Effect of Naturally Occurring DNA Modifications on DNA Structure and Packaging

Zhe Li

Trinity College

Department of Chemistry

University of Cambridge

September 2018

Supervised by Professor Sir Shankar Balasubramanian

This dissertation is submitted for the degree of Doctor of Philosophy

Effect of Naturally Occurring DNA Modifications on DNA Structure and Packaging

Zhe Li (ZL330@CAM.AC.UK)

Abstract

In eukaryotes, the genomic double-stranded DNA (dsDNA) coils around histones to form nucleosomes. Arrays of these nucleosomes bundle together to generate chromatin. Most DNA-related processes require interactions between chromatin-protected DNA and cellular machinery. Access of cell machinery to genomic DNA is partially regulated by the position and stability of nucleosomes, which may be influenced by changes in nucleosomal DNA. DNA is composed of adenine (A), guanine (G), cytosine (C), thymine (T) nucleotides and their derivatives. It has been shown that some C derivatives participate in directing multiple biological processes, and aberrant modification patterns are often linked to diseases. It has been proposed that T derivatives exhibit similar effects. This thesis focuses on elucidating the effect of naturally occurring DNA modifications on the properties of dsDNA and nucleosomes.

dsDNA sequences systematically modified with various T derivatives were characterized using classical biophysical techniques to assess the effect of these DNA modifications. The results indicate that in the sequence context studied, 5-hydroxymethyluracil modifications destabilize dsDNA, while dense symmetrical 5-formyluracil (fU) modifications alter the dsDNA structure. These effects may provide clues to the differential protein recruitment observed in previous research.

In vitro studies on nucleosome occupancy and stability revealed that 5-formylcytosine (fC) modifications have positive effects on nucleosome formation and stability compared to the unmodified counterpart by influencing the intrinsic biochemical and biophysical properties of the nucleosomes. These results provide casual links for the observation *in vivo* between fC and the increased nucleosome occupancy and positioning. In order to further understand the positional effect of fC on the nucleosomes, a method was developed for quick and reliable incorporation of C derivatives into dsDNA at desired positions.

The positive effect of fC modifications on nucleosome occupancy and stability observed here has necessitated further studies to gain deeper insights into the biological functions of fC in the nucleosome context. Cryo-EM can be used to elucidate the structural foundation for the changes fC posts to nucleosome, and protein interacting assays will identify the cellular machineries specifically recruited/repulsed by fC-modified nucleosomes.

The effect of DNA modifications elucidated by the above studies advances our understanding on the role that DNA modifications play in regulating cellular processes.

Declaration

The work described in this thesis was carried out by the author in the Department of Chemistry, University of Cambridge, under the supervision of Professor Sir Shankar Balasubramanian, between October 2014 and September 2018. Any work carried out in collaboration with other research groups or researchers is clearly indicated in the text.

The work described here is original except where indicated by references, and has not been submitted for any other degree at this or any other university.

This thesis contains fewer than 60,000 words.

Zhe Li, Trinity College

Signature

Date

Acknowledgements

First, I would like to take this opportunity to express my profound gratitude to Professor Sir Shankar Balasubramanian for giving me the valuable opportunity to do research in such an exciting and creative group, and for his inspirational and warm guidance over the past four years. His genuine pursuit and keen insight in uncovering nature's secrets will always be a great role model for me to follow. His broad knowledge and very detailed memory of key research always amaze me. He can always quickly look through where I got lost and enlighten me with the simplest and most helpful words. He always looks ahead for me and makes decisions that are always best for me in the long run. Shankar has built such a high-level and comprehensive team with experts on so many fields. Their brilliant minds and knowledge systems make me see more angles to questions, and enable help from all fields. Shankar also has many collaborators and demonstrated to me how to interact and communicate with researchers from other labs.

I would also like to give my sincere thanks to Dr Eun-Ang Raiber, Dr Pierre Murat and Dr Fumiko Kawasaki for their patient guidance, helpful advice and insightful questions whenever I am in need. Without them this dissertation would not have been possible. It really has been a great pleasure and a rare privilege to work with them for such a long time. They are exceptionally great teachers and have shown me how to do research. I always admire their passion for important questions, sharp insight into experimental results and ingenious ideas for tackling research questions, and I strive to achieve these myself one day. Dr Sergio Martinez-Cuesta undertook all the bioinformatical analysis for our experiments and did a fantastic job. He also taught me valuable statistical knowledge with great patience and clear explanation. He always covers details I have overlooked and asks great questions.

During my studies, I have been indulged with the luxury of working with many excellent researchers. It is always very helpful and inspiring to discuss obstacles I encountered with Jane, Robyn, Vicki and Areeb. When I needed help, Shiqing, David, Laurent, Gordon, Chloe, Barbara, Dhaval, Robert, Guillem, Marco, Tobias, Kit and Louis offered valuable suggestions. Whenever I am in doubt of English, Kim, Max and Darcie help me with great fun. Xiaoyun, Dr Aleksandr and Karen Sahakyan and Olivia make it such a relaxation to talk about everything and anything I think of and listen with great interests. I would also like to thank the group as a whole for providing such a stimulating and pleasant environment to study and work in. Watching Vicki and Robyn dancing to Step's songs on Friday afternoons in the Bio-Lab will always be my treasured memory. They and other Bio-Lab teammates have made the lab into such a happy and vibrant place that I deeply missed during thesis writing. I am really glad I can soon return to it and resume experiments.

Whenever I am in trouble, I can count on Chris, Jo, Jane, Areeb, Vicki, Robyn, Euni and Leanne to listen to my venting and give very constructive suggestions. When I am down I especially miss the hot chocolate Vicki took me to drink in Cyber Café.

I am grateful to Euni, Jo, Chris, Jane, Chloe, Pierre, Lucy Sixsmith and Shankar for the meticulous proofreading of this thesis. I accept full responsibility for any remaining or reintroduced errors. I am also very grateful to my mother and father and ASDA for preparing ready-to-eat meals so that I did not have to spend time cooking and could focus on pounding out this thesis on my computer.

I have the best fortune to have the constant and warm support of my loving father, mother and my boyfriend Dr Yang Jiong. When I am sad they always know how to cheer me up. When I was stressed about writing this thesis, my mother flew all the way to the UK without any delay to keep me company, help me with everyday life and provide me with warm encouragement. I would also like to give a special thank to Dr Ren Linyong. When I was asked by my scholarship funding body to provide a new local surety only days before my scheduled embarkment for PhD study, he graciously risked his entire family for a complete stranger so that I could come to UK and study. He also provided unreserved wisdom on various aspects of life from his own experience. He is really a dear family member to me.

During my time in Cambridge, I have also met many valuable and considerate friends outside of our research group. They immensely widened my perspective of the world, and make my life so much more colourful that I dared to imagine. They let me into numerous interesting worlds of music, poetry, education, theology, economics, law, architecture, medical research, physics, insect study, mathematics, sports, etc., so many that I cannot even list them all.

I would also like to thank the Department of Chemistry, Trinity College, the University Language Centre and the University of Cambridge for their constant support, useful training and extensive facilities. I also want to thank the technical and customer service teams from all the science companies I have harassed over the past years, especially NEB, ThermoFisher, Active Motif and Sigma. They always provided clear answers to my questions, and kindly arranged lightening-speed shipments when I was in a hurry for experiments.

Finally I would like to express my deep appreciation to A*STAR, Singapore, for providing generous funding so that I did not have to worry about money, as well as for their efforts to help me along my road of research.

Abstract

In eukaryotes, the genomic double-stranded DNA (dsDNA) coils around histones to form nucleosomes. Arrays of these nucleosomes bundle together to generate chromatin. Most DNA-related processes require interactions between chromatin-protected DNA and cellular machinery. Access of cell machinery to genomic DNA is partially regulated by the position and stability of nucleosomes, which may be influenced by changes in nucleosomal DNA. DNA is composed of adenine (A), guanine (G), cytosine (C), thymine (T) nucleotides and their derivatives. It has been shown that some C derivatives participate in directing multiple biological processes, and aberrant modification patterns are often linked to diseases. It has been proposed that T derivatives exhibit similar effects. This thesis focuses on elucidating the effect of naturally occurring DNA modifications on the properties of dsDNA and nucleosomes.

dsDNA sequences systematically modified with various T derivatives were characterized using classical biophysical techniques to assess the effect of these DNA modifications. The results indicate that in the sequence context studied, 5-hydroxymethyluracil modifications destabilize dsDNA, while dense symmetrical 5-formyluracil (fU) modifications alter the dsDNA structure. These effects may provide clues to the differential protein recruitment observed in previous research.

In vitro studies on nucleosome occupancy and stability revealed that 5-formylcytosine (fC) modifications have positive effects on nucleosome formation and stability compared to the unmodified counterpart by influencing the intrinsic biochemical and biophysical properties of the nucleosomes. These results provide casual links for the observation *in vivo* between fC and the increased nucleosome occupancy and positioning. In order to further understand the positional effect of fC on the nucleosomes, a method was developed for quick and reliable incorporation of C derivatives into dsDNA at desired positions.

The positive effect of fC modifications on nucleosome occupancy and stability observed here has necessitated further studies to gain deeper insights into the biological functions of fC in the nucleosome context. Cryo-EM can be used to elucidate the structural foundation for the changes fC posts to nucleosome, and protein interacting assays will identify the cellular machineries specifically recruited/repulsed by fC-modified nucleosomes.

The effect of DNA modifications elucidated by the above studies advances our understanding on the role that DNA modifications play in regulating cellular processes.

Abbreviations

A	adenine
Å	angstroms
A.U.	arbitrary unit
ACF	ATP-utilizing Chromatin Assembly Factors
ACN	acetonitrile
AID	Activation-Induced Deaminase
AML	Acute Myeloid Leukemia
APOBEC	Apolipoprotein B RNA-editing catalytic component
ATP	adenosine triphosphate
B	protected base
b.p.	boiling point
BER	Base Excision Repair
bp	base pairs
BS	bisulfite
C	cytosine
caC	5-carboxycytosine
cadCtp	5-carboxy-2’-deoxycytidine-5’-triphosphate
caU	5-carboxyluracil
CD	Circular Dichroism
CGI	CpG Islands
ChIP	Chromatin ImmunoPrecipitation
CMS	cytosine-5-methylsulfonate
CPG	controlled-pore glass
Cryo-EM	Cryo-Electron Microscopy
cryo-TEM	cryo-transmission electron microscope
Da	Dalton
dCtp	2’-deoxycytidine-5’-triphosphate
DMT	4,4'-dimethoxytrityl
DNA	deoxyribonucleic acid
DNMT	DNA MethylTransferase
dNTP	deoxyNucleoside TriPhosphate
ds	double-stranded
DTT	dithiothreitol
dTtp	deoxythymidine triphosphate
dUtp	deoxyuridine triphosphate

E. coli	*Escherichia coli*
EDTA	ethylenediaminetetraacetic acid
fC	5-formylCytosine
fdCtp	5-formyl-2’-deoxycytidine-5’-triphosphate
FRET	Förster Resonance Energy Transfer
fU	5-formylUracil
G	guanine
GF	Gel Filtration
GL	gap-filling ligation
H-bond	hydrogen bonding
H3K27ac	H3 lysine27 acetylation
H3K27me3	H3 lysine27 trimethylation
H3K4me1	H3 lysine4 monomethylation
HFIP	1,1,1,3,3,3-hexafluoro-2-propanol
hmC	5-hydroxymethylcytosine
hmdCtp	5-hydroxymethyl-2’-deoxycytidine-5’-triphosphate
hmU	5-hydroxymethylUracil
HP	Histone-PGA
HPLC	high-performance liquid chromatography
HS	Human α Satellite
JBP	J-binding protein
L	litre
LC-MS	liquid chromatography mass spectrometry
M	molar
MBD	methyl-CpG binding domain
mC	5-methylCytosine
MD	Molecular Dynamics
mdCtp	5-methyl-2’-deoxycytidine-5’-triphosphate
mESCs	mouse embryonic stem cells
MM	master mix
MNase	micrococcal nuclease
mol	mole
MQ H_2O	MilliQ$^{®}$ water
MW	molecular weight
$NaCNBH_3$	sodium cyanoborohydride
NAP-1	Nucleosome Assembly Protein 1
NGS	next generation sequencing

NMR	Nuclear Magnetic Resonance
nt	nucleotide
Nuc%	nucleosome occupancy
oxBS	oxidative bisulfite
PCR	Polymerase Chain Reaction
PGA	polyglutamic acid
phage	bacteriophage T2 virus
PolStop	Polymerase Stop
qPCR	quantitative real-time polymerase chain reaction
R	rough
redBS	reduced bisulfite
RNA	ribonucleic acid
ROS	reactive oxygen species
S	smooth
SAM	*S*-Adenosyl Methionine
SD	standard deviation
SELEX	systematic evolution of ligands by exponential enrichment
seq	sequencing
SMC	structural maintenance of chromosomes
SMRT	small molecule real time
SMUG1	single-strand selective monofunctional UDG1
ssDNA	single strand DNA
T	thymine
T. brucei	*Trypanosoma brucei*
TBE	tris-borate-EDTA
TDG	Thymine DNA Glycosylase
TEA	triethylamine
TET	Ten-Eleven Translocation
TF	Transcription Factor
Tm	melting temperatures
TSS	transcription start site
U	uracil
UDG	uracil DNA glycosylase
UV	Ultra-Violet
V	volume
xCs	mC, hmC, fC and caC

Table of Contents

1 Introduction

Table of Contents

1.1 A Brief History of Heredity and Genes

In 1866, Gregor Johann Mendel, the “Father of Modern Genetics”, published his paper “Experiments in Plant Hybridization”[1, 2], in which he described his observations on the inheritance of traits in peas. He proposed the principles of heredity with three basic laws of inheritance:

- the law of segregation, stating that the paired hereditary determinants segregate in equal probability during the formation of gametes;
- the law of dominance, introducing the concept of dominant and recessive genes;
- the law of independent assortment, suggesting all heredity factors worked independently from each other.

However, this great work was forgotten until 1900, when Hugo de Vries, Carl Correns and Erich von Tschermak independently rediscovered and verified Mendel’s observations in a variety of species. Hugo de Vries named the units responsible for inheritance of different traits “pangenes” in his book “Intracellular Pangenesis”. This was later shortened to “genes”, the term that is still used today.[3, 4]

Meanwhile, as early as 1875, Eduard Strasburger, Walther Flemming and Edouard van Beneden, described a cellular matter that was subsequently called “chromosome” by W. von Waldeyer-Hartz as this material absorbed basophilic aniline dyes strongly (chromo- for “color”, -some for “body”).[5] In 1902, Walter Sutton and Theodore Boveri observed that chromosomes separated and halved in number during meiosis.[6-8] This observation provided molecular explanations for Mendel’s law of segregation and law of independent assortment, establishing chromosomes as the basis of heredity.[2] Subsequently, Thomas Morgan and Alfred Sturtevant proposed gene theory stating that genes were in a linear arrangement on the chromosome, and constructed the first linear map of genes in Drosophila[9]. In 1928, Frederick Griffith used two different strains (R and S) of *Streptococcus pneumoniae* with differing pathogenicity to study the basis of heredity in mice. The fatal S strain formed “smooth” colonies through its production of a polysaccharide coating that protected it against the host’s immune system and caused pneumonia. The nonpathogenic R strain lacked the ability to produce this coating and instead formed a “rough” colony of nonvirulent bacteria. When Griffith injected a mouse with either a heat-killed S strain or a live R strain, the mouse did not develop pneumonia, however, when a combination of these two strains was injected, the mouse became infected and died. As a result, Griffith suggested that there was a component transfer (termed “transforming principle” by Griffith) from the heat-killed S strain to the live R strain which lead to the production of a polysaccharide coating and subsequent development of fatal pneumonia.[10] However, it was not clear whether the hereditary material was deoxyribonucleic acid (DNA), ribonucleic acid (RNA) or protein from the bacteria. Using an elegant *in vitro* enzyme digestion assay, Oswald Avery, Colin MacLeod and Maclyn McCarty demonstrated that Griffith’s bacterial transforming principle was only pathogenic when digested with ribonuclease and protease, and activity was lost when digested with deoxyribonuclease.[11] They concluded the hereditary material in bacteria is DNA. Furthermore, in 1952, Alfred Hershey and Martha Chase demonstrated that DNA was the hereditary material in a virus by using isotopic labelling experiments with bacteriophage T2 virus (phage) infecting *Escherichia coli* (*E. coli*).[12] The phage was known to contain protein, rich in sulfur, and DNA, rich in phosphorus. Therefore, to determine whether protein or DNA was the hereditary material, they cultured the phage and *E. coli* in both ^{32}P media and ^{35}S media separately. After the progeny phages were labelled with the respective isotopes, they were incubated with new *E. coli*, where the labelled phages infected unlabelled *E. coli* by injecting their heredity material into the *E. coli* for propagation. The old phage coats left outside infected *E. coli* were then separated from the media using a kitchen blender in order to discern them from the hereditary material inside *E. coli*. Finally when the new progeny phages broke out of *E. coli*, they were analyzed and revealed that the phage originating from the ^{32}P media culture retained over 30% of the original ^{32}P, while the phage originating from ^{35}S media culture retained less than 1% of the original labelling. As a result, they concluded that for the bacteriophage T2 virus, the DNA was the hereditary material, which supported the discovery made by Avery *et al.* in bacteria[11]. These discoveries further suggested that DNA could be the universal hereditary material.

1.2 A Brief History of DNA Research

Alongside the pursuit of the hereditary material identity, scientists have made amazing progress to advance our understanding of DNA. As early as 1869, Friedrich Miescher isolated an acidic material he named “nuclein” from the white blood cells in the pus on soiled bandages he recovered from a nearby surgical clinic. Using combustion experiments, a compositional analysis technique commonly used in that era, he discovered that the nuclein consisted of elements commonly found in protein and other organic molecules: carbon, hydrogen, nitrogen and oxygen. To his surprise, he also found phosphorus, which is not a major component in protein, but did not find sulphur, which is common in protein. By protease digestion and solubility tests with acid/base addition, Miescher was further convinced that nuclein is a distinct category of substance from known types of protein or any other known molecules.[13]

Following Miescher’s work, Albrecht Kossel identified the building blocks of nucleic acid as bases, sugars and phosphoric acids.[14] Using hydrolysis and combustion experiments, he and Julius Bodo Unger discovered the four different DNA bases and named them according to their sample sources: guanine (from sea birds faeces, known as “guano”)[15], adenine (as discovered in ox pancreas, from the Greek prefix “aden-”)[16], thymine (as from calf thymus)[17], and cytosine (as from cells, thus "cyto-" prefix)[18, 19], while the suffix "-ine" stands for nitrogen-containing compounds. The structures of the bases were proposed and confirmed by chemical synthesis.

In 1929, Phoebus Levene proposed a "tetranucleotide" structure based on the molecular formula determined by various chemical experiments, in which the four different nucleotides of DNA arranged sequentially[20], at the time this was widely accepted by the scientific society. The later discovery by Avery *et al.*[11] and Hershey *et al.*[12] that DNA was the hereditary material for the complex system of life came as quite a surprise and posed the question: how can genetic information ensure accurate replication and transmission to the next generation with such a simple tetranucleotide system rather than proteins with 20 different amino acids?

Roger and Colette Vendrely, together with their mentor Andre Boivin, provided further support for the proposal that DNA is the hereditary material. They demonstrated that all somatic cells of the same animal contain the same amount of DNA, and twice the amount of DNA in the nuclei of sperm cells[21-23], agreeing with previously observed chromosome action during gametogenesis.

In 1950, Erwin Chargaff disproved Levene's "tetranucleotide" hypothesis when he showed that the amounts of adenine (A) and thymine (T), and cytosine (C) and guanine (G) always occur in equal amounts using paper chromatography, but the relative amount of ATGC was not 1:1:1:1 in many species. This observation paved the way for the deduction of the correct DNA structure.[24, 25] Chargaff also observed that the G+C content of DNA varied from 22 to 73% depending on different species, but remained constant in all cells of the same organism.

Following the beautiful X-ray crystallography results of B-form DNA obtained by Rosalind Franklin in 1952[26], James Watson and Francis Crick at the University of Cambridge combined the vital evidence from previous work to solve the puzzle of DNA structure. They proposed the famous DNA double helix conformation in the Eagle Pub in 1953.[27] In this model, A pairs with T, and C pairs with G though hydrogen bonding, which provides a good explanation for Chargaff's observation of equal amounts of A:T and C:G. As Watson and Crick correctly pointed out in their 1953 Nature paper: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material".[27]

Since then, numerous DNA structures have been solved under different conditions by X-ray crystallography. The most common structure for double-stranded DNA (dsDNA) is a right-handed B-form formed under high humidity and relatively low salt condition. In this structure, the bases are perpendicular to the helical axis, with a wide major groove (side of bases) and a narrow minor groove (side of sugars and phosphate backbone). Each helical turn is composed of 10.5 base pairs (bp).[28] When the DNA crystals are grown under dehydrating and high salt conditions, A-form DNA can be observed.[29] With 11 bp per helical turn, this DNA is right-handed with the bases tilted with respect to the helical axis. Compared to B-DNA, the major groove of A-DNA is deep and narrow, while the minor groove is shallow and broad. Although considered not commonly formed by DNA in physiological conditions, A-form is the conformation of most dsRNA due to the 2'-hydroxyl group on the sugar.[28] In 1979, Alexander Rich and colleagues observed that with specific sequence context or under certain extreme conditions (such as 3 M $MgCl_2$ or NaCl, or with addition of alcohol), DNA can also assume a left-handed conformation with an extremely shallow major groove and a very deep and narrow minor groove with 12 bp per turn.[30] As its phosphate backbones are in zigzag lines rather than the smooth lines present in A- and B-form DNA, this novel DNA conformation was named Z-DNA.[30] Due to the demanding conditions for formation, the biological relevance of Z-DNA was doubted until several families of proteins were identified to have high specificity towards this structure, hinting its biological role.[31-33]

Gradually other forms of dsDNA such as C-[34], D-[35], E-[36] and F-forms[37] were also discovered under unique conditions. With the exception of Z-DNA, all variants of DNA conformation discovered to date are right-handed helices. In addition, DNA has been observed to form other structures, for example, threestranded[38] or four-stranded[39, 40] conformations, hairpins[41], cruciform four-way junctions[42, 43]. Continuous efforts have been dedicated to elucidate the biological relevance of these non-canonical DNA structures.

1.2.1 Structural Basis of DNA for Protein Recognition

In cells, genomic DNA exists in the double helical structure, and serves as the primary substrate for the maintenance and transfer of genetic information. The information embedded in DNA needs to be read out by the proteins in order to regulate downstream cellular activities. The special protein recognition was accomplished through a combination of direct and indirect readouts.[28, 44, 45]

1.2.1.1 Direct Protein Recognition

Direct protein readout is achieved by the structure complementarity specified by hydrogen bonding (H-bond) and van der Waals interactions between protein side chains and mainly the major groove of DNA base pairs.

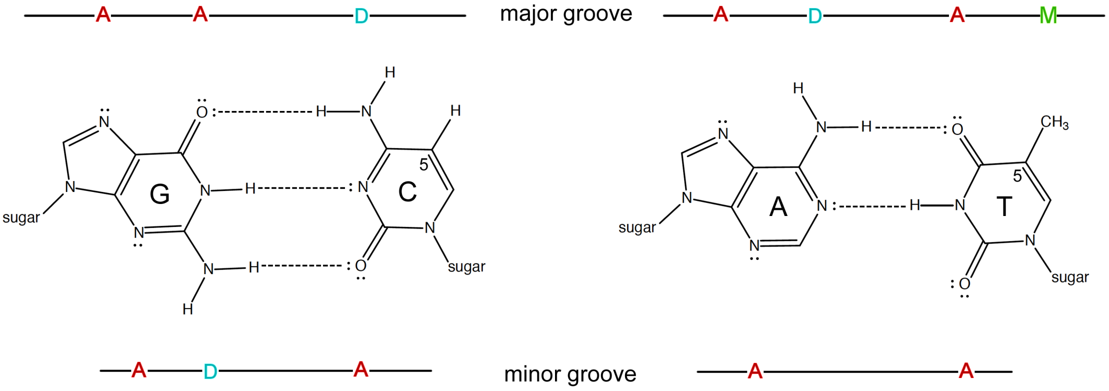


Figure 1-1. The hydrogen bond donor/acceptor pattern in the major and minor grooves of DNA. The H-bond donors were marked by blue “D”, the H-bond acceptors were marked by red “A”, while the hydrophobic methyl group was marked by green “M”.

Proteins are able to interrogate the unique H-bond donor/acceptor pattern exhibited by the DNA functional groups that are pointed towards the DNA major groove[46] by inserting a “recognition” alpha helix into the major groove of B-DNA[47]. As shown in Figure 1-1, the H-bond donor/acceptor pattern changes with the identity and the directionality of the base pairs. The $–CH_3$ group of T is a hydrophobic group that contributes to van der Waals interactions; the hydrogens of the amine groups of cytosine-C4 and adenine-C6 are H-bond donors (D in Figure 1-1); while N7 and the carbonyl oxygen of C6 of guanine, N7 of adenine, and the carbonyl oxygen of C4 of thymine are H-bond acceptors (A in Figure 1-1). The H-bond pattern is unique for each base pair in both directions.

The minor groove patterns are less variable, and only contribute to distinguish G:C/C:G base pairs from A:T/T:A base pairs through the protruding N2 of G as an H-bond donor. Superimposed crystal structures indicate that the directionality of G:C/C:G and A:T/T:A cannot be distinguished due to the almost identical locations of H-bond acceptors in both directions for both base pairs.[46]

1.2.1.2 Indirect Protein Recognition

Indirect protein recognition specificity is achieved through the easiness of assuming the correct conformation necessary for binding. This depends on the DNA flexibility and structure (groove width, basepair twist, etc.) of dsDNA, which are fundamentally modulated by DNA sequence context.[28, 44, 45]

1.3 Genome Organization in Cells

In most eukaryotic cells, the long genomic DNA (about 2 meters for a human diploid genome) needs to overcome the natural repulsion from the negatively charged phosphate backbones of DNA to fit into the confined space of the nucleus (diameter 10 µm for human)[48], while segments of genetic information need to be correctly and easily retrievable when needed. To achieve this intricate balance, the DNA is packaged into different levels of compaction with the assistance of different architectural proteins (Figure 1-2).

The first level of compaction is achieved by wrapping genomic DNA around a histone octamer to form a disk like structure called a nucleosome core particle. Histone proteins are rich in basic amino acids such as lysine and arginine and therefore carry positive charges in physiological conditions, balancing the negative charges of DNA. The histone octamer consists of two copies of each of the H2A, H2B, H3 and H4 subunits. The C-terminals of all subunits contain a histone fold domain which allows subunits to pair by docking in a “handshake” fashion: H3 and H4 associate to form a dimer and further into a tetramer, while H2A and H2B form a dimer and bind to the peripheral region of $(H3\text{-}H4)_2$ to form an octamer. The genomic DNA wraps around the histone octamer in 1.67 left-handed superhelical turns with about 9.4-10.9 bp per helical turn, 147 bp in total length.[49-51] The entry and exit points of DNA can further interact with the linker histone H1/H5, and become ready to be further compacted.[52, 53] Using electron microscopy, Amram Scheinfeld, and later Pierre Chambon, observed a beads-on-a-string pattern with uniform-sized particles (nucleosomes) evenly spaced in eukaryotic genome.[53-57] This was called "euchromatin" as these regions appeared light-colored when stained with Giemsa banding observed under a light microscope.[58] The genomic DNA in this level of compaction can be quickly made available to protein machineries such as DNA and RNA polymerases and regulatory proteins, allowing active transcription in these areas.

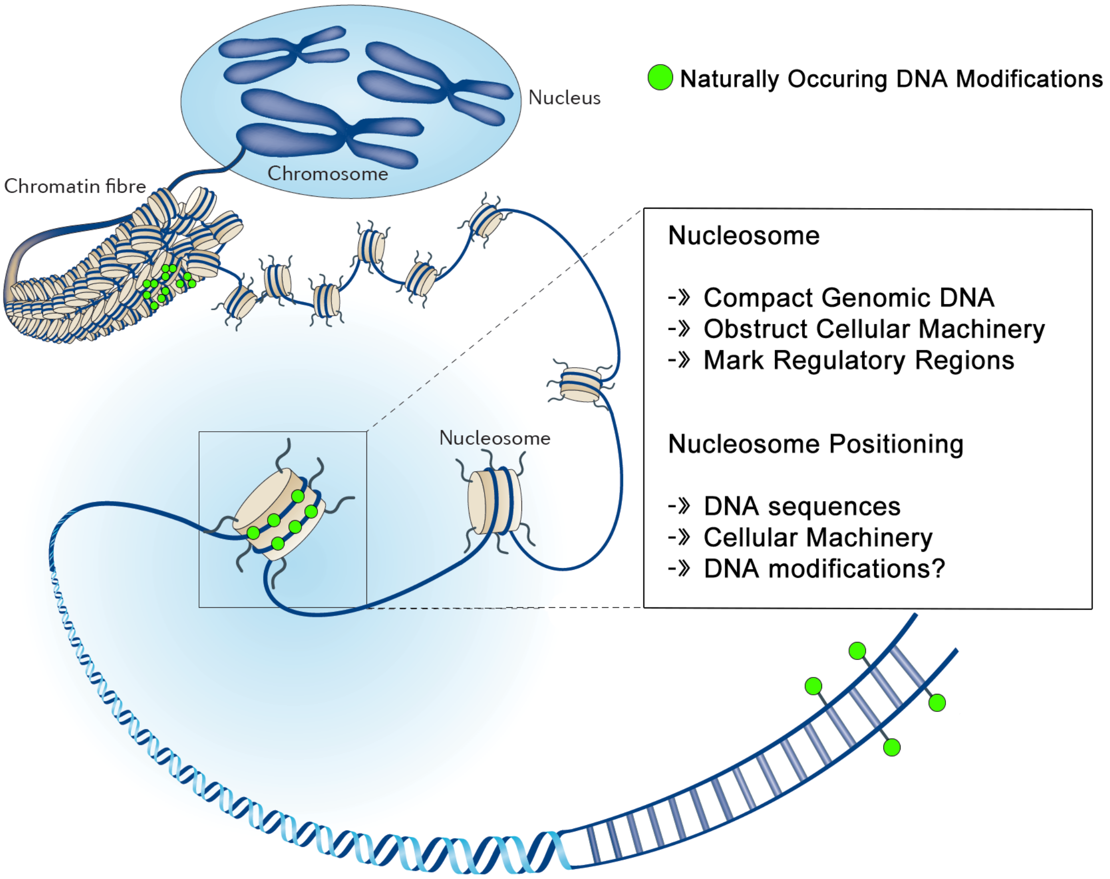


Figure 1-2. Schematic illustration of different levels of genomic DNA compaction. The figure is adapted from Arrowsmith *et al.*[59].

Most areas that are transcriptionally repressed are further packed into heterochromatin (so named as these regions appeared dark when stained). Heterochromatin is mostly located near the nuclear envelope on the peripheral region of the nucleus, as opposed to euchromatin.[60] Areas such as repressed genes, centromeres, telomeres and satellite sequences are packed into heterochromatin. With the help of adenosine triphosphate (ATP)-driven condensin complexes, part of the structural maintenance of chromosomes (SMC) family proteins, the chromatin is further compacted into chromosomes during mitosis and meiosis.[61-63]

1.4 The Definition and Mechanism of Epigenetic Regulation

Although all somatic cells contain almost identical DNA, it has been shown that the landscape of chromatin is dynamically reprogrammed at different cell phases and developmental stages. Additionally, different cell types display different phenotypical traits such as shape, size, function and lifespan. These differences are caused by epigenetic variations in the cells. The term "epigenetics" was originally coined in 1942 by Conrad Waddington, where the prefix "epi-" adds the meaning of "outside of" or "in addition to" to genetics.[64, 65] Previously epigenetics was defined to be the study of heritable changes in gene expression induced by chromatin architecture changes without alterations in the primary DNA sequence[66]. However, some histone modifications are not transmissible between generations of cells, e.g. the histone modifications in neuronal cells, which are non-dividing.[2, 67, 68] Therefore the latest definition of epigenetics encompasses both heritable and non-heritable epigenetic marks as "both heritable changes in gene activity and expression (in the progeny of cells or of individuals) and also stable, long-term alterations in the transcriptional potential of a cell that are not necessarily heritable".[2,69]

Epigenetic regulation is controlled by many factors, including DNA and RNA modifications, histone modifications and non-coding RNA[70-74]. Moreover, there is accumulating evidence for cross talk between different epigenetic modifications, adding another layer of complexity[75-77].

1.5 Widely Existing DNA Modifications

It was thought that DNA was solely composed of A, G, C and T, until 1899 when 5-methylcytosine (mC) was discovered in *tubercle bacillus*[78, 79]. The existence of mC in *tubercle bacillus* was confirmed by Johnson and Coghill in

1925[80], and subsequently in calf thymus by Hotchkiss in 1948[81]. Since then, a wide variety of modified DNA bases has been discovered in eukaryotes, prokaryotes and viruses.[82] To date, all of the known bases are variants of A, G, C and T (Figure 1-3).

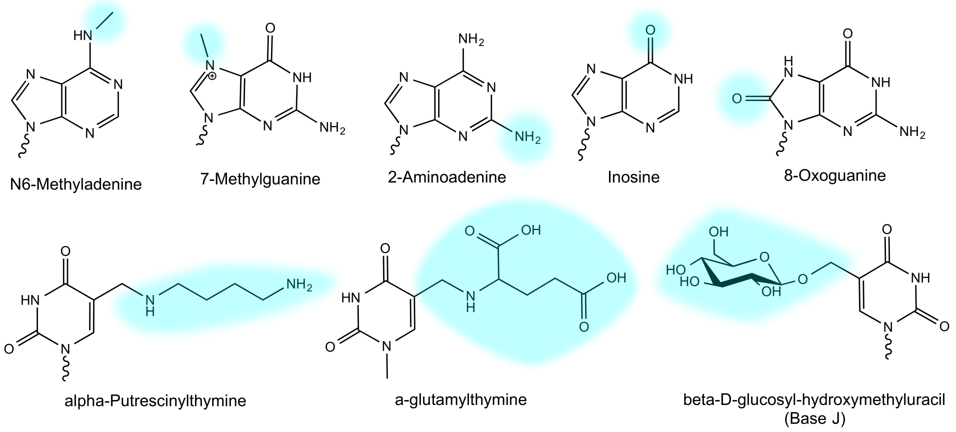


Figure 1-3. Examples of chemical structures of different DNA modifications found in nature.

Naturally occurring DNA modifications are generated and removed by enzymes. The natural abundance of modified bases varies across species, cell type, stage of development and throughout the cell cycle[82]. The modifications confer different biophysical and biochemical properties to the DNA, which can have a downstream effect on various cellular processes. The potential functions of DNA modifications have been inferred by their genomic loci, abundance at different developmental stage/tissue type/disease, and by colocalized histone modifications and interacting proteins.

1.6 DNA Modifications in Mammals

The discovery of naturally occurring nucleobases has stimulated studies to elucidate their pathway of generation, density and influence on biological processes. The C and T derivatives discussed in the next two sections are among the best characterized naturally occurring DNA modifications due to their biological relevance and natural abundance.

1.6.1 C derivatives: mC, hmC, fC and caC

1.6.1.1 Generation and Removal

mC is generated by the transfer of a methyl group from *S*-adenosyl methionine (SAM) to the C5 position of cytosine, catalysed by the DNA methyltransferase (DNMT) family of enzymes (Figure 1-4)[83-86]. DNMT3a and

DNMT3b are thought to methylate mammalian DNA *de novo* during early development and gametogensis[87], while DNMT1 maintains methylation by methylating cytosines in hemimethylated CpGs (that is, only one side of CpG is methylated) throughout the cell cycle[85, 88, 89]. Recently, Barau *et al.* discovered another *de novo* DNA methyltransferase, DNMT3c, in rodents, which is vital for mouse male fertility by methylating the promoter of evolutionarily young retrotransposons in the male germ line.[90]

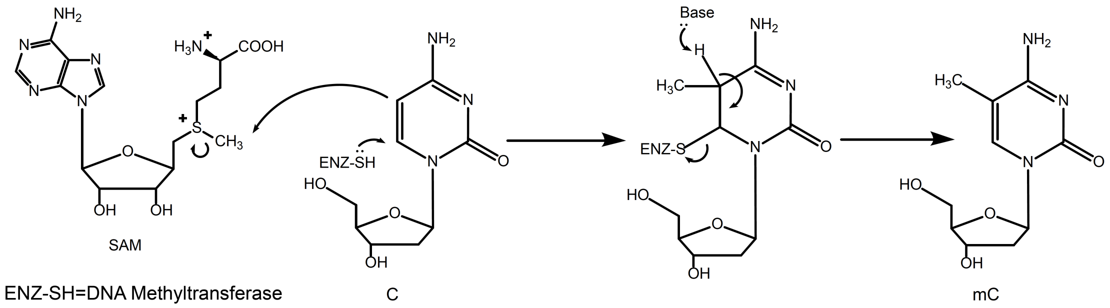


Figure 1-4. The mechanism of methylation of C to mC by DNMT methyltransferases and SAM.

Researchers have investigated the influential factors shaping the epigenomic methylation profile. It was observed that the GC ratio and CpG density of the primary DNA sequence have an influence on the DNA methylation level and histone modifications at some CpGs in the genome, although the exact mechanism has yet to be elucidated.[91, 92] In addition, factors like prenatal environments[93] and memory formation and learning process[94, 95] have also been shown to influence methylation level.

In plants, mC bases are excised by mC DNA glycosylase, and subsequently repaired via the Base Excision Repair (BER) pathway.[96] Mammals do not possess homologous mC DNA glycosylase, therefore it is inferred that mC is removed by alternative mechanisms.[96]

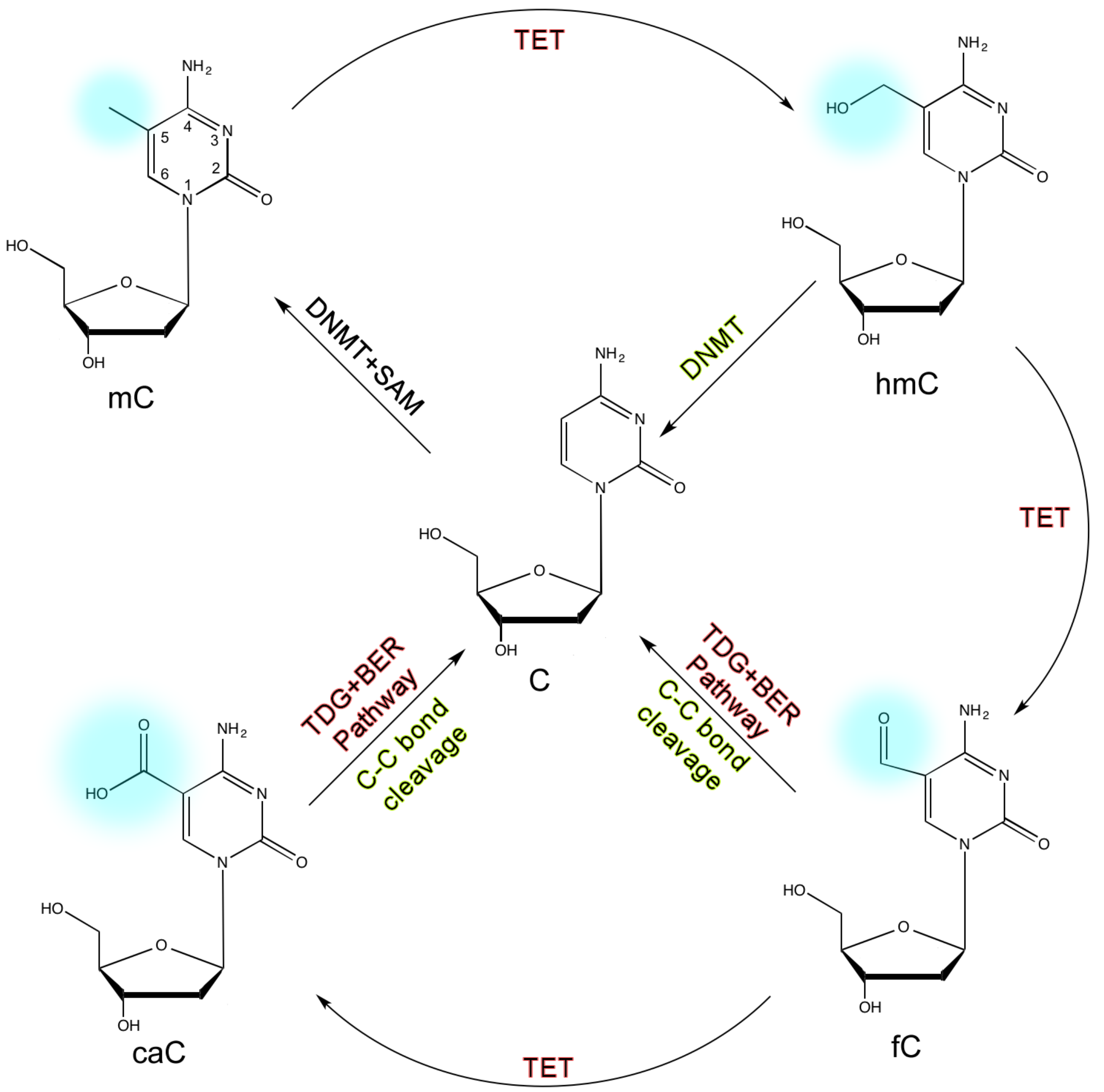


Figure 1-5. Schematic figure of mC generation and proposed active demethylation pathway (red) and alternative proposed demethylation pathway (green). The blue circles highlight the sites of the oxidation process.

In 2009, the oxidised derivatives of mC were (re)discovered.[82, 97-99] They are 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC) and 5-carboxycytosine (caC). Coupled with the discovery of 2-oxoglutarate Fe(II)-dependent dioxygenases, the Ten-Eleven Translocation (TET) enzyme family,[97, 98, 100-104] a hypothesis for an “active demethylation” pathway in mammalian cells was proposed, whereby mC is oxidized in a step-wise manner to hmC then to fC and caC which can be removed subsequently by the BER pathway (Figure 1-5).[105]

It has been shown that the TET3 enzyme is highly expressed in oocytes and zygotes after fertilization. It is hypothesized that this is in order to actively reprogram the DNA methylation landscape, but the concentration of TET3 drops rapidly while progressing to the two-cell stage. Conversely, the expression levels of TET1 and 2 are very low in oocytes and zygotes, yet increase during pre-implantation development, and become very high at the blastocyst stage.[106] The TET family is only found in eukaryotic cells, except plants.[101, 107]

The fC and caC in DNA can be recognized and excised by Thymine DNA Glycosylase (TDG), resulting in abasic sites, which can be subsequently repaired through the BER pathway.[105] *In vitro* biochemistry experiments showed that TDG binds to caC tighter than fC due to the negative charge, but excises fC faster than caC[99, 108-110]. An X-ray crystal structure revealed that TDG recognizes caC:A specifically, bends the DNA backbone towards the active site and flips out the base into a binding pocket stabilized by hydrogen bonding and van der Waals contact.[110] However, the mechanism of fC recognition by TDG is still not clear. There are several possible mechanisms proposed based on the caC excision mechanism. Since TDG bends DNA towards the active site, an increase in DNA flexibility may facilitate the substrate recognition. Ngo *et al.* have shown that the presence of even a single fC increases the DNA flexibility significantly, and that a caC modification slightly decreases the flexibility, potentially contributing to the preferential excision of fC over caC.[111, 112] In addition, base excision requires interrupting the base pair to flip the base out of the DNA groove into the active site, which can be facilitated by weaker Watson-Crick hydrogen bonding. The electron-withdrawing ability of –CHO (fC) and –COOH (caC) attenuates the electron-density of N3, reducing the base-pairing strength, and increasing the rate of base pair opening for TDG recognition.[113] Moreover, Hashimoto *et al.* proposed that fC/caC may form a tautomeric form that facilitates the base-flipping[109], and Maiti *et al.* proposed that –CHO and –COOH modifications reduce the N-glycosidic bond stability[108, 114], both contributing to the recognition and excision of fC and caC. More recently, Raiber *et al.* have proposed that the recognition of fC-DNA may be based on a DNA global structure change caused by dense symmetrical fCpG modification.[37]

To direct pluripotent stem cells to differentiate into distinct tissue lineages, genome-wide demethylation happens immediately after fertilization in mammalian cells, likely through a combination of both active demethylation and passive demethylation, which dilutes mC modifications through replication.[106, 115, 116] The proposed active DNA demethylation pathway could be a key process to program epigenetic information for mammalian development[117, 118] as the loss of TET1-3 together or TDG has been shown to cause devastating effects in mouse embryonic stem cells (mESCs) differentiation and is lethal for mouse embryos.[119, 120]

It is possible that mC removal through TDG excision and the BER pathway can lead to single strand DNA (ssDNA) breaks, which would result in a risk to genome integrity. In line with this it has been shown that the parental methylation reprogramming in mouse pronuclear zygotes is independent of TDG.[116] Therefore, an alternative demethylation pathway that does not require TDG and strand repair has been investigated.

It has been shown that hmC can be directly converted to C *in vitro* with the assistance of DNMT, and the inverse of this reaction is also possible in the presence of formaldehyde. This provides a potential pathway for hmC generation and removal.[121] It was observed by *in vivo* isotopic labelling experiments that fC and caC can directly revert to C through C-C covalent bond cleavage.[122, 123] Although the precise enzyme and mechanism have not yet been identified for fC, several potential candidates for proteins responsible for decarboxylation have been proposed[124-126].

In addition to the demethylation pathway detailed above, it has been shown that mC may be removed by cellular damage, such as spontaneous hydrolytic deamination to form T[127, 128], and by UV damage to T, C, and a series of C derivatives with C5 modifications[129], leading to potentially mutagenic results.

1.6.1.2 Distribution and Function of mC

The discovery of these naturally occurring DNA modifications has inspired substantial research to understand their distribution, density and potential functions in epigenetic regulation.

mC can be found across all domains of life[80-82, 130, 131]. In mammalian cells, the density of mC is on average 2-5% of all cytosine species[132]. The level of mC in mouse Embryonic Stem Cells (mESCs) and adult mouse cortex cells is 3.57% and 4.29% of all cytosine species respectively[130]. mC is enriched at CpG dinucleotides located outside CpG islands (regions enriched for CpG dinucleotides which present at most gene promoter regions) with the methylation mostly symmetrical on both strands[84, 133, 134]. At CpG islands, the methylation level is typically either fully methylated or fully unmethylated.[135] Methylated CpGs in promoter regions correlate with gene supression[136], while unmethylated promoters are more complicated. If active histone modifications such as histone H3 acetylation and H3K4 methylation are colocalized with methylated CpG, the genes tend to be transcriptionally active.[137] However, if bivalent histone modifications (that is both active and repressive histone modifications) are present, the promoters are termed “poised” and ready to be activated[138]. There have been observations that DNA methylation in the first exon of a gene correlates even more tightly with gene repression than in promoters, while downstream intragenic methylation is not associated with the magnitude of gene expression.[139] The transcriptional repression effect of mC has been hypothesised to be critical for maintaining chromatin structure, cellular functions and genome stability, such as heterochromatin formation at the pericentromeric area, X-chromosome silencing for dosage compensation, transposon and repetitive region silencing, cell pluripotency and genomic imprinting for marking parental-origin alleles[84, 136, 140-149]. There is an increasing amount of evidence of cross-talk between DNA methylation and histone modifications involving various chromatin remodelers during the aforementioned processes.[137, 138, 147, 150]

The link between DNA methylation and gene silencing has been rigorously studied. It has been shown that numerous transcription factors can no longer bind to DNA upon methylation[151]. Another mechanism for repression is the specific recognition of methylated DNA by transcription repressor methyl-CpG binding domain (MBD) family proteins, which can further recruit co-repressor complexes and induce transcription repression[76, 77, 152-154].

Changes in the DNA methylation profile can lead to many diseases.[155] In cancer cells, the DNA methylation pattern is aberrant and mostly occurs at specific sites.[156, 157] Regions like tumour suppressor genes which are normally unmethylated can become methylated and silenced and therefore no longer prevent cells from becoming cancerous.[158] In contrast, some normally silenced areas, such as transposons, are often demethylated in cancer cells and as such become actively transcribed, compromising transcription fidelity.[159] DNA methylation pattern alteration has been observed for other diseases as well, such as neurological disorders including Alzheimer's disease and X-linked mental retardation as well as autoimmune diseases.[160] Understanding the causality between the methylation level changes and disease state could provide new therapeutic directions.

mC needs to be removed to enable the expression of silenced genes, with hmC, fC and caC (xCs) proposed as the intermediates of the TET/TDG demethylation pathway.[143, 161, 162] Accumulating evidence of modification distribution, persistency and protein interactions indicates these proposed intermediates may also actively participate in regulating cellular functions[163]. All xCs (x = m, hm, f and ca unless indicated otherwise) have been detected in DNA extracted from various mouse tissues and mESCs, with the levels of xCs changing during differentiation.[164, 165] In mESCs all xCs have been shown to cluster in active enhancers, which are distal regulatory elements that assist in transcription initiation marked by active histone marks H3K27 acetylation (H3K27ac) and H3K4 monomethylation (H3K4me1), and poised enhancers, which are marked by both active histone marks and repressive histone marks H3K27 trimethylation (H3K27me3).[117, 164] All xCs have demonstrated differential protein binding and chromatin remodeller/transcription factor recruitment ability *in vitro*[163, 166, 167, 168], suggesting the addition of another layer of regulation for shaping chromatin architecture and directing cellular activities.

1.6.1.3 Distribution and Function of hmC

The density of hmC in mESCs and adult mouse cortex cells is 0.36% and 0.57% of all C species respectively[130]. Besides active and poised enhancers, hmC is also enriched at transcription start sites[169, 170] and exons[171] of actively transcribed genes and promoter regions of mESCs[161, 172]. The density of hmC varies drastically with tissue type with the highest density found in the central nervous system[173-176] where it increases during brain development[175, 177]. To determine whether hmC accumulates through oxidative damage, Munzel *et al.* showed that the level of 8-oxoguanine, a typical oxidative stress marker, did not correlate with hmC levels, indicating that hmC is not generated by oxidative stress.[175]

Isotopic labelling experiments showed that hmC is a predominantly stable modification in cultured cells and *in vivo*[176], which indicates that hmC is actively maintained and distinct from mC. Moreover, hmC has shown distinctly different protein-binding ability from mC[178, 179] such as helicases (Harp, Recql, etc.) and DNA glycosylases (Neil1 and Neil3)[163, 166, 180] indicating that hmC is involved in a demethylation pathway involving DNA-repair. Protein factors in neuronal progenitor cells (NPC) and brain have shown affinity to hmC, such as Wdr76, Thy28, and Uhrf2, suggesting that hmC might play a role in epigenetic regulation in these tissues.[166] Protein pull-down experiments with hmC also revealed that hmC specifically interacts with replication factor C (Rfc1-5) which suggests that hmC may also participate in replication regulation.[166] Additionally, hmC displayed a small blocking effect on transcription, while no mutagenic effect has been observed *in vitro* or *in vivo* with hmC modifications.[181]

hmC and TET enzymes have been shown to correlate with pluripotency markers to maintain pluripotency and regulate cell differentiation. Reduced hmC levels and TET-deficiency have been identified as biomarkers for Acute Myeloid Leukemia (AML) and melanoma.[182, 183] Aberrant hmC levels have also been correlated with degenerative neurological diseases such as Alzheimer's disease, Huntington's disease, and psychiatric disorders.[184, 185] Therefore hmC may play a fundamental role in the epigenetic regulation of transcription[161], cell proliferation[176], brain development[166, 175, 177], cancer progression[182, 183] and cognitive function maintenance[184, 185].

1.6.1.4 Distribution and Function of fC

The level of fC in mESCs and adult mouse cortex cells is around 0.0048% and 0.00019% of all cytosine species respectively[130], with fC clustering at certain genomic loci to a level comparable with hmC[98, 186]. In mESCs, besides active and poised enhancers[37, 117, 165, 187], fC is enriched in the CpG islands of promoters, exons and introns of gene bodies associated with transcription, cell differentiation and development.[162, 165, 188] As differentiation of embryonic stem cells proceeds, the density of fC drops sharply, indicating its involvement in epigenetic reprogramming[98]. Isotopic labelling experiments demonstrated that fC is a predominantly stable modification in cultured cells and in the brain[177], with semi-permanent modifications found at certain genomic loci[189], indicating that fC is actively maintained for yet unknown functions. Single base resolution sequencing of fC revealed that fC is clustered in the $(CpG)_n$ areas ($n \geq 3$) in mESCs and mouse two cell embryos and is symmetrically distributed in complementary strands at certain genomic loci.[37] This aligns with the observation that TET prefers to generate and maintain consecutive and symmetrical fCpG on both strands.[37, 104, 118, 187, 190] fC distribution also overlaps with mC and hmC at some genomic loci.[186, 188] In addition, fC-modified DNA selectively binds to chromatin remodelers (such as NuRD complex) and transcription factors (such as FOXK1) *in vitro*, suggesting fC may participate in genome regulation [163, 165]. Because fC is enriched in the CpG islands of promoters of transcriptionally active genes that are frequently bound to RNA polymerase II and correlates with elevated levels of active histone marks H3K4me3 and H3K27ac[191], it is considered to be linked to active transcription[163, 165].

1.6.1.5 Distribution and Function of caC

The density of caC in mESCs is about 0.00029% of all cytosine species, but has not been detected in mouse cortex DNA to date[130]. *In vitro* studies have shown that caC specifically recruits various cellular machineries[166, 168], such as BAF170, a subunit of a chromatin remodeller complex[166], suggesting that caC may also participate in cellular regulation. Both fC and caC have been observed to slow down RNA transcription polymerase Pol II *in vitro* and in human cells with a marginal mutagenic effect when located on the transcribed strand. This may have implications in transcription. [181, 192]

1.6.2 T derivatives: U, hmU and fU

1.6.2.1 Generation and Removal

Uracil (U) in DNA can be generated by various pathways via enzyme activity or DNA damage.[193] For example, deoxyuridine triphosphate (dUtp) can be erroneously incorporated into genomic DNA instead of deoxythymidine triphosphate (dTtp) during replication.[194] Spontaneous or enzyme-driven hydrolytic deamination of C can also produce dU, resulting in a mutagenic UG mismatch as U preferentially base-pairs with A rather than G during replication and transcription[82, 193, 195-197]. For this reason, U is classified as a T derivative in this thesis.

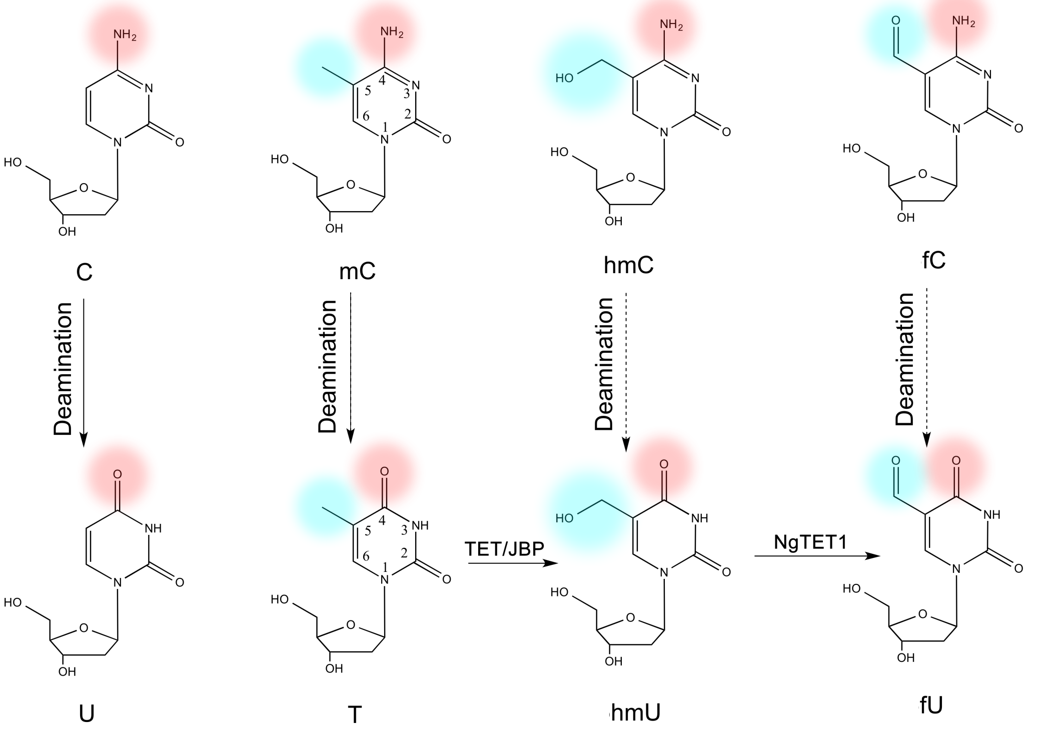


Figure 1-6. The inter-conversion between C derivatives and T derivatives. The solid arrow indicates experimentally observed biological processes, and the dashed arrows indicate proposed biological processes not yet observed experimentally[98, 130]. The blue circles highlight sites of the oxidation process, and the red circles highlight the sites of the deamination process.

Analogous to xCs, oxidized T derivatives 5-hydroxymethyluracil (hmU) and 5-formyluracil (fU) have been detected in eukaryotic DNA, while 5-carboxyluracil (caU) has not been detected *in vivo* to the best of my knowledge.

hmU can be formed by oxidative damage or by enzymatic processes.[130, 198-200] The enzymatic conversion of T to hmU is catalysed by TET enzymes in mESCs and by the J-binding protein (JBP) family (TET homologs) in trypanasomatids (Figure 1-6)[130, 198]. Researchers have attempted to identify an alternative active demethylation pathway in which hmC deaminates to form hmU catalysed by Activation-Induced Deaminase (AID) or Apolipoprotein B RNA-editing catalytic component (APOBEC), analogous to the enzymatic deamination of C to produce U. However, conflicting results were observed[201, 202]. Isotopic labelling experiments with TDG knockdown cells led to the observation of trace amounts of hmU originating from hmC; while in wild-type cells, all hmU bases were generated from T. Therefore the hmU derived from hmC may be quickly removed by TDG and therefore not detectable *in vivo*.[130]

fU is the subsequent oxidation product of hmU (Fig 1-4). Since the enzymes responsible for such an oxidation reaction *in vivo* have not yet been identified, fU is traditionally considered as an oxidative lesion produced by reactive oxygen species (ROS).[200] Recently, *in vitro* oxidation from hmU to fU and further to caU by NgTET1, a TET/JBP-like protein from *Naegleria gruberi*, has been identified.[203]

U, hmU and fU (xUs) can be excised by several DNA glycosylases and subsequently repaired through the BER pathway. When paired with A, the xUs can be excised by uracil DNA glycosylase (UDG) and single-strand selective monofunctional UDG1 (SMUG1), but when mis-paired with G, xU:G mismatches can be recognized and repaired by TDG and MBD4.[193, 204-207]

1.6.2.2 Density and Distribution

U has been detected *in vivo* in the genomic DNA of prokaryotes and eukaryotes[82, 195]. The average density of U has been reported to be 400-600 bases per human or murine genome[208]. The measured U density is dramatically different according to different reports, which is likely due to spontaneous deamination during DNA extraction and digestion prior to quantification, leading to significant experimental variation.[79, 208]

hmU and fU were first discovered in eukaryotic cells[130]. hmU has been observed to exist in relatively high densities in eukaryotic parasites, such as *Leishmania* (0.01% of all T)[209] and *Trypanosoma brucei (T. brucei,* 0.02% of all T)[210, 211]*.* The level of hmU in mESCs and adult mouse cortex cells is about 0.00017% and 0.00003% of all thymine species respectively[130]. Initially hmU was considered as an oxidative lesion of T induced by ROS and ionizing radiation[212, 213], however the density of hmU is higher in embryonic cells than in adult cells, suggesting that there are active processes which regulate the hmU density *in vivo*[130]. The level of hmU is correlated with TET expression level and changes upon cell differentiation.[130, 203] In addition, hmU is specifically recognized by chromatin remodelers and transcription factors, such as AP-1, a transcription factor related to stress response.[130, 214] It has been shown that hmU enhances transcription by interacting with bacterial RNA polymerases at some promoters in *E. coli*.[215] Therefore, hmU may be involved in transcription regulation in bacteria. In humans, the increased level of hmU mononucleoside has been observed in invasive breast cancer and could potentially be used for breast cancer prognosis in liquid biopsy.[216]

In *T. brucei*, the fU density is around 0.032% of T.[210] The density of fU in mESCs and adult mouse cortex cells is 0.00086% and 0.00069% of all thymine species respectively[130]. *In vitro* experiments have demonstrated that the presence of fU inhibits the interaction of AP-1 with DNA[217] and therefore suggested fU may influence transcription. The tautomeric form of fU may form a wobble base pair with G, inducing mutagenic results.[218, 219]

1.6.3 Effect of DNA Modifications on Protein Recognition

DNA modifications add another layer of dynamically reprogrammable information without altering the underlying DNA sequence and the introduction of DNA modifications may either facilitate or block the protein functions. In addition, it has been shown that DNA modifications enable specific protein recruitment/repulsion and influence genomic functions at the loci of modifications.[163, 166-168] This may be accomplished through influencing both direct and indirect readout.[44]

1.6.3.1 Direct Protein Recognition

Naturally occurring DNA modifications add new chemical functionalities into the major grooves of DNA that generate variations of the major groove code, which may alter their protein recognition; the minor groove code however is not influenced directly (Figure 1-7). The hydrophobic $–CH_3$ group of mC contributes to van der Waals interactions. By contrast, the $–CH_2OH$ group of hmC and hmU, and the –COOH group of caC are H-bond donors (D in Figure 1-7); while the –CHO group of fC and fU is a H-bond acceptor (A in Figure 1-7). The only duplicating H-bond pattern is Acceptor-Donor-Acceptor-Acceptor formed by fC:G and A:fU basepairs. Since the fC:G is a pyrimidine:purine basepair, while A:fU is a purine:pyrimidine basepair, the exact positions of H-bond donor and acceptor differ between the two basepairs.

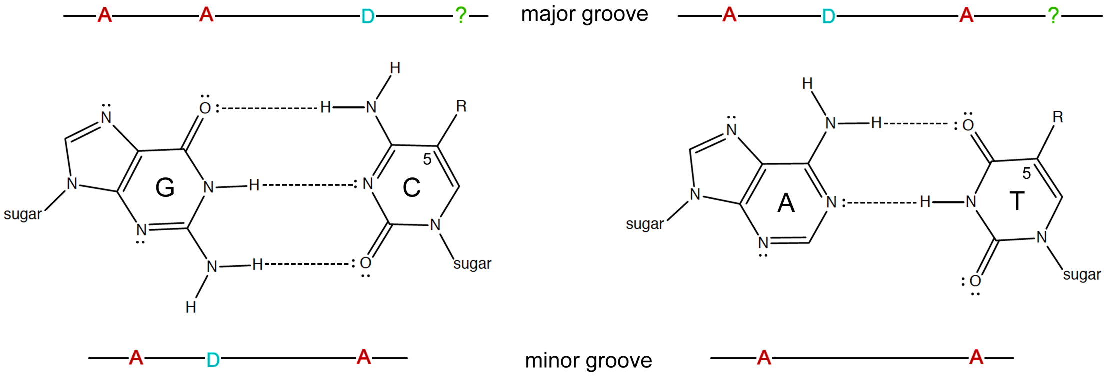


Figure 1-7. The hydrogen bond donor/acceptor pattern in major and minor groove of DNA. The H-bond donors were marked by blue "D", the H-bond acceptors were marked by red "A", while the C5 positions where additional functional groups are located were marked with green "?" mark.

1.6.3.2 Indirect Protein Recognition

The additional functional groups of DNA modifications may confer different structure and flexibility to the DNA double helix and either facilitate or block indirect protein recognition. In addition, DNA modifications may regulate cellular processes through influencing the stability of the DNA substrate for proteins. For example, transcription and replication both require initial strand separation for the polymerases to bind.[220, 221] Hence changes in the stability of dsDNA by DNA modifications may serve as an additional level of regulation for these fundamental biological processes.

1.7 Objectives of PhD Project

From the dynamic nature of DNA modification distribution, persistency and interacting cellular machineries throughout development, it is evident that naturally occurring DNA modifications are of fundamental importance to cellular activity regulation, and aberrant DNA modification patterns correlate with diseased states. However, the molecular basis for the causal link between the epigenetic DNA modifications and phenotypical changes of complex cellular processes awaits further elucidation by strictly controlled *in vitro* experiments in model systems (Figure 1-8).

Changes to the DNA stability and structure by naturally occurring DNA modifications can influence protein binding and subsequent cellular processes. Therefore understanding how DNA modifications impact the biophysical properties of DNA is fundamental to comprehend the link between DNA modifications and biological functions. There are numerous studies on the effect of naturally occurring C modifications on the biophysical properties of DNA, however the effect of T modifications is less understood. Chapter 2 of this thesis reports on the investigation into the effect of T modifications on dsDNA by Ultraviolet (UV) thermal denaturation experiments and Circular Dichroism (CD) spectroscopy in various DNA sequence contexts and modification density in order to address these questions.

Changes to the biophysical properties of the DNA double helix, discussed in Chapter 2, may impact the formation and stability of nucleosomes and downstream cellular functions. Chapter 3 of this thesis includes a systematic evaluation of the effects of naturally occurring DNA modifications on nucleosome occupancy and stability to understand how DNA modifications influence the nucleosome organization in cells.

The aim of this thesis is to investigate the effects of DNA modifications in dsDNA and nucleosomes to better understand their influence on chromatin architecture and downstream biological processes.

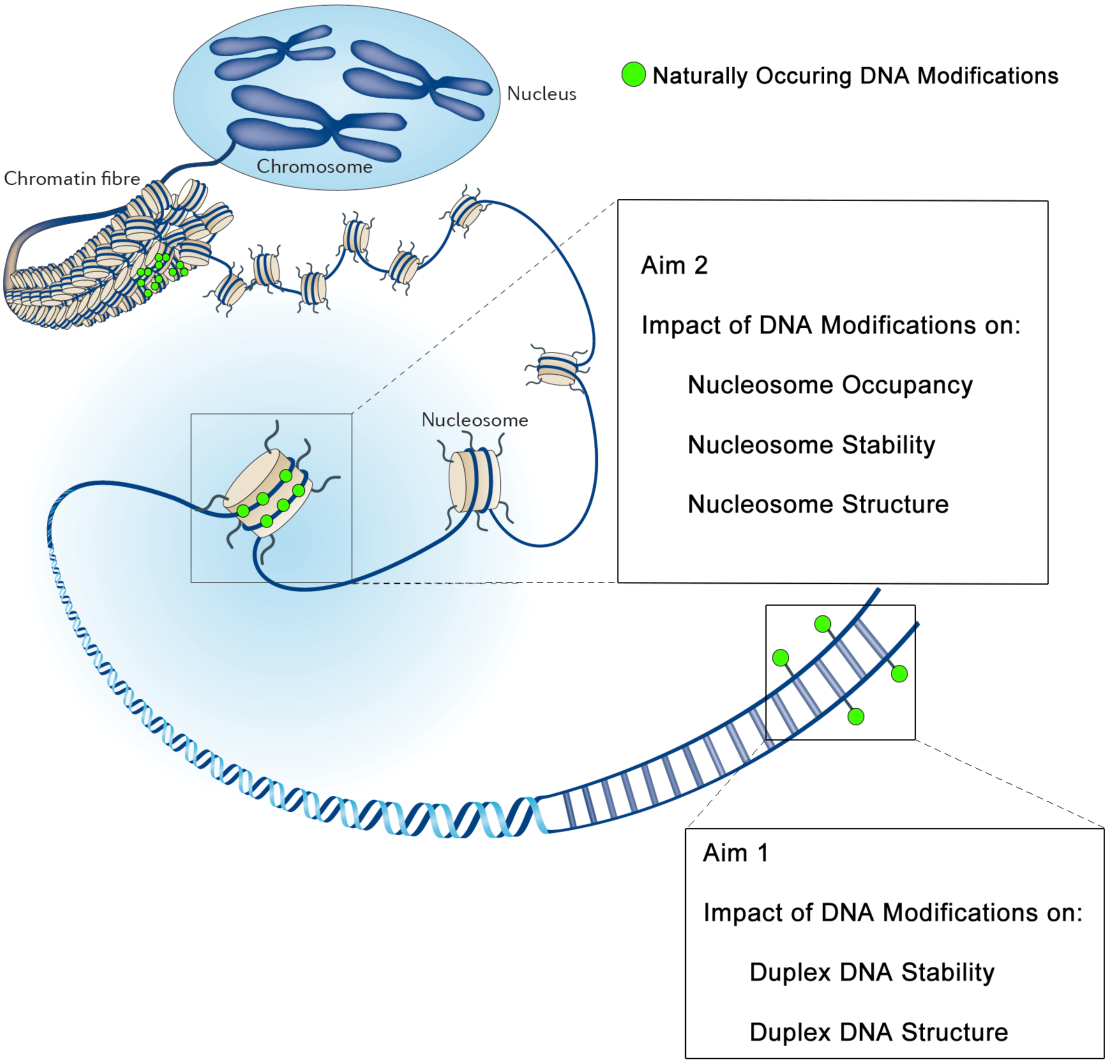


Figure 1-8. The aims of the DNA modification effect study. The figure is adapted from Arrowsmith *et al.*[59].

2 Effect of Naturally Occurring DNA Modifications on Duplex DNA Stability and Structure

Table of Contents

2.1 Background

Naturally occurring DNA modifications in mammalian DNA have been implicated in the regulation of gene expression. Considering the dynamic nature of DNA modifications throughout development and the implication of aberrant DNA modification pattern in disease development, it is important to elucidate the fundamental linkage between DNA modifications, protein recognition and downstream cellular processes by studying the effect of DNA modifications on DNA biophysical properties.

2.1.1 Effect of DNA Modifications on the Biophysical Properties of DNA

Owing to its important role in gene expression regulation and its relatively high natural abundance levels, the effect of mC on duplex DNA has been extensively studied by various biophysical techniques. Nathan *et al.* showed that the presence of mC modifications caused under-winding of duplex DNA from 10.5 to 11 base-pairs (bp) per turn[222]. Due to the contribution of hydrophobicity to the base stacking energy[223, 224], the mC-modified duplexes are more stable than the unmodified counterparts[37, 225]. It has also been reported that in very specific sequence context, mC modifications can make the duplex DNA structure convert from B-form to Z-form, E-form and A-form under different modification patterns and crystallization conditions[30, 36, 226, 227].

The effects of more recently discovered modifications hmC, fC and caC, on DNA structure and stability have been studied in various sequence contexts and modification patterns. hmC-modified sequences retained a B-form conformation like the unmodified sequences.[37, 228] The effect of hmC on thermal stability, however, varies considerably[37, 225, 229]. Studies showed that caC does not influence the general structure of duplex DNA, but the caC modification can be stabilizing or destabilizing depending on the sequences context.[37, 229, 230]

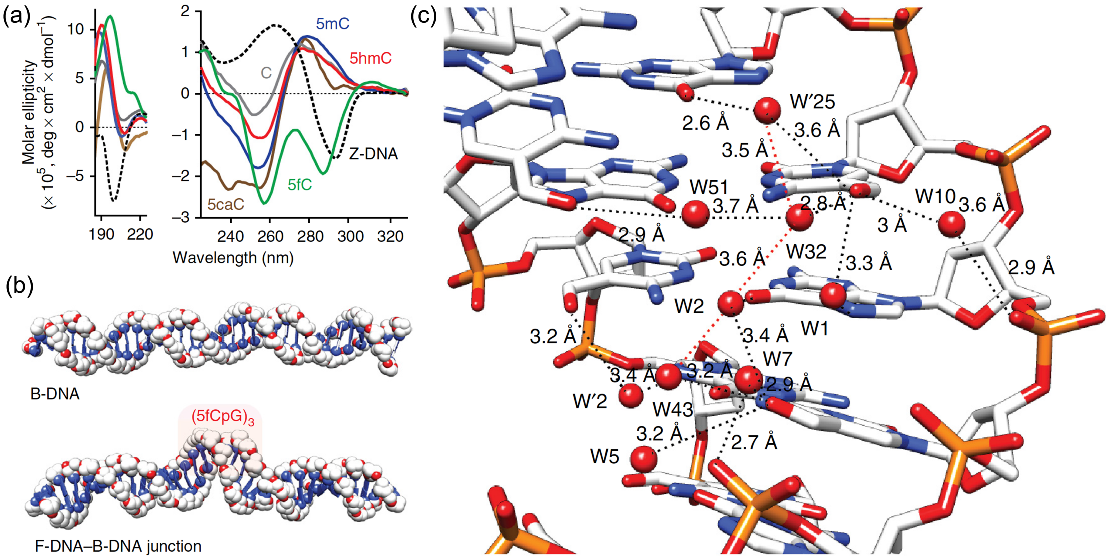


Figure 2-1. (a) CD spectra of C/mC/hmC/fC/caC-containing oligonucleotides in comparison with that of Z-DNA (dashed); (b) modeling of a 36-mer with B-DNA geometry (upper) and F-DNA geometry with flanking ideal B-form DNA helices (lower), demonstrating the alteration of the helical trajectory and local variation of the grooves induced by fC modifications; (c) H-bonding network (marked by dashed lines) facilitated by the formyl groups of fCs, O6 of guanines and water molecules, resulting in the unusual twist of the helix. Figures were adapted from Raiber *et al.*[37].

Thermodynamic studies indicate that fC can also be stabilizing or destabilizing depending on the sequence and modification density.[37, 229] With low density and non-consecutive fC modifications (one fC on the self-complementary Dickerson-Drew Dodecamer sequence, and 4 bp between the two fCpGs), the X-ray crystal structure indicated the DNA remained in the B-form[230]. Notably, Raiber *et al.* demonstrated that sequences containing three consecutive symmetrical (fC)pGs, with three fC modifications on each strand, displayed unique CD spectral characteristics, different from that of B-form DNA. The subsequent X-ray crystal structure determination confirmed a novel F-form structure at 1.4 Å resolution. The formyl groups formed an intricate hydrogen bonding network with the O6 of the neighbouring guanines and water molecules (Figure 2-1), resulting in a half-unwound structure compared to the canonical B-form structure, suggesting that fC may directly influence protein recognition through a critical DNA conformational change.[37] Later it was also reported that the same sequence formed an A-form structure at 2.3 Å resolution, however under different crystallization conditions.[231]

DNA modifications may also affect protein binding through changes in the stability of the DNA double helix. For example, Dai *et al.* have observed that the presence of fC and caC decreased DNA stability as measured by UV spectroscopy.[113] Their studies revealed that the electron-withdrawing –CHO and –COOH groups of fC and caC decreased electron density at N3 of the DNA base resulting in weakened hydrogen bonding for xC-G base-pairing thereby potentially facilitating protein recognition for downstream cellular activity. Indeed, important chromatin remodelers and transcription regulating proteins such as NuRD complex and FOXK1 have been identified to bind to fC specifically, but not to the other C modifications.[163]

The effects of fC on the biophysical properties of DNA raise the question if fU, another naturally occurring DNA modification containing an aldehyde group, could also induce changes to the DNA. To the best of my knowledge, there is only one X-ray study of short dsDNA containing two fU modifications, and the structures remain unaltered in the B-form[232]. Furthermore, other T derivatives known in eukaryotic cells remain largely unexplored. Thus a more systematic study is needed for fU and other T derivatives that interrogate the effects of modification densities and sequence contexts on the biophysical properties of the DNA, considering their potential biological relevance.

2.2 Results and Discussion

To investigate the impact of T-modifications on the DNA, we used three different oligonucleotide sequences. Previous literature and preliminary data obtained in our laboratory suggest that T modifications are found at high density in telomeric and intergenic regions of the *Trypanosomatid* genome and suggesting that these modifications tend to cluster.[209, 233, 234] Therefore, sequences with different modification contexts and densities were designed as follows (Table 2-1): a 10bp duplex (ODN1) containing one modification, a 12bp non-self-complementary duplex (ODN2) containing three modifications on one strand, and a 12bp self-complementary duplex (ODN3) containing three modifications on each strand. The higher modification intensity may amplify the influence of base modifications to distinguish from the background, and reduce the influence of sequence context. The sequences were designed to favour the formation of duplex DNA and hinder the formation of other secondary structures that could lead to a misinterpretation of the results. The sequences studied were synthesized commercially using phosphoramidite chemistry[235]: the 5fU phosphoramidite was chemically synthesized by Dr Fumiko Kawasaki from our group[236], while the others were obtained from commercial sources.

ODN 1 (H=U,T,hmU, fU)	number of H	10 mer
5'-ATCGCA**H**GTA-3'	1	forward strand
3'-TAGCGT**A**CAT-5'		reverse strand
ODN 2 (H=U,T,hmU, fU)	number of H	12mer non self complementary
5’-GAAC**H**G**H**C**H**GAG-3’	3	forward strand
3'-CTTG**A**C**A**G**A**CTC-5'		reverse strand
ODN 3 (H=U,T,hmU, fU)	number of H	12mer self complementary
5'-CG**HA**C**HA**G**HA**CG-3'	6	forward strand
3'-GC**AH**G**AH**C**AH**GC-5'		reverse strand

Table 2-1. Sequences used to study the effect of T derivatives on duplex DNA. H is either T, U, hmU or fU.

2.2.1 *Thermal stability of oligonucleotides containing T derivatives*

The effect of T derivatives on the base-pairing strength and base stacking of duplex DNA was assessed by UV thermal denaturing experiments, which rely on the change of UV absorbance in the process of dsDNA dissociating into ssDNA with elevated temperature. In dsDNA, π-π interactions from the base stacking influence the transition dipoles of the bases, and lower the UV absorbance. When dsDNA is dissociated into ssDNA under high temperature, the ordered base stacking is disrupted, resulting in increased UV absorbance (called hyperchromicity). The UV melting experiments were performed with buffer composition similar to cellular conditions, and were repeated for three cycles composed of heating and cooling processes. The UV data generated are used to calculate the melting temperature of dsDNA, defined as the temperature at which 50% of the dsDNA is dissociated as shown in Figure 2-2.[237-239]

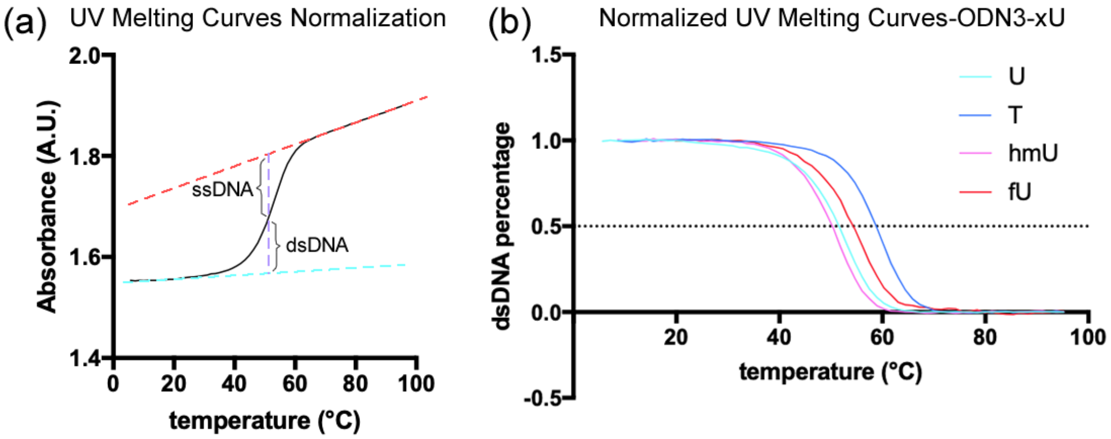


Figure 2-2. Illustration of the method (a) for normalizing the raw data obtained from UV melting experiments and the results (b). The baselines of melting curves were determined by treating the plateau at the lower temperature as 100% of the DNA strands being in duplex form, and the plateau at the high temperature as 0% of the DNA strands in duplex form (top). The melting temperature (Tm) of a given duplex, reflecting its thermal stability, was defined as the temperature where 50% of the DNA strands were in duplex form. Examples used were data obtained by UV melting experiments for ODN3-U for (a) and ODN-xU for (b).

As shown in Figure 2-3 (melting curves summarized in Figure 6-1 and Figure 6-2), the T- and U-containing duplexes showed comparable thermal stabilities in the ODN1 (1 modification) and ODN2 (3 modifications) sequence context. In the ODN1, T and U showed melting temperatures (Tm) of 44.1 ± 0.6°C and 44.8 ± 0.8°C (+0.7°C compared to T), respectively. In ODN2, T and U displayed Tm of 51.7 ± 0.2°C and 51.5 ± 0.1°C (-0.2°C compared to T), respectively. In the ODN3 (6 modifications) context, the Tm of T was 53.3 ± 0.7 °C, and 50.7 ± 0.1°C for U (-2.6°C compared to T). Overall, the results revealed that within the sequence context used for the study, U does not significantly change the DNA stability compared to T.

Notably, when the effect of hmU on dsDNA stability was measured, it was observed that the presence of hmU at a higher density significantly decreased the DNA melting temperature. ODN2 decreased the DNA denaturation temperature by 4.3°C (p values 0.0275, unpaired t-test with Welch's correction, two tailed) and ODN3 by 4.8°C (p values 0.0082, unpaired t-test with Welch's correction, two tailed) compared to T.

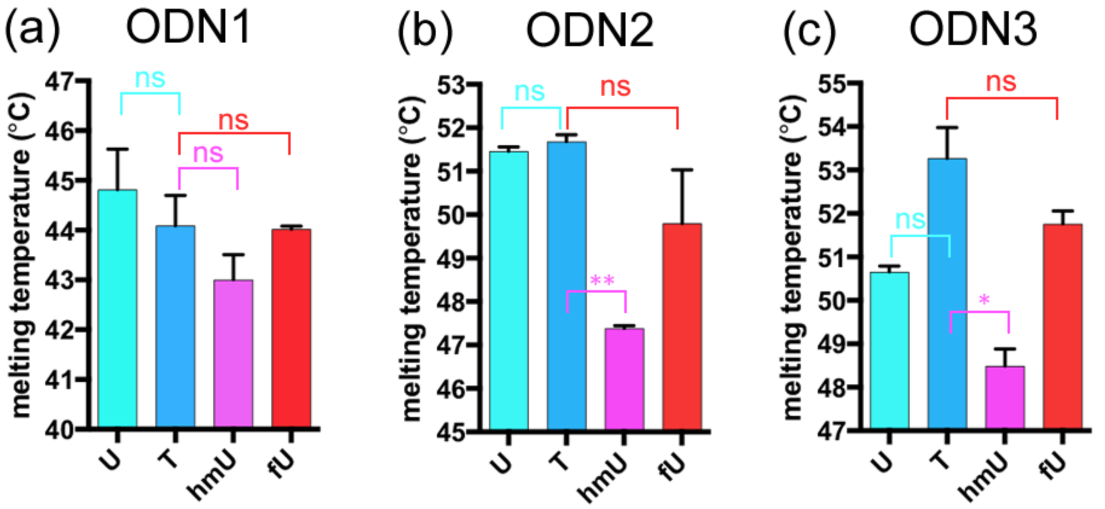


Figure 2-3. Summary of the results for the thermal stability study: UV melting temperature comparison of (a) ODN1, (b) ODN2 and (c) ODN3 sequences with T derivatives. The dsDNA concentration was 5 μM and the salt condition was 10 mM PBS buffer at pH 7.2 with 3 mM magnesium chloride. The melting experiments were performed in triplicate and the reported melting temperatures are the average of the three experiments, plotted as mean ± SD values. The significance of the data was analyzed by unpaired t-test with Welch's correction (two tailed), with p value represented in the New England Journal Medicine (NEJM) style, with 0.12 (ns), 0.033 (*), 0.002 (**). The raw data for the UV melting experiments are summarized in Figure 6-1, and normalized data for Tm extraction is summarized in Figure 6-2.

It was also observed that the thermal stabilities of the fU-containing duplexes were quite close to the unmodified counterparts. The Tm of fU was 44.0 ± 0.1°C (-0.1°C compared to T) for ODN1, 49.8 ± 1.2°C (-1.9°C compared to T) for ODN2, 51.8 ± 0.3°C (-1.5°C compared to T) for ODN3, respectively. Thus the fU modifications do not change the thermal stability of the dsDNA significantly within the sequence context and modification density studied.

Therefore the Tm increases in the order of hmU < T ≈ U ≈ fU. The lowered thermal stability by hmU modifications may be explained by the electron-donating nature of hydroxymethyl group lowering the acidity of N3-H, and therefore weakening the hydrogen-bonding for hmU:A. However, the result of comparable thermal stability amongst T, U and fU was unexpected. The electron-donating methyl group slightly decreases the acidity of N3-H for T (pKa 9.34 for U, and 10.04 for T, estimated using first principle quantum mechanics by Jang *et al.*[240]), and thus weakens the hydrogen bonding of T:A. The electron-withdrawing ability of –CHO group increases the acidity of N3-H (pKa 7.96 and 7.28 for fU (*trans* and *cis* conformation respectively), and 10.04 for T[240]), and therefore should strengthen the hydrogen bonding for base pairing for fU:A. Hence on this basis the order of base pairing strength should be T<U<fU. The comparable thermal stability may stem from other factors such as steric effect of C5 modifications, base stacking and the global DNA structure alteration.

The decreased thermal stability of hmU may contribute to the observed strong enhancement of transcription with bacterial RNA polymerases at some promoters[215] by facilitating the transcription bubble formation. In addition, it may contribute to the specific protein recognition by regulatory proteins such as Uhrf2 and chromatin remodellers Chd1 and 9 and further influence the cellular processes[130].

2.2.2 Structural characterization of oligonucleotides containing T derivatives

The structure of T derivatives containing duplexes was assessed with CD spectroscopy. CD spectroscopy takes advantage of the property that chiral molecules absorb right-handed and left-handed polarized light differently, and therefore detect the asymmetry of dsDNA. The position and amplitudes of the peaks are influenced by both the chromophore composition (DNA sequences) and the chirality posted by the DNA conformation.[241] Comparable CD spectra may indicate a similar secondary structure, and a change in the CD spectral signature may reflect a structure alteration. CD signature characteristics have been summarized empirically for A-, B-, F- and Z-form DNA secondary structures (Table 2-2)[37, 241, 242]. The CD spectra of A-form DNA typically contain a very deep negative band at around 210 nm, and a dominant and broad positive band at near 260 nm. In B-form DNA, the characteristic CD spectra have negative bands at around 210 and 245 nm, and positive bands at 220 nm and around 260-280 nm. It is noteworthy that the base pairs in the B-form DNA are perpendicular to the helix axis; this is not the case in the A-form. Thus the base pairs in the A-form DNA display significantly more chirality and the CD intensity of A-form DNA is much higher.[241] The F-form DNA shows CD characteristics of a positive peak at 195 nm, and negative bands at 260 and 290 nm.[37] The CD spectra of left-handed Z-form DNA tends to show positive peaks at 220 and 260 nm, and negative bands at 200 and 290 nm.[242]

DNA form	λ_{max} (nm)	
	positive peak	negative peak
A	260	210
B	220, long band or bands at 260-280	210, 245
F	195	260, 290
Z	220, 260	200, 290

Table 2-2. Summary of typical CD characteristics of A-, B-, F- and Z-form DNA.[37, 241, 242]

All sequences described in Section 2.2 were explored with CD spectroscopy to provide structural insights into the effect of different base modifications on DNA duplex at different modification density (1/3/6 modifications per 10 or 12 bp dsDNA).

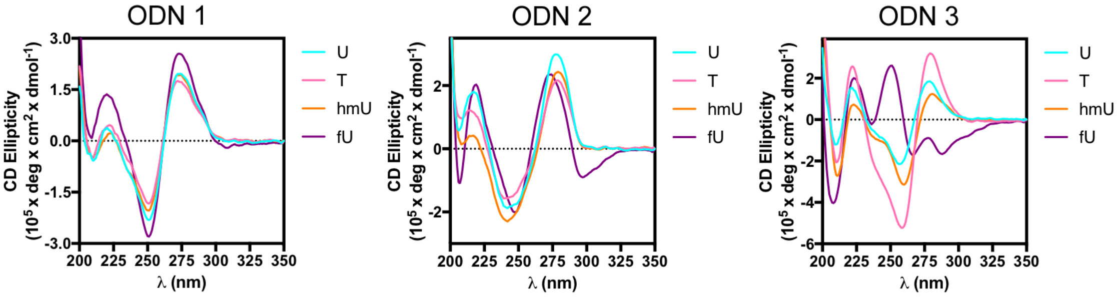
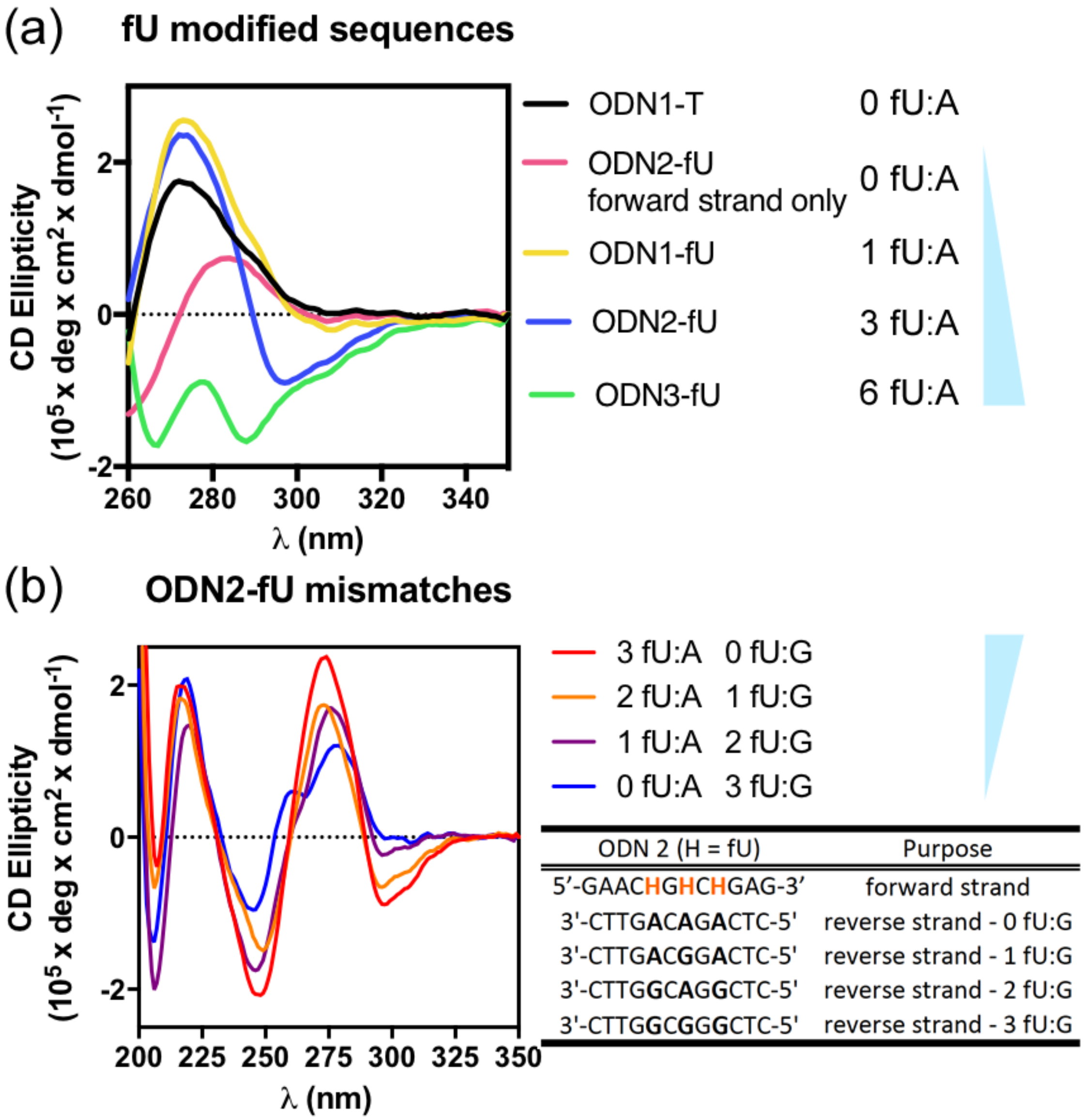


Figure 2-4. CD spectra of all sequences modified with different T derivatives. All samples were measured in 10 mM PBS buffer at pH 7.2. The samples were scanned in triplicate across the range of 200 nm to 350 nm, averaged and corrected using a buffer spectrum and absorbance at 320 nm to produce the final CD spectrum.

In the ODN1 sequence with one modification, all of the DNA investigated displayed a CD spectrum characteristic of the B-form of DNA, with negative bands around 210 and 250 nm and positive bands around 220 and 270 nm[241] (Figure 2-4). Similarly, in the ODN2 sequence with three modifications on one strand, all DNA showed typical B-form DNA CD spectra, with negative bands around 206 and 245 nm and positive bands around 215 and 245 nm[241] (Figure 2-4). The CD characteristics of ODN3 (six modifications, three on each side) were not as unanimous. DNA containing U, T or hmU still exhibited CD signatures of B-form DNA, with negative bands at around 210 nm and 255 nm (with a shoulder at lower wavelength), and positive bands at around 220 and 278 nm. Thus the CD spectra suggest U and hmU modifications do not alter the general structure of DNA in the sequence contexts and modification patterns studied. This result agrees with the finding of Delort *et al.*, who demonstrated by NMR spectroscopic measurements that U modifications within the sequence context d(GTACGXAC), X=T or U did not alter the global structure of the DNA[243].

In contrast, ODN3-fU displayed different CD spectroscopic characteristics compared to the other modifications, with negative bands at 207, 267 and 287 nm, and positive bands at 225 and 250 nm, which is not a typical B-form DNA CD signature. Interestingly, the CD signature was similar to that observed for the self-complementary fC-containing dodecamer (5'-CTA(fC)G(fC)G(fC)GTAG-3')[37], with two negative bands at around 260 nm and 290 nm, and positive CD ellipticity around 200 nm. However, the inversion of polarity at 205 - 260 nm between fC-containing dodecamer and ODN3-fU indicates that their respective structures may not be entirely the same. The CD signature of ODN3-fU also bore some resemblance to that of Z-DNA[242], a left-handed double helical structure, in that they both possess positive bands at 225 nm, and negative bands at 290 nm (the same negative band observed for F-DNA). Nevertheless, the Z-form DNA displays negative ellipticity at 200 nm, which is reverse to ODN3-fU; and ODN3-fU does not show the positive 260 nm band characteristic to Z-form DNA. Thus based on CD signature, ODN3-fU is less likely to be left-handed like Z-DNA. Due to the lack of theoretical and empirical evidence for determining the structure directly from the wavelength of CD ellipticity alteration, CD can only indicate the formation of a unique structure, while the exact structure of ODN3-fU has to be elucidated by techniques such as NMR spectroscopy and X-ray crystallography.



ODN 2 (H = fU)	Purpose
5’-GAACHGHCHGAG-3’	forward strand
3'-CTTG**A**C**A**G**A**CTC-5'	reverse strand - 0 fU:G
3'-CTTG**A**C**G**G**A**CTC-5'	reverse strand - 1 fU:G
3'-CTTG**G**C**A**G**G**CTC-5'	reverse strand - 2 fU:G
3'-CTTG**G**C**G**G**G**CTC-5'	reverse strand - 3 fU:G

Figure 2-5. (a) CD spectra of sequences with different 5fU modification pattern and density, ODN1-T is used as baseline of comparison; (b) ODN2-fU with different numbers of fU:G mismatches. All samples were measured in 10 mM PBS buffer at pH 7.2. The samples were scanned in triplicate across the range of 200 nm to 350 nm, averaged and corrected using a buffer spectrum and absorbance at 320 nm to produce the final CD spectrum.

It was noticed that CD spectra of all the fU-containing duplex DNA sequences showed a common negative ellipticity around 300 nm, but not in the spectra of the DNA containing other T modifications (Figure 2-5a). The negative band deepened and gradually shifted to a shorter wavelength with an increasing number of A-complemented fU bases (fU:A) in the duplex. The negative band could result from either fU itself, or from fU:A basepair. To identify which, CD spectra were recorded for the uncomplemented forward strand of ODN2-5fU ssDNA. The negative band was not observed, thus indicating that the negative band is not from fU itself. In addition, reverse strands of ODN2 were designed where the A complementing fU was replaced with G as shown in Figure 2-5b. As the number of fU:A basepairs replaced by fU:G increased, the negative ellipticity gradually diminished near 300 nm, suggesting the characteristic negative ellipticity as a special attribute of the fU:A base pair. This unique CD signature could indicate a local structural alteration near the fU:A base pair, which may further influence the binding and recognition by proteins.

2.3 Conclusion and Future Work

In this project, the effect of T-modifications on the thermal stability and structure of the DNA was investigated using three different sequence contexts. With hmU, although no structural change was observed, the modifications significantly reduced the thermal stability of modified DNA. Furthermore, it was observed that fU, compared to the unmodified dsDNA, slightly reduced the thermal stability although not significantly, however the CD analysis revealed characteristics that were distinct from that of B-form DNA. It is feasible that the formyl group of fU may induce structure change by modulating the hydrogen bonding network in a similar way to fC as seen in Raiber *et al.*[37]*.* More structural analysis by X-ray crystallography (preliminary work shown in Section 4.2.1) or NMR spectroscopy however will be needed to elucidate the high-resolution structure and fully understand the impact of fU on the DNA double helix structure.

Overall it was showed that hmU and fU within certain sequence context could both influence the biophysical properties of the DNA. Changes to the DNA stability or overall DNA structure may be relevant for DNA packaging and protein recognition, and thus affect biological processes in cells.

3 Effect of C Modifications on Nucleosome Occupancy and Stability

Table of Contents

3.1 Background

3.1.1 Function of Nucleosomes

In cells, the genomic DNA wraps around histone proteins to form the core unit of chromatin known as the nucleosome. Nucleosomes primarily serve to compact the genetic material into a higher order chromatin structure. Furthermore, nucleosomes control the temporary access of cellular machinery, such as DNA and RNA polymerases and transcription factors, to the genomic DNA for the relevant cellular function.[244-246]

Over the last decade, genome-wide studies by high throughput sequencing have provided us with a detailed map of the nucleosome organization in various organisms.[247-252] The use of a non-specific nuclease, the micrococcal nuclease (MNase), coupled with next generation sequencing (NGS) have enabled the identification of nucleosome positions with respect to the underlying genomic DNA sequences as they are protected from MNase digestion.[253-255] One major finding of these studies was that although the nucleosome positioning was globally variable, a subset of well-positioned nucleosomes was identified that was crucial for cellular activity. For example, the strongly positioned -1 (first nucleosome upstream of transcription start site (TSS)) and +1 nucleosomes (first nucleosome downstream of TSS) are believed to be important chromatin marks that help modulate RNA polymerase II dynamics.[256, 257] Since nucleosome positioning plays an important role in the regulation of gene expression, there has been a wide interest in understanding what determines the nucleosome organization in cells.

3.1.2 DNA Sequence as a Determinant of Nucleosome Organization

The DNA sequence itself has been demonstrated to be a determinant of nucleosome positioning.[245, 258-262] The ability to bend around the histone core and adopt the nucleosome structure greatly differs between DNA sequences.[263, 264] In 1998, Lowry *et al.* used a systematic evolution of ligands by exponential enrichment (SELEX) approach to understand the rules that govern the affinity of DNA sequences towards histone proteins. This positive selection was done by assembling nucleosomes starting with a pool of 5 x 10^{12} different randomly synthesized DNA sequences so that only 10% of the DNA was incorporated into nucleosomes. The selection was repeated 15 times, and selected DNA sequences were analysed bioinformatically to study common sequence features. A 10 bp periodicity for the dinucleotide TA/AA that favored histone-DNA interactions was observed. The highest-affinity DNA sequence identified from this study, known as the Widom 601 sequence, is now widely used in *in vitro* nucleosome studies because of this high affinity, and its ability to form homogenous nucleosomes[51, 258, 259, 265].

Sequence analysis of genome-wide nucleosome maps from chicken and yeast support the trend observed with the *in vitro* SELEX study using synthetic DNA[260, 261, 264]. A common 10 bp periodicity was identified by MNase sequencing (MNase-seq), with an enrichment of AA/AT/TA/TT dinucleotides occupying the minor groove facing towards the histone core, while CC/CG/GC/GG dinucleotides facing inwards in the major groove due to their preference to bend towards major groove[266, 267]. Since the minor grooves facing inwards to histone core need to be compressed to 3.0±0.55 Å, about half of the uncompressed width, TA dinucleotide and AT basepairs naturally enrich at these positions. This is due to their increased flexibility and endurance to local helix overwinding to enable the interaction with arginine residue inserted into the minor grooves via salt bridge.[259, 266, 268-270]

DNA sequences with the lowest histone affinity were shown to contain $T(G)_nA$ ($n \geq 1$) repeats by negative selection with SELEX experiments.[271] Several telomeric sequences with the $(G_nT_mA_{0\text{-}1})_x$ sequence motif have also shown low nucleosome formation propensity.[272] In addition, it has been noticed that the homopolymeric sequences poly(dA:dT) and poly(dG:dC) tracts were generally depleted of nucleosomes, due to their stiffness and alternative secondary structure formation, and are therefore difficult to wrap around histone proteins.[261, 273-277] It has been suggested that poly(dA:dT) is enriched in promoter regions of some organisms to keep promoters depleted of nucleosomes for protein machinery access and transcription initiation.[274, 277]

It is noteworthy that the accuracy of predicting nucleosome positioning based on these rules decreases from yeast to human[261, 278-281], indicating that with increased complexity of the genome, additional factors participate in regulating nucleosome positioning.

3.1.3 Cellular Machinery as a Determinant of Nucleosome Organization

Although the DNA sequence itself can be a predictor of nucleosome location, it cannot alone explain the nucleosome organization in cells.[282, 283] Particularly at regulatory regions that are tissue-specific or change throughout mammalian development, the nucleosome landscape is shaped by chromatin remodelers that control the access to the genetic information by moving, evicting or forming nucleosomes.[261, 274]

Early evidence supporting the role of cellular machineries in nucleosome positioning came from a functional evolutionary experiment[284], where a large portion of genomic DNA from a foreign species of yeast was introduced into *S.cerevisiae*, and the resultant nucleosome landscape was compared with the endogenous landscapes of both species. Since the two yeast species exhibit distinctly different nucleosome positioning, the contribution to nucleosome positioning due to DNA sequence context and cellular machineries can therefore be discerned. The resultant nucleosome profiling of foreign DNA displayed the characteristic nucleosome spacing of *S.cerevisiae* rather than the donor yeast species, therefore demonstrating the importance of cellular machineries in directing nucleosome positioning. Subsequent studies have identified a variety of protein machineries, including ATP-dependent chromatin remodelers, which actively reshape the chromatin landscape in accordance with cellular activities.[274, 283, 285, 286]

Chromatin remodelers use the energy from ATP hydrolysis to slide nucleosomes and influence nucleosome positioning *in vivo*.[285, 287] The loss of these chromatin remodelers has dire effects on cellular activities. For example, Whitehouse *et al.* has shown that the loss of Isw2 resulted in inappropriate transcription of both coding and noncoding areas, because Isw2 directs nucleosomes to position at vital positions for correct directionality and initiation sites for transcription.[288] Gkikopoulos *et al.* have demonstrated that the nucleosome landscape was altered significantly in the coding regions downstream of the +1 nucleosome in the absence of Isw1 and Chd1.[289] The disturbed positioning may be detrimental for the intricate cell system, as even a few base pairs shift in nucleosome positioning can change chromatin configurations[290] and protein interactions[291, 292], and further influence transcription[288] and DNA replication[293].

It is noteworthy that the *in vivo* studies investigating the role of DNA sequences and cellular machineries on nucleosome positioning have used extracted genomic DNA that already carry endogenous DNA modifications. Therefore it is important to understand how DNA modifications contribute to the regulation of the nucleosome landscape that is vital for cellular function.

3.1.4 Effect of xC on Nucleosome Positioning and Occupancy in Chromatin

The correlation between DNA modifications and genome-wide nucleosome occupancy has been studied in the context of mC, hmC and fC by comparing the genome-wide *in vivo* nucleosome footprint with DNA modification profiles in model systems such as *Arabidopsis thaliana* and mESCs.

Generally, nucleosomal DNA is linked to higher DNA methylation levels than flanking DNA in both *Arabidopsis thaliana* and human cells[294, 295]. Chodavarapu *et al.* have also shown that the methylation pattern displayed a 10 bp periodicity, coinciding with the number of base pairs in each helical turn of nucleosomal DNA. Therefore they proposed that DNA methyltransferases may preferentially target nucleosomal DNA[294].

Teif *et al.* have demonstrated that in mESCs, TET1 binding sites that had low levels of hmC (>25%) were slightly enriched with MNase-sensitive nucleosomes. Notably, higher density hmC sites (>50% or >90%) were associated with nucleosome depletion. Upon cell differentiation, the nucleosomes were depleted in mESCs but highly enriched in mouse embryonic fibroblasts cells.[296] The cell-type-dependent difference suggests that as differentiation progresses, hmC levels decrease, causing an increase in nucleosome occupancy.[296, 297]

The effect of fC on genome-wide nucleosome positioning and occupancy in embryonic mouse tissues was studied by Dr Eun-Ang Raiber (Balasubramanian group).[298] By comparing the genome-wide *in vivo* nucleosome footprint with the fC sites, it was observed that naturally existing nucleosomes tend to colocalize with fC peaks. This study also revealed that genomic regions, including CpG islands, which are generally depleted of nucleosomes showed increased nucleosome occupancy at fC-containing CpG islands. Furthermore it showed that fC contributed to the tissue-specific organization of nucleosomes. Collectively, the findings from this study suggest a role of fC in establishing distinct regulatory regions that control transcription.

3.1.5 Effects of DNA Modifications on DNA Flexibility

To understand the molecular basis for the correlation between DNA modifications and nucleosome occupancy/positioning, *in vitro* biophysical studies have been done. Due to the complexity of chromatin, researchers have used the core unit of chromatin, the nucleosome core particle, as an *in vitro* model system for investigating the effect of DNA modifications. The nucleosome is a 200 kDa disk-shaped molecule, formed by 147bp DNA wrapped left-handedly around histone proteins, with one side of the DNA facing towards the histone, interacting with the histone core and tails through hydrogen-bonding interactions and the electrostatic interactions[299]. Nucleosomal DNA is significantly deformed when wrapped around histone proteins, therefore the DNA flexibility plays an important role in the nucleosome formation and stability.[300] Therefore the impact of DNA modifications on the flexibility of the dsDNA was of particular interest. The effect of C derivatives on DNA flexibility has been studied mainly by DNA cyclization experiments (Figure 3-1a), which measure the time a strand of DNA needs for the complementary ends to anneal.

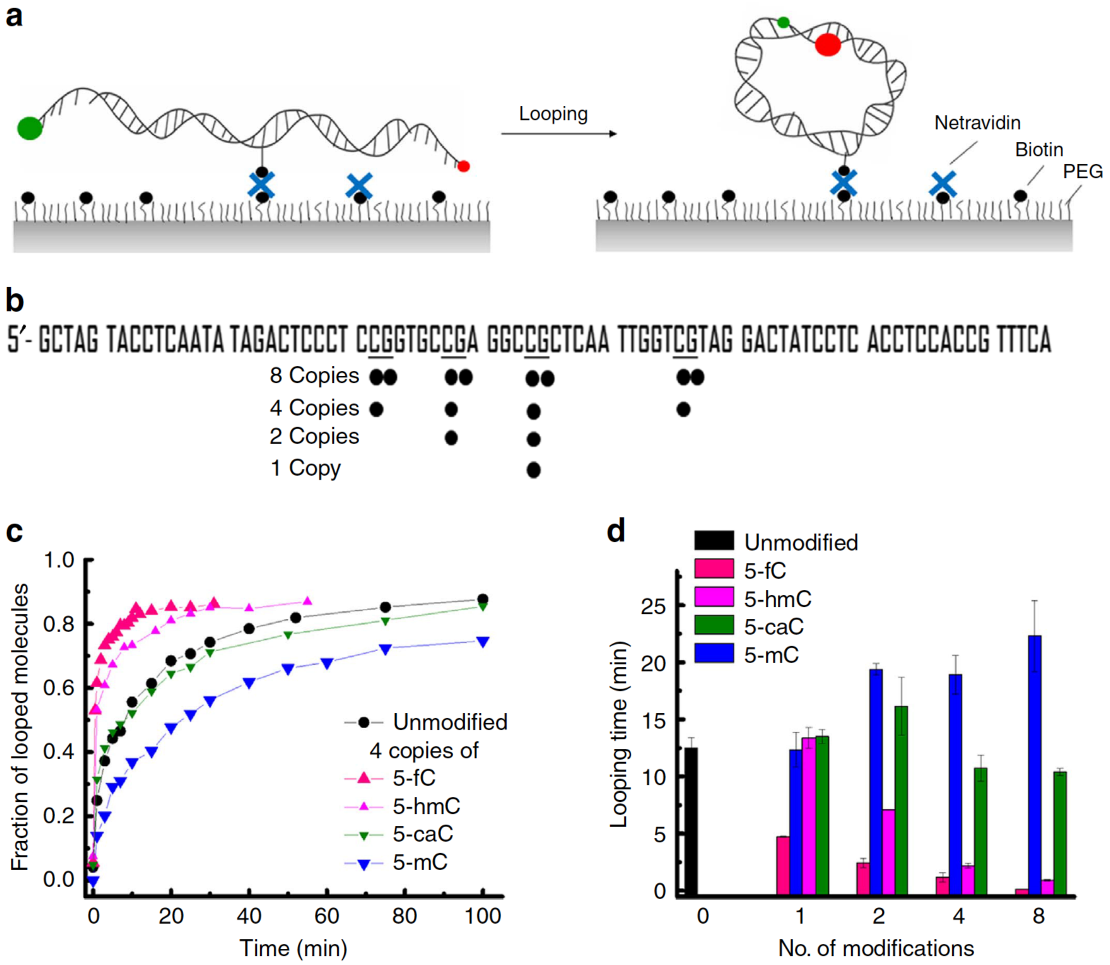


Figure 3-1. (a) schematic illustration of DNA cyclization assay; (b) DNA sequence and the modifications sites (indicated by black dots) investigated; (c) fraction of looped molecules as a function of time for DNA containing different xCs at four copies per dsDNA; (d) looping time for DNA containing different numbers and types of modifications. Figure was taken from Ngo *et al*.[111].

The cyclization times needed for DNA carrying xC modifications at different densities consistently showed that fC greatly increased the flexibility of the DNA strand compared to the unmodified counterpart (Figure 3-1). It is noteworthy that even a single fC modification was enough to make the DNA more flexible as compared to the unmodified DNA. Although not as effective as fC, multiple hmC modifications also made the DNA more flexible.[111] mC, on the other hand, rendered the DNA increasingly rigid as the number of modifications increased.[111, 222, 301]

3.1.6 Effects of DNA Modifications on Nucleosome Formation and Stability in vitro

The DNA sequence has been shown to influence the biophysical properties of the nucleosome, such as reducing nucleosome sliding on the DNA sequence, reducing nucleosome breathing (the transient opening of DNA ends) and therefore reducing the accessibility of nucleosomal DNA.[280] The DNA sequence alone can make the nucleosome stability vary over a thousand fold.[302] On top of the DNA sequence, DNA modifications add another layer of tuning nucleosome biophysical properties, such as compactness and stability.

Using Förster Resonance Energy Transfer (FRET), researchers observed that the two ends of nucleosomal DNA stayed in closer proximity upon methylation, indicating DNA methylation induces nucleosome compaction and may therefore contribute to a repressive chromatin structure.[303, 304]

The effect of DNA methylation on nucleosome stability has been studied in the context of the mC:G base pair location. As DNA faces histones through alternating major and minor grooves, the location of the mC:G has been classified based on the orientation of the groove. It has been observed that mC:G basepairs in minor groove positions destabilize the nucleosome more than those in major groove positions, both by FRET and by computation on different sequences.[305, 306] FRET experiments have also shown that methylation in central dyad positions does not influence nucleosome stability significantly.[305] The reduced tolerance of CpG methylation at both the major and minor grooves has been proposed to influence nucleosome positioning in the genome.[305-307] Through optical tweezers experiments which measure the forces needed to physically unwrap nucleosomes monitored with FRET, Ngo *et al.* have shown that mC modifications mechanically destabilize nucleosomes by assisting in the early unwrapping of the DNA termini but not the final unwrapping of the inner turn. They suggested that the destabilizing effect might be caused by the rigidity of the mC-containing DNA[111].

However, depending on sequence context and modification pattern, mC has also been observed to promote nucleosome formation.[294] To account for the influence of sequence context on nucleosome positioning and stability, a nucleosome reconstitution experiment was done with genomic DNA with or without DNA methylation by Collings *et al.* They discovered that upon methylation, normally unmethylated and nucleosome-depleted CpG island regions near the TSS were enriched with nucleosome.[308] They also noticed that the methylated CpGs preferentially located in minor grooves facing towards histone proteins[308], which agrees with the studies of Chodavarapu *et al.* in *Arabidopsis*[294] but contradicts the *in vitro* and *in silico* experiments mentioned above[305, 306]. This may suggest that the sequence context plays an important role in the geometric positional preference of mC and its influence on nucleosome stability.

Little is known about the effects of hmC and fC on the biophysical properties of nucleosomes *in vitro*. Mendonca *et al.* showed using salt titration experiments that hmC modifications increased overall nucleosome stability by increasing the affinity specifically towards the histone $(H3\text{-}H4)_2$ tetramer, but decreasing the interaction with the H2A-H2B dimer.[309]

Ngo *et al.* have shown by optical tweezers experiments that with only two fC modifications on one side of Widom DNA, the nucleosome's mechanical stability was increased compared to the unmodified counterpart. They suggested that this may be caused by the increase in flexibility of the fC-containing DNA[111].

To deepen the understanding of the phenomena observed in the complex genomic system with *in vivo* experiments, further systematic *in vitro* studies on the effect of naturally occurring DNA modifications on nucleosome occupancy and stability are needed, considering their potential importance in the regulation of cellular processes. DNA and nucleosomes containing various DNA modifications have demonstrated specific recruitment/exclusion of chromatin remodelers and transcription factors, suggesting the participation of DNA modifications in cellular activities such as transcription[163, 167].

3.2 Results and Discussion

For the biophysical studies, the Widom 601 sequence (detailed in section 3.1.2) and the Human α Satellite (HS) sequence were used (sequences listed in Table 5-1). Both sequences are commonly used for *in vitro* nucleosome studies since they are strong nucleosome positioning sequences and therefore form homogenous nucleosomes. Furthermore, both nucleosome structures have been both well-characterized by X-ray crystallography.[51, 267]

The HS sequence is a 146bp palindromic sequence derived from human α-satellite DNA[310]. The HS and Widom sequences are different in aspects such as GC% (39.7% for HS and 55.8% for Widom sequence), and the number of CpG sites (0 for HS and 30 for Widom). High GC% has been shown to favor the nucleosome formation *in vitro* but not *in vivo*, likely due to the repositioning effect of cellular machineries. The CpG content correlated positively with nucleosome occupancy at AT-rich promoters, but negatively with CpG rich promoters.[261, 276, 311, 312]

DNA modifications for this study were generally introduced by Polymerase Chain Reaction (PCR) using modified deoxynucleoside triphosphates (dNtps) (Figure 3-2). Because the primers used in the PCR were not modified, the modifiable cytosines were 70 out of 82 in both strands, and the modifiable thymines were 51 out of 65 in total.

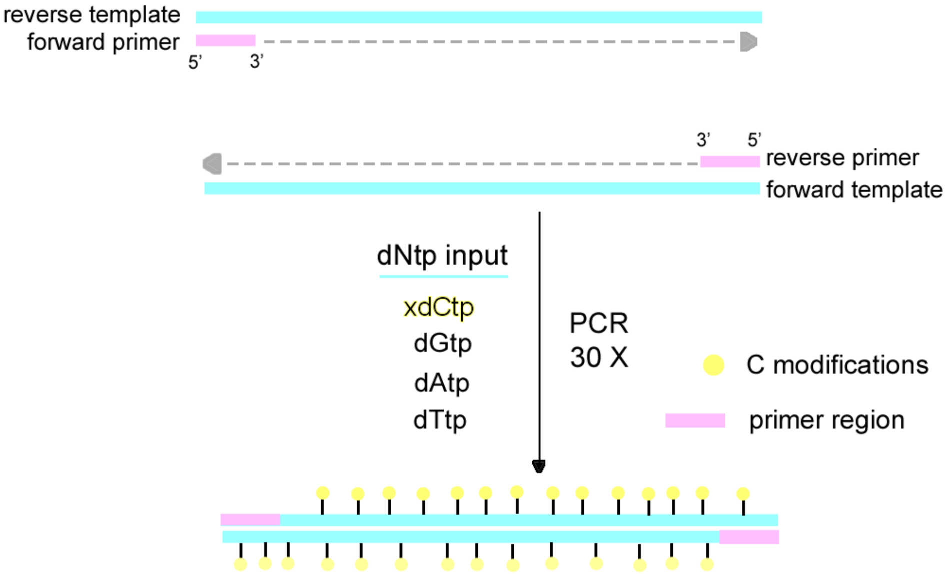


Figure 3-2. PCR schematics for generating fully xC- and xU-modified DNA (showed xC as an example); the pink primer regions do not have any modification.

3.2.1 In vitro Nucleosome Assembly

If the positively charged histone (from the protonated amino acids lysine and arginine) and negatively charged DNA (from the phosphate backbone) are directly mixed together, they tend to aggregate non-specifically and fall into a kinetic trap rather than forming the thermodynamically stable nucleosome product. Thus assisting factors are needed to gradually deposit DNA onto the histone to form the defined structure of the nucleosome.[313]

Traditionally the nucleosome is assembled by mixing histone proteins and DNA under a high salt concentration, which is enough to shield the non-specific interactions and avoid aggregate formation; the salt is then gradually removed by dialysis so that the DNA and histone can form the most thermodynamically stable nucleosome.[314] Besides inorganic salts, it has been reported that negatively charged organic molecules, such as RNA[315] and polyglutamic acid (PGA)[316, 317] can also assist in slowly depositing DNA onto the histone.

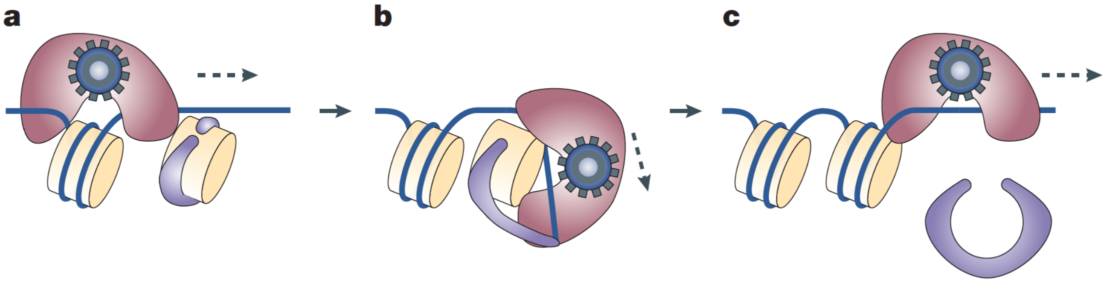


Figure 3-3. Model for ACF and NAP-1 assisted nucleosome assembly. ACF translocates along DNA and associates with a histone-NAP-1 complex, and then ACF dissociates histone-NAP-1 interactions while establishing histone-DNA interactions. Upon nucleosome formation, NAP-1 and ACF are dissociated from the nucleosome. Figure was taken from Haushalter *et al.*[318].

However, it was shown that nucleosome positioning using the above-mentioned methods was rather random, while incubation with cell extract produced regularly spaced nucleosomes.[319] Later ATP-utilizing chromatin assembly factor complex (ACF) was identified from a cellular extract to work in synergy with histone chaperones such as nucleosome assembly protein 1 (NAP-1) to deposit and regularly space nucleosomes *in vitro* by eliminating the non-nucleosomal interactions.[320-322] The multiple acidic amino acid patches allow NAP-1 to carry negative charges and interact with the positively charged histone.[323] ACF is composed of an Acf1 subunit and an ISWI subunit containing an ATPase domain that hydrolyzes ATP into ADP. The resultant energy is used in nucleosome assembly.[320] Deletion experiments have shown that the Acf1 subunit is also indispensible for the nucleosome assembly ability of ACF.[324] NAP-1 and ACF work together to load the DNA onto the $(H3\text{-}H4)_2$

tetramer[318], then the H2A-H2B dimer binds to the peripheral region of (H3-H4)$_2$ tetramer to form the nucleosome (Figure 3-3)[321]. The combination of ACF and NAP-1 has been shown to produce *in vivo* like nucleosome positioning, therefore we mainly used these biological assisting factors to asses the effect of xC on nucleosome occupancy.

3.2.2 Nucleosome Assembly of Modified DNA using Biological Assisting Factors

The nucleosomes were assembled with biological assisting factors as shown in Figure 3-4a. Nucleosomes are larger in size than free DNA, and the negative charge of nucleosomal DNA is partially balanced by the positive charge of the histone. As a result, the nucleosome migrates slower than free DNA on native gel, producing an upshifted band that runs around 400 bp compared to the DNA control as shown in Figure 3-4b.

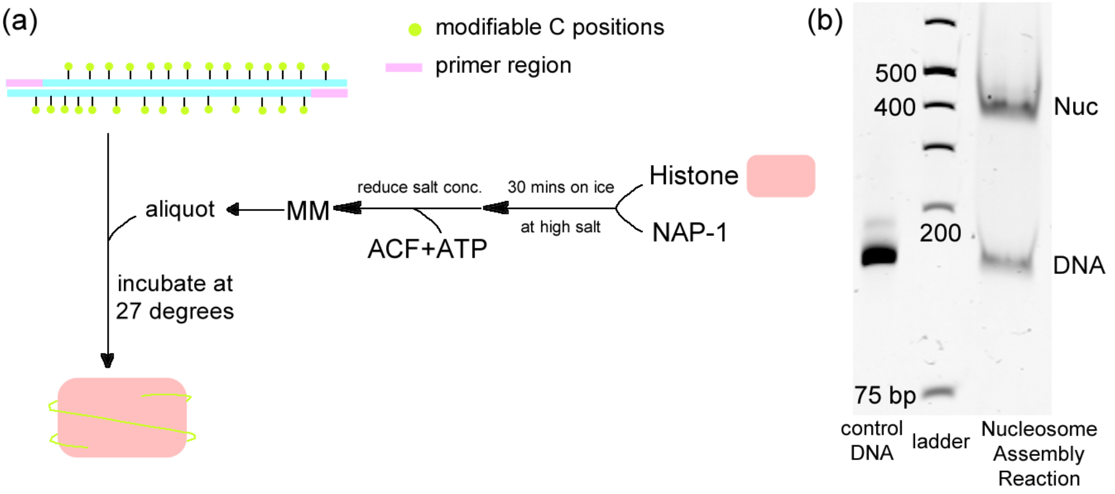


Figure 3-4. (a) Schematic illustration of nucleosome assembly experiments. MM stands for master mix; (b) Gel image of nucleosome assembly experiment with C-Widom DNA, control DNA and ladder.

To confirm that the higher band was the nucleosome fraction and not other protein-DNA aggregates (such as NAP-1 or ACF forming complex with DNA), the upshift band was excised and checked by proteomics, which identified all four histone subunits. Additional control experiments were performed, where all the components except histone proteins were mixed and incubated. No upshift band was observed, confirming again that the upshift band resulted from the nucleosome (Figure 3-5).

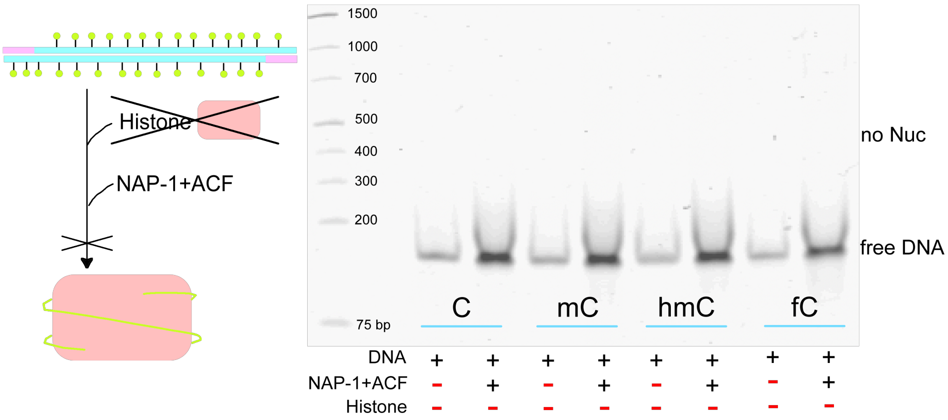


Figure 3-5. Schematic illustration and gel image for the control experiment setup with only NAP-1 and ACF but no histone proteins. There was only free DNA band with no upshift band for nucleosome.

3.2.3 Condition Optimization

In order to capture changes in nucleosome occupancy accurately, the incubation time and the histone:DNA ratio were optimized to obtain the nucleosome and free DNA fraction in equilibrium and at a similar intensity on the gel. This was undertaken to prevent the band of nucleosome or free DNA from being too faint, rendering the changes in nucleosome occupancy unquantifiable.

3.2.3.1 Incubation Time

Incubation time was modulated to ensure the nucleosome assembly reached equilibrium before evaluation. To do this, DNA and histone proteins were incubated at 27°C for three different time periods, namely 4 h, 15 h and 63 h. Figure 3-6 shows that after 4 h, the assembly was still not complete, while the assembly reaction extending beyond 15 h did not further increase the yield of nucleosome. Hence the effect of DNA modifications on nucleosome assembly was assessed at 15 h incubation and at 27°C.

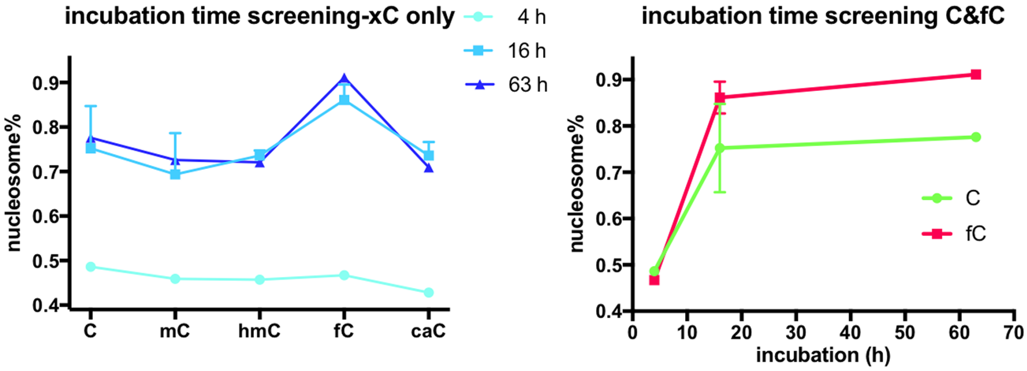


Figure 3-6. Plot of nucleosome occupancy with different incubation times in the nucleosome assembly experiment.

3.2.3.2 Histone:DNA ratio screen

Different ratios of histone to DNA concentrations were screened for condition optimization. The amount of histone proteins was fixed whereas the amount of the Widom DNA was varied to test histone to DNA ratio between 1:0.12 and 1:0.59 (equivalent to 40 to 200 ng DNA input per 15 µL reaction). The quantification results revealed that the ratio 1:0.35 (equivalent to 120 ng DNA input) produced a nucleosome occupancy of around 50% (Figure 3-7 upper panel).

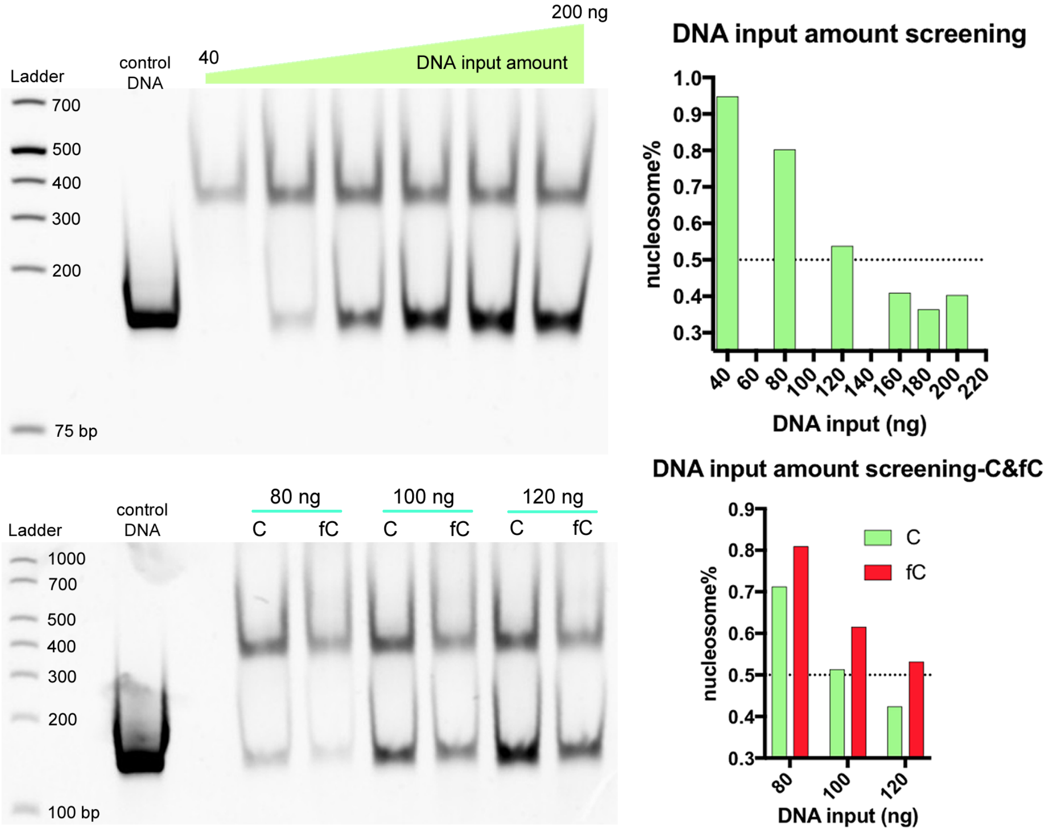


Figure 3-7. Gel images and plots of quantification for nucleosome assembly experiment with histone:DNA molecular ratio screening between 1:0.12 and 1:0.59 (with fixed MM concentration and C-Widom DNA input varying between 40 to 200 ng); and further screening with C- and fC-Widom DNA, with histone:DNA molecular ratio between 1:0.30 and 1:0.35 (with fixed MM concentration and DNA input varying between 80 to 120 ng).

Further screening with C-Widom and fC-Widom between ratios of 1:0.30 and 1:0.35 (equivalent to 80 ng to 120 ng DNA input) showed that 100 ng gave around 50% nucleosome occupancy for both reactions (Figure 3-7 lower panel), which is ideal for our study. Thus 100 ng DNA input was used to study the influence of DNA modifications on nucleosome assembly.

The results showing the optimal condition for nucleosome occupancy assessment are summarized in Table 3-1. Components were mixed in the following molar ratio: histone: DNA : NAP-1 : ACF = 1: 0.30 : 5.04 : 0.08 with DNA input of 100 ng, and incubation at 27°C for at least 15 hours.

Components	conc. (mg/mL)	volume (uL)	amount (ng)	MW (Da)	n (pmole)	ratio
histone	1.5	0.27	405.00	108768	3.72	1.00
DNA	/		100.00	90873	1.10	0.30
NAP-1	5	0.21	1050.00	56000	18.75	5.04
ACF	0.25	0.375	93.75	325000	0.29	0.08

Table 3-1. Relevant information of the components in the optimized condition of nucleosome assembly reaction. The molecular weight of NAP-1 and ACF was taken from Fyodorov *et al.* [322].

3.2.4 Effect of High Density xC Modifications on Nucleosome Occupancy

The optimized condition was used to assess the influence of DNA modifications on nucleosome occupancy. DNA fully modified with individual xC modifications was generated by PCR and assembled into nucleosomes as shown in Figure 3-4. Since the DNA-histone ratio is crucial for the efficiency of nucleosome assembly, slight changes in the amount of the input DNA may influence the observed nucleosome occupancy. To ensure equal amounts of input DNA for each assembly reaction, we used quantitative real-time PCR (qPCR) to quantify the modified DNA, since dense DNA modifications in the Widom sequence (70 modifications in 147 bp) may cause a change in the UV extinction coefficient at A_{260} and lead to misrepresentation of the DNA amount as measured by UV spectroscopy.

After nucleosome assembly, the reaction mixture was separated by native gel electrophoresis, and the gel was post-stained by GelRed^{TM} Nucleic Acid Gel Stain (Figure 6-3), which binds to dsDNA through both intercalation and electrostatic interaction. Upon binding, the dye exhibits a large fluorescence enhancement and thus allows quantitative detection of GelRed bound nucleic acid. The nucleosome band and DNA band were quantified to calculate the nucleosome occupancy and therefore determine the promoting/suppressing effect brought about by DNA modifications. Nucleosome occupancy (Nuc%) was calculated as the ratio:

$$Nuc\ \% = \frac{nucleosomal\ DNA}{nucleosomal\ DNA\ +\ free\ DNA} \times 100\%$$

Interestingly, the results (Figure 3-8) show that fC significantly increased the nucleosome occupancy compared to C (unpaired t-test with Welch's correction two tailed p value < 0.0001). Conversely, mC and caC modifications showed very similar nucleosome occupancy to C (unpaired t-test with Welch's correction two tailed p value = 0.9885 and 0.8399), whereas hmC modifications slightly decreased nucleosome occupancy (unpaired t-test with Welch's correction two tailed p value = 0.0151).

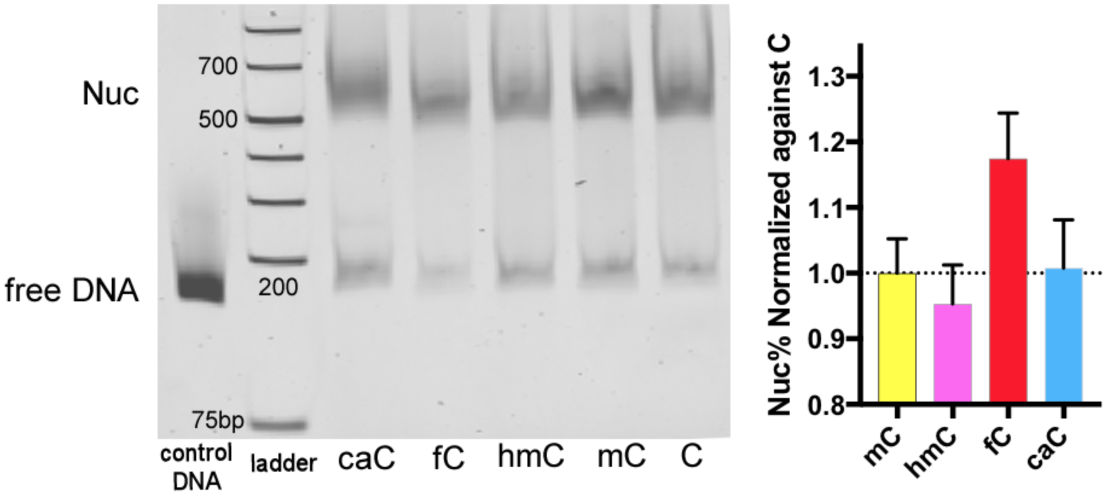


Figure 3-8. Gel image and plot of nucleosome occupancy with different modifications with respect to that of the unmodified DNA (ordinary one-way ANOVA test p value < 0.0001, each xC measurements were repeated for at least 10 times). The gels were imaged with either Typhoon or GBox with EtBr filter (excites at 532 nm and measure emission at 610 nm).

This result is consistent with earlier observations from genome-wide studies that linked fC to increased nucleosome occupancy in mouse tissues. This *in vitro* experiment provides direct evidence that the preference of fC-DNA for nucleosomes *in vivo* is a result of intrinsic fC-DNA-histone interaction. Since the formyl group of fC may either act as an additional acceptor for H-bond interactions with histone proteins, or form a covalent Schiff base with the primary amines of the histone tail (further investigated in Section 3.2.12), it may stabilize the DNA-histone interactions resulting in increased nucleosome occupancy.

3.2.4.1 Effect of T derivatives on Nucleosome Occupancy

The assessment of the influence of T derivatives on nucleosome occupancy revealed that fU greatly promotes nucleosome formation compared to the unmodified nucleosome control (unpaired t-test with Welch's correction two-tailed test p = 0.0333, Figure 3-9). In contrast, the presence of U decreased nucleosome occupancy (unpaired t-test with Welch's correction two-tailed test p = 0.0295). hmU-modified DNA exhibited similar nucleosome occupancy with unmodified DNA (unpaired t-test with Welch's correction two-tailed test p = 0.9192 > 0.05).

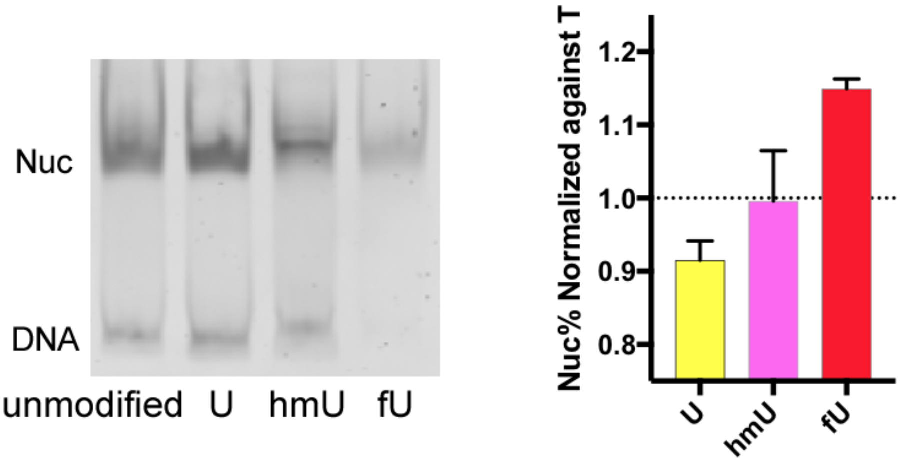


Figure 3-9. Schematics of nucleosome assembly, and gel image and result plot of nucleosome occupancy with different modifications with respect to that of the unmodified DNA (ordinary one-way ANOVA test p value < 0.0001). The gels were imaged with Typhoon with EtBr filter (excited at 532 nm and measured emission at 610 nm).

Considering the structural similarity between fC and fU, the mutual promoting effect could stem from the common formyl group. A previous study from our group showed that the formyl group of fU is more reactive than that of fC[325], suggesting a higher probability of Schiff base formation. Furthermore, structural changes to the DNA double helical structure introduced by symmetrical fU modifications (as discussed in Chapter 2) may also contribute to the promotion of nucleosome formation. DNA flexibility is another major factor influencing nucleosome formation, however we are unaware of any study on the DNA flexibility of fU-modified DNA. Our observation that U decreases nucleosome occupancy as compared to unmodified DNA suggests that the $-CH_3$ group (as present in T) positively impacts nucleosome occupancy either through additional interaction with the histone or potentially by increasing the flexibility of the DNA. The presence of $-CH_2OH$ group (as in the hmU-DNA) did not change the nucleosome occupancy as compared to the unmodified DNA, suggesting that at least in Widom sequence context hmU does not affect overall DNA-histone interactions.

As C derivatives are more abundant and the biological functions are better understood, we prioritized the investigation of the effect of C derivatives on nucleosomes for the rest of this study.

3.2.4.2 Nucleosome Formation with HS Sequence

It is possible that the sequence context of Widom DNA may stimulate the promoting effect of fC on nucleosome occupancy. Therefore, the effect of xC on nucleosome occupancy was assessed with an additional sequence, the HS sequence (detailed in Section 3.2).

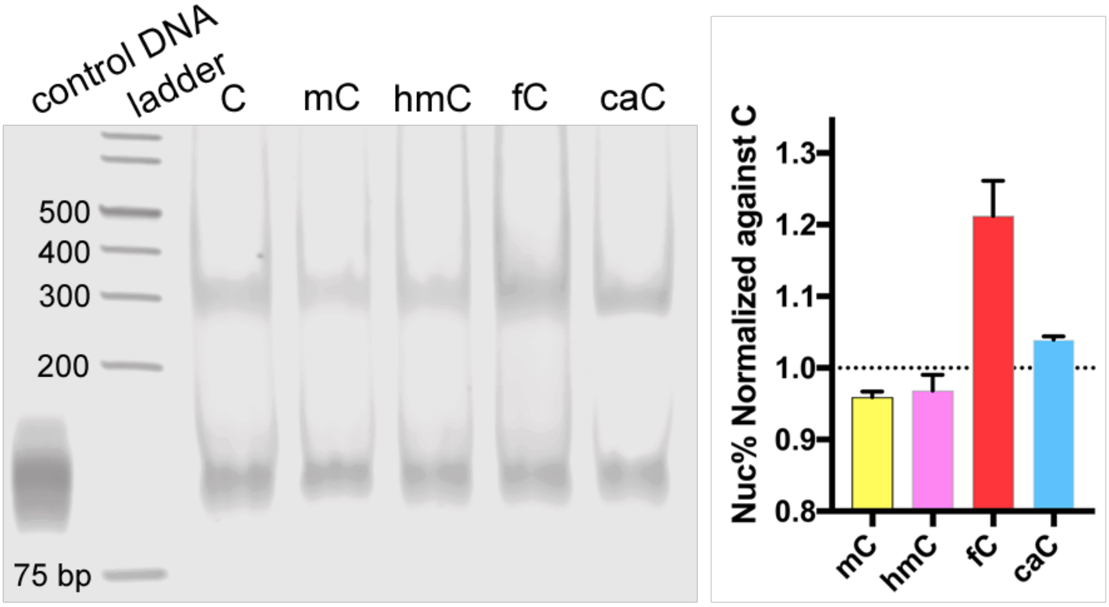


Figure 3-10. Gel image and quantification of nucleosome occupancy assessment with xC-HS normalized against the unmodified counterpart (ordinary one-way ANOVA test p value = 0.0007).

In this sequence context, fC demonstrated increased nucleosome occupancy (Figure 3-10, unpaired t-test with Welch's correction two-tailed test p = 0.1037). The promoting effect of fC agrees with our observation of Widom DNA, despite the drastic difference in GC% and number of CpG site, which have been shown to influence nucleosome formation in previous literature reports.[261, 276, 311, 312] Therefore the promoting effect of fC is to a certain extent independent of the sequence context.

On the other hand, the differences in nucleosome occupancy we observed with mC, hmC and caC modifications in the HS sequence compared to the Widom sequence suggested that the effect is rather dependent on sequence context and not modification specific.

3.2.4.3 Nucleosome Occupancy Measured by Single Fluorescent Labelling

During our initial experiments we observed that the intensity of the combined nucleosome and free DNA band were different depending on the DNA modification, suggesting that the staining of the gel by GelRed itself may be affected by the DNA modifications. To ensure that the changes in nucleosome occupancy we observed were genuine, Cy3- and Cy5-labelled DNA were used for nucleosome assembly and quantification for confirmation.

Fluorescence labels Cy3 and Cy5 (Figure 6-4) were chosen because of their high extinction coefficient, as well as the low non-specific interaction with biomolecules. Both fluorophores were used within the linear responding range between the concentration of fluorophores and the signal intensity.[326] This ensured that the intensity of fluorescence linearly reflected the amount of DNA present, as each strand of dsDNA carries only one label. The fluorescent labelling eliminates the need for post-staining, and therefore makes the values from different gels more comparable. The Cy3 and Cy5 were placed at the 5' end of primer sequence (forward strand and reverse strand, respectively, as shown in Figure 3-11) and introduced to the template by PCR.

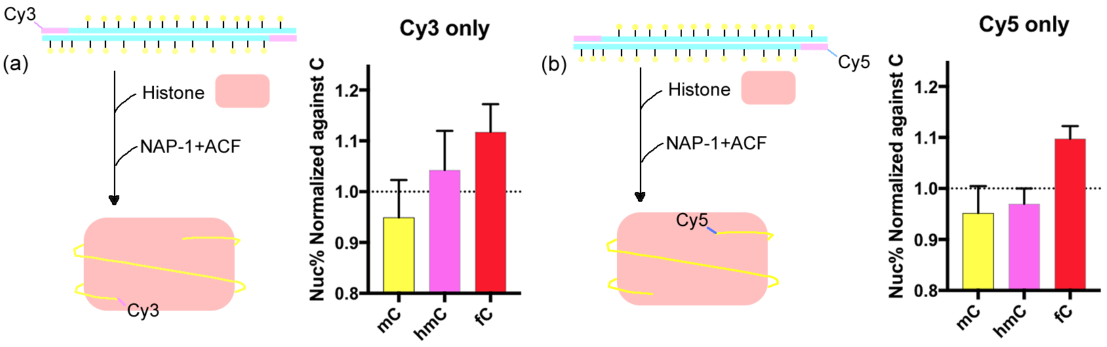


Figure 3-11. (a) nucleosome occupancy measured with Cy3 fluorescence excitation at 532 nm and emission measured at 580 nm. (b) nucleosome occupancy measured with Cy5 fluorescence excitation at 633 nm and emission measured at 670 nm.

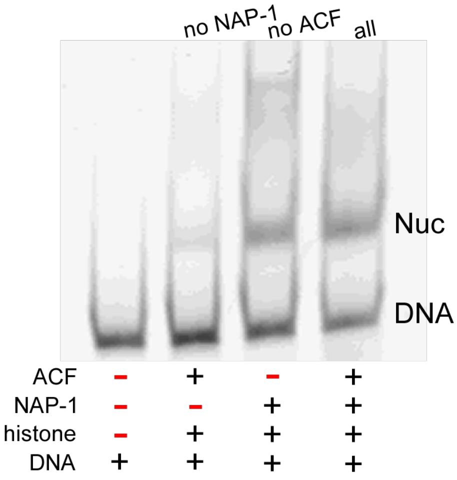
Results obtained by single fluorescent labelling imaging (Figure 3-11) confirmed the observation with post-staining imaging that fC significantly increased nucleosome occupancy compared to unmodified DNA (unpaired t-test with Welch's correction two-tailed test p value for Cy3 imaging 0.0002; Cy5 imaging < 0.0001). DNA modified with mC instead decreased nucleosome occupancy, while hmC showed comparable nucleosome occupancy with the unmodified DNA although not significantly for either imaging method.

3.2.4.4 Investigating the Interactions between fC and assisting factors

Our observation that certain DNA modifications change the nucleosome occupancy may be caused by the interaction between biological assisting factors with the DNA modifications rather than the intrinsic preference of nucleosomes for certain DNA modifications. To understand the molecular basis for the effect of DNA modifications on nucleosome occupancy, we separately investigated the roles of NAP-1 and ACF in nucleosome formation. The nucleosome occupancy was quantified after assembly using organic polymers (PGA) and inorganic salts (NaCl) as alternative chaperones to confirm the effect of DNA modifications observed in previous sections.

3.2.5 Nucleosome Assembly by NAP-1 or ACF

Nucleosomes were formed in the presence of either NAP-1 or ACF (Figure 3-12) to study whether these factors preferentially interact with certain DNA modifications. The nucleosome occupancy (Nuc%) ratio was calculated to assess the preference.



$$\text{Nuc\% ratio (no NAP-1/all)} = \frac{\text{Nuc\% (no NAP-1)}}{\text{Nuc\% (all)}}$$

$$\text{Nuc\% ratio (no ACF/all)} = \frac{\text{Nuc\% (no ACF)}}{\text{Nuc\% (all)}}$$

Figure 3-12. An example of the gel image of the nucleosome formation experiment with either NAP-1 or ACF, and both factors; and the equations to assess the preference of NAP-1 and ACF towards xC-modified DNA.

The Nuc% ratio was compared across different xC modifications (Figure 3-13). We observed that in the absence of NAP-1, the Nuc% ratio slightly decreased in the order C>mC>hmC>fC. When ACF was absent, the Nuc% ratio showed marginal decreased in the order C>fC>mC>hmC. Neither NAP-1 nor ACF demonstrate a significant preference towards fC-Widom DNA compared to the unmodified counterpart (unpaired t-test with Welch's correction two-tailed test p value for NAP-1 is 0.2665>0.05, for ACF is 0.1312>0.05). Therefore the biological assisting factors used for nucleosome assembly do not exhibit a significant effect on the promoting effect observed for fC.

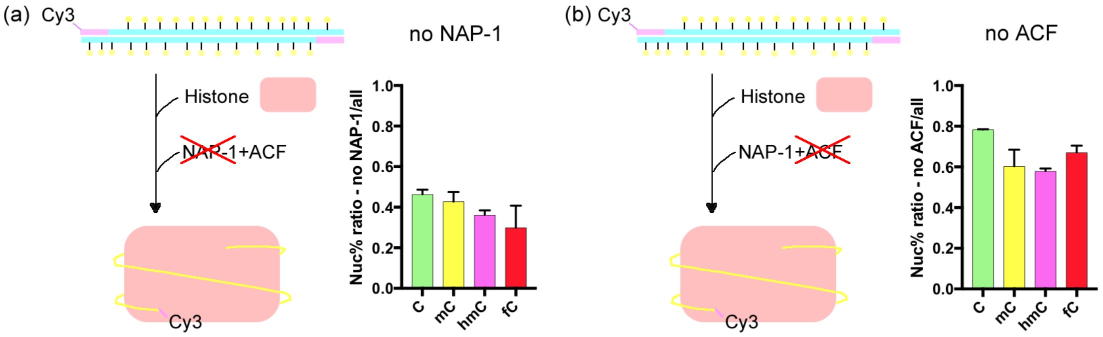


Figure 3-13. The schematic illustration of experimental set up and the quantified results for nucleosome assembly reaction without (a) NAP-1; (b) ACF.

To further confirm that the promoting effect of fC comes from the intrinsic changes that fC modifications post on nucleosomes, the nucleosome assembly experiment was repeated using PGA and NaCl as alternative chaperones.

3.2.6 *Nucleosome Assembly by PGA*

Acidic polypeptides such as PGA and polyaspartic acid have been reported to assemble histone and DNA into the nucleosome at physiological salt concentration (Figure 3-14).[316] PGA contains multiple carboxylate side chains, making it highly negatively charged and can thus interact with the positively charged histone. The PGA-histone complex prevents the nucleosome assembly reaction from falling into the kinetic trap by forming a non-specific aggregation. By gradually displacing PGA, DNA slowly wraps around the histone to form nucleosomes.[317] The histone interaction with acidic proteins and polypeptides has been suggested as a potential mechanism for cells to prevent the excessive histone from aggregating and keeping the histone available for chromatin formation.[316]

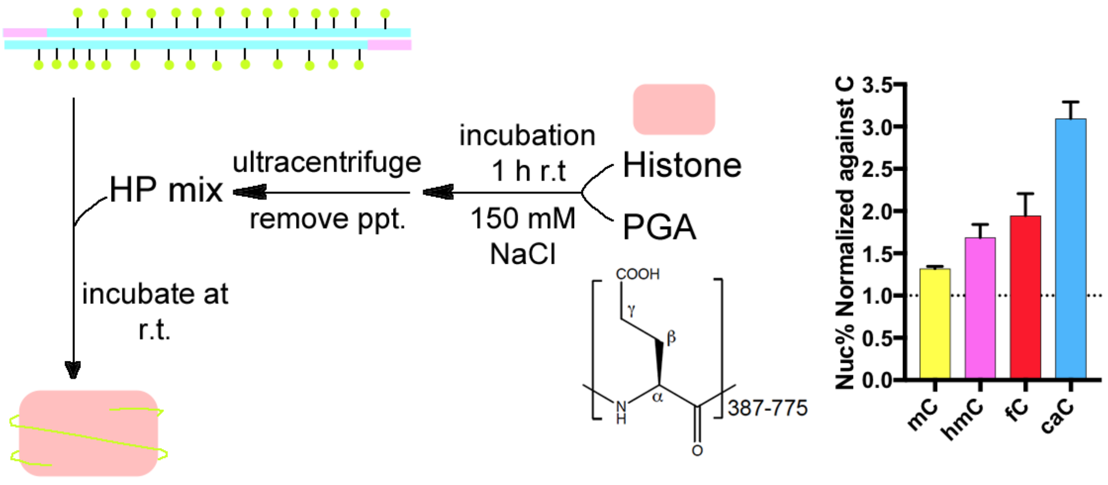


Figure 3-14. Nucleosome assembly with L-polyglutamic acid (PGA, MW 50 kDa – 100 kDa) assistance: left: the workflow of nucleosome assembly. Right: the nucleosome occupancy of xC-Widom DNA normalized against the unmodified counterpart, quantified from native gel electrophoresis experiments (ordinary one-way ANOVA test p value < 0.0001), imaged with either Typhoon with EtBr filter (excited at 532 nm and emission measured at 610 nm).

As shown in Figure 3-14, despite the different assisting factors used, fC demonstrated a similar promoting effect for nucleosome occupancy in PGA-assisted nucleosome assembly to that observed with biological assisting factors. Interestingly, in this method, caC exhibits increased nucleosome occupancy compared to unmodified DNA in the Widom DNA context, while in nucleosome assembly with biological assisting factors, caC showed a slightly repressed nucleosome occupancy. This is likely caused by the nature of this assembly method: nucleosomes were assembled by DNA competing with and replacing the negatively charged PGA molecule from histone-PGA complex. The caC-Widom DNA is significantly more negatively charged than other xC-modified DNA, because it has negative charges from both the phosphate backbone, as with other xC-Widom DNA, and the additional carboxylate groups (70/147 bp) from the base modification itself. Therefore caC-Widom DNA naturally competes more strongly for histones from the histone-PGA complex to form nucleosomes.

Since fC promotes nucleosome occupancy with the alternative chaperone PGA, this supports the conclusions, from Section 3.2.5, that fC contributes to nucleosome formation by a molecular mechanism independent of chaperones.

3.2.7 Competing Nucleosome Assembly

Having demonstrated that nucleosomes preferentially formed with fC-DNA in various assembly conditions, the ability of fC-DNA to increase the nucleosome occupancy in the presence of C-, mC- or hmC-DNA was then investigated. To do this, a competition assay for nucleosome formation was set up where the competing DNA was labelled with two different fluorophores, Cy3 and Cy5 (schematic Figure 3-15, excitation and emission profile Figure 6-5). By measuring the emission at different wavelengths, the nucleosome occupancy of individual competing DNA modifications can be accurately measured. The promoting/suppressing effect can be shown by nucleosome occupancy (Nuc%) ratio:

$$Nuc\%\ ratio\ (a/b) = \frac{nucleosome\%\ from\ channel\ a}{nucleosome\%\ from\ channel\ b}$$

To ensure that the position and type of fluorescence labelling did not interfere with the nucleosome formation, the nucleosome assembly was performed with both combinations of DNA modifications and fluorescence labels, e.g. the relative nucleosome occupancy of mC and fC was compared by both mC-Cy3 vs. fC-Cy5, and mC-Cy5 vs. fC-Cy3.

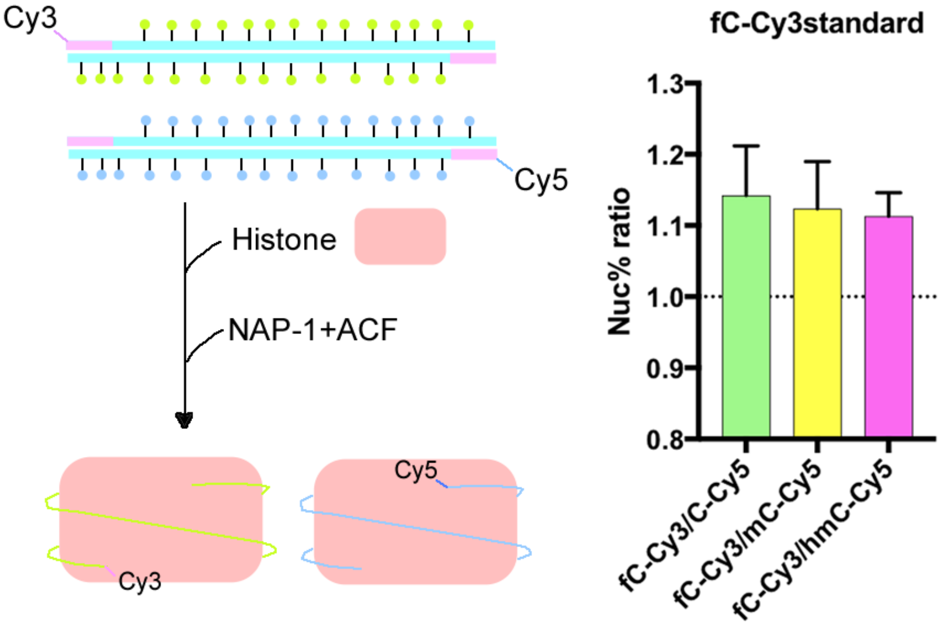


Figure 3-15. Schematic illustration for orthogonal fluorescence labelling experimental set up, and nucleosome occupancy measurement of DNA containing different xC modifications and orthogonal fluorescence labelling, (Cy3 fluorescence excited at 532 nm and measured emission at 580 nm, and Cy5 fluorescence excited at 633 nm and measured emission at 670 nm).

First, the robustness of the method was tested by comparing the nucleosome occupancy of C-Cy3 and C-Cy5. The Nuc% ratio obtained was very close to 1, confirming that the nucleosome occupancy from both channels can be well-represented.

Then the nucleosome occupancy amongst the C derivatives was compared. The ratios of nucleosome occupancy of fC-Widom-Cy3 DNA over C-, mC- and hmC-containing DNA labelled with Cy5 are all higher than 1 (Wilcoxon Signed Rank Test p value = 0.0156), therefore demonstrating that fC is the most occupancy-promoting DNA modification of all xC investigated in the Widom sequence context. This observation was confirmed by using other combinations of C-Cy5 vs. xC-Cy3 (x = m, hm and f) and fC-Cy5 vs. xC-Cy3 (x = unmodified, m and hm) labelled sequences.

This method compares the ratio of nucleosome occupancy for the two different modifications involved, and since the nucleosome occupancy is an equilibrium state that is independent of DNA input, this method provides the fairest comparison out of the three imaging methods detailed in this thesis. However, the number of samples is restricted by the number of orthogonal fluorescent labels used. In the current experiment set up, only two kinds of DNA can be compared in each experiment, and the number of experiments needed is exponentially related to number of samples, while the GelRed and single fluorescence methods are linearly related to the number of samples.

Overall, the results obtained by post-staining, single fluorescence quantification and orthogonal fluorescence competition point to the same conclusion that fC-modified DNA promotes nucleosome formation in the context of Widom and HS sequence compared to other C derivatives studied.

3.2.8 Effect of xC at Low Density on Nucleosome Formation

Although high modification density may amplify the influence of modifications from background noise, the fully modified DNA does not reflect the modification density observed in genomic DNA.[186, 327] Therefore, it was next investigated whether low-density modifications are sufficient to influence nucleosome occupancy. To do this, the PCR condition for preparing the DNA with the correct modification density was first optimized. The DNA modified with low-density xC was subsequently used to evaluate the impact of these modifications on nucleosome occupancy.

3.2.8.1 Relationship Between PCR Input and Incorporation

To generate DNA containing the desired density of modification by PCR, the relationship between xdCtp input and xC incorporation efficiency was first established by high-performance liquid chromatography (HPLC) (Figure 3-16).

First, Widom DNA with 100% xC modification was produced by PCR, and digested into nucleosides. The digested reaction mixture was resolved into individual nucleoside peaks by HPLC. The peaks were identified by synthetic standards and by mass. The peak area for all bases was integrated, and the peak ratio was used as the 100% incorporation standard.

$$peak\ ratio\ = \frac{peak\ area\ of\ xC}{peak\ area\ of\ A\ or\ G\ or\ T}$$

In the 100% xC input standard, a small amount of C nucleoside can still be seen at around 5 minutes (Figure 3-16) which originates from the primer region of the PCR product.

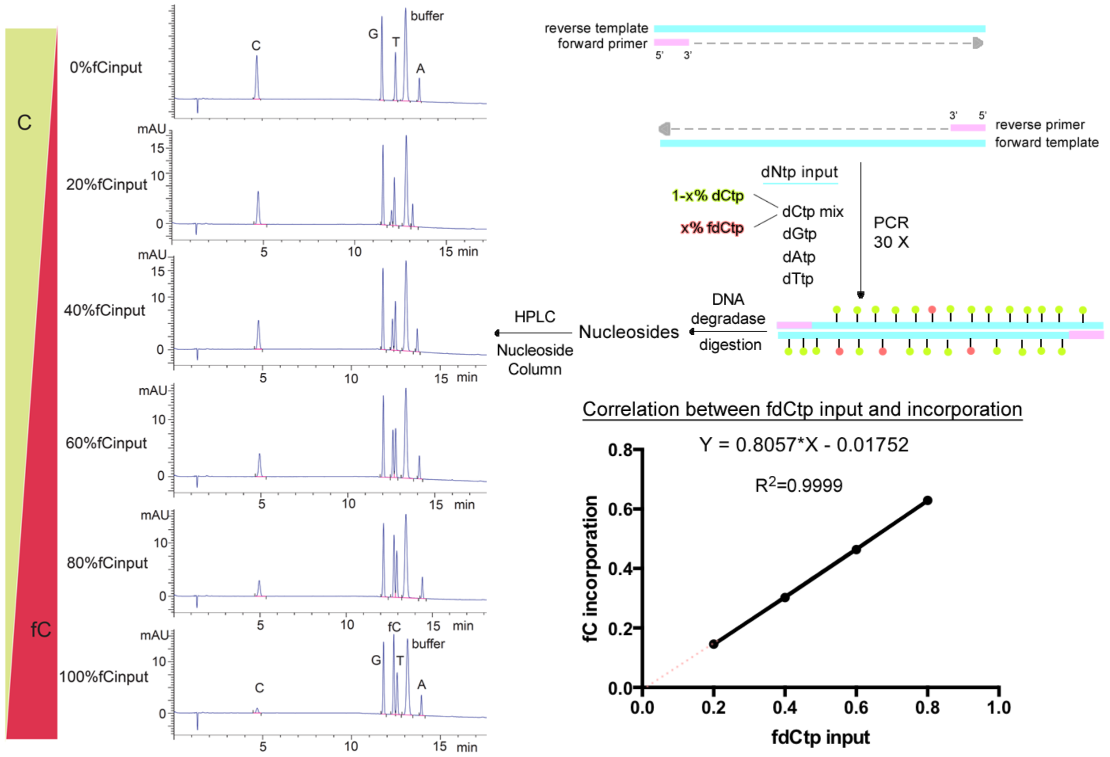


Figure 3-16. Schematic illustration for generating DNA with different densities of fC modification, and quantification of fC incorporation level by HPLC; the pink primer regions do not have any modification.

To obtain lower-density modified Widom DNA, we screened the xdCtp/dCtp ratio during the PCR step from 100% down to 0% xdCtp input percentage. The DNA was subsequently digested and the peaks were analysed by HPLC. Because all DNA strands were generated from the same template, the ratio between all C species to A/G/T remains constant. Therefore the incorporation percentage of xC can be deduced from the ratio between the peak ratios using the following:

$$incorporation\% = \frac{peak\ ratio\ (\%input)}{peak\ ratio\ (100\%\ input)} \times 100\%$$

The xC incorporation percentage was quantified through internally referencing to the corresponding peak area of A/G/T from the same run, thereby eliminating the risk of experimental error introduced by the spike-in reference method for quantification. To confirm that the modifications do not interfere with DNA digestion, the incorporation percentage of modified C was calculated with the peak area of A, G and T separately, and consistent results were obtained. In addition, the sum of incorporation percentages for all C species was very close to 100%, demonstrating that the digestion was indeed complete, and the quantification method was consistent and accurate.

When the xC incorporation level of the PCR product was plotted against the xdCtp input ratio, the incorporation level showed a linear relationship with the input ratio (Figure 3-16), allowing the generation of any density-modified DNA using the calculated DNA input.

3.2.8.2 Nucleosome Formation with DNA Containing Low-Density C Modifications

Using the relationship established between xdCtp input and incorporation, Widom DNA modified with 1% xC (x=m, hm and f) was generated (Figure 3-17), corresponding to around 1 modified C per dsDNA. Since 1% modification cannot be quantified by normal HPLC, LC-MS/MS with spike-in standards were used to confirm this low incorporation rate.

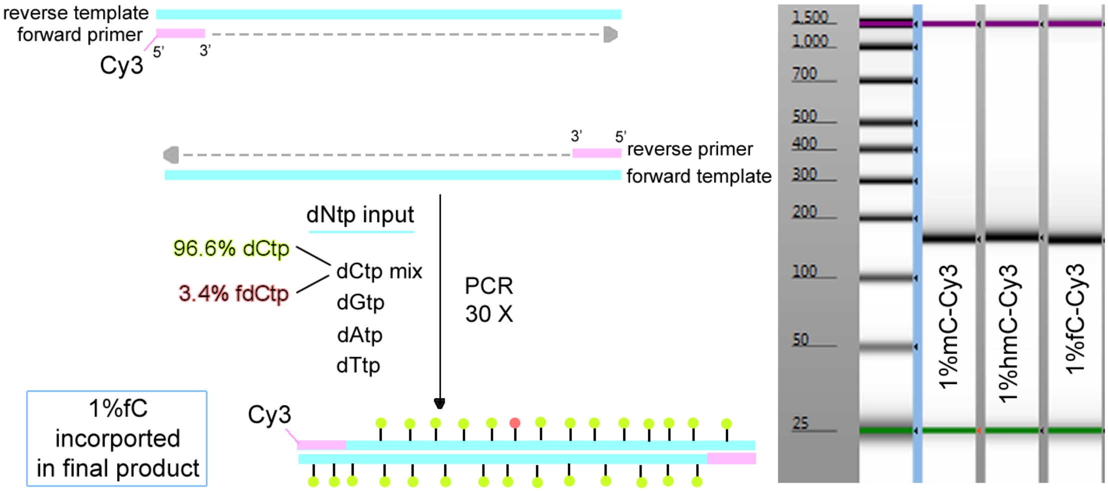


Figure 3-17. Schematic illustration for generating 1% incorporation xC-modified DNA; and the tapestation image examining the purity of product.

As shown in Figure 3-18, in order to compare the effect of modification density on the affinity of DNA to histone, DNA modified with low-density xC labelled with Cy3 was competed with corresponding fully modified DNA labelled with Cy5 to form the nucleosome. Surprisingly, comparable nucleosome occupancy was observed for low- and high-density modification for all the xC studied. This indicates that one xC modification per copy of DNA is enough to promote nucleosome formation to the same extent as full modified DNA.

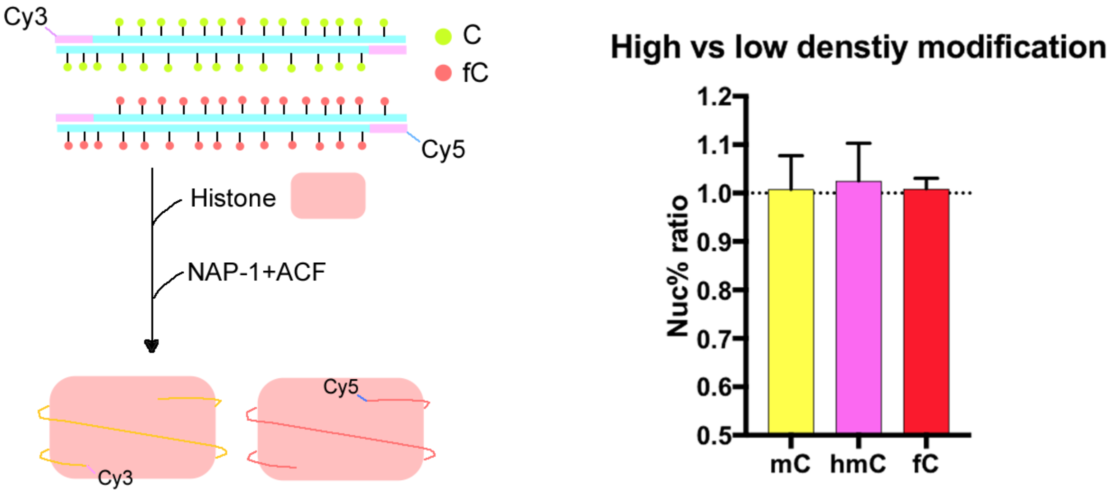


Figure 3-18. Nucleosome occupancy comparison between DNA containing low- and high-density modifications (Cy3 fluorescence excited at 532 nm and measured emission at 580 nm, and Cy5 fluorescence excited at 633 nm and measured emission at 670 nm).

To confirm that the modification densities between 1% and 100% also exhibit similar nucleosome promoting capacity, four intermediate modification densities (30%, 10%, 5% and 2%) of fC labelled with Cy3 were prepared and competed with C-Cy5 to form the nucleosome (Figure 3-19). The nucleosome occupancy ratio of low percentage fC-DNA over unmodified DNA was consistently around 1.1, which indicates that fC modification density ranging from 2% to 30% increased DNA affinity to histone compared to unmodified DNA with comparable potency. This result was confirmed by single fluorescence imaging and post-staining.

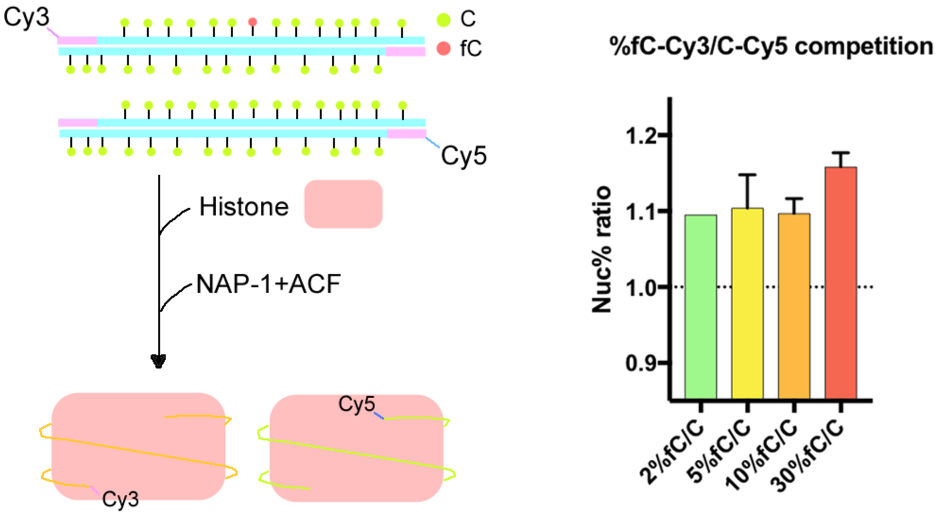
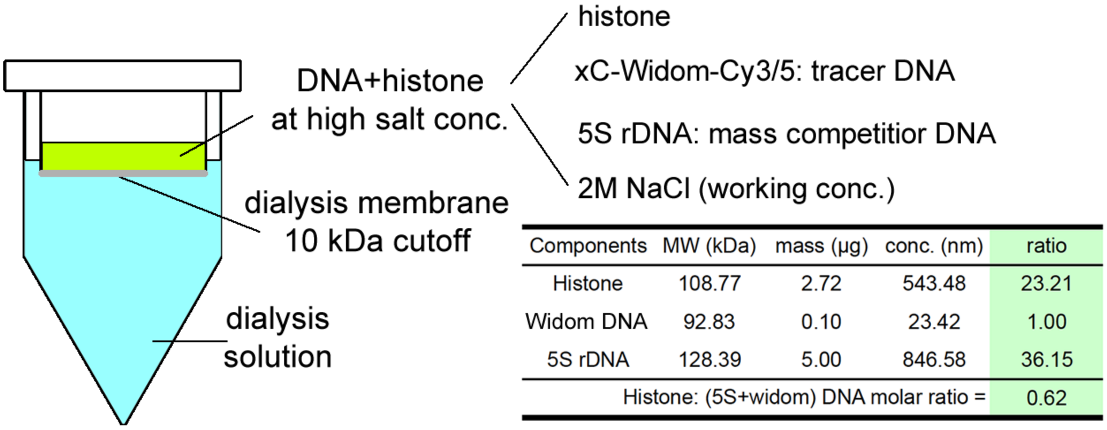


Figure 3-19. Example of nucleosome occupancy measurements for the DNA containing fC modification different at densities, measured by Orthogonal Fluorescence Competition method. Cy3 fluorescence excited at 532 nm and measured emission at 580 nm, and Cy5 fluorescence excited at 633 nm and measured emission at 670 nm.

In most genomic regions only small clusters of modified bases can be found, therefore it is interesting to observe that even a low abundance of modified bases can cause a change in DNA affinity for the histone, which suggests that naturally occurring levels of modifications may be enough to participate in directing nucleosome positioning and regulating downstream processes.

3.2.9 Effect of DNA Modifications on Nucleosome Stability

After demonstrating that the C modifications can influence nucleosome formation, the relative stability of different xC-modified nucleosomes compared to the unmodified counterparts was evaluated by measuring the relative affinities of histone-DNA interactions in nucleosome in terms of free energy change. The experimental setup followed a design by Thåström *et al.*.[302] Fluorescently labelled tracer DNA containing the DNA modifications of interest competed with a large excess of unlabelled competitor DNA to form nucleosomes with a limited amount of histone (Figure 3-20). The reaction was set up at 2 M NaCl, sufficient to shield all DNA-histone interactions to ensure robust competition.[302, 328] The shielding effect of high salt concentration was gradually decreased through dialysis until the final salt concentration reached 0.25 M (Figure 3-21).



Components	MW (kDa)	mass (µg)	conc. (nm)	ratio
Histone	108.77	2.72	543.48	23.21
Widom DNA	92.83	0.10	23.42	1.00
5S rDNA	128.39	5.00	846.58	36.15
Histone: (5S+widom) DNA molar ratio =				0.62

Figure 3-20. The experimental set up for salt dialysis method to measure the nucleosome stability.

The reaction mixture was resolved by native gel electrophoresis and the nucleosome and free DNA fraction of the tracer DNA was quantified and used for calculation of the relative stability.

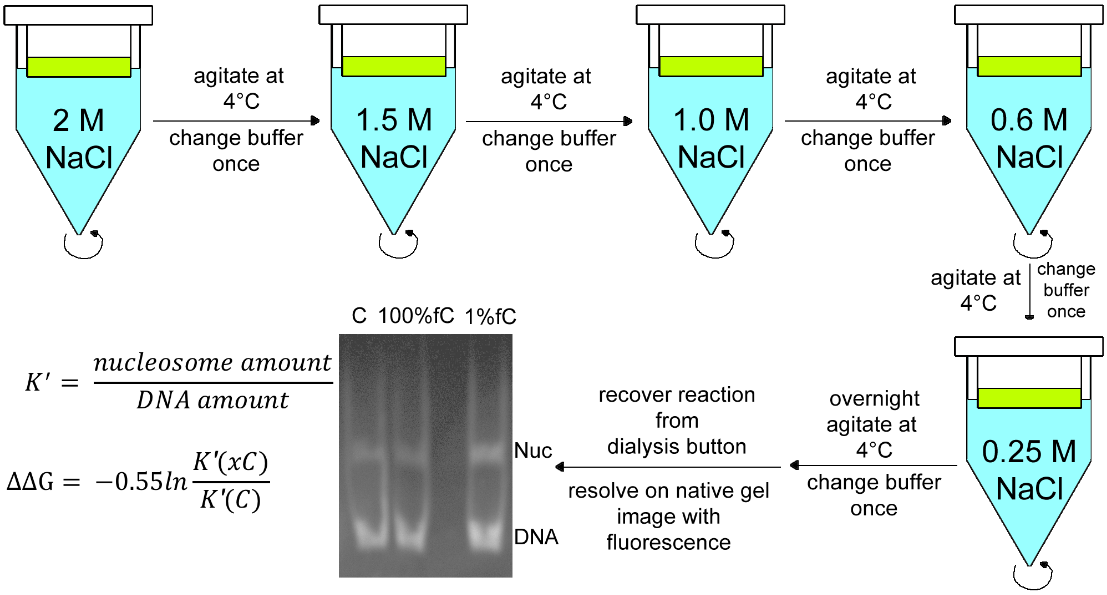


Figure 3-21. Schematic illustration of the salt dialysis experiment setup.

The reactions were set up from the same master mix containing everything but tracer DNA to ensure a fair competition. By normalizing the quantification results obtained for the unmodified control nucleosome measured in the same batch, the batch difference was also accounted for. Tracer DNA with Cy3 or Cy5 labelling at either 5' end was used to confirm that the fluorescence labelling and its position did not have any effect on the stability.

3.2.9.1 Calculation of Relative Stability of Nucleosome

The relative nucleosome stability between modified and unmodified nucleosomes was calculated as follows:

The equilibrium constant of the nucleosome assembly is defined as:

$$\text{DNA} + \text{Histone} \rightleftharpoons \text{Nucleosome} \qquad Keq = \frac{[nucleosome]}{[DNA][histone]}$$

As the volumes (V) were the same for nucleosome and DNA:

$$Keq = \frac{[nucleosome]\cdot V}{[DNA]\cdot V\,[histone]} = \frac{nucleosome\ amount}{DNA\ amount\cdot[histone]}$$

The Gibbs free energy can be calculated for the equilibrium as:

$$\Delta G = -\mathrm{RTlnKeq} = -\mathrm{RTln}\frac{nucleosome\ amount}{DNA\ amount\ \cdot[histone]}$$

$$= -\mathrm{RTln}\frac{nucleosome\ amount}{DNA\ amount}\cdot\frac{1}{[histone]}$$

$$= -\mathrm{RTln}\frac{nucleosome\ amount}{DNA\ amount} - RTln\frac{1}{[histone]}$$

To simplify the equation, K' was defined as

$$K' = \frac{nucleosome\ amount}{DNA\ amount}$$

The relative stability (ΔΔG) of the modified tracer DNA compared to the unmodified tracer DNA is

$$\begin{aligned}\Delta\Delta\mathrm{G} &= \Delta\mathrm{G}(\mathrm{xC}) - \Delta\mathrm{G}(\mathrm{C}) \\ &= -RTln\mathrm{K}'(\mathrm{xC}) - \mathrm{RTln}\frac{1}{[xC\ histone]} \\ &- \left(-RTln\mathrm{K}'(\mathrm{C}) - \mathrm{RTln}\frac{1}{[C\ histone]}\right) \\ &= -RT[lnK'(xC) - lnK'(C)] - RT(ln\frac{1}{[xC\ histone]} \\ &- ln\frac{1}{[C\ histone]})\end{aligned}$$

Since the competitor DNA was present in large excess compared to the tracer DNA (molar ratio 36.15:1), and the experiments from the same batch were set up from the same master mix, the final histone concentration in different nucleosome reactions were equal, and the equation was simplified to be:

$$\Delta\Delta\mathrm{G} = -RTln\frac{K'(xC)}{K'(C)}$$

Since

R = gas constant = 1.987×10^{-3} kcal·mole^{-1}K^{-1}

T = 4°C = 277.15 K

The ΔΔG calculation was further simplified to (with unit of kcal/mole):

$$\Delta\Delta\mathrm{G} = -0.55ln\frac{K'(xC)}{K'(C)}$$

For K' calculation, the DNA amount and nucleosome amount was quantified by the fluorescence intensity of the labelling. Since each DNA strand was labelled with one fluorescent labelling, the intensity of fluorescence linearly relates to the number of strands of DNA.

3.2.9.2 Nucleosome Stability Measurement

As shown in Figure 3-22, the negative ΔΔG value revealed that the nucleosomes containing low- and high-density mC, hmC and fC modifications were all more stable than unmodified nucleosomes with the exception of 100%mC. In the C modifications, fC is the most stabilizing modification. 1%fC-modified nucleosome is 0.164 kcal/mole more stable than the unmodified Widom nucleosome (Wilcoxon Signed Rank Test p value = 0.0005). Our observation on the relative stability of low-density fC-Widom DNA is consistent with previous reports that two copies of fC modifications within the Widom sequence were sufficient to increase the nucleosome mechanical stability.[111] Within each C modification, the 100% modified nucleosome is always slightly less stable than the 1% modified counterpart (unpaired t-test with Welch's correction: p value = 0.7553, 0.6951 and 0.0217 for mC, hmC and fC, respectively). Similar to fC, fU modified nucleosome has demonstrated high stability compared to unmodified nucleosome (-1.123 kcal/mole).

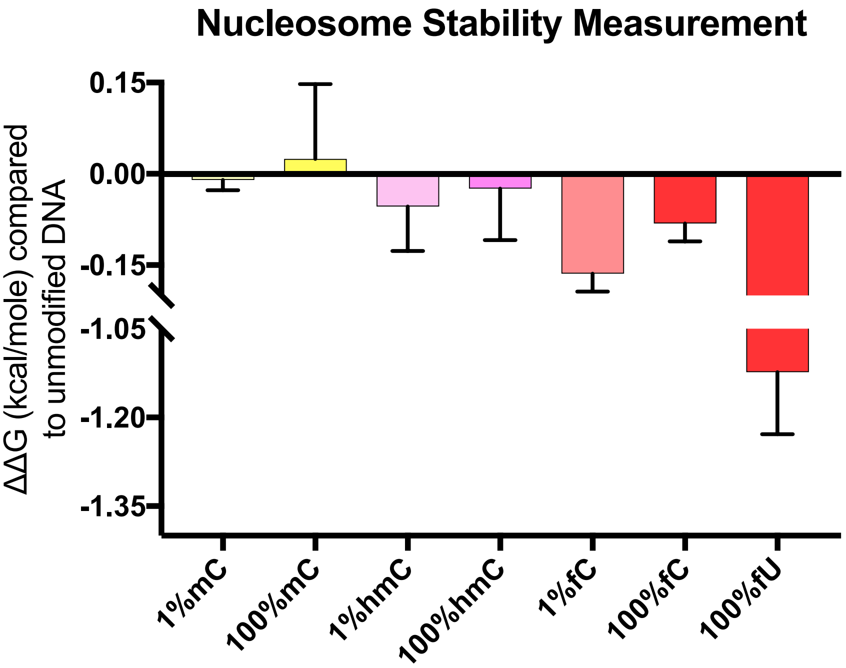


Figure 3-22. Nucleosome stability measurement with salt dialysis: ΔΔG value (kcal/mole) compared against the unmodified counterpart, quantified from native gel electrophoresis experiments (ordinary one-way ANOVA test p value < 0.0001).

mC-, hmC-, fC- and fU-Widom nucleosomes exhibit ΔΔG values that progressively increase in magnitude. This corresponds to a lower nucleosome stabilizing and occupancy-promoting effect for mC and hmC as compared to fC and fU. This is consistent with the observation for nucleosome assembly with biological assisting factors and PGA, where mC, hmC, fC and fU modifications increased the nucleosome occupancy in an increasing order. The stabilizing effect of hmC could be explained by both increased DNA flexibility[111] and potentially additional hydrogen bonding with the histone, the –

CH_2OH group being a H-bond donor. These results on the increase of nucleosome stability by hmC modifications measured by salt dialysis align with Mendonca *et al.*[309] With mC, although the rigidity of DNA has been shown to increase with mC modifications, the $–CH_3$ groups could contribute to base-stacking interactions, resulting in the comparable nucleosome stability of mC-nucleosome with the unmodified counterpart.

This nucleosome stability study assembled nucleosomes with NaCl as the chaperone, and fC continued to demonstrate a promoting effect. Together with the promoting effect of fC observed with PGA-assisted nucleosome assembly, it can be concluded that fC indeed increases both nucleosome occupancy and stability through the influence of fC. The increased stability could possibly be explained by the increased DNA flexibility induced by fC modifications[111], the covalent Schiff base formation, and non-covalent interactions with the –CHO group of fC functioning as a hydrogen bond acceptor.

3.2.10 Positional Preference of fC within the Nucleosome

(NGS data bioinformatics analyses were undertaken by Dr Sergio Martinez Cuesta, University of Cambridge)

Previous experiments have demonstrated that fC has a positive effect on nucleosome occupancy and stability by influencing the intrinsic biochemical and biophysical properties of nucleosomes regardless of the density. I proceeded to study whether there is a positional preference for fC in the nucleosome by SELEX type experiments coupled with single base resolution reduced bisulfite sequencing (redBS-seq, explained in Section 3.2.11.2)[329].

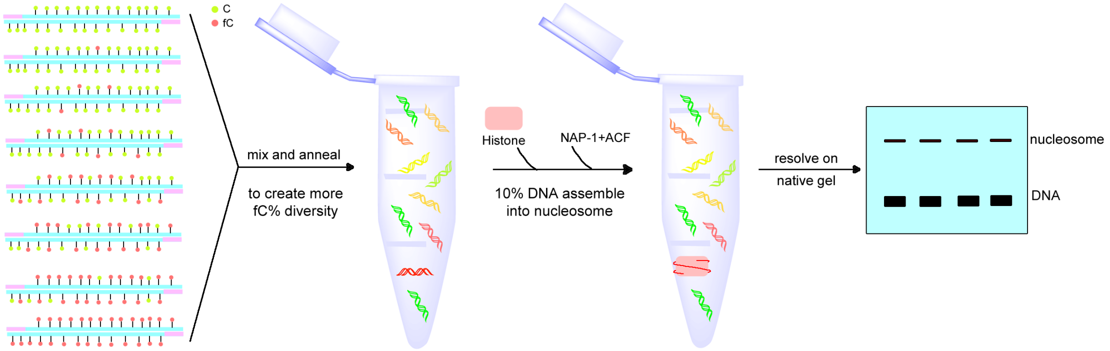


Figure 3-23. Work flow for selecting DNA with highest affinity towards histones from a pool of DNA containing different fC modification densities.

Widom DNA sequences modified with fC at eight different densities between 0% and 100% were prepared by PCR (as detailed in Section 3.2.8.1) and mixed (Figure 3-23). The DNA mixture was denatured and annealed to create a greater modification density variation than in dsDNA. The DNA mixture obtained was combined with a limited amount of histone proteins to select the top 10% modified Widom DNA sequences exhibiting the highest affinity towards histones. The optimization process for nucleosomal DNA separation and recovery was detailed in Appendix Section 6.4, and the optimized workflow is summarized in Figure 3-24.

The nucleosome band was excised and subjected to redBS-seq profiling (detailed in Section 3.2.11.2).[186, 329] The fC positions in the highest affinity DNA sequences were identified by aligning the resultant sequencing data to the Widom sequence. The average density of fC at each position was then calculated, and the density ratio between nucleosomal DNA and control DNA was plotted against DNA positions to highlight the preferred positions of fC. A ratio higher than one indicates that fC was enriched at this particular position during the competitive nucleosome assembly.

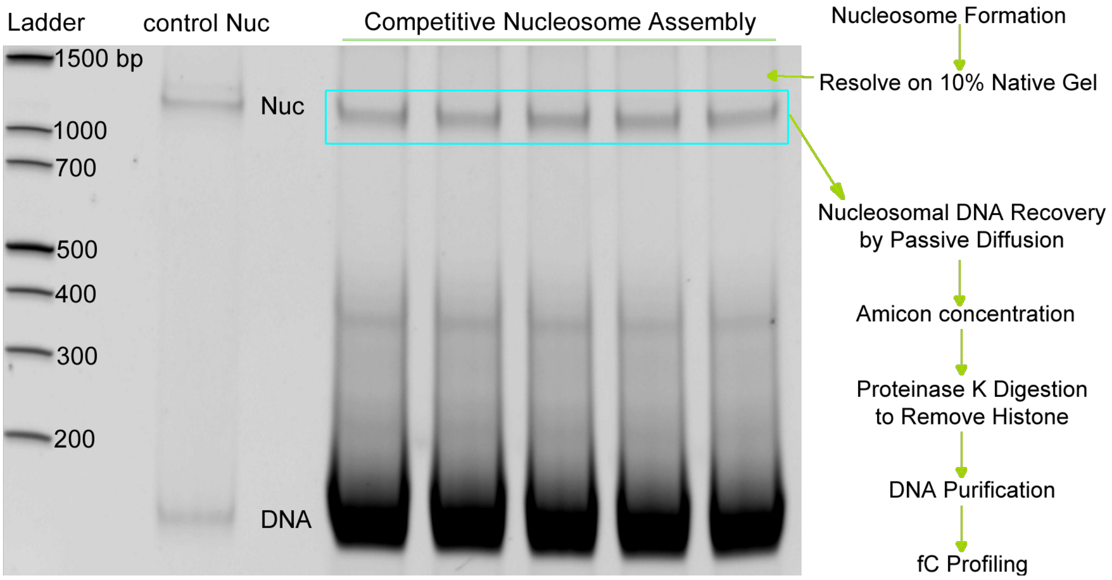

Ladder
control Nuc
Competitive Nucleosome Assembly
1500 bp
1000
700
500
400
300
200
Nuc
DNA
Nucleosome Formation
Resolve on 10% Native Gel
Nucleosomal DNA Recovery by Passive Diffusion
Amicon concentration
Proteinase K Digestion to Remove Histone
DNA Purification
fC Profiling

Figure 3-24. Nucleosomal DNA separation with 10% TBE gel and workflow to profile fC positions in the nucleosomal DNA recovered.

3.2.10.1 Profiling Preferred fC Positions in Nucleosomal DNA

As shown in Figure 3-25a, fC was evenly enriched (1.15 < ratio < 1.32) in both forward and reverse strands at all modifiable positions compared to the untreated DNA, and slightly more enriched (ratio > 1.33) in positions 94, 99-102 and 114 in the reverse strand.

The spatial relationship between the fC positions identified and the histone core (Figure 3-25b and c) was analysed in association with a previously reported high-resolution structure of unmodified Widom nucleosome (3LZ0[51]) under the assumption that the fC modifications do not alter the relative rotational and translational positioning of nucleosomal DNA significantly. This assumption was based on the strong positioning ability of Widom DNA to generate a single nucleosome conformation[51].

fC was enriched at all possible positions compared to the control DNA, therefore there is no distinct preference towards groove positions or distance from the histone core. Positions that were slightly more enriched are located near the H2B (position 114), H3 (position 99-102) and H4 (position 94 and 99-102) subunits, indicating possible enhanced interactions with these two subunits. However, considering the pseudo-2-fold symmetry of the nucleosome structure, and that the corresponding positions (positions around

56 to 74) were not more enriched, suggesting the sequence context also played a role in the slight enrichment. Position 94 and 114 are in major groove positions while position 99-102 are in minor groove positions. Position 94, 99, 100 and 114 are in positions close to histone core, while position 101 and 102 are in positions far away from histone core. Therefore the more enriched fC positions have a slight preference towards minor groove positions (4 vs. 2 major groove positions) and positions close to histone (4 vs. 2 far positions).

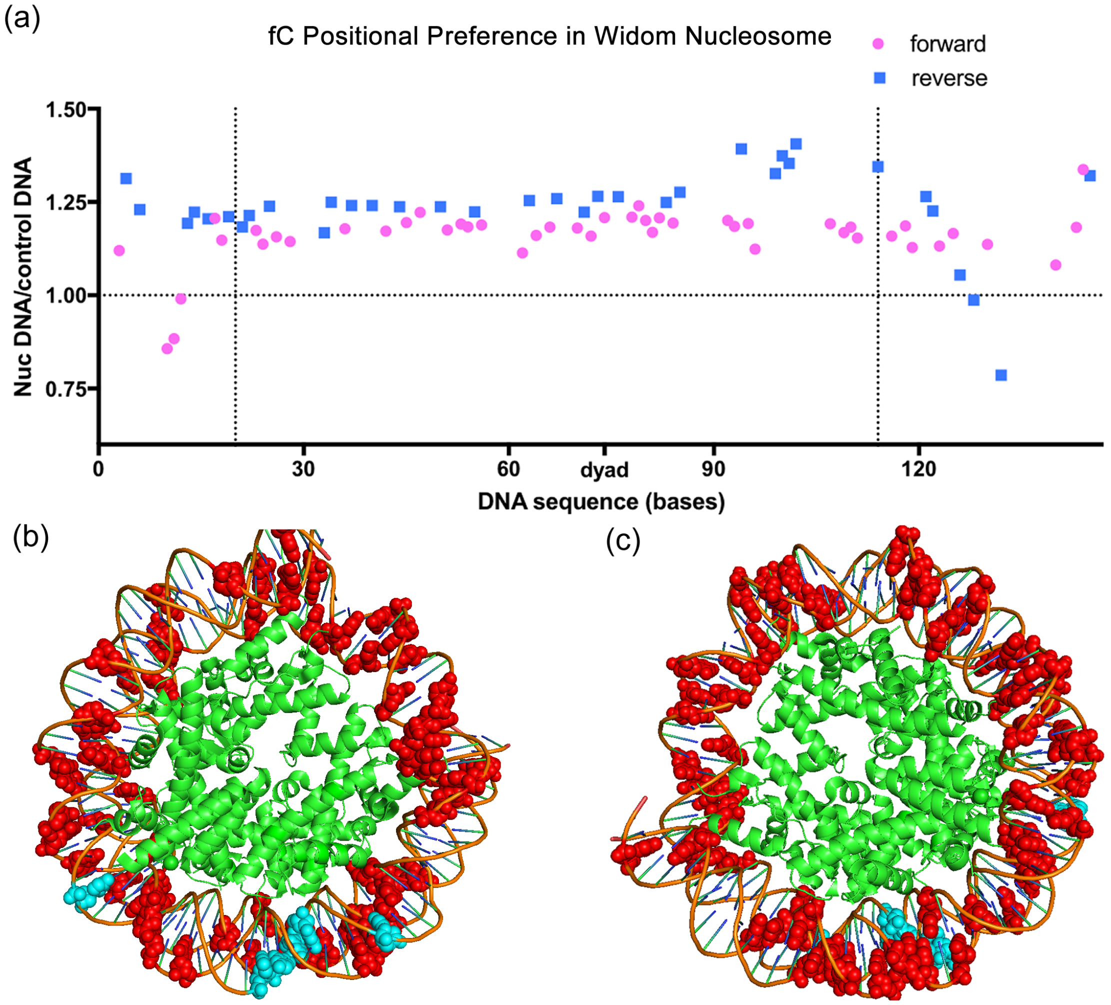


Figure 3-25. Results of enriched fC population density ratio between nucleosomal DNA and free DNA in C/fC competition experiments. The fC was identified by C readings in redBS experiments. (a) The two vertical lines mark the primer regions, where all Cs in the forward strand (magenta 5' to 3') on the left, and all Cs in the reverse strands (blue 3' to 5') on the right are not modifiable. Dots with a value higher than 1 indicate positions that fC are favoured in nucleosomal DNA, while value lower than 1 indicate positions that fC are deselected in nucleosomal DNA. All points are shown between the two primer regions, but some points are omitted for the two primer regions. (b) Top view (c) bottom view of the enriched fC positions compared to the control DNA, highlighted by red (1.15 < ratio < 1.32) and blue (ratio > 1.33) spheres. The nucleosome structure was drawn with 3LZ0[51] with PyMOL[330].

The general enrichment of fC aligns with the promoting and stabilizing effects on the nucleosome observed in previous sections. Therefore the positive effect of fC on nucleosome occupancy/stability does not depend on the position of fC significantly, except the slight preference towards positions in minor groove and close to histone core in Widom sequence context.

3.2.11 Positional Preference of xC within the Nucleosome

(NGS data bioinformatics analyses were undertaken by Dr Sergio Martinez Cuesta)

So far the effects of different DNA modifications on the biophysical properties of the nucleosome have been investigated separately. In the genome, however, the distribution of C derivatives may overlap in certain regulatory regions.[164, 165, 186, 327] For example, hmC and fC can both be found in active enhancer regions, raising the question of how the presence of different DNA modifications within the same nucleosome will impact the nucleosome formation. To address this, a SELEX type approach similar to the one described in Section 3.2.10 was used to investigate the positional preference of xC (x = m, hm and f) within the nucleosome.

Widom DNA with randomly incorporated C, mC, hmC and fC was prepared by PCR using a pool of 2'-deoxycytidine-5'-triphosphate (dCtp), 5-methyl-2'-deoxycytidine-5'-triphosphate (mdCtp), 5-hydroxymethyl-2'-deoxycytidine-5'-triphosphate (hmdCtp) and 5-formyl-2'-deoxycytidine-5'-triphosphate (fdCtp) as shown in Figure 3-26. The resulting DNA mixture was subsequently used in a competitive nucleosome reconstitution assay, where the DNA was present in large excess compared to histone proteins to select for DNA modification patterns exhibiting the highest affinity towards histone. The positions of each xC modification in the nucleosomal DNA was profiled and analysed in combination with the high-resolution nucleosome structure[51].

3.2.11.1 Relationship between PCR Input and Incorporation in Mixed xC context

Since the incorporation rate between the different xdCtp varies, we first investigated the input/incorporation ratio relationship with a pool of varying ratios of dCtp, mdCtp, hmdCtp and fdCtp by PCR followed by liquid chromatography mass spectrometry (LC-MS), similar to the method described in Section 3.2.8.1. In this way, a calibration line between xdCtp input and incorporation was built for mixed xC-DNA (Figure 3-26). Since the incorporation of each C derivative was observed to be influenced by the input percentage of other xCs, the PCR conditions were further screened around the value predicted by the calibration line. The PCR condition was finalized to be 30 mM dCtp, 17.2 mM mdCtp, 124 mM hmdCtp and 28.8 mM fdCtp input to give the 20% final incorporation rate for mC, hmC and fC, which corresponds to around 16.5 of each modified C in Widom DNA. It is noteworthy that the hmdCtp showed a very low incorporation rate compared to other xCs, thus high hmdCtp input was needed in order to achieve the same incorporation rate with other C derivatives.

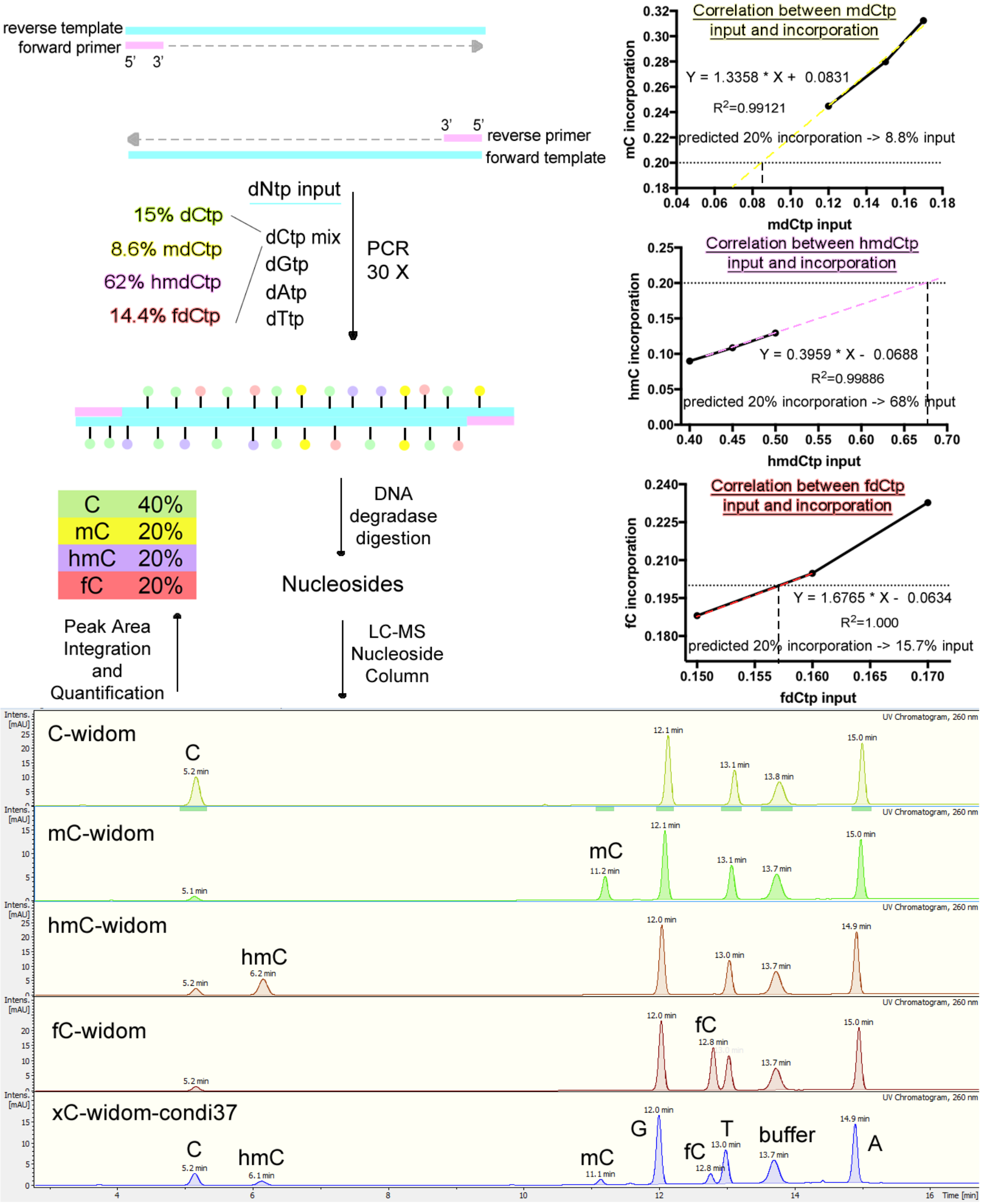


Figure 3-26. DNA with mixed xC random modifications generated by PCR and quantified by LC-MS: the work flow, calibration line between xdCtp input and incorporation, and LC-MS spectra of digested fully modified Widom DNA with single modified C (row 1-4: fully C/mC/hmC/fC modified Widom DNA, respectively) and mixed modified C-Widom DNA with 20% incorporation rate of each modified C (row 5). Peak positions in the LC condition used: C 5.2 min, hmC 6.1 min, mC 11.1 min, G 12.0 min, fC 12.8 min, T 13.0 min, digestion buffer 13.7 min, A 14.9 min. In the PCR reaction, the pink primer regions do not have any modification.

3.2.11.2 Methods to Identify xC Positions in Nucleosomal DNA

The golden standards for profiling C derivatives at single base resolution are bisulfite sequencing (BS-seq), its oxidative derivative, oxidative bisulfite sequencing (oxBS-seq), and its reductive derivative, redBS-seq.[329] BS treatment causes the deamination of C, fC and caC that upon sequencing leads to a C to T change (conversion pattern summarized in Table 3-2, reactions summarized in Figure 6-6). Since C and fC both deaminate upon BS treatment they are not distinguishable. Similarly, mC and hmC are not distinguishable in BS-seq as neither deaminate upon BS treatment. Consequently two different sequencing method based on BS-seq have been developed. By oxidizing hmC to fC in oxBS-seq and comparing the results with BS-seq, hmC can be distinguished from mC. By reducing fC to hmC in redBS-seq and comparing the results with BS-seq, fC can be identified from C. caC cannot be distinguished from C with the combination of BS/oxBS/redBS-seq and is therefore excluded from this study.

Sequencing Method		**BS**	**oxBS**	**redBS**
C	Reaction	C -> U	C -> U	C -> U
	Read Out	T	T	T
mC	Reaction	no reaction	no reaction	no reaction
C-reading(oxBS)	Read Out	C	C	C
hmC	Reaction	CMS formation	oxidized to fC -> deformylate -> C -> U	CMS formation
C-reading(BS-oxBS)	Read Out	C	T	C
fC	Reaction	deformylate -> C -> U	deformylate -> C -> U	reduce to hmC -> CMS
C-reading(redBS-BS)	Read Out	T	T	C
caC	Reaction	decarboxylate -> C -> U	decarboxylate -> C -> U	decarboxylate -> C -> U
	Read Out	T	T	T

Table 3-2. The reaction and response of different modified bases to BS/oxBS/redBS treatment. CMS stands for cytosine-5-methylsulfonate. To aid interpretation, the red blocks highlight the modifications that give a C reading after treatment, and the yellow blocks highlight the modifications that give a T reading after treatment.

It is noteworthy that although oxBS-seq and redBS-seq enable the single base resolution of hmC and fC sites, they can only provide information on the position as a population average rather than at the single DNA molecule level because their identification needs comparison between two datasets.

3.2.11.3 Profiling Preferred xC Positions in Nucleosomal DNA

To select the top 10% of modified Widom DNA sequences exhibiting the highest affinity towards histone, nucleosomes were assembled with randomly modified xC-Widom DNA, and the SELEX experiments were carried out as described in Section 3.2.10. The nucleosome band was excised and used for BS-, oxBS- and redBS-seq. The resulting sequencing data was aligned to the Widom sequence to determine the xC modification pattern of the high affinity DNA sequences. The positional preference of xCs identified was further analysed in combination with the high-resolution structure (3LZ0[51]) to show the spatial relationship between the enriched positions and the histone, under the assumption that the xC modifications do not alter the relative rotational and translational positions of nucleosomal DNA significantly. Interesting distribution patterns emerged when comparing the relative distance and geometric positions between xCs and histone core.

The analysis showed that mC was highly enriched in the reverse strand in all positions, weakly enriched in the forward strand between positions 23 to 62, and disfavoured in the remaining positions (Figure 3-27a). It became apparent that highly enriched mC positions (represented by yellow spheres) were mostly located at positions located furthest away from the histone core (examples highlighted by black arrows in Figure 3-28a), while the weakly enriched positions (represented by white spheres) occupied some positions near the histone core (examples highlighted by pink arrows in Figure 3-28b). The enriched mC was found more in the major groove positions (20 vs. 13 minor groove positions for highly enriched mC, 7 vs. 6 minor groove positions for weakly enriched mC, entry and exit turn mC omitted due to high mobility).

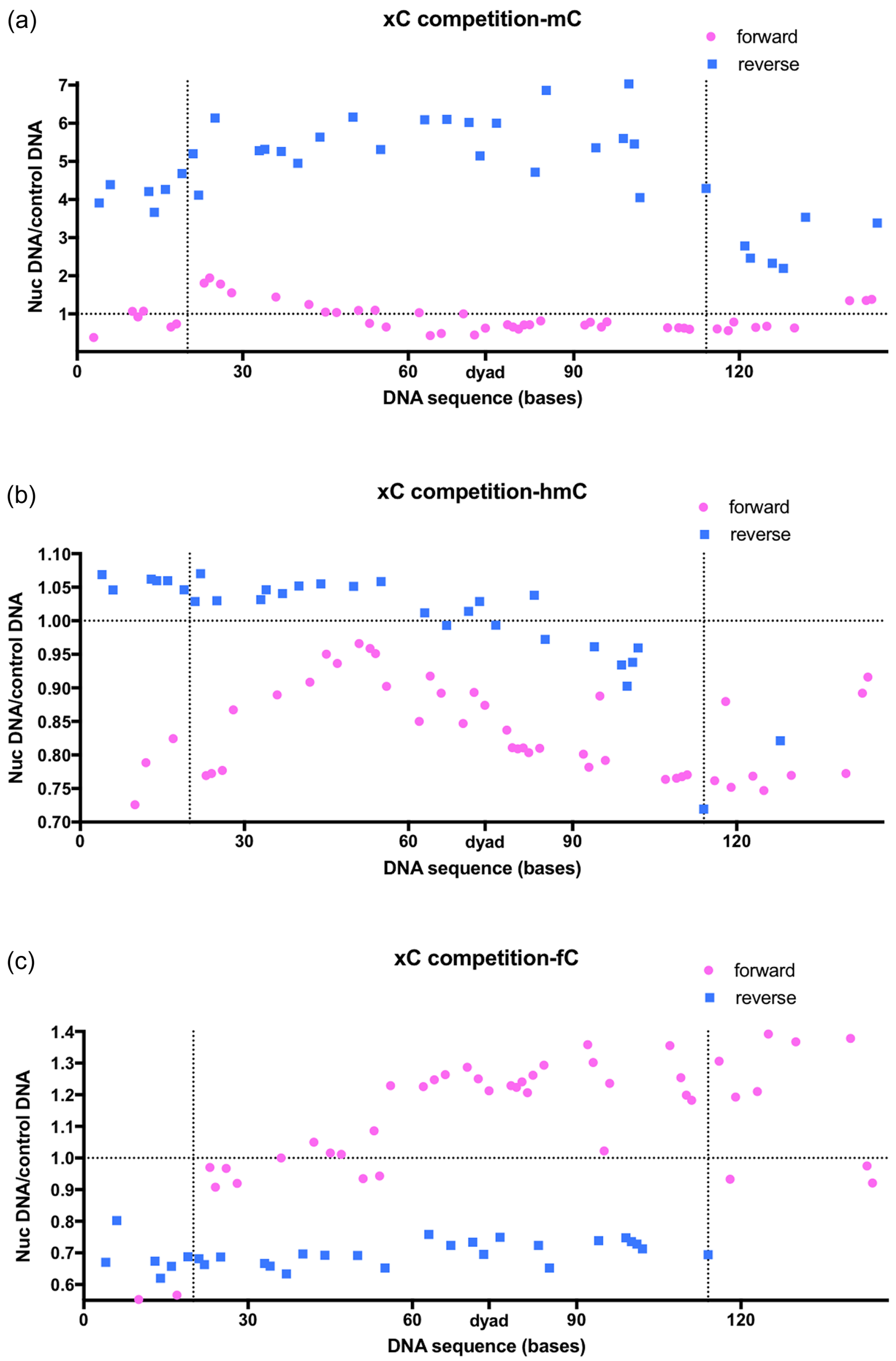


Figure 3-27. Results of enriched xCs population density ratio between nucleosomal DNA and free DNA in xC competition experiments. The two vertical lines mark the primer regions, where all Cs in the forward strand (magenta 5' to 3') on the left, and all Cs in the reverse strands (blue 3' to 5') on the right are not modifiable. Dots with a value higher than 1 indicate positions that xCs are favoured in nucleosomal DNA, while value lower than 1 indicate positions that xCs are deselected in nucleosomal DNA. All points are shown between the two primer regions, but some points are omitted for the two primer regions. (a) Results of enriched mC population density ratio between nucleosomal DNA and free DNA in xC competition experiments. The mC was identified by C readings in oxBS results. (b) Results of enriched hmC population density ratio between nucleosomal DNA and free DNA in xC competition experiments. The hmC was identified by subtracting the C readings in oxBS reads from those in BS reads. (c) Results of enriched fC population density ratio between nucleosomal DNA and free DNA in xC competition experiments. The fC was identified by subtracting the C readings in BS reads from those in redBS reads.

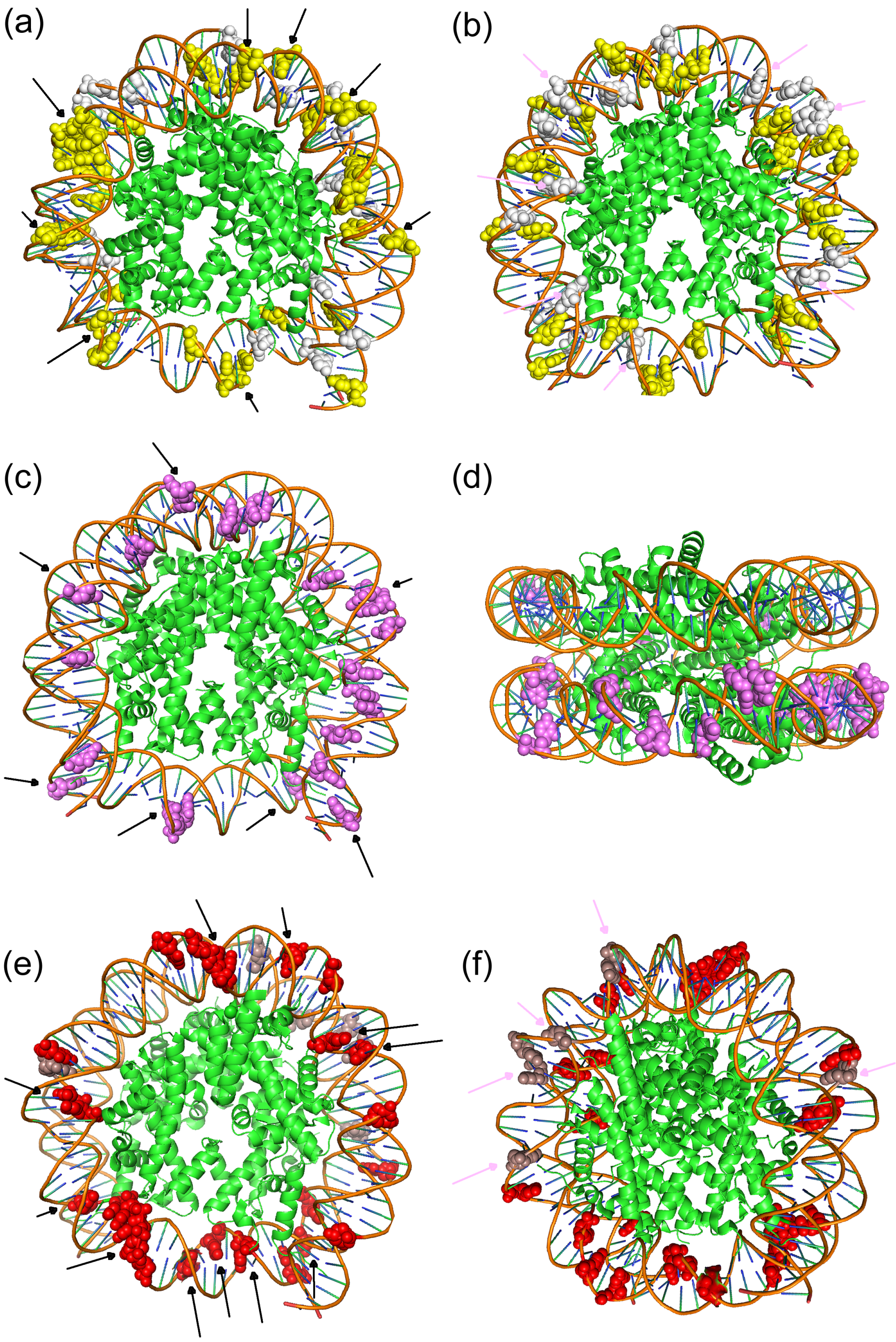


Figure 3-28. Positions of enriched xC in nucleosomal DNA by xC competition experiments. The nucleosome structure was drawn with 3LZ0[51] with PyMOL[330]. (a) Top view of the enriched mC positions, examples of highly enriched mC positions (ratio > 2) are pointed out by black arrows. (b) Bottom view of the enriched mC positions, examples of weakly enriched mC positions (1 < ratio < 2) are pointed out by pink arrows. The enriched mC positions are highlighted by yellow (highly enriched) and white (weakly enriched) spheres. (c) Top view of the enriched hmC positions, examples of enriched hmC positions (ratio > 1) are pointed out by black arrows. (d) Side view of the enriched hmC positions to show that enriched hmC positions are mainly located in the first half of nucleosomal DNA. The enriched hmC positions are highlighted by purple spheres; (e) Top view of the enriched fC positions, examples of highly enriched fC positions (ratio > 1.18) are pointed out by black arrows. (f) Bottom view of the enriched fC positions, examples of weakly enriched fC positions (1 < ratio < 1.05) are pointed out by pink arrows. The enriched fC positions are highlighted by red (highly enriched) and maroon (weakly enriched) spheres.

The data showed that hmC was generally disfavoured in the forward strand, while enriched in the first half of the reverse strands between positions 4 to 83 (Figure 3-27b). The enriched hmC positions are favoured in major groove positions (15 vs. 5 for minor groove positions, entry turn hmC omitted), and located near the groove alternation points from the perspective of histone (examples highlighted by the black arrow in Figure 3-28c), occupying positions closer than highly enriched mC positions from histone core. It is also quite interesting to observe hmC enriched only in the first half of nucleosomal DNA (Figure 3-28d), which may be an indication of a preference for DNA sequence context but not histone subunits because of the pseudo-2-fold symmetry of the nucleosome particle.

fC was disfavoured in the reverse strand of Widom DNA, while weakly enriched in the first half of the forward strand between positions 36 to 53, and strongly enriched in the second half of the forward strand between positions 56 to 140 (Figure 3-27c). The highly enriched fC positions (represented by red spheres) were mostly immediately next to the histone core (Figure 3-28e, examples highlighted by black arrows), while the weakly enriched positions (represented by maroon spheres) were located further away from histone core than the highly enriched positions (Figure 3-28f, examples highlighted by pink arrows). The highly enriched fC, reverse to mC and hmC, was found more in the minor groove positions (14 vs. 10 for major groove positions, exit turn fC omitted). The trend observed here agrees with what was observed for the slightly enriched fC positions in Section 3.2.10.1.

The relative distance between DNA modifications and the histone core may suggest the strength of interaction. fC seems to interact most strongly with the histone core, possibly through potential covalent Schiff base formation and non-covalent hydrogen bonding interactions. hmC seems to interact with histone cores less strongly than fC as it occupies positions further from histone core. The mC mainly occupied the positions that are furthest away from the histone core, therefore the interactions between mC and histone core are likely to be the weakest of all three C derivatives investigated. The relative distances between enriched positions of xC from histone cores also agree with the order of nucleosome occupancy and measured stability. All xCs were enriched in selective positions in nucleosomal DNA, aligning with our previous observation that all three C derivatives promote nucleosome formation to different extent.

The positions of enriched mC located preferentially in the major groove rather than minor groove agrees with previously reported FRET experimental observations and by computation[305, 306], but conflicted with observations by Collings *et al.* and Chodavarapu *et al.*[294, 308]. This may suggest that rather than groove type, mC influences the nucleosome occupancy and stability through the relative distance to histone, considering the clear strand preference combined with the CpG modification pattern of Widom DNA. The mCs in the forward strand in the CpG dinucleotide basepairs were not as enriched as the mC in the reverse strand even through they reside in the same type of major groove, suggesting that it is really the distance of the base modification from the histone core that governs the selection.

Using the current method, with mixed C derivatives present, hmC and fC can only be determined on an averaged population level, thus the modification pattern of different C derivatives is missed in the same strand of DNA. In addition, there is no way of relating the forward and reverse strand of the same duplex DNA with the current method, thus the modification pattern of duplex DNA cannot be correlated using this method. Introducing randomized indexing enables identification of both stands of dsDNA, and allows differentiation of the signals from dsDNA with the same modification pattern from that from PCR replicates.[331] In addition, newly emerged sequencing technologies such as small molecule real time (SMRT) sequencing (PacBio)[332] and Nanopore sequencing (Oxford)[333] may one day enable profiling all xC in the dsDNA simultaneously without destroying the strands.

3.2.12 Covalent Interaction between fC/fU and histone proteins

(Condition optimization for reduction capture and PolStop Assay were completed by Dr Robyn Hardisty, University of Cambridge, and the NGS data bioinformatics analyses were undertaken by Dr Sergio Martinez Cuesta)

The aldehyde group of fC can potentially react with the primary amines in histone tails (such as the ε-NH2 group of lysine residues and the α-NH2 group of N-terminal amino acids) to form Schiff base (imine, Figure 3-29), providing a very interesting chemical mechanism for promoting nucleosome formation and stability, and further directing nucleosome positioning in genomic DNA. We set out to explore whether Schiff bases can form, and if so, in what positions fC modifications are more likely to form such interactions. In addition, our group has previously shown that the –CHO group of fU is more reactive than that of fC[325], therefore the Schiff base formation ability of fU was also investigated.

3.2.12.1 Proteomics to Identify the Cross-linked Histone Subunits

As the Schiff base formation is highly reversible by hydrolysis, the C=N bond formed was captured by reduction to CH-NH with sodium cyanoborohydride ($NaCNBH_3$), which cannot be hydrolysed as easily. The captured Schiff base was characterized by denaturing gel electrophoresis, which eliminates all non-covalent interactions.

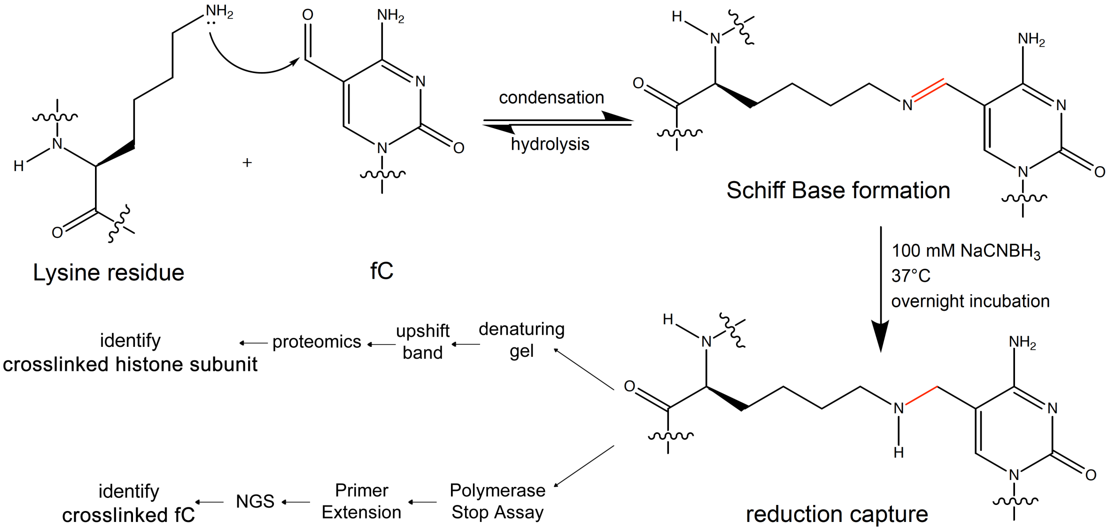


Figure 3-29. Proposed Schiff Base formation between lysine from histone and fC from DNA, and capture by reduction with 100 mM $NaCNBH_3$ incubated at 37°C overnight. The captured histone subunit was identified by proteomics of the excised upshift band in denaturing gel; and the captured fC was identified by PolStop assay and NGS.

The molecular weight (MW) of ssDNA is around 46 kDa for fC- and fU-Widom DNA, whereas single histone subunits range between 11.3 and 15.3 kDa (average 13.6 kDa). Upon reduction, we observed an upshift band in fC lanes migrated between the 50 and 70 kDa markers that are not present in C- and caC-Widom DNA lanes, corresponding to the molecular weight of ssDNA+one histone subunit (~59.6 kDa, Figure 3-30). Fainter higher upshift bands were also observed with a molecular weight corresponding to ssDNA cross-linked to multiple histone subunits.

The band corresponding to the ssDNA+one histone subunit from the fC lane was excised and analysed by proteomics. All four histone subunits could be identified by mass spectrometry, indicating that Schiff base was formed within the nucleosome context. However, H2A only showed a very low number of reads compared to other subunits. This could stem from the fact that fewer fC modifications are present near H2A because both copies of H2A subunit in nucleosome particle locate close to the unmodified primer regions of the nucleosomal DNA on both 5' ends. Expectedly, no histone proteins were identified from the same migration position for the C lanes. At the time of our analysis, two papers were published reporting the existence of Schiff bases between fC and lysine in *in vitro* reconstituted nucleosomes[334] and also in human embryonic kidney cells[335], detected by denaturing gel electrophoresis and nanoLC-MS.

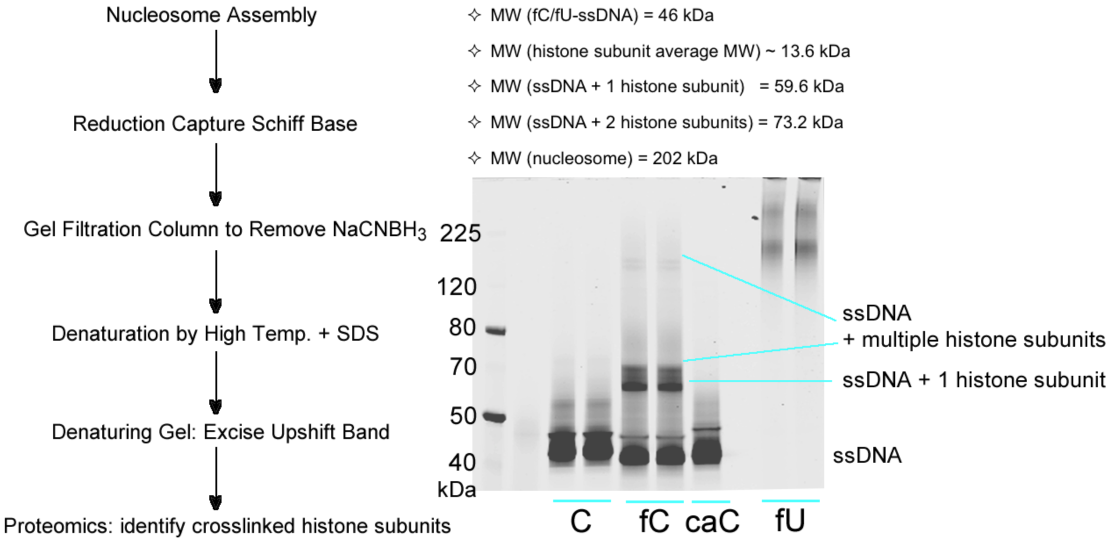

Nucleosome Assembly
Reduction Capture Schiff Base
Gel Filtration Column to Remove NaCNBH3
Denaturation by High Temp. + SDS
Denaturing Gel: Excise Upshift Band
Proteomics: identify crosslinked histone subunits
✧ MW (fC/fU-ssDNA) = 46 kDa
✧ MW (histone subunit average MW) ~ 13.6 kDa
✧ MW (ssDNA + 1 histone subunit) = 59.6 kDa
✧ MW (ssDNA + 2 histone subunits) = 73.2 kDa
✧ MW (nucleosome) = 202 kDa
225
120
80
70
50
40
kDa
ssDNA
+ multiple histone subunits
ssDNA + 1 histone subunit
ssDNA
C
fC
caC
fU

Figure 3-30. The schematic illustration of capturing Schiff base formation between fC- and fU-Widom DNA and histone proteins in nucleosome context with SDS denaturing protein gel, and identifying the cross-linked histone subunits by proteomics with the upshift gel band. Two replicates were shown for C, fC and fU, and one was shown for caC. The DNA was labelled with Cy3 and Cy5 fluorophores and stained with GelRed to compensate for the fluorophores damage during reduction, and this gel was imaged with both Cy3 channel (to identify the position of other ladders, result not shown) and Cy5 channel (shown in this figure).

In contrast, the fU lanes only showed two well-defined bands with a molecular weight slightly above and below 225 kDa, without any free DNA bands present between the 40 and 50 kDa markers. As the molecular weight of the fU-nucleosome is 202 kDa, the lower band likely corresponds to the densely cross-linked nucleosome, suggesting that even the harsh denaturing conditions used did not separate the two ssDNA and eight histone subunits. The molecular weight of the upper band suggests that this could be the cross-linked product of nucleosome+NAP-1(56 kDa), as experiments with model fU-DNA did show cross-linking ability with non-close interacting proteins in the buffer. The cross-linked product is less likely to be nucleosome + either ACF subunit, as the molecular weight of Acf1 and ISWI are 185 and 140 kDa respectively. Proteomic analysis was performed to identify the two bands, however possibly due to the high degree of cross-linking between fU and histone subunits, no histone fragments were identified from the excised band.

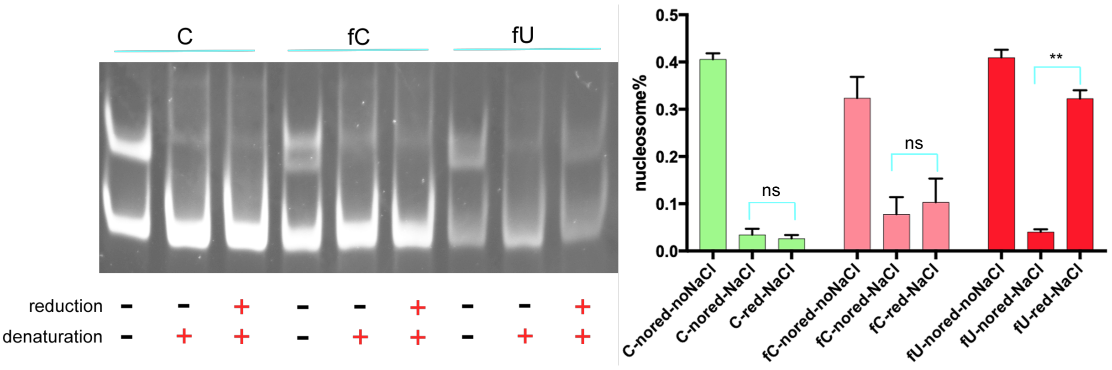


Figure 3-31. The reduction capture of Schiff base formed in fC- and fU-nucleosome with 0.1 M $NaCNBH_3$ at 37 °C incubation and subsequent disturbance of non-covalent bond with 2 M NaCl. C-nucleosome was used as the control.

The existence of Schiff bases within the fU-nucleosome was further investigated with a salt-disturbance experiment (Figure 3-31). The nucleosomes were assembled and the resultant Schiff base was reduced. The non-covalent intramolecular interactions within the nucleosome were then disturbed with 2 M NaCl, enough to compete out ionic interactions between histones and DNA. Samples that were not reduced, with/without salt denaturation were prepared as controls. C- and fC-nucleosomes were also prepared as controls. The reaction mixtures were resolved by native gel electrophoresis, so that if the crosslinking observed by denaturing gel in Figure 3-30 happens between fU-DNA and histone subunits, the cross-linked nucleosome would migrate to the same position as the native nucleosome control.

As shown in Figure 3-31, compared with untreated nucleosomes, the nucleosome percentage dropped significantly for nucleosomes with NaCl only treatment, suggesting that the DNA-histone interactions were mostly diminished without reduction capture. With reduction, the nucleosomes migrated to the same position as the untreated controls, suggesting that the cross-linking likely happened between DNA and histone, rather than with other proteins in the solution.

The ratio of nucleosome occupancy with and without reduction was used to gauge the extent of the covalent Schiff base bond formation. With reduction, the nucleosome percentages compared to the non-reduced samples were increased for both fC and fU, but not for C. The normalized Nuc% percentages of samples with NaCl treatment without vs. with reduction compared to the respective nucleosome percentages of untreated samples are as follows: 8.5% vs. 6.7% for C (p value 0.1606, paired t-test); 23.5% vs.33.0% for fC (p value 0.5825); and 9.8% vs. 78.8% for fU (p value 0.0020). The increase in nucleosome percentage of the reduced fU-nucleosome compared to the non-reduced sample suggests that the covalent Schiff base is responsible for maintaining DNA-histone interactions through salt disturbance. Combined with the denaturing gel results (Figure 3-30), it can be inferred that the Schiff bases occurred at multiple sites within the fU-nucleosome to link the two ssDNA and eight histone subunits together.

The nucleosome percentage is only slightly increased in the fC-nucleosome with reduction capture, showing that the Schiff base indeed anchors fC-DNA onto the histone, but most nucleosomes were still dissociated, likely because the fC-Schiff base is highly reversible and occurs with low probability as shown by the denaturing gel (Figure 3-30). In addition, the reduction capture was done for only two hours rather than overnight, and therefore small amount of Schiff bases were captured, in contrast with fU-nucleosome. This again demonstrated that the –CHO group of fU is more reactive than that of fC, agreeing with previously literature[325]. In addition, this result aligns with the previous observation that fU-nucleosome is more stable than fC-nucleosome (-1.123 kcal/mole and -0.164 kcal/mole, respectively, unpaired t-test with Welch's correction: p value = 0.0019).

The observation of Schiff base formation in this experiment suggests a possible mechanism for directing and locking the nucleosome positioning even with a single copy of fC or fU modification. As the biological function and genomic distribution of fU is not yet clear, the fC study was initially prioritized. Considering the high density of fC modification used (70/147 bp) and the number of primary amines (lysine residues and N-terminal amino acids) in eight histone subunits, and that the band intensity for ssDNA+1 histone subunit was much more intense than that for ssDNA+multiple histone subunits, it is fair to deduce that the majority of the cross-linking happened only between one fC and one primary amine residue in each nucleosome under the conditions used. As the Schiff base forms at various positions within a nucleosome, it is pertinent to identify the preferred fC positions for the Schiff base formation, to understand its effect on directing nucleosome positioning.

3.2.12.2 PolStop Assay to Identify the Cross-linked fC Positions

The fC-histone positions were identified through a Polymerase Stop (PolStop) Assay coupled with NGS as illustrated in Figure 3-32. After Schiff base formation and trapping the Schiff base by reduction, the histone subunit was irreversibly covalently linked to the DNA. During the primer extension, the bulk size of the cross-linked histone subunit stalled the procession of polymerase before or at the cross-linked sites and resulted in short DNA fragments. The elongation phase of the primer extension experiment was restricted to five minutes to reduce polymerase bypass caused by long incubation. In order to account for the naturally existing polymerase-stalling events, the free fC-Widom DNA was used as a no-reduction control.

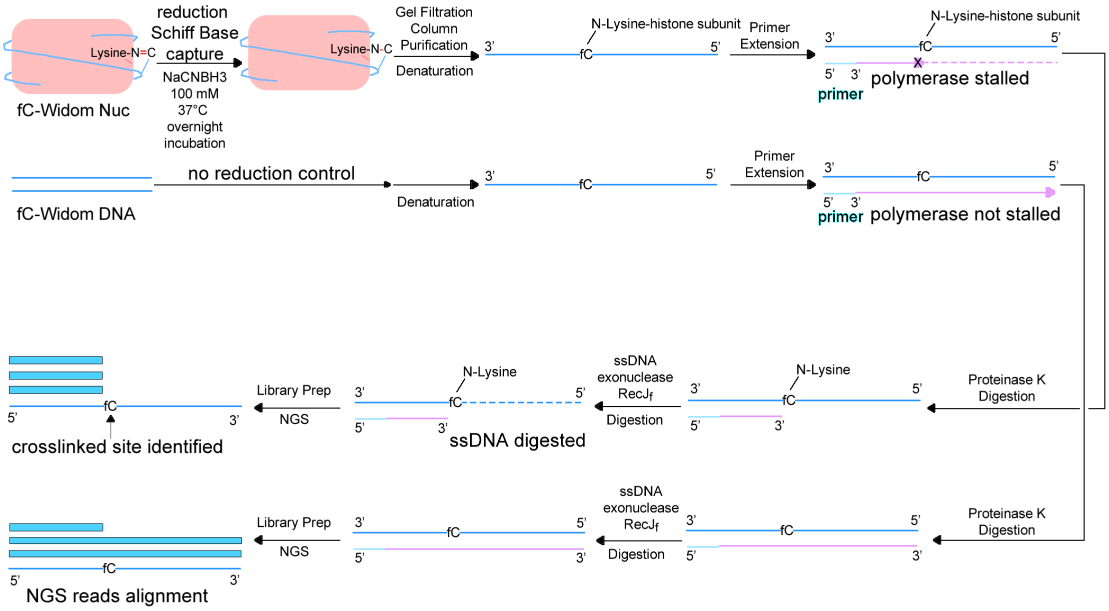


Figure 3-32. Schematic illustration of PolStop Assay and NGS to identify the positions of fC that preferentially form Schiff bases.

The number of reads of polymerase stalling at each position was counted for both the reduction capture sample and the no-reduction control (adjusted to the library size), and the ratio of the two numbers represents the true stalling events at each position:

$$Fold\ change\ (position\ x) = \frac{reads\ stalled\ at\ x\ (reduction\ capture)}{reads\ stalled\ at\ x\ (no\ reduction\ control)}$$

A higher ratio of fold change indicates a higher frequency of stalling at this site, and thus more probable and stable Schiff base formation at this base position. The fold change results obtained from bioinformatics analysis ($\log_2$ value) were plotted against DNA positions as shown in Figure 3-33. As expected, the results indicated that the peak of stalling always locates at the exact fC positions or one to three bases before. The results demonstrated a clear pattern of ~11 bp periodicity of Schiff base formation.

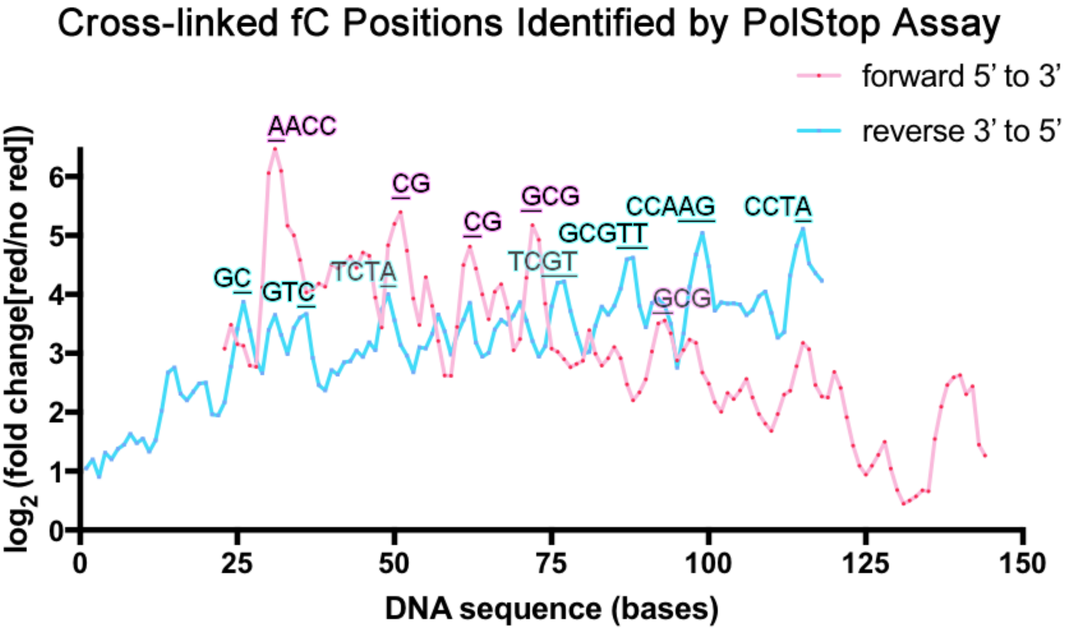


Figure 3-33. Log_2 fold change ratio between the numbers of reads stalled at each DNA position identified from the reduction capture nucleosome sample and that from the no-reduction control DNA sample, for forward strand (pink) and reverse strand (blue).

The fC positions identified were further analysed in the context of the previously reported high-resolution structure of unmodified Widom nucleosome (3LZ0[51]) under the assumption that the fC modifications do not alter the relative rotational and translational positioning of nucleosomal DNA significantly. From Figure 3-34 it is clear that almost all the fC positions (18/19) identified are located in the major groove positions of nucleosome, while the number of fC positions close to or far away to histone are about the same (9 vs. 10).

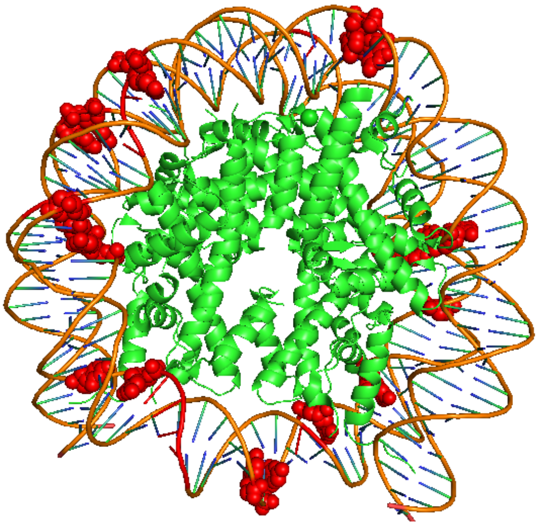


Figure 3-34. The fC positions (highlighted by red spheres) identified by PolStop Assay are mostly located in the major groove positions (top view). The nucleosome structure was drawn with 3LZ0[51] with PyMOL[330].

3.2.12.3 Significance of fC-lysine Schiff base formation

Needing only a few copies of fC, the Schiff base can restrict the DNA mobility and increase the local concentration of DNA. In this way the equilibrium shifts towards the nucleosome and the nucleosome occupancy and stability are thus increased, and may further direct nucleosome positioning in genomic DNA. The detection of Schiff bases *in vivo*[298, 335] may suggests that Schiff base formation plays a part in the regulation of cellular processes.

The DNA-protein covalent interactions naturally raise questions about the possible hindrance to biological processes such as DNA replication and transcription. Recently Ji *et al.* reported that when the fC-H2A/fC-H4 Schiff base is positioned at the beginning of replication (i.e. the exact binding site of DNA polymerase), the kinetics of polymerases are severely reduced, with an increase in C to T mutations and deletions at the position opposite to the fC-Schiff base.[336] However, in a cellular environment the Schiff bases are not reduced and therefore highly reversible. In our PolStop Assay setup we observed that with long incubation time (30 minutes) at the primer extension step, the polymerase eventually passed through the reduced cross-linked DNA-histone position and resulted in minimal stalling. Therefore, our results suggested that as long as the Schiff base is away from the exact replication/transcription initiation base pair, even with the DNA-protein covalent linkage, the genetic information should be available for readout when needed, albeit at a potentially slower rate, rather than causing catastrophic results.

Since Schiff bases can be formed between fC and lysines, considering the plethora of reports on lysine modifications/mutations[337, 338] implicated in cellular activities[339, 340] and even disease states[341], such as cancer[338] and developmental disorders[342, 343], the Schiff base may play a role in these observations and is worth further investigation.

3.3 Incorporating Desired C Modifications at Designated Positions

(NGS data bioinformatics analyses were undertaken by Dr Sergio Martinez Cuesta, University of Cambridge)

In the studies using fully modified DNA and randomly modified low-density DNA modifications in previous sections, the results obtained on nucleosome occupancy and stability represented a combined effect from modifications at different positions. In addition, the base positions identified in the positional preference study (Section 3.2.11) and Schiff-base formation study (Section 3.2.12) need to be interrogated separately to confirm the positional preferences observed on a population level. Therefore a simple, quick and reliable method for incorporating DNA modifications at designated positions is required.

3.3.1 Existing Methods for Producing xC Modified DNA

Several methods have been traditionally used to introduce xC modifications into DNA strands, including solid phase synthesis, enzymatic modification and polymerase incorporation. The advantages and disadvantages of these methods are explained as follows.

3.3.1.1 Solid phase synthesis

Solid phase synthesis uses phosphoramidite chemistry to synthesize ssDNA from the 3’ to 5’ end nucleotide by nucleotide in a stepwise fashion going through cycles of coupling, oxidation and detritylation steps (Figure 3-35)[235]. Therefore, there is neither limitation of sequence context nor ambiguity regarding the position of the modified bases. In addition, as long as the phosphoramidite for the target modification can be synthesized and protected throughout the synthesis and purification procedures, there is virtually no limitation on the types of modification for incorporation. However, this method has drawbacks such as yield and product length restriction, and the preparation can be both costly and time consuming.

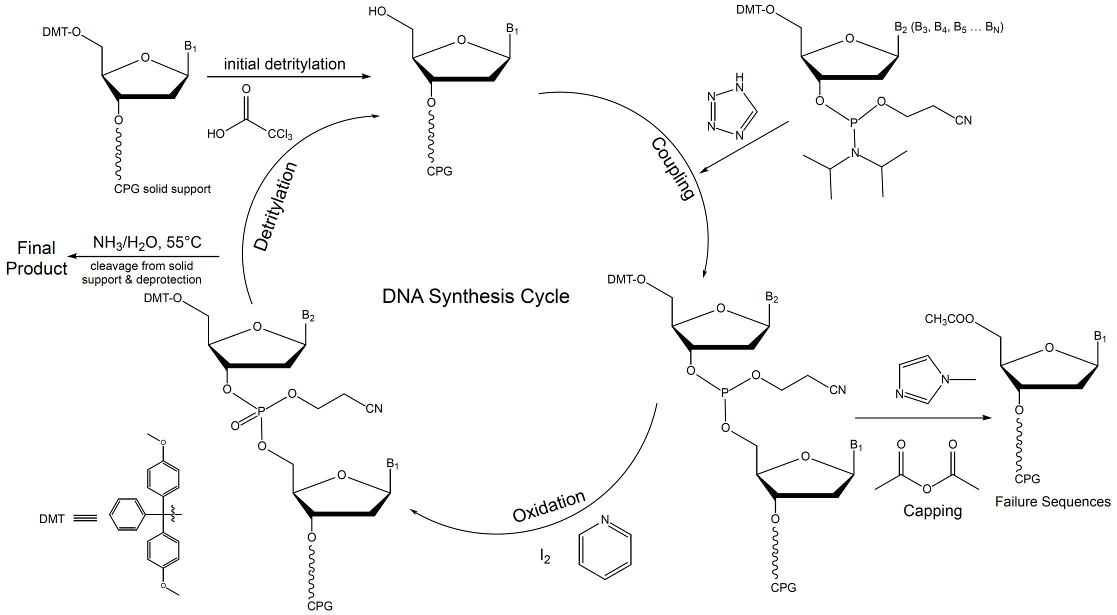


Figure 3-35. The schematic illustration for phosphoramidite chemistry for solid phase DNA synthesis. DMT group stands for 4,4'-dimethoxytrityl group, B stands for protected base, CPG stands for controlled-pore glass. The reaction condition was taken from [344].

As the DNA is synthesized in a stepwise manner, the final product yield depends exponentially on the coupling yield as a function of the product length. For example, to synthesize a 100-nucleotide (nt) ssDNA with a 90% coupling efficiency for each step, the overall yield is merely 0.0027%. Even with a high 99% coupling efficiency for each synthesis cycle, the yield for a 200-nt product synthesis is only 13%. Therefore manufacturers commonly restrict the ssDNA synthesis to 150- or 200-nt to ensure a sufficient yield.

In addition, after solid-phase synthesis, the DNA needs to be cleaved from the solid support and deprotected using heated liquid ammonia, which can be harsh for some DNA modifications. As the modified phosphoramidites need to be specially prepared[345-349], and they may be incompatible with standard conditions for DNA synthesis and purification, the long ssDNA preparation with DNA modifications can be very expensive.

To circumvent the yield and cost issue with long modified DNA preparation, Li *et al*[334] used an elegant method reminiscent of DNA origami[350]. Short fragments of DNA are directed in the correct order and in close proximity by partially complementary scaffold DNA; a DNA ligase was used to stitch neighbouring short fragment DNA strands together into a single piece of long ssDNA. In this way, DNA modifications are free to be placed at desired positions within a short DNA prepared by solid phase synthesis, and all DNA fragments can be prepared in high yield at lower cost. However, since this involves annealing multiple fragments together in one pot, the molecular ratio needs to be meticulously controlled. Imprecise input of any fragments will lead to incomplete products requiring further separation.

3.3.1.2 Introducing DNA Modifications via Enzymatic Modification

Various modifying enzymes have been developed to introduce DNA modifications *in vitro*. For example, DNA methylation at CpG sites can be achieved by CpG methyltransferase M.SssI[351], and further oxidation of mC to hmC, fC and caC can be achieved by TET[100-104].

However, introducing DNA modifications via enzymatic reactions can be difficult to reach completion. Additionally, it is difficult to halt the stepwise oxidation of TET with high efficiency, to obtain pure hmC- or fC-modified DNA, for example. The positions of modifications are also quite dependent on the enzyme specificity and sequence context biases. If there are multiple suitable substrate positions, it is difficult to control the sequence context of the site of modification. The types of DNA modifications possible are also limited to the capability of currently available modification enzymes.

3.3.1.3 Introducing DNA Modifications via Polymerase

DNA modifications can also be incorporated into the DNA strand using modified nucleotide triphosphates during PCR or primer extension. This method can be comparatively cheap and fast as long as the modified nucleotide triphosphate can be generated and is compatible with the experimental conditions for strand elongation. Therefore, the DNA used for the nucleosome occupancy and stability study was prepared using this method. However, the exact positions of incorporation cannot be controlled easily except in the preparation of fully modified dsDNA strands (excluding primer regions) in which all canonical Cs are replaced by modified Cs (Figure 3-2). While it is possible to control the global proportion of base modifications in this approach by fixing the xdCtp to dCtp ratio (as detailed in Section

3.2.8.1), it is not possible to selectively incorporate a base modification in a specific sequence context in the presence of the canonical base.

Considering the disadvantages in cost, product purity and modification position ambiguity, these methods are not satisfactory for the systematic studies of various nucleosomal DNA positions identified in previous sections. Therefore, it is imperative to design a method to incorporate modified nucleotides at designated positions with a fast and simple procedure that achieves a high yield.

3.3.2 Design and Optimization of a Gap Filling-Ligation Method

The inspiration for such a method arose from the development of a molecular inversion probe (also known figuratively as the padlock probe) used for sequencing single nucleotide polymorphism analysis or loci capture at selected regions.[352, 353] Probes are designed to complement the flanking areas of genomic DNA of interest. Upon annealing, the complementary part anchors the probe to the correct genomic region, leaving a gap between the areas under investigation. Polymerase and dNtps are then added to fill in the gap, with the 3' –OH of the probe ligated with the 5' phosphate group. In this way, the genetic information from only the area of interest can be enriched and sequenced.

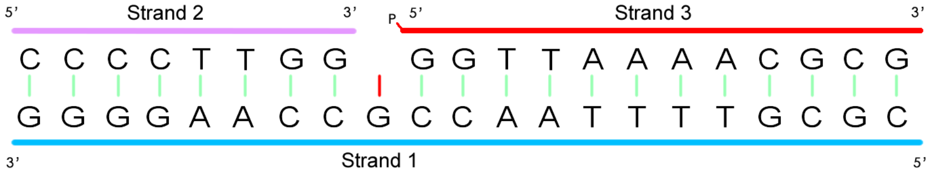


Figure 3-36. Schematic for the 3-strand system for the GL (gap-filling ligation) platform.

This design inspired the development of a method for introducing a target modified nucleotide into DNA using polymerase extension over the gap at designated positions with a 3-strand system (Figure 3-36). The method design was initiated for the incorporation of C derivatives as the mC and oxidative derivatives are among the most abundant and biologically relevant DNA modifications as described in Chapter 1.

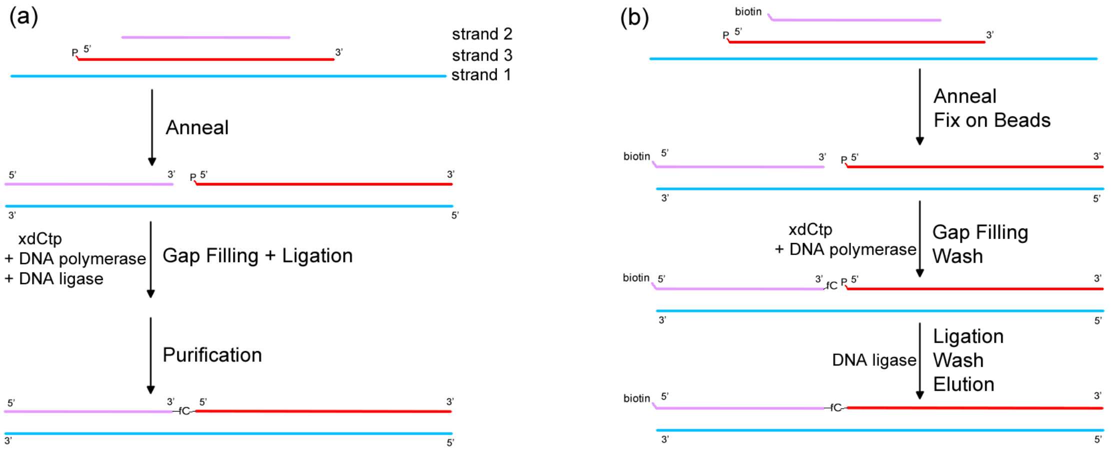


Figure 3-37. Schematic for the workflow for GL reaction using fC as an example (a) without biotin (b) with biotin.

Strands 1, 2 and 3 were synthesized by solid-phase synthesis and the 5' end of strand 3 is phosphorylated. Strand 1 (template strand) is annealed together with strands 2 and 3, leaving a 1-nt gap where the base modification is required (Figure 3-37a). A DNA polymerase is then used to fill the gap opposite to G with xdCtp. DNA ligase then covalently links the 3'-OH of the incorporated modified nucleotide with the 5' phosphate group of strand 3 to form a phosphodiester bond. The resultant DNA can be purified from unreacted fragments by separation by either polyacrylamide gel electrophoresis or on size exclusion/ion exchange columns. Alternatively, strand 2 can be biotinylated at the 5' end (Figure 3-37b) to facilitate purification through streptavidin bead binding, where unreacted DNA strands can be removed by washing. However, biotinylation is not always possible if the 5' end of strand 2 needs to be labelled with other modifications, such as Cy3 fluorescence labelling, therefore experimental conditions for both scenario were developed.

This workflow was named as the Gap filling-Ligation (GL) method. Each step of the workflow was optimized as follows.

3.3.2.1 Annealing Step Optimization

The concentration of strands 1, 2 and 3 was carefully quantified by UV absorbance. In order to drive the equilibrium towards duplex DNA formation with three strands, strand 3 was added in slight excess. Strand 3 is a better choice than strands 1 and 2 because strand 1, when in excess, can form incomplete annealing products, strand 1+2 or strand 1+3. These can be difficult to purify due to their similar length to the final product, and can create confusion for the efficiency evaluation detailed in Section 3.3.3. With strand 2 in excess, the quantification of ligation efficiency will be skewed to a lower value (detailed in Section 3.3.3.3). Conversely, with strand 3 in excess, the only incomplete annealing product possible is strands 1+3, which do not influence the subsequent efficiency evaluation, and can be easily washed off streptavidin beads if strand 2 is biotinylated. Adding strand 3 in slight excess rather than strand 1 and biotinylated strand 2 is more economical since strand 1 is invariably the longest sequence of all three strands, making it the most expensive to purchase, and strand 2 can be more expensive than strand 3 due to the 5' biotin modification. Therefore, the molar ratio of strands 1:2:3 = 0.95:0.95:1.0 was used for annealing.

3.3.2.2 Gap Filling Step Optimization

Commercially available polymerases were selected for evaluation with the criteria that the polymerase cannot have 5' to 3' exonuclease activity and strand replacement activity, otherwise strand 3 is liable for degradation during reaction. In addition, polymerases with A-tailing ability were also avoided to minimize the alteration of DNA other than at the position of interest.

Many commercially available enzymes satisfy these criteria, such as T4 polymerase, Sulfolobus DNA Polymerase IV and Bst Large Fragment Polymerase. T4 polymerase was chosen as it can also function robustly at a 12°C incubation temperature, the lowest working temperature amongst all the commercially available polymerases that satisfy the selection criteria. The low working temperature is favoured in order to reduce the possible strand displacement by partial dsDNA denaturation.

The ability of T4 ligase to incorporate each DNA modifications was evaluated by separate primer extension experiments with a dNtp mixture containing dCtp, mdCtp, hmdCtp, fdCtp or 5-carboxy-2'-deoxycytidine-5'-triphosphate (cadCtp) as illustrated in Figure 3-38. After the primer extension experiments, all reaction mixtures produced DNA with the same migration as the control

DNA generated by PCR, while the control experiment without any dNtp input did not show any band near the target product, indicating the band observed in previous lanes are not the annealed product of template strand and primer before polymerase and ligation activity. Thus T4 polymerase is able to incorporate these modifications (Figure 3-38).

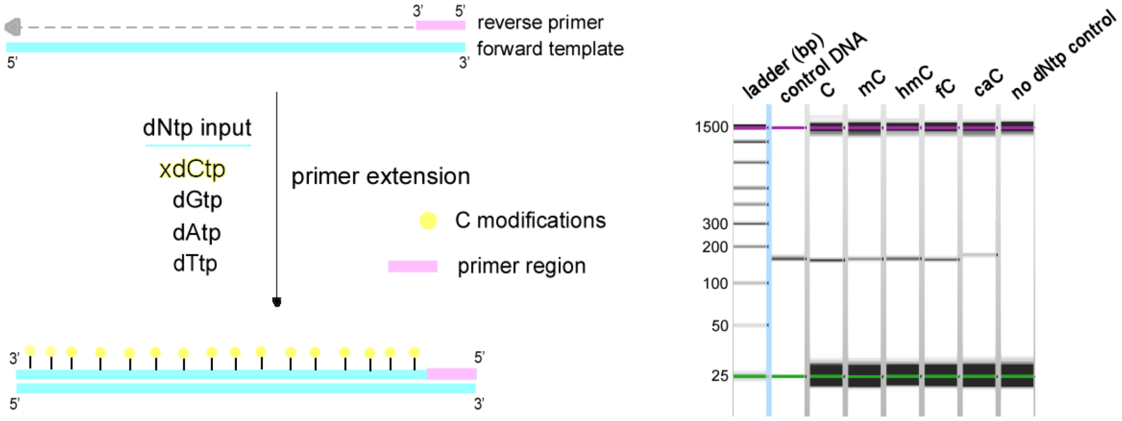


Figure 3-38. (left) Schematic illustration for the primer extension experiment to demonstrate the ability of T4 polymerase to incorporate xdCtps. (right) Tapestation image showing that T4 polymerase is able to incorporate various modified nucleotides in primer extension experiments. Control DNA was generated by PCR with 1% modification, and no dNtp control was generated by the control primer extension experiment without dNtp.

3.3.2.3 Ligation Step Optimization

As T4 polymerase does not incorporate ATP used in T4 ligase buffer (according to manufacturer testing), and the buffer conditions and incubation temperature for both enzymes are quite compatible, the GL reaction with non-biotinylated strand 2 can be completed in a one-pot reaction with incubation at 12°C to avoid an additional purification step and reduce hands-on time for preparation. However, since T4 polymerase has strong 3’ to 5’ exonuclease ability that was only suppressed by low temperature[354], it is more sensible to remove the T4 polymerase as soon as possible to avoid DNA degradation and a subsequent drop in yield.

Separation of T4 polymerase from DNA can easily be achieved using biotinylated strand 2 where, immediately after gap filling reaction, DNA anchored to magnetic streptavidin beads through a biotin label can be quickly separated from T4 polymerase by a magnet. The T4 ligase can then be added and incubated for a longer period to push the ligation to completion.

3.3.3 Method Validation with Short and Long DNA Sequences

3.3.3.1 Validation with Short DNA Sequences

The workflow was first validated using five short sequences (28-35 bp, Figure 3-39 for the sequences near the gap position). Different sequence contexts flanking the gap were explored to examine the compatibility of the GL method with mC, hmC and fC incorporation. The annealing, gap filling and ligation steps were performed using the optimized conditions.

name	sequence near gap (5' to 3')
set f	CC C GGTG
set g	CT C AATT
set h	CA C GTAC
set i	CC C TAGT
set j	GT C AGATA

Figure 3-39. Different base environments investigated for short sequences. The gap position is highlighted in red. The full sequences are listed in Table 5-2.

The products obtained from GL treatment were resolved using LC-MS to confirm the DNA identity by elution time monitored with UV and mass by MS (Figure 3-40). It was observed that the elution time of the peak changed upon GL treatment compared to starting materials, suggesting the formation of a new product. By calling peaks with predicted mass values (MW for target product for set h is 10433.8, n=10 fragmentation peak mass for correct product is 1042.4, for mis-ligated product detailed in Section 3.3.3.3 is 1013.5), a prominent peak was detected with the correct product mass (M^{10-} ion) at the migration position of GL product, while the mass corresponding to the "mis-ligated" product was not detected. Therefore, the GL method worked well in the short sequences studied.

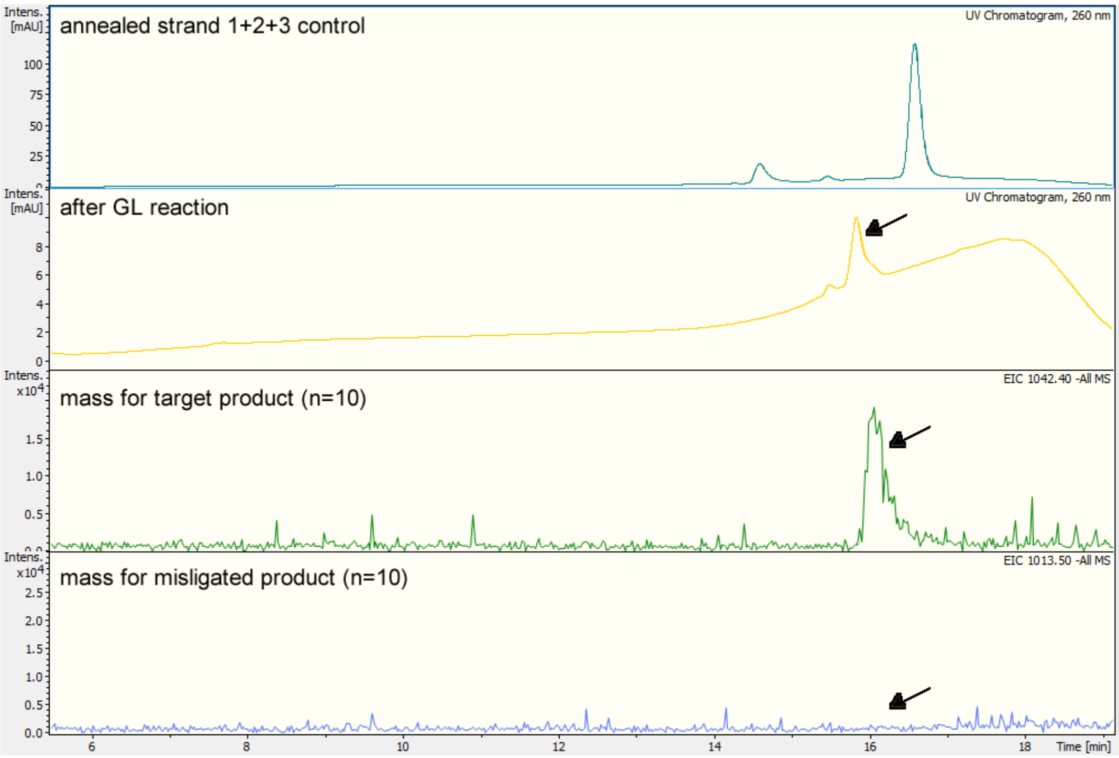


Figure 3-40. Examples of LC-MS UV trace and peak calling for sequence set 'h' with C incorporation. Annealed strand 1+2+3 is a control without GL treatment. Peak calling was performed with both predicted mass value of the correct product and the product ligated without gap filling (MW for target product for set h is 10433.8, n=10 fragmentation peak mass for correct product is 1042.4, for mis-ligated product is 1013.5).

3.3.3.2 Validation with Long DNA Sequences

Widom DNA with a single mC, hmC, fC or caC modification at various positions was prepared for further validation of the GL method and for future studies on nucleosome occupancy and stability. Modification positions were chosen within the sequence based on the positions identified from PolStop data in Section 3.2.12.2. Only C sites in a CpG context (Figure 3-41) were studied due to their biological relevance in mammals.

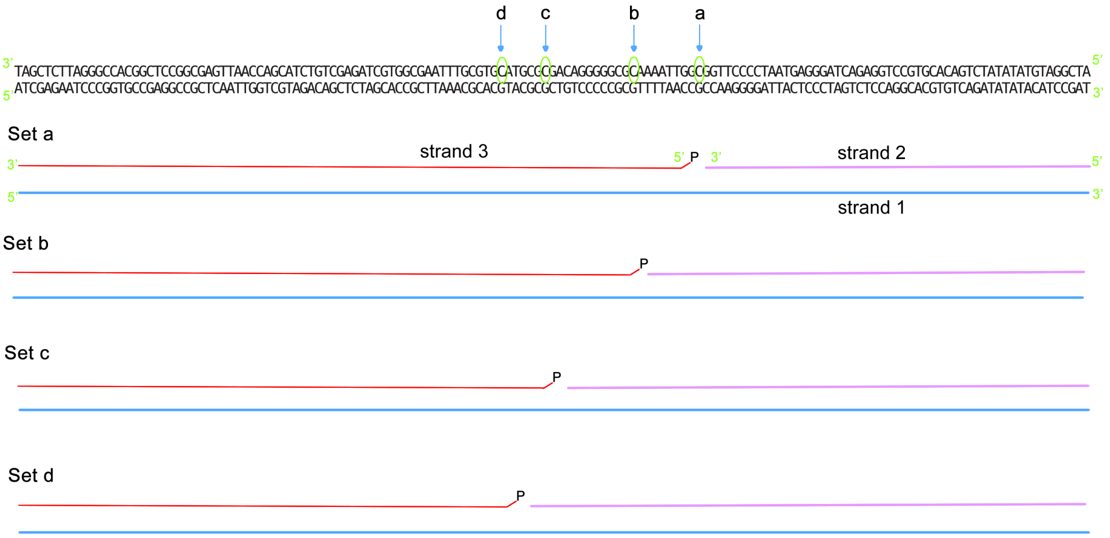

d
c
b
a
3'
TAGCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATCTGTCGAGATCGTGGCGAATTTGCGTGCATGCGCGACAGGGGGCGCAAAATTGGCGGTTCCCCTAATGAGGGATCAGAGGTCCGTGCACAGTCTATATATGTAGGCTA
5'
5'
ATCGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAGGGGATTACTCCCTAGTCTCCAGGCACGTGTCAGATATATACATCCGAT
3'
Set a
3'
strand 3
5'
P
3'
strand 2
5'
5'
3'
strand 1
Set b
P
Set c
P
Set d
P

Figure 3-41. Sequence Design for incorporating modified C in the Widom sequence at various CpG positions. The full sequences are listed in Table 5-3.

3.3.3.3 Validation by Gel Electrophoresis

The experiments were carried out using the optimized conditions. Because 147 bp DNA is too long to be directly characterized with LC-MS, the products were first visualized by denaturing gel electrophoresis to confirm the success of the reaction.

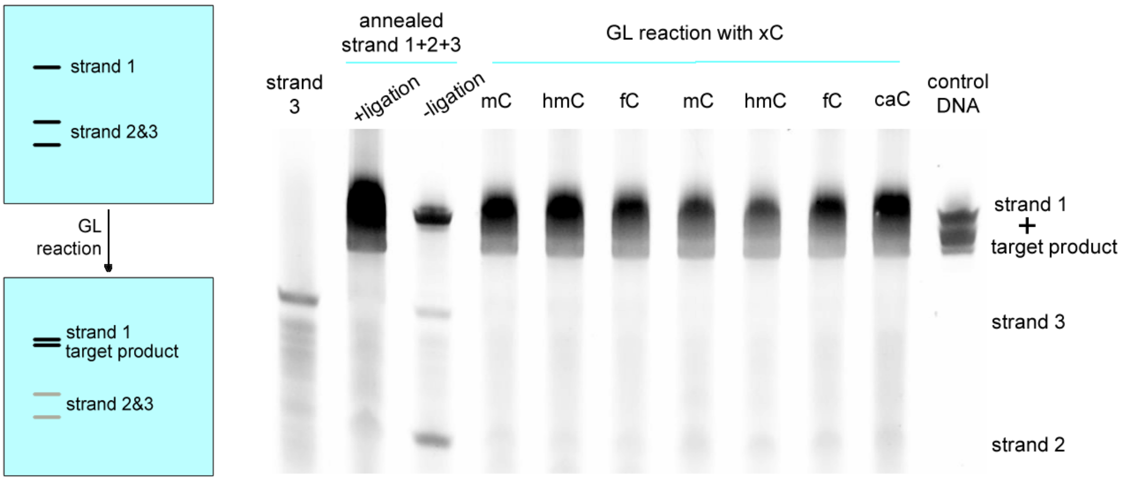


Figure 3-42. Monitoring the GL expt progress with denaturing gel electrophoresis stained by 1XSybr Gold (example shown is set a). Control DNA was generated by PCR.

Successful gap filling and ligation covalently joins strands 2 and 3, therefore the experiment progress can be monitored through the relative band intensity changes using denaturing gel electrophoresis as demonstrated by Figure 3-42. As strand 1 and the target product (complementary strand of strand 1) have a very similar migration position after denaturation, it is ambiguous to directly conclude whether the reverse strand has been produced successfully;

it can only be inferred by the relative band intensity reduction of strands 2 and 3.

Control experiments were performed where annealed strands 1+2+3 were directly ligated without the gap-filling step (Figure 3-42, “annealed strand 1+2+3” lanes “+ligation” compared to “-ligation”). Surprisingly, the bands for strands 2 and 3 also disappeared similar to the lanes of GL reactions, suggesting that T4 ligase is capable of ligation over the 1-nt gap to give a mis-ligated product. This phenomenon was also observed in the literature[355] and by the manufacturer. Therefore the denaturing gel electrophoresis can only provide information about the success rate of the ligation step but cannot differentiate the correct product from the mis-ligated product due to the similarity in product length. Thus NGS was used to further evaluate the error rate for mis-ligated product formation.

3.3.3.4 Validation by NGS

Since the gap filling step by T4 polymerase and ligation step by T4 ligase are independent of each other, there are four possible products (summarized in Figure 3-43). The efficiency of the ligation step was evaluated by denaturing gel electrophoresis; NGS was then used to measure the efficiency of the gap-filling step. The GL reaction was performed with biotinylated strand 2 attached to streptavidin beads, and the product was denatured by 0.1 M sodium hydroxide to break all H-bonds. Therefore, strand 1 and unligated strand 3 were washed away as they were not covalently bonded to the biotinylated strand 2. This excludes the interference from strand 1 in the subsequent PCR to the efficiency evaluation for the gap-filling step. For the library preparation for sequencing, PCR was used to generate blunt ended dsDNA. Only one type of xdCtp was added for each reaction, therefore a C reading from NGS is enough to confirm the success of gap filling without the need to use BS/redBS/oxBS-seq. Analysis of the DNA sequence from NGS data provides information about whether the gap position was filled with the desired C modifications (correct) or was mis-ligated (deletion). Since there is only 1-nt difference between these two products types, PCR biases are expected to be minimal, and the number of reads for each type of product represents the success and failure rate for the gap-filling step.

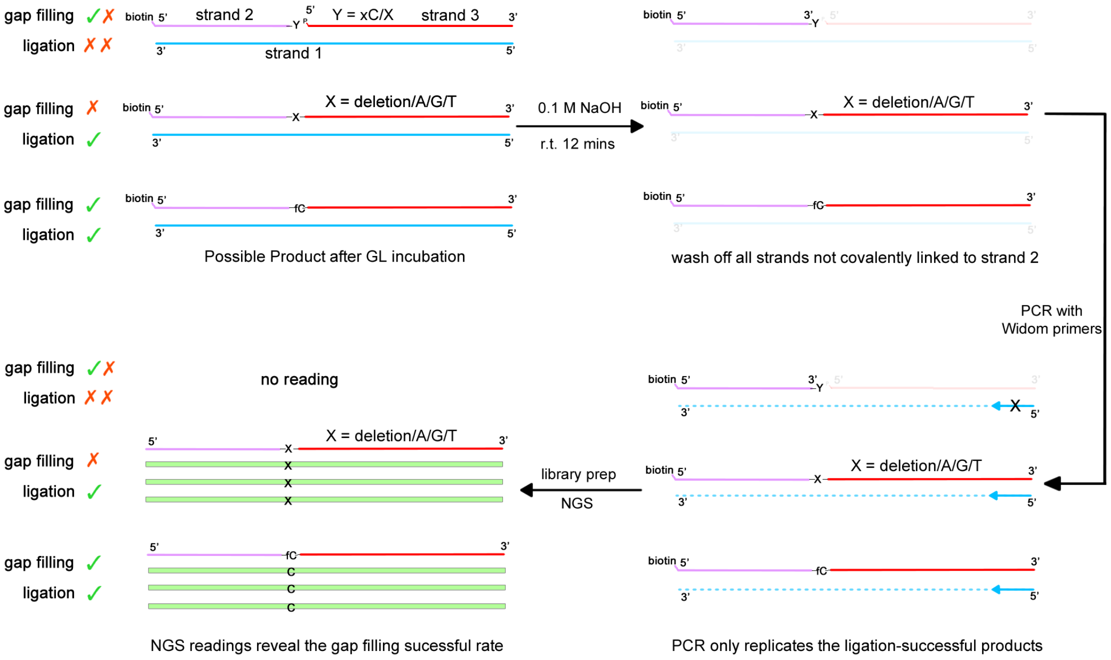


Figure 3-43. Schematic illustration for evaluating the efficiency of gap filling by NGS.

The NGS data showed that the percentage of mis-gap product (deletion) is mostly below 2% for mC, hmC and fC, which should be satisfactory for most applications of this approach. caC however showed 5.3% to 8.6% mis-ligation rates for sets B, C and D, indicating that T4 polymerase takes longer to incorporate cadCtp into the gap position so that the ligation before the gap-filling happened more frequently. Similar phenomena have been observed during previous PCR experiments to prepare caC-containing DNA. DreamTaq and Taq polymerases, although able to incorporate dCtp, mdCtp, hmdCtp and fdCtp, could not incorporate cadCtp to produce the PCR product with standard PCR conditions. A special polymerase, KOD XL, had to be used for preparing caC containing DNA.[356]

Gap-insert xC	number of reads		percentage of total reads (corrected with baseline error rate)			
	total	correct	correct	deletion	mutation	insertion
setA-mC	78983	78238	99.06	0.56	-0.06	-0.01
setB-mC	68148	67763	99.44	0.28	-0.19	0.12
setC-mC	88643	86632	97.73	1.77	0.01	0.01
setD-mC	100482	98558	98.09	1.24	-0.12	0.00
setA-hmC	61594	60902	98.88	0.71	-0.06	0.03
setB-hmC	79085	78495	99.25	0.43	-0.18	0.19
setC-hmC	64545	62837	97.35	2.16	0.03	0.08
setD-hmC	85025	82751	97.33	1.90	-0.09	0.05
setA-fC	31629	31250	98.80	0.81	-0.09	0.04
setB-fC	57924	57432	99.15	0.47	-0.14	0.09
setC-fC	92236	89987	97.56	1.92	0.06	0.00
setD-fC	99335	97250	97.90	1.25	-0.07	0.01
setA-caC	72533	71247	98.23	1.17	0.02	0.03
setB-caC	78531	73574	93.69	5.32	-0.01	0.11
setC-caC	193519	176363	91.13	8.15	0.15	0.07
setD-caC	73372	65173	88.83	8.64	0.08	0.07

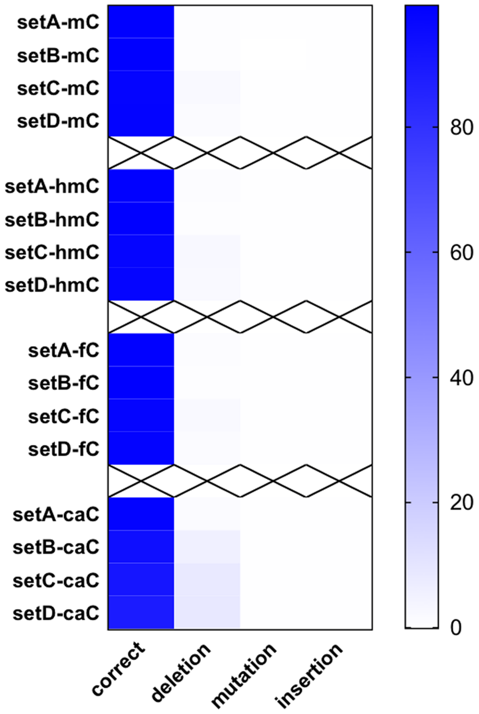


Figure 3-44. NGS results summary for set a to D with mC, hmC, fC and caC modifications.

The bioinformatic analysis also revealed that in addition to the target product and mis-ligated product (deletion), there are a small number of random mutations and insertions to the DNA sequences that may arise during the PCR amplification and sequencing (Table 6-1). A baseline error rate was constructed to compare the error rates for the gap positions with the rest of positions (Table 6-2). As positions other than the gap position were synthesized by solid-phase phosphoramidite chemistry, and theoretically should be error-free, the deletion/mutation/insertion rates from NGS readings were calculated as the baseline error rate. Corrected with the baseline error rate obtained from the same data set, the percentage of mismatches and insertions are both close to 0% (Figure 3-44).

3.3.4 Conclusion and Future Work

A simple, fast and reliable method for generating DNA with single C derivative modifications at designated positions has been developed (the comparison with other methods is summarized in Table 3-3), and the robustness of this method has been demonstrated on both short sequences with various sequence contexts and Widom sequences at different positions. With the current sequences explored, limiting conditions such as sequence context were not observed. This method will be used to prepare DNA with modifications at precise loci to scrutinize their effect on nucleosome occupancy and stability rather than measuring an average effect. This method is especially suitable for the studies on comparing the effect of different DNA modifications in the same sequence context. As opposed to solid phase synthesis that requires separate preparation of all strands with high cost, the GL method only requires preparation of a common set of strands 1, 2 and 3, and the sequences with different modifications can then be easily prepared in parallel.

	GL Method	Solid Phase Synthesis	Enzymatic Modification	PCR
Base Modifications Compatibility	limited by the availability of modified deoxynucleotide triphosphate and suitable polymerase	limited by the availability of properly protected modified phosphoramidite	limited by the availability of modifying enzyme	limited by the availability of modified deoxynucleotide triphosphate and suitable polymerase
Incorporation Position	Precise	Precise	depending on sequence context	ambiguous
Single Product?	Yes	Yes	depending on sequence context and enzyme specificity	depending on Modification Density
Sequence Context Requirement	No	No	Yes	No
Cost	Low	High	Middle	Low
Hands-on time	Short	Long	Long	Short
Condition	Mild	Harsh	Mild	Mild
Yield	High	Low	High	High

Table 3-3. Comparison between the developed GL method and other methods used to introduce DNA modifications.

A potential shortcoming of the GL method is that mis-ligated products, although at a very low percentage (< 2.2% for mC, hmC and fC, < 8.7% for caC), are difficult to separate from the target product. A few strategies can potentially tackle this problem. In GL reactions with non-biotinylated strand 2, polymerase extension and ligation can be carried out in two separate reactions to ensure no ligation can occur before the gap filling takes place. Alternatively, a Taq ligase[357] can be explored to help eliminate this issue. This ligase cannot ligate over the 1-nt gap[358, 359], and therefore ensures the ligation only occurs after the gap-filling step is completed. The failed product in this case will always be significantly shorter than the target product, and thus much easier to purify.

In the future, it would be interesting to study the effectiveness of this GL method in other sequence contexts. This method can be expanded to other modifications such as T derivatives, as long as the modified nucleotide triphosphates and suitable polymerases are available. This method can also be adapted to the preparation of DNA with mixed DNA modifications at multiple positions. In addition, with wider gaps, this method allows the preparation of DNA with a patch of dense modifications at various lengths. The length of DNA that the GL method can directly prepare is restricted only by the length of DNA required to achieve correct annealing and the cost of solid phase synthesis to prepare strands 1, 2 and 3. This issue could be circumvented by combining the GL method with restriction enzymes, widening the range of DNA lengths this method can be applied to.

3.4 Conclusion

In summary, it was shown that mC, hmC and fC modifications increased nucleosome occupancy and stability in an increasing order, and that even low-density modifications are sufficient to induce such changes. With mC and hmC present, fC preferentially occupied the positions facing towards the histone in the forward strands with a slight preference for minor groove positions, while hmC mainly occupied the positions on the first half of reverse strand further away from histone than fC with a preference for major groove positions, and mC primarily dominated the positions furthest away from histone core on the reverse strand with a preference for major groove positions. This provides an interesting pattern for nucleosome positioning in DNA with mixed C derivatives. Without mC and hmC present, fC was enriched in both forward and reverse strands of DNA both facing towards and away from histone core compared to the control DNA, indicating combined mechanisms of promoting nucleosome formation. Several possible factors could contribute towards the promoting effect of fC (summarized in Figure 3-45), and they were examined separately. The results revealed that independent of chaperones, the effects of fC on the biochemical and biophysical properties of the nucleosome.

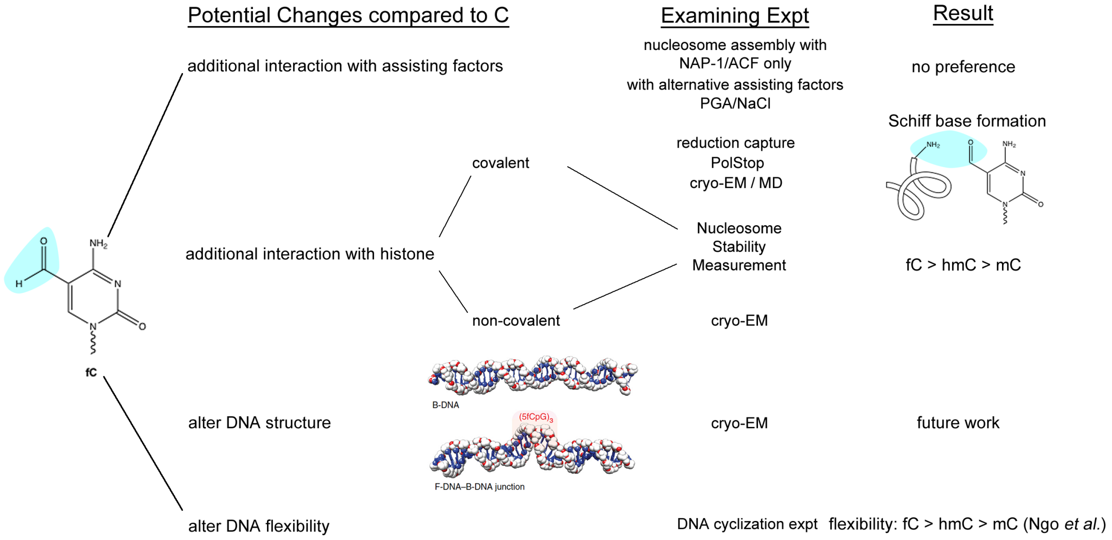

Potential Changes compared to C
Examining Expt
Result
additional interaction with assisting factors
nucleosome assembly with NAP-1/ACF only
with alternative assisting factors PGA/NaCl
no preference
Schiff base formation
covalent
reduction capture
PolStop
cryo-EM / MD
O
H
NH2
N
N
O
fC
additional interaction with histone
Nucleosome Stability Measurement
fC > hmC > mC
non-covalent
cryo-EM
B-DNA
(5fCpG)3
F-DNA–B-DNA junction
alter DNA structure
cryo-EM
future work
alter DNA flexibility
DNA cyclization expt
flexibility: fC > hmC > mC (Ngo et al.)

Figure 3-45. Summary of hypotheses for the potential changes caused by fC contributing to the increased nucleosome formation, and the proposed experiments to examine the hypotheses. The F-DNA structure figure is taken from Raiber *et al.*[37]. DNA cyclization experiment result reported by Ngo *et al.*[111].

The link between fC and nucleosomes provides a new conceptual foundation for the understanding of nucleosome formation and positioning by DNA modifications. Upon formation, the Schiff base anchors the DNA to the histone and stabilizes the nucleosome. The fC positions identified with a high probability of Schiff base formation showed a clear pattern of periodicity. The spatial relationship with histone proteins and the exact amino acid residues cross-linked to fC may be identified by high-resolution cryo-Electron Microscopy (cryo-EM) or Molecular Dynamics (MD) simulation. Well-positioned nucleosomes may impede the access of protein factors, or mark regulatory elements, or recruit pioneer factors to initiate transcription. Since fC peaks in mESCs and embryonic tissues overlap with active histone markers such as H3K27ac and H3K4me1[164], it may be plausible that fC promotes and stabilizes nucleosomes at necessary positions and influences protein factors recruitment for the regulation of transcription. Some preliminary proteomics work has been performed to explore this hypothesis (summarized in Section 4.2.3).

In addition, fC may cause biophysical changes to the nucleosomal DNA in terms of DNA structure and flexibility, similar to the effect observed with dsDNA detailed in Section 2.1.1 and 3.1.5[37, 111]. In the restrained environment of the nucleosome, the Widom DNA containing dense symmetrical fCpG modifications may form the intricate hydrogen-bonding network observed in F-DNA[37], and relax the nucleosomal DNA to enable more histone-DNA interactions. In addition, the nucleosomal DNA may become more flexible as observed with dsDNA[111], and allow extra interactions with histone proteins. These potential changes can be examined by comparing the high-resolution cryo-EM structure of fC-modified and unmodified nucleosomes (preliminary work shown in Section 4.2.2). These potential structural changes may also contribute to the increased nucleosome occupancy and stability, and may even influence the length of helical turns on the nucleosome, and further lead to changes in DNA packaging in cells with clustered fC modifications.

4 Reflection and Further Questions

4.1 Reflection

The focus of this thesis has been on the effect of naturally occurring DNA modifications on duplex DNA and nucleosomes. Several interesting phenomena were observed during this research and are worthy of further investigation.

The systematic studies of T derivatives by CD spectroscopy in Chapter 2 revealed that the presence of symmetrical dense fU modifications, but not other T derivatives, in a DNA duplex is able to produce a similar structural change to the previously reported for fC-containing DNA duplex[37]. All fU-containing complementary duplexes investigated showed a characteristic negative band around 300 nm, which is likely attributable to the fU:A base-pair.

The systematic *in vitro* evaluation of the effects of naturally occurring DNA modifications on nucleosome in Chapter 3 suggested that fC modification is apt to increase nucleosome occupancy and stability regardless of modification density. Additionally, it has been shown that this effect is the result of intrinsic changes to the nucleosome upon fC modifications, independent of chaperones. The fC positional preference study (Section 3.2.10) showed that fC was enriched in all groove positions, facing both towards and away from the histone core, suggesting multiple mechanisms of promoting nucleosome formation. Possible mechanisms include the Schiff base formation between fC in the major grooves and primary amines from histone tails. Upon formation, the Schiff base anchors the DNA to the histone and stabilizes the nucleosome. Possible biophysical changes including increased dsDNA flexibility[111] and potential nucleosomal DNA structural changes similar to those previously observed in dsDNA[37] could also contribute to the promoting effect of fC.

The changes to dsDNA and nucleosome properties induced by naturally occurring C and T derivatives advance our understanding of the role DNA modifications play in reshaping chromatin architecture and thus regulating cellular activities such as DNA replication and transcription. These results encourage the pursuit of further insights into their effect and functions through structural and proteomics studies.

4.2 Preliminary Studies to Address Further Questions

4.2.1 Towards Obtaining a High-Resolution Structure of ODN3-fU dsDNA

The high resolution structure of ODN3-fU needs to be elucidated along with its unmodified counterpart as the control to confirm the structural similarity between ODN3-fU and F-DNA[37] observed in CD spectra, and also to understand the changes in intramolecular interactions caused by fU modifications resulting in the structural deviation from B-form DNA. Previous studies have demonstrated that several protein factors can be preferentially recruited by fU, therefore the high-resolution dsDNA structure may also provide structural insights for specific recognition.

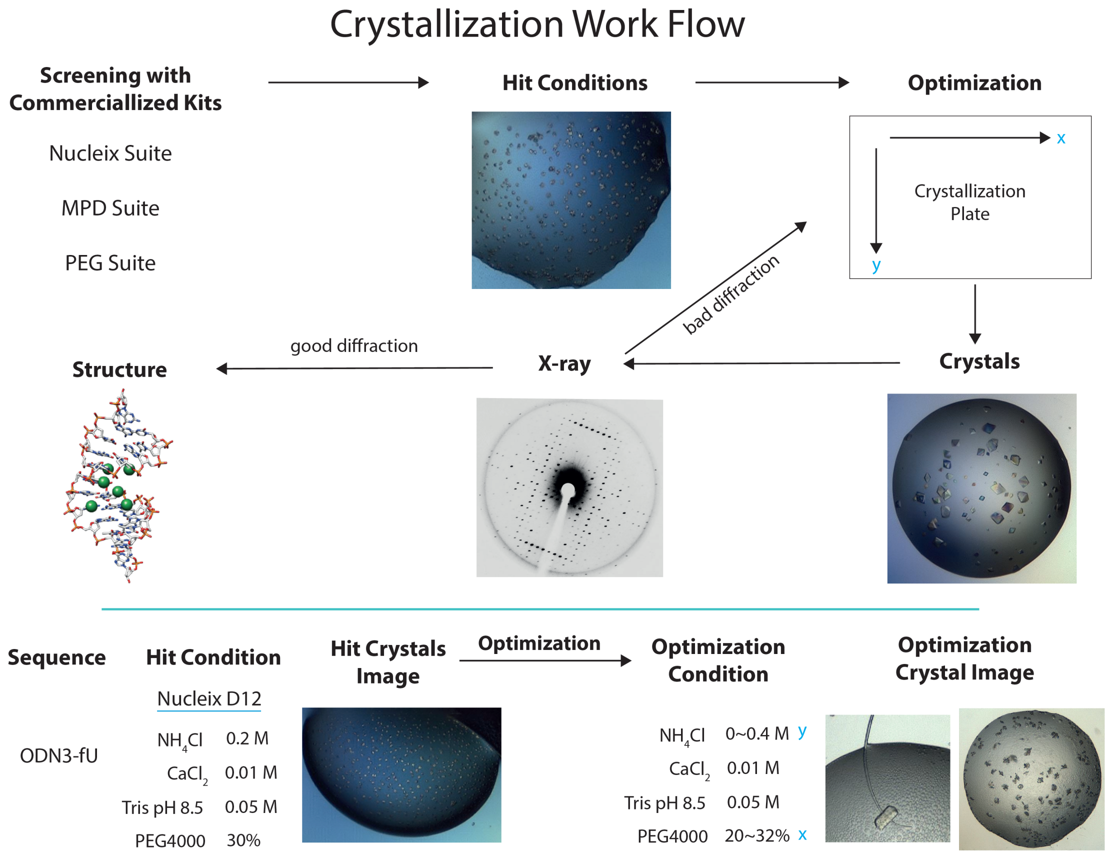


Figure 4-1. Workflow of X-ray crystallization and examples of initial conditions and current conditions after optimization. The optimization was done in 96-well plates, where the concentration of precipitants from the initial condition was gradually varied in both the horizontal (x) and vertical (y) directions of the crystallization plate.

X-ray crystallography involves incubating DNA molecules with precipitants to aid crystallization by promoting the orderly assembly of DNA molecules. The X-ray diffraction patterns of DNA crystals are solved to produce high-resolution structures[27, 360, 361]. Following an initial screen of crystallisation conditions (Figure 4-1), further optimization is required in order to obtain sufficiently large crystals for high-resolution structure determination.

4.2.2 Towards a High-Resolution Structure of the fC-Nucleosome

(DNA and nucleosomes were prepared by me; grid preparation, imaging and structural determination was done by Dr Ben Luisi, Dr Dimitri Chirgadze and Dr Kotryna Bloznelyte, Department of Biochemistry, University of Cambridge)

The observed positive effect of fC modifications on nucleosome formation and stability prompted me to initiate a structural study to understand the basis of these effects. Cryo-EM was utilized to elucidate the high-resolution nucleosome structure by bombarding samples with an electron beam and detecting the interference between the electron beam and the sample. Initially, the nucleosome preparation conditions were optimized (summarized in Figure 4-2, detailed in Appendix Section 6.6) to produce large quantities of well-isolated fC-nucleosomes (Figure 4-3), which were imaged to solve the high-resolution structures of nucleosome (Figure 4-4) to enable the investigation of the changes imposed by fC on the nucleosome. Potential changes of particular interest include differential histone-DNA interactions, and nucleosomal DNA conformation changes compared to the unmodified nucleosome.

After reconstruction, the resolution of the fC-containing nucleosomes reached 3.1 Å, and the preliminary map of the fully modified fC nucleosome is shown in Figure 4-4. From the model, the nucleosome was found to be a disk-shaped molecule with right-handed DNA wrapping left-handed helically around the histone octamer in about 1.7 turns. The nucleosome has a diameter of 106 Å and a height of 65 Å. All these parameters agree with previous studies[49, 50], suggesting well-formed nucleosomes.

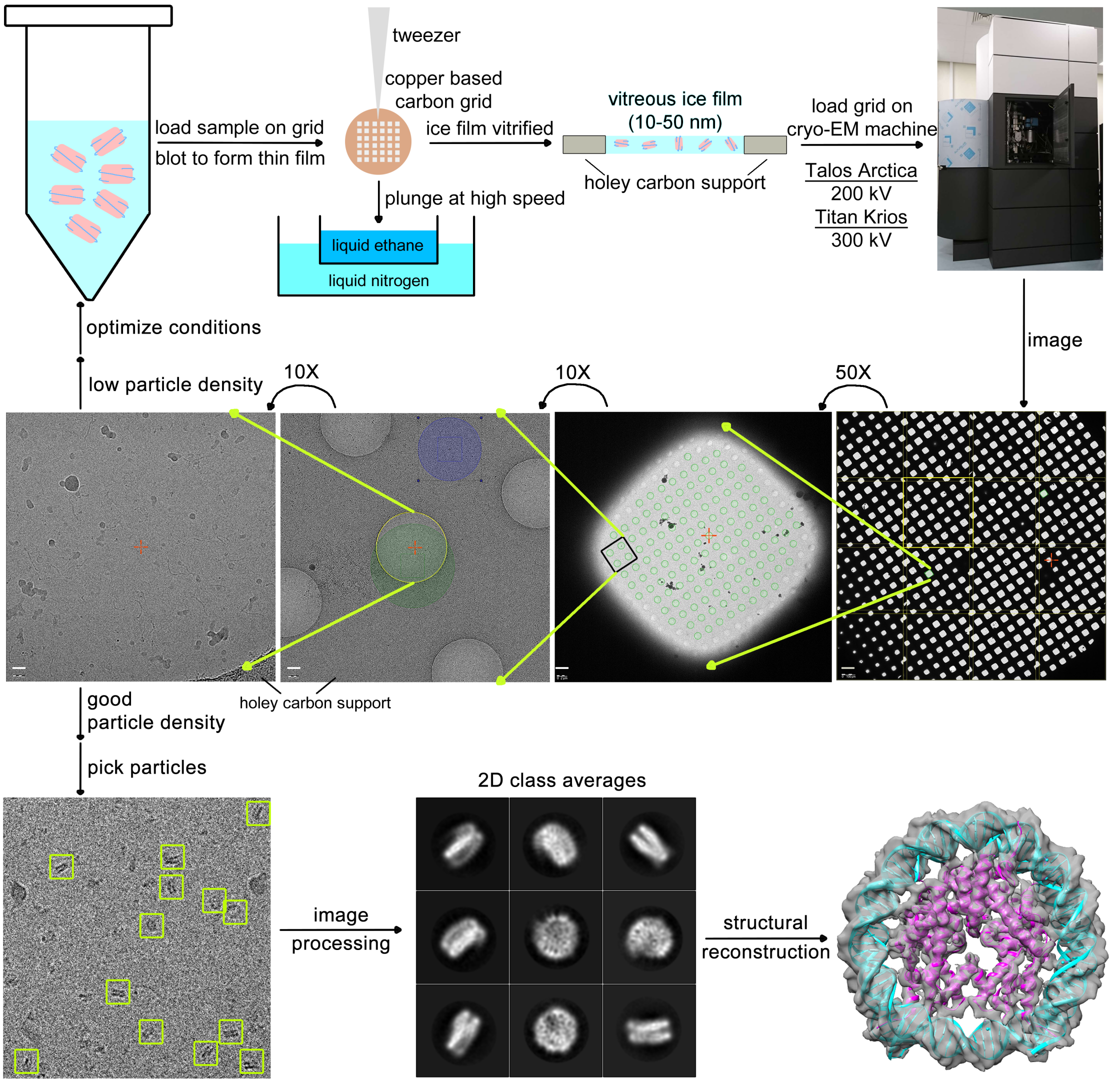
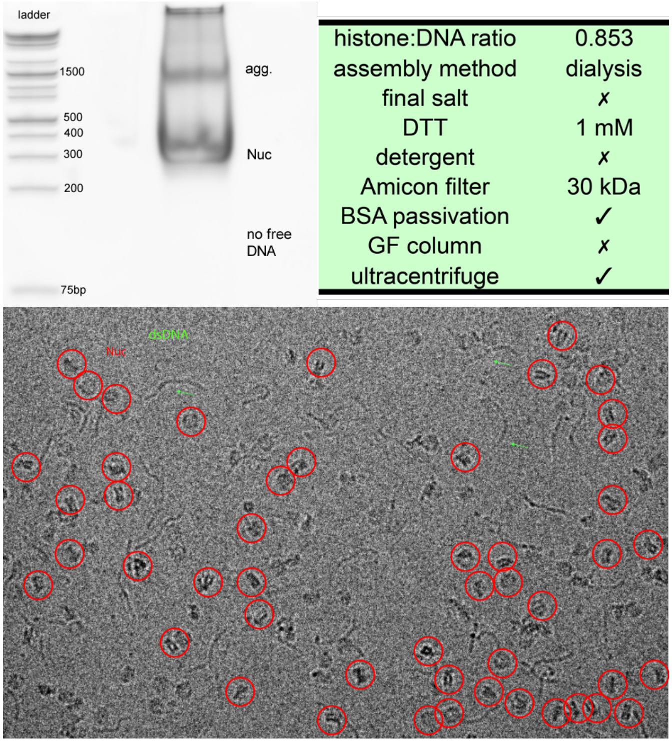


Figure 4-2. Illustration of the cryo-EM workflow from sample preparation to structure reconstruction. The second row of images depicts an example of a grid at sequentially higher magnification from right to left. The scale bars in the bottom left corner of each figure represent (from right to left) 100 µm, 2 µm, 200 nm and 20 nm, respectively.

The fC positions identified to form Schiff bases with the highest frequency will be analysed in the context of the high-resolution cryo-EM structure of the fC-Widom nucleosome to spatially identify the corresponding lysine residues. The movements of histone tails are not restricted in aqueous solution, and therefore invisible due to the nature of cryo-EM structure determination. However, if a Schiff base forms between DNA and histone tails, the histone tail will be restricted to one region with higher probability, resulting in (partially) visible histone tails. With high enough resolution, the Schiff base might be observed directly or indirectly through changes in density around fC sites. With the exact arrangement of DNA and histone proteins available, the fC positions identified were further analysed in aspects such as DNA helical parameters and spatial relationship between fC modifications and histone tails.



histone:DNA ratio	0.853
assembly method	dialysis
final salt	✗
DTT	1 mM
detergent	✗
Amicon filter	30 kDa
BSA passivation	✓
GF column	✗
ultracentrifuge	✓

Figure 4-3. Nucleosome produced by final condition: summary of final sample condition decided; and the results of nucleosomes checked by native gel electrophoresis and cryo-EM.

With the GL technique developed for incorporating DNA modifications at desired positions, it will now be interesting to prepare nucleosomes with single fC or short patches of consecutive fCpG modifications to elucidate how only a few modifications are enough to influence the nucleosome structure and interactions and cause the significant changes in nucleosome properties as presented in Section 3.2.8.

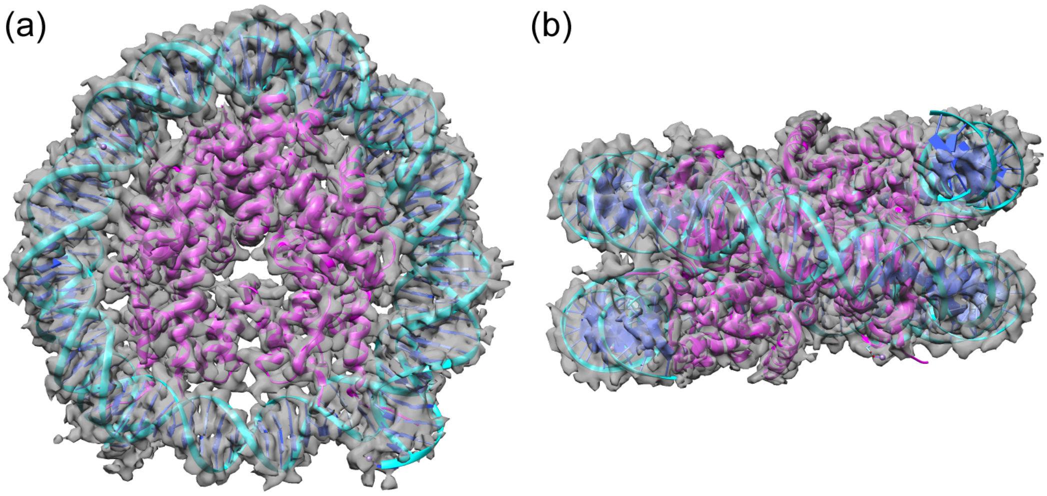


Figure 4-4. The model of nucleosome structure solved to 3.1 Å. (a) top view; (b) side view. The DNA is in cyan and the histone is in magenta.

4.2.3 Towards Identifying Protein Factors Specifically Interacting with Modified Nucleosome(s)

xC modifications have been demonstrated to cause changes to the nucleosome's biophysical and biochemical properties, which may influence protein machinery recruitment and add another layer for shaping chromatin architecture and regulating cellular activities.

Previous studies on the protein recruitment of xCs were mostly done with dsDNA carrying xC modifications, while only Bartke *et al.* studied the effect of DNA methylations on protein recruitment in the nucleosome context.[163, 166-168] Since chromatin remodelers and pioneer transcription factors have been shown to bind to nucleosomes[362, 363], it would be interesting to study how xC-modified nucleosomes can impact protein-nucleosome interactions to further understand their biological functions.

Due to the importance of linker DNA, it would also be interesting to study strings of several nucleosomes with both modified histone and modified linker DNA to see what chromatin remodelers can be specifically recruited.

In addition, as it was shown in Iurlaro *et al.*[164], genome-wide fC distribution overlaps nicely with active histone marks such as H3K4me1 and H3K27ac. Therefore, it would be interesting to see whether fC-nucleosomes containing active histone marks can specifically attract chromatin remodelers and transcription factors to regulate transcription. In our preliminary results for nucleosome pull-down with HeLa nuclear extract, we have noticed that fC-modified nucleosomes specifically attracted proteins related to histone acetylation such as NAT10[364, 365], while unmodified nucleosomes can specifically recruit HDAC1, which deacetylates the histone[366]. This preliminary proteomics data aligns with *in vivo* observations of fC colocalization with H3K27ac, and suggests the hierarchy of DNA formylation and histone modification generation. It would be interesting to follow up with strictly controlled *in vitro* biochemistry experiments to examine such directionality of DNA and histone modifications generation and removal.

5 Experimental Methodology

Table of Contents

All DNA concentration measurements in this thesis were quantified on a NanoDrop One Microvolume UV-Vis Spectrophotometer (Thermo Fisher Scientific, Massachusetts, USA). All the nucleoside and short oligo analyses were done by HPLC (Agilent Technology 1200 Series) or by LC-MS (Thermo Scientific Dionex Ultimate RS3000 LC coupled with Bruker AmaZonX MS, Massachusetts, USA). All PCR and incubation procedures were done with peqSTAR thermal cyclers (PEQLAB, Germany) unless otherwise stated. All water used in this study was ultrapure grade Milli-Q® water (MQ H_2O, Millipore, Merck, Germany). All values listed in this chapter are working concentrations/quantities.

5.1 Duplex DNA Study in Chapter 2

5.1.1 Oligo DNA Preparation

The fU- and hmU-modified sequences were ordered from ATDbio (Southampton, UK), and all other sequences were ordered from Invitrogen (California, USA), IBT (Germany) and Sigma-Aldrich (Missouri, USA). These sequences were synthesized by phosphoramidite chemistry[235] and purified by HPLC. Samples were dissolved in MQ H_2O to 1 mM (ssDNA concentration) stock solution and stored at -20°C.

5.1.2 UV Denaturation for Duplex DNA Stability Studies

The samples were prepared by mixing the forward and reverse strands in a 1:1 ratio at 5 μM (dsDNA concentration) in 10 mM PBS buffer at pH 7.2 (10 mM sodium phosphate salt, 137 mM sodium chloride and 2.7 mM potassium chloride) and 3 mM magnesium chloride. Samples were placed in quartz cuvettes with 1 cm path length and covered with 200 μL of mineral oil. UV melting was done using a Cary 100 UV-Vis Spectrophotometer (Agilent Technologies, California, USA). Samples were heated at 95°C for 10 min and annealed in the instrument from 95°C to 5°C at 10°C/min. After equilibrating at 5°C for 10 min, samples were heated to 95°C and cooled to 5°C at a rate of 1°C/min for 3 cycles, with data collected at every 1°C interval during both melting and cooling processes.

The baselines of melting curves were determined by treating the plateau at the lower temperature as 100% of the DNA strands being in duplex form (bottom), and the plateau at the high temperature as 0% of the DNA strands in duplex form (top). The melting temperature (Tm) of a given duplex, reflecting its thermal stability, was defined as the temperature where 50% of the DNA strands were in duplex form. The melting curves were fitted with the Boltzmann sigmoid function in Prism (GraphPad)[367] using the following equation to calculate Tm:

$$percent\ folding = bottom + \frac{top - bottom}{1 + e^{(\frac{Tm - temperature}{slope})}}$$

The melting experiments were performed in triplicate, and the reported Tms are the average of three experiments.

5.1.3 Circular Dichroism for Duplex DNA Structure Studies

The samples were prepared by mixing the forward and the reverse strand of duplex in a 1:1 ratio at 10 μM (duplex concentration) in 10 mM PBS at pH 7.2. Samples were annealed by heating to 75°C and slowly cooled down to 5°C over a period of 3 h. CD spectra were recorded with a Chirascan Plus spectropolarimeter (Applied Photophysics, UK) in quartz cuvettes with 0.1 cm path length. Scans were performed across the range of 200-350 nm at 20°C. Each sample was scanned three times with a step size of 1 nm, time per point of 1 s and a bandwidth of 1 nm; the three scans were averaged to provide the final result. The spectrum of pure buffer was subtracted from the final result, and baseline corrected at 320 nm.

5.1.4 X-Ray Crystallography for Duplex DNA Structure Studies

The oligonucleotides for crystallization were desalted using PD-10 columns (GE Healthcare Life Sciences, New Jersey, USA) and annealed by heating to 95°C for 5 min and slowly cooling down to room temperature over a period of 3 h. Crystallization was done using the vapour-diffusion sitting-drop method in 96-well MRC 2-drop crystallization plates (Molecular Dimension, UK). Each drop was made up with 200 nL of 1 mM DNA solution (strand concentration) mixed with 200 nL reservoir solution from commercialized screening kits (Nucleix Suite, MPD suite and PEG Suite (Qiagen, California, USA)) using a Mosquito Crystallization Robot (TTP Labtech, UK). The crystallization plates were incubated at 20°C, and monitored with a Rock Imager 1000 (Formulatrix, Massachusetts, USA).

5.2 Nucleosome Study in Chapter 3

5.2.1 Nucleosome Occupancy Comparison

5.2.1.1 Nucleosomal DNA Preparation

DNA for nucleosome assembly experiments was generated by standard PCR amplification using Taq polymerase (NEB, Massachusetts, USA) for C-, mC-, hmC- and fC-modified DNA, and KOD XL Polymerase (Novagen, Wisconsin, USA) for caC-modified DNA. For a 50 µL reaction, the template (0.9 ng, 0.02 pmole) was mixed with deoxynucleotide triphosphates (dNtps, 200 µM), primers (0.5 µM) and polymerases (2.5 U). For Widom DNA, the annealing condition was 55.5°C for 30 s; the elongation condition was 68°C for 30 s. After 30 cycles, the reaction mixture was slowly cooled down to 4 °C. For HS DNA, the annealing condition was 64.5°C for 30 s; the elongation condition was 68°C for 30 s. After 29 cycles, the reaction mixture was snap cooled to 4°C to avoid the formation of unwanted concatenated product.

The deoxynucleoside triphosphates used were unmodified dNtps (Thermo Scientific) and/or modified deoxycytidine triphosphates or modified deoxyuridine triphosphates (TriLink, California, USA). The DNA template and primers were synthesized by Sigma Aldrich (Missouri, USA) and IBA (Germany). The sequence of template and primers are as shown in Table 5-1.

Sequence Name		Sequence (5' to 3')
widom	template	ATCGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAAACGCACGTACGCG CTGTCCCCCGCGTTTTAACCGCCAAGGGGATTACTCCCTAGTCTCCAGGCACGTGTCAGATATATACATCCGAT
	forward primer-Cy3	(Cy3)-ATC GAG AAT CCC GGT GCC GA
	forward primer-biotin	(bio)-ATC GAG AAT CCC GGT GCC GA
	reverse primer-Cy5	(Cy5)-ATC GGA TGT ATA TAT CTG ACA CGT GCC TGG AGA
	forward primer	ATC GAG AAT CCC GGT GCC GA
	reverse primer	ATC GGA TGT ATA TAT CTG ACA CGT GCC TGG AGA
HS	template	ATCAATATCCACCTGCAGATTCTACCAAAAGTGTATTTGGAAACTGCTCCATCAAAAGGCATGTTCAGCTGAATTC AGCTGAACATGCCTTTTGATGGAGCAGTTTCCAAATACACTTTTGGTAGAATCTGCAGGTGGATATTGAT
	primer	ATC AAT ATC CAC CTG CAG ATT CTA C

Table 5-1. Sequence table of all sequences and their primers used in this study.

For fully modified DNA, the dCtp was fully replaced by the respective xdCtp; for partially modified DNA, the PCR was done with a mixture of dCtp and xdCtp, with the ratio determined by the calibration curve specified in Figure 3-16. The 1% xC-modified DNA for mC, hmC and fC, was generated with the input percentage of 2% mdCtp, 10% hmdCtp and 3.4% fdCtp respectively, with the rest being dCtp. The percentage of xC incorporation was confirmed by LC-MS/MS.

After PCR synthesis, reaction mixtures were purified using the GeneJet PCR Purification Kit (Life Technologies, California, USA), and eluted with MQ H_2O. The DNA was quantified by Nanodrop. The DNA quality was checked using a Tapestation 2200 (Agilent, California, USA) with D1000 tapes and reagents, where the DNA is pre-stained with an intercalating dye (Sybr Green I), separated on a gel tube composed of agarose and polyacrylamide gel and imaged to show the migration pattern and purity.

5.2.1.2 LC-MS/HPLC Condition for xC Incorporation Quantification

The DNA (500 ng) was digested into nucleosides with 2.5 U DNA Degradase Plus (Zymo, California, USA) in a 25 µL reaction by incubation at 37°C (block) 45°C (lid) for 3 h in a peqSTAR thermal cycler, and cleaned using an Amicon Ultra-0.5 Centrifugal Filter Unit (10 kDa) to remove the degradase. The solution in the collection tube was separated on a Pursuit 5 C18 column (150 x 4.6 mm, Agilent) by HPLC monitored at 280 nm or LC-MS monitored at 260 nm. Solvent A (10 mM ammonium formate in MQ H_2O) and B (acetonitrile containing 0.1% formic acid) were used for the nucleoside separation. The percentage of B is shown in Figure 5-1. Flow rate is 1.5 mL/min.

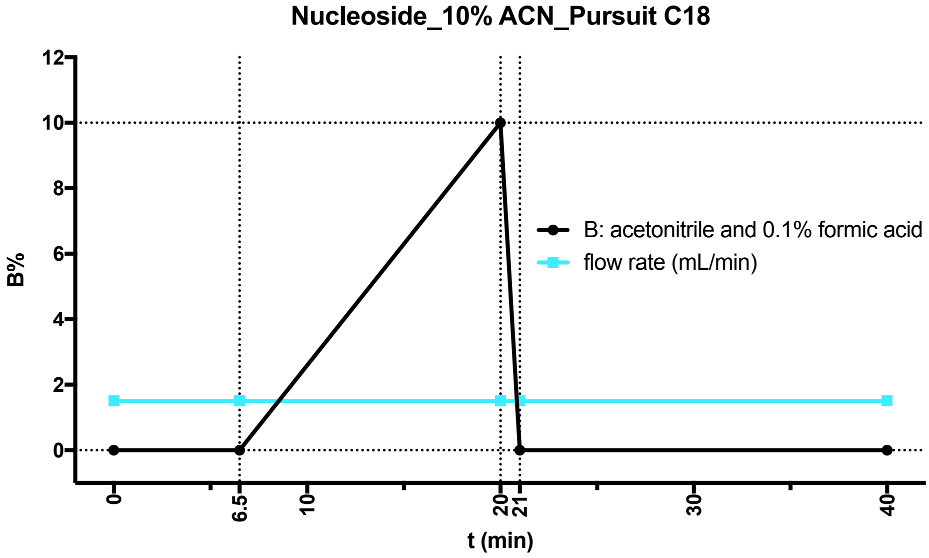


Figure 5-1. LC-MS and HPLC condition to separate nucleosides for xC incorporation quantification. Solvent A (10 mM ammonium formate in MQ H_2O) and B (acetonitrile with 0.1% formic acid) were used. Flow rate is 1.5 mL/min.

5.2.1.3 Nucleosome Preparation

5.2.1.3.1 Assisted by Biological Assisting Factors NAP-1 and ACF

All chaperones and histone proteins were from the Chromatin Assembly Kit (Active Motif, Belgium). A master mix (MM) of all the assembly components except DNA was prepared following the manufacturer's instructions with some modifications and aliquoted into individual tubes before adding DNA to avoid condition variance and handling small reagent volumes. For a 15 µL assembly 1.5 µL High Salt Buffer was incubated with 0.21 µL h-NAP-1 and 0.27 µL HeLa Core Histones. The mixture was incubated on ice for 30 minutes before the addition of 3.65 µL Low Salt Buffer, 0.38 µL ACF complex and 1.5 µL freshly prepared complete 10X ATP Regeneration System. Complete 10X ATP Regeneration System was prepared by mixing 0.1 µL Creatine Kinase with 1.65 µL 10X ATP Regeneration System. The mixture was gently agitated after addition of each component. DNA (100ng) was diluted with MQ H_2O to 7.5 µL, mixed with 7.5 µL master mix and incubated at 27°C (block) and 50°C (lid) overnight in a peqSTAR thermal cycler. The reaction was then cleaned up using a Bio-Spin 6 Tris column (P6 column, BioRad, California, USA). For competing nucleosome formation, the DNA carrying different modifications and fluorescence labels was mixed in equal amounts and incubated with master mix.

5.2.1.3.2 Assisted by PGA

Poly-L-Glutamic Acid sodium salt (6.6 µg, MW 50 – 100 kDa, polymerized exclusively through the α-COOH group) was mixed with HeLa histone (13.2 µg, Active Motif) in 2 M sodium chloride and diluted to 150 mM in TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA) to give 120 µL HP mix at 110 ng/µL. The mixture was mixed by gentle pipetting and incubated at room temperature for 1 hour. The mixture was ultra-centrifuged at 13 k rcf for 10 minutes to remove possible aggregates. The HP mix was adjusted to 80 ng/µL with TE buffer with 150 mM NaCl, and then mixed with DNA (35 ng) and BSA (0.8 µg) at total volume of 14 µL. The reaction mixture was incubated at room temperature for 5.5 hours. Then the nucleosome was cross-linked with formalin solution (0.75%) and incubated at room temperature for 5 minutes, then quenched with glycine solution (125 mM) for 5 minutes. The reaction was subsequently cleaned up using a P6 column.

5.2.1.4 Gel Electrophoresis and Imaging

The result of the nucleosome assembly procedure was checked by gel electrophoresis using 6% DNA Retardation gel or 4-12% TBE gel (Life Technologies) in 0.4X TBE buffer (Life Technologies) at 4°C at 125 V using the XCell SureLock Mini-Cell Electrophoresis tank (Life Technologies). The gels were pre-run for at least 1.5 hours. GeneRuler 1kb Plus DNA ladder (Thermo Scientific) was used as a reference. The gels were imaged with a Typhoon Trio Imager (Amersham Biosciences, UK) or GBOX (Syngene, Cambridge, UK).

For the unlabelled samples, the DNA was imaged using either GelRed or SybrGold. For GelRed staining, the gels were stained with 1X GelRed Nucleic Acid Gel Stain (Biotium, California, USA) for 10 min while shaking. The gels were imaged with 532 nm green light and a 610 nm band pass emission filter. For SybrGold staining, the gels were stained with 1X SybrGold (Thermo Fisher) for 10 min while shaking, and imaged by excitation at 526 nm and emission measured at 532 nm. Cy3 labelled gels were excited at 532 nm and emission measured at 580 nm with 30 nm filter bandwidth, while Cy3 labelled gels were excited at 633 nm and emission measured at 670 nm with 30 nm filter bandwidth. The band intensity was quantified using Image Studio Lite (Li-COR Biosciences, Nebraska, USA) and confirmed by ImageQuant (GE Healthcare, UK).

5.2.2 Nucleosome Stability Comparison

5.2.2.1 Nucleosomal DNA Preparation

The DNA was prepared in the same way as described in Section 5.2.1.1 with either Cy3 or Cy5 labelled primers.

5.2.2.2 Nucleosome Stability Measurement

Fluorescently labelled xC-containing Widom DNA (100 ng) was mixed with 5 µg competitor 5S rDNA (NEB) without any fluorescence labelling and 2.72 µg recombinant histone proteins (NEB) at 2 M NaCl (total sample volume 50 μL). The resultant mixture was dialyzed in a Slide-A-Lyzer^{TM} MINI Dialysis Unit (MW 10 kDa cutoff, ThermoFisher Scientific) in a 2 mL eppendorf tube filled with dialysis buffer composed of 1XDTE buffer (20 mM Tris-HCl pH 8.0, 1 mM EDTA, 1 mM DTT) and NaCl concentrations decreasing from 2 M to 0.25 M via three intermediate salt concentrations (1.5, 1.0, 0.6 M NaCl) over 24 hours at 4 °C. The dialysis buffer was changed at least once per salt concentration. The mixture was recovered from the dialysis unit and prepared for analysis by gel electrophoresis as described in Section 5.2.1.4.

5.2.3 fC Positional Preference in the Nucleosome

5.2.3.1 Nucleosomal DNA Preparation

The DNA was prepared in the same way as described in Section 5.2.1.1. fdCtp input percentages of 5%, 10%, 20%, 40%, 60%, 80%, 100% with the rest being dCtp were used to prepare Widom DNA with different fC densities.

5.2.3.2 Nucleosome Preparation

The nucleosome was prepared as described in Section 5.2.1.3.1, with the following alterations: the MM concentration is 60% of that described before, and the DNA input was 600 ng, composed of 50 ng of DNA with fC at the aforementioned seven modification densities and 250 ng unmodified Widom DNA.

5.2.3.3 Nucleosomal DNA Recovery

The nucleosome assembly mixture was resolved on 10% native TBE gel (conditions as described in Section 5.2.1.4), and imaged by GelRed staining to show the position of the nucleosome band (position confirmed with the control nucleosome as shown in Figure 3-24). The gel slice containing nucleosomes was cut out and crushed finely into a gel slurry, and repetitively soaked with 1XTE buffer (pH 8.0) with 0.05% SDS and extracted at least three times, while vigorously shaking at 800 rpm at 4 °C. The extraction solution was filtered using a Spin-X centrifuge tube filter (0.22 μm pore size, Corning Costar, Sigma Aldrich) and concentrated down with an Amicon 10 kDa concentrator. The resulting solution was subjected to Proteinase K

(Active Motif) digestion at 55°C for 30 mins, and purified with the GeneJet PCR Purification Kit, and subsequently with a P6 column.

5.2.3.4 Profiling fC Positions by redBS-seq

The recovered nucleosomal DNA was subjected to standard library prep procedure using the NEBNext Ultra Library Prep Kit for Illumina (NEB), and then the redBS treatment. The DNA was reduced with 0.25 M sodium borohydride, and incubated at room temperature for 1 hour without light. Then 25 mM sodium acetate (pH 5) was used to quench the reaction mixture. The mixture was purified with a P6 column, then subjected to BS treatment with the EZ DNA Methylation-Gold Kit (Zymo). The obtained library was then subjected to 20 cycles of PCR amplification with Veraseq Ultra Polymerase (Enzymatics, Massachusetts, USA), and quantified using a KAPA Library Quantification Kit (KAPA Biosystems, Roche, Switzerland) prior to sequencing using a MiSeq (Illumina, California, USA) and the MiSeq Reagent Kit v2 (300-cycle, Illumina).

5.2.4 xC Positional Preference in the Nucleosome

5.2.4.1 Nucleosomal DNA Preparation

The DNA was prepared in the same way as described in Section 5.2.1.1, with the exception that the xdCtp input percentage is 15% dCtp, 8.6% mdCtp, 62% hmdCtp and 14.4% fdCtp, to generate Widom DNA modified with 20% of mC, hmC and fC. The percentage of xC incorporation was confirmed by LC-MS.

5.2.4.2 LC-MS/HPLC condition for xC Quantification

The conditions used are as described in Section 5.2.1.2.

5.2.4.3 Nucleosome Preparation

The nucleosome was prepared in the same way as described in Section 5.2.1.3.1, with the following alterations: the MM concentration is 60% of that described before, and the DNA input was 600 ng.

5.2.4.4 Nucleosomal DNA Recovery

The conditions used are the same as described in Section 5.2.3.3.

5.2.4.5 Profiling xC position by BS/oxBS/redBS-seq

The recovered nucleosomal DNA was subjected to standard library prep procedure using the NEBNext Ultra Library Prep Kit for Illumina (NEB), and then subjected to BS/oxBS/redBS treatment. BS was carried out using the EZ DNA Methylation-Gold Kit; redBS was performed as described in Section 5.2.3.4. oxBS was carried out using a TrueMethyl Seq Kit (Cambridge Epigenetics) according to the manufacturer’s instructions, with the exception that only one round of BS conversion was done. Then the obtained library was subjected to 20 cycles of PCR amplification with Veraseq Ultra Polymerase, and quantified using the KAPA Library Quantification Kit prior to sequencing using the MiSeq with the MiSeq Reagent Kit v2 (300-cycle).

5.2.5 Schiff Base Capture

5.2.5.1 DNA and Nucleosome Preparation

The 100% fC modified Widom DNA labelled with both Cy3 and Cy5 at 5’ ends was prepared as described in Section 5.2.1.1. After PCR synthesis, the reaction mixtures were purified initially using the GeneJet PCR Purification Kit, and subsequently on a 1.5% agarose gel pre-stained with GelRed, then extracted using a GeneJet Gel Extraction Kit (Life Technologies, California, USA) and eluted with MilliQ water.

The nucleosome was prepared as described in Section 5.2.1.3.1 and Section 5.4.2.1 by both methods, with the exception that the DNA input is 300 ng.

5.2.5.2 Reduction Capture of the Schiff Base

Nucleosomes were treated with 100 mM sodium cyanoborohydride (Sigma Aldrich) and incubated at 37°C for 18 h. The mixture was then purified using a P6 column.

5.2.5.3 Denaturing Gel Electrophoresis and Proteomics Analysis

The purified reduced nucleosome was mixed with 4X NuPAGE^{TM} LDS Sample Buffer (ThermoFisher Scientific) and heated at 95°C for 5 minutes. The resultant mixture was subsequently run on a 4-20% Bis-Tris Protein gel in 1X MES SDS buffer (ThermoFisher Scientific). The upshifted gel bands were excised and submitted to the Cambridge Centre for Proteomics, University of Cambridge, for commercial proteomics analysis using the procedure described in Section 5.2.5.4.

5.2.5.4 Proteomics Analysis

1D gel bands were transferred into a 96-well PCR plate. The bands were cut into 1 mm^2 pieces, destained, reduced (DTT) and alkylated (iodoacetamide) and subjected to enzymatic digestion with chymotrypsin overnight at 37°C. After digestion, the supernatant was pipetted into a sample vial and loaded onto an autosampler for automated LC-MS/MS analysis.

All LC-MS/MS experiments were performed using a Dionex Ultimate 3000 RSLC nanoUPLC (Thermo Fisher Scientific) system and a Q Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific). Separation of peptides was performed by reverse-phase chromatography at a flow rate of 300 nL/min and a Thermo Scientific reverse-phase nano Easy-spray column (Thermo Scientific PepMap C18, 2 mm particle size, 100 Å pore size, 75 mm i.d. x 50 cm length). Peptides were loaded onto a pre-column (Thermo Scientific PepMap 100 C18, 5 mm particle size, 100 Å pore size, 300 mm i.d. x 5 mm length) from the Ultimate 3000 autosampler with 0.1% formic acid for 3 minutes at a flow rate of 10 mL/min. After this period, the column valve was switched to allow elution of peptides from the pre-column onto the analytical column. Solvent A was water + 0.1% formic acid and solvent B was 80% acetonitrile, 20% water + 0.1% formic acid. The linear gradient employed was 2-40% B in 30 minutes.

The LC eluent was sprayed into the mass spectrometer by means of an Easy-Spray source (Thermo Fisher Scientific Inc.). All *m/z* values of eluting ions were measured in an Orbitrap mass analyzer, set at a resolution of 70000 and was scanned between *m/z* 380-1500. Data dependent scans (Top 20) were employed to automatically isolate and generate fragment ions by higher energy collisional dissociation (HCD, NCE: 25%) in the HCD collision cell and measurement of the resulting fragment ions was performed in the Orbitrap analyser, set at a resolution of 17500. Singly charged ions and ions with unassigned charge states were excluded from being selected for MS/MS and a dynamic exclusion window of 20 seconds was employed.

Post-run, the data was processed using Protein Discoverer (version 2.1., ThermoFisher). Briefly, all MS/MS data were converted to mgf files and the files were then submitted to the Mascot search algorithm (Matrix Science, London UK) and searched against the Uniprot human database (151984 sequences; 47833598 residues) and a common contaminant sequences (115 sequences, 38274 residues). Variable modifications of oxidation (M), deamidation (NQ) and carbamidomethyl were applied. The peptide and fragment mass tolerances were set to 5ppm and 0.1 Da, respectively. A significance threshold value of $p<0.05$ and a peptide cut-off score of 20 were also applied.

5.2.5.5 PolStop Assay

The procedure was performed by Dr Robyn Hardisty from Balasubramanian lab.[298] The nucleosome was first subjected to reduction treatment as described in Section 5.2.5.2, and the nucleosome was denatured by heat. The resulting ssDNA was then used as the template for single primer extension experiments with the forward and reverse primers of Widom DNA, with the elongation step restricted to five minutes. The histone subunits were then digested by Proteinase K, and the 3' overhang from the template strand was subsequently blunted by ssDNA exonuclease RecJ$_f$ (NEB). The resulting blunt dsDNA was subjected to a standard library prep procedure using NEBNext Ultra Library Prep Kit for Illumina (NEB), quantified with KAPA Library Quantification Kit and sequenced using a MiSeq with the MiSeq Reagent Kit v3 (150-cycle, Illumina).

5.3 GL Method in Chapter 3

5.3.1 DNA Preparation

All DNA sequences were purchased from Sigma and Biomers (Singapore). The DNA was dissolved in MilliQ water and the concentration was adjusted with UV absorbance of Nanodrop measurements and stored at -20°C.

Sequence Name		Sequence (5' to 3', P'- stands for phosphate group)
Set f	Strand 1	GAGCGGCCTCGGCACCGGGATTCTCGAT
	Strand 2-bio	biotin-ATCGAGAATCC
	Strand 3-P'	P’-GGTGCCGAGGCCGCTC
Set g	Strand 1	GAGCTGTCTACGACCAATTGAGCGGCCTCGGC
	Strand 2-bio	biotin-GCCGAGGCCGCT
	Strand 3-P'	P’-AATTGGTCGTAGACAGCTC
Set h	Strand 1	CGCGGGGGACAGCGCGTACGTGCGTTTAAGCGG
	Strand 2-bio	biotin-CCGCTTAAACGCA
	Strand 3-P'	P’-GTACGCGCTGTCCCCCGCG
Set i	Strand 1	CGTGCCTGGAGACTAGGGAGTAATCCCCTTGGCGG
	Strand 2-bio	biotin-CCGCCAAGGGGATTACTCC
	Strand 3-P'	P’-TAGTCTCCAGGCACG
Set j	Strand 1	CGGATGTATATATCTGACACGTGCCTGGAGACTAG
	Strand 2-bio	biotin-CTAGTCTCCAGGCACGTGT
	Strand 3-P'	P’-AGATATATACATCCG

Table 5-2. Short sequences used for validating the GL method.

Sequence Name		Sequence (5' to 3', P'- stands for phosphate group)
Strand 1		ATCGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAGGGGATTACTCCCTAGTCTCCAGGCACGTGTCAGATATATACATCCGAT
Set a	Strand 2-Cy3	Cy3-ATCGGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCCCTTGG
	Strand 2-bio	biotin-ATCGGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCCCTTGG
	Strand 3-P'	P’-GGTTAAAACGCGGGGGACAGCGCGTACGTGCGTTTAAGCGGTGCTAGAGCTGTCTACGACCAATTGAGCGGCCTCGGCACCGGGATTCTCGAT
Set b	Strand 2-Cy3	Cy3-ATCGGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCCCTTGGCGGTTAAAA
	Strand 2-bio	biotin-ATCGGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCCCTTGGCGGTTAAAA
	Strand 3-P'	P’-GCGGGGGACAGCGCGTACGTGCGTTTAAGCGGTGCTAGAGCTGTCTACGACCAATTGAGCGGCCTCGGCACCGGGATTCTCGAT
Set c	Strand 2-Cy3	Cy3-ATCGGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCCCTTGGCGGTTAAAACGCGGGGGACAG
	Strand 2-bio	biotin-ATCGGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCCCTTGGCGGTTAAAACGCGGGGGACAG
	Strand 3-P'	P’-GCGTACGTGCGTTTAAGCGGTGCTAGAGCTGTCTACGACCAATTGAGCGGCCTCGGCACCGGGATTCTCGAT
Set d	Strand 2-Cy3	Cy3-ATCGGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCCCTTGGCGGTTAAAACGCGGGGGACAGCGCGTA
	Strand 2-bio	biotin-ATCGGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCCCTTGGCGGTTAAAACGCGGGGGACAGCGCGTA
	Strand 3-P'	P’-GTGCGTTTAAGCGGTGCTAGAGCTGTCTACGACCAATTGAGCGGCCTCGGCACCGGGATTCTCGAT

Table 5-3. Long sequences used for validating the GL method.

5.3.2 GL Method

5.3.2.1 With Non-biotinylated Strand 2

Strands 1 and 2 and 3 were mixed in the molar ratio of 0.95:0.95:1 in 1XTES buffer (20 mM Tris pH 8.0, 1 mM EDTA and 50 mM NaCl). Then the mixture was annealed by heating to 95°C and slowly cooled down to 4°C over a period of 1.2 hours. Then the annealed DNA was subjected to one-pot gap filling and ligation reaction with 1.5 U T4 polymerase (NEB), 100 μM xdCtp and 400 U T4 ligase (NEB) for 6 pmole annealed DNA at 12°C incubation for 1 hour.

5.3.2.2 With Biotinylated Strand 2

Strands 1 and 2 and 3 were mixed and annealed as described in Section 5.3.2.1. Then the annealed DNA was immobilized on Dynabeads^{TM} MyOne^{TM} Streptavidin T1 beads (30 µL for 6 pmole annealed DNA, ThermoFisher Scientific) in 30 µL BB-3 (5 mM Tris-HCl pH 8.0, 1000 mM NaCl, 0.5 mM EDTA and 0.05% Triton X-100, all working concentration) by incubating at room temperature for 20 minutes or 4°C overnight while rotating in a 250 µL PCR tube. The DNA on beads was washed twice with 150 µL BB-4 (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 1 mM EDTA and 0.05% Triton X-100, all working concentration). Then the washing buffer was removed and 20 µL gap filling reaction mixture with 1.5 U T4 polymerase (NEB), 100 µM xdCtp was added to beads. The beads were quickly and gently resuspended, and incubated at 12°C for 20 minutes. Then the solution was separated from the beads using a magnet, and the beads were washed with BB-4 once. Then 10 µL ligation reaction mixture containing 400 U T4 ligase was added to the beads. The beads were quickly and gently resuspended, and incubated at 16°C for 45 minutes. The reaction was stopped by again separating the solution from beads using a magnet; the beads were washed twice with 150 µL BB-4.

5.3.3 LC-MS Validation

LC-MS validation uses short DNA sequences with biotinylated strand 2 prepared as described in Section 5.3.2.2. The DNA was eluted from streptavidin beads by 25 µL 95% formamide + 10 mM EDTA and incubated at 65°C for 5 minutes. Then the beads were pelleted using a magnet, and the supernatant was collected and cleaned using a P6 column and injected into the LC-MS.

The DNA was separated on a XTerra@MS C18 column (2.5 µm, 2.1mm x 50 mm, Waters) on a LC-MS monitored at 260 nm. Solvent A (10 mM triethylamine (TEA, Fluka) and 100 mM 1,1,1,3,3,3-hexafluoro-2-propanol (HFIP, Fluka) in MQ H_2O) and B (methanol) were used for the DNA separation. The percentage of B is shown in the Figure 5-2. Flow rate is 0.2 mL/min.

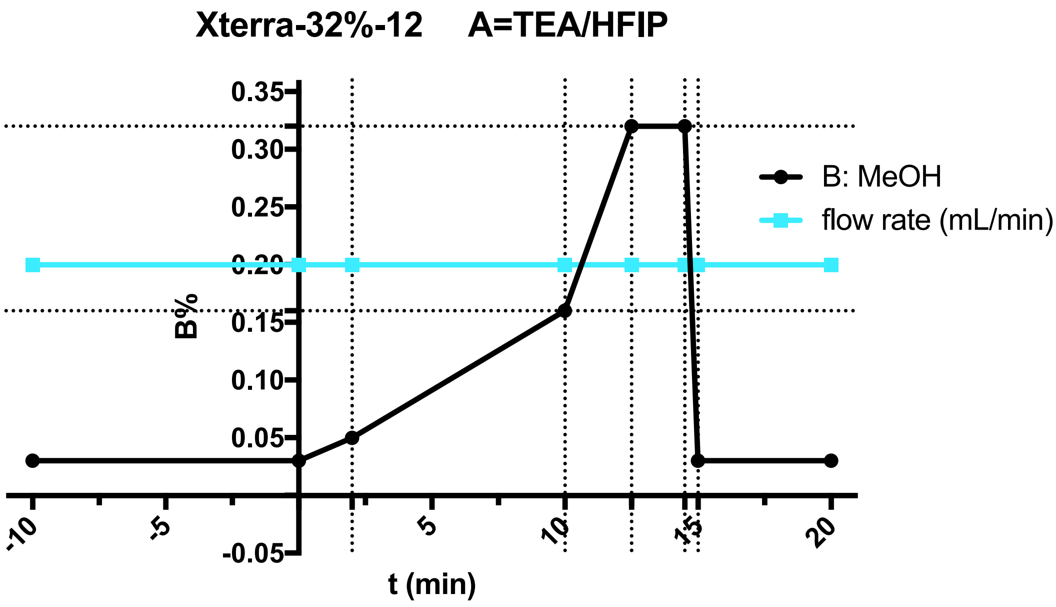


Figure 5-2. LC-MS conditions to confirm product identity for GL method validation. Solvent A (10 mM TEA and 100 mM HFIP) and B (100% methanol) were used. The percentage of B is shown in the figure. Flow rate is 0.2 mL/min.

5.3.4 Gel Electrophoresis Validation

Gel electrophoresis validation uses long DNA sequences with a non-biotinylated strand 2 prepared as described in Section 5.3.2.1. After ligation incubation, the DNA was directly mixed with 2X TBE Urea Sample Buffer (ThermoFisher Scientific) and heated at 95°C for 5 minutes. Then the reaction was loaded into 6% Novex TBE-Urea gel (ThermoFisher Scientific) and run in 1XTBE buffer. For GL reaction with unlabelled strand 2, the gel was stained with 1XSybr Gold for 10 min while shaking, and the gels were imaged with 526 nm excitation and 532 nm emission. For GL reaction with Cy3-labelled strand 2, the gel was directly imaged by excited at 532 nm and measured emission at 580 nm with 30 nm filter bandwidth without staining.

5.3.5 NGS Validation

NGS validation uses long DNA sequences with biotinylated strand 2 prepared as described in Section 5.3.2.2. The streptavidin beads with DNA after GL reaction was incubated with freshly prepared 0.1 M sodium hydroxide at room temperature for 12 minutes, then the beads were pelleted and supernatant was discarded. The beads were washed with 150 µL BB-4 twice and used as a template for on-bead PCR as described in Section 5.2.1.1 for 12 cycles. The beads were pelleted and the supernatant was recovered for standard library prep procedure using NEBNext Ultra Library Prep Kit for Illumina (NEB). The obtained library was subjected to 12 cycles of PCR amplification with Taq Polymerase, and quantified with KAPA Library Quantification Kit (KAPA Biosystems, Roche) for single-end sequencing on a MiSeq using the MiSeq Reagent Kit v3 (150-cycle, Illumina).

5.4 Cryo-EM Study in Chapter 4

5.4.1 DNA Preparation

DNA preparation was as described in Section 5.2.1.1 with the following addition: after purification with the GeneJet Kit, the DNA was concentrated using an Amicon 10 kDa column until the DNA concentration is above 1000 ng/µL.

5.4.2 Nucleosome Preparation

5.4.2.1 Assembled by Salt Dilution

DNA (213 pmole) was mixed with recombinant histone proteins (181 pmole, NEB) in the molar ratio of histone: DNA = 0.853 at 2 M NaCl at 1XDTE buffer with 0.01% Triton X-100 (1XDTTE buffer) with a total sample volume of 85 µL. After incubating at room temperature for 30 minutes, the mixture was sequentially diluted with 1XDTTE buffer with 30 minutes incubation between two dilutions. The salt concentration was reduced gradually from 2 M NaCl to 1.5, 1.0, 0.6, 0.25 and finally to 0.12 M NaCl. The mixture was then ready for concentrating prior to cryo-EM grid preparation.

5.4.2.2 Assembled by Salt Dialysis

DNA (260 pmole) was mixed with recombinant histone proteins (222 pmole) in the molar ratio of histone: DNA = 0.853 at 2 M NaCl at 1XDTE buffer (total sample volume 400 µL), and dialyzed in the Slide-A-Lyzer^{TM} MINI Dialysis Devices (MW 10 kDa cut off, 0.5 mL) in 15 mL conical tubes filled with dialysis buffers (1XDTE buffer and NaCl concentration successively decreased from 2 M to 0 M). The salt concentration was gradually reduced to 0 M in four intermediate salt concentrations (1.5, 1.0, 0.6 and 0.25 M NaCl) over 24 hours at 4°C. The dialysis buffer was changed at 1.5 hour interval, and dialyzed overnight at 0 M NaCl, and then changed to another tube of fresh 0 M dialysis buffer and repeat dialysis for 1.5 hours. The mixture was recovered from the dialysis device in preparation for concentrating.

5.4.2.3 Nucleosome Concentration for Cryo-EM Sample Preparation

Nucleosomes were concentrated with a BSA-passivized Amicon 10 or 30 kDa column. The passivation was done by filling the Amicon filter with 550 µL freshly dissolved 1% BSA in the nucleosome assembly buffer, and incubating either overnight at 4°C or 2 hours at room temperature. The BSA solution was then discarded and the column filter was washed vigorously with 550 µL nucleosome assembly buffer three times to remove non-binding BSA. The nucleosome solution was then added into the filter for concentrating.

5.4.3 Nucleosome Quality Check with Native Gel Electrophoresis

The procedure was the same as described in Section 5.2.1.4.

5.4.4 Cryo-EM Sample Preparation and Data Collection

The procedure of grid preparation, imaging and structural determination was performed by Dr Ben Luisi, Dr Dimitri Chirgadze and Dr Kotryna Bloznelyte, Department of Biochemistry, University of Cambridge. Quantifoil R1.2/1.3 holey carbon cupper grids were used for cryo-EM sample preparation. 3 µL samples were loaded on the face of the freshly glow discharged grid (2 mins at 25 mAmp with PELCO easiGlow machine), and the grid was attached onto a Vitrobot (Thermo Fisher Scientific) for the sample vitrification. The chamber was kept constant at 4°C and 100% humidity. After blotting for 3 s with blot force -5 or 0, the grid was plunge-frozen in liquid ethane, cooled with liquid nitrogen. The grids were stored in liquid nitrogen before screening and data collection. Grid screening was done using a Talos Arctica 200kV FEG cryo-transmission electron microscope (cryo-TEM) with autoloader from FEI (Thermo Fisher Scientific), and data collection was done using a Titan Krios 300kV FEG cryo-transmission electron microscope from FEI (Thermo Fisher Scientific) equipped with sample autoloader, Falcon 3 and Gatan's K3 detectors, and GIF Quantum LS imaging filter. Electron micrographs were recorded by exposing the grid for 2s on a Titan Krios at 300 kV at a nominal magnification of 75kX, giving a resolution of 1.1 Å/pixel. Data was collected with defocus range of -1.9 to -3.3 with step 0.2, giving a dose rate of 42.6

e/Å/sec, and 2 exposures per hole.

5.5 Proteomics Study in Chapter 4

5.5.1 DNA Preparation

DNA preparation was as described in Section 5.2.1.1 with the biotinylated forward primer.

5.5.2 Nucleosome Preparation

Nucleosome preparation was as described in Section 5.4.2.1.

5.5.3 Protein Pull-Down Experiments and Proteomics Analysis

The guard nucleosomes and bait nucleosomes (both 2.52 μg) were immobilized separately on Dynabeads^{TM} MyOne^{TM} Streptavidin T1 beads washed three times with WB-1 (10 mM Tris-HCl pH 8.0, 125 mM NaCl, 0.2 mM EDTA and 0.01% Triton X-100, all working concentration) by incubating at 4°C for 2 hours while rotating in a 250 μL PCR tube. The guard nucleosomes on beads were washed three times with 125 μL BB-2 (20 mM HEPES pH 8.0, 100 mM NaCl, 0.2 mM EDTA, 20% Glycerol, 1 mM DTT, 0.1% Triton X-100 and 1X cOmplete^{TM} mini EDTA-free protease inhibitor cocktail (Roche), all working concentration). The supernatant from the guard nucleosomes on streptavidin beads was removed and 2.52 μg HeLa S3 nuclear extract (Abcam, Cambridge, UK) diluted with BB-2 to 190 μL was added and incubated at 4°C for 2 hours. The nuclear extract was subsequently recovered and added to the bait nucleosome already washed three times with 125 μL BB-2. After incubating at 4°C overnight, the bait nucleosome was recovered by pelleting the beads using a magnet, and washed five times with BB-2. The proteins pulled down by the bait nucleosome were eluted by 16 μL 1XLDS buffer (diluted with MQ H_2O from 4 X NuPAGE^{TM} LDS Sample Buffer) and heated at 95°C for 5 minutes. The supernatant was run 2 cm into a 4-12% Bis-Tris SDS protein gel in 1 X MOPS SDS buffer, and the protein bands were visualized by InstantBlue^{TM} Coomassie Protein Stain (Expedeon, Cambridge, UK), excised and submitted to proteomics for LC-MS/MS for protein identification as described in Section 5.2.5.4.

6 Appendix

6.1 UV Melting Curves for All Samples

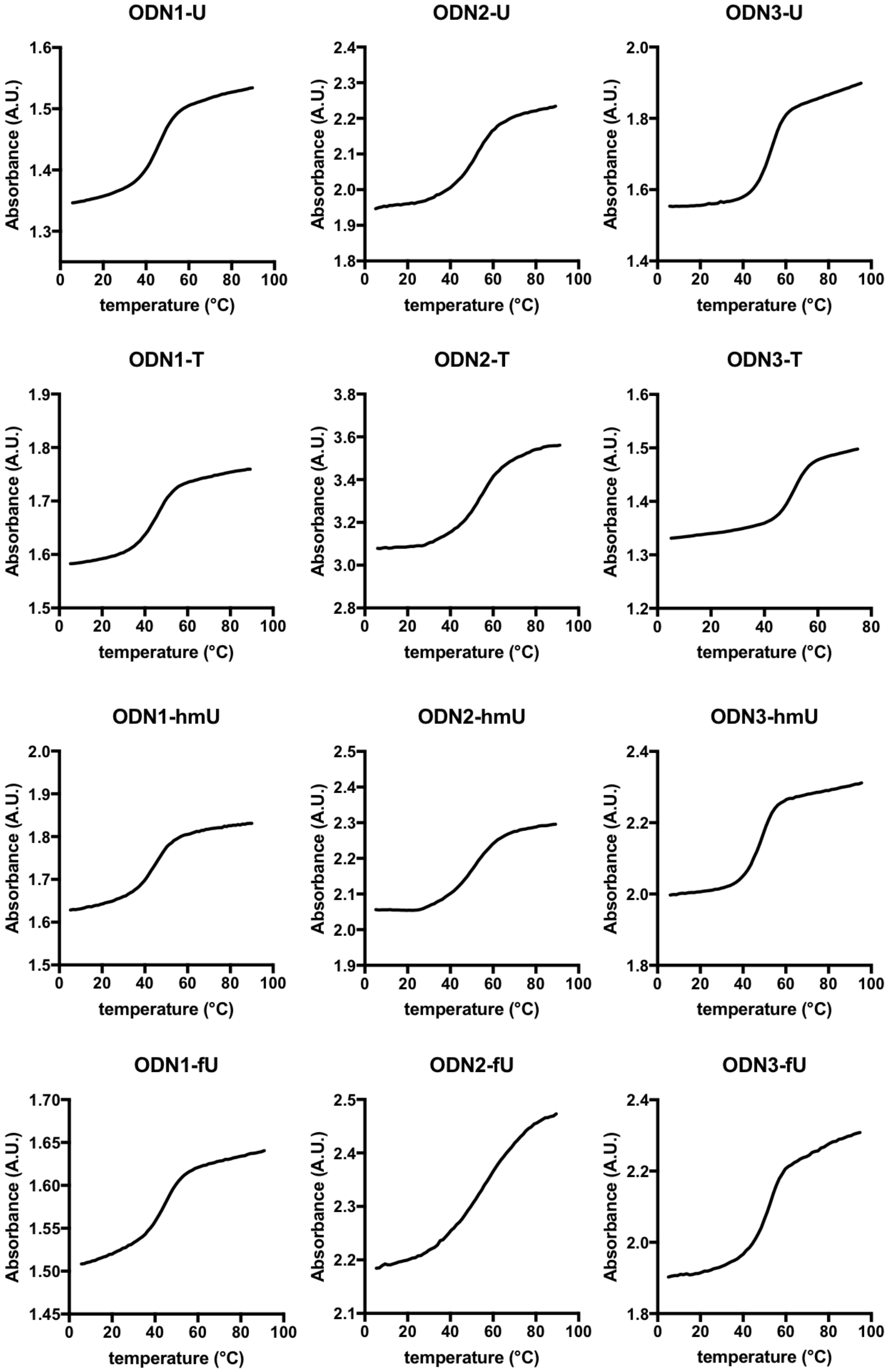


Figure 6-1. Examples of melting curves (raw data) for UV melting experiments detailed in Section 2.2.1. A.U. stands for arbitrary unit.

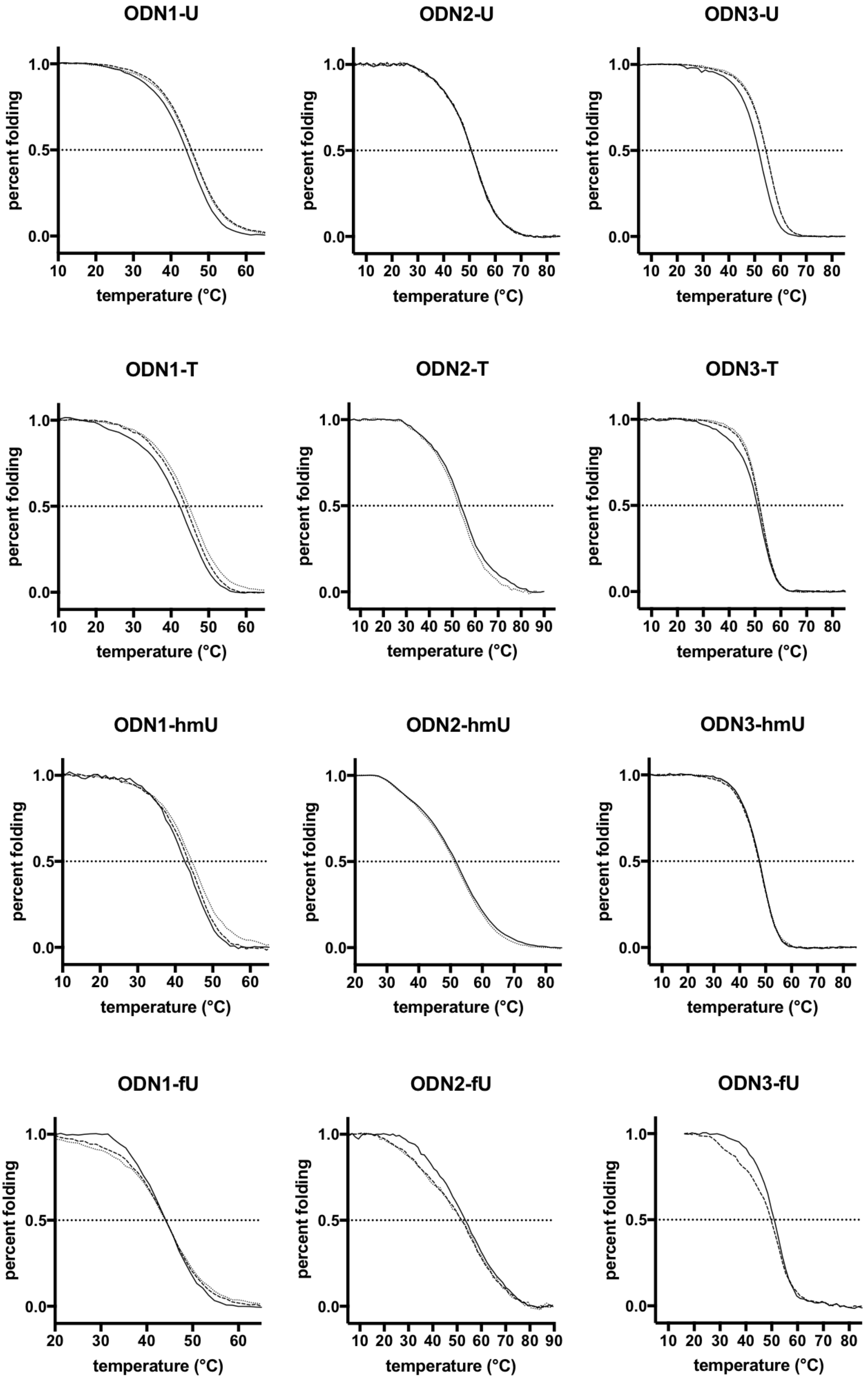


Figure 6-2. Normalized melting curves corresponding to Tm values in Figure 2-3. Curves obtained from three measurements are shown in solid, dashed and dotted lines.

6.2 Nucleosome Occupancy Measurements

Upon binding, the dye exhibits a large fluorescence enhancement and thus allows quantitative detection of GelRed bound nucleic acid. The staining time with GelRed was screened from 5 minutes to 2 hours at 4 °C, and the nucleosome% calculated were the same.

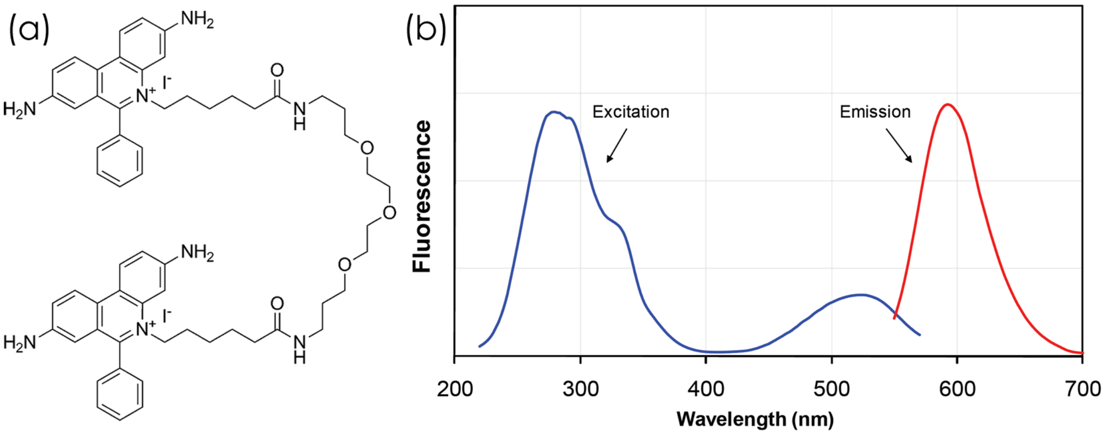


Figure 6-3. (a) Structure of GelRed; (b) Excitation (left) and emission (right) spectra of GelRed bound to dsDNA in TBE buffer (taken from Biotium product information sheet).

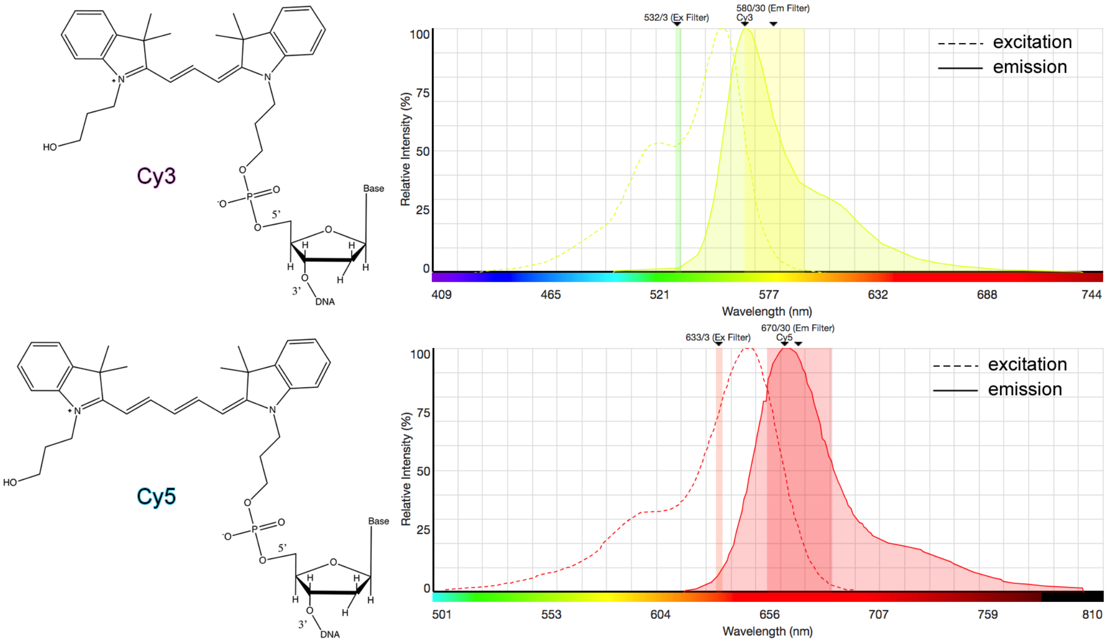


Figure 6-4. Chemical structure of Cy3 and Cy5 fluorescence labelling, attached covalently to the 5' end of the sugar of the first nucleoside. The excitation and emission spectra were created by Fluorescence SpectraViewer (ThermoFisher Scientific), with the excitation and emission filter wavelength taken from Typhoon setting. Cy3 fluorescence was excited at 532 nm and measured emission at 580 nm with 30 nm filter bandwidth; Cy5 fluorescence was excited at 633 nm and measured emission at 670 nm with 30 nm filter bandwidth.

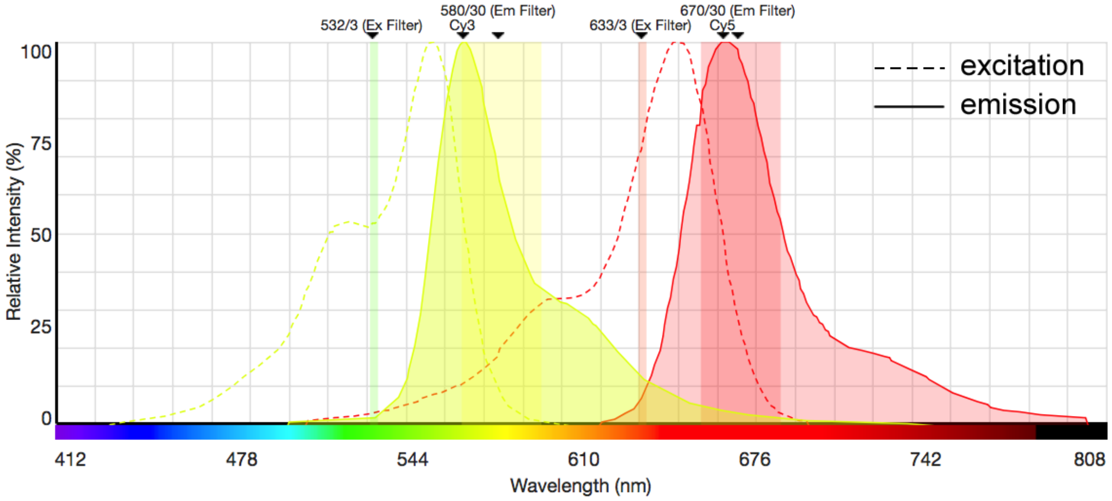


Figure 6-5. The excitation and emission spectra of Cy3 and Cy5 fluorescence labelling. Cy3 fluorescence was excited at 532 nm and emission measured at 580 nm with 30 nm filter bandwidth; Cy5 fluorescence was excited at 633 nm and emission measured at 670 nm with 30 nm filter bandwidth (figure created by Fluorescence SpectraViewer, ThermoFisher Scientific, with the excitation and emission filter wavelength taken from Typhoon setting).

6.3 Reactions Related to BS/oxBS/redBS-seq for xC Position Profiling

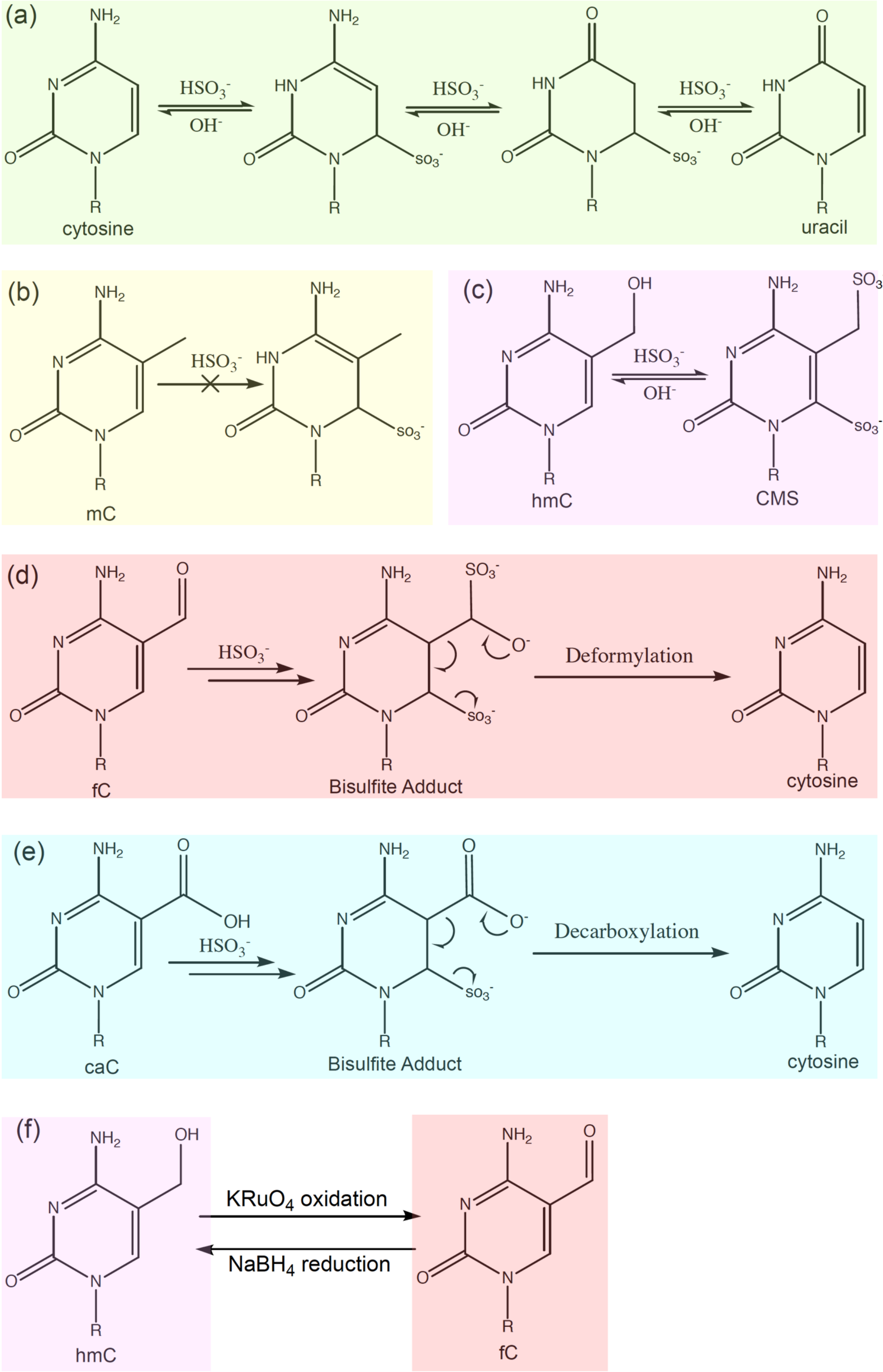


Figure 6-6. Reactions related to BS/oxBS/redBS-seq for xC position profiling in Section 3.2.11.2. Reactions were drawn according to Booth *et al.*[329].

6.4 Condition Optimization for Competition and Nucleosomal DNA Recovery

First, the nucleosome formation conditions were screened to select for the top 10% of total DNA with the highest affinity towards histone. With fixed histone + NAP-1 + ACF (MM, shown in Figure 3-4 and Table 3-1) input, the DNA input amount was screened between 250 to 650 ng per 15 µL assembly reaction (Figure 6-7). However, even 650 ng DNA input cannot achieve 10% selection. To avoid overloading, the DNA input amount was set to be 600 ng, and the MM concentration in the nucleosome assembly reaction was screened to further reduce nucleosome percentage. The screening results indicate that around 60% MM and 600 ng DNA input in each assembly reaction will give 10% nucleosome.

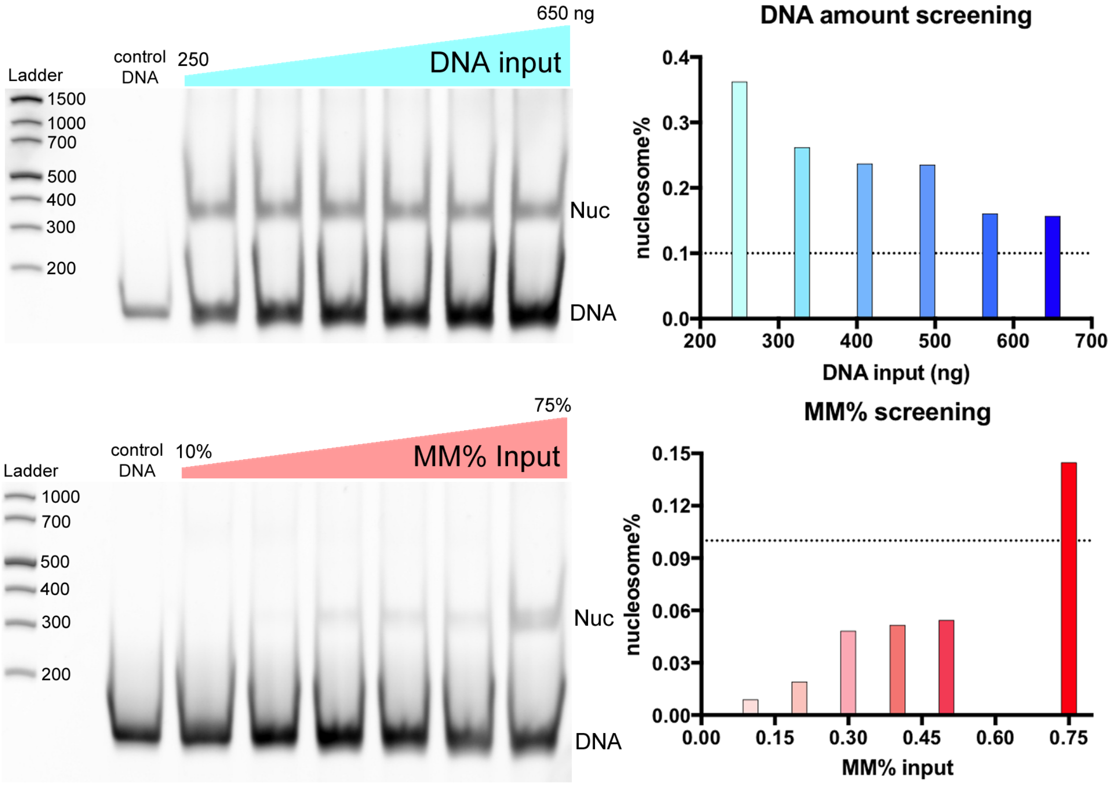


Figure 6-7. DNA amount and MM% input screening for xC competition for 10% DNA incorporation into the nucleosome. DNA amount was screened between 250 to 650 ng, and MM% input was screened between 10% to 75%.

During the screening process, it was observed that the free DNA and nucleosome band was streaking across the entire lane when run on 6% DNA retardation gel. This had to be eradicated, as once the nucleosomal DNA is contaminated with free DNA, which was not selected for nucleosome-formation, they cannot be distinguished due to their identical sequence, and thus give misleading results during xC profiling. Separating free DNA from nucleosomal DNA is essential. In order to do this, different percentages (4 – 12%) of native gel were tested as shown in Figure 6-8.

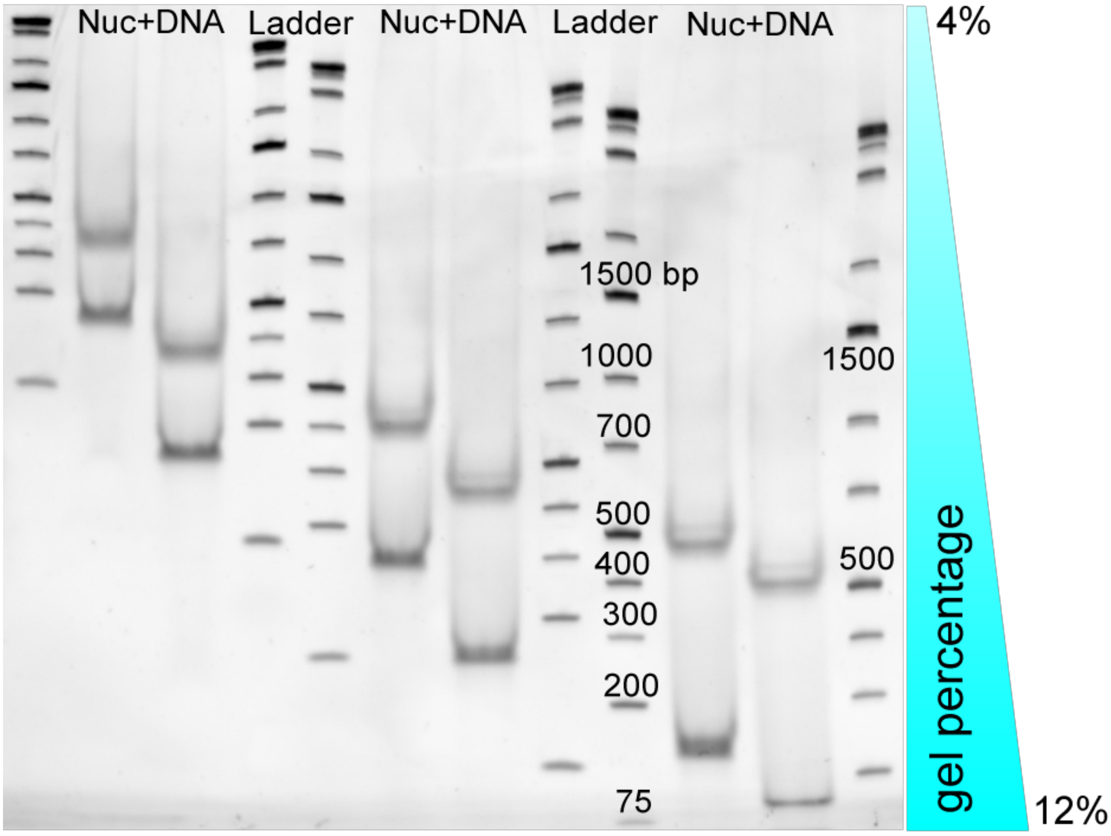


Figure 6-8. Gel percentage screening for optimum condition to separate nucleosomal DNA and free DNA cleanly.

It was observed that an increase in gel percentage caused the nucleosome band to gradually shift up with respect to the ladder bands (from below 400 bp to above 500 bp), despite the constant position of free DNA with respect to the 200 bp marker. Therefore, the distance between the nucleosome and DNA gradually increased, and separation became cleaner. It appears that 10% native gel can cleanly separate free DNA from nucleosomal DNA.

To ensure fair and thorough competition, the nucleosome assembly reaction was assembled together in siliconized tubes in an eppendorf incubator, with 300 rpm shaking to facilitate the effective diffusion of components. The nucleosome assembly reaction was incubated overnight to allow the Widom DNA bearing the most nucleosome-favouring xC combination to become incorporated into the nucleosome.

The nucleosome assembly mixture was resolved on 10% native gel, with the control nucleosome run alongside to indicate the correct nucleosome band position, since with the increase of gel percentage from 6% to 10%, the nucleosome no longer migrates between the 300 and 400 bp markers. As shown in Figure 3-24, the position of nucleosomal DNA was elucidated with GelRed staining and the control nucleosome position, and recovered by cutting out the gel slice, finally crushing it into a gel slurry. It was repetitively soaked with 1XTE buffer pH 8.0 (at least 3 stages of extraction to increase the recovery yield) while vigorously shaking at 800 rpm at 4°C to prevent fC from reacting and reducing DNA degradation. SDS was added into the extraction buffer to deform the gel matrix and nucleosome, allowing the nucleosomal DNA to diffuse out. The extracted DNA was concentrated and digested with Proteinase K, to remove histone proteins and biological assisting factors. The resultant mixture was further purified using a silica-based column to remove GelRed, Proteinase K and amino acid fragments from previous digestion, to reduce possible interference with subsequent xC profiling. The DNA was further purified using a gel filtration column, as the DNA extracted from gel tends to have an unidentified high peak at around 230 nm, interfering with UV quantification.

6.5 GL Method Validation

6.5.1 NGS Results

Gap-insert xC	number of reads				
	total	correct	deletion	mutation	insertion
setA-mC	78983	78238	577	168	22
setB-mC	68148	67763	320	65	109
setC-mC	88643	86632	1754	257	51
setD-mC	100482	98558	1766	158	24
setA-hmC	61594	60902	550	142	37
setB-hmC	79085	78495	505	85	177
setC-hmC	64545	62837	1523	185	86
setD-hmC	85025	82751	2125	149	69
setA-fC	31629	31250	315	64	25
setB-fC	57924	57432	403	89	80
setC-fC	92236	89987	1963	286	41
setD-fC	99335	97250	1880	205	30
setA-caC	72533	71247	1061	225	48
setB-caC	78531	73574	4698	259	112
setC-caC	193519	176363	16353	803	224
setD-caC	73372	65173	7929	270	75

Table 6-1. Number of reads for each gap position and different C modification incorporation for the correct incorporations (C), deletions (mis-ligated product), mutations (T/G/A rather than C), and insertions (more than one C incorporated).

Gap-insert xC	deletion percentage			mutation percentage			insertion percentage		
	uncorrected %	baseline error %	corrected %	uncorrected %	baseline error %	corrected %	uncorrected %	baseline error %	corrected %
setA-mC	0.73	0.17	0.56	0.21	0.27	-0.06	0.03	0.03	-0.01
setB-mC	0.47	0.19	0.28	0.10	0.28	-0.19	0.16	0.04	0.12
setC-mC	1.98	0.21	1.77	0.29	0.28	0.01	0.06	0.05	0.01
setD-mC	1.76	0.52	1.24	0.16	0.27	-0.12	0.02	0.02	0.00
setA-hmC	0.89	0.18	0.71	0.23	0.29	-0.06	0.06	0.03	0.03
setB-hmC	0.64	0.21	0.43	0.11	0.29	-0.18	0.22	0.04	0.19
setC-hmC	2.36	0.20	2.16	0.29	0.26	0.03	0.13	0.05	0.08
setD-hmC	2.50	0.60	1.90	0.18	0.27	-0.09	0.08	0.03	0.05
setA-fC	1.00	0.19	0.81	0.20	0.29	-0.09	0.08	0.04	0.04
setB-fC	0.70	0.23	0.47	0.15	0.29	-0.14	0.14	0.04	0.09
setC-fC	2.13	0.21	1.92	0.31	0.25	0.06	0.04	0.04	0.00
setD-fC	1.89	0.64	1.25	0.21	0.28	-0.07	0.03	0.02	0.01
setA-caC	1.46	0.29	1.17	0.31	0.29	0.02	0.07	0.03	0.03
setB-caC	5.98	0.67	5.32	0.33	0.34	-0.01	0.14	0.03	0.11
setC-caC	8.45	0.30	8.15	0.41	0.27	0.15	0.12	0.04	0.07
setD-caC	10.81	2.17	8.64	0.37	0.29	0.08	0.10	0.03	0.07

Table 6-2. Percentage of total reads for each gap position and different C modification incorporation for the correct incorporation (C), deletions (mis-ligated product), mutations (T/G/A rather than C), and insertions (more than one C incorporated).

6.6 Cryo-EM Initial Sample Screen and Condition Optimization

(DNA and nucleosomes were prepared by me; grid preparation, imaging and structural determination was done by Dr Ben Luisi, Dr Dimitri Chirgadze and Dr Kotryna Bloznelyte, Department of Biochemistry, University of Cambridge)

6.6.1 Cryo-EM Background and Workflow

Cryo-EM solves molecular structure by bombarding samples with an electron beam and detecting the interference between the electron beam and the sample. In order to prepare samples for the bombardment, the sample is loaded onto a grid blotted to remove excessive solution to form a thin film. Then the grid is rapidly plunged into liquid ethane (boiling point (b.p.) 184.6 K) cooled by liquid nitrogen (b.p. 77.36 K) for rapid freezing. Liquid ethane was chosen for its high heat exchange rate. Under these conditions, the thin film of aqueous solution forms vitreous (non-crystalline) ice, where the molecular structure can be maintained. The grid is loaded onto the electron microscope, which is maintained under vacuum and cooled by liquid nitrogen. An electron beam subsequently bombards the sample to produce terabytes of 2D movie data of the molecules with free orientations. These 2D images are subjected to computational analysis to reconstruct the 3D structure.

6.6.2 Advantages and Limitations of Cryo-EM

The advantages of cryo-EM to resolve nucleosome structure over other conventional structure determination techniques such as EM and X-ray crystallization, are as the following:

During the electron bombardment for imaging, the majority of electrons will only change direction while maintaining the same energy. A small fraction of electrons, however, will transfer the energy to the sample, causing radiation damage to the sample. In traditional EM performed at room temperature, a low dosage of electron beam has to be used to avoid excessive damage, and thus compromises the resolution. In order to increase the sensitivity, dyes with heavy atoms are used to coat the surface of the molecule to intensify the contrast. However, it has been reported that the dyes can cause sample flattening and structure distortion.[368] In contrast, cryo-EM is imaged at the cold temperature maintained by liquid nitrogen, which effectively reduces the radiation damage towards the sample during bombardment, and thus a higher dose of electron beam can be employed to elucidate the structure without the need to use staining as in traditional EM, while producing higher resolution.

comparison	cryo-EM	crystallization
amount	µg	mg
buffer	no glycerol	strict
crystals	no need	vital
purity	medium to high	high
conformations	multiple OK	single
sample state	aqueous	solid
result	direct imaging	indirect imaging
screening	slow	fast

Table 6-3. Comparison between the cryo-EM and X-ray crystallization in sample preparation and structure determination.

Cryo-EM exceeds X-ray crystallography in aspects such as sample preparation, molecular environment and imaging results (Table 6-3). For cryo-EM, each grid only requires 3-4 μL of sample with a concentration about 0.2-2 mg/mL, while crystallization requires significantly more sample and so is more expensive and laborious to prepare. In addition, the buffer choice for cryo-EM is relatively flexible, allowing examination of the structure under various conditions. Only glycerol cannot be used as it decreases the contrast between the nucleosome and solvent, and leads to the undesirable signal-to-noise ratio decrease. In contrast, crystallization has a high requirement on solution composition in order to crystallize. The search for suitable crystallization condition is mostly empirical and requires meticulous screening, patience and time for crystal growing, and luck. Thus obtaining the crystal alone can be a daunting and time-consuming mission. Growing crystals and solving structures by X-ray diffraction often require exceptionally high homogeneity of the molecule, so the molecules need to be of high purity and in one conformation. However, with cryo-EM, impurities can be cleaned *in silico*, and different confirmations can be classified and analysed separately. Without the need for growing crystals, the sample preparation for cryo-EM is relatively simple and fast. Finally, the results obtained by cryo-EM resemble the structures in aqueous solution, which is more relevant to provide explanation for phenomena observed in other aqueous-based experiments; whereas in X-ray crystallography, the crystalline samples might be in different conformations and hydration states from molecules in aqueous solution. In terms of imaging results, cryo-EM provides direct magnified images of target molecules, whereas this is not possible for X-ray crystallography, in which the phase and amplitude of diffraction are measured, and the electron cloud of the molecule can be computed.

Regardless, the cryo-EM does have its shortcomings compared to X-ray crystallography in aspects such as resolution and condition screening process. X-ray crystallography can reach near-atomic resolution, which can be rarely achieved by current cryo-EM techniques. Nevertheless, with the development of more powerful microscopes and computing algorithms, the resolution of cryo-EM is increasing rapidly. As for condition screening, thanks to the long history of X-ray crystallography, the crystallization condition screening has been largely automated (such as Dragon Fly, Phoenix and Mosquito), which allows high-throughput finecombing without too much hands-on time, whereas cryo-EM samples need to be loaded on grids separately, different conditions need to be manually prepared individually. Additionally, checking for the existence of crystals for X-ray crystallization is very swift, requiring only a few seconds inspection under a common light microscope. Moreover, with automated imaging system (e.g. Rock Maker), image checking schedule can be set, and the images will be automatically uploaded onto a website, eliminating the need for a light microscope. On the contrary, sample screening for cryo-EM requires access to a high-power electron microscope such as Talos Arctica (200 kV), maintained at demanding conditions (vacuum and cooled by liquid nitrogen), and the screening of even a single grid requests both expertise and hours for each condition.

Due to the length of time needed for each screening condition conflicting with large demand of the machine, the screening opportunities on Talos Arctica are very precious and scarce. Thus the search for suitable nucleosome conditions was first explored by native gel electrophoresis before examination on Talos Arctica.

6.6.3 Nucleosome Condition Screening and Optimization

6.6.3.1 Initial Screening and Conditions

Nucleosomes containing fully fC modified Widom DNA were assembled by salt assistant method (Figure 6-9), as the biological chaperones and PGA are too close to the size of the nucleosome, adding complexity to particle picking for structural reconstruction.

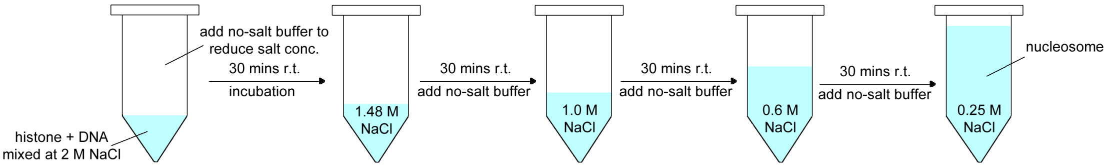
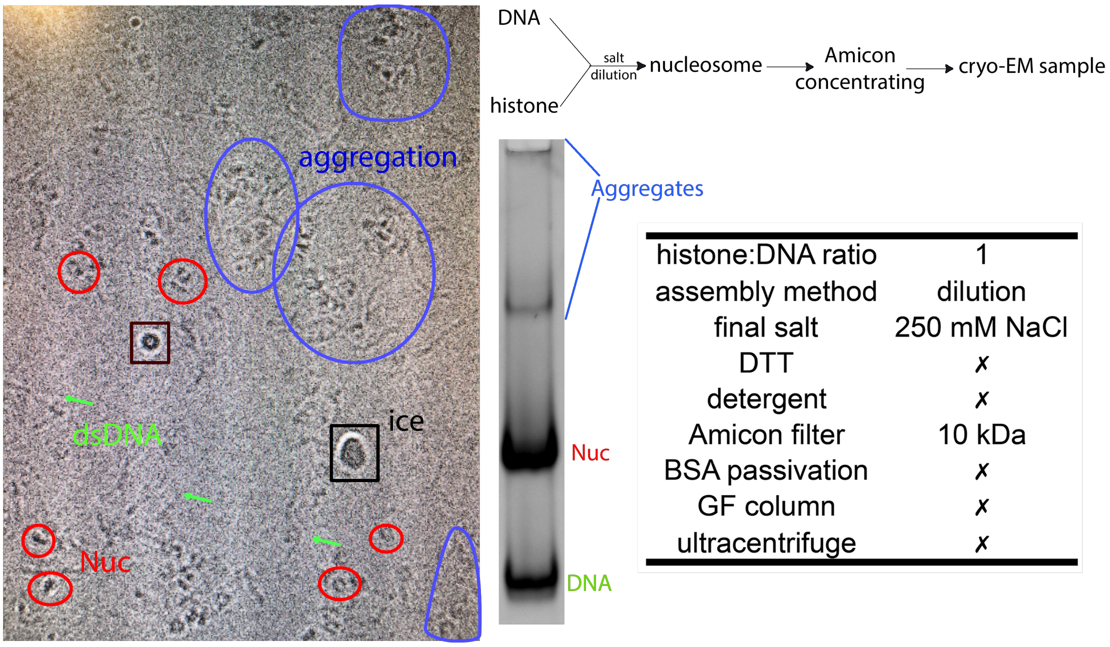


Figure 6-9. Schematic illustration for nucleosome assembly by salt dilution. The figure is not drawn to scale.

Recombinant histones were used to prevent the possible structural heterogeneity caused by various histone PMTs that exist in histone extracted from chromatin, and post difficulty for structural reconstruction. DNA and recombinant histone were combined at 2 M NaCl, completely shielding the electrostatic interaction between DNA and histone. The salt concentration was gradually reduced by diluting the reaction mixture with no salt buffer until the final salt concentration reached 0.25 M. During this process, DNA was gradually loaded onto histones to form nucleosomes in 63 ng/µL, 0.31 µM. In order to obtain enough particles per cryo-EM scan, the nucleosome was concentrated down by Amicon Ultra-0.5 centrifugal filter unit (working principle shown in Figure 6-11f) with MW cutoff of 10 kilo Dalton (kDa) (fC-nucleosome MW 201.6 kDa), until the sample volume reduced from 400 µL to about 30 µL. Then the nucleosome was checked on gel and inspected on Talos Arctica to obtain cryo-EM images.



histone:DNA ratio	1
assembly method	dilution
final salt	250 mM NaCl
DTT	✗
detergent	✗
Amicon filter	10 kDa
BSA passivation	✗
GF column	✗
ultracentrifuge	✗

Figure 6-10. workflow and conditions for nucleosome preparation, and results of native gel electrophoresis and initial cryo-EM attempt with examples of nucleosome (red), free DNA (green), aggregation (blue) and crystalline ice (black).

The gel image shows nucleosomes, DNA and a small amount of high-order aggregation, which migrates slower than nucleosomes due to its large size, as well as a small quantity of massive aggregation that remains in the loading well due to its size and charge.

Preliminary screening with cryo-EM (Figure 6-10) confirmed the observations obtained from gels. Side and top views of nucleosome particles (red circle) can be seen. The free DNA (green arrow) is also plainly visible on the grid, stretching approximately 450 angstroms (Å) end-to-end, which corresponds to the length of Widom DNA (147 bp, theoretical length 464.8 Å assuming perfect B-form DNA). The free DNA can be picked out *in silico*, and does not interfere with analysis. However, the aggregation of nucleosomes (blue circle) will affect structure determination, as these nucleosomes are not in well isolated states.

The sample condition required further optimization to increase the density of well-isolated nucleosome for 3D structural reconstruction. This issue was addressed from the following two angles: increasing nucleosome concentration and decreasing the aggregation.

6.6.3.2 Proposed Strategies for Condition Optimization

In order to increase the nucleosome concentration, the following methods were proposed (summarized in Figure 6-11a and b).

As there is free DNA present, the amount of nucleosomes might be increased by adding more histones to push the equilibrium towards the side of the nucleosomes (Figure 6-11c). In addition, nucleosome concentration can be increased by simply preparing more nucleosome sample and concentrating to a smaller volume. Moreover, nucleosomes could also be lost through non-specific interactions during sample preparation and concentrating. Such loss could be reduced by using siliconized tubes during nucleosome assembly, and passivized Amicon columns for concentrating. The passivation was done by filling the Amicon filter with freshly dissolved 1% BSA in the nucleosome assembly buffer, and incubating either overnight at 4°C or 2 hours at room temperature. The BSA solution was then discarded and the column filter was washed vigorously with the nucleosome assembly buffer three times to remove non-binding BSA. The nucleosome solution can be subsequently added into the filter for concentrating.

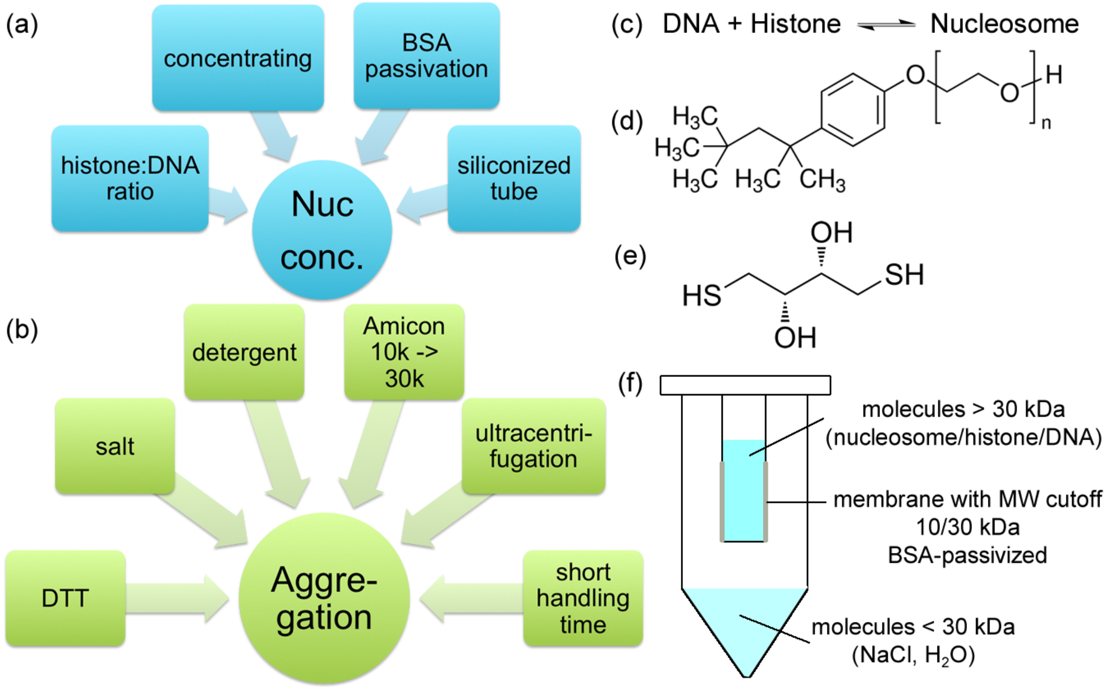


Figure 6-11. Proposed strategies and relevant information for nucleosome condition optimization: (a) to increase the nucleosome concentration; (b) to prevent/reduce aggregation formation; (c) equilibrium between DNA, histone and nucleosome;(d) structure of Triton X-100; (e) structure of DTT (f) Amicon column working principle.

Simply increasing the nucleosome concentration may exacerbate unwanted nucleosome aggregation, thus the following methods were attempted (Figure 6-11b).

First, 1 mM dithiothreitol (DTT, structure shown in Figure 6-11e) was added to break any disulfide bonds that formed between nucleosomes and undesirably strengthened the aggregation. Salt was also tested to reduce the aggregation by shielding the electrostatic interactions between nucleosomes. In addition, detergents such as Triton X-100 (structure shown in Figure 6-11d) could be employed to break up the aggregation, as well as reduce the nucleosome loss by non-specific interactions. Furthermore, the Amicon filter cut-off was increased from 10k to 30k to accelerate the concentrating step and reduce possible aggregation formed by local high concentration during spinning. An ultracentrifugation step for 10 minutes at 4°C at 10 krcf was added at the end of sample preparation to pellet down the precipitation formed by massive aggregation. Additionally, as the nucleosomes are possibly prone to form aggregation in the concentrated state, it could be vital to prepare grids immediately after the concentrating and ultracentrifugation.

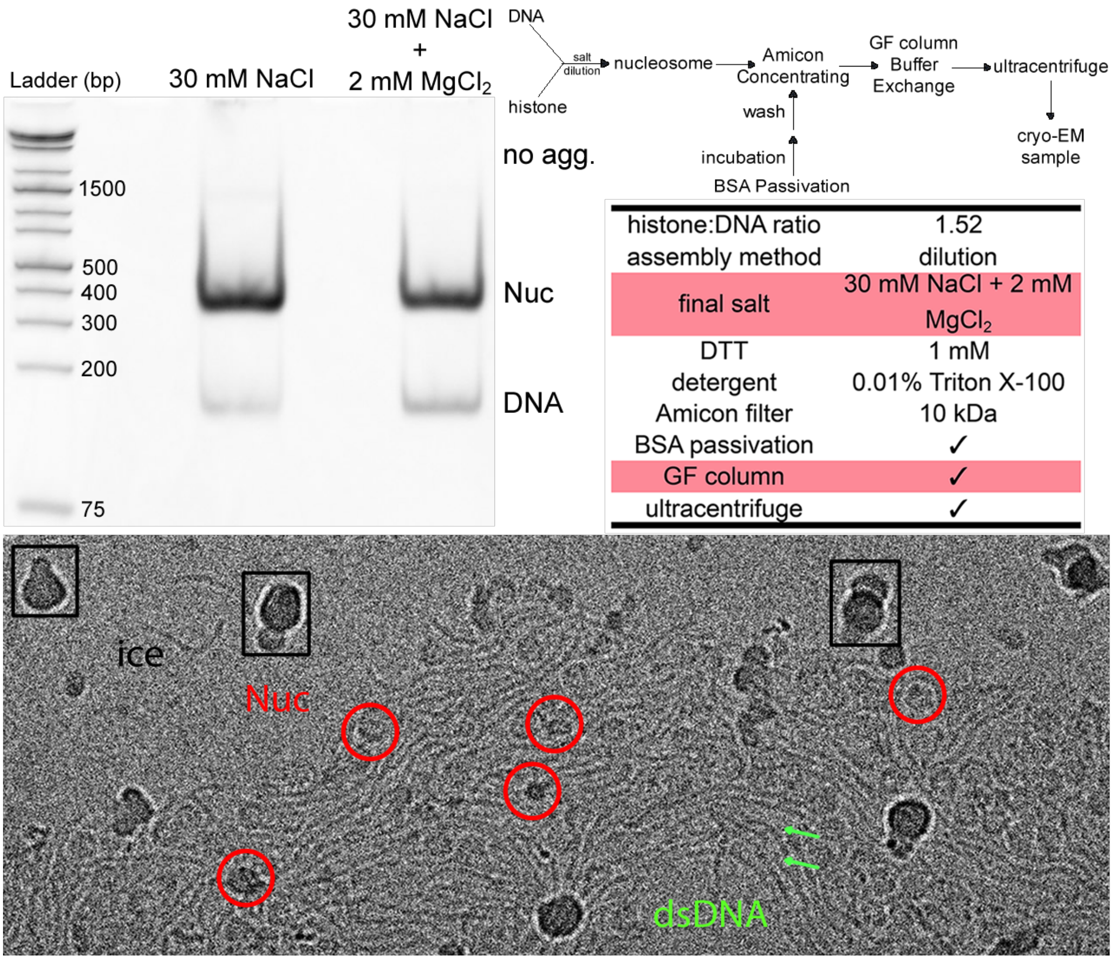
6.6.3.3 Strategy Validation with Native Gel Electrophoresis and Cryo-EM

All the strategies were first screened with 6% native DNA Retardation gels to ensure the nucleosome remains intact after implementing the changes, and then screened by cryo-EM Talos Arctica to see whether these strategies increased the sample condition.

6.6.3.3.1 Salt

Mg^{2+}, reported to compact nucleosome in FRET study[369], may help the nucleosomes stay well isolated for cryo-EM particle picking. Additionally, as the initial screening sample contained 250 mM NaCl, which might be the cause for aggregation, a concentration ranging from 0-250 mM was investigated. The salt concentration was adjusted by passing the concentrated nucleosomes through Gel Filtration (GF) columns equilibrated with desired final buffers as shown in the flowchart in Figure 6-12.

Native gel electrophoresis indicates nucleosomes were well formed with both Mg^{2+} and Na^+ without severe aggregation. However, cryo-EM revealed that upon concentrating, Mg^{2+} seems to act as glue and caused a gel like aggregation formed by free DNA and nucleosomes (Figure 6-12). Since the Mg^{2+} addition cannot produce isolated nucleosomes for structure reconstruction, Mg^{2+} was not used in the final condition. On the other hand, the decrease of NaCl concentration appeared to alleviate the aggregation. Thus no salt was used in the final condition.



histone:DNA ratio	1.52
assembly method	dilution
final salt	30 mM NaCl + 2 mM $MgCl_2$
DTT	1 mM
detergent	0.01% Triton X-100
Amicon filter	10 kDa
BSA passivation	✓
GF column	✓
ultracentrifuge	✓

Figure 6-12. The effect of salt addition on aggregation elimination. Although native gel electrophoresis indicated nucleosome is intact with Na^+ and Mg^{2+}, the cryo-EM showed that the Mg^{2+} can induce gel-like aggregation formation.

6.6.3.3.2 Detergent and GF column

Triton X-100 at concentration of 0.01% was added to break up the aggregation. Triton may also reduce the nucleosome lost through non-specific interaction during nucleosome assembly and concentrating.

Triton addition in the nucleosome solution has demonstrated favourable effects on reducing the aggregation checked by native gel electrophoresis (Figure 6-13a). The sample with Triton shows significant reduction of both the deposit in well (likely the massive aggregation that is too large or too positive to enter the well) as well as the in-gel aggregation present at around 1000 bp ladder. Triton has also made the reaction solution less prone to bubble during pipetting, and thus reduced the possible denaturation. However, upon checking with Talos Arctica, the nucleosome density of the sample with 0.01% Triton did not improve as compared to the no-Triton counterpart. Higher Triton concentration cannot be used as it may form micelle and interfere with the grid preparation and image analysis. As a result, Triton was not used in the final condition.

Besides the initial purpose of buffer exchange, GF column treatment has also shown the reduction of both the deposit in the well and in-gel aggregation (Figure 6-13a). However, after GF column buffer exchange, both with and without Triton samples have shown a decrease on the absolute nucleosome signal intensity (Figure 6-13b), despite the same volume of samples being applied on the gel, indicating that there might be sample dilution and/or sample loss during the GF treatment.

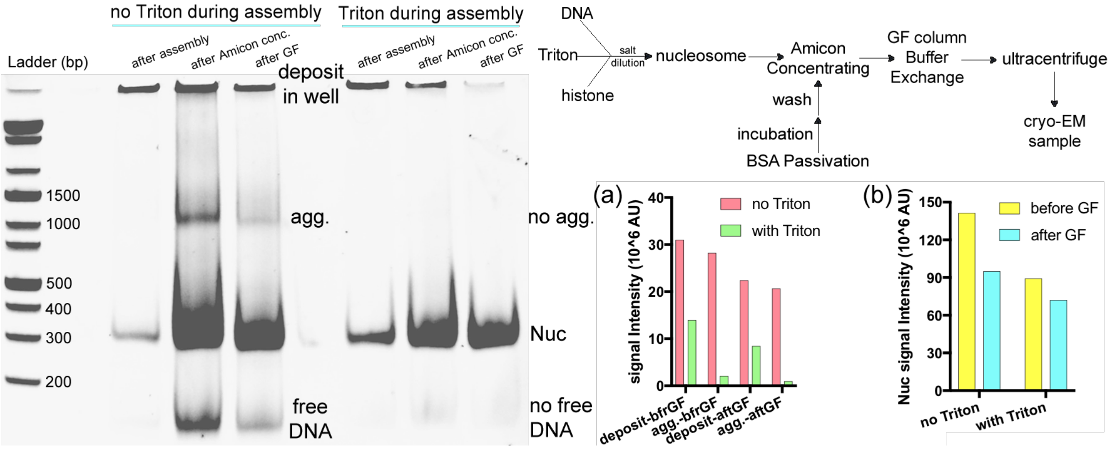


Figure 6-13. The effect of detergent, 0.01% Triton, and GF column on aggregation elimination: (a) plot of the quantification for the effect. Deposit refers to the deposit in the well that is too large to enter the gel; agg. refers to the aggregation present slightly around 1000 bp ladder. The effect was quantified before and after the GF column purification. The quantifications have taken into consideration both the sample dilution after GF column as well as the sample concentration difference between “no Triton” sample and “with Triton” sample; (b) effect of GF column treatment on the nucleosome concentration.

6.6.3.3.3 Histone:DNA Ratio

The equilibrium between histone, DNA and nucleosome can be pushed towards nucleosome formation by adding more DNA or histone, creating more nucleosome particles for structure reconstruction. Both histone in excess and DNA in excess have been tried.

A rise in histone input did increase the incorporation percentage of free DNA into nucleosome checked by native gel electrophoresis, however the absolute amount of nucleosome reduced when more histones were added. This indicates that the nucleosome might have been lost during sample preparation. Indeed, the cryo-EM (Figure 6-14) image indicated that the nucleosome particle density was really low for the histone-surplus samples.

Surprisingly, reducing histone input actually improved the absolute amount of nucleosome compared to the other two ratios tested, despite the same amount of DNA being used in all three samples. Cryo-EM has also shown that nucleosomes prepared with less histone than DNA are in a well-isolated state. Thus the final condition was set to be DNA in excess for nucleosome assembly.

This observation agrees with previous publication[265], that excessive histone can induce chromatin aggregation, while DNA in excess prevents the aggregation.

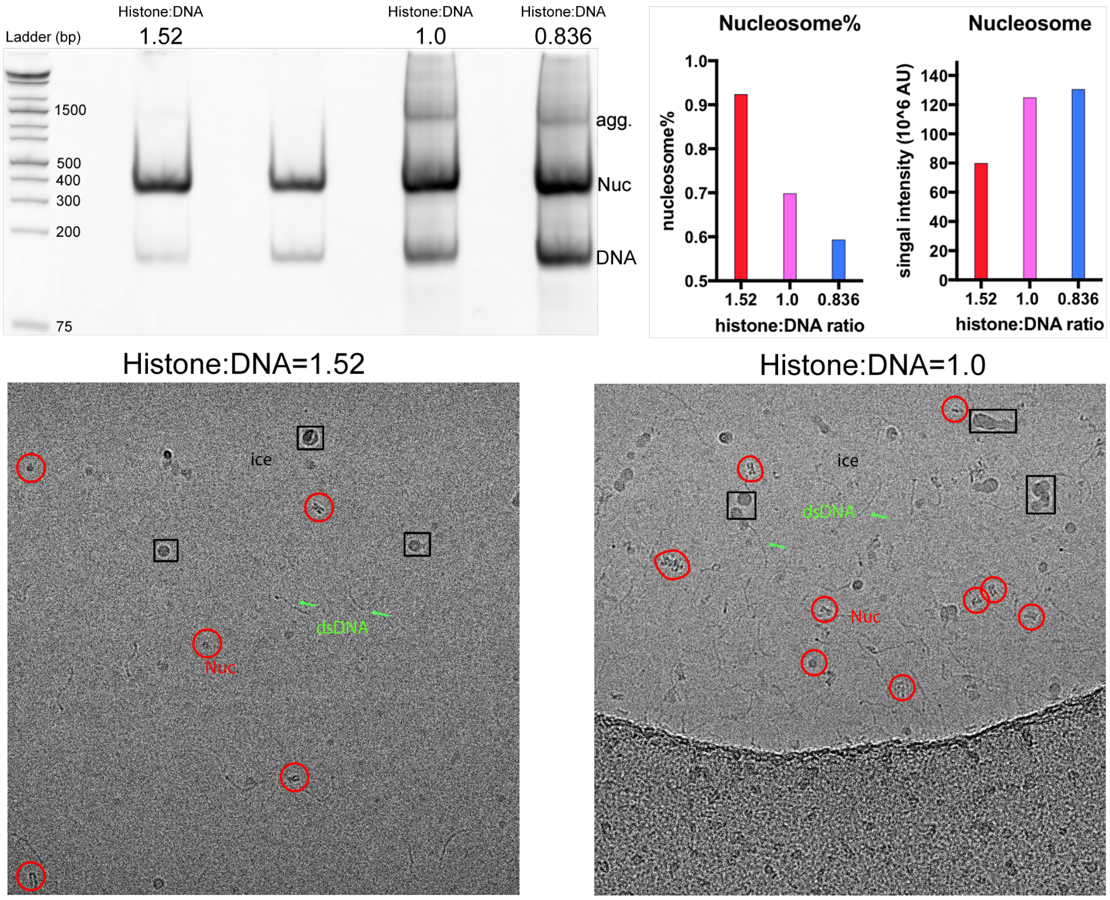


Figure 6-14. The effect of histone:DNA input ratio on nucleosome amount and particle density. In this experiment the DNA input was kept constant and histone input was varied for ratio screening.

6.6.3.3.4 Nucleosome Concentrating

For structure reconstruction, maximum nucleosome particles are required. However if the nucleosome concentration is too high, the basic histone tails may potentially interact with the acidic patch of the neighbouring nucleosomes[370, 371], and form aggregation. Thus different nucleosome concentrations were screened, and the cryo-EM results revealed that the concentration slightly below 2 mg/mL managed to give a good particle density without aggregation (Figure 6-15).

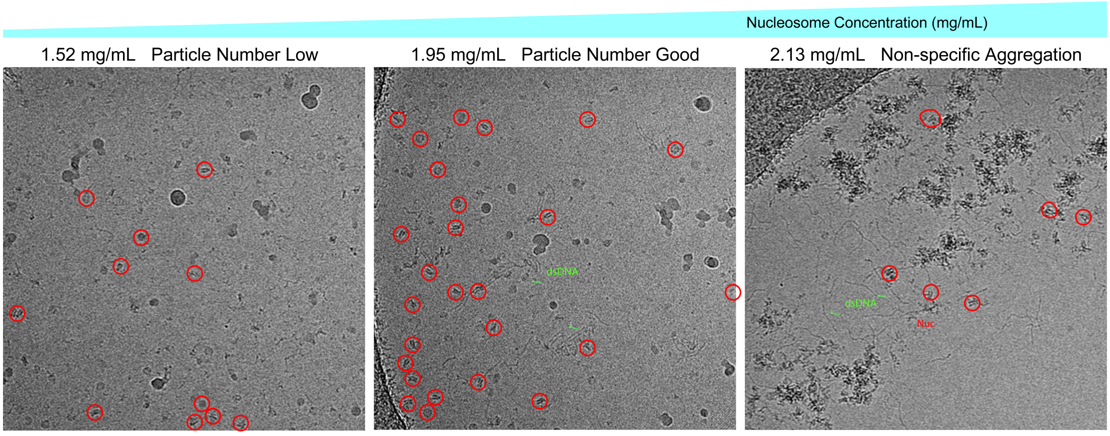

Nucleosome Concentration (mg/mL)
1.52 mg/mL Particle Number Low
1.95 mg/mL Particle Number Good
2.13 mg/mL Non-specific Aggregation
dsDNA
dsDNA
Nuc

Figure 6-15. The electron micrographs collected for nucleosome samples at different concentrations. 1.95 mL/mL gives the best particle number without massive aggregation. 1.52 mg/mL has too few particles present, while 2.13 mg/mL is too concentrated and nucleosomes are severely aggregated.

Aggregation could also form during concentrating with Amicon column. As nucleosomes tend to crowd at the bottom of the filter with prolonged concentrating spinning, a high local concentration can lead to aggregation formation. The concentrating step, therefore, needs to be kept to a minimum. The concentrating time is significantly shortened when using Amicon 30k instead of 10k MW cut-off, with the solution inside the filter pipette mixed every 2 minutes, to reduce local over-concentrating.

However, simply increasing the nucleosome concentration by shrinking the volume does have its limitations. It was observed that when the nucleosome was concentrated above 1.49 mg/mL, the precipitation started to emerge, and further concentrating even by a few microlitres (to 23 μL for 1.95 mg/mL and 21 μL for 2.13 mg/mL) had a big effect on the amount of precipitation. Precipitation also leads to decreased nucleosome density, thus during nucleosome concentrating it needs to be monitored closely for the precipitation once the concentration increased beyond 1.4 mg/mL.

6.6.3.3.5 Nucleosome Assembly by Dialysis vs. Dilution

In order to further reduce the time between the start of concentrating and grid preparation, the dialysis method (Figure 6-16) was used to assemble nucleosomes instead of the dilution method, for the following reasons:

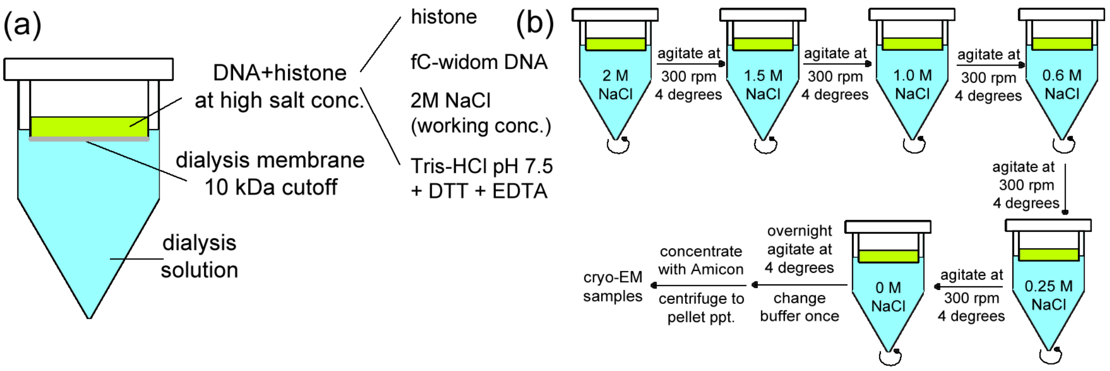


Figure 6-16. Schematics for (a) experimental setup and (b) workflow for nucleosome assembly by salt dialysis method.

The final nucleosome volume of salt dilution is drastically higher than salt dialysis due to the experimental setup. As shown in Figure 6-9, in order to reduce the salt concentration from 2 M NaCl to 0.25 M by dilution, the final sample volume will be eight times of the initial volume. On the contrary, the dialysis method allows the sample volume to stay virtually constant throughout the assembly. Thus the concentrating time for dialysis sample is significantly shorter than that for dilution sample, and thus less liable to aggregation formation.

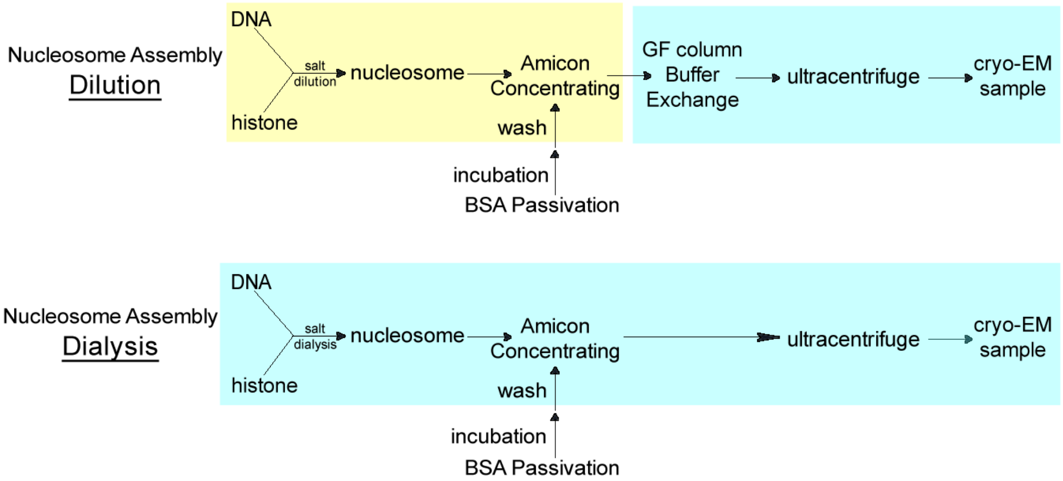


Figure 6-17. Comparison between the workflow of nucleosome assembly by salt dilution and salt dialysis. In the salt dilution method, the nucleosome is assembled in assembly buffer (marked in yellow), and changed to the desired buffer for cryo-EM (marked in blue) by GF column buffer exchange. In contrast, the salt dialysis method enables the nucleosome to remain in the desired buffer throughout.

In addition, the screening of different buffer conditions for the dilution method was achieved by passing the concentrated nucleosome through GF column equilibrated with desired buffer (Figure 6-17). In contrast, with the dialysis method, the buffer can be directly adjusted via dialysis, shortening the period between concentrating and grid preparation, and further reducing the chance for aggregation formation. The sample loss/dilution during the GF column treatment can also be avoided. In addition, since the buffer change in the dialysis method is done gradually, the dilution method is completed within a very short time and less environment shock will be posted on the nucleosomes.

The aggregation is also likely to develop during the period after concentrating but before freezing in liquid ethane. It is imperative to prepare grids immediately after concentrating.

6.6.3.3.6 Summary of Screening Results for Proposed Strategies

After screening various parameters (summarized in Table 6-4), the final condition was set (summarized in Figure 4-3). Nucleosomes were assembled with histones and DNA at the ratio of 0.853. Dialysis was used to gradually remove salt from the assembly mixture and deposit the DNA onto histone. The nucleosome was then dialyzed against no salt and no detergent buffer.

The nucleosome obtained was concentrated down with Amicon column with 30k MW cut-off, and then ultracentrifuged to remove the possible precipitate. The nucleosome sample was loaded onto grids and plunge frozen in liquid ethane. The grids were finally loaded onto Titan Krios to collect data for structure reconstruction (Section 4.2.2).

Parameters		rationale	native gel	cryo-EM
detergent	Triton X-100 0.01%	prevent aggregation formation	reduced aggregation	no improvement
salt	$MgCl_2$	reported to compact nucleosome,may help structure reconstruction	✓	severe aggregation
	NaCl	may help to balance the charge and stabilize individual nucleosome to prevent aggregation	✓	less aggregation is observed with low salt
histone:DNA molar ratio	1.52	histone in excess to push equilibrium towards nucleosome	✓	too much aggregation
	0.853	DNA in excess to push equilibrium towards nucleosome	✓	well separated nucleosome + free DNA
assembly method	Dilution	easy to prepare and less prone to contamination	✓	dissociation and aggregation observed
	Dialysis	can remove all salt from nucleosome solution	✓	well separated nucleosome + free DNA
BSA passivation	1%	to block non-specific interaction between nucleosome and Amicon membrane	increased yield	✓
Amicon	10 k	concentrate down nucleosome to increase partile density for imaging	✓	✓
	30 k	shorten the concentrating time and reduce aggregation formation	✓	well separated nucleosome + free DNA
gel filtration column		to get rid of salt and aggregation	less aggregation but less concentrated nucleosome	nucleosome particle in low density

Table 6-4. Summary of screening results for proposed strategies. The conditions highlighted in green were used as final condition to prepare nucleosomes for cryo-EM data collection for structure reconstruction.

7 References

1. Mendel, G. Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn.* **Bd.4 (1865-1866)**, 3-47 (1865).

2. Gayon, J. From Mendel to epigenetics: History of genetics. *C R Biol* **339**, 225-30 (2016).

3. Stamhuis, I.H., Meijer, O.G. & Zevenhuizen, E.J. Hugo de Vries on heredity, 1889-1903. Statistics, Mendelian laws, pangenes, mutations. *Isis* **90**, 238-67 (1999).

4. De Vries, H. Intracellulare pangenesis. *Chicago, the Open Court publishing Co.* (1889).

5. Flemming, W. Zur Kenntniss der Zelle und ihrer Theilungs-Erscheinungen. *Schriften des Naturwissenschaftlichen Vereins für Schleswig-Holstein*, 23–27. (1878).

6. Boveri, T. Ergebnisse über die Konstitution der chromatischen Substanz des Zellkerns (G. Fischer, Jena, 1904).

7. S. Sutton, W. On the morphology of the chromosome group in Brachystola magna (Biological Bulletin, 1902).

8. Sutton, W.S. The chromosomes in heredity. *The Biological Bulletin* **4**, 231-250 (1903).

9. Sturtevant, A.H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology* **14**, 43-59 (1913).

10. Griffith, F. The Significance of Pneumococcal Types. *J Hyg (Lond)* **27**, 113-59 (1928).

11. Avery, O.T., MacLeod, C.M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *The Journal of Experimental Medicine* **79**, 137 (1944).

12. Hershey, A.D. & Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**, 39-56 (1952).

13. Dahm, R. Friedrich Miescher and the discovery of DNA. *Developmental Biology* **278**, 274-288 (2005).

14. Xie, C. & O'Leary, J.P. Vignettes in Medical History - DNA and the Brains Behind Its Discovery. *The American Surgeon* **62**, 979-980 (1996).

15. Unger, B. Bemerkungen zu obiger Notiz. *Justus Liebigs Annalen der Chemie* **58**, 18-20 (1846).

16. Kossel, A. Ueber eine neue Base aus dem Thierkörper. *Berichte der deutschen chemischen Gesellschaft* **18**, 79-81 (1885).

17. Kossel, A. & Neumann, A. Ueber das Thymin, ein Spaltungsproduct der Nucleïnsäure. *Berichte der deutschen chemischen Gesellschaft* **26**, 2753-2756 (1893).

18. Kossel, A. & Steudel, H. in Hoppe-Seyler´s Zeitschrift für physiologische Chemie 49 (1903).

19. Kossel, A. & Neumann, A. Darstellung und Spaltungsprodukte der Nucleïnsäure (Adenylsäure). *Berichte der deutschen chemischen Gesellschaft* **27**, 2215-2222 (1894).

20. Levene, P.A. & London, E.S. The structure of thymonucleic acid. *Journal of Biological Chemistry* **83**, 793-802 (1929).

21. Leuchtenberger, C., Vendrely, R. & Vendrely, C. A Comparison of the Content of Desoxyribosenucleic Acid (DNA) in Isolated Animal Nuclei by Cytochemical and Chemical Methods. *Proceedings of the National Academy of Sciences of the United States of America* **37**, 33-38 (1951).

22. Vendrely, R. & Vendrely, C. La teneur du noyau cellulaire en acide désoxyribonucléique à travers les organes, les individus et les espèces animales. *Experientia* **4**, 434-6 (1948).

23. Boivin, A., Vendrely, R. & Vendrely, C. L'acide désoxyribonucléique du noyau cellulaire, dépositaire des caractères héréditaires; arguments d'ordre analytique. *C R Hebd Seances Acad Sci* **226**, 1061-3 (1948).

24. Chargaff, E., Lipshitz, R. & Green, C. Composition of the desoxypentose nucleic acids of four genera of sea-urchin. *Journal of Biological Chemistry* **195**, 155-160 (1951).

25. Chargaff, E. Structure and function of nucleic acids as cell constituents. *Fed Proc* **10**, 654-9 (1951).

26. Franklin, R.E. & Gosling, R.G. The structure of sodium thymonucleate fibres. I. The influence of water content. *Acta Crystallographica* **6**, 673-677 (1953).

27. Watson, J.D. & Crick, F.H.C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737-738 (1953).

28. Dickerson, R.E. The DNA Helix and How it is Read. *Scientific American* **249**, 94-111 (1983).

29. Fuller, W., Wilkins, M.H., Wilson, H.R. & Hamilton, L.D. The molecular configuration of deoxyribonucleic acid. IV. X-ray diffraction study of the A form. *J Mol Biol* **12**, 60-76 (1965).

30. Wang, A.H., Quigley, G.J., Kolpak, F.J., Crawford, J.L., van Boom, J.H., van der Marel, G. & Rich, A. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* **282**, 680-6 (1979).

31. Herbert, A.G., Spitzner, J.R., Lowenhaupt, K. & Rich, A. Z-DNA binding protein from chicken blood nuclei. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 3339-3342 (1993).

32. Kim, Y.-G., Lowenhaupt, K., Oh, D.-B., Kim, K.K. & Rich, A. Evidence that vaccinia virulence factor E3L binds to Z-DNA in vivo: Implications for development of a therapy for poxvirus infection. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 1514 (2004).

33. Wittig, B., Wölfl, S., Dorbic, T., Vahrson, W. & Rich, A. Transcription of human c-myc in permeabilized nuclei is associated with formation of Z-DNA in three discrete regions of the gene. *The EMBO Journal* **11**, 4653-4663 (1992).

34. Marvin, D.A., Spencer, M., Wilkins, M.H. & Hamilton, L.D. The molecular configuration of deoxyribonucleic acid. III. X-ray diffraction study of the C form of the lithium salt. *J Mol Biol* **3**, 547-65 (1961).

35. Premilat, S. & Albiser, G. A new D-DNA form of poly(dA-dT).poly(dA-dT): an A-DNA type structure with reversed Hoogsteen pairing. *European Biophysics Journal* **30**, 404-410 (2001).

36. Vargason, J.M., Eichman, B.F. & Ho, P.S. The extended and eccentric E-DNA structure induced by cytosine methylation or bromination. *Nat Struct Biol* **7**, 758-61 (2000).

37. Raiber, E.A., Murat, P., Chirgadze, D.Y., Beraldi, D., Luisi, B.F. & Balasubramanian, S. 5-Formylcytosine alters the structure of the DNA double helix. *Nat Struct Mol Biol* **22**, 44-9 (2015).

38. Peters, M., Rozas, I., Alkorta, I. & Elguero, J. DNA Triplexes: A Study of Their Hydrogen Bonds. *The Journal of Physical Chemistry B* **107**, 323-330 (2003).

39. Gehring, K., Leroy, J.-L. & Guéron, M. A tetrameric DNA structure with protonated cytosine-cytosine base pairs. *Nature* **363**, 561 (1993).

40. Gellert, M., Lipsett, M.N. & Davies, D.R. Helix formation by guanylic acid. *Proceedings of the National Academy of Sciences* **48**, 2013 (1962).

41. Chou, S.H., Chin, K.H. & Wang, A.H.J. Unusual DNA duplex and hairpin motifs. *Nucleic Acids Research* **31**, 2461-2474 (2003).

42. Murchie, A.I.H. & Lilley, D.M.J. in Methods in Enzymology 158-180 (Academic Press, 1992).

43. Hays, F.A., Watson, J. & Ho, P.S. Caution! DNA crossing: crystal structures of Holliday junctions. *J Biol Chem* **278**, 49663-6 (2003).

44. Frederick, C.A., Grable, J., Melia, M., Samudzi, C., Jen-Jacobson, L., Wang, B.C., Greene, P., Boyer, H.W. & Rosenberg, J.M. Kinked DNA in crystalline complex with EcoRI endonuclease. *Nature* **309**, 327-31 (1984).

45. Kool, E.T. Hydrogen bonding, base stacking, and steric effects in dna replication. *Annu Rev Biophys Biomol Struct* **30**, 1-22 (2001).

46. Seeman, N.C., Rosenberg, J.M. & Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A* **73**, 804-8 (1976).

47. Zubay, G. & Doty, P. The isolation and properties of deoxyribonucleoprotein particles containing single nucleic acid molecules. *Journal of Molecular Biology* **1**, 1-IN1 (1959).

48. Lehninger, A.L., Nelson, D.L. & Cox, M.M. Lehninger principles of biochemistry. *New York W.H. Freeman* (2008).

49. Richmond, T.J., Finch, J.T., Rushton, B., Rhodes, D. & Klug, A. Structure of the nucleosome core particle at 7 Å resolution. *Nature* **311**, 532 (1984).

50. Uberbacher, E.C. & Bunick, G.J. Structure of the Nucleosome Core Particle at 8 Å Resolution. *Journal of Biomolecular Structure and Dynamics* **7**, 1-18 (1989).

51. Vasudevan, D., Chua, E.Y.D. & Davey, C.A. Crystal Structures of Nucleosome Core Particles Containing the ‘601’ Strong Positioning Sequence. *Journal of Molecular Biology* **403**, 1-10 (2010).

52. Zheng, C. & Hayes, J.J. Structures and interactions of the core histone tail domains. *Biopolymers* **68**, 539-46 (2003).

53. Thoma, F., Koller, T. & Klug, A. Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *J Cell Biol* **83**, 403-27 (1979).

54. Scheinfeld, A. & Schweitzer, M.D. You and heredity. *New York, Frederick A. Stokes Co.* (1939).

55. Finch, J.T., Noll, M. & Kornberg, R.D. Electron microscopy of defined lengths of chromatin. *Proceedings of the National Academy of Sciences of the United States of America* **72**, 3320-3322 (1975).

56. Langmore, J.P. & Wooley, J.C. Chromatin Architecture: Investigation of a Subunit of Chromatin by Dark Field Electron Microscopy. *Proceedings of the National Academy of Sciences of the United States of America* **72**, 2691-2695 (1975).

57. Oudet, P., Gross-Bellard, M. & Chambon, P. Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell* **4**, 281-300 (1975).

58. International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931 (2004).

59. Arrowsmith, C.H., Bountra, C., Fish, P.V., Lee, K. & Schapira, M. Epigenetic protein families: a new frontier for drug discovery. *Nat Rev Drug Discov* **11**, 384-400 (2012).

60. Nielsen, A.L., Oulad-Abdelghani, M., Ortiz, J.A., Remboutsika, E., Chambon, P. & Losson, R. Heterochromatin Formation in Mammalian Cells: Interaction between Histones and HP1 Proteins. *Molecular Cell* **7**, 729-739 (2001).

61. Hirano, T. Condensin-Based Chromosome Organization from Bacteria to Vertebrates. *Cell* **164**, 847-57 (2016).

62. Wood, A.J., Severson, A.F. & Meyer, B.J. Condensin and cohesin complexity: the expanding repertoire of functions. *Nature Reviews Genetics* **11**, 391 (2010).

63. Ono, T., Losada, A., Hirano, M., Myers, M.P., Neuwald, A.F. & Hirano, T. Differential contributions of condensin I and condensin II to mitotic chromosome architecture in vertebrate cells. *Cell* **115**, 109-21 (2003).

64. Waddington, C.H. The epigenotype. *Int J Epidemiol* **41**, 10-3 (1942).

65. Waddington, C.H. Towards a theoretical biology. *Nature* **218**, 525-7 (1968).

66. Berger, S.L., Kouzarides, T., Shiekhattar, R. & Shilatifard, A. An operational definition of epigenetics. *Genes Dev* **23**, 781-3 (2009).

67. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693-705 (2007).

68. Bird, A. Perceptions of epigenetics. *Nature* **447**, 396-8 (2007).

69. Roadmap Epigenomics Project (2010) http://www.roadmapepigenomics.org/

70. Richards, E.J. & Elgin, S.C.R. Epigenetic Codes for Heterochromatin Formation and Silencing: Rounding up the Usual Suspects. *Cell* **108**, 489-500 (2002).

71. Nightingale, K.P., O'Neill, L.P. & Turner, B.M. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Current Opinion in Genetics & Development* **16**, 125-136 (2006).

72. Guil, S. & Esteller, M. DNA methylomes, histone codes and miRNAs: Tying it all together. *The International Journal of Biochemistry & Cell Biology* **41**, 87-95 (2009).

73. Goldberg, A.D., Allis, C.D. & Bernstein, E. Epigenetics: A Landscape Takes Shape. *Cell* **128**, 635-638 (2007).

74. Chahwan, R., Wontakal, S.N. & Roa, S. The multidimensional nature of epigenetic information and its role in disease. *Discov Med* **11**, 233-43 (2011).

75. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet* **10**, 295-304 (2009).

76. Moore, L.D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38**, 23-38 (2013).

77. Nan, X., Ng, H.H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N. & Bird, A. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386-9 (1998).

78. Ruppel, W.G. in Hoppe-Seyler´s Zeitschrift für physiologische Chemie 218 (1899).

79. Raiber, E.-A., Hardisty, R., van Delft, P. & Balasubramanian, S. Mapping and elucidating the function of modified bases in DNA. **1**, 0069 (2017).

80. Johnson, T.B. & Coghill, R.D. Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus1. *Journal of the American Chemical Society* **47**, 2838-2844 (1925).

81. Hotchkiss, R.D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *Journal of Biological Chemistry* **175**, 315-332 (1948).

82. Gommers-Ampt, J.H. & Borst, P. Hypermodified bases in DNA. *FASEB J* **9**, 1034-42 (1995).

83. Smith, S.S., Kaplan, B.E., Sowers, L.C. & Newman, E.M. Mechanism of human methyl-directed DNA methyltransferase and the fidelity of

cytosine methylation. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 4744-4748 (1992).

84. Goll, M.G. & Bestor, T.H. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* **74**, 481-514 (2005).

85. Svedruzic, Z.M. Mammalian cytosine DNA methyltransferase Dnmt1: enzymatic mechanism, novel mechanism-based inhibitors, and RNA-directed DNA methylation. *Curr Med Chem* **15**, 92-106 (2008).

86. Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet* **19**, 219-20 (1998).

87. Okano, M., Bell, D.W., Haber, D.A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-57 (1999).

88. Goyal, R., Reinhardt, R. & Jeltsch, A. Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. *Nucleic Acids Research* **34**, 1182-1188 (2006).

89. Fatemi, M., Hermann, A., Pradhan, S. & Jeltsch, A. The activity of the murine DNA methyltransferase Dnmt1 is controlled by interaction of the catalytic domain with the N-terminal part of the enzyme leading to an allosteric activation of the enzyme after binding to methylated DNA11Edited by J. Karn. *Journal of Molecular Biology* **309**, 1189-1199 (2001).

90. Barau, J., Teissandier, A., Zamudio, N., Roy, S., Nalesso, V., Herault, Y., Guillou, F. & Bourc'his, D. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science* **354**, 909-912 (2016).

91. Wachter, E., Quante, T., Merusi, C., Arczewska, A., Stewart, F., Webb, S. & Bird, A. Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *eLife* **3**, e03397 (2014).

92. Krebs, A.R., Dessus-Babus, S., Burger, L. & Schubeler, D. High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *Elife* **3**, e04094 (2014).

93. Perera, F. & Herbstman, J. Prenatal environmental exposures, epigenetics, and disease. *Reprod Toxicol* **31**, 363-73 (2011).

94. Day, J.J. & Sweatt, J.D. Cognitive neuroepigenetics: a role for epigenetic mechanisms in learning and memory. *Neurobiol Learn Mem* **96**, 2-12 (2011).

95. Day, J.J. & Sweatt, J.D. DNA methylation and memory formation. *Nature Neuroscience* **13**, 1319 (2010).

96. Gehring, M., Reik, W. & Henikoff, S. DNA demethylation by DNA repair. *Trends Genet* **25**, 82-90 (2009).

97. Kriaucionis, S. & Heintz, N. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science* **324**, 929-930 (2009).

98. Pfaffeneder, T., Hackner, B., Truss, M., Munzel, M., Muller, M., Deiml, C.A., Hagemeier, C. & Carell, T. The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew Chem Int Ed Engl* **50**, 7008-12 (2011).

99. He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.-X., Zhang, K., He, C. & Xu, G.-L. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science* **333**, 1303-1307 (2011).

100. Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A., Rappsilber, J. & Helin, K. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-8 (2011).

101. Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L. & Rao, A. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-5 (2009).

102. Wu, H. & Zhang, Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes Dev* **25**, 2436-52 (2011).

103. Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C. & Zhang, Y. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300-3 (2011).

104. Hu, L., Li, Z., Cheng, J., Rao, Q., Gong, W., Liu, M., Shi, Y.G., Zhu, J., Wang, P. & Xu, Y. Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation. *Cell* **155**, 1545-1555 (2013).

105. Kohli, R.M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472-9 (2013).

106. Tan, L. & Shi, Y.G. Tet family proteins and 5-hydroxymethylcytosine in development and disease. *Development* **139**, 1895-902 (2012).

107. Iyer, L.M., Tahiliani, M., Rao, A. & Aravind, L. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle* **8**, 1698-710 (2009).

108. Maiti, A. & Drohat, A.C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem* **286**, 35334-8 (2011).

109. Hashimoto, H., Hong, S., Bhagwat, A.S., Zhang, X. & Cheng, X. Excision of 5-hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase domain: its structural basis and implications for active DNA demethylation. *Nucleic Acids Res* **40**, 10203-14 (2012).

110. Zhang, L., Lu, X., Lu, J., Liang, H., Dai, Q., Xu, G.L., Luo, C., Jiang, H. & He, C. Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nat Chem Biol* **8**, 328-30 (2012).

111. Ngo, T.T.M., Yoo, J., Dai, Q., Zhang, Q., He, C., Aksimentiev, A. & Ha, T. Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat Commun* **7** (2016).

112. Yang, W. Structure and mechanism for DNA lesion recognition. *Cell Res* **18**, 184-97 (2008).

113. Dai, Q., Sanstead, P.J., Peng, C.S., Han, D., He, C. & Tokmakoff, A. Weakened N3 Hydrogen Bonding by 5-Formylcytosine and 5-Carboxylcytosine Reduces Their Base-Pairing Stability. *ACS Chemical Biology* **11**, 470-477 (2016).

114. Maiti, A., Michelson, A.Z., Armwood, C.J., Lee, J.K. & Drohat, A.C. Divergent mechanisms for enzymatic excision of 5-formylcytosine and 5-carboxylcytosine from DNA. *J Am Chem Soc* **135**, 15813-22 (2013).

115. Bogdanović, O., Smits, A.H., Calle, M.E., Tena, J.J., Ford, E. & Williams, R. Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat Genet.* **48** (2016).

116. Guo, F., Li, X., Liang, D., Li, T., Zhu, P., Guo, H., Wu, X., Wen, L., Gu, T.P., Hu, B., Walsh, C.P., Li, J., Tang, F. & Xu, G.L. Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell* **15**, 447-459 (2014).

117. Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, A.C., Fung, H.L., Zhang, K. & Zhang, Y. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692-706 (2013).

118. Wang, L., Zhang, J., Duan, J., Gao, X., Zhu, W., Lu, X., Yang, L., Zhang, J., Li, G., Ci, W., Li, W., Zhou, Q., Aluru, N., Tang, F., He, C., Huang, X. & Liu, J. Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**, 979-991 (2014).

119. Cortázar, D., Kunz, C., Selfridge, J., Lettieri, T., Saito, Y. & MacDougall, E. Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature.* **470** (2011).

120. Dawlaty, M.M., Breiling, A., Le, T., Barrasa, M.I., Raddatz, G., Gao, Q., Powell, B.E., Cheng, A.W., Faull, K.F., Lyko, F. & Jaenisch, R. Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Dev Cell* **29**, 102-11 (2014).

121. Liutkeviciute, Z., Lukinavicius, G., Masevicius, V., Daujotyte, D. & Klimasauskas, S. Cytosine-5-methyltransferases add aldehydes to DNA. *Nat Chem Biol* **5**, 400-2 (2009).

122. Iwan, K., Rahimoff, R., Kirchner, A., Spada, F., Schroder, A.S., Kosmatchev, O., Ferizaj, S., Steinbacher, J., Parsa, E., Muller, M. & Carell, T. 5-Formylcytosine to cytosine conversion by C-C bond cleavage in vivo. *Nat Chem Biol* (2017).

123. Schiesser, S., Hackner, B., Pfaffeneder, T., Muller, M., Hagemeier, C., Truss, M. & Carell, T. Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing. *Angew Chem Int Ed Engl* **51**, 6516-20 (2012).

124. Xu, S., Li, W., Zhu, J., Wang, R., Li, Z., Xu, G.L. & Ding, J. Crystal structures of isoorotate decarboxylases reveal a novel catalytic mechanism of 5-carboxyl-uracil decarboxylation and shed light on the search for DNA decarboxylase. *Cell Res* **23**, 1296-309 (2013).

125. Schiesser, S., Pfaffeneder, T., Sadeghian, K., Hackner, B., Steigenberger, B., Schroder, A.S., Steinbacher, J., Kashiwazaki, G., Hofner, G., Wanner, K.T., Ochsenfeld, C. & Carell, T. Deamination, oxidation, and C-C bond cleavage reactivity of 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxycytosine. *J Am Chem Soc* **135**, 14593-9 (2013).

126. Liutkevičiūtė, Z., Kriukienė, E., Ličytė, J., Rudytė, M., Urbanavičiūtė, G. & Klimašauskas, S. Direct Decarboxylation of 5-Carboxylcytosine by DNA C5- Methyltransferases. *Journal of the American Chemical Society* **136**, 5884-5887 (2014).

127. Illingworth, R.S. & Bird, A.P. CpG islands--'a rough guide'. *FEBS Lett* **583**, 1713-20 (2009).

128. International Human Genome Sequencing, C. Initial sequencing and analysis of the human genome. *Nature* **409**, 860 (2001).

129. Privat, E. & Sowers, L.C. Photochemical deamination and demethylation of 5-methylcytosine. *Chem Res Toxicol* **9**, 745-50 (1996).

130. Pfaffeneder, T., Spada, F., Wagner, M., Brandmayr, C., Laube, S.K., Eisen, D., Truss, M., Steinbacher, J., Hackner, B., Kotljarova, O., Schuermann, D., Michalakis, S., Kosmatchev, O., Schiesser, S., Steigenberger, B., Raddaoui, N., Kashiwazaki, G., Müller, U., Spruijt, C.G., Vermeulen, M., Leonhardt, H., Schär, P., Müller, M. & Carell, T. Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. *Nat Chem Biol* **10**, 574-581 (2014).

131. Wyatt, G.R. Occurrence of 5-Methyl-Cytosine in Nucleic Acids. *Nature* **166**, 237 (1950).

132. Millar, D.S., Holliday, R. & Grigg, G.W. in The Epigenome 1-20 (Wiley-VCH Verlag GmbH & Co. KGaA, 2005).

133. Ehrlich, M. & Wang, R.Y. 5-Methylcytosine in eukaryotic DNA. *Science* **212**, 1350-7 (1981).

134. Deaton, A.M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010-22 (2011).

135. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**, 484-92 (2012).

136. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**, 6-21 (2002).

137. Yan, C. & Boyd, D.D. Histone H3 Acetylation and H3 K4 Methylation Define Distinct Chromatin Regions Permissive for Transgene Expression. *Molecular and Cellular Biology* **26**, 6357-6371 (2006).

138. Siegfried, Z. & Simon, I. DNA methylation and gene expression. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **2**, 362-371 (2009).

139. Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A.J., Socci, N.D. & Scandura, J.M. DNA Methylation of the First Exon Is Tightly Linked to Transcriptional Silencing. *PLOS ONE* **6**, e14524 (2011).

140. Suzuki, M.M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**, 465-76 (2008).

141. Reik, W. & Allen, N.D. Genomic Imprinting: Imprinting with and without methylation. *Current Biology* **4**, 145-147 (1994).

142. Robertson, K.D. DNA methylation and chromatin - unraveling the tangled web. *Oncogene* **21**, 5361-79 (2002).

143. Hackett, J.A., Sengupta, R., Zylicz, J.J., Murakami, K., Lee, C., Down, T.A. & Surani, M.A. Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. *Science* **339**, 448-52 (2013).

144. Gaston, K. & Jayaraman, P.S. Transcriptional repression in eukaryotes: repressors and repression mechanisms. *Cell Mol Life Sci* **60**, 721-41 (2003).

145. Palacios, D., Summerbell, D., Rigby, P.W.J. & Boyes, J. Interplay between DNA Methylation and Transcription Factor Availability: Implications for Developmental Activation of the Mouse Myogenin Gene. *Molecular and Cellular Biology* **30**, 3805-3815 (2010).

146. Barlow, D.P. & Bartolomei, M.S. Genomic imprinting in mammals. *Cold Spring Harb Perspect Biol* **6** (2014).

147. Rose, N.R. & Klose, R.J. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1839**, 1362-1372 (2014).

148. Gama-Sosa, M.A., Midgett, R.M., Slagel, V.A., Githens, S., Kuo, K.C., Gehrke, C.W. & Ehrlich, M. Tissue-specific differences in DNA methylation in various mammals. *Biochim Biophys Acta* **740**, 212-9 (1983).

149. Yoder, J.A., Walsh, C.P. & Bestor, T.H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* **13**, 335-40 (1997).

150. Du, J., Johnson, L.M., Jacobsen, S.E. & Patel, D.J. DNA methylation pathways and their crosstalk with histone methylation. *Nature Reviews Molecular Cell Biology* **16**, 519 (2015).

151. Watt, F. & Molloy, P.L. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev* **2**, 1136-43 (1988).

152. Nan, X., Meehan, R.R. & Bird, A. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Res* **21**, 4886-92 (1993).

153. Bird, A.P. & Wolffe, A.P. Methylation-induced repression--belts, braces, and chromatin. *Cell* **99**, 451-4 (1999).

154. Jones, P.L., Veenstra, G.J., Wade, P.A., Vermaak, D., Kass, S.U., Landsberger, N., Strouboulis, J. & Wolffe, A.P. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet* **19**, 187-91 (1998).

155. Baylin, S.B. & Jones, P.A. A decade of exploring the cancer epigenome — biological and translational implications. *Nature Reviews Cancer* **11**, 726 (2011).

156. Varley, K.E., Gertz, J., Bowling, K.M., Parker, S.L., Reddy, T.E., Pauli-Behn, F., Cross, M.K., Williams, B.A., Stamatoyannopoulos, J.A., Crawford, G.E., Absher, D.M., Wold, B.J. & Myers, R.M. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res* **23**, 555-67 (2013).

157. Ushijima, T. & Asada, K. Aberrant DNA methylation in contrast with mutations. *Cancer Sci* **101**, 300-5 (2010).

158. Teng, I.W., Hou, P.C., Lee, K.D., Chu, P.Y., Yeh, K.T., Jin, V.X., Tseng, M.J., Tsai, S.J., Chang, Y.S., Wu, C.S., Sun, H.S., Tsai, K.D., Jeng, L.B., Nephew, K.P., Huang, T.H., Hsiao, S.H. & Leu, Y.W. Targeted methylation of two tumor suppressor genes is sufficient to transform mesenchymal stem cells into cancer stem/initiating cells. *Cancer Res* **71**, 4653-63 (2011).

159. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics* **8**, 286 (2007).

160. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nat Biotechnol* **28**, 1057-68 (2010).

161. Branco, M.R., Ficz, G. & Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet* **13**, 7-13 (2012).

162. Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Anselmi, F., Parlato, C., Medana, C., Dal Bello, F. & Oliviero, S. Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics. *Cell Reports* **10**, 674-683.

163. Iurlaro, M., Ficz, G., Oxley, D., Raiber, E.A., Bachman, M., Booth, M.J., Andrews, S., Balasubramanian, S. & Reik, W. A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol* **14**, R119 (2013).

164. Iurlaro, M., McInroy, G.R., Burgess, H.E., Dean, W., Raiber, E.-A., Bachman, M., Beraldi, D., Balasubramanian, S. & Reik, W. In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome Biology* **17**, 141 (2016).

165. Raiber, E.-A., Beraldi, D., Ficz, G., Burgess, H., Branco, M., Murat, P., Oxley, D., Booth, M., Reik, W. & Balasubramanian, S. Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biology* **13**, R69 (2012).

166. Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W., Bauer, C., Munzel, M., Wagner, M., Muller, M., Khan, F., Eberl, H.C., Mensinga, A., Brinkman, A.B., Lephikov, K., Muller, U., Walter, J., Boelens, R., van Ingen, H., Leonhardt, H., Carell, T. & Vermeulen, M. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146-59 (2013).

167. Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S.C., Mann, M. & Kouzarides, T. Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* **143**, 470-84 (2010).

168. Wang, D., Hashimoto, H., Zhang, X., Barwick, B.G., Lonial, S., Boise, L.H., Vertino, P.M. & Cheng, X. MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma. *Nucleic Acids Research* **45**, 2396-2407 (2017).

169. Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S. & Reik, W. Dynamic regulation of 5-

hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402 (2011).

170. Xu, Y., Wu, F., Tan, L., Kong, L., Xiong, L., Deng, J., Barbera, A.J., Zheng, L., Zhang, H., Huang, S., Min, J., Nicholson, T., Chen, T., Xu, G., Shi, Y., Zhang, K. & Shi, Yujiang G. Genome-wide Regulation of 5hmC, 5mC, and Gene Expression by Tet1 Hydroxylase in Mouse Embryonic Stem Cells. *Molecular Cell* **42**, 451-464 (2011).

171. Khare, T., Pai, S., Koncevicius, K., Pal, M., Kriukiene, E., Liutkeviciute, Z., Irimia, M., Jia, P., Ptak, C., Xia, M., Tice, R., Tochigi, M., Morera, S., Nazarians, A., Belsham, D., Wong, A.H., Blencowe, B.J., Wang, S.C., Kapranov, P., Kustra, R., Labrie, V., Klimasauskas, S. & Petronis, A. 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. *Nat Struct Mol Biol* **19**, 1037-43 (2012).

172. Wu, H., D'Alessio, A.C., Ito, S., Wang, Z., Cui, K., Zhao, K., Sun, Y.E. & Zhang, Y. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev* **25**, 679-84 (2011).

173. Globisch, D., Munzel, M., Muller, M., Michalakis, S., Wagner, M., Koch, S., Bruckl, T., Biel, M. & Carell, T. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* **5**, e15367 (2010).

174. Munzel, M., Globisch, D. & Carell, T. 5-Hydroxymethylcytosine, the sixth base of the genome. *Angew Chem Int Ed Engl* **50**, 6460-8 (2011).

175. Münzel, M., Globisch, D., Brückl, T., Wagner, M., Welzmiller, V., Michalakis, S., Müller, M., Biel, M. & Carell, T. Quantification of the Sixth DNA Base Hydroxymethylcytosine in the Brain. *Angewandte Chemie International Edition* **49**, 5375-5377 (2010).

176. Bachman, M., Uribe-Lewis, S., Yang, X., Williams, M., Murrell, A. & Balasubramanian, S. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat Chem* (2014).

177. Bachman, M., Uribe-Lewis, S., Yang, X., Burgess, H.E., Iurlaro, M., Reik, W., Murrell, A. & Balasubramanian, S. 5-Formylcytosine can be a stable DNA modification in mammals. *Nat Chem Biol* **11**, 555-7 (2015).

178. Tardy-Planechaud, S., Fujimoto, J., Lin, S.S. & Sowers, L.C. Solid phase synthesis and restriction endonuclease cleavage of oligodeoxynucleotides containing 5-(hydroxymethyl)-cytosine. *Nucleic Acids Res* **25**, 553-9 (1997).

179. Jin, S.G., Kadam, S. & Pfeifer, G.P. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res* **38**, e125 (2010).

180. Lercher, L., McDonough, M.A., El-Sagheer, A.H., Thalhammer, A., Kriaucionis, S., Brown, T. & Schofield, C.J. Structural insights into how 5-hydroxymethylation influences transcription factor binding. *Chem Commun (Camb)* **50**, 1794-6 (2014).

181. You, C., Ji, D., Dai, X. & Wang, Y. Effects of Tet-mediated Oxidation Products of 5-Methylcytosine on DNA Transcription in vitro and in Mammalian Cells. *Sci. Rep.* **4** (2014).

182. Koh, K.P., Yabuuchi, A., Rao, S., Huang, Y., Cunniff, K., Nardone, J., Laiho, A., Tahiliani, M., Sommer, C.A., Mostoslavsky, G., Lahesmaa, R., Orkin, S.H., Rodig, S.J., Daley, G.Q. & Rao, A. Tet1 and Tet2 Regulate 5-Hydroxymethylcytosine Production and Cell Lineage Specification in Mouse Embryonic Stem Cells. *Cell Stem Cell* **8**, 200-213 (2011).

183. Ko, M., Huang, Y., Jankowska, A.M., Pape, U.J., Tahiliani, M., Bandukwala, H.S., An, J., Lamperti, E.D., Koh, K.P., Ganetzky, R., Liu, X.S., Aravind, L., Agarwal, S., Maciejewski, J.P. & Rao, A. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839 (2010).

184. Sherwani, S.I. & Khan, H.A. Role of 5-hydroxymethylcytosine in neurodegeneration. *Gene* (2015).

185. Al-Mahdawi, S., Virmouni, S.A. & Pook, M.A. The emerging role of 5-hydroxymethylcytosine in neurodegenerative diseases. *Front Neurosci* **8**, 397 (2014).

186. Booth, M.J., Marsico, G., Bachman, M., Beraldi, D. & Balasubramanian, S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat Chem* **6**, 435-440 (2014).

187. Song, C.-X., Szulwach, Keith E., Dai, Q., Fu, Y., Mao, S.-Q., Lin, L., Street, C., Li, Y., Poidevin, M., Wu, H., Gao, J., Liu, P., Li, L., Xu, G.-L., Jin, P. & He, C. Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. *Cell* **153**, 678-691 (2013).

188. Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat Biotech* **32**, 1231-1240 (2014).

189. Su, M., Kirchner, A., Stazzoni, S., Müller, M., Wagner, M., Schröder, A. & Carell, T. 5-Formylcytosine Could Be a Semipermanent Base in Specific Genome Sites. *Angewandte Chemie International Edition* **55**, 11797-11800 (2016).

190. Xu, L., Chen, Y.C., Chong, J., Fin, A., McCoy, L.S., Xu, J., Zhang, C. & Wang, D. Pyrene-based quantitative detection of the 5-formylcytosine loci symmetry in the CpG duplex content during TET-dependent demethylation. *Angew Chem Int Ed Engl* **53**, 11223-7 (2014).

191. Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaöz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., James, K.D., Lefebvre, G.C., Bruce, A.W., Dovey, O.M., Ellis, P.D., Dhami, P., Langford, C.F., Weng, Z., Birney, E., Carter, N.P., Vetrie, D. & Dunham, I. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Research* **17**, 691-707 (2007).

192. Kellinger, M.W., Song, C.X., Chong, J., Lu, X.Y., He, C. & Wang, D. 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nat Struct Mol Biol.* **19** (2012).

193. Krokan, H.E., Drablos, F. & Slupphaug, G. Uracil in DNA--occurrence, consequences and repair. *Oncogene* **21**, 8935-48 (2002).

194. Krokan, H.E., Standal, R. & Slupphaug, G. DNA glycosylases in the base excision repair of DNA. *Biochemical Journal* **325**, 1 (1997).

195. Olinski, R., Jurgowiak, M. & Zaremba, T. Uracil in DNA--its biological significance. *Mutat Res* **705**, 239-45 (2010).

196. Lindahl, T. & Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**, 3405-10 (1974).

197. Conticello, S.G. The AID/APOBEC family of nucleic acid mutators. *Genome Biol* **9**, 229 (2008).

198. Cliffe, L.J., Siegel, T.N., Marshall, M., Cross, G.A.M. & Sabatini, R. Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of Trypanosoma brucei. *Nucleic Acids Research* **38**, 3923-3935 (2010).

199. Mouret, J.F., Polverelli, M., Sarrazini, F. & Cadet, J. Ionic and radical oxidations of DNA by hydrogen peroxide. *Chem Biol Interact* **77**, 187-201 (1991).

200. Bjelland, S., Eide, L., Time, R.W., Stote, R., Eftedal, I., Volden, G. & Seeberg, E. Oxidation of Thymine to 5-Formyluracil in DNA: Mechanisms of Formation, Structural Implications, and Base Excision by Human Cell Free Extracts. *Biochemistry* **34**, 14758-14764 (1995).

201. Nabel, C.S., Jia, H., Ye, Y., Shen, L., Goldschmidt, H.L., Stivers, J.T., Zhang, Y. & Kohli, R.M. AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nat Chem Biol* **8**, 751-8 (2012).

202. Guo, J.U., Su, Y., Zhong, C., Ming, G.L. & Song, H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* **145**, 423-34 (2011).

203. Pais, J.E., Dai, N., Tamanaha, E., Vaisvila, R., Fomenkov, A.I., Bitinaite, J., Sun, Z., Guan, S., Corrêa, I.R., Noren, C.J., Cheng, X., Roberts, R.J., Zheng, Y. & Saleh, L. Biochemical characterization of a Naegleria TET-like oxygenase and its application in single molecule sequencing of 5-methylcytosine. *Proceedings of the National Academy of Sciences* **112**, 4316 (2015).

204. Masaoka, A., Matsubara, M., Hasegawa, R., Tanaka, T., Kurisu, S., Terato, H., Ohyama, Y., Karino, N., Matsuda, A. & Ide, H. Mammalian 5-formyluracil-DNA glycosylase. 2. Role of SMUG1 uracil-DNA glycosylase in repair of 5-formyluracil and other oxidized and deaminated base lesions. *Biochemistry* **42**, 5003-12 (2003).

205. Bauer, N.C., Corbett, A.H. & Doetsch, P.W. The current state of eukaryotic DNA base damage and repair. *Nucleic Acids Res* **43**, 10083-101 (2015).

206. Jacobs, A.L. & Schar, P. DNA glycosylases: in DNA repair and beyond. *Chromosoma* **121**, 1-20 (2012).

207. Papaluca, A., Wagner, J.R., Saragovi, H.U. & Ramotar, D. UNG-1 and APN-1 are the major enzymes to efficiently repair 5-hydroxymethyluracil DNA lesions in C. elegans. *Sci Rep* **8**, 6860 (2018).

208. Galashevskaya, A., Sarno, A., Vågbø, C.B., Aas, P.A., Hagen, L., Slupphaug, G. & Krokan, H.E. A robust, sensitive assay for genomic uracil determination by LC/MS/MS reveals lower levels than previously reported. *DNA Repair* **12**, 699-706 (2013).

209. Kawasaki, F., Beraldi, D., Hardisty, R.E., McInroy, G.R., van Delft, P. & Balasubramanian, S. Genome-wide mapping of 5-hydroxymethyluracil in the eukaryote parasite Leishmania. *Genome Biology* **18**, 23 (2017).

210. Bullard, W., Lopes da Rosa-Spiegler, J., Liu, S., Wang, Y. & Sabatini, R. Identification of the Glucosyltransferase That Converts Hydroxymethyluracil to Base J in the Trypanosomatid Genome. *Journal of Biological Chemistry* **289**, 20273-20282 (2014).

211. Liu, S., Ji, D., Cliffe, L., Sabatini, R. & Wang, Y. Quantitative mass spectrometry-based analysis of beta-D-glucosyl-5-hydroxymethyluracil in genomic DNA of Trypanosoma brucei. *J Am Soc Mass Spectrom* **25**, 1763-70 (2014).

212. Cooke, M.S., Evans, M.D., Dizdaroglu, M. & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J* **17**, 1195-214 (2003).

213. Teebor, G.W., Frenkel, K. & Goldstein, M.S. Ionizing radiation and tritium transmutation both cause formation of 5-hydroxymethyl-2'-

deoxyuridine in cellular DNA. *Proceedings of the National Academy of Sciences of the United States of America* **81**, 318-321 (1984).

214. Rogstad, D.K., Liu, P., Burdzy, A., Lin, S.S. & Sowers, L.C. Endogenous DNA lesions can inhibit the binding of the AP-1 (c-Jun) transcription factor. *Biochemistry* **41**, 8093-102 (2002).

215. Janouskova, M., Vanikova, Z., Nici, F., Bohacova, S., Vitovska, D., Sanderova, H., Hocek, M. & Krasny, L. 5-(Hydroxymethyl)uracil and -cytosine as potential epigenetic marks enhancing or inhibiting transcription with bacterial RNA polymerase. *Chem Commun (Camb)* **53**, 13253-13255 (2017).

216. Djuric, Z., Heilbrun, L.K., Lababidi, S., Berzinkas, E., Simon, M.S. & Kosir, M.A. Levels of 5-hydroxymethyl-2'-deoxyuridine in DNA from blood of women scheduled for breast biopsy. *Cancer Epidemiol Biomarkers Prev* **10**, 147-9 (2001).

217. Rogstad, D.K., Heo, J., Vaidehi, N., Goddard, W.A., 3rd, Burdzy, A. & Sowers, L.C. 5-Formyluracil-induced perturbations of DNA function. *Biochemistry* **43**, 5688-97 (2004).

218. Privat, E.J. & Sowers, L.C. A proposed mechanism for the mutagenicity of 5-formyluracil. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **354**, 151-156 (1996).

219. Liu, P., Burdzy, A. & Sowers, L.C. Repair of the mutagenic DNA oxidation product, 5-formyluracil. *DNA Repair (Amst)* **2**, 199-210 (2003).

220. Liu, C. & Martin, C.T. Fluorescence characterization of the transcription bubble in elongation complexes of T7 RNA polymerase. *J Mol Biol* **308**, 465-75 (2001).

221. Huberman, J.A. & Riggs, A.D. On the mechanism of DNA replication in mammalian chromosomes. *J Mol Biol* **32**, 327-41 (1968).

222. Nathan, D. & Crothers, D.M. Bending and flexibility of methylated and unmethylated EcoRI DNA. *J Mol Biol* **316**, 7-17 (2002).

223. Acosta-Silva, C., Branchadell, V., Bertran, J. & Oliva, A. Mutual relationship between stacking and hydrogen bonding in DNA. Theoretical study of guanine-cytosine, guanine-5-methylcytosine, and their dimers. *J Phys Chem B* **114**, 10217-27 (2010).

224. Tretyakova, N., Guza, R. & Matter, B. Endogenous cytosine methylation and the formation of carcinogen carcinogen-DNA adducts. *Nucleic Acids Symp Ser (Oxf)*, 49-50 (2008).

225. Thalhammer, A., Hansen, A.S., El-Sagheer, A.H., Brown, T. & Schofield, C.J. Hydroxylation of methylated CpG dinucleotides

reverses stabilisation of DNA duplexes by cytosine 5-methylation. *Chem Commun (Camb)* **47**, 5325-7 (2011).

226. Mooers, B.H., Schroth, G.P., Baxter, W.W. & Ho, P.S. Alternating and non-alternating dG-dC hexanucleotides crystallize as canonical A-DNA. *J Mol Biol* **249**, 772-84 (1995).

227. Tippin, D.B., Ramakrishnan, B. & Sundaralingam, M. Methylation of the Z-DNA decamer d(GC)5 potentiates the formation of A-DNA: crystal structure of d(Gm5CGm5CGCGCGC). *J Mol Biol* **270**, 247-58 (1997).

228. Renciuk, D., Blacque, O., Vorlickova, M. & Spingler, B. Crystal structures of B-DNA dodecamer containing the epigenetic modifications 5-hydroxymethylcytosine or 5-methylcytosine. *Nucleic Acids Res* **41**, 9891-900 (2013).

229. Munzel, M., Lischke, U., Stathis, D., Pfaffeneder, T., Gnerlich, F.A., Deiml, C.A., Koch, S.C., Karaghiosoff, K. & Carell, T. Improved synthesis and mutagenicity of oligonucleotides containing 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine. *Chemistry* **17**, 13782-8 (2011).

230. Szulik, M.W., Pallan, P.S., Nocek, B., Voehler, M., Banerjee, S., Brooks, S., Joachimiak, A., Egli, M., Eichman, B.F. & Stone, M.P. Differential stabilities and sequence-dependent base pair opening dynamics of watson-crick base pairs with 5-hydroxymethylcytosine, 5-formylcytosine, or 5-carboxylcytosine. *Biochemistry* **54**, 1294-305 (2015).

231. Hardwick, J.S., Ptchelkine, D., El-Sagheer, A.H., Tear, I., Singleton, D., Phillips, S.E.V., Lane, A.N. & Brown, T. 5-Formylcytosine does not change the global structure of DNA. *Nat Struct Mol Biol* **advance online publication** (2017).

232. Tsunoda, M., Kondo, J., Karino, N., Ueno, Y., Matsuda, A. & Takenaka, A. Water mediated Dickerson-Drew-type crystal of DNA dodecamer containing 2′ -deoxy-5-formyluridine. *Biophysical Chemistry* **95**, 227-233 (2002).

233. Genest, P.-A., ter Riet, B., Cijsouw, T., van Luenen, H.G.A.M. & Borst, P. Telomeric localization of the modified DNA base J in the genome of the protozoan parasite Leishmania. *Nucleic Acids Research* **35**, 2116-2124 (2007).

234. Kawasaki, F., Martinez Cuesta, S., Beraldi, D., Mahtey, A., Hardisty, R.E., Carrington, M. & Balasubramanian, S. Sequencing 5-Hydroxymethyluracil at Single-Base Resolution. *Angew Chem Int Ed Engl* (2018).

235. Beaucage, S.L. & Iyer, R.P. Advances in the Synthesis of Oligonucleotides by the Phosphoramidite Approach. *Tetrahedron* **48**, 2223-2311 (1992).

236. Kawasaki, F., Murat, P., Li, Z., Santner, T. & Balasubramanian, S. Synthesis and biophysical analysis of modified thymine-containing DNA oligonucleotides. *Chemical Communications* **53**, 1389-1392 (2017).

237. Ansevin, A.T., Vizard, D.L., Brown, B.W. & McConathy, J. High-resolution thermal denaturation of DNA. I. Theoretical and practical considerations for the resolution of thermal subtransitions. *Biopolymers* **15**, 153-74 (1976).

238. Doty, P. The physical chemistry of deoxyribonucleic acids. *Journal of Cellular and Comparative Physiology* **49**, 27-57 (1957).

239. Marmur, J. & Doty, P. Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. *J Mol Biol* **5**, 109-18 (1962).

240. Jang, Y.H., Sowers, L.C., Çağin, T. & Goddard, W.A. First Principles Calculation of pKa Values for 5-Substituted Uracils. *The Journal of Physical Chemistry A* **105**, 274-280 (2001).

241. Kypr, J., Kejnovska, I., Renciuk, D. & Vorlickova, M. Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res* **37**, 1713-25 (2009).

242. Sutherland, J.C., Griffin, K.P., Keck, P.C. & Takacs, P.Z. Z-DNA: vacuum ultraviolet circular dichroism. *Proc Natl Acad Sci U S A* **78**, 4801-4 (1981).

243. Delort, A.M., Neumann, J.M., Molko, D., Herve, M., Teoule, R. & Tran Dinh, S. Influence of uracil defect on DNA structure: 1H NMR investigation at 500 MHz. *Nucleic Acids Res* **13**, 3343-55 (1985).

244. Workman, J.L. & Kingston, R.E. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu Rev Biochem* **67**, 545-79 (1998).

245. Sekinger, E.A., Moqtaderi, Z. & Struhl, K. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol Cell* **18**, 735-48 (2005).

246. Kornberg, R.D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285-94 (1999).

247. Johnson, S., Tan, F., McCullough, H., Riordan, D. & Fire, A. Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res* **16**, 1505 - 1516 (2006).

248. Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I. & Pugh, B.F. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* **18**, 1073-83 (2008).

249. Mavrich, T., Jiang, C., Ioshikhes, I., Li, X., Venters, B., Zanton, S., Tomsho, L., Qi, J., Glaser, R., Schuster, S., Gilmour, D., Albert, I. & Pugh, B. Nucleosome organization in the Drosophila genome. *Nature* **453**, 358 - 362 (2008).

250. Schones, D.E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G. & Zhao, K. Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* **132**, 887-898 (2008).

251. Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J., Costa, G., McKernan, K., Sidow, A., Fire, A. & Johnson, S. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res* **18**, 1051 - 1063 (2008).

252. Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J. & Rando, O.J. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309**, 626-30 (2005).

253. Axel, R. Cleavage of DNA in nuclei and chromatin with staphylococcal nuclease. *Biochemistry* **14**, 2921-2925 (1975).

254. Clark, R.J. & Felsenfeld, G. Structure of chromatin. *Nat New Biol* **229**, 101-6 (1971).

255. Nedospasov, S.A. & Georgiev, G.P. Non-random cleavage of SV40 DNA in the compact minichromosome and free in solution by micrococcal nuclease. *Biochem Biophys Res Commun* **92**, 532-9 (1980).

256. Jimeno-González, S., Ceballos-Chávez, M. & Reyes, J.C. A positioned +1 nucleosome enhances promoter-proximal pausing. *Nucleic Acids Research* **43**, 3068-3078 (2015).

257. Radman-Livaja, M. & Rando, O.J. Nucleosome positioning: How is it established, and why does it matter? *Developmental Biology* **339**, 258-266 (2010).

258. Lowary, P.T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol* **276**, 19-42 (1998).

259. Thåström, A., Lowary, P.T., Widlund, H.R., Cao, H., Kubista, M. & Widom, J. Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences1. *Journal of Molecular Biology* **288**, 213-229 (1999).

260. Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I., Wang, J. & Widom, J. A genomic code for nucleosome positioning. *Nature* **442**, 772 - 778 (2006).

261. Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J. & Segal, E. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362-366 (2009).

262. Chua, E.Y.D., Vasudevan, D., Davey, G.E., Wu, B. & Davey, C.A. The mechanics behind DNA sequence-dependent properties of the nucleosome. *Nucleic Acids Research* (2012).

263. Trifonov, E.N. Sequence-dependent deformational anisotropy of chromatin DNA. *Nucleic Acids Res* **8**, 4041-53 (1980).

264. Satchwell, S.C., Drew, H.R. & Travers, A.A. Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology* **191**, 659-675 (1986).

265. Chua, E.Y.D., Vogirala, V.K., Inian, O., Wong, A.S.W., Nordenskiöld, L., Plitzko, J.M., Danev, R. & Sandin, S. 3.9 Å structure of the nucleosome core particle determined by phase-plate cryo-EM. *Nucleic Acids Research* **44**, 8013-8019 (2016).

266. Dickerson, R.E., Goodsell, D.S. & Neidle, S. "...the tyranny of the lattice...". *Proc Natl Acad Sci U S A* **91**, 3579-83 (1994).

267. Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F. & Richmond, T.J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-60 (1997).

268. Johnson, R.C., Stella, S. & Heiss, J.K. in Protein-Nucleic Acid Interactions: Structural Biology 176-220 (The Royal Society of Chemistry, 2008).

269. Davey, C.A. & Richmond, T.J. DNA-dependent divalent cation binding in the nucleosome core particle. *Proceedings of the National Academy of Sciences* **99**, 11169 (2002).

270. Richmond, T.J. & Davey, C.A. The structure of DNA in the nucleosome core. *Nature* **423**, 145-50 (2003).

271. Cao, H., Widlund, H.R., Simonsson, T. & Kubista, M. TGGA repeats impair nucleosome formation. *J Mol Biol* **281**, 253-60 (1998).

272. Cacchione, S., Cerone, M.A. & Savino, M. In vitro low propensity to form nucleosomes of four telomeric sequences. *FEBS Letters* **400**, 37-41 (1997).

273. Kunkel, G.R. & Martinson, H.G. Nucleosomes will not form on double-stranded RNa or over poly(dA).poly(dT) tracts in recombinant DNA. *Nucleic Acids Res* **9**, 6869-88 (1981).

274. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat Struct Mol Biol* **20**, 267-73 (2013).

275. McCall, M., Brown, T. & Kennard, O. The crystal structure of d(G-G-G-G-C-C-C-C) a model for poly(dG) · poly(dC). *Journal of Molecular Biology* **183**, 385-396 (1985).

276. Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z. & Sidow, A. Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516-20 (2011).

277. Segal, E. & Widom, J. Poly(dA:dT) Tracts: Major Determinants of Nucleosome Organization. *Current opinion in structural biology* **19**, 65-71 (2009).

278. Liu, H., Wu, J., Xie, J., Yang, X., Lu, Z. & Sun, X. Characteristics of nucleosome core DNA and their applications in predicting nucleosome positions. *Biophys J* **94**, 4597-604 (2008).

279. Gupta, S., Dennis, J., Thurman, R.E., Kingston, R., Stamatoyannopoulos, J.A. & Noble, W.S. Predicting human nucleosome occupancy from primary sequence. *PLoS Comput Biol* **4**, e1000134 (2008).

280. Eslami-Mossallam, B., Schiessel, H. & van Noort, J. Nucleosome dynamics: Sequence matters. *Advances in Colloid and Interface Science* **232**, 101-113 (2016).

281. Liu, H., Zhang, R., Xiong, W., Guan, J., Zhuang, Z. & Zhou, S. A comparative evaluation on prediction methods of nucleosome positioning. *Briefings in Bioinformatics* **15**, 1014-1027 (2014).

282. Padinhateeri, R. & Marko, J.F. Nucleosome positioning in a model of active chromatin remodeling enzymes. *Proc Natl Acad Sci U S A* **108**, 7799-803 (2011).

283. Hartley, P.D. & Madhani, H.D. Mechanisms that Specify Promoter Nucleosome Location and Identity. *Cell* **137**, 445-458 (2009).

284. Hughes, A.L., Jin, Y., Rando, O.J. & Struhl, K. A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern. *Mol Cell* **48**, 5-15 (2012).

285. Längst, G., B.Teif, V. & Rippe, K. in Genome Organization and Function in the Cell Nucleus (2011).

286. Yadav, T. & Whitehouse, I. Replication-Coupled Nucleosome Assembly and Positioning by ATP-Dependent Chromatin-Remodeling Enzymes. *Cell Reports* **15**, 715-723 (2016).

287. Yamada, K., Frouws, T.D., Angst, B., Fitzgerald, D.J., DeLuca, C., Schimmele, K., Sargent, D.F. & Richmond, T.J. Structure and mechanism of the chromatin remodelling factor ISW1a. *Nature* **472**, 448 (2011).

288. Whitehouse, I., Rando, O.J., Delrow, J. & Tsukiyama, T. Chromatin remodelling at promoters suppresses antisense transcription. *Nature* **450**, 1031-5 (2007).

289. Gkikopoulos, T., Schofield, P., Singh, V., Pinskaya, M., Mellor, J., Smolle, M., Workman, J.L., Barton, G.J. & Owen-Hughes, T. A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. *Science* **333**, 1758-60 (2011).

290. Koslover, E.F., Fuller, C.J., Straight, A.F. & Spakowitz, A.J. Local geometry and elasticity in compact chromatin structure. *Biophys J* **99**, 3941-50 (2010).

291. Li, G. & Widom, J. Nucleosomes facilitate their own invasion. *Nat Struct Mol Biol* **11**, 763-9 (2004).

292. Mao, C., Brown, C.R., Griesenbeck, J. & Boeger, H. Occlusion of regulatory sequences by promoter nucleosomes in vivo. *PLoS One* **6**, e17521 (2011).

293. Simpson, R.T. Nucleosome positioning can affect the function of a cis-acting DMA elementin vivo. *Nature* **343**, 387 (1990).

294. Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J., Casero, D., Bernal, M., Huijser, P., Clark, A.T., Kramer, U., Merchant, S.S., Zhang, X., Jacobsen, S.E. & Pellegrini, M. Relationship between nucleosome positioning and DNA methylation. *Nature* **466**, 388-92 (2010).

295. Kelly, T.K., Liu, Y., Lay, F.D., Liang, G., Berman, B.P. & Jones, P.A. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* **22**, 2497-506 (2012).

296. Teif, V.B., Beshnova, D.A., Vainshtein, Y., Marth, C., Mallm, J.P., Hofer, T. & Rippe, K. Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Res* **24**, 1285-95 (2014).

297. Tan, L., Xiong, L., Xu, W., Wu, F., Huang, N., Xu, Y., Kong, L., Zheng, L., Schwartz, L., Shi, Y. & Shi, Y.G. Genome-wide comparison of DNA hydroxymethylation in mouse embryonic stem cells and neural

progenitor cells by a new comparative hMeDIP-seq method. *Nucleic Acids Research* (2013).

298. Raiber, E.A., Portella, G., Martinez Cuesta, S., Hardisty, R., Murat, P., Li, Z., Iurlaro, M., Dean, W., Spindel, J., Beraldi, D., Liu, Z., Dawson, M.A., Reik, W. & Balasubramanian, S. 5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells. *Nat Chem* (2018).

299. Luger, K. & Richmond, T.J. DNA binding within the nucleosome core. *Current Opinion in Structural Biology* **8**, 33-40 (1998).

300. Drew, H.R. & Travers, A.A. DNA bending and its relation to nucleosome positioning. *Journal of Molecular Biology* **186**, 773-790 (1985).

301. Perez, A., Castellazzi, C.L., Battistini, F., Collinet, K., Flores, O., Deniz, O., Ruiz, M.L., Torrents, D., Eritja, R., Soler-Lopez, M. & Orozco, M. Impact of methylation on the physical properties of DNA. *Biophys J* **102**, 2140-8 (2012).

302. Thåström, A., Lowary, P.T. & Widom, J. Measurement of histone–DNA interaction free energy in nucleosomes. *Methods* **33**, 33-44 (2004).

303. Choy, J.S., Wei, S., Lee, J.Y., Tan, S., Chu, S. & Lee, T.-H. DNA Methylation Increases Nucleosome Compaction and Rigidity. *Journal of the American Chemical Society* **132**, 1782-1783 (2010).

304. Lee, J.Y. & Lee, T.H. Effects of DNA methylation on the structure of nucleosomes. *J Am Chem Soc* **134**, 173-5 (2012).

305. Jimenez-Useche, I., Ke, J., Tian, Y., Shim, D., Howell, S.C., Qiu, X. & Yuan, C. DNA Methylation Regulated Nucleosome Dynamics. *Sci. Rep.* **3** (2013).

306. Portella, G., Battistini, F. & Orozco, M. Understanding the Connection between Epigenetic DNA Methylation and Nucleosome Positioning from Computer Simulations. *PLoS Computational Biology* **9**, e1003354 (2013).

307. Davey, C., Pennings, S. & Allan, J. CpG methylation remodels chromatin structure in vitro. *J Mol Biol* **267**, 276-88 (1997).

308. Collings, C.K., Waddell, P.J. & Anderson, J.N. Effects of DNA methylation on nucleosome stability. *Nucleic Acids Research* **41**, 2918-2931 (2013).

309. Mendonca, A., Chang, E.H., Liu, W. & Yuan, C. Hydroxymethylation of DNA influences nucleosomal conformation and stability in vitro. *Biochim Biophys Acta* **1839**, 1323-9 (2014).

310. Yang, T.P., Hansen, S.K., Oishi, K.K., Ryder, O.A. & Hamkalo, B.A. Characterization of a cloned repetitive DNA sequence concentrated on the human X chromosome. *Proc Natl Acad Sci U S A* **79**, 6593-7 (1982).

311. Minary, P. & Levitt, M. Training-free atomistic prediction of nucleosome occupancy. *Proceedings of the National Academy of Sciences* **111**, 6293 (2014).

312. Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I. & Andrau, J.C. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res* **22**, 2399-408 (2012).

313. Das, C., Tyler, J.K. & Churchill, M.E.A. The histone shuffle: histone chaperones in an energetic dance. *Trends in Biochemical Sciences* **35**, 476-489 (2010).

314. Luger, K., Rechsteiner, T.J., Flaus, A.J., Waye, M.M. & Richmond, T.J. Characterization of nucleosome core particles containing histone proteins made in bacteria. *J Mol Biol* **272**, 301-11 (1997).

315. Nelson, T., Wiegand, R. & Brutlag, D. Ribonucleic acid and other polyanions facilitate chromatin assembly in vitro. *Biochemistry* **20**, 2594-2601 (1981).

316. Stein, A., Whitlock, J.P., Jr. & Bina, M. Acidic polypeptides can assemble both histones and chromatin in vitro at physiological ionic strength. *Proc Natl Acad Sci U S A* **76**, 5000-4 (1979).

317. Stein, A. Reconstitution of chromatin from purified components. *Methods Enzymol* **170**, 585-603 (1989).

318. Haushalter, K.A. & Kadonaga, J.T. Chromatin assembly by DNA-translocating motors. *Nat Rev Mol Cell Biol* **4**, 613-620 (2003).

319. Glikin, G.C., Ruberti, I. & Worcel, A. Chromatin assembly in Xenopus oocytes: In vitro studies. *Cell* **37**, 33-41 (1984).

320. Ito, T., Bulger, M., Pazin, M.J., Kobayashi, R. & Kadonaga, J.T. ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. *Cell* **90**, 145-55 (1997).

321. Andrews, A.J., Chen, X., Zevin, A., Stargell, L.A. & Luger, K. The Histone Chaperone Nap1 Promotes Nucleosome Assembly by Eliminating Nonnucleosomal Histone DNA Interactions. *Molecular Cell* **37**, 834-842 (2010).

322. Fyodorov, D.V. & Kadonaga, J.T. Chromatin assembly in vitro with purified recombinant ACF and NAP-1. *Methods Enzymol* **371**, 499-515 (2003).

323. Fujii-Nakata, T., Ishimi, Y., Okuda, A. & Kikuchi, A. Functional analysis of nucleosome assembly protein, NAP-1. The negatively charged COOH-terminal region is not necessary for the intrinsic assembly activity. *J Biol Chem* **267**, 20980-6 (1992).

324. Ito, T., Levenstein, M.E., Fyodorov, D.V., Kutach, A.K., Kobayashi, R. & Kadonaga, J.T. ACF consists of two subunits, Acf1 and ISWI, that function cooperatively in the ATP-dependent catalysis of chromatin assembly. *Genes Dev* **13**, 1529-39 (1999).

325. Hardisty, R.E., Kawasaki, F., Sahakyan, A.B. & Balasubramanian, S. Selective Chemical Labeling of Natural T Modifications in DNA. *Journal of the American Chemical Society* (2015).

326. Shi, L., Tong, W., Su, Z., Han, T., Han, J., Puri, R.K., Fang, H., Frueh, F.W., Goodsaid, F.M., Guo, L., Branham, W.S., Chen, J.J., Xu, Z.A., Harris, S.C., Hong, H., Xie, Q., Perkins, R.G. & Fuscoe, J.C. Microarray scanner calibration curves: characteristics and implications. *BMC Bioinformatics* **6 Suppl 2**, S11 (2005).

327. Booth, M.J., Branco, M.R., Ficz, G., Oxley, D., Krueger, F., Reik, W. & Balasubramanian, S. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934-7 (2012).

328. Widom, J. Equilibrium and dynamic nucleosome stability. *Methods Mol Biol* **119**, 61-77 (1999).

329. Booth, M.J., Raiber, E.-A. & Balasubramanian, S. Chemical Methods for Decoding Cytosine Modifications in DNA. *Chemical Reviews* **115**, 2240-2254 (2015).

330. The PyMOL Molecular Graphics System, Version 2.1.1 Schrödinger, LLC.

331. Kennedy, S.R., Schmitt, M.W., Fox, E.J., Kohrn, B.F., Salk, J.J., Ahn, E.H., Prindle, M.J., Kuong, K.J., Shen, J.C., Risques, R.A. & Loeb, L.A. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**, 2586-606 (2014).

332. Rhoads, A. & Au, K.F. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* **13**, 278-289 (2015).

333. Jain, M., Olsen, H.E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* **17**, 239 (2016).

334. Li, F., Zhang, Y., Bai, J., Greenberg, M.M., Xi, Z. & Zhou, C. 5-Formylcytosine Yields DNA–Protein Cross-Links in Nucleosome Core Particles. *Journal of the American Chemical Society* **139**, 10617-10620 (2017).

335. Ji, S., Shao, H., Han, Q., Seiler, C.L. & Tretyakova, N.Y. Reversible DNA‐Protein Cross‐Linking at Epigenetic DNA Marks. *Angewandte Chemie International Edition* **56**, 14130-14134 (2017).

336. Ji, S., Fu, I., Naldiga, S., Shao, H., Basu, A.K., Broyde, S. & Tretyakova, N.Y. 5-Formylcytosine mediated DNA–protein cross-links block DNA replication and induce mutations in human cells. *Nucleic Acids Research*, gky444-gky444 (2018).

337. Zhao, Y. & Garcia, B.A. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harb Perspect Biol* **7**, a025064 (2015).

338. Funato, K. & Tabar, V. Histone Mutations in Cancer. *Annual Review of Cancer Biology* **2**, 337-351 (2018).

339. Lewis, P.W., Muller, M.M., Koletsky, M.S., Cordero, F., Lin, S., Banaszynski, L.A., Garcia, B.A., Muir, T.W., Becher, O.J. & Allis, C.D. Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma. *Science* **340**, 857-61 (2013).

340. Herz, H.M., Morgan, M., Gao, X., Jackson, J., Rickels, R., Swanson, S.K., Florens, L., Washburn, M.P., Eissenberg, J.C. & Shilatifard, A. Histone H3 lysine-to-methionine mutants as a paradigm to study chromatin signaling. *Science* **345**, 1065-70 (2014).

341. Kim, J.H., Lee, J.H., Lee, I.S., Lee, S.B. & Cho, K.S. Histone Lysine Methylation and Neurodevelopmental Disorders. *Int J Mol Sci* **18** (2017).

342. Parkel, S., Lopez-Atalaya, J.P. & Barco, A. Histone H3 lysine methylation in cognition and intellectual disability disorders. *Learn Mem* **20**, 570-9 (2013).

343. Faundes, V., Newman, W.G., Bernardini, L., Canham, N., Clayton-Smith, J., Dallapiccola, B., Davies, S.J., Demos, M.K., Goldman, A., Gill, H., Horton, R., Kerr, B., Kumar, D., Lehman, A., McKee, S., Morton, J., Parker, M.J., Rankin, J., Robertson, L., Temple, I.K. & Banka, S. Histone Lysine Methylases and Demethylases in the Landscape of Human Developmental Disorders. *The American Journal of Human Genetics* **102**, 175-187 (2018).

344. https://www.atdbio.com/content/17/Solid-phase-oligonucleotide-synthesis accessed on 2018/Sep/09.

345. Karino, N., Ueno, Y. & Matsuda, A. Synthesis and properties of oligonucleotides containing 5-formyl-2′-deoxycytidine: in vitro DNA polymerase reactions on DNA templates containing 5-formyl-2′-deoxycytidine. *Nucleic Acids Research* **29**, 2456-2463 (2001).

346. Schröder, A.S., Steinbacher, J., Steigenberger, B., Gnerlich, F.A., Schiesser, S., Pfaffeneder, T. & Carell, T. Synthesis of a DNA Promoter Segment Containing All Four Epigenetic Nucleosides: 5-Methyl-, 5-Hydroxymethyl-, 5-Formyl-, and 5-Carboxy-2′-Deoxycytidine. *Angewandte Chemie International Edition* **53**, 315-318 (2013).

347. Münzel, M., Lischke, U., Stathis, D., Pfaffeneder, T., Gnerlich, F.A., Deiml, C.A., Koch, S.C., Karaghiosoff, K. & Carell, T. Improved Synthesis and Mutagenicity of Oligonucleotides Containing 5-Hydroxymethylcytosine, 5-Formylcytosine and 5-Carboxylcytosine. *Chemistry – A European Journal* **17**, 13782-13788 (2011).

348. Xuan, S., Wu, Q., Cui, L., Zhang, D. & Shao, F. 5-Hydroxymethylcytosine and 5-formylcytosine containing deoxyoligonucleotides: Facile syntheses and melting temperature studies. *Bioorganic & Medicinal Chemistry Letters* **25**, 1186-1191 (2015).

349. Dai, Q. & He, C. Syntheses of 5-Formyl- and 5-Carboxyl-dC Containing DNA Oligos as Potential Oxidation Products of 5-Hydroxymethylcytosine in DNA. *Organic Letters* **13**, 3446-3449 (2011).

350. Nangreave, J., Han, D., Liu, Y. & Yan, H. DNA origami: a history and current perspective. *Current Opinion in Chemical Biology* **14**, 608-615 (2010).

351. Nur, I., Szyf, M., Razin, A., Glaser, G., Rottem, S. & Razin, S. Procaryotic and eucaryotic traits of DNA methylation in spiroplasmas (mycoplasmas). *J Bacteriol* **164**, 19-24 (1985).

352. Hardenbol, P., Banér, J., Jain, M., Nilsson, M., Namsaraev, E.A., Karlin-Neumann, G.A., Fakhrai-Rad, H., Ronaghi, M., Willis, T.D., Landegren, U. & Davis, R.W. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotechnology* **21**, 673 (2003).

353. Krishnakumar, S., Zheng, J., Wilhelmy, J., Faham, M., Mindrinos, M. & Davis, R. A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 9296-9301 (2008).

354. Tabor, S., Struhl, K., Scharf, S.J. & Gelfand, D.H. DNA-dependent DNA polymerases. *Curr Protoc Mol Biol* **Chapter 3**, Unit3.5 (2001).

355. Riedl, J., Fleming, A.M. & Burrows, C.J. Sequencing of DNA Lesions Facilitated by Site-Specific Excision via Base Excision Repair DNA Glycosylases Yielding Ligatable Gaps. *Journal of the American Chemical Society* **138**, 491-494 (2016).

356. Steigenberger, B., Schiesser, S., Hackner, B., Brandmayr, C., Laube, S.K., Steinbacher, J., Pfaffeneder, T. & Carell, T. Synthesis of 5-Hydroxymethyl-, 5-Formyl-, and 5-Carboxycytidine-triphosphates and Their Incorporation into Oligonucleotides by Polymerase Chain Reaction. *Organic Letters* **15**, 366-369 (2013).

357. Takahashi, M., Yamaguchi, E. & Uchida, T. Thermophilic DNA ligase. Purification and properties of the enzyme from Thermus thermophilus HB8. *J Biol Chem* **259**, 10041-7 (1984).

358. Barany, F. Genetic disease detection and DNA amplification using cloned thermostable ligase. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 189-193 (1991).

359. Barany, F. The ligase chain reaction in a PCR world. *PCR Methods Appl* **1**, 5-16 (1991).

360. Ducruix, A. & Giegé, R. Crystallization of nucleic acids and proteins: a practical approach (IRL Press at Oxford University Press, 1992).

361. McGregor, H.C. & Gunderman, R.B. X-ray crystallography and the elucidation of the structure of DNA. *AJR Am J Roentgenol* **196**, W689-92 (2011).

362. Iwafuchi-Doi, M., Donahue, G., Kakumanu, A., Watts, J.A., Mahony, S., Pugh, B.F., Lee, D., Kaestner, K.H. & Zaret, K.S. The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Mol Cell* **62**, 79-91 (2016).

363. Dann, G.P., Liszczak, G.P., Bagert, J.D., Muller, M.M., Nguyen, U.T.T., Wojcik, F., Brown, Z.Z., Bos, J., Panchenko, T., Pihl, R., Pollock, S.B., Diehl, K.L., Allis, C.D. & Muir, T.W. ISWI chromatin remodellers sense nucleosome modifications to determine substrate preference. *Nature* **548**, 607-611 (2017).

364. Lv, J., Liu, H., Wang, Q., Tang, Z., Hou, L. & Zhang, B. Molecular cloning of a novel human gene encoding histone acetyltransferase-like protein involved in transcriptional activation of hTERT. *Biochem Biophys Res Commun* **311**, 506-13 (2003).

365. Chi, Y.H., Haller, K., Peloponese, J.M., Jr. & Jeang, K.T. Histone acetyltransferase hALP and nuclear membrane protein hsSUN1 function in de-condensation of mitotic chromosomes. *J Biol Chem* **282**, 27447-58 (2007).

366. Verdin, E., Dequiedt, F. & Kasler, H.G. Class II histone deacetylases: versatile regulators. *Trends Genet* **19**, 286-93 (2003).

367. Boltzmann sigmoid function of Prism http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?reg_classic_boltzmann.htm (accessed 2015/June/30).

368. Feng, H.-p. Early cryo-EM work. *Nature Structural Biology* **7**, 22 (2000).

369. Poirier, M.G., Oh, E., Tims, H.S. & Widom, J. Dynamics and function of compact nucleosome arrays. *Nature structural & molecular biology* **16**, 938-944 (2009).

370. Bilokapic, S., Strauss, M. & Halic, M. Cryo-EM of nucleosome core particle interactions in trans. *Sci Rep* **8**, 7046 (2018).

371. Kalashnikova, A.A., Porter-Goff, M.E., Muthurajan, U.M., Luger, K. & Hansen, J.C. The role of the nucleosome acidic patch in modulating higher order chromatin structure. *Journal of the Royal Society Interface* **10**, 20121022 (2013).