBT domain got its name. L3MBTL (Figure 11) was the first MBT protein from which the crystal structure was described and it was found that the MBT domains contain both a 30-50 residue N-terminal arm motif followed by a 60-80 residue β-barrel motif. The methyllysine binding pocket is present in one of the 2-4 MBT domain repeats found in proteins. This MBT domain has a binding pocket composed of a conserved triad of Phe-Trp-Tyr residues complemented with a conserved aspartate residue, which can hydrogen bond with the Kme1-2 methylammonium function.

![Figure 11](image_url) | L3MBTL1 H4K20me2 binding MBT domain. The H4K20me2 peptide is shown in yellow. PDB ID: 2PQW.
group inducing Kme1-2 binding specificity.[81] Further specificity over Kme3 and Kme0 is generated by steric exclusion of the bulky Kme3 moiety as the aromatic cage is situated in a pocket lined with hydrophobic residues which preferentially interact with the more hydrophobic Kme1-2 residues relative to unmethylated lysine. Noteworthy is the fact that the residues flanking the Kme1-2 residue in the histone sequence do not form extensive interactions with the MBT domain, as is the case in surface-groove binding domains. This results in lower sequence specificity of these domains in vitro.

Finally: it is worthy to note that the chromo-, Tudor-, PWPP- and MBT-domains share a structurally similar chromobox motif.[72] This motif contains the aromatic residues responsible for methyllysine binding of chromodomains, although these residues are not conserved amongst the other domains and methyllysine recognition by such domains does not necessarily occur solely via the chromobox motif.

**V) Not of Royal Blood: PHD Zinc Fingers**

Plant Homeo Domain (PHD) Zinc Fingers (Figure 8D-F) can be divided in several subsets: H3K4me3 binders, H3K4me0 binders, H3K9me3 binders and H3K19ac binders of which the H3K4me3 binding subset will be described in more detail below. [82, 83] The general PHD finger structure comprises approximately 50-80 amino acid residues with low sequence conservation, although several highly conserved cysteine residues and a conserved histidine residue are present in all PHD fingers. These residues chelate two zinc ions in a cross-braced manner generally using the Cys4-His-Cys3 PHD finger motif, where the fourth and third cysteine of the PHD finger motif are involved in the zinc chelation. Despite the low sequence similarity between the various proteins, they do share a conserved fold structure which is stabilized by the chelated zinc ions. Usually the PHD domain contains a short double stranded anti-parallel \( \beta \)-sheet, one or two short \( \alpha \)-helices and several loops that connect the zinc binding regions.

In the case of H3K4me3 binding PHD fingers, which is the most predominant type, the binding mechanism appears to be quite conserved.[82] Upon binding, the histone peptide forms a third antiparallel \( \beta \)-strand which pairs with the double stranded \( \beta \)-sheet of the PHD domain. The methylammonium moiety is then inserted into an aromatic cage which can range from a full cage structure, as is the case for Bromodomain PHD finger Transcription Factor (BPTF), to a minimal two tryptophan half cage structure as can be seen for Jumonji/ARID Domain-Containing Protein 1A (JARID1A also known as KDM5A).[84, 85] An intermediate structure can be seen for IA1A box binding protein Associated factor 3 (IAf-3) which contains a Met, Irp, Irp aromatic cage where the second tryptophan functions as a “lid” on the cage.[26]
INTRODUCTION

All cages have a single residue in common: a conserved tryptophan residue, which separates the H3K4me3 binding pocket from the adjacent H3R2 binding area (Figure 8D-F). Here, the H3R2 guanidinium moiety interacts with less conserved Asp, Gln or Glu residues. The aforementioned tryptophan residue serves to generate additional sequence specificity as it prevents binding of H3K9me3 and H3K27me3 sequences. Indeed, deletion of T3 in the ARTKme3QTARKS histone 3 sequence prevents binding to the PHD domain. Additionally, the N-terminal alanine, H3A1, is coordinated towards one or several backbone carbonyls which stem from the PHD finger loop regions. Together, the K4me3 binding pocket, H2 binding groove, A1 coordination and the conserved Trp residue in the aromatic cage appear to bestow sequence selectivity upon the PHD Finger domain.

It is noteworthy that specificity for H3K4me3 generally decreases as H3R2 becomes methylated. There are only few cases where R2me does not negatively affect binding affinity for H3K4me3, example are RAG2 (Recombination Activating Gene 2) and PYGO (PYgopus family PHD finger). Here PYGO is completely unaffected by R2 methylation as its R2 binding area is no longer functional, whereas a tyrosine residue which is present in the R2 binding groove of RAG2 bestows a minor preference for R2me2s over R2 or R2me2a.

The other subgroups of PHD fingers, which include Kme0 and Kac readers, are beyond the scope of this introduction and have therefore been omitted. Briefly, H3K4me0 binding PHD fingers do not have an aromatic cage. Instead they have a binding pocket composed of acidic and hydrophobic residues that interact with the K4 residue. In the case of H3K4 methylation, binding is abrogated, bestowing the readers with H3K4me0 selectivity. This immediately underscores the importance of an aromatic cage in the readout of trimethyllysine residues.

1.4. Crosstalk between epigenetic marks

After describing a variety of PTMs and the proteins that write, read and/or erase such modifications, it is time to place them in a broader context. The three domains, readers, writers and erasers, generally do not function as individual entities.[29, 56, 86] Instead many proteins contain combinations of modules that work in concert in order to achieve complex downstream functions.[87] Reader domains may be linked to other readers, to writers or to erasers in order to facilitate a downstream effect mediated by a specific PTM (or in some cases, in the absence of such a modification). The PTMs may then work together in combinatorial or mutually exclusive ways. Tandemly linked reader modules, for example, may work together to achieve higher
affinity and specificity. By these multivalent mechanisms, PTMs that are associated with transcriptional repression, for example, may recruit a reader-effector protein to facilitate downstream functions such as the methylation of DNA.

The simplest mode of multivalence, with regard to histone PTMs, is the binding of multiple PTMs on the same histone tail (Figure 12), also referred to as cis binding.[29] An example of this is the DPF3b (D4, zinc and double PHD Fingers, family 3b) protein, which contains 2 tandem PHD fingers and is able to multivalently recognize H3K4 and H3K14ac.[88] Mechanistically: the first PHD finger domain contains the K14ac binding module, whereas the second PHD finger contains binding sites for H3A1,

H3R9 and unmethylated H3K4. Together they impart the DPF3b tandem PHD finger with binding affinity ($K_d$) of approximately 0.6 μM whereas the affinity for the unacetylated peptide was only 2.3 μM. Interestingly, this binding affinity was decreased upon methylation of H3K4, indicating a cooperative mode between the two types of PTM.

Another possibility of histone PTM binding is the trans mode where two PTMs that are present on two different histone tails may be recognized. BPTF contains a C-terminal PHD-bromo domain cassette which is capable of binding H3K4me3 in the aromatic cage of its PHD domain. Its bromodomain is able to bind the H4K16ac PTM which, in combination with the PHD finger bound H3K4me3, results in increased binding affinity for H3K4me3-H4K16ac containing nucleosomes.[89] This is a valuable finding as BPTF is a member of the NURF (NUcleosome Remodeling Factor) complex which is involved in transcriptional activation.

More complex is the multivalent readout of PTMs in combination with DNA.[88] UHRF1 (Ubiquitin-Like with PHD and RING finger domains 1) is implied to be involved in targeting DNMT1 to replication forks and contains N-Terminal tandem Tudor domains and a PHD finger, plus a C-terminal SRA (SET- and RING-Associated) domain which are of interest with regard to its DNA-PTM reading ability. The protein’s SRA domain is able to recognize hemimethylated DNA whereas the PHD finger binds the N-terminal H3 tail via H3R2 recognition. The tandem Tudor then binds H3K9me3 via an aromatic cage structure.

As mentioned earlier, epigenetic readout is not the only case of epigenetic crosstalk. Certain proteins facilitate writing or erasing of PTMs based upon the presence or absence of other PTMs. An example is the MLL1 protein which can generate a H3K4me3 mark which can be subsequently read by one of its PHD fingers.[88] Another of such mechanisms belongs to the lysine demethylase PHF8 (PHD Finger protein 8).[90] PHF8 contains an N-terminal PHD finger capable of reading H3K4me3 upon which its C-terminal Jumonji domain demethylates H3K9me2 or H3K27me2. It has been shown that binding to the H3K4me3 activation mark increased the proteins enzymatic activity greatly.

**Protein complexes**

Of course, even individual proteins that contain many different domains often do not work alone. The molecular biology of the cell is complex and the effects caused by one protein will directly or indirectly influence its environment. Complex clusters of different proteins have evolved which work together in order to generate complex and daunting downstream functions within a cell.
The NURF complex is an example of such a protein complex.[91] From experiments with D. melanogaster NUH1 it has been shown that this complex is able to move nucleosomes in 10bp steps along the DNA strand.[92] The complete mechanism function of the complex is still unknown, however it is an evolutionary conserved complex, that can be found in almost all eukaryotes and has been implicated in the regulation of higher order chromatin. In H. sapiens the complex is composed of BPTF, SNF2L and RBBP4/7. Of these proteins BPTF and SNF2L are believed to interact with the DNA and various histone PTMs whereas RBBP4/7 have been shown to directly interact with H4.

COMPASS (COMplex of Proteins ASsociated with Set1) and COMPASS-like complexes are KMT complexes which often function as coactivators of nuclear receptor transcription factors.[93] In yeast a single COMPASS complex has been identified, vertebrates however have a variety of six COMPASS and COMPASS-like complexes. The complexes have a set of four conserved core subunits and a variable set of complex specific subunits. Of these COMPASS complexes the two Set1A-B/COMPASS complexes are responsible for the majority of H3K4 di- and tri-methylation while the remaining four COMPASS-like KMT2A-D complexes are mainly associated with mono- and di-methylation of genes associated with development. In mammalian organisms the KMTs present in the four COMPASS-like complexes are MLL protein family members. Besides the SET1 and KMT domains which are responsible for the complex's main catalytic activity, the surrounding subunits are involved in localization of the complex, increasing the complex's catalytic functionality and function as erasers. Together a total minimum of five core subunits are required for accurate functioning of the complex in humans.

The SAGA (Spt-Ada-Gcn5 Acetyltransferase) complex is a large multi-subunit complex with both acetyltransferase and deubiquitination activity composed of 21 subunits in yeast.[94, 95] The complex's subunits are widely conserved across species and are involved in activation of transcription and promoting transcription elongation. In yeast approximately 10% of all genes are activated by the SAGA complex although most of its genes can be activated by SAGA if TFIID (Transcription Factor II D) is not expressed. Without going in too much detail, the SAGA complex can activate genes after it becomes recruited by transcriptional activators upon which the bromodomain of yGcn5 binds the acetylated histone N-termini, initiating the histone acetyltransferase activity of the SAGA complex. The ySpt3 subunit subsequently recruits TATA Box Binding Protein (TBP) in order to aid the assembly of the preinitiation transcription complex.
Lastly the nucleosome will be discussed (Figures 2 & 13). As has been described before, the nucleosome is a structure composed of a 146 bp DNA strand wrapped around a histone octamer consisting of a H3/H4 tetramer and two H2A/H2B dimers. As the histones that make up the nucleosome have unstructured histone “tails” that protrude from the nucleosome assembly these tails are accessible for PTM in various ways. Of course the nucleosome itself is also affected by these PTMs. It has been described, for example, that di-methylation of H3K79 has an effect on the nucleosome’s structure. By chemically modifying recombinant H3 using cysteine chemistry, a useful technique pioneered by Shokat and co-workers, the site specific incorporation of the dimethyllysine mark (H3K₇⁹me₂) was achieved. This allowed for the assembly of designer nucleosomes core particles (NCP) containing the H3K₇⁹me₂ mark. Upon the refinement of NCP crystal structures, it was shown that the di-methylation of K₇⁹ causes the K₇⁹me₂ residue to adopt an alternate conformation which results in the partial uncovering of a small hydrophobic pocket. This change results in the reshaping of the nucleosome surface near the C-terminal end of alpha-helix 1 (a1) of H3.

Other PTMs that affect the nucleosome are the trimethylation of H4K20 and H4K16 acetylation, as they appear to have a marked impact on chromatin compaction. H4 serves a crucial role in the condensation of nucleosome arrays (a short series of linked nucleosomes forming a bead on a string structure) as was proven when proteolysis of the H4 tail prevented the condensation of such an array. The H4 tail is capable of interacting with neighboring nucleosomes by binding to the nucleosomes acidic patch, a region composed of 6 H2A and 2 H2B acidic residues which form a negatively charged groove on the nucleosome. The acidic patch is further complemented by a series of 3 hydrophobic residues lining the bottom of the groove. Due to the basic nature of the H4 tail, H4 residues 16-24 are able to bind to the acidic patch. When H4K16 is acetylated, however, a strong ionic bond between the H4 tail and the nucleosome is abrogated which has been shown to abrogate chromatin condensation. Additionally, the trimethylation of H4K20 has been shown to enhance the condensation of nucleosome arrays and has been correlated with transcriptional repression. Due to the roles of both H3K16Ac and H4K20me3 in chromatin function it is not surprising that a reduction in both PTMs has been associated with human cancers.
CHAPTER 1

1.5. Aim of this thesis

The previous paragraphs illustrate the importance of the regulatory mechanisms behind the epigenetic language. The proteins and macromolecular interactions responsible for the proper functioning of critical cellular mechanisms, such as gene activation, rely on the proper placement and readout of many epigenetic marks. Even minor changes to the smallest P1M associated protein domain could have a huge impact on the large and complex environment in which the protein functions. This is why we
INTRODUCTION

need to understand, in detail, how the protein-PTM interactions are driven. By understanding how the fundamental interactions between PTMs and their associated reader, writer and eraser modules work, we will glean valuable information. This will allow us to understand the complete picture of epigenetic regulation and the biochemistry of a living cell more accurately. It will give us the knowledge to explain the biochemistry of life, but also to heal it, in the event where the complex mechanisms in our cells go awry. A thing which unfortunately happens all too often.

This is why the aim of the work described in this thesis was to elucidate the chemical basis of the interaction between trimethyllysine containing histones and readers of trimethyllysine. Here, recombinant H3K4me3 reader domains were used in combination with H3-related molecules, ranging from relatively short peptides to the complete histone protein and the more complex nucleosome core particle. The central question that was addressed in this thesis is therefore:

*What are the chemical mechanisms behind the recognition of trimethyllysine by epigenetic reader domain proteins?*
1.6. Outline of this thesis

Chapter 1 | Introduction

Chapter 1, the introduction of this thesis has been written in order to provide the reader with the basic concept of epigenetics, followed by an in-depth overview of the epigenetic mechanisms that were studied in this thesis: the recognition of trimethylated histone 3 lysine 4 by so-called reader domain proteins.

Chapter 2 | The effect of the length of histone H3K4me3 on the recognition by reader proteins

Chapter 2 describes the effects of variation in the length of the histone 3 peptide that was used for binding experiments with reader domain proteins on the protein’s affinity for H3K4me3.

Chapter 3 | The role of electrostatic interactions in binding of histone H3K4me2/3 to the SGF29 tandem tudor domain

Chapter 3 addresses the role of the aspartic acid residue, which is present in SGF29’s aromatic cage, in its binding to di- and trimethylated histone 3 at lysine 4 (H3K4me2 and H3K4me3). By mutating the aromatic cage, insight has been obtained in the way H3K4me2/3 is recognized by the SGF29 reader domain.

Chapter 4 | The chemical origin of the recognition of trimethyllysine by epigenetic reader proteins

Chapter 4 addresses the question “how is trimethyllysine recognized by reader proteins?” from the most fundamental point of view. The involvement of cation-π interactions in the methyllysine-reader protein binding event is investigated on an atomic level, by chemically modifying the methylammonium group of the methyllysine ligand and by modifying the JARID1A reader domain by incorporating unnatural amino acids into its aromatic cage.
Chapter 5 | Novel methyllysine analogs for investigating the chemical-biological role of lysine methylation

Chapter 5 describes new tools that can be used to study cation-π interactions using full length histone proteins and even whole nucleosomes. A set of novel methyllysine analogs is presented which can be selectively introduced at defined positions of histone proteins using cysteine-based chemistry. In addition, it is shown that these histones can be assembled into the octameric histone core.

Chapter 6 | General discussion

Chapter 6 places the work presented in chapters 1-5 into a broader context, discusses the major findings in an epigenetic context and provides perspectives for future work.
1.7. References

2

The effect of the length of histone H3K4me3 on the recognition by reader proteins

B.J.G.E. Pieters
R. Belle
J. Mecinović

CHAPTER 2

Abstract

Objective: Short histone 3 peptides are often used for experiments with proteins that recognize trimethylated lysine on histone 3 (H3K4me3). In this chapter it is investigated what the effect of the length of such peptides on recognition by a selection of H3K4me3 reader domains is.

Approach: H3K4me3 reader domains of JARID1A, BPTF, JMJD2A, TAF3 and SGF29 were recombinantly expressed and purified. Additionally, ten peptide variants which mimic the N-terminal region of H3 (Xenopus laevis) have been synthesized by solid phase peptide synthesis. Variations on peptide length (3-15 AA) and charge of the peptides C-terminus have been used to identify the effect of such modifications on peptide recognition by the H3K4me3 reader domains, using isothermal titration calorimetry (ITC) to obtain a complete thermodynamic description of the reader-peptide binding event.

Results: ITC has shown that the shortest peptide that was readily recognized by the five readers was a 4-mer peptide: ARTKme3. Interestingly, peptides bearing a C-terminal amide functionality were recognized with higher affinity than their naturally occurring C-terminal carboxilic acid counterparts. The N-terminal A1 residue proved to be crucial in the efficient recognition of the H3K4me3 peptides by BPTF, SGF29 and TAF3 due to the presence of a A1 binding pocket. Both the deletion of A1 and the incorporation of an N-terminal glycine, G0, resulted in reduced binding affinities for all readers except JMJD2A

Conclusions: The 4-mer ARTKme3 is recognized by the five readers with affinity equal to their 15-mer counterparts and can therefore be used as a H3K4me3 mimic in reader-H3K4me3 binding studies.
2.1 - Introduction

In eukaryotic cells, DNA is wrapped around histone proteins. The flexible, unstructured histone tails protruding from these histone proteins undergo extensive post-translational modifications (PTMs), which in turn have been associated with the regulation of eukaryotic DNA function and structure.[1] Among them, histone lysine methylation has been found in transcriptionally active as well as silent regions of genome, depending on the position and type of modification. Histone lysine methylation is a dynamic process that is regulated by three sets of proteins: i) writers (histone methyl-transferases) that install methyl groups on lysine, ii) erasers (histone demethylases) that remove them, and iii) readers that recognise methylated lysines.[2–4]

A diverse set of conserved protein domains, including the double chromodomain, double/tandem tudor, and PHD finger, have been functionally and structurally characterised to recognise trimethylated lysine 4 in histone 3 (i.e. H3K4me3). All reader domains possess an aromatic cage consisting of 2–4 aromatic amino acids for the binding of trimethyllysine. In addition to the trimethyllysine binding site, readers of H3K4me3 require the appropriate sequence around Kme3, in order to achieve high specificity, which affects the binding of H3K4me3 via the following mechanisms: i) induced β-sheet intermolecular hydrogen bonding between reader modules and histone peptides, ii) an additional binding pocket for terminal residues in histone 3, and iii) the presence of tryptophan that separates K4me3 and R2, and generates specificity for H3K4me3 over H3K9me3.[5] Studies on the methylation-state selectivity demonstrated that readers of H3K4me3 typically recognise histone 3 in the order H3K4me3 > H3K4me2 > H3K4me > H3K4.[6] Recent investigations, however, have not addressed the question: how does the length of the histone tail affect the biomolecular recognition by reader domain proteins?. Herein, we present thermodynamic analyses on the influence of the length of the H3K4me3 histone tail on the interactions with reader domain proteins, which reveal that readers of H3K4me3 require only the first four amino acids of histone 3 in order to achieve a strong recognition.

In order to study the effect of the length of histone H3 on the recognition by readers, histone peptides containing the natural sequence of H3 were synthesised via solid-phase peptide synthesis employing FMOC-protected amino acids. We have chosen the 15-mer H3K4me3 peptide (residues 1-15; ARTKme3QTARKSTGGKA) as the longest histone mimic, because biochemical and biophysical studies have shown that histone peptides that contain more than 10 amino acids are efficiently recognised by reader domain proteins. Shorter analogues of H3K4me3 included the 10-mer, 5-mer, 4-mer, and 3-mer peptides, all starting with the first amino acid, A1, in histone 3. Five representative reader domain proteins were selected, namely: BPTF,[7] JARID1A,[8]
and TAF3[9] that belong to the PHD finger family, and JMJD2A,[10] and SGF29[11] that contain tudor domains and belong to the Royal family. These readers were chosen for the following reasons: i) their crystal or solution structures in the presence of H3K4me3 have been solved (Figure 1), ii) they recognise H3K4me3 with binding affinities (Kd < 10 μM) that can be measured by isothermal titration calorimetry (ITC), iii) they can be readily expressed in E.coli, and iv) they provide an opportunity to compare two different subfamilies of readers of trimethyllysine, namely PHD fingers and tudor domains.

Figure 1: Reader-H3K4me3 complex as presented by the solution structure of TAF3-H3K4me3. Among all five reader proteins investigated in this study, TAF3-H3K4me3 structure is the only one that includes positions of residues 9-13 in histone 3. TAF3 and H3K4me3 (ARTTAKSTGG) are shown in dark gray and light gray, respectively. Zinc ions are shown as spheres.

2.2 - Results & Discussion

Thermodynamics of the interactions between readers and histone peptides have been measured by ITC, which provided full thermodynamic descriptions of the binding events (i.e. dissociation constant Kd, free energy of binding ΔG⁰, enthalpy of binding ΔH⁰, entropy of binding -TΔS⁰, Table 1). In agreement with previous studies, the dissociation constants for binding of the 15-mer histone peptide to readers were observed to be between 0.24 and 2.0 μM (i.e. it gives ΔG⁰ = -7.7 – -9.0 kcal mol⁻¹). [7–11] Calorimetric data on the 15-mer histone peptide also demonstrated that its
recognition by all five readers is enthalpy-driven ($\Delta H^0 = -7.9 - -12.2 \text{ kcal mol}^{-1}$), while the entropy of binding is slightly unfavourable ($-T\Delta S^0 = 0.2 - 4.2 \text{ kcal mol}^{-1}$). The 10-mer peptide binds to readers with indistinguishable thermodynamic signature relative to the 15-mer peptide. These results indicate that residues 11-15 in histone 3 do not play any significant role, i.e. do not directly interact with the reader domain proteins, in the complexation event.

Further thermodynamic analyses of shorter histone 3 peptides revealed that the 5-mer peptide (i.e. ARTKme3Q) bound with a similar affinity as the 15-mer analogue for all five readers that have been tested. Enthalpy-entropy compensation was observed for binding of the 5-mer peptide to reader proteins: $\Delta H^0$ becomes less favourable, whereas $-T\Delta S^0$ more favourable. This result can be rationalised by the lack of residues 6-10 that are potentially involved in non-covalent stabilisation of the complex and the fact that shorter peptides are entropically less restricted than longer peptides upon the formation of reader-histone complex. These ITC-based results are in agreement with crystallographic analyses of reader-histone complexes that illustrate that residues T6 and A7 are within van der Waals distance to the surface of readers and could be involved in the recognition event. The shortest histone 3 peptide that was observed to bind (Kd < 7 $\mu$M) to all five reader proteins was 4-mer ARTKme3 that contains a C-terminal amide (Figure 2). In all cases except TAF3, enthalpy was found to be less favourable and entropy more favourable when compared to the 15-mer analogous peptide.

Interestingly, TAF3 and BPTF formed stronger complexes with 4-mer than with 15-mer H3K4me3, with the value of incremental free energy of binding $\Delta\Delta G^0 = -0.6 \pm 0.1 \text{ kcal mol}^{-1}$. Comparison of the thermodynamics of interactions of ARTKme3 peptides that bear C-terminal amide and acid functionalities, respectively, show that readers recognise the former with slightly higher or equal affinity than the latter, possibly due to different hydrogen-bonding properties between the acid and amide functionalities. Surprisingly, JMJD2A did not bind ARTKme3 with the terminal COOH group. As expected, none of the readers formed a complex with histone 3 peptide ART that contains only the first three amino acids, hence verifying the essential contribution of trimethyllysine at the position 4. Overall, calorimetric studies revealed that the shortest histone 3 peptide that is well recognised by readers of H3K4me3 is the 4-mer ARTKme3. Although structural analyses of reader-histone complexes suggested that residues 5-7 could possibly contribute to the overall stabilisation of the reader-histone complexes, thermodynamic analyses demonstrate that readers require only the first four residues in histone 3 for binding.
Table 1 | Thermodynamic parameters for binding of histone peptides to reader domain proteins.\textsuperscript{a}

<table>
<thead>
<tr>
<th>Histone peptide</th>
<th>$K_i$</th>
<th>$\Delta G^\circ$</th>
<th>$\Delta H^\circ$</th>
<th>$-T\Delta S^\circ$</th>
<th>$K_i$</th>
<th>$\Delta G^\circ$</th>
<th>$\Delta H^\circ$</th>
<th>$-T\Delta S^\circ$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[µM]</td>
<td>[kcal mol(^{-1})]</td>
<td>[kcal mol(^{-1})]</td>
<td>[kcal mol(^{-1})]</td>
<td>[µM]</td>
<td>[kcal mol(^{-1})]</td>
<td>[kcal mol(^{-1})]</td>
<td>[kcal mol(^{-1})]</td>
</tr>
<tr>
<td>BPTF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARTKme3QTARKSTG GKA-OH</td>
<td>1.5 ± 0.3</td>
<td>-8.0 ± 0.1</td>
<td>-12.2 ± 0.1</td>
<td>4.2 ± 0.2</td>
<td>0.61 ± 0.08</td>
<td>-8.5 ± 0.1</td>
<td>-11.1 ± 0.1</td>
<td>2.6 ± 0.1</td>
</tr>
<tr>
<td>ARTKme3QTARKS-OH</td>
<td>1.4 ± 0.2</td>
<td>-8.0 ± 0.1</td>
<td>-12.0 ± 0.1</td>
<td>4.0 ± 0.2</td>
<td>0.48 ± 0.02</td>
<td>-8.6 ± 0.1</td>
<td>-11.6 ± 0.1</td>
<td>3.0 ± 0.1</td>
</tr>
<tr>
<td>ARTKme3Q-NH(_3)</td>
<td>0.94 ± 0.03</td>
<td>-8.2 ± 0.1</td>
<td>-10.5 ± 0.3</td>
<td>2.3 ± 0.2</td>
<td>1.5 ± 0.03</td>
<td>-8.0 ± 0.1</td>
<td>-10.1 ± 0.1</td>
<td>2.1 ± 0.1</td>
</tr>
<tr>
<td>ARTKme3-NH(_2)</td>
<td>0.53 ± 0.1</td>
<td>-8.6 ± 0.1</td>
<td>-10.1 ± 0.1</td>
<td>1.5 ± 0.2</td>
<td>1.1 ± 0.03</td>
<td>-8.1 ± 0.1</td>
<td>-9.0 ± 0.1</td>
<td>0.8 ± 0.2</td>
</tr>
<tr>
<td>ARTKme3-OH</td>
<td>2.2 ± 0.2</td>
<td>-7.7 ± 0.1</td>
<td>-9.1 ± 0.2</td>
<td>1.4 ± 0.3</td>
<td>2.1 ± 0.08</td>
<td>-7.7 ± 0.1</td>
<td>-9.2 ± 0.1</td>
<td>1.5 ± 0.1</td>
</tr>
<tr>
<td>ART-OH</td>
<td>n.b.\textsuperscript{b}</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
</tr>
<tr>
<td>RTKme3Q-NH(_2)</td>
<td>10.7\textsuperscript{c}</td>
<td>-6.8</td>
<td>-4.8</td>
<td>-2.0</td>
<td>3.8 ± 0.6</td>
<td>-7.4 ± 0.09</td>
<td>-3.9 ± 0.4</td>
<td>-3.5 ± 0.4</td>
</tr>
<tr>
<td>RTKme3-NH(_2)</td>
<td>10.2\textsuperscript{c}</td>
<td>-6.8</td>
<td>-3.3</td>
<td>-3.5</td>
<td>4.4 ± 1.0</td>
<td>-7.3 ± 0.1</td>
<td>-2.9 ± 0.3</td>
<td>-4.4 ± 0.4</td>
</tr>
<tr>
<td>TKme3Q-OH</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
</tr>
<tr>
<td>GARTKme3QTARKS-OH</td>
<td>4.0 ± 0.2</td>
<td>-7.4 ± 0.1</td>
<td>-10.7 ± 0.2</td>
<td>3.4 ± 0.3</td>
<td>3.8 ± 0.2</td>
<td>-7.4 ± 0.1</td>
<td>-9.0 ± 0.1</td>
<td>1.6 ± 0.1</td>
</tr>
<tr>
<td>JMJ D2A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARTKme3QTARKSTG GKA-OH</td>
<td>0.24 ± 0.01</td>
<td>-9.0 ± 0.1</td>
<td>-11.3 ± 0.2</td>
<td>2.3 ± 0.2</td>
<td>1.5 ± 0.1</td>
<td>-8.0 ± 0.1</td>
<td>-11.8 ± 0.4</td>
<td>3.8 ± 0.4</td>
</tr>
<tr>
<td>ARTKme3QTARKS-OH</td>
<td>0.086 ± 0.01</td>
<td>-9.6 ± 0.1</td>
<td>-11.0 ± 0.2</td>
<td>1.4 ± 0.3</td>
<td>1.6 ± 0.04</td>
<td>-7.9 ± 0.1</td>
<td>-14.1 ± 0.2</td>
<td>6.2 ± 0.2</td>
</tr>
<tr>
<td>ARTKme3Q-NH(_3)</td>
<td>0.086 ± 0.03</td>
<td>-9.7 ± 0.2</td>
<td>-12.2 ± 0.2</td>
<td>2.5 ± 0.2</td>
<td>5.3 ± 0.2</td>
<td>-7.2 ± 0.1</td>
<td>-8.3 ± 0.6</td>
<td>1.1 ± 1.0</td>
</tr>
<tr>
<td>ARTKme3-NH(_2)</td>
<td>0.092 ± 0.0</td>
<td>-9.6 ± 0.1</td>
<td>-12.5 ± 0.3</td>
<td>2.9 ± 0.3</td>
<td>6.9 ± 0.3</td>
<td>-7.0 ± 0.2</td>
<td>-5.3 ± 0.5</td>
<td>-1.7 ± 0.8</td>
</tr>
<tr>
<td>ARTKme3-OH</td>
<td>0.24 ± 0.03</td>
<td>-9.0 ± 0.1</td>
<td>-12.9 ± 1.1</td>
<td>3.9 ± 0.3</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
</tr>
<tr>
<td>ART-OH</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Data were derived from isothermal titration calorimetry experiments.

\textsuperscript{b} n.b. indicates not binding.

\textsuperscript{c} The subscripts \textsuperscript{c} indicate the binding strength.
<table>
<thead>
<tr>
<th>Histone peptide</th>
<th>$K_s$ [μM]</th>
<th>$\Delta G^\circ$ [kcal mol$^{-1}$]</th>
<th>$\Delta H^\circ$ [kcal mol$^{-1}$]</th>
<th>$-T\Delta S^\circ$ [kcal mol$^{-1}$]</th>
<th>$K_s$ [μM]</th>
<th>$\Delta G^\circ$ [kcal mol$^{-1}$]</th>
<th>$\Delta H^\circ$ [kcal mol$^{-1}$]</th>
<th>$-T\Delta S^\circ$ [kcal mol$^{-1}$]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TAF3</td>
<td></td>
<td></td>
<td></td>
<td>JMJD2A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTKme3Q-NH$_2$</td>
<td>1.0±0.2</td>
<td>$-8.2\pm0.1$</td>
<td>$-9.1\pm0.4$</td>
<td>0.88±0.6</td>
<td>3.4$^c$</td>
<td>$-7.5$</td>
<td>$-5.3$</td>
<td>$-2.2$</td>
</tr>
<tr>
<td>RTKme3-NH$_2$</td>
<td>1.5±0.1</td>
<td>$-8.0\pm0.1$</td>
<td>$-8.2\pm0.2$</td>
<td>0.24±0.3</td>
<td>10.2$^{d}$</td>
<td>$-6.8$</td>
<td>$-3.3$</td>
<td>$-3.5$</td>
</tr>
<tr>
<td>TKme3Q-OH</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
</tr>
<tr>
<td>GARTKme3QTARKS-OH</td>
<td>3.1±0.3</td>
<td>$-7.5\pm0.1$</td>
<td>$-6.1\pm0.2$</td>
<td>$-1.4\pm0.2$</td>
<td>1.9±0.4</td>
<td>$-7.8\pm0.1$</td>
<td>$-12.9\pm0.3$</td>
<td>$5.1\pm0.2$</td>
</tr>
<tr>
<td><strong>Sgf29</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARTKme3QTARKSTG GKA-OH</td>
<td>2.0±0.3</td>
<td>$-7.8\pm0.1$</td>
<td>$-7.9\pm0.2$</td>
<td>0.16±0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARTKme3QTARKS-OH</td>
<td>2.1±0.6</td>
<td>$-7.8\pm0.2$</td>
<td>$-7.8\pm0.2$</td>
<td>0.04±0.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARTKme3Q-NH$_2$</td>
<td>4.4±0.1</td>
<td>$-7.3\pm0.1$</td>
<td>$-5.2\pm0.1$</td>
<td>$-2.1\pm0.4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARTKme3-NH$_2$</td>
<td>4.2±0.5</td>
<td>$-7.3\pm0.1$</td>
<td>$-6.1\pm0.5$</td>
<td>$-1.2\pm0.6$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARTKme3-OH</td>
<td>4.9±0.4</td>
<td>$-7.2\pm0.1$</td>
<td>$-4.3\pm0.1$</td>
<td>$-2.9\pm0.1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ART-OH</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTKme3Q-NH$_2$</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTKme3-NH$_2$</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TKme3Q-OH</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GARTKme3QTARKS-OH</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td>n.b.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

One of the proposed important factors that govern sequence specificity of readers of H3K4me3 is the binding pocket for the positively charged nitrogen of the alanine residue at the N-terminus of histone 3.[5,12] Histone peptide RTKme3Q that lacks the N-terminal alanine, bound to BPTF, TAF3, and SGF29 with significantly decreased binding affinity than the counterpart that bears alanine at the N-terminus (BPTF ~ 11-fold, TAF3 ~ 12-fold, SGF29 > 40-fold). The observed enthalpic penalty for

Figure 2 | Representative ITC experiments displaying the titration of histone peptides to TAF3. A) 10-mer ARTKme3QTARKS-OH, B) 4-mer ARTKme3-NH2, C) 11-mer GARTKme3QTARKS-OH, and D) 3-mer RTKme3-NH2. It is clear that both the addition of G0 or deletion of A1 cause a decrease of the enthalpy of binding (as seen in smaller values on the y-axis).
binding of RTKme3Q implies that the favourable electrostatic interactions are lost in the absence of A1, thus confirming the existence of the H3A1 binding pocket. Our data for SGF29 are in agreement with the recent work by Min and co-workers that illustrated the importance of the N-terminus of H3K4me3 (Figure 3).[11] The absence of the first alanine in histone 3 caused 3-fold weaker binding to JARID1A. Notably, JMJD2A was insensitive to deletion of A1, resulting in indistinguishable values for binding affinities for ARTKme3Q and RTKme3Q. This result is supported by crystallographic analysis of JMJD2A-H3K4me3 structure that shows that A1 is positioned away from the surface of JMJD2A and that only residues RTKme3QT directly interact with the protein (Figure 3).[10] Similarly, all readers except JMJD2A bound to RTKme3 with substantially decreased affinity than they bound to ARTKme3 (BPTF ~ 19-fold, JARID1A ~ 4-fold, TAF3 ~ 16-fold, SGF29 > 40-fold). The histone peptide TKme3Q which contained an additional deletion of R2 did not form any complex with all five readers, a result that is supported by structures of reader-histone complexes which show that R2 forms a salt-bridge with the negatively charged side chain of aspartic acid in reader proteins.

To additionally probe the importance of the pocket that recognises the positively charged α-amino group of A1, we investigated how the extention of histone tail on the N-terminus affects the thermodynamics of interaction. An additional glycine residue

---

**Figure 3** View on the H3A1 binding pocket of A) TAF3 (PDB ID: 2K17), B) SGF29 (PDB ID: 3MEA), and C) JMJD2A (PDB ID: 2GFA). Reader domain proteins and H3K4me3 histone peptides are shown in dark grey and light grey, respectively.
positioned in front of alanine in the 10-mer histone peptide (i.e. GARTKme3QTARKS) resulted in a substantial reduction of binding affinity for TAF3 (~ 36-fold, Figure 2) and SGF29 (> 50-fold). 3-Fold and 8-fold decrease in binding affinities was also observed in cases of BPTF and JARID1A, respectively. ITC data demonstrated that the decreased values of $\Delta G^0$ upon addition of G0 are a result of less favourable enthalpy of binding ($\Delta H^0 = 2.9 \pm 1.4$ kcal mol$^{-1}$), a result that can be explained by the lack of favourable electrostatic interactions between the terminal amino group (as in A1) and the H3A1 binding pocket. In contrast to other reader domain proteins, JMJD2A recognised the 10-mer and 11-mer histones with indistinguishable binding affinities, again demonstrating that there is no fixed binding pocket in JMJD2A for the recognition of the terminal amino group. This result is in agreement with the results on the deletion of A1 in 4-mer and 5-mer histone peptides (see above); in both cases, the length and the chemical nature of the N-terminus does not play a crucial role in the peptide-JMJD2A recognition event.

### 2.3 - Conclusions

We have demonstrated that readers of H3K4me3 recognise the histone 3 peptide that contains the first four amino acids ARTKme3 with virtually the same binding affinity as they recognise longer 15-mer or 10-mer analogues. Calorimetric analyses on H3 peptides that lack A1 or contain an additional G0 residue, furthermore, showed that readers BPTF, TAF3 and SGF29 bind them with significantly decreased binding affinities caused by unfavourable enthalpy of binding, whereas JMJD2A is insensitive to the N-terminal addition or truncation. These results emphasise the importance of the H3A1 binding pocket in molecular readout and demonstrate that there is a selectivity between readers of H3K4me3 for histone peptides that contain the N-terminal amino group at the appropriate position. Such a difference might be exploited in the development of highly active and selective small molecule inhibitors of readers of H3K4me3 that have not been characterised yet.[13,14] Our study provides a good starting point by illustrating that readers only bind a relatively short, N-terminal histone 3 sequence.

Comparison of histone peptide length that readers of H3K4me3 request for recognition with the length required by erasers of H3K9me2/3 provides another important remark. A minimal length of 4 amino acids was found for all five readers evaluated. In contrast, recent studies indicated that the erasers JMJD2C, JMJD2A and PHF8 catalysed the removal of methyl group of histone peptides that contain at least 4, 8, and 14 amino acids, respectively.[15] In addition, TDRD3, a reader of asymmetric dimethylarginine located at the N-terminus of histone 4 (H4R3me2a), does not recognise the 8-mer
histone, whereas the 10-mer and 15-mer analogues bind with micromolar binding affinity.[16,17] A combination of these results on TDRD3 and our data on readers of H3K4me3 imply that there is an obvious distinction for minimum requirements between various readers of histones that contain post-translationally modified amino acids at the N-termini. Further investigations are needed, to unveil what the shortest histone sequences required for the recognition by readers of other methylated lysines and arginines in histone 3 and 4 are. Detailed molecular level understanding, of reader-histone interactions that includes the requirements for amino acid sequence and histone tail length, will facilitate the rational design of small molecule inhibitors of the reading processes that might possess a great therapeutic potential.

2.4 - Experimental Section

Protein expression and purification
Tagged BPTF, JMJD2A, JARID1A, TAF3 and SGF29 proteins were expressed in E.coli Rosetta BL21 DE3 pLysS hosts, using Terrific Broth medium containing the appropriate antibiotics. The bacteria were cultured to OD600 ~ 0.6 at 37 °C after which they were induced with IPTG overnight at 16 °C. Proteins were purified using Ni-NTA beads for 6xHis tagged proteins or glutathione sepharose beads for GST tagged proteins, respectively. After purification, the 6xHis tag was cleaved from JMJD2A and SGF29 using TEV-protease and the GST tag was cleaved from TAF3 using thrombin. Size exclusion chromatography using a superdex 75 column was used as a final purification step. SGF29 was eluted in 25 mM TRIS, 50 mM NaCl, 1 mM DTT at pH 7.5; JMJD2A and TAF3 were eluted in 50 mM TRIS at pH 7.5; BPTF and JARID1A were eluted in 50 mM TRIS, 20 mM NaCl at pH 7.5. All proteins were made filter sterile and stored at 4 °C until further use.

Solid-phase peptide synthesis
FMOC-protected trimethyllysine was synthesised following the reported procedures. [18,19] Peptides were made on solid phase using semi-automated solid phase peptide synthesizer employing FMOC-chemistry. Peptides that contain the carboxylic acid at the C-terminus where made on the Wang resin, whereas peptides that contain the amide group at the C-terminus were made on the Breipohl resin. Couplings where done in DMF using 3.0 equiv. of Fmoc-protected amino acid, 3.3 equiv. of DIPCIDI, and 3.6 equiv. of HOBt. Completion of the reaction was determined with the Kaiser test and the removal of the FMOC was done by adding a large excess of 20% piperidine in DMF for 20-30 minutes. After completion of the peptide syntheses, the desired peptide on the resin was washed 3x with diethylether. The peptide including the acid labile protecting groups was cleaved by a mixture of 92.5 % TFA, 2.5 % H₂O,
2.5 % triisopropylsilane, and 2.5 % ethane-1,2-dithiol. After mixing/shaking for 4-5 hours, the product was precipitated in diethylether (3x) and the diethylether was decanted off after spinning the suspension down in the centrifuge for 3 minutes at 5000 rpm. Peptides were identified by LC-MS (Thermo Finnigan LCQ-Fleet ESI-ion trap, 2x Shimadzu LC 20AD, SIL 20 AC HT, CTO 20AC, column: RP18 200 mm, 160 bar, 35 °C, ACN/H₂O, (Thermo Fischer, Breda, The Netherlands) (for peptides with Mw < 600 Da) and Bruker Biflex MALDI-TOF (Bruker Daltronik, Bremen, Germany) (for peptides with Mw > 600 Da).

**Isothermal titration calorimetry**

Reader domain proteins obtained after purification by size exclusion chromatography were directly used in ITC experiments (see Supporting information for the composition of buffers). The same buffers were used for the preparation of histone peptides. Protein concentrations were measured by UV-Vis spectroscopy at 280 nm. Typically, a reader protein (20 μM) was titrated with histone peptide (250 μM). For histone peptides that exhibited weak binding affinity, 200 μM of protein was titrated with 2 mM of histone peptide. Each ITC titration consisted of 19 injections. Experiments were repeated 3 to 5 times. ITC experiments were performed on the fully automated Microcal Auto-iTC200 (GE Healthcare Life Sciences, USA). Curve fitting was performed by Origin 6.0 (Microcal Inc., USA) using one-site model.

**2.5 - Acknowledgements**

We thank C. David Allis (BPTF, JARID1A), Jinrong Min (SGF29), Georges Mer (JMJD2A) and Marc Timmers (TAF3) for kindly providing constructs. Furthermore we thank Roman Belle for synthesizing and purifying the histone peptides.
2.6 - References

17. INVALID CITATION !!! {}
## Appendix

### Supplementary table 1 | Buffers used for ITC experiments.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Buffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP1F</td>
<td>50 mM TRIS-HCl, 20 mM NaCl at pH 7.5</td>
</tr>
<tr>
<td>JARID1A</td>
<td>50 mM TRIS-HCl, 20 mM NaCl at pH 7.5</td>
</tr>
<tr>
<td>TAF3</td>
<td>50 mM TRIS-HCl at pH 7.5</td>
</tr>
<tr>
<td>JMJD2A</td>
<td>50 mM TRIS-HCl at pH 7.5</td>
</tr>
<tr>
<td>SGF29</td>
<td>25 mM TRIS-HCl, 50 mM NaCl, 1 mM DTT at pH 7.5</td>
</tr>
</tbody>
</table>

### Supplementary table 2 | Calculated and observed mass spec data for histone peptides.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Exact mass (calc.) [M + H]^+</th>
<th>ESI-MS [M + H]^+</th>
<th>ESI-MS [M + 2H]^{2+}</th>
<th>MALDI-TOF [M]^+</th>
</tr>
</thead>
<tbody>
<tr>
<td>ART-OH</td>
<td>347.20</td>
<td>347.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTKme3-NH₂</td>
<td>445.32</td>
<td>445.36</td>
<td>223.24</td>
<td></td>
</tr>
<tr>
<td>TKme3Q-OH</td>
<td>418.28</td>
<td>418.40</td>
<td>209.80</td>
<td></td>
</tr>
<tr>
<td>ARTKme3-OH</td>
<td>517.35</td>
<td>517.32</td>
<td>259.32</td>
<td></td>
</tr>
<tr>
<td>ARTKme3-NH₂</td>
<td>516.36</td>
<td>516.40</td>
<td>258.76</td>
<td></td>
</tr>
<tr>
<td>RTKme3Q-NH₂</td>
<td>573.38</td>
<td>573.56</td>
<td>287.52</td>
<td></td>
</tr>
<tr>
<td>ARTKme3Q-NH₂</td>
<td>644.42</td>
<td></td>
<td></td>
<td>643.51</td>
</tr>
<tr>
<td>ARTKme3QTARKS-OH</td>
<td>1188.72</td>
<td>1187.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GARTKme3QTARKS-OH</td>
<td>1245.74</td>
<td>1244.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARTKme3QTARKSTGGKA-OH</td>
<td>1602.94</td>
<td>1602.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplementary figure 1 | SDS-PAGE of expressed and purified reader domain proteins.

Supplementary figure 2 | MALDI-TOF spectrum of ARTKme3QTARKSTGGKA-OH.
Supplementary figure 3 | MALDI-TOF spectrum of GARTKme3QTARKS-OH.

Supplementary figure 4 | MALDI-TOF spectrum of ARTKme3QTARKS-OH.
Supplementary figure 5| MALDI-TOF spectrum of ARTKme3Q-NH₂.

Supplementary figure 6| ESI-MS of ARTKme3-OH.
CHAPTER 2

Supplementary figure 7 | ESI-MS of TKme3Q-OH.

Supplementary figure 8 | ESI-MS of ARTKme3-NH₂.
Supplementary figure 9 | ESI-MS of the RTKme3-NH$_2$.

Supplementary figure 10 | ESI-MS of RTKme3Q-NH$_2$. 
CHAPTER 2

Supplementary figure 11 | ESI-MS of ART-OH.

Supplementary figure 12 | ESI-MS of ARTKme3Q-NH$_2$. 
3

The role of electrostatic interactions in binding of histone H3K4me2/3 to the SGF29 tandem tudor domain

B.J.G.E. Pieters
E. Meulienbroeks
R. Belle
J. Mecinović

PLoS One, 2015;10(9):e0139205
Abstract

Objective: Saga Associated Factor 29 (SGF29) has a histone 3 di/trimethyllysine (H3K4me2/3) recognition domain containing an aromatic cage. Such an aromatic cage contains aromatic residues which form cation-π interactions with the positively charged methylammonium group of Kme2/3. In the case of SGF29 this cage is supplemented with a negatively charged aspartic acid residue. In this chapter it is investigated what the molecular basis for the involvement of the D266 residue in H3K4me2/3 recognition is.

Approach: Wild type (WT) SGF29 H3K4me2/3 reader domain protein and mutant variants were recombinantly expressed and purified. The effect of these mutations on H3K4me2 and H3K4 me3 peptide binding has been investigated using circular dichroism (CD), differential scanning fluorimetry (DSF) and isothermal titration calorimetry (ITC).

Results: Analysis by ITC has shown that all induced mutations except the D266E variant resulted in a marked reduction of binding affinity of SGF29 for H3K4me2/3 peptides, stressing the importance of the negative charge. The Y238F mutant has further shown that no water mediated hydrogen bond is present between D266-Y238 that contributes to H3K4me2/3 binding. Binding experiments with peptides lacking R8 have also shown that no direct salt bridge persists between D266 of SGF29 and R8 of the H3K4me3 peptide. Finally, structural analysis of the SGF29 aromatic cage has revealed that the favourable interaction between D266 and Kme2/3 is likely due to electrostatic interactions and not due to favourable hydrogen bonding.

Conclusions: In SGF29 the general mechanism for the recognition of H3K4me2 and H3K4me3 is similar. Additionally, the molecular basis of methyllysine recognition by the D266 residue appears to be electrostatic in nature.
3.1 - Introduction

In order to structure the vast amount of genetic material stored within the cell, several mechanisms of DNA condensation have evolved in eukaryotic organisms. The first level of condensation is achieved by wrapping DNA around a protein assembly of eight histone proteins to form the nucleosome.[1] The histone tails protruding from this assembly are especially accessible for various posttranslational modifications (PTM) such as acetylation, phosphorylation and methylation.[2] These PTMs have been found to be involved in processes such as gene activation, gene transcription and chromatin condensation. Not surprisingly, due to this close relation of epigenetics to cellular function, alterations in function of proteins involved in epigenetic gene regulation have been associated with human disease, most notably various cancers.[3,4]

One of the most widespread histone PTMs is lysine methylation, which can be found in three different methylation states: monomethylated (Kme1), dimethylated (Kme2) and trimethylated (Kme3). The process of lysine methylation is controlled by three types of proteins, namely histone lysine methyltransferases (writers), histone lysine demethylases (erasers), and histone lysine reader domain proteins (also known as chromatin effectors), that are capable of specifically recognizing posttranslationally-modified histones.[5,6] A common feature of readers that specifically recognize Kme3 and Kme2 is that they possess a methyllysine recognition site referred to as the aromatic cage. Such an aromatic cage typically consists of 2-4 aromatic amino acid residues, which are capable of interacting with the positively-charged Kme2 and Kme3 side-chain, presumably via a combination of cation-π, hydrophobic and electrostatic interactions.[7] Aromatic cages flanking the Kme2/3 on two or three sides are referred to as half cages, whereas those flanking the Kme2/3 at four sides are referred to as full cages.[7,8] Half cages usually contain an additional negatively-charged amino acid (aspartic acid or glutamic acid) instead of an aromatic amino acid, as is the case with full cage binding sites. Comparative analyses determined that, in general, binding affinities follow the trend of Kme3 > Kme2 > Kme1 > K.[7] Substitution studies on residues that constitute the readers’ aromatic cages demonstrated the involvement of individual amino acids in associations with Kme2 or Kme3.[7,8]

SGF29 (SAGA-associated Factor 29) is a subunit of the SAGA (Spt-Ada-Gcn5 acetyltransferase) complex, a chromatin modifying complex, which is capable of acetylating and deubiquitinating histone proteins.[9,10] Crystallographic work has shown that SGF29 contains a tandem tudor domain with each tudor domain having a negatively-charged pocket capable of binding H3A1 and an aromatic cage for the recognition of H3K4me2/3 residues (Figure 1).[11] The aromatic cage is composed of 3 aromatic amino acid residues and a negatively-charged aspartate (Y238, Y245,
CHAPTER 3

F264, and D266), with F264 being positioned at the edge of the recognition site. Thermodynamic analyses have confirmed the requirement of the H3A1 recognition site for adequate binding of H3K4me3 peptides to SGF29 and that only the first four N-terminal amino acid residues (ARTKme3) are essentially required for binding of H3K4me3 peptides to the SGF29 reader domain.[12]

This study examines the molecular basis for the involvement of the negatively-charged aspartate/glutamate in the biomolecular recognition of Kme2/3-containing histone peptides using SGF29 as a reader domain protein. We chose the tandem tudor domain of SGF29 for this study for the following reasons: 1) the Kme2/3 recognition site contains three aromatic residues and one negatively-charged Asp266; 2) X-ray crystal or NMR solution structures for SGF29, SGF29-H3K4me2 and SGF29-H3K4me3 complexes have been solved (Figure 1); 3) binding affinities in low micromolar range can be determined by isothermal titration calorimetry (ITC); 4) wild-type SGF29 and its variants can be obtained in quantities necessary for further biochemical and

Figure 1 | Crystal structure of the apo form of SGF29 (A, PDB ID: 3MEW), SGF29-H3K4me2 (B, PDB ID: 3MET) and SGF29-H3K4me3 (C, PDB ID: 3ME9). SGF29 and H3K4me2/3 histone peptide are shown in cyan and yellow, respectively.
biophysical studies. In order to probe the role of Asp266 in the SGF29-H3K4me2/3 interactions, the D266 residue was substituted by various amino acids using site directed mutagenesis. Firstly, the role of D266 in the SGF29-H3K4me2/3 binding event was investigated by D266 substitution, using amino acid residues that vary in length and charge. Secondly, D266 was substituted by various aromatic residues in order to study the effect of the potential full cage-like structure on the association between SGF29 and H3K4me2/3. Here we report the systematic thermodynamic analyses for the association of H3K4me2/3 histone peptides to SGF29 and its aromatic cage variants for the first time, with a special focus on elucidating the role of Asp266 in the SGF29-H3K4me2/3 interactions.

3.2 - Results & discussion

Protein biophysical characterization
Fmoc-Kme2-OH and Fmoc-Kme3-OH were incorporated in 10-mer histone 3 peptides (peptide sequences: ARTKme2QTARKS and ARTKme3QTARKS) on position 4 using solid-phase peptide synthesis. The two peptides were purified by Rf-C18 preparative HPLC and the lyophilized peptides were used for binding studies (Supplementary figures 1 & 2). Wild-type SGF29 and its recognition site variants were expressed in E. coli and subsequently purified (Supplementary figure 3). In order to determine whether the variations introduced at the D266 site of SGF29 caused structural perturbations, CD experiments were initially carried out on the recombinantly expressed wild-type SGF29 (hereafter referred to as WT SGF29) and its variants (Figure 2). Multiple batches of WT SGF29 and each of its variants were produced in order to take into the account the batch-to-batch variation of proteins (Figure 2). Collectively, the CD data showed that WT and D266E SGF29 have very similar CD spectra, whereas D266N has a slightly altered spectrum. D266A, D266F, D266Y and D266W variants displayed significantly different spectra when compared to WT SGF29. These results are possibly due to the fact that D266 is positioned in an unstructured loop region of the reader domain and has a potential to interact with the adjacent Y238 via a water-mediated hydrogen bond, as suggested by structural analyses of the apo and holo forms of SGF29 (Figure 1). Upon abrogation of this hydrogen bond, the unstructured D266 loop region might adopt alternative conformations resulting in altered tertiary protein structure. In order to test this hypothesis, we mutated the Y238 residue, which is also located in an unstructured region of SGF29, into phenylalanine, which does not have the ability to directly participate in hydrogen bonding. From the seven batches of Y238F proteins we prepared, a broad range of CD spectra have been observed, ranging from wild-type like to a distinctly altered spectrum.
Figure 2: CD spectra of various expressions of wild-type SGF29 its D266 variants, and the Y238F and Y245F variants. CD experiments were carried out at the concentration of 0.1 mg ml⁻¹ in 10 mM sodium phosphate buffer (pH 7.5).
Additionally, differential scanning fluorimetry (DSF) was employed as a secondary method to determine the structural stability of the SGF29 variants (Figure 3, Table 1). Using DSF, WT SGF29 was shown to have a Tm of 57.8 °C. Variant proteins have similar or decreased Tm values when compared to WT SGF29, ranging from 53.9-57.7 °C. The introduction of the negatively-charged Glu at the D266 site gave indistinguishable Tm values relative to WT SGF29. In addition, the introduction of aromatic residues only had a small influence on Tm, as substitution of D266 with Trp, Phe or Tyr only decreased Tm by 2.3, 3.3 or 3.9 °C, respectively. Collectively, all Tm values can be considered to be within close proximity to the Tm for WT SGF29.

These data, combined with CD spectra indicate that D266A, F, Y and W variants have altered tertiary structures, which do not substantially affect the stability of the variant proteins. Because the D266 residue is located in an unstructured region on the surface of the reader domain, alterations made to this residue do not appear to affect the more structurally and thermodynamically stable β-sheet rich regions. These structured regions make up the major part of the SGF29 reader domain, explaining the subtle changes in Tm values for the D266X variants. Similarly, we observed similar Tm values for all Y238F variants that display altered and unchanged CD spectra relative to WT SGF29, suggesting that different conformations of the D266 containing loop region do not substantially affect the stability of this variant. Y245F variant displays similar Tm value as the related Y238F.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Tm values for SGF29 and its variants as determined by DSF displayed as Tm± SD (Measured in triplicate).</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGF29</td>
<td>Tm (°C)</td>
</tr>
<tr>
<td>WT</td>
<td>57.8 ± 0.1</td>
</tr>
<tr>
<td>D266E</td>
<td>57.7 ± 0.1</td>
</tr>
<tr>
<td>D266A</td>
<td>55.8 ± 0.3</td>
</tr>
<tr>
<td>D266N</td>
<td>56.4 ± 0.1</td>
</tr>
<tr>
<td>D266W</td>
<td>55.5 ± 0.1</td>
</tr>
<tr>
<td>D266Y</td>
<td>54.5 ± 0.2</td>
</tr>
<tr>
<td>D266F</td>
<td>53.9 ± 0.5</td>
</tr>
<tr>
<td>Y238F</td>
<td>56.5 ± 0.2</td>
</tr>
<tr>
<td>Y245F</td>
<td>56.4 ± 0.1</td>
</tr>
</tbody>
</table>
Chapter 3

ITC

Recognition of 10-mer H3K4me3 and H3K4me2 histone peptides by WT SGF29 and its D266 variants was examined by isothermal titration calorimetry (ITC). ITC provided full thermodynamic descriptions (i.e. dissociation constant Kd, free energy of binding ΔG°, enthalpy of binding ΔH°, and entropy of binding -TΔS°) for the association of reader-histone complexes (Table 2, Figure 4). Consistent with previous studies, thermodynamics of interactions between H3K4me3 and SGF29 showed that the

Figure 3 | Tm curves of A) wild-type SGF29 and its D266 variants and B) SGF29 and its Y238 and Y245 variants.
association is enthalpy-driven ($\Delta H^\circ = -8.1 \text{ kcal mol}^{-1}$), whereas the entropy of binding is slightly unfavorable ($-\Delta S^\circ = 0.6 \text{ kcal mol}^{-1}$).[12] H3K4me2 histone peptide exhibits slightly lower binding affinity than H3K4me3 for binding to WT SGF29 (Kd = 4.7 $\mu$M for H3K4me2 vs. 3.1 $\mu$M for H3K4me3). Thermodynamic analyses furthermore demonstrated that there is an enthalpy-entropy compensation for binding of H3K4me3 and H3K4me2 to WT SGF29. $\Delta H^\circ$ is more favorable for the formation of SGF29-H3K4me3 complex relative to SGF29-H3K4me2 by 0.9 kcal mol$^{-1}$, while $-\Delta S^\circ$ is less favorable by 0.7 kcal mol$^{-1}$. Similar observations have been reported for other reader domain proteins, including BPTF and JARID1A, which specifically recognize H3K4me3 and H3K4me2. [13,14]

We then investigated the importance of the negatively-charged D266 on the recognition of H3K4me2/3 histones. D266 is located in the SGF29’s Kme2/3 recognition site, complementing the aromatic cage that consists of Y238, Y245 and F264 (Figure 1). Substitution of D266 by glutamic acid (D266E), an amino acid of one-carbon longer side-chain and the same negative charge as aspartic acid, showed virtually indistinguishable binding affinities for H3K4me3 and H3K4me2, respectively. Interestingly, D266E substitution results in more favorable $\Delta H^\circ$ (by -0.9 kcal mol$^{-1}$), which is fully compensated by more unfavorable $-\Delta S^\circ$ (by 1.0 kcal mol$^{-1}$) for both H3K4me3 and H3K4me2. This result can be rationalized by increased flexibility of glutamate relative to aspartate in the unbound state, which becomes substantially restricted upon the formation of a complex with H3K4me2/3.

Substitution of the negatively-charged aspartate D266 by neutral and smaller alanine (D266A) resulted in a significant reduction in binding affinities for H3K4me3 (8-fold) and H3K4me2 (7-fold), suggesting the essential contribution of the negative charge of D266 to the overall binding affinities. The observed similar decreases in $\Delta G^\circ$ values are a result of more unfavorable $\Delta H^\circ$ for H3K4me3 (1.7 kcal mol$^{-1}$) and H3K4me2 (1.1 kcal mol$^{-1}$), implying that the underlying mechanism for the loss of binding affinity for D266A is similar for both peptides. When D266 was substituted by the uncharged asparagine (D266N), it was observed that there was a decrease, similar to the D266A substitution, in binding affinity for H3K4me3 (8-fold) and H3K4me2 (9-fold) when compared to the WT SGF29. The decrease in $\Delta H^\circ$ for H3K4me3 was observed to be 2.9 kcal mol$^{-1}$, which was compensated by an increase in entropy of 0.7 kcal mol$^{-1}$ resulting in a $\Delta G^\circ$ of -6.3 kcal mol$^{-1}$. For H3K4me2 the decrease in $\Delta H^\circ$ was observed to be 1.8 kcal mol$^{-1}$ and compensated by an increase in $-\Delta S^\circ$ of 0.5 kcal mol$^{-1}$ resulting in a $\Delta G^\circ$ of -6.0 kcal mol$^{-1}$. When compared to the WT SGF29, D266A and D266E variants, these data indicate that the overall negative charge on D266 is responsible for the observed favorable binding affinities for H3K4me2/3, the length of the side-chain being less important.
Table 2 | Thermodynamic parameters for binding of histone H3K4me2 and H3K4me3 peptides to SGF29 variant proteins.
[a] As measured by ITC ± SD (2-4 repeats). [b] NB = No Binding (Kd > 200 μM).

<table>
<thead>
<tr>
<th>SGF29</th>
<th>Kd (μM)</th>
<th>ΔG° (kcal mol⁻¹)</th>
<th>ΔH° (kcal mol⁻¹)</th>
<th>-TΔS° (kcal mol⁻¹)</th>
<th>Kd (μM)</th>
<th>ΔG° (kcal mol⁻¹)</th>
<th>ΔH° (kcal mol⁻¹)</th>
<th>-TΔS° (kcal mol⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>H3K4me2</td>
<td>H3K4me3</td>
<td></td>
<td></td>
<td>H3K4me2</td>
<td>H3K4me3</td>
<td></td>
</tr>
<tr>
<td>WT</td>
<td>4.7 ± 0.3</td>
<td>-7.3 ± 0.1</td>
<td>-7.2 ± 0.2</td>
<td>-0.1 ± 0.2</td>
<td>3.1 ± 0.3</td>
<td>-7.5 ± 0.1</td>
<td>-8.1 ± 0.2</td>
<td>0.6 ± 0.2</td>
</tr>
<tr>
<td>D266E</td>
<td>5.6 ± 1.5</td>
<td>-7.2 ± 0.2</td>
<td>-7.9 ± 0.4</td>
<td>0.7 ± 0.6</td>
<td>3.9 ± 0.3</td>
<td>-7.4 ± 0.1</td>
<td>-9.0 ± 0.1</td>
<td>1.6 ± 0.1</td>
</tr>
<tr>
<td>D266A</td>
<td>38 ± 4</td>
<td>-6.0 ± 0.1</td>
<td>-6.1 ± 0.2</td>
<td>0.1 ± 0.1</td>
<td>25 ± 3</td>
<td>-6.3 ± 0.1</td>
<td>-6.4 ± 0.1</td>
<td>0.1 ± 0.1</td>
</tr>
<tr>
<td>D266N</td>
<td>44 ± 5</td>
<td>-6.0 ± 0.1</td>
<td>-5.4 ± 0.2</td>
<td>-0.6 ± 0.2</td>
<td>25 ± 2</td>
<td>-6.3 ± 0.1</td>
<td>-6.2 ± 0.1</td>
<td>-0.1 ± 0.1</td>
</tr>
<tr>
<td>D266W</td>
<td>NB[b]</td>
<td>NB</td>
<td>NB</td>
<td>NB</td>
<td>NB[b]</td>
<td>NB</td>
<td>NB</td>
<td>NB</td>
</tr>
<tr>
<td>D266Y</td>
<td>38 ± 4</td>
<td>-6.0 ± 0.1</td>
<td>-3.9 ± 0.2</td>
<td>-2.1 ± 0.2</td>
<td>28 ± 4</td>
<td>-6.2 ± 0.1</td>
<td>-4.8 ± 0.2</td>
<td>-1.4 ± 0.3</td>
</tr>
<tr>
<td>D266F</td>
<td>57 ± 3</td>
<td>-5.8 ± 0.1</td>
<td>-2.6 ± 0.2</td>
<td>-3.2 ± 0.1</td>
<td>44 ± 5</td>
<td>-5.9 ± 0.1</td>
<td>-3.4 ± 0.2</td>
<td>-2.5 ± 0.1</td>
</tr>
<tr>
<td>Y238F</td>
<td>6.8 ± 0.4</td>
<td>-7.1 ± 0.1</td>
<td>-7.0 ± 0.1</td>
<td>-0.1 ± 0.2</td>
<td>3.5 ± 0.2</td>
<td>-7.4 ± 0.1</td>
<td>-8.1 ± 0.1</td>
<td>0.7 ± 0.1</td>
</tr>
<tr>
<td>Y245F</td>
<td>19 ± 2</td>
<td>-6.4 ± 0.1</td>
<td>-4.5 ± 0.1</td>
<td>-1.9 ± 0.1</td>
<td>14 ± 1</td>
<td>-6.6 ± 0.1</td>
<td>-5.9 ± 0.1</td>
<td>-0.7 ± 0.1</td>
</tr>
</tbody>
</table>
Having shown that D266 has an important role in the strong recognition of H3K4me3 and H3K4me2 histone peptides, we aimed to provide further explanation for such observations. Based on crystallographic analyses of SGF29-H3K4me3 and SGF29-H3K4me2 complexes, D266 could contribute into the overall binding affinity in three possible ways: i) favorable direct electrostatic interactions with Kme2/3, ii) favorable H-bond network that involves a water molecule and Y238, and iii) favorable salt-bridge with R8 of histone 3.

Firstly, our thermodynamic analyses show that all D266X variations affect the binding affinities for H3K4me3 and H3K4me2 in very similar fashion (i.e. $\Delta \Delta G$, $\Delta \Delta H$, and $-\Delta \Delta S$ values for association of Kme3 to D266X relative to WT SGF29 are similar to $\Delta \Delta G$, $\Delta \Delta H$, and $-\Delta \Delta S$ values for Kme2-binding). These results imply that D266 likely contributes to binding of Kme3 and Kme2 in the same way. The distance between the negatively-charged oxygen of D266 and the hydrogen atoms of the positively-charged $\text{C}_3\text{H}_7\text{N}^+$ of Kme3 and Kme2 is about 3.0 Å, which is within the van der Waals contact distance. Notably, the hydrogen atom bonded to nitrogen in $^+\text{NHMe2}$ is positioned away from the D266 side-chain, suggesting the absence of direct H-bonding or salt-bridge between $^+\text{NHMe2}$ and D266 (there are two conformations of SGF29-H3K4me2 structure, both having NH$^+$ positioned away from the D266 site; PDB ID: 3MET). Thus, in the case of SGF29, it is likely that interactions between D266 and the Kme2/3 side-chain are predominantly electrostatic in nature.

The second possibility for the observed involvement of D266 in recognition of H3K4me2/3 is that the network of H-bonding between Y238-water-D266 is abrogated in D266A, which consequently results in an overall decrease in $\Delta G$°. In order to test this hypothesis, a simple variation was introduced at the Y238 position. By substituting tyrosine by phenylalanine (Y238F) the ability of Y238 to form a hydrogen bond is abrogated. Importantly, ITC data on Y238F variants that displayed both altered or similar CD spectra relative to WT SGF29, showed indistinguishable thermodynamic profiles for H3K4me3 and H3K4me2 binding, when compared to WT SGF29 (Figure 5).

These results suggest that even if the SGF29 variants display altered conformations in their apo form, they might have the ability to adopt a properly folded aromatic cage WT-like structure in the complexed SGF29-H3K4me2/3 form. In contrast to Y238F, Y245 substitution (Y245F) resulted in a reduction in binding affinities. H3K4me2/3 binding became entropically more favorable, but this was offset by a decrease in enthalpy of binding resulting in a 4.5- and 4.0-fold reduced binding affinities for H3K4me3 and H3K4me2, respectively. This result can be explained by the fact that Y245 forms a H-bond with the Kme2/3’s backbone NH of the amide group (Figure 1). When this hydrogen bonding potential is abolished (as in
Figure 4 | Thermodynamic analyses of SGF29-H3K4me2 and SGF29-H3K4me3 interactions. Representative ITC experiments showing the titration of H3K4me3 (top row, A-D) and H3K4me2 (bottom row, E-H) peptides to SGF29 (first column, A,E), D266E (second column B,F), D266N (third column, C,G) and D266Y (fourth column, D,F).
Y245F), the enthalpy of binding is reduced, thus resulting in less favorable binding affinity. In contrast, Y238 interacts with D266 via a water-mediated hydrogen bond. This interaction likely helps to stabilize the unstructured D266 loop, whereas the tyrosine’s hydroxyl group is not directly involved in H3K4me3 and H3K4me2 peptide
binding. Overall, these data, supported by structural analyses of the free and complexed forms of SGF29, illustrate that the water-mediated H-bonding between D266 and Y238 helps to stabilize SGF29 and enables the formation of the static aromatic cage for the recognition of H3K4me2/3. The lack of D266-water-Y238 hydrogen bonding network, however, is not determining the binding affinity between SGF29 and H3K4me2/3, because Y238F variants that display altered CD spectra exhibit similar thermodynamics of interactions to the unaltered counterparts. These data suggest that protein variants with alternative loop folds have the ability to form a properly folded aromatic cage that resembles the WT aromatic cage in the bound SGF29-H3K4me2/3 complexes. Recent work by Ali et al. supports this view, as the crystallographic analyses of the apo and holo forms of the PHD domain of MLL5 showed that the unbound and bound forms of the protein exhibit different conformations of the N-terminal loop region.[15] The binding of H3K4me3 substrate to MLL5 is associated with a large conformational change of the unstructured loop that contains the negatively-charged D128, which is part of the aromatic cage/recognition site and directly interacts with the Kme3 side chain. Thus, the ability of the free protein to undergo the conformational change is a prerequisite for the energetically favorable association with its H3K4me3 substrate. In regard to our results on SGF29, it is possible that the D266X variants, although most of them bear different conformations of the loop region in the free state, may have an ability to adopt the conformation in the bound state that is similar to the conformation of the WT SGF29.

In line with these results, more detailed analyses of the crystal or solution structures of the apo and holo forms of epigenetic reader domain proteins that contain the aromatic cage for specific binding of methylated lysines show that in many cases aromatic residues and negatively-charged Asp/Glu residues are positioned in the unstructured loop regions (although these residues are also found in α-helices and β-sheets).7 It is noteworthy that substitutions of these residues typically result in decreased binding affinities for associations with Kme2/3-containing histones. The observed weaker binding has often been attributed to the lack of favorable non-covalent interactions, although it is important to consider that structural changes may also influence the binding affinities.

The involvement of the potential salt-bridge between the negatively-charged D266 and positively-charged R8 of 10-mer H3K4me3 and H3K4me2 peptides was also examined (PDB ID: 3ME9, 3MET). 10-mer H3K4me3 peptide that contains G instead of R at position 8 bound to WT SGF29 with a Kd of 4 μM, similar to the native 10-mer peptide (Figure 6A). In addition, 7-mer peptide ARTKme3QTA that lacks the last three residues, including the eighth residue R8, bound to WT SGF29 with Kd value of 9 μM, demonstrating the 3-fold decrease of the binding affinity relative to 10-mer H3K4me3
peptide (Figure 6B). These data suggest that it is unlikely that potential D266-R8 salt-bridge contributes to a significant extent to the overall favorable association between H3K4m2/3 and SGF29.

Next, with the aim to construct a full cage-like reader module, D266 was substituted by the aromatic amino acids tryptophan, tyrosine or phenylalanine (D266W, D266Y, or D266F, respectively). When a D266W substitution was introduced, however, binding of H3K4me3 and H3K4me2 was abrogated. These results can be explained by the destabilizing effect of D266W substitution on the overall protein structure as was observed by CD, possibly caused by the increased size of the tryptophan residue, when compared to the native aspartic acid. D266Y bound to H3K4me2 and H3K4me3 with 9- and 10-fold decreased binding affinities relative to WT SGF29. Enthalpy decreased markedly for both H3K4me2 (3.3 kcal mol\(^{-1}\)) and H3K4me3 (3.3 kcal mol\(^{-1}\)), but was only partially compensated by an increase in entropy of about 2.1 kcal mol\(^{-1}\) for both methylation states. D266F substitution displayed even greater reduction in binding affinities for H3K4me3 (14-fold) and H3K4me2 (12-fold). The observed decrease in \(\Delta H^\circ\) of 4.6 kcal mol\(^{-1}\) and 4.7 kcal mol\(^{-1}\) for H3K4me3 and

![Figure 6](image)

**Figure 6** | ITC experiments showing binding of A) ARTKme3QTAGKS and B) ARTKme3QTA to WT SGF29. Thermodynamics of binding for A) \(K_d = 4.0 \pm 0.6 \mu M, \Delta G^\circ = -7.4 \pm 0.1\) kcal mol\(^{-1}\), \(\Delta H^\circ = -8.0 \pm 0.1\) kcal mol\(^{-1}\), -T\(\Delta S^\circ = 0.6 \pm 0.1\) kcal mol\(^{-1}\) and for B) \(K_d = 9.0 \pm 0.5 \mu M, \Delta G^\circ = -6.9 \pm 0.1\) kcal mol\(^{-1}\), \(\Delta H^\circ = -7.8 \pm 0.1\) kcal mol\(^{-1}\), -T\(\Delta S^\circ = 0.9 \pm 0.1\) kcal mol\(^{-1}\).
H3K4me2, respectively, was only partially compensated by more favorable $-\Delta S^°$ (3.1 kcal mol$^{-1}$) for both peptides. Observations that hydrophobic Tyr and Phe substituents at the D266 site exhibit more favorable entropy of binding relative to native D266 possibly derives from more favorable desolvation of the aromatic amino acids and a higher degree of flexibility of the protein or histone in the complex. Collectively, mutagenesis data on SGF29 show that the presence of aromatic residues that have a potential to form a full cage-like module does not exhibit advantage for the natural sequence of H3K4me2/3 over the half-cage arrangement. Apart from the structural changes observed by CD, the increase in bulk when exchanging aspartic acid for the aromatic residues phenylalanine, tyrosine or tryptophan may also affect binding affinity due to steric hindrance, which may have been introduced into the aromatic cage.

Together with recent binding studies on BPTF, L3MBTL1, TAF3, and HP1 reader proteins our study suggests that the underlying mechanisms for the specific recognition of Kme2 by negatively-charged Asp or Glu vary in each case, and that the recognition does not obey an involvement of the same type of favorable interactions. For instance, BPTF Y17E variant exhibits selectivity for binding to Kme2 over Kme3 due to direct H-bonding with hydrogen of $^{+}$NHMe2 side-chain.[16] L3MBTL1, a reader domain protein that specifically recognizes lower methylation states of lysine (i.e. Kme and Kme2), also makes a direct H-bond with negatively-charged D355 located in the recognition site. In addition, D877 residue of TAF3 is not directly involved in intermolecular interactions with the Kme2 side-chain, but forms a hydrogen bond with the hydroxyl group of Thr6 in the H3K4me2 peptide.[17] In contrast, E52 residue of chromodomain HP1 interacts with the positively-charged $^{+}$NHMe2 via water-mediated H-bonding.[18]

### 3.3 - Conclusions

Our thermodynamic results on SGF29’s recognition site variants, supported by crystallographic analyses, suggest that favorable H-bonding interaction between D266 and $^{+}$NHMe2 is unlikely to be present, thus emphasizing the existence of favorable direct electrostatic interactions between negatively-charged D266 and positively-charged Kme2 or Kme3. The magnitude of these interactions for associations with Kme2 and Kme3 is similar, which results in comparable selectivities (1.3–1.9) for the SGF29 variants that were investigated. Additionally, our work suggests that the negative charge on the D266 residue appears to play an important role in maintaining the structural integrity of the loop region that constitutes one part of the aromatic cage of SGF29 in its apo state. We furthermore showed that H-bonding between
Y245 and backbone K4 amide provides an additional driving force (~1 kcal mol\(^{-1}\)) for the specific readout of H3K4me2/3 by SGF29. This work, together with our previous studies that confirmed the existence of the H3A1 recognition site, [12] highlights the importance of electrostatic interactions and hydrogen bonding in specific recognition of H3K4me2/3 by the SGF29 tandem tudor domain.

3.4 - Experimental Section

Solid-phase peptide synthesis
Fmoc-protected dimethyllysine and trimethyllysine were synthesized by modifications of reported procedures [12]. The ARTKme3QTARKS and ARTKme2QTARKS peptides were made on solid phase by using manual solid phase peptide synthesis applying Fmoc chemistry. Peptides containing a carboxylic acid at the C terminus were made on Wang resin (± 0.6 mmol g\(^{-1}\)) and couplings were done in DMF with Fmoc-protected amino acid (3.0 equiv.), diisopropylcarbodiimide (DIPCDI) (3.3 equiv.) and hydroxybenzotriazole (HOBt) (3.6 equiv.). Completion of the reaction was determined with the Kaiser test, and removal of Fmoc was achieved by treatment with a large excess of piperidine (20 %) in DMF for 20–30 min. Every wash step was performed with 3 \(\times\) DMF and after building completion the Fmoc was removed followed by wash 3 \(\times\) DMF and 3 \(\times\) Et2O continued by drying of the resin in vacuo. The peptide (including acid-labile protecting groups) were cleaved by a mixture of TFA (92.5 %), \(\text{H}_2\text{O}\) (2.5 %), tri-isopropylsilane (TIPS) (2.5 %) and ethane-1,2-dithiol (EDT) (2.5 %). After mixing / shaking for 4–5 h, the product was precipitated in \(\text{Et}_2\text{O}\), and the \(\text{Et}_2\text{O}\) was decanted off after centrifugation (3500 rpm, 3 min, Hermle 220.72 v04). Peptides were identified by LC-MS and MALDI-TOF. Purification was performed on preparative HPLC.

Site directed mutagenesis
Primers were designed using PrimerX software (http://bioinformatics.org/primerx/) and obtained from Biolegio (S1 Table). After DNA amplification by PCR, the reaction mixtures were subjected to DpnI treatment and the reaction products were subsequently transformed into competent \(E.\text{coli}\) top 10 cells. After DNA isolation, the presence of the desired mutation was verified by sequence analysis.

Protein expression and purification
Wild-type and variant SGF29 reader domain proteins (\(Homo\ sapiens\), residues 115-293) were expressed in terrific broth (TB) medium containing 50 \(\mu\text{g}\) mL\(^{-1}\) kanamycin, using \(E.\text{coli}\) Rosetta BL21 DE3 PlysS as expression host. Bacteria were cultured to OD600 \(\approx\) 0.6 at 37 °C, 200 rpm after which the cultures were induced using 1 mM IPTG (final conc.). Cultures were then incubated at 16 °C for 16 hours.
Cells were harvested and lysed by sonication in 1x PBS, 250 mM NaCl, 5 % glycerol, 0.2 % CHAPS, 10mM β-mercaptoethanol. Lysate was centrifuged and the supernatant purified using Ni-NTA beads. The 6xHis tag was cleaved off overnight at 4 °C using TEV-protease. Size exclusion chromatography, using a superdex 75 column, was employed as a final purification step with 25 mM TRIS pH 8.0, 50 mM NaCl, 1 mM DTT as eluent. Protein concentrations were determined by UV/Vis spectroscopy at 280 nm.

**CD Spectroscopy**

CD measurements were conducted using a J-815 circular dichroism spectropolarimeter (JASCO) at 0.1 mg ml⁻¹ protein concentration in a 1 mm cuvette using 10 mM NaH₂PO₄ pH 7.5 as buffer. The wavelength range scanned was 180-260 nm using a bandwidth of 1 nm at a time constant of 0.5 second and scan rate of 50 nm sec⁻¹. Obtained spectra are averages of 10 measurements and have been smoothed using a Savitzky-Golay filter with a convolution width of 11.

**Differential Scanning Fluorimetry (DSF)**

SGF29 protein melt curves were obtained as described by Reinhard et. al. [19] using a StepOne-Plus Real-Time PCR system (Applied Biosystems) and MicroAmp fast optical 96-well reaction plates (Applied Biosystems). SYPRO-Orange protein gel stain (Invitrogen) was used as a reporter dye, emitting fluorescence in the FAM channel. Total reaction volume was 25 μl; 20 μl buffer (25 mM TRIS pH 7.5, 50 mM NaCl, 1 mM DTT), 2.5 μl of 25 μM protein and 2.5 μl SYPRO-Orange dye (diluted 1:100 in ddH₂O). Melt curve data were obtained in triplicate, in a temperature range of 20-95 °C at a step-wise temperature increment of 1 °C min⁻¹. Obtained data were analyzed using DSF Analysis v3.0.2 software, designed by Niesen et. al. (available via ftp://ftp.sgc.ox.ac.uk/pub/biophysics/).[20]

**Isothermal Titration Calorimetry (ITC)**

ITC experiments were conducted using an automated Microcal AutoITC200 (GE Healthcare). Generally 1.2 mM of peptide was titrated to 100 μM of protein. For variants that displayed low binding affinities, concentrations of 2.4 mM peptide were titrated to 200 μM of protein. The buffers used for ITC experiments were the same as the elution buffer used for SEC; 25 mM TRIS pH 8.0, 50 mM NaCl, 1 mM DTT. Heats of dilution were subtracted from the titration binding data curve fitting. Curve fitting was performed with Origin 6.0 (Microcal Inc., Northampton, MA) using a one-site model.
3.5 - Acknowledgements

This work has been conducted in part by Roman Belle (peptide synthesis) and Erik Meulbroeks (mutagenesis, protein expression and purification). We thank Professor Ger Pruijn for encouragement.
3.6 - References

## Appendix

### Supplementary table 1 | Primers used for SGF29 mutagenesis.

Mutations are underscored and codons are indicated in grey.

<table>
<thead>
<tr>
<th>SGF29 mutation</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>D266E fwd.</td>
<td>5’ CGGTCTCTTTGAGAGAACCTCTCTATGCAGATGG 3’</td>
</tr>
<tr>
<td>D266E rev.</td>
<td>5’ CCATCTGCATAGGAGGTTTCTTCAACAGGACCG 3’</td>
</tr>
<tr>
<td>D266A fwd.</td>
<td>5’ CGGTCTCTTTGAGGCCACCTCCTATGCAG 3’</td>
</tr>
<tr>
<td>D266A rev.</td>
<td>5’ CTGCACTAGGAGGCTCTCAAACAGGACCG 3’</td>
</tr>
<tr>
<td>D266N fwd.</td>
<td>5’ CTGGTCTCTTTGAGAACCTCCTATGCAG 3’</td>
</tr>
<tr>
<td>D266N rev.</td>
<td>5’ CTGATAGGAGGTTTTCAACAGGACCG 3’</td>
</tr>
<tr>
<td>D266W fwd.</td>
<td>5’ TACTCGGTCTCTTTGAGACCTCCTATGCAGATGGGC 3’</td>
</tr>
<tr>
<td>D266W rev.</td>
<td>5’ GCCATCTGCATAGGAGGTTTCTTCAACAGGACCGAGTAG 3’</td>
</tr>
<tr>
<td>D266Y fwd.</td>
<td>5’ CTGGTCTCTTTGAGAACCTCCTATGCAG 3’</td>
</tr>
<tr>
<td>D266Y rev.</td>
<td>5’ CTGCACTAGGAGGCTCTCAAACAGGACCG 3’</td>
</tr>
<tr>
<td>D266F fwd.</td>
<td>5’ CTACTCGGTCTCTTTGAGATCCACCTCCTATGCAGGAG 3’</td>
</tr>
<tr>
<td>D266F rev.</td>
<td>5’ CCATCTGCATAGGAGGCTCTTCAACAGGACCGAGTAG 3’</td>
</tr>
<tr>
<td>Y238F fwd.</td>
<td>5’ CTGGCCCTGTCCCCACACTAC 3’</td>
</tr>
<tr>
<td>Y238F rev.</td>
<td>3’ CACGACCGGACAAAGGGG 5’</td>
</tr>
<tr>
<td>Y245F fwd.</td>
<td>5’ ACCTGCTCTTTCCCGCCTG 3’</td>
</tr>
<tr>
<td>Y245F rev.</td>
<td>3’ GATGGACGAAGAGGCGG 5’</td>
</tr>
</tbody>
</table>
Supplementary figure 1 | MALDI-TOF spectrum of HPLC-purified ARTKme2QTARKS peptide. Calculated [M] 1173.69, found [M+H]+ 1174.6.

Supplementary figure 3 | 15% SDS-PAGE gel displaying Ni-NTA purified recombinant WT SGF29 and its variants. Amino acid substitutions are displayed in single letter code. From left to right: SGF29 wild-type, D266A, D266E, D266F, D266N, D266W, D266Y, Y238F and Y245F, respectively.
4

The chemical origin of the recognition of trimethyllysine by epigenetic reader domains


CHAPTER 4

Abstract

Objective: The trimethyllysine (Kme3) posttranslational histone modification is recognized by reader proteins containing an aromatic cage. This aromatic cage is capable of forming cation-π interactions with the methylammonium moiety of Kme3, which is suggested to be a major driving force in the Kme3 recognition event. To what extent this interaction actually contributes to overall binding has not been extensively investigated. Here we investigate the contribution of cation-π interactions to the association between histone 3 trimethyllysine 4 (H3K4me3) and epigenetic reader domains.

Approach: A dual approach was employed in which five H3K4me3 reader domain proteins, JARID1A, BPTF, JMJD2A, SGF29 and TAF3, were used to investigate the contribution of cation-π interactions on the reader-H3K4me3 recognition event. Isothermal titration calorimetry (ITC) was used as a tool to gain a full thermodynamic description of reader-histone association. In the first approach, we probed the effect of the trimethylammonium’s positive charge on the recognition by reader domains using an uncharged carbon methyllysine analog (K^-me3) as a ligand. In the second approach, we investigated the effect of the electronegativity of JARID1A’s aromatic cage by incorporating fluorinated tryptophan residues into the cage. Both approaches were complemented by structural and computational support.

Results: ITC analysis has indicated that binding of H3K^-4me3 to the five reader domains did not result in major reductions in binding affinities when compared to H3K4me3. Computational and structural support indicate that upon abrogation of the trimethylammonium’s charge, a compensatory effect takes place due to more favourable desolvation of the K^-me3 ligand when compared to Kme3. Complementary results were observed for the fluorinated JARID1A reader domain-H3K4me3 binding event. Here the reduction in electronegativity of the aromatic cage also did not result in reduced binding affinities or free energy of binding. Computational and structural support show that the charge reduction of the aromatic cage results in a compensatory mechanism where more favourable desolvation of the aromatic cage results in more favourable free energy of binding.

Conclusions: Compensation for the loss of favourable enthalpy, due to less favourable cation-π interactions, takes place when the charge of either the aromatic cage or Kme3 are reduced or abrogated. This compensation is due to more favourable desolvation of the aromatic cage or ligand, which results in more favourable free energy of binding and (partially) compensates the reduction in favourable enthalpy due to diminished cation-π interactions. These results indicate that both cation-π and desolvation of the aromatic cage are important contributors to reader domain-H3K4me3 association.
### 4.1 - Introduction

The 3 billion base pairs that make up the human genome are condensed and structured in such ways that optimal compaction is achieved while maintaining accessibility to the genome. The first level of such DNA compaction is achieved by wrapping the DNA around a histone octamer which consists of a H3-H4 tetramer and two H2A-H2B dimers, in order to yield a nucleosome.[1] The histone proteins that make up the nucleosome assembly are not fully structured into a tertiary or quaternary protein structure, however. Stretches of unstructured histone, dubbed “histone tails”, protrude from the nucleosome and are a canvas where various posttranslational modifications (PTM) such as acetylation, phosphorylation and methylation can be deposited.[2] The regulation of such histone PTMs is controlled by three types of proteins; reader proteins, writer proteins and eraser proteins. One of the most common histone PTMs is the methylation of lysine residues by lysine methyltransferases, as lysines can be dynamically methylated to a monomethylated (Kme1), dimethylated (Kme2) and trimethylated (Kme3) state. These methylation states can then be recognized with high specificity by reader domain containing proteins, which can facilitate downstream functions such as gene activation.[3, 4] In the past years it has become evident that due to the close relationship between epigenetics and gene control a variety of diseases, such as mixed-lineage leukemia (MLL), can occur if the epigenetic histone pattern is aberrant.[5-7] The JARID1A protein, for example, contains a histone 3 trimethyllysine 4 (H3K4me3) reader domain which is known to be involved in such MLL development.[7]

A recurring feature in reader proteins which specifically recognize Kme3 and Kme2 is a methyllsine recognition site called the aromatic cage. This structural element is able to recognize methyllsine with binding affinities dependent on the methylation status: Kme3 > Kme2 > Kme1 > Kme0.[8] An aromatic cage is usually composed of 2-4 aromatic residues which are capable of interacting with the Kme2/3 moiety via a combination of cation-π, hydrophobic and electrostatic interactions.[9] Readers with a cage flanking the Kme2/3 moiety on four sides are usually referred to as full cages, whereas cages flanking the Kme2/3 moiety at only two or three sides are referred to as half cages.[8] Cages which recognize Kme1/Kme2 are often supplemented with an acidic residue that is able to form a hydrogen bond with the charged nitrogen of the methyllsine residue. The position of the trimethylammonium group inside the aromatic cage, as demonstrated by structural determination of several reader-Kme3 complexes, suggests that the specific readout process is primarily driven by cation-π interactions, although charge independent interactions may also contribute to the overall binding. [9-12]
Protein-ligand association is a widely studied phenomenon and has been the subject of investigation for a broad spectrum of protein-ligand pairs. Generally, the nature of such interactions can be examined by either modifying the ligand or the protein itself. In the case of the association between a reader domain’s aromatic cage and the histones’ trimethyllysine, this indicates modifying the aromatic cage’s residues and the trimethylammonium’s chemical-physical properties.

Various methods, ranging from mutagenesis experiments to more complex approaches, have been employed in order to investigate protein-ligand interactions on protein level. With the more classical mutagenesis approach, the importance of single amino acids for a protein-ligand association can be investigated by introducing subtle or profound mutations.[13] Unfortunately, when employing such an approach one depends on the natural repertoire of 20 amino acids which are encoded for by our DNA. In a sense this provides plenty of opportunities for investigating subtle effects, but unfortunately such possibilities are limited by the relatively small number of canonical amino acids. Even though plenty of information may be gleaned from a Trp to Phe mutation, one is often confronted with a result induced by a combination of effects such as changes in charge, hydrophobicity, protein structure etc. Therefore, possibilities to introduce more convenient unnatural or non-canonical amino acids (NCAA) into proteins have been developed and several high quality reviews have been written on this topic.[14-16] One complex but increasingly more accessible method to incorporate NCAA into proteins has been achieved through expansion of the genetic code by using the amber stop codon, TAG, as a codon for a NCAA. Upon the generation of a highly efficient and orthogonal tRNA and tRNA synthethase pair, designed for a single specific NCAA, this method can be employed to incorporate NCAA at defined positions in a polypeptide chain by introducing the amber codon at the desired location(s). A more classical method to incorporate NCAA is the use of auxotrophic bacteria. Such bacteria lack the ability to produce a specific nutrient, in this case an amino acid, and are therefore dependent on the uptake of these nutrients from their surroundings. By restricting access to that specific natural amino acid and exposing the bacteria to a slightly modified unnatural amino acid variant, one can “force” the bacteria to accept these unnatural amino acid variants as substitutes for its natural paralogs. Of course this system has several drawbacks amongst which the fact that not all NCAA are readily accepted and incorporated, and the fact that the amino acid variant will replace all natural variants in a protein. Despite the complexity of these systems, incorporation of NCAA has been successfully used to investigate protein functions.[17, 18] Dougherty et al. have made extensive use of fluorinated amino acids in order to subtly probe the effect of cation-π interactions in the acetylcholine receptor, for example.[19] In our own investigations, which are described in this chapter, we have used auxotrophic bacteria for the incorporation of NCAA.
into the JARID1A reader domain in order to explore the chemical features of reader-H3K4me3 interactions.

In this study, a combination of two complementary chemical-biological approaches was used to probe the chemical basis of the recognition of H3K4me3 by reader proteins. The first approach focuses on employing an uncharged trimethyllysine carbon analog (K^cme3) in order to investigate the effect of the trimethyllysine's charge on the ability of reader proteins to bind H3K4me3. The second, and complementary, approach involves the incorporation of fluorinated tryptophan residues into the aromatic cage of JARID1A. By incorporating such fluorinated tryptophan residues, the charge or the aromatic cage is subtly affected without causing perturbations other than the desired charge effect. Both approaches probe the effect of charge (the trimethyllysine moiety’s positive charge and the π-electrons present in the aromatic rings of JARID1A’s aromatic cage, respectively) in order to address the question to what extent cation-π interactions are involved in the H3K4me3-reader binding event. These studies reveal that trimethyllysine-aromatic cage association is driven by both an electrostatic component, the cation-π interaction between the trimethyllysine moiety and the π-electrons in the aromatic cage, and the trimethyllysine mediated release of high energy water molecules, which are present in the aromatic cage.

4.2 - Results

4.2.1 – Association of reader domains with an uncharged carbon analogue of trimethyllysine

Specific favourable binding of the positively-charged side chain of Kme3 to the aromatic cage of reader proteins could, in principle, be a result of i) favourable solute-solute interactions (cation-π and CH-π interactions), ii) partial desolvation of the Kme3 side chain of histone tails (via the hydrophobic effect), and/or iii) desolvation of the aromatic cage of reader proteins. In order to elucidate the underlying chemical basis for the recognition of natural Kme3 by reader proteins, we have carried out detailed comparative studies for binding of 10-mer histone peptides that contain the positively charged Kme3, its neutral carbon analogue K^cme3, or a glycine residue that lacks the entire side chain at the 4-position of histone 3 (H3K4me3, H3K^c4me3, and H3G4, respectively; Figure 1A).[20]

We have chosen the simplest uncharged analogue, K^cme3, to directly probe the involvement of the proposed cation-π interactions in reader-histone associations, because it has virtually the same size, shape, and the polarizability as the positively-charged Kme3, but lacks the presence of the fixed positive charge.[10] Values for
volumes of Kme3 (160.2 Å³) and Kćme3 (158.2 Å³) indicate that, in the case that the binding mode is the same for both side chains, they should displace the same amount of water molecules from the protein site upon binding. We have chosen the 10-mer H3G4 peptide to explore the importance of the entire side chain of Kme3 on the association with reader proteins. The H3G4 variant directly probes the significance of the potential displacement of water molecules that are localized inside the aromatic cage of reader proteins.

**Thermodynamic analyses of reader-H3K4me3/H3Kć4me3 association**

We used isothermal titration calorimetry (ITC) to obtain full thermodynamic descriptions for binding of H3K4me3, H3KC4me3 and H3G4 peptides to five representative reader domain proteins that specifically recognize H3K4me3 (the PHD zinc fingers of JARID1A, BPTF, TAF3 and the tudor domains of SGF29 and JMJD2A) (Table 1, Figure 1B-C and data not shown).[7, 21-24] The five reader pockets are different in the aromatic cage composition and architecture, which allows us to examine the effect of individual constitution of the aromatic cage on binding differences. Comparative ITC experiments for the associations of H3K4me3 and H3KC4me3 showed that: i) the positively-charged H3K4me3 binds 2 to 33-fold stronger than the neutral H3KC4me3 to four out of five reader proteins (JARID1A, TAF3, BPTF, JMJD2A, Table 1). These four readers all contain a Trp residue as part of the aromatic cage, which has been shown to form the strongest cation-π interactions when compared to the other aromatic amino acids. ii) association of the Kme3 side chain with the aromatic cage is on average about 4.3 kcal mol⁻¹ more favourable in enthalpy than the association of the neutral KCme3 group to the same cage. iii) association of the Kme3 side chain is about 3.1 kcal mol⁻¹ less favourable in entropy than the association of the Kćme3 group to the same aromatic pocket. Collectively, these data provide evidence for the presence of the favourable cation-π interactions in the natural readout process, as exemplified by the enthalpy-driven association of the naturally-occurring Kme3 with the electron-rich aromatic cage of reader proteins. In contrast to other readers that contain at least one Trp residue, H3K4me3 and H3Kć4me3 bound to the tandem tudor domain of SGF29 with virtually indistinguishable thermodynamics of association, indicating the lack (or at least a minor contribution) of cation-π interactions in the association of Kme3 by the Tyr/Phe-containing half-cage of SGF29 (Table 1). This result is consistent with the well-established observation that the strength of cation-π interactions depends on the type of aromatic amino acid.[25-34] Studies on the related protein systems showed that Trp forms significantly stronger cation-π interactions with quaternary ammonium ions than do Phe or Tyr residues.[25, 26] For SGF29, the electrostatic interactions between Kme3 and D266, and between the positively-charged α-amino group of the N-terminal Ala residue and the H3A1-binding pocket importantly contribute to the overall binding affinity of H3K4me3.[35, 36]
Table 1 | Thermodynamic parameters for the associations of 10-mer H3K4me3, H3K4me3 and H3G4 peptides (ARTKme3QTRAKS, ARTKme3QTARKS and ARTGQTARKS) to five reader proteins.

<table>
<thead>
<tr>
<th></th>
<th>H3K4me3*</th>
<th></th>
<th></th>
<th>H3K4me3*</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_d$ (μM)</td>
<td>$ΔG^\circ$ (kcal mol$^{-1}$)</td>
<td>$ΔH^\circ$ (kcal mol$^{-1}$)</td>
<td>$-TΔS^\circ$ (kcal mol$^{-1}$)</td>
<td>$K_d$ (μM)</td>
<td>$ΔG^\circ$ (kcal mol$^{-1}$)</td>
</tr>
<tr>
<td>JARID1A</td>
<td>0.094</td>
<td>-9.6 ± 0.1</td>
<td>-11.0 ± 0.1</td>
<td>1.4 ± 0.1</td>
<td>0.34</td>
<td>-8.8 ± 0.1</td>
</tr>
<tr>
<td>TAF3</td>
<td>0.024</td>
<td>-10.4 ± 0.1</td>
<td>-10.9 ± 0.1</td>
<td>0.5 ± 0.2</td>
<td>0.79</td>
<td>-8.3 ± 0.1</td>
</tr>
<tr>
<td>BPTF</td>
<td>0.49</td>
<td>-8.6 ± 0.1</td>
<td>-13.1 ± 0.1</td>
<td>4.5 ± 0.1</td>
<td>0.76</td>
<td>-8.3 ± 0.1</td>
</tr>
<tr>
<td>SGF29</td>
<td>1.7</td>
<td>-7.9 ± 0.1</td>
<td>-7.7 ± 0.1</td>
<td>-0.2 ± 0.1</td>
<td>1.4</td>
<td>-8.0 ± 0.1</td>
</tr>
<tr>
<td>JMJD2A</td>
<td>0.94</td>
<td>-8.2 ± 0.1</td>
<td>-13.1 ± 0.2</td>
<td>4.9 ± 0.2</td>
<td>16</td>
<td>-6.5 ± 0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>H3G4†</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_d$ (μM)</td>
<td>$ΔG^\circ$ (kcal mol$^{-1}$)</td>
<td>$ΔH^\circ$ (kcal mol$^{-1}$)</td>
</tr>
<tr>
<td>JARID1A</td>
<td>88 ± 2</td>
<td>-5.5 ± 0.01</td>
<td>-2.1 ± 0.2</td>
</tr>
<tr>
<td>TAF3</td>
<td>36 ± 10</td>
<td>-6.1 ± 0.2</td>
<td>-2.5 ± 0.1</td>
</tr>
<tr>
<td>BPTF</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>SGF29</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>JMJD2A</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>

* Values obtained from 5-7 experiments. †Values obtained from triplicate measurements. Stoichiometry (histone peptide:reader protein, n) = 0.95-1.05. ND: not determined
Figure 1 | Binding analysis of H3 peptide variants to JARID1A. A) Structures of the positively-charged Kme3 and neutral K<sup>4</sup>me3 and G analogues; B) ITC curves of 10-mer H3K4me3, H3K<sup>4</sup>me3 and H3G4 histone peptides binding to the PHD3 domain of JARID1A; C) ITC curves of 10-mer -H3K4me3, H3K<sup>4</sup>me3 and H3G4 histone peptides binding to the PHD domain of TAF3.
We further examined the contribution of the entire Kme3 side chain to the overall binding associations with the aromatic cage of five reader proteins. ITC data showed that binding of 10-mer H3G4 to all five reader proteins was dramatically reduced (>500-fold) when compared to binding of the H3K4me3 counterpart, highlighting the importance of the entire side chain in the complexation process. More detailed thermodynamic analyses were only possible with JARID1A and TAF3, because both proteins bind to the reference H3K4me3 peptide with \( K_d \) values in sub-micromolar range and the H3G4 peptide had sufficient residual affinity for ITC characterization (Figure 1B-C, table 1). Overall, thermodynamic data revealed that i) binding of the entire side chain of the Kme3 contributes about -4 kcal mol\(^{-1}\) (i.e. about 40%) to the overall Gibbs free energy (\( \Delta G^\circ \)); ii) favourable enthalpy provides a dominant contribution (\( \sim -8.5 \) kcal mol\(^{-1}\)) to the binding of the entire Kme3 side chain to the aromatic cage; and iii) entropy of binding becomes more favourable (\( -T\Delta S^\circ = -4.5 \) kcal mol\(^{-1}\)) for H3G4 relative to H3K4me3. In addition to the thermodynamic results for binding of H3K4me3, these results indicate that favourable cation-π interactions are not solely responsible for strong binding affinity of H3K4me3, but that other types of solute-solute interactions and reader/histone desolvation could also play an important role in the specific readout of Kme3.

### 4.2.2 - Incorporation of fluorinated tryptophan residues into JARID1A's aromatic cage

The half cage reader domain of JARID1A specifically recognizes H3K4me3 with sub-micromolar affinity.[7] Its aromatic cage is composed of only two tryptophan residues which bind the Kme3 moiety in a surface-groove binding mode. Since these are the only two tryptophans in the entire JARID1A-PHD domain, an auxotrophic E. coli strain could be used to incorporate various fluorinated tryptophan residues directly into its aromatic cage. The presence of only two tryptophans eliminates the risk for perturbation of the reader domains’ structure as a result of additional fluorinated tryptophan residues outside the aromatic cage. This allowed us to investigate the effect of the aromatic cage’s \( \pi \)-electrons on trimethyllysine recognition by fluorinating the tryptophans indole rings at position 5 (5F), position 6 (6F) or positions 5 and 6 (5,6-dF) simultaneously (Figure 2A-B). The reason why we chose fluorination of the tryptophan residues was because i) fluorine’s electronegativity can be exploited in order to reduce the electronegativity of tryptophan’s indole ring, ii) fluorine and hydrogen are comparable in size, resulting in minimal structural perturbations of the protein and iii) the JARID1A-PHD domain’s aromatic cage is composed of only two tryptophan residues, which are the only tryptophan residues in the JARID1A reader domain.
Figure 2 | Crystal structures, and biophysical characteristics of JARID1A protein expressed in E. coli Castellani and Chalmers. **A)** JARID1A-H3K4m3e peptide complex (PDB ID: 3GL6). The H3K4me3 peptide is shown in grey, tryptophan residues present in the aromatic cage are shown in red and Zn$^{2+}$ ions are depicted as grey spheres. **B)** JARID1A aromatic cage (PDB ID: 3GL6). The tryptophan residues which constitute the aromatic cage are indicated in red with their respective protons in white. The trimethyllysine residue is shown in grey-blue. **C)** 15% Tris-Tricine SDS-PAGE showing the purified, untagged JARID1A PHD domain. From left to right: Wild type-JARID1A, 5F-JARID1A, 6F-JARID1A and 5,6-diF-JARID1A. **D)** ESI mass spectra show the masses of the purified JARID1A PHD domains. The wild type protein has a mass of 7170.1 Da whereas the 5F and 6F alloproteins show a mass increase of +35.9 and +35.8 Da respectively and 5,6-di-F shows an increase of +71.9 Da. **E)** CD spectra taken for the purified PHD domains over a range of 260-190 nm. Black: Wild type protein, red: 5F-JARID1A, green: 6F-JARID1A and blue: 5,6-diF-JARID1A.
Preparation of alloproteins
We successfully expressed and purified the JARID1A reader domain in the auxotrophic, tryptophan deficient *E. coli* Castellani and Chalmers strain in a method similar to the procedure described by Budisa et al.[37]. Briefly, the auxotrophs containing the GST-JARID1A construct were grown in well-defined new minimal medium (NMM) composed of nutrients, trace metals and the 20 naturally occurring individual L-amino acids. Upon reaching their optimal growth phase, the cells were centrifuged, washed, supplemented with fresh NMM and protein expression was induced. By selectively adding culture medium devoid of tryptophan instead of regular medium during induction, the auxotrophic strain accepted unnatural tryptophan variants as a nutrient. These unnatural amino acids were then efficiently incorporated into the proteins which were produced. The GST tag was cleaved off with TEV-protease and the JARID1A alloproteins were further purified using standard biochemical techniques in order to obtain proteins of high purity in reasonable yields (Figure 2C). Interestingly though, the wild type construct expressed markedly less well than its alloprotein counterparts. We successfully produced the protein variants with the three tryptophan analogs: 5F-tryptophan, 6F-tryptophan and 5,6-diF-tryptophan into the JARID1A reader domain. Additional attempts to incorporate other tryptophan analogs, namely 4,5,6,7-tetraF-tryptophan and tryptophan in which the indole ring’s nitrogen had been replaced with sulfur were not successful, the latter is in agreement with previous observations.[14, 38]

Alloprotein characterization
After purifying the JARID1A proteins, we evaluated the obtained proteins using ESI-MS (Figure 2D). ESI mass spectra confirmed that the wild-type JARID1A PHD domain protein indeed had a mass of 7170.1 Daltons and that the incorporation of fluorinated tryptophan residues led to the expected mass increase of 18 Da per fluorine (7206.0; 7205.9 and 7242.0 Da for 5F-, 6F- and 5,6-diF-tryptophan respectively). Additionally, our circular dichroism (CD) spectra (Figure 2E) suggested that the structures of the proteins containing the various tryptophan analogs are identical to that of the wild type (WT) protein.

These results lead us to conclude that the JARID1A alloproteins all have similar folds and that no structural perturbations have been introduced by incorporating fluorinated tryptophan. This finding is further supported by the fact that differential scanning fluorimetry (DSF) showed no marked decrease in the alloprotein’s melting temperature when compared to the wild type protein with melting temperatures of 51.0 ± 1.7 °C; 51.8 ± 0.7 °C; 53.9 ± 0.1 °C and 51.9 ± 0.4 °C for WT, 5F-, 6F- and 5,6-diF-tryptophan respectively. Together, these data show that the produced alloproteins have similar biophysical properties when compared to the WT protein, suggesting that the
incorporation of fluorinated tryptophan analogs does not have a negative impact on JARID1A’s structure.

As mentioned above, the WT protein expressed markedly less well in the auxotrophic strain when compared to its alloprotein counterparts. CD-spectroscopy and ESI-MS analysis conducted on the WT protein expressed in both *E. coli* Hosetta BL21 (DE3) pLysS (hereafter referred to as BL21) and *E. coli* Castellani and Chalmers (hereafter referred to as AUX) showed that both expression strains produced proteins with identical masses and tertiary structures (Supplementary figures 1 & 2). It can therefore be concluded that both expression systems produce WT protein with identical structural properties. Due to the marked better expression of the WT protein in the BL21 strain when compared to the auxotrophic strain, it was decided to use the WT protein expressed in the BL21 strain for our thermodynamic analyses.

**Thermodynamic analyses of alloprotein-H3K4me3 association**

After characterizing the JARID1A proteins, the effect of JARID1A fluorination on binding to H3K4me3 peptides (sequence: ARTK_{me}QTARKS; Table 2) was assessed by ITC measurements. First, a comparison was made between WT protein expressed in the auxotrophic strain and WT protein expressed in the BL21 strain, which had also been used for JARID1A expression in our previous studies.[36] No significant differences in K_{eq} were observed, with a K_{eq} of 54 ± 6 nM for BL21 wild type and 48 ± 8 nM for auxotrophic wild type protein. Furthermore, thermodynamic parameters such as ΔG° (BL21: -9.92 ± 0.07 vs. AUX: -9.99 ± 0.10), ΔH° (BL21: -11.59 ± 0.12 vs. AUX: -11.91 ± 0.01) and -ΔS° (BL21: 1.68 ± 0.18 vs. AUX: 1.92 ± 0.11) did not differ significantly. In conjunction with the previously described CD and ESI-MS data this further shows that both WT proteins are identical with respect to their biophysical and binding properties (Supplementary figures 1 & 2).

Next, we analyzed the alloproteins. 5F-JARID1A showed a decreased enthalpy when compared to BL21 WT-JARID1A with a ΔΔH° of +1.01 ± 0.14 kcal mol⁻¹. The decrease in enthalpy was completely compensated by an increase in entropy, however, with a -ΔΔS° of -1.02 ± 0.28 kcal mol⁻¹. This compensation resulted in indistinguishable ΔG° values with a ΔΔG° of 0.07 ± 0.16 kcal mol⁻¹ and identical binding affinities of 48 ± 8 nM for BL21-JARID1A and 49 ± 12 nM for 5F-JARID1A, respectively.

Also for 6F-JARID1A no large differences in thermodynamic binding parameters were observed when compared to BL21 WT: ΔK_{eq} of +3 ± 2 nM, ΔΔG° of +0.01 ± 0.19 kcal mol⁻¹. Nevertheless, a small decrease in ΔΔH° of -0.47 ± 0.18 kcal mol⁻¹ and a small increase of -ΔΔS° of +0.49 ± 0.32 kcal mol⁻¹ were found. As observed for 5F-JARID1A, the change in enthalpy was compensated by a change in entropy,
leading to ΔG° and Kₐ values of -9.91 ± 0.18 kcal mol⁻¹ and 57 ± 17 nM, respectively, which are again identical to those obtained for the wild type protein.

The 5,6-diF-JARID1A alloprotein displayed the largest change in Kₛ and ΔG°, however, these changes were still minor as the Kₛ was only increased by a factor of 1.3 and ΔΔG° was only -0.20 ± 0.15 kcal mol⁻¹ when comparing BL21 WT and 5,6-diF-JARID1A. Furthermore, ΔΔH° and -TΔS°, relative to BL21 WT, were found to be only -0.06 ± 0.16 kcal mol⁻¹ and +0.2 ± 0.32 kcal mol⁻¹, respectively. Similar to the data obtained for 5F- and 6F-JARID1A, the values obtained for 5,6-diF-JARID1A-H3K4me3 association are indistinguishable from the BL21 WT data.

Taken together, these results show that the reduction in electronegativity of the indole rings, which make up JARID1A’s aromatic cage, does not reduce the protein’s binding affinity for H3K4me3. This suggests that cation-π interactions are not solely responsible for the binding of JARID1A to H3K4me3. A comparison of these data with those obtained for the H3K³me3-JARID1A interaction indicates that ΔG° values are not drastically affected by a reduction of the electrostatic potential of the trimethyllysine moiety nor by that of the aromatic cage. Considering the fact that the trimethylation status of H3K4 increases the affinity of JARID1A for the H3K4me3 sequence over 27-fold compared with H3K4me0, the 3.6-fold loss of binding affinity due to the loss of the ligand’s charge in the case of H3K²⁴me3 binding is small.[7]
4.2.3 - Structural assessment of reader-histone complexes

The thermodynamic information described above is in part the result of the docking of the trimethyllysine side chain into the aromatic cage of the various reader proteins, and the interaction of several of the other H3K4me3 peptide amino acid residues with the reader domain. In order to further understand and explain the thermodynamic data, we aimed to obtain ultrastructural information regarding the reader domains complexed with H3K4me3 peptide as well as the alloprotein complexed with H3K4me3 peptide.

X-ray structure determination of reader-H3K4me3 complexes

To aid rationalization of the loss of binding affinity resulting from the loss of positive charge at trimethyllysine 4 of H3, we solved the crystal structures for complexes of H3K4me3 with JARID1A, TAF3 and SGF29 by X-ray diffraction. A resolution of 1.6-2.8 Å was obtained (Figure 3 and supplementary figure 3.) and the obtained crystal structures of JARID1A PHD-H3K4me3, TAF3 PHD-H3K4me3, and SGF29 tandem Tudor-H3K4me3 complexes have been deposited into the Protein Data Bank under accession codes 5C11, 5C13 and 5C0M respectively.

All three reader-H3K4me3 structures clearly illustrated that the uncharged side chain of K4me3 is positioned well inside the aromatic cages of JARID1A, TAF3, and SGF29, in virtually the same binding mode as the positively-charged Kme3 (Figure 3A-C). The calculated average values of the root-mean-squared deviation (RMSD) for binding of “K4me3” and “Kme3”-aromatic cage pairs were: 0.124 Å for JARID1A, 0.261 Å for TAF3, and 0.108 Å for SGF29, respectively, suggesting essentially the same complexation mode engaging in aromatic pocket residues upon binding to neutral K4me3 (Figure 3D-F). In all three complexes, the K4me3 histone peptide binds to an electrostatically negative surface with the long side chain positioned in a surface groove formed by the caging residues (Figure 3G-I). Upon the formation of the JARID1A-H3K4me3 complex, the buried solvent accessible surface area (SASA) of K4me3 (hydrogen atoms added) is 160.6 Å², which accounts for 38.8% of the total SASA of K4me3, as compared with K4me3 binding to JARID1A with a buried SASA of 163.8 Å², which equals 39.5% of total SASA. Similar features have also been observed for binding of H3K4me3 and H3K4me3 to TAF3 with 48.3% buried SASA for H3K4me3 and 50.1% buried SASA for H3K4me3, and to SGF29 with 48.6% and 47.3% buried SASA for H3K4me3 and H3K4me3, respectively. Therefore, a difference in ligand binding mode is not the cause of the reduction in binding affinity for readers regarding K4me3 when compared to Kme3.
Similar to the work with H3K4me3, we attempted to obtain crystal structures for the alloprotein-H3K4me3 complexes. Unfortunately, the crystallization process did not yield crystals which were suitable for X-ray diffraction. Therefore, we employed $^{19}$F-NMR spectroscopy to compare the free and H3K4me3-complexed forms of the JARID1A alloproteins (Figure 4 and supplementary figures 4 and 5). NMR spectra
w
ere obtained for 5F- and 6F-JARID1A complexed with 10-mer H3K4me3 peptide. Upon bindin
go of H3K4me3 to 5F-JARID1A we observed down-field shifts with one peak shifting approximately +3 ppm whereas its second peak shifted +6 ppm when compared to free 5F-JARID1A. For 6F-JARID1A we observed down-field shifts of approximately +1 ppm upon H3K4me3 binding. These results indicate that that the positively charged trimethyllysine moiety present in the H3K4me3 peptide is likely to interact with the fluorinated tryptophans incorporated into the aromatic cage.[39, 40]

Interestingly, the 5F-JARID1A protein signal resided at approximately -126 ppm whereas the 6F-JARID1A signal resided at -122 ppm. The chemical shift observed for 5F-JARID1A was also larger than that for 6F-JARID1A. This difference can be rationalized by the fact that the crystal structure for the wild type JARID1A-H3K4me3 complex (Figure 2) shows that the H5 protons of the tryptophan residues point towards the reader protein whereas the H6 protons are directed away from the reader protein. Therefore, it is likely that, upon JARID1A-H3K4me3 complexation, both the 5F- and 6F- fluorinated tryptophans will be differently affected by the trimethyllysine moiety.

Additional CD spectroscopy analysis confirmed that the proteins remained stable during the NMR measurements performed at 15°C (Supplementary figure 6). Additionally, the CD analyses indicated a small change in the protein's structure upon alloprotein-H3K4me3 complexation. A small shift in mean residual ellipticity (MRE) between 215-240 nm corresponds to a more extensive β-sheet conformation.[41] This is in line with the finding that H3K4me3 peptides form a third antiparallel β-strand when complexed with PHD domains, this is also visible in the reported JARID1A-H3K4me3 structure (Figure 1E).[7, 42]
4.2.4 - Quantum-chemical analyses in the gas and aqueous phase

In order to provide further support for the thermodynamic data described above and to explore the individual thermodynamic parameters which are involved in the reader-Kme3 interaction, computational analyses for the JARID1A reader domain were conducted. The individual energetic components of the reader-Kme3 interaction were calculated from the crystal structures previously obtained. The reasons why JARID1A was chosen for the quantum-chemical analyses are i) the JARID1A-H3K4me3 and JARID1A-H3K4me3 crystal structures have been solved, ii) the JARID1A aromatic cage is composed of only 2 residues and iii) these residues are located in a β-sheet region and are therefore structurally quite rigid.

**JARID1A-H3K4me3 analysis**

Our aim was to elucidate the nature and selectivity of the non-covalent interactions between the JARID1A aromatic cage that consists of two tryptophan residues (hereafter designated as the TRP2 fragment) and the Kme3 versus K+me3 side chain of the histone peptide. To this end, we have quantum-chemically characterized the energetics and bonding mechanism in two model complexes. The model complexes consist of those moieties of the JARID1A-H3K4me3 and JARID1A-H3K4me3 X-ray structures that give rise to the intermolecular interaction in the reader-histone complexes with respect to the aromatic cage. To simulate the structural rigidity that is imposed by the protein backbone in the full protein system, the TRP2 fragment is kept frozen to the X-ray structure, both as a separate fragment and in the complexes. Geometries of the optimized model systems differ only very slightly from the X-ray structures.

Our computations show that, in line with experimental data, there is an energetic preference of approximately 2 kcal mol⁻¹ for the JARID1A-Kme3 over the JARID1A-K+me3 model complex with bond energies ΔE(aq) of -10.2 and -8.4 kcal mol⁻¹, respectively (Table 3). Bonding analyses reveal that the bond energies ΔE(aq) associated with the molecular recognition processes of Kme3 versus K+me3 in water are essentially identical with the corresponding instantaneous interaction energies ΔE_int(aq) of -10.3 and -8.7 kcal mol⁻¹, respectively. The reason is that complexation only very slightly changes the geometry of the Kme3 and K+me3 side chains as a result of which the associated deformation strain is negligible, i.e., 0.1 and 0.3 kcal mol⁻¹, respectively. The intrinsic preference for Kme3 over K+me3, that is, the interaction energy ΔE_int between the same structures but in the absence of the solvent, is even more in favor of the former with values of -27.6 and -10.9 kcal mol⁻¹, respectively (Table 3). The significantly stronger interaction energy of Kme3 is, however, strongly attenuated by the desolvation incurred upon binding, which is significantly more favourable for Kme3. Thus, solvent effects destabilize the JARID1A-Kme3 complex by +17.3,
Table 3 | Quantum-chemical bonding analysis in the JARID1A TRP2 + H3K4me3 and TRP2 + H3K4me3 systems in aqueous solution.

<table>
<thead>
<tr>
<th></th>
<th>TRP2 + H3K4me3</th>
<th>TRP2 + H3K-4me3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔE(aq)</td>
<td>-10.2</td>
<td>-8.4</td>
</tr>
<tr>
<td>ΔE(aq) strain</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>ΔE(aq) int</td>
<td>-10.3</td>
<td>-8.7</td>
</tr>
<tr>
<td>ΔE(desolv) int</td>
<td>17.3</td>
<td>2.2</td>
</tr>
<tr>
<td>ΔE int</td>
<td>-27.6</td>
<td>-10.9</td>
</tr>
<tr>
<td>ΔE Pauli</td>
<td>20.8</td>
<td>14.1</td>
</tr>
<tr>
<td>ΔV stat</td>
<td>-15.0</td>
<td>-5.4</td>
</tr>
<tr>
<td>ΔE el</td>
<td>-13.0</td>
<td>-3.6</td>
</tr>
<tr>
<td>ΔE disp</td>
<td>-20.4</td>
<td>-16.0</td>
</tr>
<tr>
<td>d(H_Na-C_TRP-EM)</td>
<td>3.38</td>
<td>3.15</td>
</tr>
<tr>
<td>d(H_Na-C_TRP-EM)</td>
<td>2.78</td>
<td>3.16</td>
</tr>
</tbody>
</table>

Energies presented in kcal mol⁻¹, distances in Å

whereas the desolvation penalty in the JARID1A-K⁺me3 complex is only +2.2 kcal mol⁻¹. The reason for this large difference can be attributed to the removal of solvent (desolvation) around the positive charge of the Kme3 side-chain ammonium group. Note that the stronger binding in JAHID1A-Kme3 causes a reduction in the bond distances, resulting in a computed Pauli repulsion energy between closed-shells that is +6.7 kcal mol⁻¹ more repulsive for this more stable JARID1A-Kme3 complex.

Interestingly, although dispersion ΔEdisp is the largest contributor to the reader-histone interaction, it contributes only 4.4 kcal mol⁻¹ to the 16.7 kcal mol⁻¹ difference in ΔEint between JARID1A-Kme3 (-27.6 kcal mol⁻¹) and JARID1A-K⁺me3 (-10.9 kcal mol⁻¹) (Table 3). Instead, the difference in stability between JARID1A-Kme3 and JAHID1A-K⁺me3 mainly originates from the electrostatic (ΔVeStat) and orbital interaction (ΔEel) terms that favor the complex with Kme3 by 9.6 and 9.4 kcal mol⁻¹, respectively. The more attractive ΔVeStat in the case of Kme3 goes hand in hand with the significantly more positive charge on all atoms in the Kme3 ammonium.[43] The nitrogen atom in Kme3 carries a positive charge of +59 milli-a.u., which has to be compared with the negative charge of -40 milli-a.u. on the structurally analogous carbon atom in the overall neutral K⁺me3. Importantly, the hydrogen atoms of the trimethylammonium group of Kme3 are also significantly more positively charged than the corresponding
ones of the tert-butyl group in K\textsuperscript{e}me3. For example, the hydrogen atom closest to the reader’s TRP2 fragment has an atomic charge of +84 and +29 milli-a.u. in Kme3 and K\textsuperscript{e}me3, respectively.

**Alloprotein-H3K4me3 analysis**

We have performed a series of analyses, analogous to those described above, for 5F-, 6F- and 5,6-diF-TRP2 with Kme3 (Table 4). First, bond energies hardly change for 5F-TRP2-Kme3, 6F-TRP2-Kme3, and 5,6-diF-TRP2-Kme3 systems (\(\Delta E(\text{aq}) = -10.2 - -10.4\) kcal mol\(^{-1}\)). The same has been observed for both the deformation strain and interaction energies, with a maximum change of 0.4 kcal mol\(^{-1}\). Thus, also in the presence of fluoro substituents on TRP2, complexation only very slightly changes the geometry of the Kme3 side chain.

However, changes become apparent when these interactions are analysed without water. Fluorination of TRP2 causes a destabilization of the intrinsic (i.e., in vacuum) interaction between Kme3 and TRP2 which varies from \(\Delta E_{\text{int}} = -27.6\) to -25.4, -24.9 and -22.4 kcal mol\(^{-1}\) for TRP2-Kme3, 5F-TRP2-Kme3, 6F-TRP2-Kme3 and 5,6-diF-TRP2-Kme3, respectively (Table 4). This destabilizing trend is counteracted by the less and less unfavourable desolvation energies \(\Delta E(\text{desolv})_{\text{int}}\), which range from +17.3

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Quantum-chemical bonding analysis of TRP2-Kme3 and fluorinated TRP2-Kme3 JARID1A systems in aqueous solution.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TRP2 + Kme3</td>
</tr>
<tr>
<td>(\Delta E(\text{aq}))</td>
<td>-10.2</td>
</tr>
<tr>
<td>(\Delta E(\text{aq})_{\text{strain}})</td>
<td>0.1</td>
</tr>
<tr>
<td>(\Delta E(\text{aq})_{\text{int}})</td>
<td>-10.3</td>
</tr>
<tr>
<td>(\Delta E(\text{desolv})_{\text{int}})</td>
<td>17.3</td>
</tr>
<tr>
<td>(\Delta E_{\text{int}})</td>
<td>-27.6</td>
</tr>
<tr>
<td>(\Delta E_{\text{Pauli}})</td>
<td>20.8</td>
</tr>
<tr>
<td>(\Delta V_{\text{Pauli}})</td>
<td>-15.0</td>
</tr>
<tr>
<td>d(H\text{Me}<em>C</em>{(TRP/5MR)})</td>
<td>3.38</td>
</tr>
<tr>
<td>d(H\text{Me}<em>C</em>{(TRP/5MR)})</td>
<td>2.78</td>
</tr>
</tbody>
</table>

Energies in kcal mol\(^{-1}\), distances in Å
to +14.7, +14.1 and +11.9 kcal mol⁻¹ along the same series. In all complexes, the largest contribution to the desolvation energy stems from removal of solvent around the positively charged ammonium group of Kme3. But this term is constant. The trend of less and less unfavourable desolvation energies comes from the fact that the aromatic cores of fluorinated TRP2 are less negative as charge is pulled out of the core and into the electronegative fluoro substituents. The latter are and remain solvated upon complexation. It is the surface of the aromatic core that is desolvated when Kme3 forms a complex. As the net charge on these cores is reduced upon fluorination, the interaction with the solvent that is expelled becomes weaker and desolvation thus less destabilizing.

As stated above, ΔE_{\text{rel}} weakens from -27.6, -25.4 to -24.9 and -22.4 kcal mol⁻¹ for TRP2-Kme3, 5F-TRP2-Kme3, 6F-TRP2-Kme3, and 5,6-diF-TRP2-Kme3, respectively.

Next, we have carried out the energy decomposition analysis of the interaction energy. First, it is observed that aforementioned weakening is not the result of the Pauli repulsion term, which remains quite constant (ΔE_{\text{Pauli}} = 19.3 - 20.2 kcal mol⁻¹), with a maximum difference of 1.5 kcal mol⁻¹ compared to unfluorinated TRP2 system. The same constant behaviour is given by the dispersion correction term, which also gives a maximum difference of 0.3 kcal mol⁻¹ with fluorination. As previously observed for TRP2-Kme3, this ΔE_{\text{disp}} term contributes the largest to the interaction, however it is observed that it is not determinant.

Thus, it turns out that the trend in the interaction energy originates from the electrostatic attraction ΔV_{\text{elect}}. This attraction is less favourable by 2.9 - 3.3 kcal mol⁻¹ for 5F- and 6F-TRP2, and by 6.1 kcal mol⁻¹ for 5,6-diF-TRP2. The weakening of the electrostatic potential is caused by the fact that the electronegative fluoro substituents pull electronic charge density away from the aromatic core, thus causing these rings to be less negatively charged. This is clearly observed by comparison of the two extreme systems, TRP2-Kme3 and 5,6-diF-TRP2-Kme3, that present the strongest effect. The former only presents one positively charged carbon atom in the six-membered ring, whereas the latter presents four (data not shown). Such more positively charged six-membered ring causes a less favourable electrostatic interaction with trimethyllysine. The effect is intermediately strong for the 5F-TRP2-Kme3 and 6F-TRP2-Kme3 systems which have two positively charged carbon atoms in the ring skeleton.

Taken together, the computational analyses for both the JARID1A-H3K4me3 and fluorinated JARID1A-H3K4me3 interactions are in line with the obtained thermodynamic data. From the thermodynamic analyses we have seen that little to no change in binding affinity was observed upon JARID1A-H3K4me3 and al洛protein-H3K4me3 complexation when compared to the JARID1A-H3K4me3 binding event. These
results have been confirmed by quantum-chemical analysis: upon reduction of the electrostatic component, the loss in binding energy is compensated by more favourable desolvation energy resulting in retention of binding affinities.

4.2.5 - WaterMap calculations
As the quantum-chemical analysis experiments revealed that favourable desolvation energy was mainly responsible for the preservation of the JARID1A reader-ligand interaction despite the reduction of the electrostatic component, a further analysis was performed on the presence of water in the reader’s aromatic cage in order to further explain the contribution of water in the reader-ligand binding event.

Header-K^{+}-me3 WaterMap calculations
WaterMap calculations were performed for JARID1A, BPTF, JMJD2A, TAF3 and SGF29 readers in order to evaluate the contribution of aromatic cage desolvation to the affinity of Kme3 and K^{+}-me3 for reader proteins (Figure 5 and supplementary table 1). WaterMap computes thermodynamic quantities (free energy, enthalpy and entropy).

Figure 5 | WaterMap calculations for the solvation of aromatic cages of reader proteins. a) JARID1A, b) SGF29, c) TAF3, d) JMJD2A, e) BPTF. Superimposed Kme3 side chain and water molecules are presented as green stick and grey spheres. Numbers adjacent to grey spheres represent the value of the free energy (DG) for individual water molecule; (f) Thermodynamic parameters for the solvation of the aromatic cages of five reader proteins used in this study.
for simulated water molecules around a protein binding site using explicit solvent molecular dynamics (MD) simulation and thermodynamic characterization. In short, regions of high solvent density from the MD simulations are clustered into “hydration sites”, and thermodynamic quantities for these sites are calculated using inhomogeneous solvation theory.[44, 45] For all five reader proteins, two to four high-energy hydration sites were identified within the aromatic cage (Figure 5A-E). These hydration sites are displaced from the aromatic cage by both the Kme3 and K2me3 side chain, but not by the H3G4 peptide. The total free energy contributed by desolvating the aromatic cage (determined as the difference in WaterMap scores between Kme3 and Gly) ranges from 4.3 kcal mol-1 for JARID1A to 8.7 kcal mol-1 for SGF29. Depending on the composition of the cage, this free energy reward can be both entropically and enthalpically driven (Figure 5F, Supplementary Figures 7 and 8). For example, both TAF3 and JMJD2A contain an Asp residue that can form hydrogen bonds with the binding site water molecules, resulting in more favourable enthalpy of the hydration sites in the cage, hence more unfavourable change in enthalpy upon displacing those waters upon Kme3/K2me3 binding. On the other hand, the BPTF cage is completely surrounded by aromatic residues, producing an enthalpically unfavourable environment for water and therefore a favourable free energy change from water displacement upon Kme3/K2me3 binding.

**Alloprotein WaterMap calculations**

Similar to the WaterMap computations performed on JARID1A, JMJD2A, BPTF, TAF3 and SGF29 with K2me3 peptides, the fluorinated JARID1A alloproteins were further investigated by WaterMap analysis (Figure 6, supplementary table 2 and supplementary figures 9 & 10). For WT JARID1A four high energy hydration sites were identified whereas the alloproteins have only three hydration sites which can be displaced upon Kme3 complexation. The total free energy contributions from desolvation were calculated to be -4.9 kcal mol-1 for WT protein and -8.0, -7.8 and -6.6 kcal mol-1 for 5F-, 6F-, and 5,6-diF-JARID1A, respectively (Supplementary table 2). This increase in free energy contribution appears to result from less favourable enthalpy of solvation. This result indicates that fluorination of the aromatic cage results in a more favourable free energy change upon displacement of water molecules by Kme3. These results are supporting the quantum-chemical analysis of the JARID1A alloprotein-ligand binding event as these computations predicted a compensation mechanism due to less favourable electrostatics and more favourable desolvation (Table 4). It should be noted, however, that the quantum-chemically computed trend of increasingly more favourable desolvation upon 5,6-difluorination was not reflected by the WaterMap calculations. Despite this fact, the WaterMap calculations do support the general notion of energetically more favourable desolvation upon fluorination of the tryptophan residues constituting JARID1A’s aromatic cage.
4.3 - Discussion

The advances in experimental and theoretical tools developed in the past decade have enabled more extensive analysis of the origins of several genuinely important biomolecular recognition phenomena, including the molecular basis of the hydrophobic effect(s) in protein-ligand interactions and the fundamentals of the receptor-neurotransmitter interactions in neurochemistry.[32, 46]

The work described in this chapter comprehensively examined the origin of the biomolecular association between naturally-occurring trimethyllysine-containing histone proteins and several H3K4me3 reader domains that are involved in epigenetic gene regulation processes. We used a combinatorial approach, supported by high-resolution structural analyses and theoretical calculations, to elucidate the molecular basis of this reader-histone interaction.

Based on studies of proteins that possess the aromatic cages for the recognition of positively-charged methylammonium groups, it has been suggested that epigenetic readers recognize Kme3 via cation-π interactions.[10, 32, 47, 48] Our integrated thermodynamic, structural, and computational studies confirm the presence of favourable cation-π interactions in the readout of H3K4me3 by reader domains of JARID1A, TAF3, BPTF and JMJD2A. Previous examination of the recognition of
neutral K\textsuperscript{C}me3 by the HP1 chromodomain, a reader of H3K9me3 which contains an aromatic cage comprising two tyrosine and one tryptophan residues, revealed that HP1 bound H3K\textsuperscript{C}9me3 with substantially lower affinity than H3K9me3.[11] This suggests that the positive charge of Kme3 is crucial for the HP1-H3K9me3 interaction. Comprehensive structural data on JARID1A, TAF3 and SGF29 in complex with H3K\textsuperscript{C}4me3, as described in this chapter, provides clear evidence that the K\textsuperscript{C}me3 side chain is positioned inside the aromatic cages of these three reader proteins in the same manner as the positively-charged Kme3 (Figure 3) and thus enable us to interpret the calorimetric data (Table 1). Out of three possible mechanisms (that should always be considered in the interpretation of any protein-ligand system) that govern the recognition of Kme3 by reader proteins, namely: solute-solute interactions, desolvation of the ligand (in this case Kme3) and desolvation of the protein (in this case the aromatic cage), we can exclude desolvation of the Kme3 side chain, because charged residues are highly soluble in aqueous media and have to pay a big desolvation penalty to become desolvated. In this regard, it is essential that the energetically unfavourable desolvation of Kme3 is fully compensated (or more correctly overcompensated) by energetically favourable protein-ligand interactions and protein desolvation, in order to provide a strong binding force for the specific recognition of Kme3 by reader proteins. Based on ITC experiments, our observed enthalpy-driven association of positively-charged Kme3 (relative to K\textsuperscript{C}me3) to the electron-rich aromatic cage of several reader proteins has its molecular origin in strong cation-π interactions. In addition, the methylene groups of the side chain of Kme3 located within van der Waals distance of the aromatic cages, contribute to the overall binding affinity via weaker, but still favourable, CH-π interactions.[49, 50]

Furthermore, our results have shown that the incorporation of fluorinated tryptophan residues into aromatic cage of the JARID1A reader domain yielded stable proteins with structures and biophysical properties similar to those of the wild type protein. Our NMR data indicated that the alloprotein aromatic cage-Kme3 association is likely to be similar to that of the WT protein studied previously.[7, 20, 36] This notion was supported by CD spectroscopy which indicated that upon alloprotein-ligand complexion a more β-sheet like structure was formed, a common feature for PHD finger reader proteins.[7, 42] ITC data have shown that, upon fluorination of JARID1A’s aromatic cage, no reduction in binding affinity and free energy of binding occured. In support, our computational work suggests a desolvation driven compensatory mechanism when fluorinated JARID1A alloproteins bind to H3K4me3. These results indicate that a reduction in the electrostatic component of the cation-π interaction is compensated by more favourable desolvation of the aromatic cage. This result is complementary to the results obtained with the JARID1A-K\textsuperscript{C}me3 binding event.
Quantum mechanical studies, furthermore, reveal that reader-Kme3 association has the strongest dispersion contribution (similar to reader-K$^2$me3), but that the differences in binding affinities between Kme3 and K$^2$me3 are primarily a result of disparities in electrostatic interactions and orbital interactions (data not shown).[20] Biophysical studies yielded similar conclusions for 5F-, 6F- and 5,6-diF-JARID1A: while electrostatic interactions are reduced upon fluorination of the aromatic cage, the reduction in energy of solute-solute interactions is compensated by more favourable desolvation energies.

Even though most biologically relevant interactions occur in aqueous media and water is known to have an important contributing factor in many biochemical interactions, the contribution of water is often difficult to investigate.[51, 52] Despite the difficulties involved, however, many researchers are focusing on the contribution of this important molecule to biochemical interactions.[53-56] Our observations that binding affinities of H3G4 with JARID1A and TAF3 are drastically reduced when compared to H3K4me3 led to the hypothesis that the aromatic cages are occupied by high-energy water molecules. Although difficult to confirm experimentally, WaterMap calculations performed on five representative reader proteins (both in tree and complexed forms) provided evidence that water molecules located inside the aromatic cages exhibit significant unfavourable free energy (Figure 6, supplementary table 1).[20] These high-energy water molecules are displaced by Kme3 side chain upon binding, which consequently makes a substantial contribution to Kme3 binding. Indeed, Water-Map calculations for 5F-, 6F- and 5,6-diF-JARID1A have shown similar results (Figure 7 and supplementary table 2). Upon the reduction of the electronegativity of the aromatic cage, the electrostatic component of the Kme3-aromatic cage interaction is reduced. This reduction in favourable enthalpy of binding is calculated to be offset, however, by more favourable desolvation of the aromatic cage. The amount of free energy of the aromatic cage’s hydration sites increases upon fluorination of the aromatic cage, resulting in more favourable free energies of binding upon the displacement of the water molecules by the Kme3 side chain. Together, both the K$^2$me3 and alloprotein approaches point towards a compensation mechanism which indicates that the reduction of the electrostatic component of the cation-π interaction of the reader-Kme3 binding event is compensated by a hydrophobic interaction component, namely more favourable desolvation.

This work employed, for the first time, fluorinated amino acids to address a fundamental question in the field of reader domain epigenetics. Furthermore, our data may aid the development of small molecules which may inhibit the JARID1A-PHD domain as the complete preservation of the cation-π interaction did not appear to be required for adequate binding of H3K4me3 to JARID1A. Indeed, a hydrophobic
moiety may bind the aromatic cage with adequate affinity providing more room for the development of small molecule drugs.

Concluding, the work presented in this chapter suggests that the aromatic cage-trimethyllysine association is driven by energetically favourable cation-π interactions and the favourable release of high-energy unstructured water molecules which are present in the aromatic cage. This result emphasizes the essential contribution of water in the Kme3 readout process. It is envisioned that our fundamental insight into the readout of Kme3 will aid the development and design of H3K4me3 reader antagonists, an area of research which has seen limited advances due to the innate difficulties with Kme3 reader druggability.[57-59]

4.4 - Methods

Synthesis of Fmoc-L-H3Cme3 and solid-phase peptide synthesis
Fmoc-L-H3Cme3 synthesis and solid phase peptide synthesis of H3 peptides was performed as described by Kamps et al. and Pieters et al.[20, 36]

Preparation and purification of reader proteins for K4me3 studies
Reader proteins were prepared and purified according to previously reported procedures.[36] Briefly, the reader domains of BPTF, JMJD2A, JARID1A, TAF3 and SGF29 were expressed in E. coli Rosetta BL21 (DE3) pLysS hosts, using Terrific Broth medium. The bacteria were cultured to OD600~0.6 at 37 °C after which they were induced with 0.4 mM IPTG overnight at 16 °C. Proteins were purified using Ni-NTA beads for His6-tagged proteins or glutathione sepharose beads for GST-tagged proteins, respectively. After purification, the His6-tag was cleaved from JMJD2A and SGF29 using TEV-protease and the GST-tag was cleaved from TAF3 using thrombin whereas BPTF and JARID1A remained tagged. Proteins were purified by size exclusion chromatography using a Superdex 75 column (GE Healthcare). SGF29 was eluted in 25 mM Tris-HCl, 50 mM NaCl, 1 mM DTT, pH 7.5; JMJD2A and TAF3 were eluted in 50 mM TRIS-HCl at pH 7.5; BPTF and JARID1A were eluted in 50 mM TRIS-HCl, 20 mM NaCl at pH 7.5. All proteins were made filter sterile and stored at 4 °C until further use.

JARID1A protein expression for auxotrophic studies
WT-JARID1A protein (Homo sapiens, residues 1600-1663) expressed in Rosetta BL21 (DE3) pLysS E. coli containing the JARID1A construct was cultured and expressed in TB medium supplemented with the appropriate antibiotics. At OD600~0.6, expression was induced with 0.4 mM IPTG and 0.1 mM ZnCl2 (final concentration) and cultured ON at 16°C. Cells were then harvested, lysed and purified.
using GST affinity. The GST tag was cleaved off with TEV-protease under reducing conditions (10 mM dithiothreitol) and the protein was subsequently purified by size exclusion chromatography on a Superdex 75 column using 20 mM TRIS-HCl pH 7.5, 50 mM NaCl, 1 mM DTT as running buffer. Protein concentration was measured spectrophotometrically using a Denovix DS-11 spectrophotometer and protein masses were confirmed using ESI-MS.

WT-JARID1A (*Homo sapiens*, residues 1600-1663) and alloproteins expressed in the auxotrophic *E. coli* (Migula) Castellani and Chalmers strain were cultured in either New Minimal Medium (NMM) or in Unnatural amino acid New Minimal Medium (UNMM) supplemented with appropriate antibiotics, respectively. NMM was prepared as described by Budisa et al. [37] In brief; NMM contained 100 mM K$_2$HPO$_4$, 55 mM KH$_2$PO$_4$, 20 mM D-glucose, 8.5 mM NaCl, 7.5 mM (NH$_4$)$_2$SO$_4$, 1 mM MgSO$_4$, 10 mg L$^{-1}$ biotin, 10 mg L$^{-1}$ thiamine-HCl, 1 mg L$^{-1}$ CaCl$_2$ and FeCl$_3$, 1 µg L$^{-1}$ CuSO$_4$, MnCl$_2$, ZnCl$_2$, NaMoO$_4$ and 50 mg L$^{-1}$ of each individual amino acid. UNMM was prepared similarly except that tryptophan was substituted with the desired fluorinated tryptophan analogue, at a final concentration of 25 mg/l.

For the expression of WT JARID1A, *E. coli* (Migula) Castellani and Chalmers containing the JARID1A construct was cultured in NMM at 37°C. At OD$_{600}$ ~0.6, the NMM medium was refreshed by harvesting the cells after which they were resuspended in fresh NMM. Expression was then induced with 0.1 mM IPTG and 0.1 mM ZnCl$_2$ (final concentrations). The cells were subsequently cultured for 3 hours at 37°C, after which the culture was harvested, lysed and purified as described above.

In order to incorporate fluorinated tryptophan analogues JARID1A was expressed as follows: *E. coli* (Migula) Castellani and Chalmers containing the JARID1A construct was initially cultured in NMM at 37°C. At OD$_{600}$ ~0.6 the cells were harvested and subsequently washed three times with 0.9% NaCl at room temperature. Following the washing steps, the cells were resuspended in fresh UNMM. Expression was then induced with 0.1 mM IPTG and 0.1 mM ZnCl$_2$ (final concentrations). The cells were subsequently cultured for 3 hours at 37°C after which the culture was harvested, lysed and purified as described above.

**Circular dichroism**

CD was measured at a protein concentration of 0.1 mg/ml in 10 mM phosphate buffer, pH 7.5, on a J-815 circular dichroism spectropolarimeter. The samples were measured over a range of 180-260 nm with normal sensitivity and a bandwidth of 1 nm. Scanning was performed at 50 nm per minute, a data integration time (D.I.T.) of 0.5 seconds and a data pitch of 0.5 nm. The spectra for each protein are a result of 10 accumulations.
**Differential Scanning Fluorimetry (DSF)**

Protein melt curves were obtained as described by Heinrich et. al. using a StepOne-Plus Real-Time PCR system (Applied Biosystems) and MicroAmp fast optical 96-well reaction plates (Applied Biosystems).[60] SYPRO-Orange protein gel stain (Invitrogen) was used as a reporter dye, emitting fluorescence in the FAM channel. Total reaction volume was 25 μl: 20 μl buffer (25 mM TRIS-HCl pH 7.5, 50 mM NaCl, 1 mM DTT), 2.5 μl of 25 nM protein and 2.5 μl SYPRO-Orange dye (diluted 1:100 in ddH₂O). Melt curve data were obtained in triplicate, in a temperature range of 20-95 °C at a step-wise temperature increment of 1 °C min⁻¹. Obtained data were analyzed using DSF Analysis v3.0.2 software, designed by Niesen et al. (available via ftp://ftp.sgc.ox.ac.uk/pub/biophysics/).[61]

**Isothermal titration calorimetry**

Generally, 350-600 μM of H3K4me3 or H3K²⁴me3 peptides were titrated to 25-40 μM of protein, except for JMJD2A-H3K²⁴me3 in which case 3 mM H3K²⁴me3 was titrated to 200 μM JMJD2A. In the case of the H3G4 titrations, 5 mM peptide was titrated to 330 μM JARID1A and 3 mM peptide was titrated to 200 μM TAF3. In the case of the alloproteins studies, 400 μM of peptide was titrated to 20 μM of protein and the buffers used for ITC experiments were the same as the elution buffer used for SEC; 20 mM TRIS-HCl pH 7.5, 50 mM NaCl, 1 mM DTT.

Each ITC titration consisted of 19 injections. ITC experiments were performed on the fully automated Microcal Auto-iTC200 (GE Healthcare Life Sciences, USA). Heats of dilution for histone peptides were determined in control experiments, and were subtracted from the titration binding data before curve fitting. Curve fitting was performed by Origin 6.0 (Microcal Inc., USA) using one set of sites binding model. For each reader-histone system, 5-7 independent ITC experiments were carried out.

**X-ray crystallography**

The tandem tudor domain of human SGF29 (residues 115-293) was cloned into a pET-28a-MHL vector, and is expressed, purified as described before.[23] The purified SGF29 was concentrated to 20 mg ml⁻¹ as a stock and frozen at -80 °C for future use. Purified SGF29 (15mg ml⁻¹) was mixed with histone peptide H3K²⁴me3 in a molecular ratio of 1:3, and the complex was crystallized in a buffer containing 0.1 M Bis-Iris, pH 5.5, 27% PEG3350, 200 mM ammonium sulphate and 5 mM strontium chloride. Before being flash-frozen in liquid nitrogen the crystals were soaked in a cryoprotectant buffer containing 88% reservoir solution and 12% glycerol.

Human JARID1A PHD finger (residues 330-380) was PCR-amplified, and cloned into a modified pET28b vector (Novagen) with an N-terminal His₆-SUMO tandem tag.
JARID1A PHD finger used for crystallization was expressed in E. coli BL21 (Novagen) induced overnight by 0.2 mM isopropyl β-D-thiogalactoside at 25 °C in the LB medium supplemented with 0.1 mM ZnCl₂. The harvested cells were suspended in 500 mM NaCl, 20 mM Tris-HCl, pH 8.5. After cell lysis and centrifugation, the supernatant was applied to a HisTrap (GE Healthcare) column and the protein was eluted with a linear imidazole gradient from 20 mM to 500 mM, followed by tag cleavage using ULP1. A HisTrap column was used to remove the cleaved His₆-SUMO tag after removal of imidazole by desalting. The JARID1A-PHD sample flow-through was then pooled, concentrated and purified by size-exclusion chromatography on a Superdex™75 16/60 column (GE Healthcare) using 150 mM NaCl, 20 mM Tris-HCl, pH 8.5 as elution buffer. The collected fractions containing the JARID1A PHD-finger was then concentrated to ~17 mg ml⁻¹, split into small aliquots and snap frozen in liquid nitrogen for future use.

The human TAF3 PHD-finger construct 885-915 was cloned, expressed and purified using the same strategy as used for the JARID1A PHD-finger. The TAF3 PHD-finger was concentrated to ~25 mg ml⁻¹, split into small aliquots and snap frozen in liquid nitrogen for future use.

Crystallization was performed via the sitting drop vapor diffusion method under 4 °C by mixing equal volumes (0.2-1.0 µl) of JARID1A PHD-H3K4me3 complex (1:1.8 molar ratio, 14-16 mg ml⁻¹) and reservoir solution containing 0.02 M sodium l-glutamate, 0.02 M DL-alanine, 0.02 M glycine, 0.02 M DL-lysine HCl, 0.02 M DL-serine, 0.1 M Tris-HCl, 0.1 M Bicine, pH 8.5, 12.5% MPD, 12.5% PEG 1K, 12.5% PEG 3350. The TAF3 PHD-H3K4me3 complex (1:1.4 molar ratio, 22-24 mg ml⁻¹) crystal was grown in reservoir solution containing 0.03 M magnesium chloride, 0.03 M calcium chloride, 0.1 M MES, 0.1 M imidazole, pH 6.5, 15% PEGMME 550, 15% PEG 20K. The complex crystals were directly flash frozen in liquid nitrogen with reservoir solution as cryoprotectant for data collection. The diffraction data were collected at the beamline BL17U of the Shanghai Synchrotron Radiation Facility at 0.9793 Å. All diffraction images were indexed, integrated, and merged using HKL2000/57. The structure was determined by molecular replacement using MOLREP58 with the free JARID1A PHD-finger (PDB ID: 2KGG) and free TAF3 PHD-finger (PDB ID: 2K16) as the search model. Structural refinement was carried out using PHENIX59, and iterative model building was performed with COOT60. Structural figures were created using the PYMOL (http://www.pymol.org/) program.

**19F-NMR**

Measurements were obtained on a Bruker AVANCE III 400 MHz system equipped with a BBFO probe capable of 19F nucleus detection with 1H decoupling. Samples
CHAPTER 4

were prepared using 5 mm Shigemi tubes matched to D_2O to minimize solvent volume required. \(^{19}\text{F}\)-NMR experiments were performed in 10 mM H_3KPO_4 pH 7.5, at a concentration of 450 \(\mu\text{M}\) protein and 1 mM H3K4me3 peptide for 6F-JARID1A and at 150 \(\mu\text{M}\) protein and 300 \(\mu\text{M}\) H3K4me3 peptide for 5F-JARID1A. All measurements were performed at 288 K. After samples were inserted into the magnet, the sample was shimmed using the lock nucleus in D_2O and a \(^1\text{H}\) spectrum was acquired to assess the quality of the shims. The probe was then manually tuned and matched to \(^{19}\text{F}\) to optimize \(^{19}\text{F}\) detection. A 15 \(\mu\text{s} @ 23\) Watts 90-degree pulse was used. \(^{19}\text{F}\{\text{^1}\text{H}\}\) spectra were then acquired with the following parameters: \(\text{NS} = 1.5k – 28k, d1 = 1, aq = 1.09\ s, sw = 20.1\ \text{ppm and o1p near -120 ppm.}\)

Quantum-chemical analysis

Quantum-chemical analyses were performed similarly as described by Kamps et al.[20]

WaterMap calculations

WaterMap calculations were performed similarly as described by Kamps et al.[20]

4.5 - Acknowledgements

We thank W. Tempel for the data collection of SGF29 and the staff members at beamlines BL17U of the Shanghai Synchrotron Radiation Facility for their assistance in data collection of JARID1A and TAF3 sum of the free energies of hydration sites displaced by the ligand upon binding. We would like to thank Paul White for his assistance with the \(^{19}\text{F}\) NMR studies, Maud Wuts for her extensive effort regarding investigations concerning the expression of alloproteins. Additionally, we would like to thank Professor Jan van Hest for kindly donating reagents.

4.6 - Author contributions

JM conceived and supervised the project. BP executed part of the experiments, performed data analysis, wrote, together with JM and MW, a manuscript entitled “The recognition of histone H3K4me3 by JARID1A is insensitive to fluorination of the aromatic cage”, which will be submitted for publication and which was integrated in this chapter, and wrote this chapter. MW also expressed and purified wild-type and fluorinated JARID1A. GP provided expert input, materials and reagents, and contributed to writing of the manuscripts. JK synthesized Fmoc-Cme3 and prepared histone peptides. JK and JM carried out thermodynamic studies and analysed the data for the “Chemical basis for the recognition of trimethyllysine by epigenetic reader
proteins” manuscript, which also was integrated in this chapter. JH and HL performed structural analyses of JARID1A and TAF3. JP and FB carried out quantum-chemical analyses. CX, AD and JMin performed structural experiments with SGF29. WS and TB carried out WaterMap calculations and interpreted the results. JM wrote the “Chemical basis for the recognition of trimethyllysine by epigenetic reader proteins” manuscript with contributions from WS, TB, FB, HL, AD, JH, JK, HL, JM, JMin, BP, JP and CX.
4.7 - References

42. Gatchalian, J. and T.G. Kufateladze, PHD Fingers as Histone Headers, in Histone Recognition, M.-M. Zhou, Editor, 2015, Springer: Cham, Switzerland.
## 4.8 - Appendix

### Supplementary table 1 | WaterMap calculations of thermodynamic parameters (in kcal mol⁻¹) for the solvation of aromatic cages of five reader proteins.

<table>
<thead>
<tr>
<th></th>
<th>Free protein</th>
<th></th>
<th>Complexed protein (Kme3)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PDB ID</td>
<td>-TΔS°</td>
<td>ΔH°</td>
<td>ΔG°</td>
</tr>
<tr>
<td>JARID1A</td>
<td>2KGG</td>
<td>1.4</td>
<td>0.8</td>
<td>2.2</td>
</tr>
<tr>
<td>SGF29</td>
<td>3MEW</td>
<td>2.9</td>
<td>6.8</td>
<td>9.7</td>
</tr>
<tr>
<td>TAF3</td>
<td>2K16</td>
<td>1.6</td>
<td>2.5</td>
<td>4.1</td>
</tr>
<tr>
<td>JMJD2A</td>
<td>2GF7</td>
<td>6.0</td>
<td>-0.8</td>
<td>5.2</td>
</tr>
<tr>
<td>BPTF</td>
<td>2F6N</td>
<td>3.5</td>
<td>3.6</td>
<td>7.1</td>
</tr>
</tbody>
</table>

Calculations are based on the crystal structures of free reader proteins or complexed reader proteins. Structures of reader proteins are obtained from PDB (see PDB IDs for free and complexed forms).

### Supplementary table 2 | WaterMap calculations of thermodynamic parameters (in kcal mol⁻¹) for the solvation of aromatic cage of WT and fluorinated JARID1A.

<table>
<thead>
<tr>
<th></th>
<th>ΔG°</th>
<th>ΔH°</th>
<th>-TΔS°</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>4.9</td>
<td>1.8</td>
<td>3.1</td>
</tr>
<tr>
<td>5F†</td>
<td>8</td>
<td>5.4</td>
<td>2.6</td>
</tr>
<tr>
<td>6F</td>
<td>7.8</td>
<td>5.3</td>
<td>2.5</td>
</tr>
<tr>
<td>5,6-di-F</td>
<td>6.6</td>
<td>4</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Calculations are based on the crystal structures of free reader proteins or complexed reader proteins. Structures of reader proteins are obtained from PDB ID: 2KGJ and 2KGJ.
CHAPTER 4

Supplementary figure 1 | JARID1A CD spectra. Wild type JARID1A reader domain expressed in *E. coli* Rosetta BL21 (DE3) pLysS is displayed in black and the auxotrophic *E. coli* Castellani and Chalmers in grey respectively.

Supplementary figure 2 | JARID1A WT ESI-MS. JARID1A reader domain expressed in **A** *E. coli.* Rosetta BL21 (DE3) pLysS and **B** the auxotrophic *E. coli* Castellani and Chalmers strain respectively.
Supplementary figure 3] Comparison of free (unbound) aromatic cages and Kme3-bound (complexed) aromatic cages illustrates that aromatic cages are predominantly preformed for binding of the Kme3 ligand. a) BPTF PHD finger, b) JARID1A PHD finger, c) TAF3 PHD finger, d) JMJD2A Tudor, e) SGF29 Tudor. Coordinates are taken from Protein Data Bank with entry numbers 2F6N (free) and 2F6J (complex) for BPTF, 2KGG (free) and 3GL6 (complex) for JARID1A, 2K16 (free) and 2K17 (complex) for TAF3, 2GF7 (free) and 2GFA (complex) for JMJD2A, 3LX7 (free) and 3MEA (complex) for SGF29, respectively.

Supplementary figure 4] JARID1A 19F-NMR spectra. 19F-NMR spectra for 5F- and 6F-JARID1A with (+) and without (-) ligand (10-mer H3K4me3: ARTKme3QTARKS). An asterisk indicates the 19F-NMR signal of free fluoride.
Supplementary figure 5 | JARID1A $^{19}$F-NMR spectra overlay. $^{19}$F-NMR spectra for 5F- and 6F-JARID1A overlay with (+) and without (-) ligand (10-mer H3K4me3: ARTKme3QTARKS). Blue lines correspond to 5F-Trp and red lines to 6F-trp JARID1A. An asterisk indicates the $^{19}$F-NMR signal of free fluoride.

Supplementary figure 6 \| Continued.

Supplementary figure 7 | WaterMap calculations for the solvation enthalpies of aromatic cages of reader proteins. a) JARID1A, b) SGF29, c) TAF3, d) JMJD2A, e) BPTF. Superimposed Kme3 side chain and water molecules are presented as green stick and grey spheres. Numbers (red) adjacent to grey spheres represent the value of the enthalpy ($\Delta H^\circ$) for individual water molecule. f) Thermodynamic parameters for the solvation of the aromatic cages of five reader proteins.
Supplementary figure 8 | WaterMap calculations for the solvation entropies of aromatic cages of reader proteins. a) JARID1A, b) SGF29, c) TAF3, d) JMJD2A, e) BPTF. Superimposed Kme3 side chain and water molecules are presented as green stick and grey spheres. Numbers (blue) adjacent to grey spheres represent the value of the entropy (-TΔS°) for individual water molecule. f) Thermodynamic parameters for the solvation of the aromatic cages of five reader proteins.
Supplementary figure 9 | WaterMap calculations for the solvation enthalpies of the aromatic cages of JARID1A. a) Wild type, b) 5F-JARID1A, c) 6F-JARID1A, d) 5,6-di-F-JARID1A, e) Thermodynamic parameters for the solvation of the aromatic cages of JARID1A.

Supplementary figure 10 | WaterMap calculations for the solvation entropies of the aromatic cages of JARID1A. a) Wild type, b) 5F-JARID1A, c) 6F-JARID1A, d) 5,6-di-F-JARID1A, e) Thermodynamic parameters for the solvation of the aromatic cages of JARID1A.
Novel methyllysine analogs for investigating the chemical-biological role of lysine methylation

B.J.G.E. Pieters
Y. Grobben
A.H.K. Al Temimi
J.A.G. Kamps
J. Mecinović

[Manuscript in preparation]
CHAPTER 5

Abstract

Objective: Tools are required in order to investigate the effects of the many epigenetic histone posttranslational modifications (PTM) from a chemical/biological perspective. In the case of the methyllysine PTM, tools have previously been developed which allow for the directed deposition of mono-, di-, and tri-methyllysine analogs (MLA) on cysteines using simple alkylation reagents. Here we expand upon the relatively limited repertoire of available MLAs, providing the scientific community with tools to study the effects of lysine methylation upon various biological processes.

Approach: Novel alkylation reagents, designed for the investigation of electrostatic and steric effects in biomolecular recognition in epigenetics, were prepared and used to alkylate recombinant full length Xenopus laevis histone H3. Furthermore, using isothermal titration calorimetry (ITC) the binding features of the JARID1A reader domain to H3K4 variants, generated by alkylation of a K4C variant, were determined in order to verify the applicability of the designed MLAs.

Results: Alkylation reagents were successfully prepared and used to install a variety of MLAs on full length histone H3. Western blotting indicated that the K4me3 epitope is still recognized in the case of the installed H3K_c4 MLAs with the exception of the uncharged H3K_c4me3 analog. Further investigations using ITC have shown the effects of changes in the MLAs methylammoniums charge, charge density, bulk and chain length on the JARID1A-H3K4me3 binding event. Here we observed a trend which suggested adequate binding for the MLAs. Furthermore, we clearly observed an enthalpy-entropy compensation mechanism when performing binding experiments with the uncharged H3K_c4me3 analog, as has also been discussed in Chapter 4. This indicates that our MLAs have potential as a novel tool for the investigation of lysine methylation effects.

Conclusions: Protocols and reagents have been successfully developed for the installation of a variety of MLAs at a defined position in full-length histone H3, which can be used for the assembly of higher order histone octamer complexes. Using ITC, it has been shown that the developed analogs have strong potential as novel tools for the investigation of the methyllysine PTM in epigenetic processes.
5.1- Introduction

Biochemical sciences have provided the scientific community with a wealth of tools in order to study the important molecular processes of life.[1, 2] When studying proteins, for example, it is possible to employ classical mutagenesis in order to exchange a single natural amino acid for one of the remaining 19 canonical amino acids. This type of studies has proven to be invaluable when attempting to elucidate protein structure and function, molecular pathways, the involvement of specific single nucleotide polymorphisms in disease and many cellular processes. Unfortunately, these classical biochemical tools have the limitation that they rely on the repertoire of biomolecules that occur naturally within a living cell, such as the 20 naturally occurring amino acids. This complicates detailed mechanistic studies which require the manipulation of specific physico-chemical properties of single amino acids.

Fortunately, several techniques have been developed in order to circumvent this innate limitation imposed on the chemical and biochemical communities by the occurrence of only 20 canonical amino acids.[3] The incorporation of non-canonical amino acids (NCAA) into proteins is possible, for example, by making use of an auxotrophic bacteria strain. When such a strain is cultured in the absence of the amino acid it is unable to synthesize, it can be forced to use slightly modified variants of this amino acid as a substrate. This strategy results in the exchange of all naturally occurring amino acids by its unnatural variant during protein synthesis. A more advanced method for the incorporation of NCAAs into proteins is based on the amber codon suppression strategy. [4, 5] Here the amber stop codon (TAG) is made to encode for an unnatural amino acid, essentially leading to an expansion of the genetic code. The advantage of amber suppression is that an NCAA can be introduced at a very specific site without coincidentally exchanging other amino acids. Recent advances even allow the specific, simultaneous incorporation of multiple, distinct NCAAs.

Such methods for the incorporation of NCAAs, however, rely on the incorporation of unnatural amino acids during protein synthesis. Some major drawbacks of these systems include the fact that auxotrophic strains are severely limited in the amino acid variants they accept as substrate and that the development of tRNA/tRNA synthetase pairs for amber suppression is laborious, time consuming and has to be repeated for each amino acid variant that is to be introduced.[6, 7] Therefore, several chemical protein modification methods have been developed that allow for the incorporation of NCAAs after the translation process.[8-10] One such method is the alkylation of cysteines in order to introduce modifications that mimic natural amino acids.[11] This process is especially useful for the manipulation of histones, as H3 is the only canonical histone variant that contains a cysteine residue (Cys110), which
can readily be mutated to an alanine residue.[12] The freedom to position cysteines at any desired position along the intact H3 protein (and also histones H2A, H2B and H4) creates the possibility to deposit modifications at any location. It has been shown that the alkylation of cysteines into lysine or methylated lysine analogs (MLA) does not adversely change the residues’ properties when compared to natural lysine or methyllysine.[13] For example, histone lysine methyltransferases and readers of specific methylation states recognize the MLAs with similar affinity as their natural counterparts.

Unfortunately, only a few reagents which can be used to prepare MLAs have been developed: lysine, monomethyllysine, dimethyllysine and trimethyllysine analogs are currently the only MLAs that can be introduced using cysteine alkylation. These MLAs mimic the natural lysine state which allow for the investigation of the function of the natural posttranslational modifications (PTM) present on lysine residues. When one wants to probe more specific effects of these methyllysine PTMs, however, more analogs are required. This demand is what has been addressed in the current work.

Here, we present a variety of trimethyllysine analogs, which can be used to study the influence of charge, charge density, chain length and size of the trimethyllysine moiety. We have been able to successfully incorporate these MLAs into intact H3 protein using cysteine alkylation strategies and subsequently used these alkylated histones to assemble histone octamers. Furthermore, ITC studies with alkylated histone proteins and the H3K4me3 reader domain protein JARID1A have been performed.

5.2 – Results

Histone expression and alkylation

In order to perform alkylation reactions on full length histone 3, a H3 construct was obtained containing a C110A and K4C mutation. This protein, designated H3C4, was then expressed in *E. coli* and subsequently purified from the pellet fraction.[14] After expression and purification of recombinant H3C4, the protein was alkylated using a variety of MLA reagents (Figure 1A-B).

Each of the prepared MLAs is a specific variant of the natural Kme3 (Figure 1 C). The lysine (K4) and trimethyllysine (K4me3) variants were developed previously and have already been shown to mimic natural lysine and trimethyllysine well.[12] Our novel MLAs expanded this work and contain modifications with regard to the lysine side chain. Two of our MLAs have been designed in order to probe the effect of the
Figure 1 | Histone 3 structure, cysteine alkylation reaction and methyllysine analogs. A) Nucleosome and histone 3 (H3) structures. Green: H2A, cyan: H2B, yellow: H3 and red: H4. Highlight: Lysine 4 on histone 3 (H3K4). B) Conversion of cysteine 4 on histone 3 (H3C4) into a methyllysine analog (MLA) by alkylation using bromide alkylation reagents. C) H3C4 and the MLAs described in this chapter and their corresponding abbreviations.
methylammonium moieties’ charge density, namely a phosphonium (K\textsubscript{c}^\textsuperscript{+}me3) and arsonium (K\textsubscript{c}^\textsuperscript{+}me3) analog. Furthermore, an uncharged carba (K\textsubscript{c}^\textsuperscript{0}me3) analog was designed in order to investigate the effect of the methylammonium’s positive charge. Three additional MLAs, propane-trimethyllysine (K\textsubscript{c}^\textsuperscript{0}propane\textsubscript{3}me3), butane-trimethyllysine (K\textsubscript{c}^\textsuperscript{0}butane\textsubscript{4}me3) and pentane-trimethyllysine (K\textsubscript{c}^\textsuperscript{0}pentane\textsubscript{5}me3), were devised in order to explore the effect of variation of the side chain’s length and lastly, a pyridine analog (K\textsubscript{c}^\textsuperscript{0}pyridine\textsubscript{4}me3) was prepared in order to investigate the effect of steric. After performing the alkylation reactions on intact H3C4 protein, denaturing ESI-MS experiments were performed on the alkylated histone products in order to confirm successful alkylation (Figure 2, Supplementary table 1 and Supplementary Figures 1 & 2).

The unalkylated H3C4 protein mass corresponds to the expected calculated mass minus 2 Da. This shift in mass is likely caused by the deprotonation of the H3 histidines during ESI-MS measurements. A separate peak which corresponds to a mass increase of 43 Da was also visible. Peaks corresponding to this mass have been reported previously and likely correspond to the addition of a carbamyl group (O\textsuperscript{\text{-}}C\textsuperscript{\text{-}}NH\textsubscript{3})\textsubscript{+}.[12] Such mass increases are commonly observed when proteins are purified using urea as this compound slowly decomposes into isocyaninic acid, via an ammonium cyanate intermediate, which is capable of reacting with free amines. The H3K\textsubscript{c}^\textsuperscript{0}4 protein showed only minimal amounts of unreacted (non-alkylated) species, similarly: H3K\textsubscript{c}^\textsuperscript{0}4me3, H3K\textsubscript{c}^\textsuperscript{0}4\textsuperscript{Ac}4me3, H3K\textsubscript{c}^\textsuperscript{4}4me3, H3K\textsubscript{c}^\textsuperscript{0}propane4me3 and H3K\textsubscript{c}^\textsuperscript{0}pyridine4me3 showed good conversion with mass increases corresponding to the expected molecular masses (Figure 2, Supplementary figures 1 and 2, Supplementary table 1). H3K\textsubscript{c}^\textsuperscript{0}4me3, H3K\textsubscript{c}^\textsuperscript{0}butane4me3 and H3K\textsubscript{c}^\textsuperscript{0}pentane4me3 also showed efficient conversion, although over-alkylation was also observed for these three products. The reagents used for preparation of these products proved to be more reactive, which led to a fraction of the alkyalted histones containing multiple adducts despite our optimization efforts. The incorporation of bulkier triethyllysine proved unsuccessful despite our optimization efforts.

Next, we assessed the alkylated histones by SDS-PAGE and western blotting (Figure 3). The results showed that the alkylated histones have been obtained in high purity (Figure 3A) and still contained the epitopes required for recognition by antibodies directed against unmethylated histone 3 (Figure 3B). Interesting results were obtained when the purified alkylation products were used for western blot analysis using antibodies specifically directed against H3K4me3 (Figure 3C). Here, we see that the antibodies are indeed specific for the trimethylated status of K4 on intact H3, despite the presence of a sulfur atom in the side chain as H3K\textsubscript{c}^\textsuperscript{4}4 was not recognized whereas H3K\textsubscript{c}^\textsuperscript{4}4me3 was recognized. This result is consistent with previous studies on the H3K9 lysine analogs.[12]
Figure 2 | Alkylated histone ESI-MS spectra. A) H3C4. B) H3K4. C) H3K4me3. D) H3Kc4me3 and several minor peaks indicating the deposition of multiple trimethylphosphonium groups. E) H3KcAs4me3. F) H3Kc4me3. G) H3KcPropan4me3. H) H3KcButane4me3 and a minor peak indicating the deposition of 2 butane-trimethyllysine groups. I) H3KcPentane4me3 and several peaks indicating the deposition of multiple pentane-trimethyllysine groups. J) H3KcPyridine4. Peaks indicated with a double dagger correspond to a mass of +16 Da (ammonium adduct) and peaks indicated with an asterisk correspond to a mass of +43 Da (carbamyl adduct).
Interestingly, a replacement of the nitrogen at the lysine side-chain by phosphorus (H$^3$K$_c^{4\text{me3}}$), or arsenic (H$^3$K$_c^{4\text{Asme3}}$), or elongation of the side-chain by an additional methylene group (H$^3$K$_c^{4\text{Propane4me3}}$), did not (or hardly) affect the recognition by the anti-H3K4me3 antibodies. This observation indicates that small changes in charge density or chain length do not affect the ability of antibodies to recognize the specific Kme3 epitope. Further increases in the lysine chain length resulted in weaker recognition of the epitope, although full abrogation of binding has not been observed. Even the installation of a bulky, positively charged pyridine group did not decrease recognition by the antibodies beyond the limits of detection. Elimination of the trimethylammonium’s positive charge by the homologous neutral carba analogue, on the other hand, completely interfered with recognition of the epitope. Together, these results indicate that the positive charge situated on the Kme3 moiety is important for recognition of the epitope(s) by the polyclonal antibody.

Figure 3| SDS-PAGE and western blot analysis of alkylated histones. A) SDS-PAGE of purified alkylation products. B) Western blot of the purified alkylation products using antibodies directed against H3. C) Western blot of purified alkylation products using antibodies specifically directed against H3K4me3.
used. In summary: with the exception of H3K\textsubscript{c} and H3K\textsubscript{c}4me3, all incorporated MLAs resulted in H3 variants that were recognized by the anti-H3K4me3 antibody, albeit with different affinities.

**Thermodynamic analysis**

After characterizing the alkylated histone proteins, these variants were used to study the thermodynamic features of the JARID1A-H3 interaction by ITC (Supplementary Figure 3). It was found that all MLAs were able to function as binding partners for JARID1A with dissociation constants within a range of 0.13-0.6 \(\mu\)M and Gibbs free energies in the range of -8.5 to -9.5 kcal mol\(^{-1}\) (Figure 4 and Table 1). Unexpectedly the H3K\textsubscript{c} 4 variant also bound to the JAHID1A reader domain with reasonable binding affinities, albeit with less favorable enthalpy and more favorable entropy than H3K\textsubscript{c}4me3.

From the obtained data it is clear that the association between JARID1A and H3K\textsubscript{c}4me3 is enthalpy driven while entropy is slightly unfavorable, as had been observed previously with peptide based investigations.[15, 16] The H3K\textsubscript{c}4me3 and

![Figure 4](image-url) **Figure 4| Alkylated histone ITC data.** JARID1A-PHD domain was titrated with alkylated histone protein at a concentration of 30 \(\mu\)M reader and 300 \(\mu\)M histone. Gibbs free energy is displayed in blue, enthalpy in red and entropy in dark green.
Table 1 | Alkylated histone ITC data ± SD. ITC data obtained from experiments conducted with alkylated histone MLAs and the JARID1A H3K4me3 reader domain.

<table>
<thead>
<tr>
<th></th>
<th>Kd (uM)</th>
<th>ΔG° (kcal)</th>
<th>ΔH° (kcal)</th>
<th>-TΔS° (kcal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4</td>
<td>0.6</td>
<td>-8.5</td>
<td>-8.2</td>
<td>-0.4</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>0.13 ± 0.1</td>
<td>-9.5 ± 0.5</td>
<td>-10.9 ± 0.2</td>
<td>1.3 ± 0.3</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>0.23 ± 0.1</td>
<td>-9.1 ± 0.2</td>
<td>-9.4 ± 0.1</td>
<td>0.3 ± 0.1</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>0.29 ± 0.1</td>
<td>-8.9 ± 0.1</td>
<td>-10.1 ± 0.6</td>
<td>1.2 ± 0.5</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>0.21 ± 0.1</td>
<td>-9.1 ± 0.1</td>
<td>-7.9 ± 0.5</td>
<td>-1.2 ± 0.6</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>0.29 ± 0.1</td>
<td>-8.9 ± 0.3</td>
<td>-11.1 ± 2.6</td>
<td>2.2 ± 2.9</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>0.30 ± 0.1</td>
<td>-8.9 ± 0.1</td>
<td>-10.5 ± 1.2</td>
<td>1.6 ± 1.3</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>0.27 ± 0.1</td>
<td>-9.0 ± 0.1</td>
<td>-11.0 ± 1.0</td>
<td>2.0 ± 1.1</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>0.60 ± 0.1</td>
<td>-8.5 ± 0.1</td>
<td>-11.1 ± 0.7</td>
<td>2.6 ± 0.7</td>
</tr>
</tbody>
</table>

Data was obtained in duplicate. *Single measurement

H3K4me3 MLAs showed a trend similar to H3K4me3 binding with an enthalpy driven association. The binding affinities and free energies for H3K4me3, H3K4me3 and H3K4me3 are similar, but differences in ΔH° and -TΔS° were observed. Both H3K4me3 and H3K4me3 binding events display less favorable, less favourable enthalpies (ΔΔH°) of +1.5 and +0.8 kcal mol⁻¹ and more favourable entropies - ΔΔS° of -1.0 and -0.2 kcal mol⁻¹, respectively, when compared to H3K4me3. This trend is in line with the expected reduction in cation-π interactions as the cationic charge component becomes less dense, which results in a weaker electrostatic component and hence lower energies of binding. The entropic compensation may be explained by a mechanism of more favorable desolvation of the methylammonium moiety, as had been proposed by us before.[17] Further work will need to be performed in order to establish a more conclusive picture of the H3K4me3 and H3K4me3 binding events. When examining the H3K4me3 binding event, we observed a result identical to that obtained in our previous work, with peptides containing a neutral carba variant of trimethyllysine: a clear enthalpy-entropy compensation mechanism with a ΔΔH° of +2.9 kcal mol⁻¹ and -TΔS° of -2.6 kcal mol⁻¹ relative to H3K4me3.[16]

H3K4me3, H3K4me3 and H3K4me3 binding events show a trend which suggests that binding to JARID1A is enthalpically more favorable while the entropic penalty becomes larger. It is tempting to speculate that due to the increased
NOVEL METHYLLYSINE ANALOGS

Figure 5: Representative histone octamer size exclusion chromatograms. A) H3K_cme3. B) H3K_cPme3. C) H3K_cCme3. D) H3K_cbutyrylme3. E) H3K4_cpyridine. The first peak in each chromatogram contains higher order aggregates, whereas the major peak at 13 ml (*) corresponds to the assembled histone octamer and the subsequent peaks contain H3_c-H4 Eaton tetramers and H2A-H2B dimers, respectively. Absorbance is given in arbitrary units.

Over the length of the methylammonium group is able to come in closer contact with the aromatic cage, resulting in stronger enthalpy of binding. Due to steric hindrance and/or due to additional strain on the MLA side chains, however, a higher entropic penalty must be paid in order to bind the methylammonium moiety into the aromatic cage. Finally, H3K4_cPyridine4 resulted in a marked but overall modest increase in \( \Delta G^0 \) of 1.0 and 0.6 kcal mol\(^{-1}\) when compared to H3K_c4me3 and H3K_cP4me3 respectively. Overall, \( \Delta G^0 \) was slightly lower when compared to the other MLAs. Also, a relatively large unfavorable \( -T \Delta S^0 \) of +2.6 kcal mol\(^{-1}\) was observed in the case of H3K4_cPyridine4-JARI-D1A binding.

Histone octamer assembly
One of the advantages of installing MLAs on intact histone proteins is that these chemically modified histones can be used to study higher order effects mediated by a specific PTM. The methylation status of H3K79, for example, alters the nucleosome structure and H4K20 methylation affects higher order chromatin structures. Using MLAs, the mechanistic details behind the effects of these simple, but important
modifications could potentially be unveiled. After the successful installation of a variety of MLAs on H3C4, the purified alkylated histones were used for the assembly of histone octamers according to reported procedures.\[19\]

Size exclusion chromatograms and SDS-PAGE analysis (Figures 5 and 6A) show that the alkylated H3C4 histones can be assembled into complete histone octamers composed of recombinant H2A, H2B, H3 and H4 proteins (Supplementary figure 4), regardless of the MLA installed. Further analysis by western blotting and dot blotting (Figure 6B-C and supplementary figure 5) shows that antibodies can readily distinguish between the individual histones without displaying any cross-reactivity under both denatured and native assembly conditions. This indicates that the MLA containing histone proteins can be readily incorporated into higher order structures and may be used for the investigation of such complexes.

**Figure 6| Histone octamer SDS-PAGE and western blotting.** A) SDS-PAGE of assembled histone octamers containing various H3 MLA’s. B) Western blot analysis using an antibody directed against histone 3. C) Western blot analysis using an antibody directed against H3K4me3.
5.3 - Discussion

Advances in the field of epigenetics push the scientific community towards developing methods to study epigenetic modifications at the molecular level. As the limits imposed by naturally occurring amino acids and PTMs are exhaustive and biomolecules containing specific modifications may be difficult to obtain in reasonable quantity and purity, techniques have been developed to incorporate NCAAs into proteins and peptides.[3, 8, 9] Such techniques have been used successfully in investigations ranging from elucidating the functionality of neuron receptors to bestowing new biophysical properties upon existing proteins.[20, 21] With regard to an epigenetic histone context, Shokat and co-workers have developed a relatively simple method to site-specifically install MLAs onto histone proteins using chemoselective cysteine alkylation.[12]

In our current work we expanded the toolbox for the incorporation of MLAs by cysteine alkylation. We developed simple MLAs which differ from trimethyllysine in charge, charge density, chain length and steric properties. Such trimethyllysine analogs can be used to investigate the function of specific proteins of the methyllysine moiety without causing major perturbations.[16, 22] A major advantage of the developed alkylation reagents is the ease by which they may be synthesized and used. In comparison: it is otherwise difficult to incorporate NCAAs or PTMs into larger polypeptides as solid phase peptide synthesis is no longer an optimal option for the preparation of biomolecules of such size. Other methods such as native chemical ligation may be used, but are laborious in comparison to the use of cysteine alkylation. [23] By using simple alkylation reagents, it is possible to incorporate trimethyllysine analogs onto cysteines leading to the desired MLA containing polypeptide. Such innovations open doors for researchers who do not have the required expertise or means for complex peptide syntheses or methods such as amber codon suppression.

As the results in this chapter indicate, our alkylation reagents can be used to investigate the effects of the methyllysine moiety on reader-histone association in manners which have already been shown to be invaluable in the elucidation of such interactions.[16, 22] For example, in our previous work (Chapter 4), we have shown the influence of water on H3K4me3 binding by reader domain proteins using a neutrally charged carba analog. This has shed light upon an enthalpy-entropy compensation mechanism due to more favorable desolvation of the neutral MLA while the cation-π component in the H3K4c4me3-reader association event was reduced. This compensation mechanism has now also clearly been observed with our H3Kc4me3 MLA. Furthermore, the H3Kc4me3 and H3Kas4me3 MLAs also appear to display a similar compensation mechanism when the charge density is
progressively reduced in the trimethylphosphonium and trimethylarsonium moieties. Additional investigations will be needed to provide more conclusive evidence for the binding mode of JMARID1A to H3K$^{+4}$me3 and H3K$^{+}$as$^{4}$me3. The effects of chain length on H3K4me3 binding was less clear, and future investigations will be needed to explore the importance of this parameter in detail. It is important to note, however, that our MLAs bind to JARID1A with affinities and thermodynamic parameters similar to previous observations using 10-mer histone peptides.

5.4 - Conclusions

In conclusion, we have developed reagents that can be employed for the site specific incorporation of various MLAs into intact histones, using a relatively simple cysteine alkylation procedure. These reagents are a novel addition to the already existing methods for studying histone modifications.[13] We have shown that these MLAs can be used to probe the effect of charge and bulk properties of the methylammonium moiety of trimethyllysine in a sequence specific context and that they have the potential to be employed to study higher order nucleosome and chromatin assemblies. Furthermore, our preliminary results are in line with results that are described in Chapter 4, which substantiates their reliability.

5.5 – Experimental section

Preparation of full-length histone proteins

Full length histone proteins (X. laevis) were prepared as described previously. [12, 14] The constructs encoding for the full length histone proteins were generously donated by Professor Christopher Schofield, Oxford University. The H3 construct containing a C110A mutation was adjusted with a K4C mutation using PCR, which was subsequently verified using sequence analysis. The obtained histone constructs were transformed into competent E. coli. Rosetta BL21 DE3 PlyS cells and used for expression in TB medium. H2A and H2B were expressed by inducing the cultured cells at OD$_{600}$ = 0.6 with 0.1 mM IPTG whereas H3 was induced with 0.4 mM and H4 with 1 mM IPTG. The cells were then cultured for 3 hours at 37°C after which the cells were harvested and stored at -80°C until further use.

The expressed histone proteins were then purified from the pellet fraction by soaking the pellet in DMSO for 30 minutes followed by a 1-hour extraction with 6 M guanidinium HCl, 20 mM NaAc pH 5.2 and 1 mM DTT. The extracted protein was further purified by gel filtration using a HiPrep Sephacryl S-200 HR column. H2A, H2B and H4 protein
fractions were eluted in 7 M deionized urea, 20 mM NaAc pH 5.2, 200 mM NaCl and 2 mM 2-mercaptoethanol at 3 ml min⁻¹ at room temperature. H3 was purified using 6 M guanidinium HCl, 20 mM NaAc pH 5.2, 1 mM DTT as eluent, in order to more effectively separate the H3 protein from DNA contaminants. Fractions containing the histone proteins were collected, dialyzed against 1 mM 2-mercaptoethanol and subsequently lyophilized for prolonged storage at -20°C.

Expression and purification of reader domain proteins
JARID1A reader domain protein was expressed and purified as described by Pieters et al. [15]

Synthesis of alkylation reagents
The synthesis and purification of the butyl-, pentyl- and pyridyl-based alkylation agents were carried out based on previously reported protocols and purifications were performed by recrystallisation to their pure anhydrous forms.[24-26] The syntheses of the phosphonium- and arsonium-based alkylation reagents were carried out as described in literature.[27] Purifications were performed by preparative-HPLC.

SDS-PAGE and western blotting
SDS-PAGE analysis was performed on the assembled histone octamers using a 15% PAA gel. PAGE was run for 30 minutes at 100 V and subsequently run for 1 hour at 150 V. Gels were stained with coomassie brilliant blue for visualization of protein. Western blotting was performed according to standard procedures. Briefly an SDS-PAGE gel was prepared as described above, after which it was blotted onto a nitrocellulose membrane for 1 hour at 250 mA. The histone proteins were visualized using antibodies directed against H3 and H3K4me3 respectively (Abcam, Ab1791 and Novus Biologicals, NB21-1023).

Alkylation of full length histone H3
The K₅, K₅me3, K₇me3, K₇me3, K₇pyridine, K₇propane-m3, K₇butane-m3 and K₇pentane-m3 groups were deposited according to our own optimized conditions as described in table 2 based on the protocols described by Simon et al.[12, 28] Briefly; 10 mg of purified and lyophilized histone 3 was dissolved in 980 μl alkylation buffer (4 M GuHCl, 1 M Hepes pH 7.8 and 10 mM D/L-methionine) and allowed to incubate for 1h at 37° C under reducing conditions by adding 20 μl 1 M DTT. The desired alkylation reagent was directly dissolved into the reaction mixture and allowed to react at the specified temperatures described in table 2. After 2.5 h reaction time, 10μl 1 M DTT was added to the reaction mixture and the reaction was allowed to proceed for another 2.5 h. The reaction was quenched by incubating the reaction mixture with 50 μl 2-mercaptoeth-
### Table 2 | Alkylation reaction conditions for the preparation of MLAs using H3C4 protein.

<table>
<thead>
<tr>
<th>Modification</th>
<th>Abbreviation</th>
<th>Reagent</th>
<th>Solvent</th>
<th>Molar ratio reagent: H3</th>
<th>T (°C)</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lysine*</td>
<td>K_c</td>
<td>2-bromo-N,N,N-trimethylethanaminium bromide</td>
<td>Alkylation buffer</td>
<td>75:1</td>
<td>21</td>
<td>2x 2.5 hrs.</td>
</tr>
<tr>
<td>Trimethyllysine*</td>
<td>K_cme3</td>
<td>(2-bromoethyl)-trimethyl ammoniumbromide</td>
<td>Alkylation buffer</td>
<td>300:1</td>
<td>50</td>
<td>2x 2.5 hrs.</td>
</tr>
<tr>
<td>Phosphonium-trimethyllysine</td>
<td>K_cPme3</td>
<td>(2-bromoethyl)-trimethylphosphonium bromide</td>
<td>Alkylation buffer</td>
<td>50:1</td>
<td>50</td>
<td>2x 2.5 hrs.</td>
</tr>
<tr>
<td>Arsonium-trimethyllysine</td>
<td>K_cAsme3</td>
<td>(2-bromoethyl)-trimethylarsonium bromide</td>
<td>Alkylation buffer</td>
<td>200:1</td>
<td>50</td>
<td>2x 2.5 hrs.</td>
</tr>
<tr>
<td>Carbon-trimethyllysine</td>
<td>K_cCme3</td>
<td>1-Bromo-3,3-dimethylbutane</td>
<td>DMSO</td>
<td>300:1</td>
<td>50</td>
<td>2x 2.5 hrs.</td>
</tr>
<tr>
<td>Propane-lysine</td>
<td>K_cpropane3me3</td>
<td>3-bromo-N,N,N-trimethylpropane-1-aminium bromide</td>
<td>Alkylation buffer</td>
<td>300:1</td>
<td>37</td>
<td>2x 2.5 hrs.</td>
</tr>
<tr>
<td>Butane-lysine</td>
<td>Kcbdme3</td>
<td>4-bromo-N,N,N-trimethylbutane-1-aminium bromide</td>
<td>Alkylation buffer</td>
<td>166:1</td>
<td>37</td>
<td>2x 2.5 hrs.</td>
</tr>
<tr>
<td>Pentane-lysine</td>
<td>K_cpentane3me3</td>
<td>5-bromo-N,N,N-trimethylpentane-1-aminium bromide</td>
<td>Alkylation buffer</td>
<td>200:1</td>
<td>37</td>
<td>2x 2.5 hrs.</td>
</tr>
<tr>
<td>Pyridine-lysine</td>
<td>K_cpyride</td>
<td>1-(2-bromoethyl)-pyridinium bromide</td>
<td>Alkylation buffer</td>
<td>340:1</td>
<td>50</td>
<td>2x 2.5 hrs.</td>
</tr>
</tbody>
</table>

*Reaction performed as described by Simons et al.
anol for 30 minutes at RT after which the alkylated histones were desalted using a PD-10 column (#17-0851-01, GE Healthcare) and subsequently buffer exchanged into 1 mM 2-mercaptoethanol using a centrifugal filter unit. The alkylated histone proteins were then lyophilized and stored at -20°C.

For installing K\textsuperscript{\textcircled{C}}\textsuperscript{me3} a separate protocol was devised: 10 mg of purified and lyophilized histone 3 was dissolved in 980 µl DMSO and reduced for 1 h at 37°C by adding 20 µl 1M DTT in DMSO. 1 ml reactions were performed in 1.5 ml conical tubes by pipetting the reagent directly into the reaction mixture and incubating at 50°C. After 2.5 hrs. 10 µl 1M DTT in DMSO was added and the reaction was allowed to proceed for another 2.5 hrs. After a total of 5 h reaction time, the reaction was quenched by incubating the reaction mixture with 50 µl 2-mercaptoethanol for 30 minutes at RT. After quenching, the mixture was diluted with ddH\textsubscript{2}O and buffer exchanged into 1 mM 2-mercaptoethanol using a centrifugal filter unit. The hydrophobic reagent was removed from solution by ether extraction and the protein was subsequently lyophilized and stored at -20°C.

**ITC measurements**

ITC measurements were conducted with recombinant JARID1A H3K4me3 reader domain and full length histones at a temperature of 298 K. Buffers identical to the buffers used for reader protein purification were used for ITC experiments. Typically, 30 µM reader protein was titrated with 300 µM histone protein. Each ITC titration consisted of 19 injections and experiments were repeated 3 to 5 times. The ITC experiments were performed on a fully automated Microcal Auto-iTC200 (Malvern) and curve fitting was performed using Origin 6.0 (Microcal Inc., USA) using a one-site model.

**Histone octamer assembly**

Histone octamers were assembled as described by Luger et al.[19] Upon expression and purification of full length histones (Supplementary figure 2) the lyophilized histones were re-dissolved and allowed to unfold for maximally 3 hours at a concentration of 2 mg ml\textsuperscript{-1} in freshly prepared 6 M guanidinium chloride, 20 mM Tris pH 7.5 and 5 mM DTT solution. The histone proteins were mixed in exact equimolar amounts and diluted to a final total protein concentration of 1 mg ml\textsuperscript{-1}. The histone proteins were then dialyzed 3 times at least 6 hours against 2 M NaCl, 10 mM Tris pH 7.5, 1 mM EDTA and 5 mM 2-mercaptoethanol at 4 degrees using a 10 kD MWCO dialysis membrane. The assembled octamers were concentrated and subsequently purified by size exclusion chromatography on a superdex 200 HR column using 2 M NaCl, 10 mM Tris pH 7.5, 1 mM EDTA and 5 mM 2-mercaptoethanol as eluent. The obtained octamers were stored at 4°C.
5.6 - Acknowledgements

We thank Professor Ger Pruijn (Radboud University) for support and encouragement, Professor Christopher Schofield (University of Oxford) for kindly donating the constructs encoding for the histones, Abbas Al Temimi and Jos Kamps for synthesizing and purifying the alkylation reagents and Maud Wuts for her help during the initiation of the project. Additionally, I would especially like to thank Yvonne Grobben for her help and hard work.
5.7 - References


Appendix

Supplementary table 1 | Masses for H3C4 and its alkylation products as determined by ESI-MS.

<table>
<thead>
<tr>
<th>Alkylation product</th>
<th>Calculated mass (Da)*</th>
<th>Observed mass (Da)</th>
<th>Calculated mass increase (Da)*</th>
<th>Observed mass increase (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3C4</td>
<td>15213.7</td>
<td>15211.7</td>
<td>0.0</td>
<td>-2.0</td>
</tr>
<tr>
<td>H3K_c4</td>
<td>15255.8</td>
<td>15255.8</td>
<td>44.1</td>
<td>44.1</td>
</tr>
<tr>
<td>H3K_c4me3</td>
<td>15297.9</td>
<td>15297.8</td>
<td>86.2</td>
<td>86.1</td>
</tr>
<tr>
<td>H3K_c4me3</td>
<td>15314.8</td>
<td>15314.7</td>
<td>103.1</td>
<td>103.0</td>
</tr>
<tr>
<td>H3K_cAs4me3</td>
<td>15358.8</td>
<td>15359.8</td>
<td>147.1</td>
<td>148.1</td>
</tr>
<tr>
<td>H3K_c4me3</td>
<td>15295.9</td>
<td>15295.9</td>
<td>84.2</td>
<td>84.2</td>
</tr>
<tr>
<td>H3K_cprop4me3</td>
<td>15311.9</td>
<td>15311.8</td>
<td>100.2</td>
<td>100.1</td>
</tr>
<tr>
<td>H3K_cbutane4me3</td>
<td>15325.9</td>
<td>15326.0</td>
<td>114.2</td>
<td>114.3</td>
</tr>
<tr>
<td>H3K_cpentane4me3</td>
<td>15339.9</td>
<td>15339.6</td>
<td>128.2</td>
<td>127.9</td>
</tr>
<tr>
<td>H3K_cpyridine4me3</td>
<td>15317.9</td>
<td>15317.9</td>
<td>106.2</td>
<td>106.2</td>
</tr>
</tbody>
</table>

*Expected mass for the alkylation products is calculated relative to the observed H3C4 mass.
Supplementary figure 1 | Alkylated histone ESI-MS spectra. A) H3C4. B) H3K4. C) H3K7-4me3. D) H3K7-4me3 and several minor peaks indicating the deposition of multiple trimethylphosphonium groups. E) Peaks indicated with a double dagger correspond to a mass of +16 Da (ammonium adduct) and peaks indicated with an asterisk correspond to a mass of +43 Da (carbamyl adduct).
Supplementary figure 2 | Alkylated histone ESI-MS spectra. A) H3K<sub>c</sub><sup>4me3</sup>. B) H3K<sub>propane</sub><sup>4me3</sup>. C) H3K<sub>butane</sub><sup>4me3</sup> and a minor peak indicating the deposition of 2 butane-trimethyllysine groups. D) H3K<sub>pentane</sub><sup>4me3</sup> and several peaks indicating the deposition of multiple pentane-trimethyllysine groups. E) H3K<sub>pyridine</sub><sup>4</sup>. Peaks indicated with a double dagger correspond to a mass of +16 Da (ammonium adduct) and peaks indicated with an asterisk correspond to a mass of +43 Da (carbamyl adduct).
Supplementary figure 3 | Representative JARID1A-H3 ITC graphs. A) JARID1A-H3K4me3. B) JARID1A-H3K74me3. C) JARID1A-H3K94me3.
Supplementary figure 4 | SDS-PAGE of purified histone proteins.

Supplementary figure 5 | Dot-blot of assembled histone octamers. A) anti-Histone 3 and B) anti-H3K4me3.
General discussion and future perspectives
General discussion and future perspectives

Ever since the discovery that epigenetic marks can influence gene activity, the field of epigenetics has gained much interest. The investigations and research that have been conducted regarding the field of epigenetics have led to the development of a histone code hypothesis which implies that “multiple histone modifications, acting in a combinatorial or sequential fashion on one or multiple histone tails, specify unique downstream functions”.[1] Using this hypothesis as a guide many investigations have been conducted, not in the least due to the implications of epigenetics regarding many diseases.[2-7] A subgroup of epigenetic histone modifications is lysine methylation, a methylation state which can occur as mono-, di- or trimethyllysine.[8] The trimethyllysine state is an interesting PTM in the sense that this modification can be specifically recognized by reader proteins which contain a feature called an aromatic cage, composed of multiple aromatic residues.[9, 10] The aromatic cage is an important feature in trimethyllysine recognition as it has been hypothesized that it is capable of forming cation-π interactions with the trimethyllysine moiety.

In this thesis I mainly focused on the aromatic cage and how it contributes to binding of trimethylated lysine 4 on histone 3 (H3K4me3) to reader domain proteins. We investigated how subtle changes in trimethyllysine-containing histones and modifications of the aromatic cage of reader proteins can affect the reader-H3K4me3 association. Additionally, we developed tools for probing the involvement of cation-π interactions in the association of reader proteins with H3K4me3. In this chapter our fundamental findings are placed in a broader context and suggestions regarding future research in the chemical epigenetics field are described.

Effects of histone length on recognition by readers

Literature has shown that distinctive features in the H3 sequence are responsible for adequate recognition and sequence specificity for readers of H3K4me3.[8, 11] The presence of T3 in the H3K4me3 N-terminal sequence (ARTKme3QTARKS), for example, allows for sequence specificity over ARTKQTARKme3S due steric exclusion of the ARKme3 sequence by a tryptophan residue in the aromatic cage.[12-15] Furthermore, the ligands’ N-terminal Ala and Arg residues further increase specificity as some readers have additional binding pockets for these residues.

In our work, we have shown that the length and the composition of the H3K4me3 peptides have a marked effect on the recognition of the Kme3 mark by reader domains.[16] The shortest peptide sequence which is readily recognized by the TAF3, JARID1A, BPTF, SGF29 and JMJD2A H3K4me3 reader domains is a 4 amino acid sequence: ARTKme3. We have further shown that the N-terminal alanine is
important in generating sequence specificity for BPTF, JARID1A, TAF3 and SGF29. When our results are compared with studies which have been performed on erasers of methyllysine, it becomes obvious that the minimal sequence required for proper recognition of the methyllysine residue is not similar between readers and erasers. [17] The JMJD2A eraser domain, for example, requires an 8 amino acid H3 sequence (AA 7-14) for proper demethylation of its substrate. JMJD2C, on the other hand, only requires 4 amino acids (AA 7-10) for minimal activity, while its optimum activity was observed with an 8 amino acid peptide (AA 7-14). In contrast to readers of H3K4me3, erasers of H3K9me3 do not require the N-terminal sequence of H3 to be present for proper recognition of the substrate. When JMJD2A's reader and erased domains are linked, however, the combination of the H3K4me3 reader domain and the H3K9me3 eraser domain result in increased demethylase activity when compared to the truncated H3K9me3 eraser domain.[16, 18] This result provides an example of epigenetic cross-talk in c/s binding mode.[19]

The information obtained by our, and other, investigations should be further exploited for the development of peptide based inhibitors of reader proteins. Targeting readers of methyllysine is a research branch which is still in its infancy as only six of the many known methyllysine readers have been targeted with antagonists.[20, 21] Contrary, several methyltransterase inhibitors are already in the stages of clinical trial, clearly showing that the development of inhibitors of reader domains is lagging behind. In the past years a variety of KDM inhibitors have also been developed and several have now been enrolled into clinical trials.[22] The reasons for the slow development of reader domain inhibitors are the relatively low binding affinities, poor druggability of reader domains and a lack of complete understanding of the chemical origin of biomolecular recognition between posttranslationally modified histones and reader proteins.[23] Past research has successfully focused on more drugable targets, such as bromodomains, which led to the development of BET inhibitors.[24, 25] Future work of the scientific community should focus more closely on the development of antagonists of methyllysine readers, however, as the disease implications for this class of proteins are quite severe.[2]

**On the composition of the aromatic cage**

Some aromatic cages, such as SGF29, JMJD2A and TAF3’s, are complemented with a negatively charged residue, which may form an additional interaction with H3K4me3’s positively charged trimethylammonium moiety.[26, 27] As a general rule, mutations of the aromatic residues present in the aromatic cage of any Kme3 reader domain result in a reduction of binding affinities.[28] It has been shown though, that in some cases engineering of an aromatic cage can result in readers with novel functions.[29, 30]
Our own investigations with SGF29 have shown that the presence of aspartic acid in the aromatic cage is important for the proper recognition of both H3K4me2 and H3K4me3.[31] Abrogating the negative charge by mutation of the D266 residue resulted in marked reduction of binding affinity for both H3K4me2 and H3K4me3. Even the incorporation of an aromatic residue at the D266 site, which could have increased the cation–π interaction component, did not result in improved binding. Moreover, the D266F/Y/W mutations did not change the readers’ preference for Kme3 over Kme2, indicating the importance on the aspartic acid residue for the overall methyllysine recognition in the specific case of SGF29. By structural investigation, we further showed that hydrogen bonding of the K4me2/3 moiety to the SGF29 D266 residue is unlikely and that direct electrostatics between the methyllysine and aspartic acid residue are responsible for favorable interactions (Figure 1).

Contrary to our work on SGF29, Waters and co-workers did show that mutating the E52 residue complementing the aromatic cage of the HP1α chromodomain resulted in an increased selectivity of H3K9me3 over H3K9me2, when compared to the wild type variant.[30] Also, Li et al. showed that a single Y17E mutation of the BPTF PHD finger shifted its preferential recognition from H3K4me3 to H3K4me2 due to a reduced affinity for H3K4me3 while the affinity for H3K4me2 was maintained.[29] These

![Figure 1](image_url) **Figure 1** | **SGF29 reader domain.** Aromatic residues constituting the aromatic cage are indicated in red. The ARTKme3QTAR peptide is indicated in yellow and hydrogen bonds are indicated as blue dashes. PDBID: 3MEW.
reports in combination with our own work show that even though all readers of Kme3 possess an aromatic cage, different readout strategies within the aromatic cage are employed amongst the different readers. This feature may be exploited when developing specific Kme3 reader antagonists.

Taken together, our results show that the recognition of both H3K4me3 and H3K4me2 by SGF29 follows the same general mechanism. This conclusion is important, because changes in the aromatic cage that cause a shift in lysine methylation state preference can have drastic downstream effects. Recent work on the Dnmt3a ADD (ADD$_{3a}$), for example, has shown that a change in reader domain affinity from H3K4me0 to H3K4me3 has a marked downstream effect on the DNA methylation state in mouse embryonic stem cells (ESCs).[32] This change in DNA methylation state subsequently interfered with proper cellular differentiation leading to an aberrant phenotype. This type of work beautifully indicates the potential deleterious effects that can be invoked due to changes to the aromatic cage, which lead to aberrant functioning of the reader domain. Such negative effects have also been observed in certain cancers where reader domains are aberrantly fused to effector proteins, bestowing novel and destructive functions upon their fusion partners.[33]

**H3K4me3 recognition and cation-π interactions**

The most striking feature of H3K4me3 recognition by reader domains is the presence of the aromatic cage, which is supposed to recognize the trimethylated lysine state via a combination of cation-π and hydrophobic interactions.[9, 11] To what extent these interactions contribute to H3K4me3-reader domain association has not yet been extensively studied. Here we have shown, using several model reader domain proteins, that the cation-π component plays an important role in the K4me3-aromatic cage binding event.

In order to gain a more complete understanding of H3K4me3 recognition by the aromatic cage, we made a major effort in order to elucidate the chemical basis of H3K4me3 recognition by JARID1A, JMJD2A, TAF3, SGF29 and BPTF reader domains.[34] By replacing the positively charged trimethylammonium moiety of a H3K4me3 peptide by a neutral carba variant (K$^{\text{C}}$me3), we observed that binding affinities were not markedly affected for JARID1A, JMJD2A, BPTF and SGF29 despite the elimination of the positive charge, which is required for proper cation-π interaction. Even the binding affinity of TAF3, which was most strongly affected by the charge substitution, was only reduced by a factor 33. In contrast, replacing the Kme3 residue by glycine resulted in a $>500$-fold reduction in $K_d$. The relatively modest reductions in binding affinity for K$^{\text{C}}$me3 is an interesting result as the positive charge of the methylammonium moiety had been suggested as a major driving force for Kme3
binding by the aromatic cage. Previous work performed by Waters and co-workers has shown that in the specific case of the HP1 chromodomain, a reader of H3K9me3, cation-π interactions are indeed important in the H3K9me3 recognition event.[35] By substituting the positive charge of H3K9me3 for a neutral analog, a significant decrease in binding affinity was observed. Hence it was suggested that the positively charged methyammonium group was critical for binding the aromatic cage in the HP1-H3K9me3 association event. This conclusion was strongly supported by previous, extensive work on cation-π interactions by Dougherty and co-workers.[36, 37]

Our subsequent structural and computational investigations have shown that cation-π interactions are an important contributor to the reader domain-H3K4me3 binding. Interestingly, the abrogation of cation-π in reader domains is compensated by more favorable desolvation of the K9me3 ligand. Because K9me3 is more easily desolvated than its natural Kme3 counterpart, this results in an enthalpy-entropy compensatory mechanism which enables ligand binding to the reader domain, despite the reduced cation-π component.[34]

After obtaining this information, we used a complementary approach to investigate the role of the π-electrons present in JARID1A’s aromatic cage on the trimethyllysine readout process. By fluorinating the tryptophan residues in the half-cage, the electrostatic component of the reader-H3K4me3 cation-π interaction was reduced. Despite this reduction in the electrostatic component of the cation-π interaction, a marked decrease in reader-H3K4me3 binding affinity was not observed. Upon employing computational analyses, results similar to the data obtained from the neutral trimethyllysine analog work were obtained, namely that a desolvation-driven compensation mechanism was present. Due to the reduced electronegativity of the aromatic cage, desolvation of the aromatic cage became enthalpically more favorable (Figure 2) allowing for favorable binding of the H3K4me3 ligand. These results again confirm the importance of water in the cation-π driven binding of H3K4me3 to the aromatic cage of JARID1A.

It is understandable that the effect of water in the H3K4me3 readout has not been immediately recognized in the field of epigenetics, as it is difficult to investigate the involvement of water in biomolecular interactions. Past investigations and computational analyses have been performed on cation-π interactions where water has been excluded from the calculations, despite its known biophysical importance.[38-40] Recent advances in computational analyses of proteins, however, have allowed the scientific community to address the function of water in relation to biomolecular function.[41] It is now well recognized that water plays important roles in biomolecular processes and, as a consequence, is an important player in drug discovery.[42, 43]
CHAPTER 6

Installation of novel methyllysine analogs

The H3K4me3 PTM is a modification that is commonly studied by using short peptides that mimic histone 3. These peptides are generally synthesized using solid phase peptide synthesis (SPPS) which allows for the desired PTM to be readily incorporated at any desired location within the peptide. Unfortunately, several drawbacks of this method are: i) the requirement of a certain degree of expertise in SPPS and ii) the relatively short peptide length which can be reasonably prepared using chemical peptide synthesis. These limitations make that investigations with Kme3 and Kme3 analogs can only be performed by experienced chemists, or at the expense of great cost when peptides containing certain PTMs or PTM analogs need to be purchased. This strategy puts the more biological oriented community at a disadvantage. In order to address this problem, methods have been developed for the relatively simple inclusion of PTMs such as methyllysine and methylarginine onto

Figure 2 | Impression of the desolvation event of an aromatic cage. A) Aromatic cage containing water molecules. B) Binding of a trimethyllysine substrate to the aromatic cage. C) Binding of a trimethyllysine substrate to a fluorinated aromatic cage. PDB ID: 3GL6.
histone proteins using cysteine alkylation reagents.[44, 45] This method has been used successfully in order to place methyllysine analogs onto histone proteins which can subsequently be assembled into full nucleosomes.[46] Unfortunately, only a few Kme3 alkylation reagents have been developed: Kme3, Kme2, Kme1 and Kme0. These preexisting methyllysine analogs (MLA) unfortunately do not allow for the probing of subtle effects involved in epigenetic biomolecular processes, including the examination of the cation-π interaction.

Therefore, we have developed novel alkylation reagents, which allow for the incorporation of MLAs using the cysteine alkylation strategy (Figure 3). These reagents generate MLAs which differ in the side chains’ charge, charge density and the overall bulkiness.

**Figure 3** Methyllysine analogs. A) Methyllysine analogs as described by Shokat and co-workers: Cysteine, Kc (lysine) and Kcme3 (trimethyllysine). B) Novel methyllysine analogs: Kc^pme3 (phosphorus trimethyllysine), Kc^asme3 (arsenic trimethyllysine), Kc^cme3 (carbon trimethyllysine), Kc^pyr (pyridine lysine), Kc^propane me3 (trimethyllysine chain length increased by 1 carbon atom), Kc^butane me3 (trimethyllysine chain length increased by 2 carbon atoms) and Kc^pentane me3 (trimethyllysine chain length increased by 3 carbon atoms).
Our results have shown that these reagents can be used for the incorporation of MLAs using simple procedures similar to the procedure described by Shokat and co-workers.[44] This is an important result since these newly developed Kme3 analogues will allow more detailed future investigations on the role of trimethyllysine in epigenetics. It is known that the strength of cation-π interactions is dependent on the size, charge and location of the cation within a potential cation-π system.[36, 47] Using our MLAs, subtle effects of cation-π interactions between readers and Kme3 can be investigated without having to employ SPPS. Furthermore, the installation of MLAs on full length histones allow for the investigation of Kme3 binding on biologically relevant relevant structures such as nucleosomes and nucleosome arrays.[46, 48] By using our novel reagents, subtle events in Kme3 recognition can be probed and differences between various readers of Kme3 can be detected. These differences can later be exploited to design and synthesize specific inhibitors which may lead to novel drugs. Indeed, our previous work has already shown that cation-π interactions contribute differently to binding of H3K4me3 to various readers.[34] This variation is due to differences in the composition of the aromatic cage, which leads to subtle differences in the Kme3 recognition event.

The scientific community has often made creative use of the reactivity of cysteines in order to generate functional groups, which can be used to elucidate biochemical processes in great detail.[49, 50] Our work complements the existing repertoire of cysteine alkylation tools developed by Shokat and co-workers and the total repertoire of protein modification tools such as native chemical ligation, click chemistry and cysteine chemistry (such as cysteine alkylation and conversion of cysteines to dehydroalanine, for example).[44, 45, 51, 52] As the incorporation of MLAs using cysteine alkylation does not require in depth knowledge of chemistry, peptide synthesis or biochemical methods such as amber suppression, this is a convenient tool for the biochemical/chemical-biological community to study posttranslational modifications of histones and other proteins. Our alkylation reagents are additionally useful for the investigation of the molecular contribution of Kme3 on Kme3 recognition, Kme3 demethylation and histone modifications in a nucleosomal context and thus may have an impact on the biology-driven epigenetic field.
Future perspectives

The work described in this thesis has given us valuable insights in the recognition of trimethylated lysine 4 of histone 3. New information has been gleaned and resulted in a more thorough understanding of the most important chemical interactions involved in the recognition of the H3K4me3 mark by several reader domain proteins. Such insights are invaluable from both a fundamental point of view and for the purpose of drug discovery.

The work that has been presented in this thesis has mainly focused on a select group of representative H3K4me3 reader domains: JARID1A [53], BPTF [12], JMJD2A [13], SGF29 [27] and TAF3 [26]. These are not the only H3K4me3 reader domains and other readers such as ING2 [15], PHF8 [54], MLL5 [55], Spindlin1 [56], for example, should be considered for future investigation as well. By investigating the molecular mechanisms behind H3K4me3 recognition of more known readers it is possible to identify common and unique features amongst H3K4me3 readers. Such features should be exploited as it would allow for the development of drugs that inhibit specific readers or entire classes of readers. By exploiting the small differences in molecular recognition of Kme3, protein specific inhibitors may be designed whereas targeting the common features of Kme3 recognition within a specific class of readers may be used to design broad spectrum inhibitors.

Since the H3K4me3 mark is not only read, writers and erasers are equally important for the epigenetic regulation process. Therefore, it is imperative to elucidate how the writing and erasing processes take place in a living organism. ASH1 [57] and SETD7 [58], for example, write on H3K4, whereas the KDM1 and KDM5 families [59] are erasers of the H3K4me1/2 and H3K4me2/3 marks, respectively. The recognition of the H3K4me3 mark is critical for the function of these enzymes and it is important to understand exactly how such enzymes recognize their substrates. Similar to readers, understanding of the recognition mechanisms of these enzymes will lead to the identification of certain features, both common and unique, which may be exploited for the development of probes and inhibitors.

The combination of reader and writer/eraser domains is another point of interest. No reader or catalytic domain acts alone. It is a combinatorial effect which often leads to the desired downstream functions.[19, 56] With the information that is already known about readers, writers and erasers, future efforts should also be directed at the elucidation of the combinatorial effects of these individual domains. For example, studies on the H3 ligand length that is required for adequate binding or catalytic activity of the JMJD2A H3K4me3 reader and K3K9me3 catalytic domains respectively,
have shown that optimal binding and catalytic activity was observed with relatively short peptides (residues 1-4, ARTKme3 and 7-14, ARKme3STGGK respectively).[16, 17] Additional work by Lohse et al. has shown that, for JMJD2A, using 24 amino acid residue peptides containing both the H3K4me3 and H3K9me3 PTMs, the combination of a reader and eraser domain resulted in higher enzymatic activity than for a truncated eraser domain alone.[18] Such investigations should be continued for other proteins as well. The fact that nucleosomes that harbor specific PTMs can be assembled in vitro even allows for precise investigations of combinatorial domains that bind in cis or trans binding modes.

As already mentioned above, the molecular tools for trimethyllysine research may be used for further investigations of the recognition of this modified residue. The chemical-biological approach will allow us to investigate the function of specific readers and erasers, not only using peptides, but by employing whole reconstituted nucleosomes as well. Subtle effects of charge density, charge and bulkiness on downstream functions such as nucleosome array condensation can now be investigated using a relatively simple alkylation technique. Additionally, these tools may also be used for investigations regarding PTMs on non-histone proteins, broadening the application of our MLAs beyond histone modifications.

The results described in this thesis are not limited to readers of H3K4me3 or epigenetics in general. They are much broader than this. Water is indispensable for life on earth and most biochemical reactions and recognition events occur in an aqueous environment. Therefore, water will always be involved in the many biochemical processes that maintain life, stressing the importance to study and understand the role of this important molecule. In our work, we have shown the importance of water in the H3K4me3 recognition event by a variety of readers. Future work should aim to further elucidate the role of water in other epigenetic readout events. It is envisioned that this work cannot be solely executed by biochemists, chemical biologists and physical organic chemists, but the collective effort will undoubtedly shed light on the fundamental role of water behind epigenetic processes. The contribution of computational approaches and work performed by bio-physicists will be essential to uncover the role of this important molecule in detail.

Therefore, I urge that intensive collaborations between the various scientific disciplines are maintained and expanded, because it will not be possible to solve any of today’s major scientific challenges as a sole individual, research group or even discipline as a whole. Only by strong interdisciplinary collaborations it will be possible to make major advancements.
References

Summary in English
&
Nederlandse samenvatting
Summary

The H3K4me3 epigenetic mark has been shown to be involved in various cancers. Unfortunately, fundamental mechanistic knowledge regarding the recognition of this mark by reader proteins is still lacking. This information gap interferes with the development of drugs that may be used to treat H3K4me3 associated cancers. This is why the aim of the work described in this thesis was to elucidate the chemical basis of the interaction between trimethyllysine containing histones and readers of trimethyllysine. Citing the central question presented in chapter 1: "What are the chemical mechanisms behind the recognition of trimethyllysine by epigenetic reader domain proteins?"

In chapter 2 we have shown that the length and composition of H3K4me3 peptides had a marked effect on the recognition of the Kme3 mark by readers. It was found that the shortest peptide sequence that can still be readily recognized by TAF3, Jarid1a, BPTF, SGF29 and JMJD2A is the 4 amino acid sequence ARTKme3. We have further shown that the N-terminal alanine residue is important in generating sequence specificity for recognition by BPTF, Jarid1a, TAF3 and SGF29.

Our investigations with SGF29, as described in chapter 3, have shown that the presence of the D266 residue in SGF29’s aromatic cage is important for proper recognition of both H3K4me2 and H3K4me3. Abrogating the negative charge by mutation of this residue resulted in marked reduction of binding affinity for both H3K4me2 and H3K4me3. Structural investigations further showed that direct electrostatics between the methyllysine and aspartic acid residues are responsible for favorable interactions. Importantly, our investigations have shown that the molecular basis for the recognition of H3K4me2 and H3K4me3 by SGF29 are similar in nature.

Another important contributor to the trimethyllysine-aromatic cage binding event is the cation-π interaction. Chapter 4 describes a major effort that has been conducted in order to understand the chemical basis of the recognition of H3K4me3 by Jarid1a, JMJD2A, TAF3, SGF29 and BPTF. Upon replacing the positively charged trimethylammonium moiety of H3K4me3 by a neutral carba analogue, we observed that binding affinities were not markedly affected for Jarid1a, JMJD2a, BPTF and TAF3 despite the abrogation of the positive charge that is required for energetically favourable cation-π interactions. Supporting structural and computational investigations have shown that the abrogation of cation-π interactions is compensated by more favourable desolvation of the neutral carba moiety. This results in an entropically driven compensatory mechanism countering the loss of favourable enthalpy, indicating the important role of water in the reader-H3K4me3 interaction.
Additionally, chapter 4 presents an approach complementary to the substitution of the trimethylammonium's charge described above, in order to investigate the role of the π-electrons present in Jarid1a's aromatic cage. Instead of abrogating the trimethyllysine's positive charge, the tryptophan residues present in Jarid1a's half-cage were fluorinated using auxotrophic bacteria. This resulted in a charge reduction of the aromatic cage and therefore reduced the electrostatic component of the H3K4me3-reader cation-π interaction. Similar to the results obtained from the carba work, it was observed that an enthalpy-entropy compensatory mechanism was present. Due to the reduced electronegativity of the aromatic cage, desolvation of the aromatic cage became more favorable, compensating for the loss of favourable enthalpy. Both the carba and fluorination approaches therefore indicate that both cation-π and hydrophobic interactions are important in the reader-H3K4me3 recognition event. The results of both approaches also attest to the long known, but easily overlooked, important role of water in biological interactions.

The development and use of a tool that allows for the incorporation of novel, subtle trimethyllysine analogs (MLA) onto full length H3, is presented in chapter 5. Using a cysteine alkylation approach, we have shown that it is possible to incorporate novel MLA's into biologically relevant structures such as full length histones and histone octamers. Because the incorporation of MLA's using cysteine alkylation does not require specialist skills, this is a convenient tool for the biochemical/chemical-biological community. Using the developed MLA's, we have shown that a single carbon increase in length of the lysine side chain, a decrease in charge density of the methylammonium moiety and the abrogation of the methylammoniums positive charge appear to not markedly affect H3K4me3 recognition by Jarid1a.

The work presented in this thesis has elucidated a small part of the complex epigenetic machinery, aiding the development of novel drug substances. However, future research will remain indispensable in order to unravel the workings of epigenetics, beyond readers and beyond H3K4me3, in order to further facilitate epigenetic drug development. Future work should also aim to elucidate the role of water in epigenetic readout events beyond H3K4me3 recognition. It is recommended to pursue strong interdisciplinary collaborations in order to be able to adequately address these fundamental research lines.
Samenvatting

Het is aangetoond dat de H3K4me3 epigenetische markering betrokken is bij het ontstaan van verschillende soorten kanker. Helaas schiet de onderliggende fundamentele mechanistische kennis betreffende de herkenning van bovengenoemde markering door herkenningseiwitten nog tekort. Dit gebrek aan informatie belemmert de ontwikkeling van medicijnen die gebruikt zouden kunnen worden om dergelijke met H3K4me3 geassocieerde kankers te kunnen behandelen. Het doel van het werk dat in deze dissertatie beschreven staat, is dan ook het ophelderen van de onderliggende chemische basis aangaande de interactie tussen trimethyllysine bevattende histonen en trimethyllysine herkennende eiwitten. De kernvraag uit hoofdstuk 1 citerend: “Wat zijn de chemische mechanismes achter de herkenning van trimethyllysine door epigenetische herkenningsdomein eiwitten?”

Het werk beschreven in hoofdstuk 2 laat zien dat de lengte en samenstelling van H3K4me3 peptiden een aantoonbaar effect hebben op de herkenning van de Kme3 markering door herkenningseiwitten. Het is aangetoond dat het kortste peptid dat nog redelijkerwijs herkend wordt door TAF3, Jarid1a, BPTF, SGF29 en JMJ2A, de aminozuursequentie ARTKme3 is. Verder is aangetoond dat het alanine residu op de N-terminus belangrijk is voor het creëren van sequentie specificiteit voor de herkenning van het peptide door BPTF, Jarid1a, TAF3 en SGF29.

Het onderzoek aangaande SGF29, beschreven in hoofdstuk 3, laat zien dat de aanwezigheid van het D266 residu in de aromatische kooi van SGF29 belangrijk is voor adequate herkenning van zowel H3K4me2 als H3K4me3. Het elimineren van de negatieve lading in de kooi, door middel van een mutatie van dit specifieke residu, resulteert in een aanzienlijke reductie in bindingsaffiniteit voor zowel H3K4me2 als H3K4me3. Structuuronderzoek heeft verder aangetoond dat directe elektrostatische interacties tussen Kme3 en asparaginezuur waarschijnlijk verantwoordelijk zijn voor de gunstige interactie. Belangrijk om te melden is dat dit onderzoek heeft aangetoond dat de moleculaire basis voor de herkenning van H3K4me2 en H3K4me3 door SGF29 een gelijke oorsprong hebben.

Een andere belangrijke bijdrage aan de methyllysine-aromatische kooi binding wordt geleverd door de zogenaamde kation-π interactie. In hoofdstuk 4 wordt een aanzienlijk werk beschreven dat is uitgevoerd om de chemische basis achter de herkenning van H3K4me3 door Jarid1a, JMJ2A, TAF3, SGF29, en BPTF te achterhalen. Na de positieve lading in de trimethylammonium groep te vervangen door een neutrale koolstof variant (K4aMe3) werd geobserveerd dat de bindingsaffiniteiten van H3K4aMe3 met Jarid1a, JMJ2A, BPTF en TAF3 nauwelijks veranderden ten opzichte van het controle
peptide, ondanks het feit dat de positieve lading die een voorwaarde is voor het aangaan van een kation-π interactie was geëlimineerd. Additioneel structuur- en rekenonderzoek hebben aangetoond dat de eliminatie van de kation-π interactie wordt gecompenseerd door gunstigere dehydratatie van de methyllysine groep. Dit resulteert in een voor entropie gedreven compensatiemechanisme dat de reductie van gunstige enthalpie gedeeltelijk opheft. Dit wijst op de belangrijke rol die water speelt in het herkenningseiwit-H3K4me3 interactiomechanisme.

Hiernaast beschrijft hoofdstuk 4 een aan de koolstofvariant complementaire methode. Hierbij werd de rol van de π-elektronen in de aromatische kooi van Jarid1a onderzocht. In plaats van de eliminatie van de positieve lading van de methylammonium groep, werden de tryptofaan residuen in de aromatische kooi van Jarid1a gefluorineerd. Hierdoor wordt de lading van de aromatische kooi verminderd wat leidt tot een reductie van de elektrostatische component in de H3K4me3-herkennings-eiwit kation-π interactie. Vergelijkbaar met de resultaten van het koolstof variant onderzoek werd een enthalpie-entropie compensatorisch mechanisme waargenomen. Door het verminderd elektro-negatief potentiaal van de aromatische kooi wordt de dehydratatie van de aromatische kooi gunstiger, wat compenseert voor de reductie in gunstige enthalpie. Zowel de koolstofvariant- en fluorineeringsmethode laten dus zien dat zowel kation-π als hydrofobe interacties van belang zijn bij het Jarid1a herkenningseiwit-H3K4me3 interactiomechanisme. Deze resultaten getuigen van de reeds bekende, maar tevens makkelijk over het hoofd geziene, rol van water in biologische interactiomechanismen.

De ontwikkeling en toepassing van een methode die het mogelijk maakt om nieuwe, subtiel methyllysineanalogen (MLA) in te bouwen in volledig H3 eiwit is gepresenteerd in hoofdstuk 5. Gebruik makend van een cysteine alkylatie reactie is aangetoond dat het mogelijk is om de nieuwe MLA in te bouwen in biologisch relevante structuren zoals volledige H3 eiwitten en histon octameren. Omdat het inbouwen van de MLA aan de hand van deze methode nauwelijks het beheersen van zeer specialistische chemische vaardigheden vereist, is dit een methode die uitermate geschikt is voor gebruik door onderzoekers in het biochemisch/chemisch-biologisch werkveld. Gebruik makend van de ontwikkelde nieuwe MLA werd aangetoond dat de verlenging van de lysine zijken met één koolstoflengte, een reductie in ladingdichtheid van de methylammonium groep en de eliminatie van de positieve lading van de methylammonium groep in zijn geheel amper effect hebben op de herkenning van H3K4me3 door Jarid1a, JMJD2A, TAF3, SGF29 en BPTF.
Het werk gepresenteerd in deze dissertatie heeft een klein deel van de complexe epigenetische mechanismen in de cel opgehelderd, iets dat een bijdrage levert aan de ontwikkeling van nieuwe medicijnen. Uiteraard zal verder onderzoek onontbeerlijk blijven om de volledige werkingsmechanismen van de epigenetica te ontrafelen, om zo de ontwikkeling van medicijnen die aangrijpen op de vele epigenetische mechanismes mogelijk te maken. Toekomstig onderzoek zou zich ook moeten richten op het belichten van de rol van water in epigenetische herkenningsmechanismes, buiten de H3K4me3 hoek. Het is aan te raden om duurzame interdisciplinaire samenwerkingsverbanden aan te gaan om zo deze onderzoekslijnen adequaat te kunnen volgen.
Dankwoord,
CV & List of publications
Dankwoord

Na 4 jaar, eindelijk klaar! En nu vraag ik me af “Hoe heb ik het eigenlijk voor elkaar gekregen?”. Niet alleen, dat is zeker. Vanaf het begin van mijn promotietraject heb ik veel steun gehad van mensen om mij heen. Steun op wetenschappelijk vlak, maar zeker ook persoonlijke steun. Daarom wil ik hier diegenen bedanken die mij de afgelopen jaren hebben geholpen. Een bedankje zonder specifieke volgorde, gewoon zoals de personen in mij op komen terwijl ik schrijf. Ik wil bedanken:

**Pap & Mam.** Want zonder jullie geen Bas. Al zolang ik mij kan herinneren hebben jullie mij gesteund. Basisschool, middelbare school, hogeschool, universiteit... In het begin heb ik van jullie het zetje gekregen dat ik nodig had om te komen waar ik nu ben. Bedankt voor alles wat jullie me geleerd en gegeven hebben.

**Jasmin.** I would like to thank you for giving me the opportunity to work in your research group. Looking back and seeing all that we have achieved and learned in the past years I can tell that I have really grown as a scientist. **Abbas, Danny & Vijay.** Thanks for all the help and fun! You guys will be the engine that keeps the Mecinović group running and I am pretty sure that engine will never stop roaring! I wish you guys all the best. Just don’t forget to have fun and get some sleep every now and then (eat, sleep, science, repeat). Happy people are creative people and creative people make good scientists! **Jos & Roman:** Bedankt voor de avonden samen eten en alles waar we samen aan hebben gewerkt. **Jacky.** Jij hebt er voor gezorgd dat mijn promotie een stuk makkelijker is gelopen. Bedankt voor alle hulp.

**Ger.** Als mijn promotor wil ik je bedanken voor alle tijd die je in mijn promotietraject hebt geïnvesteerd. Daarnaast wil ik je vooral ook persoonlijk bedanken voor alle tijd en moeite. In de tweede week van mijn promotietijd heb ik je een email gestuurd met de vraag of ik eens langs zou mogen komen bij BMC om kennis te mogen komen maken, omdat ik waarschijnlijk wat werk bij jullie zou komen doen. De rest is geschiedenis. Ik heb bijna mijn hele promotietijd op de **BMC afdeling** doorgebracht. Ik werd warm ontvangen door de hele BMC groep (“tuurlijk ga ik mee naar de Aesculaaf!... Wat is de Aesculaaf eigenlijk?”) en ben vanaf die dag eigenlijk nooit meer weg gegaan. **Els, Carla, Wilma, Annemarie, Judith, Ilmar, Merel, Fleur, Sanne, Remon, Joyce, Cynthia, Chantal, Megan, Marina en Mathijs,** jullie hebben er voor gezorgd dat het een geweldige tijd is geworden! **Wilbert.** Ook jou wil ik bedanken. Tijdens mijn promotietraject stond ook jij altijd klaar met raad en advies. Daar heb ik veel aan gehad en ik heb je daarom ook altijd rijkelijk beloond met bier, kerstbomen, smileys en zilverpapier.
Els. Want zonder Els geen BMC, dat weet iedereen. Bedankt voor alle hulp en de gesprekken die we in de afgelopen jaren allemaal hebben gehad.

Erik, Maud en Yvonne. Het zou niet eerlijk zijn om jullie hier niet te noemen. Drie geweldige stagiaires die ervoor hebben gezorgd dat ik tijdens mijn promotietijd meer voor elkaar heb gekregen dan ik ooit in mijn eentje had kunnen doen. Een groot deel van dit boekje staat dankzij jullie inzet en hulp. Het was druk, maar het was vooral ook erg leuk om jullie te begeleiden en met jullie samen te werken!

The Chemical Biologists. So what happens if you take a bunch of scientists with different backgrounds, similar interests, rather dubious links to multiple research groups and put them together? They create their own group. Something a big bird would be proud of. The chemical biology group was one of the best things during my PhD career. Living proof that fun and science go hand in hand!


Ilmar & Selma. Voor de geweldige tijd die we hebben gehad! Van het opzetten van onze eigen chemical biology group tot de stedentrips die we hebben gemaakt. Ik geloof dat jullie de twee personen zijn die hier tussen de regels door het vaakst worden bedankt. En terecht.

De eio’s.

Kim, Selma, Fleur en Lianne (the Bonger Lab). De tegenhangers van de Mecinović groep. Daarnaastook nooit te beroerd voor een borrel. En altijd erg gezellig.

spelen we met de uitbreidingsset. Ik kan niet wachten om hem een keer mee te nemen naar zijn eerste “goede” concert.

**Michel.** Na meer dan 13 jaar nog steeds goede vrienden. In de zomer van 2006 belde je me op en vroeg je me het volgende: “Bas, we gaan zo naar een open dag. Ga je mee?” Dat is de *enige* reden geweest die ervoor heeft gezorgd dat mijn studentenleven is gelopen zoals hij is gelopen. Als je me toen niet had gebeld, was ik naar Groningen verhuisd om daar te gaan studeren (ik had al een kamer man!). Samen met *Erik* lijkt het erop alsof we nooit echt volwassen zijn geworden. Goeie zaak.

**Melissa.** Ook jij bent een van de belangrijkste redenen waarom mijn promotie is geslaagd. Bedankt voor alle telefoontjes, steun, vriendschap en festivals!


**Frans Allelyn.** Als er ooit een docent is geweest die er voor heeft gezorgd dat ik boven mezelf uit steeg dan was jij het wel. De dingen die ik van confetti heb geleerd gebruik ik nu nog.

**Rinse, Jeroen, Elroy en Chris.** Bedankt voor alle gezelligheid! Zonder jullie was mijn promotie een stuk minder ontspannen en veel te serieus geweest. De neomarxist-ische praatgroep waar de vampier wordt gebruikt als symbool voor de hedendaagse samenleving heeft mijn blik op de wereld vast geschaad.

**Bert.** Ein van de letste dinger wo ver ut euver höbbe gehad waor of ik noe waal of gein stropdas um mos bie mien sollicitatiegesprek in Nijmege. Ich ving ut joamer dat ste de res neet mje höbs kinne mitmake. Die baan höb ich gekrege. Zuug mich goan!

**Mariëlle.** De artiest van de familie. Degene die de omslag van dit boekje heeft gemaakt. Degene die de vrije geest heeft waar veel mensen zo jaloers op kunnen zijn.

**Kristel.** Zus nummer twee. Degene die mijn sociale leven in Nijmegen heeft gekickstart. Degene die er voor heeft gezorgd dat Nijmegen me met open armen heeft ontvangen.
Meggie
Mijn lieve Oma. Want de taarten waarmee ik tijdens de cake van de week de blits heb gemaakt waren naar jouw originele recept.

Lotte. Ik ben ontzettend blij dat ik je tegen ben gekomen. Zelfs op de meest frustrerende momenten en in de lange, late uurtjes van mijn promotie was je er voor me. Een promotie kun je halen. Maar de mooiste dingen moet je vinden.

And of course all the people of the 2nd floor, the 6th floor, from “across the street”, the people from the IMM-educational committee and all those whom I should have thanked but have not mentioned in person!

I hanks!

Bas
Curriculum vitae

Bas Johannes Gerardus Elizabeth Pieters was born on the 7th of June 1988 in Meerssen, the Netherlands. In 2006 Bas completed his higher general secondary education and graduated from Stella Maris College, Meerssen. Hereafter he started his bachelor’s education in the field of molecular life sciences at Zuyd University, specializing in biochemistry. After obtaining his bachelor’s degree in 2009, he continued his studies at Maastricht University where he obtained his master’s degree in molecular life sciences in 2011. Hereafter, Bas started working as a technician at the biotech company Pharmacell, in Maastricht for one year. Here he worked under cGMP conditions on multiple projects regarding regenerative medicine and biopharmaceuticals.

In 2012, Bas started his PhD trajectory at the Radboud University in Nijmegen under the guidance of Dr. Jasmin Mecinović. Here, he performed his studies mainly at the department of Biomolecular Chemistry headed by Prof. Dr. Ger Pruijn. During this time, Bas worked in the fields of epigenetics and supramolecular protein assemblies. His studies in the field of epigenetics concerned the elucidation of the underlying chemical mechanisms of H3K4me3 recognition by reader-domain proteins. Bas’ research in the area of supramolecular protein assemblies focused on a curiously concatenated protein assembly, CS$_2$ hydrolase. In 2013 Bas was awarded the NWO poster presentation award at the NWO-CW study group meeting for Protein research, Nucleic acids and Lipids & Biomembranes and in 2014 he was awarded the IMM award for best oral presentation at the annual sIMMposium.

During his PhD trajectory Bas supervised various bachelor and master students during their internships, coordinated projects of practical courses and gave a number of lectures for bachelor students. Beside his research and teaching activities, Bas has also been a member of the Radboud University Institute for Molecules and Materials educational committee for 3 years.

Since 2016 Bas is working as a scientist downstream processing at the contract research organization Batavia Biosciences in Leiden. Here, he is involved in multiple projects regarding the downstream process development for biologics such as vaccines and antibodies.
Publications


* These authors contributed equally to this work