The Involvement of 5-Hydroxymethylcytosine in

Two DNA Metabolism Pathways:

Transcription and double stranded DNA Breaks

Julia Robertson

Thesis for the degree of Philosophiae Doctor (PhD)

Institute of Medical Microbiology, Department of Molecular Biology

Rikshospitalet, Oslo University Hospital

University of Oslo, Norway

2014

Acknowledgements

The work presented in this dissertation was carried out at the Department of Medical Microbiology at Oslo University Hospital, Rikshospitalet, between July 2009 and May 2014. Financial support was provided by Helse Sør-Øst RHF.

I would like to express my deepest appreciation to my thesis advisor Professor Arne Klungland. He has given me: the opportunity to work on this very exciting topic; excellent guidance and fruitful discussions throughout the years; helpful and positive feedback; and the freedom to pursue and perform the experiments I felt necessary. I would also like to thank my co-supervisor Dr. Adam Robertson for his limitless help, support, scientific input, constructive criticism, guidance, and encouragement along the way. Your enthusiasm and genuine interest in science are a great inspiration to me. I would like to thank the members of the Molecular Microbiology Department and the Klungland group for their helpful input and ideas.

I would like to express my eternal appreciation towards my parents and family; no matter where I have been working, they have always supported me unconditionally. I thank my husband for his never-ending motivation and help; I couldn't have completed my PhD without you.

Julia Robertson

Oslo, 2014

Table of contents

List of papers

This thesis is based on the following papers, which will be referred to by their Roman numerals:

I.   **Robertson J.***, Robertson A.B.*, Klungland A. "The presence of 5-hydroxymethylcytosine at the gene promoter and not in the gene body negatively regulates gene expression" (* contributed equally)
Biochem Biophys Res Commun. 2011 Jul 22;411(1):40-3.

II. Robertson A.B.*, **Robertson J.***, Fusser M., Klungland A. "Endonuclease G Preferentially Cleaves 5-Hydroxymethylcytosine-Modified DNA creating a substrate for recombination" (* contributed equally)
Nucleic Acids Res., in press (2014)

Abbreviations

5hmC	5-hydroxymethylcytosine
5meC	5-methylcytosine
6meA	N6-methyladenine
A	Adenosine
AID	Activation induced cytosine deaminase
AP	Abasic site
BER	Base excision repair
C	Cytosine
CD	Catalytic domain
Dam	DNA adenine methyltransferase
DNA	Deoxyribonucleic acid
dsDNA	Double stranded deoxyribonucleic acid
E. coli	*Escherichia coli*
EndoG	Endonuclease G
G	Guanidine
iPS cell	Induced pluripotent stem cell
JBP1	J-binding protein 1
MMR	DNA mismatch repair
mRNA	Messenger ribonucleic acid
OGT	O-Linked N-Acetylglucosamine Transferase
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
T	Thymine
TDG	Thymidine DNA glycosylase
Tet	Ten-eleven translocation methylcytosine dioxygenase
VSP	Very short patch repair

Summary

5-hydroxymethylcytosine (5hmC) is involved in various cellular processes, including transcriptional regulation, demethylation of 5-methylcytosine (5meC) and stem cell pluripotency. 5hmC is formed by oxidation of 5-methylcytosine by the Tet family of enzymes. This thesis describes two functions attributed 5hmC – these are the role of 5hmC in transcription and the role of 5hmC in hepatocyte polyploidy.

Overwhelming evidence supports the notion that 5meC has a negative effect on transcription; however, only recently has the effect that 5hmC has on transcription begun to be studied. Using model substrates including the CMV_{IE} promoter and a generic gene body we have directly assessed the effect that 5hmC, both at the promoter and in the gene body, has on *in vitro* gene transcription. We show that the presence of 5hmC modifications strongly represses transcription. We also demonstrate that the inhibition of transcriptional activity is primarily due to the presence of 5hmC in the promoter and that 5hmC in the gene body has a minimal effect on transcription. Thus, we propose that the presence of 5hmC in promoter prevents the binding of essential transcription factors or recruits factors that repress transcription.

In the second part of my thesis, we describe the identification of an activity that preferentially cleaves 5hmC-modified DNA; a mammalian nuclease that can differentiate between cytosine, 5meC, and 5hmC. Using biochemical methods this activity was isolated from mouse liver extracts and the enzyme responsible for the cleavage of 5hmC-modified DNA was identified to be Endonuclease G (EndoG). Recombinant EndoG preferentially recognizes and cleaves a core sequence when one specific cytosine within that core sequence is hydroxymethylated. EndoG catalyzes the formation of double stranded DNA breaks in a 5hmC-dependent manner *in vitro*

and *in vivo*. Finally we show that EndoG and 5hmC can initiate recombination *in vitro*.

Introduction

According to biology's central dogma DNA is transcribed to RNA; and RNA is translated to proteins (Fig. 1). While this central dogma generally remains true, in the past half century, researchers have found that these processes are highly regulated [1-4]. This regulation at the DNA to RNA level is accomplished through the coordinated effort of DNA modifications – 5-methylcytosine and 5-hydxymethycytosine; proteins that bind DNA and effect transcription; and modifications to proteins that allow for accessibility of the DNA for active or repressed transcription [5]. The translation of RNA to protein is regulated by RNA splicing, RNA binding and RNA modifying proteins, and ribosomal binding proteins, among other entities [3, 6, 7].
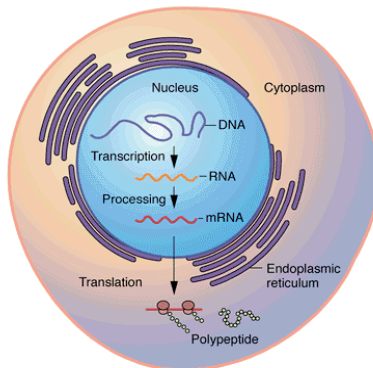
Fig. 1. Central dogma of molecular biology. DNA is transcribed to RNA; and RNA is translated to proteins.

Deoxyribonucleic Acid

Deoxyribonucleic acid (DNA) is the heritable molecule in most organisms, with the exception of several viruses that utilize ribonucleic acid (RNA) as their genetic material [8]. The DNA backbone is a polymer of alternating phosphates and

deoxyribose sugars. Attached to each of these deoxyribose sugars is a base, which is either adenosine (A), cytosine (C), guanidine (G) or thymine (T). Each strand of the DNA polymer is paired with another strand of the DNA polymer that is complementary to the first strand. The complementary strand pairs using Chargaff's Rules – where A pairs with T and C pairs with G [9, 10].

The entirety of the DNA in a normal cell is referred to as that organism's genome [11]. In general, genes are smaller units of DNA that code for proteins, the entities that carry out the bulk of the work in a cell. Transmission of alleles, different variations of a particular gene, from parental organism to daughter organism allows for the large variations of a single species seen in nature [12, 13]. In mammalian cells, most alleles are split into exons, which contain the coding sequence for proteins, and introns, which are non-coding regions between the exons. Upstream of the first intron (Fig. 2), or internal to the first intron, of most genes is a promoter – a DNA element that allows for the induction of transcription, or RNA synthesis. Even further upstream from the first exon, or in some cases downstream, are enhancer elements, which, as their name implies, act to enhance the action of the promoter [5, 14-16].
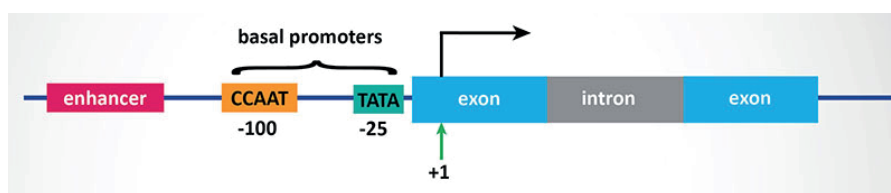
Fig. 2. Schematic structure of a eukaryotic protein-coding gene. The promoter region of the gene consists of a distal enhancer region, a CCAAT-box and a TATA-box. Most eukaryotic mRNA genes contain a basic structure consisting of coding exons and non-coding introns.

RNA Synthesis

In most organisms genes are transcribed to RNA, which is then translated to proteins, commonly referred to as gene expression. In mammalian cells there are three RNA polymerases – RNA polymerases I-III. RNA polymerase I is responsible for the synthesis of most ribosomal RNA (rRNA) [17, 18]. As rRNAs are one of the most abundant species of RNA, RNA polymerase I synthesizes more than half of the cellular RNA pool. RNA polymerase II is the enzyme chiefly responsible for the synthesis of messenger RNA (mRNA); mature mRNA is utilized by the ribosome to synthesize proteins [18, 19]. RNA polymerase II also synthesizes some small nucleolar RNAs and microRNAs [20]. RNA Polymerase III is chiefly responsible for the synthesis of transfer RNAs (tRNA); however, this polymerase also transcribes the 5S rRNA [21, 22]. Gene expression is a highly regulated process involving histones, transcription factors, and direct modification of the DNA molecule itself. These regulatory elements will be discussed in more detail below.

Processing of messenger RNA

Prior to being translated into proteins by the ribosomes, mammalian pre-mRNAs are, more often than not, spliced to remove sequences that are not part of the final coding sequence. These intervening sequences are referred to as introns whereas the coding regions are called exons. Exons are spliced together in the spliceosome protein complex [23, 24]. mRNA splicing allows for a wide variety of different final transcripts from the same DNA coding sequence as different exons can be included or excluded according to the cell's needs (Fig. 3). Spliced mRNAs have a GTP cap added to the 5' end of the molecule while the 3' end of most mRNAs is modified by the addition of multiple adenosines to the RNA molecule. These modifications, called

the 5' cap and the polyA tail, improve the stability of the mRNA molecule [25]. The processed mRNAs are exported from the nucleus to the cytoplasm to be translated into proteins. As more recent studies have shown, mRNA molecules are further modified by several modifications; among these are not only adenosine deamination but also methylation at the N6 position of riboadenosine [26-29]. While it is understood that adenosine deamination participates in RNA editing, the significance of methylated RNA remains unclear [30].
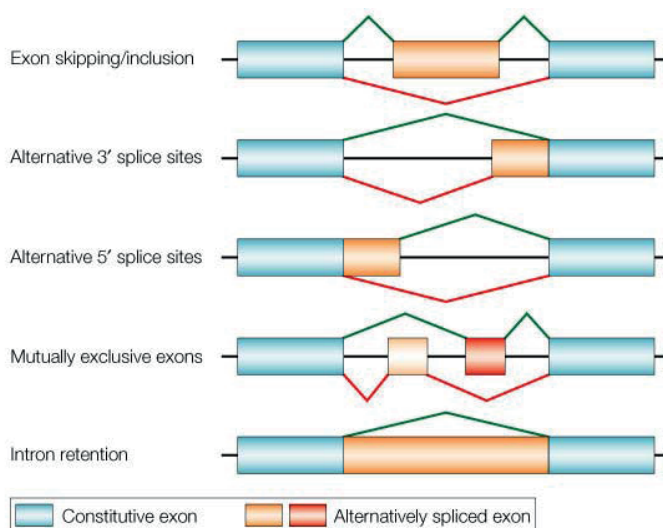
Fig. 3. The figure illustrates different types of alternative splicing: exon inclusion or skipping, alternative splice-site selection, mutually exclusive exons, and intron retention. For an individual pre-mRNA, different alternative exons often show different types of alternative-splicing patterns [31].

Translation

  mRNAs are translated from the nucleic acid three letter code into proteins via the ribosome [32, 33]. As soon as the protein begins to be synthesized the cellular machinery recognizes its initial signal sequence and the protein is sorted to its appropriate subcellular localization. In most eukaryotic cells there are two distinct

pools of ribosomes – the cytosolic ribosomes and the endoplasmic reticulum (ER) localized ribosomes. Proteins translated on cytosolic ribosomes have a distinctly different fate than mRNAs translated on ER ribosomes. Proteins synthesized on cytosolic ribosomes, in most cases, end up in the cytosol, nucleus, chloroplast or mitochondria. In contrast, proteins synthesized on ER ribosomes enter the secretory pathway and will end up in the ER, the golgi apparatus, the lysosome or will be secreted [34, 35]. Proteins carry out much of the work in a cell and are responsible for nearly all processes undertaken by an organism. Therefore, the development and maintenance of all the cells within an organism requires the differential expression of numerous proteins. This differential expression is controlled by the conversion of DNA to heterochromatin or euchromatin [36].

Heterochromatin and Euchromatin

Gene expression is principally controlled by two factors – (i) the presence of the appropriate RNA polymerase and appropriate transcription factors and (ii) the ability of these factors to access the DNA. DNA that is tightly wrapped around histones – heterochromatin – is transcriptionally inactive; DNA that is wrapped less tightly around the structural histone proteins – euchromatin – results in a state that is more transcriptionally active [37, 38]. This conversion from heterochromatin to euchromatin is controlled by two main factors – the number and types of histone modifications and DNA modifications [39, 40].

Histones

In most cells five types of histones compose the histone pool – these include Histone H1, Histone H2A, Histone H2B, Histone H3, and Histone H4 (Fig. 4). The core histones are generally hetero-octameric proteins that interact with DNA. Each histone octamer is comprised of two copies each of the histone proteins H2A, H2B, H3 and H4. 147 base pairs of DNA are wrapped around each histone octamer. Histone H1 links these octamers together resulting in significant compaction of the DNA [41]. In eukaryotic cells, this organization allows for the entire genome to fit within the cellular confines; in humans this is approximately 2 meters of DNA per cell [42]. There are several histone variants and alternative functions have been ascribed to the use of these variants. Among these variants are Histone H3.1, H3.2, H3.3, H2AX, and H2AZ [43]. Having a substantial amount of basic amino acid residues at their amino and carboxy termini, histone octamers are positively charged. This charge allows the histone octamers to interact strongly through ionic interactions with the negatively charged DNA backbone. As several studies have shown, the stronger the interaction between the histone octamer and the DNA the less accessible the DNA is to transcriptional, replicative, and repair machinery [44].
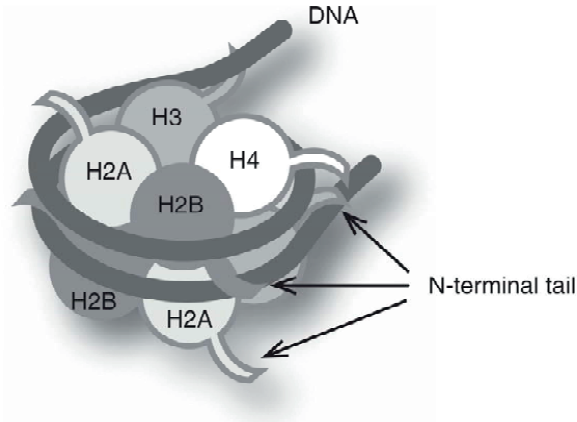
Fig. 4. Schematic nucleosome structure. A nucleosome consists of two copies of each core histone (H2A, H2B, H3 and H4) and 150bp DNA. The N-terminal tail of each histone is extruded from the nucleosome [41].

Histone modification

Typically histones are modified by a wide array of chemical groups at their carboxy and amino terminal tails, these include all of the following: acetylation; methylation of arginine and lysine residues; phosphorylation at serine and threonine residues; glycosylation, ubiquitination, sumoylation, and ADP-ribosylation (Fig. 5) [44, 45]. The wide variety of Histone octamers can be modified in well over 100 combinations of modifications. Histone modifications act to make the DNA either more accessible or less accessible to proteins – including transcription factors, RNA polymerases, and DNA repair proteins. By regulating access to DNA, histones can regulate transcription, replication, DNA repair, and almost all DNA transactions.

Fig. 5. (a) Known post-translational modifications and the amino acid residues they modify. **(b)** Residues that can undergo several different forms of post-translational modification. ac, acetylation; bio, biotinylation; cit, citrullination; me, methylation; su, SUMOylation [46].

DNA Modifications

In a variety of organisms the bases in DNA are modified. In most cases, DNA modifications are deleterious to the organism; however, certain modifications are enzymatically added to the DNA and are essential for proper development and cellular maintenance. Deleterious modifications are the result of DNA damage. Mutagenic DNA damage falls into two general classes: modifications that alter base pairing and modifications that block replication; modifications referred to as DNA damage often alter a hydrogen bond donor or acceptor site within the modified base – examples of these include deamination of cytosine resulting in uracil, an abasic site, and deamination of adenosine resulting in hypoxantine (Fig. 6) [47-49]. However, the

modification may also be mutagenic if it sterically hinders the bases ability to fit normally within the double helix – examples of these include 8-oxoguanine, 7-methylguanine, and O^4-methylthymine. Several types of modifications prevent DNA replication by blocking the relevant DNA polymerase from synthesizing DNA using the damaged base as a template [50, 51].
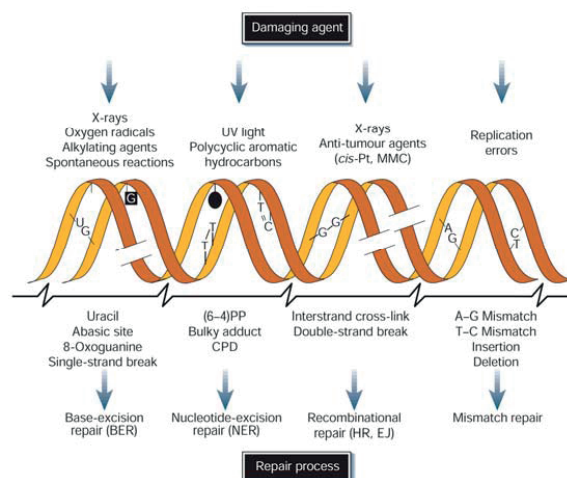
Fig. 6. DNA damage, repair mechanisms and consequences. Common DNA damaging agents (top), examples of DNA lesions induced by these agents (middle) and the DNA repair mechanism responsible for the repair of the lesions (bottom) [52].

In instances where the organism intentionally adds the modification to the DNA, these regulatory modifications are generally advantageous to the organism's survival. Below are descriptions of regulative DNA modifications identified in several organisms.

DNA Modifications in bacteriophages

Bacteriophages are viruses that infect bacteria. As the bacteriophage is an obligate intracellular parasite, it requires a bacterial host to replicate its DNA; the

bacteriophage injects its DNA into the bacteria, subsequently taking over the bacterial machinery to copy phage DNA [53, 54]. To prevent bacteriophage infection and the entry of foreign DNA, the bacterial host uses a restriction modification system [55, 56]. Briefly, the bacterium synthesizes enzymes that specifically cut foreign DNA. These enzymes either recognize a DNA sequence that is not present in the bacterial genome or the cleavage of these enzymes is blocked by modifications present in the bacterial genome. In other cases, these restriction enzymes only destroy modified DNA. To evade the bacterial host's restriction/modification defense mechanisms the bacteriophage camouflages its DNA with elaborate DNA modifications. Thus both the bacteria and the phage have co-evolved sophisticated DNA modification systems that allow for either the prevention of bacteriophage infection in the host or enhance the bacteriophage's ability to evade the host's restriction modification system.

T-even bacteriophages (T2, T4, and T6) incorporate 3'-deoxyribo-5-hydroxymethylcytosine into their DNA instead of 3'-deoxyribocytosine. Importantly, the 5-hydroxymethylcytosine modification is present prior to incorporation into the genome by the relevant DNA polymerase [57, 58]. This is a fundamentally different process than occurs in mammalian genomes, discussed later, which results in nearly all cytosines in the T-even phages being substituted with 5hmC. T2, T4, and T6 further modify their DNA by adding at least one glucose molecule to a substantial fraction of these 5hmC-modified bases. The T2, T4, and T6 bacteriophages all possess the β-glucosyltransferase; an enzyme that catalyzes the formation of β-glucosyl-5hmC. In addition the T4 bacteriophage possesses the β-glucosyltransferase, catalyzing the formation of β-glucosyl-5hmC [59]. The glucosylation of 5hmC is likely an evolutionary defense mechanism that protects the invading bacteriophage DNA from the activity of the nucleases coded for by the *E. coli rglA* and *rglB* genes.

Initially designated as restricts glucose-less (*rgl*), these gene products specifically restrict 5hmC-modified DNA; however these enzymes are ineffective at cleaving glucosylated 5hmC modified DNA. As an additional defense mechanism the T6 bacteriophage possesses a disaccharide modification on a significant fraction of its 5hmC modified DNA: this modification is gentiobiosyl [60]. It is unclear at the present time which enzyme is responsible for the synthesis of gentiobiosyl.

Most of these elaborate T-even bacteriophage DNA modifications are the result of synergistic evolution of both the bacterial host and the bacteriophage acting to prevent infection and increase infectivity, in each organism respectively.

DNA Modifications in *E. coli*

There are two principle stable DNA modifications present in *E. coli*: N6-methyladenine (6meA) and N5-methylcytosine (5meC) [61]. 6meA is deposited at almost all d(GATC) sequences following DNA replication. This methylation is catalyzed by the DNA adenine methyltransferase coded for by the *Dam* gene [62]. Dam methylation is required to correct post-replication errors through the DNA mismatch repair (MMR) pathway [63, 64]. While the daughter strand is unmethylated, the parental strand is labeled by methylation after replication; this allows the MMR machinery to distinguish the correct, parental strand from the mutated, daughter strand. After this distinction is made the MMR machinery acts to repair the unmethylated strand. Consistent with its function in MMR, a deletion of the *dam* gene results in 2-3 orders of magnitude reduction in replication fidelity.

While the function of 6meA is better described, the function of 5meC in *E. coli* genomes is less well understood. 5meC in *E. coli* is deposited by the transfer of a methyl group from S-adenosyl-methylmethionine to the second cytosine in the

sequence 5'-CCWGG-3' by the DNA cytosine methyltransferase (*Dcm*) [65]. This methylation poses an interesting problem for the bacterium. While deamination of cytosine results in a uracil:guanosine (U:G) mispair, the deamination of 5meC results in thymine:guanosine (T:G) mispair. As the U:G mispair is recognized by the uracil DNA glycosylase (UDG), uracil is efficiently excised from the DNA and the base excision repair pathway (BER) processes the resultant abasic site [66]. Unlike mammals, *E. coli* does not encode a thymidine DNA glycosylase (TDG) required to repair a T:G mispair. Interestingly, the bacterium appears to have co-evolved a distinct repair pathway that recognizes T:G mispairs in the sequence 5'-CTWGG-3': the very short patch repair pathway (VSP). When a T:G mispair is present at the second cytosine within the sequence 5'-CCWGG-3', the mispair is recognized by the Vsr endonuclease; this cleaves 5' to the mispaired thymine. The resultant nick and mispair is processed by DNA polymerase I and DNA ligase I [67]. Interestingly, both the Dcm and the Vsr genes are encoded at the same genetic loci. Typically, a mutation in either the Vsr gene or in the Dcm gene results in the inactivation of both genes. This genetic and functional arrangement indicates that the two genes evolved at nearly the same time. Indeed, without Dcm methylation the substrate for Vsr endonuclease is never present. Alternatively, in the absence of the Vsr endonuclease, Dcm methylation is mutagenic.

The purpose of methylation within *E. coli* is a matter of debate. It is known that Dam methylation is required for efficient MMR pathway [68]. While Dcm methylation is required for VSP, it seems unlikely that the bacterium evolved a mutagenic methylation system, followed by the evolution of a repair mechanism that would repair errors created by this system. In an evolutionary sense it appears easier to lose the Dcm gene to rectify the problem. Many have postulated that methylation

by both Dam and Dcm function in all of the following processes: DNA repair; prevent the integration of foreign DNA; and suppress transposon replication. These mechanisms are beyond the scope of this thesis; however for a review of the subject please see Palmer et al [69].

DNA Modifications in kinetoplastids

Important in biology because many of these organisms cause severe human diseases, Kinetoplastids and Trypanosomes possess a unique DNA modification – the J-base [70]. The J-base, or β-glucosyl-5-hydroxymethyluracil, is synthesized in these organisms in two steps: JBP1 catalyzed hydroxylation of thymidine resulting in 5-hydroxymethyluracil, followed by the glucosylation of 5-hydroxymethyluracil by an enzyme that has not yet been identified. The function of the J-base is unclear; however, it is found most prominently at telomere sequences [71]. Interestingly, the JBP1 protein binds with a relatively high affinity to the J-base in DNA. As the J-base is almost chemically identical to β -glucosyl-5hmC, JBP1 binds with a high affinity to this base as well.

DNA Modifications in Eukaryotic Cells

The most common DNA modification found in mammalian DNA is 5-methylcytosine (5meC). 5meC is composed of a deoxyribocytosine modified at the N5 position by a methyl group [72]. Methyl groups are generally added to the cytosine moiety within the CpG dinucleotide sequence; this methylation occurs most frequently at promoter regions [73]. 5meC modified promoters are frequently transcriptionally inactive and 5meC-modified DNA is preferentially bound by methyl DNA binding

domain proteins (MBD) [74]. It is widely accepted that the interaction between the MBD proteins and 5meC-modified DNA recruits histone modifying enzymes; these enzymes modify histones such that the region is transcriptionally inactive [75, 76].

In 2009, a second DNA modification was identified, which is likely key in gene regulation; this newly discovered modification was 5-hydroxymethylcytosine [77, 78] (5hmC, Fig. 7). 5hmC is created by the enzymatic oxidation of 5meC catalyzed by the Tet family of proteins. These proteins include Tet1, Tet2, and Tet3. Interestingly, all three of the Tet proteins (Tet1-3) are evolutionarily related to the kinetoplastid thymidine hydroxylase – JBP1.
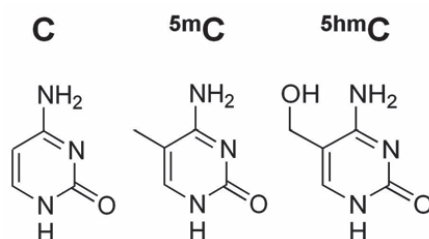
Fig. 7. Cytosine, 5-methylcytosine and 5-hydroxymethylcytosine modifications.

Since its discovery, a plethora of studies have demonstrated that 5hmC is involved in a large number of cellular processes – including (i) transcriptional regulation, (ii) active DNA demethylation, (iii) stem cell self renewal, (iv) proper development, and (v) cancer formation [79].


Transcriptional Regulation

As early reports have shown, the 5hmC modification might be an intermediate in the conversion of 5meC to cytosine; thus an enzymatic activity was identified that can potentially demethylate DNA [78]. At the same time, 5hmC was shown to be a stable DNA modification. It was found on one hand to be present in specialized

nondividing neurons and on the other hand absent in specific cancer cell lines [77]. While MBD proteins interact well with 5meC; these proteins interact poorly with 5hmC-modified DNA [80]. Therefore, it is speculated that the inability of the MBD proteins to interact with 5hmC-modified DNA prevents histone modifying enzyme recruitment; this purportedly relieves the 5meC-implemented transcriptional repression. However, the presence of 5hmC at a gene promoter has been shown not only to repress transcription but also to enhance transcription. The temporal and spatial transcriptional regulation in diverse cell types is likely to be partially accounted for by differential 5hmC and 5meC patterning [81-84]. Hydroxylation of 5-methylcytosine effectively modulates gene expression using a novel transcriptional regulatory mechanism [85].

Active Demethylation

Perhaps the most interesting observation is that 5hmC appears to be an intermediate in the active demethylation of 5meC. The formation of 5hmC from 5meC is catalyzed by the enzymes – Tet1, Tet2, and Tet3 (Tet1-3) [78, 86]. As several studies have proposed, at least three mechanisms explain the involvement of Tet1-3 in 5meC demethylation – these mechanisms are summarized in **Fig. 8**. In the first mechanism, 5hmC is a substrate for the activation induced cytosine deaminase (AID). The deamination of 5hmC by AID results in the formation of 5-hydroxymethyluracil; this is processed further by elements of the base excision repair (BER) pathway yielding cytosine. In the second mechanism, Tet1-3 oxidizes 5hmC further forming both 5-formylcytosine and 5-carboxycytosine [87]. These bases are acted on by the thymidine DNA glycosylase (TDG); this results in the formation of an abasic site. The BER machinery processes this abasic site liberating cytosine. In the third mechanism, a 5-carboxycytosine is enzymatically decarboxylated; the enzyme or enzymes responsible

for this decarboxylation reaction remain unknown [88]. All these reactions will require further investigation; however, in addition to being a stable DNA modification, it appears that 5hmC is an intermediate in 5meC active demethylation. As promoter demethylation is positively correlated with active transcription, the logical consequence of 5meC demethylation is an increase in transcriptional activation.
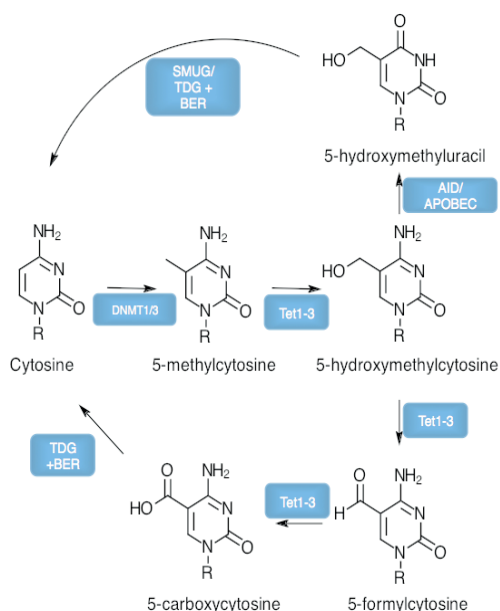
Fig. 8. Cytosine base modifications and possible demethylation pathways. Cytosine is converted to 5-methylcytosine by both DNMT1 and DNMT3. Either Tet1, Tet2, or Tet3 (Tet1-3) oxidize 5-methylcytosine to 5-hydroxymethylcytosine (5hmC). 5hmC can be further oxidized by Tet1-3; this yields 5-formylcytosine and 5-carboxycytosine. Both 5-formylcytosine and 5-carboxycytosine are depyrimidated by the thymidine DNA glycosylase (TDG) and processed by the base excision repair (BER) pathway. Alternatively, 5-hydroxymethylcytosine is enzymatically deaminated by AID or APOBEC resulting in 5-hydroxymethyluracil; this base is depyrimidated by SMUG or TDG and processed by the BER pathway. All these pathways result in unmodified cytosine.

Stem Cells and Development

The modulation of the transcriptional landscape by Tet1-3 profoundly affects cell fate. Tet1 downregulation and concomitant reduction in 5hmC levels are characteristic of stem cell differentiation [78, 86, 89]. While remaining pluripotent, stem cells deficient for Tet1 and Tet2 appear to execute defective differentiation programs; this phenotype was not completely penetrant [90]. In combination with Oct4, Sox2, Klf4, and c-Myc, the addition of exogenous Tet1 or Tet2 to embryonic fibroblasts significantly enhances induced pluripotent stem (iPS) cell colony formation [91, 92]. During fertilization, maternal Tet3 catalyzes male pronuclei demethylation [93]. In line with this finding, Tet3 deficiency is strongly correlated with reduced female fertility. As these findings indicate, both Tet3 and 5hmC are likely to be involved in early development. As these studies also suggest, the appropriate regulation of both Tet1 and Tet2 is necessary for proper stem cell differentiation; and the precise execution of developmental programs requires the concerted action of Tet1, Tet2, and Tet3.

5hmC and Cancer

The inappropriate regulation of both tumor suppressor genes and oncogenes is a well-known aberrant cellular process; such disruptions often result in cancer formation. This inappropriate regulation is often caused either by oncogene promoter hypomethylation or by tumor suppressor gene hypermethylation [94-101]. The effects that 5hmC has on protein expression is the subject of many past and current studies. As recent reports suggest, 5hmC is present in all healthy mouse tissues at a low abundance. In many tumor cell lines 5hmC is either completely absent or aberrantly patterned [77]. Aberrantly patterned 5hmC are strongly linked to not only myeloid cancers and

melanomas but also metastatic events [102-105]. A reduction in 5hmC levels is directly proportional to both the growth and the invasiveness of melanomas [106]. Indicating that appropriate Tet1-3 levels may inhibit tumor formation or progression, the reintroduction of Tet2 into 5hmC-depleted melanomas results in a significant decrease in both tumor volume and growth [106]. As these results highlight, gene regulation by the 5hmC DNA modification is important in tumorigenesis.

Unknown Tet1-3 Regulatory Mechanism

As these studies [102, 103] strongly suggest, Tet1-3 catalyzed oxidation of 5meC to 5hmC is a non-random process. The important processes tied to 5hmC placement suggest that the Tet enzyme family is regulated by upstream elements; most likely these upstream elements are different in unrelated cell types. Most studies that look at the Tet enzyme family focus on the outward effect of inappropriate 5hmC patterning and levels, the downstream effects of Tet1-3 enzyme activity are well characterized. No screen has been designed to identify Tet1-3 upstream regulatory elements. Furthermore, only a few studies have been conducted to identify proteins that directly interact with Tet1-3. One research group used a screen to identify proteins that physically interact with Tet2 and Tet3 [107]. As this study showed, Tet2 and Tet3 interact with O-Linked N-Acetylglucosamine Transferase (OGT). However, the interaction between OGT and either Tet2 or Tet3 affects neither cytosine hydroxymethylation nor Tet2 or Tet3 enzyme function. In a screen designed to find proteins that directly interact with the pluripotency gene – Nanog, Tet1 was identified. As this report showed, the interaction between Tet1 and Nanog had no appreciable effect on either 5hmC levels or 5hmC patterning [92].

Aims of study

5hmC in the mammalian genome has significant implications in development, metabolism, and disease. Therefore, the principle aim of my thesis work was to elucidate the mechanisms by which 5hmC influences these cellular processes. The two secondary aims of my thesis were (i) to determine the role of 5hmC in generalized transcription and (ii) describe a novel function of 5hmC, unrelated to transcription, in genomic DNA.

Summary of results

Paper I

In our first manuscript, we show that 5hmC strongly inhibits generalized transcription if the modification is present at promoter regions. On the contrary, even when all cytosines in the gene body are replaced with 5hmC transcription is relatively unaffected. Further in the manuscript we speculate about the mechanism(s) by which transcription may be regulated by 5hmC.

Paper II

In the second study we identified an activity present in mouse liver nuclear extracts that preferentially catalyzes the cleavage of 5hmC-modified DNA. We isolated this activity; we determined that the activity is catalyzed by Endonuclease G. We characterize the *in vitro* activity of Endonuclease G. Additionally we show that Endonuclease G catalyzes the cleavage of 5hmC-modified DNA *in vivo*, creating a substrate for recombination.

Discussion

As relatively recent reports [108-110] and unpublished data from our laboratory show, followed by promoter and enhancer regions, 5hmC is present most prominently within gene bodies. Given the strong link between 5meC and transcriptional regulation, the involvement of 5hmC in transcription has been given a priority by most laboratories including ours. 5hmC is distributed throughout the genome principally within gene bodies – outside of the promoter and enhancer regions. Interestingly, 5hmC has limited effects on transcription when present within the gene body.

The effect that 5meC has on transcriptional regulation has been extensively studied; however, the effect that 5hmC has on transcription has only been indirectly addressed [89, 111-113]. We aimed to directly evaluate the effect that the presence of 5hmC has on *in vitro* transcription using the CMV_{IE} promoter and a generic gene body in a well defined mammalian system.

We show in our first paper that the presence of 5hmC at the CMV_{IE} promoter strongly inhibits transcription in human nuclear extracts. This finding is supported by two other studies that indirectly show that the presence of 5hmC at gene promoters reduces gene expression [111, 112]. We also demonstrate that the presence of 5hmC in the gene body has negligible effects on transcription in this *in vitro* system, contrasting with a report that suggests that 5hmC in the gene body increases transcription [89]. This study used DNA modified with 5hmC in the presence of nucleosomes that may differentially affect transcription in the presence of 5hmC within the gene body. Our study was intended to measure the singular effect that 5hmC has on transcription in the promoter and/or in the gene body of naked DNA in the absence of nucleosomes. Demonstrating that 5hmC in the gene body has a limited

effect on transcription is important because this finding rules out that 5hmC provides an elongation block to RNA polymerase II. Therefore, we can conclude that elongation catalyzed by RNA polymerase II is not inhibited by the presence of 5hmC and thus 5hmC prevents transcriptional initiation.

We propose that the inhibition of transcription initiation from a gene that has 5hmC at its promoter may occur by three different mechanisms. First, the 5hmC at the promoter may directly prevent the binding of necessary transcription factors or RNA polymerase II from binding to the DNA preventing the formation of the pre-initiation complex (Fig. 9A). Secondly, 5hmC may recruit an unknown factor to the promoter that inhibits the binding of basal transcription factors or RNA polymerase II (Fig. 9B). Currently, we are designing assays to assess the mechanism by which this transcriptional inhibition is carried out. Finally, 5hmC at the promoter may allow for the binding of the basal transcription factors and RNA polymerase II; however, 5hmC may provide an inhibitory signal to the pre-initiation complex preventing the initiation of transcription despite the presence of the all the necessary transcription factors and RNA polymerase II (Fig. 9C).
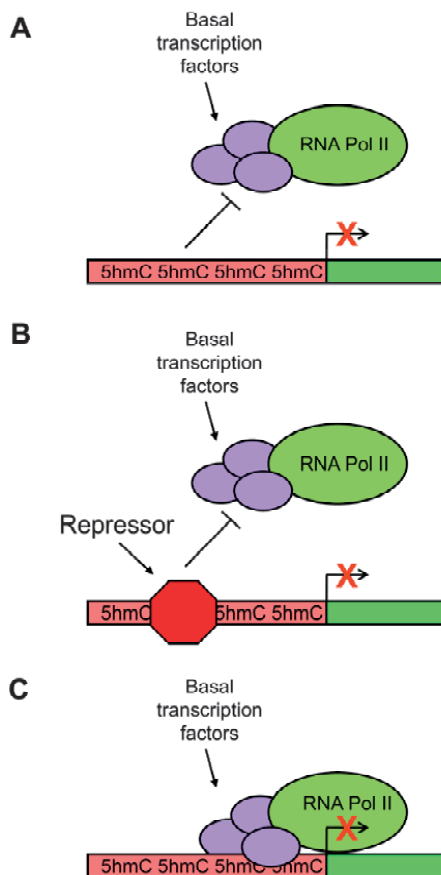
Fig. 9. Models representing the potential mechanisms of transcriptional inhibition from promoters that contain 5hmC modifications. (A) 5hmC at the promoter prevents the binding of basal transcription factors or RNA polymerase II effectively suppressing transcription. (B) 5hmC recruits another cellular factor(s) that prevents some or all of the transcription machinery from binding to the promoter suppressing transcription. (C) 5hmC at a gene promoter may allow for the basal transcription machinery to bind to the promoter; however, the 5hmC induces a conformational change in the basal transcription factors or recruits another protein that prevents the release of RNA polymerase II from the promoter inhibiting transcription initiation.

As mentioned in the opening paragraph, 5hmC resides at the greatest levels within gene bodies and not at promoter regions. These reports in combination with our result in Paper I, suggesting that 5hmC within the gene body has a limited affect on

transcription, indicates that 5hmC within gene bodies and other genetic regions is likely to have a function outside of transcription. This led us to propose that transcription is not the only function for 5hmC. Using this as a working hypothesis we endeavored to identify other proteins that interact with 5hmC modified DNA. In Paper II, we did not identify a protein that preferentially interacts with 5hmC-modified DNA; however, we found an activity that preferentially cleaves 5hmC modified DNA. That activity was found to be catalyzed by Endonuclease G.

Several studies have implicated cytosine hypermethylation as a driving force for the initiation of endogenous double stranded breaks [114, 115]. Interestingly, the techniques used in these reports to identify cytosine methylation – bisulfite sequencing – cannot distinguish between cytosine methylation and hydroxymethylation [116]. Therefore, it is plausible that the endogenous double stranded breaks correlating with hypermethylation, observed in previous studies [117], are to some extent induced by hyper-hydroxymethylation.

When the Tet2 CD is overexpressed we observed a significant increase in the number of γ-H2AX positive cells and the quantity of γ-H2AX foci per cell, signifying an increase in the number of dsDNA breaks. Importantly, the quantity of γ-H2AX foci returns to near control levels when EndoG is knocked down. The HeLa cell line overexpressing Tet2 CD shows a reduction in growth rate, which was overcome by the additional knockdown of EndoG. This slowed growth rate may be a response to the significant increase in dsDNA breaks. We suggest that these breaks must be repaired as apoptosis is not increased and cell viability is not reduced.

Our results suggest that the 5'-GGGG^{5hm}CCAG-3' sequence is preferentially cleaved by EndoG. Although we show that the 5'-GGGG^{5hm}CCAG-3' sequence can be cleaved by EndoG, we cannot rule out the possibility that EndoG can catalyze the

cleavage of other 5hmC-modified sequences. This hydroxymethylated cytosine is not in a CpG context; however, we note that 5hmC at this position is within the CHH (H = A, C, or T) sequence context, a known target of cytosine methylation [118, 119]. Depending upon the method used to quantify 5hmC content, the level of 5hmC at CHH regions is between 0.04% and 3% [120-122]. The differences in 5hmC at CHH sites in these reports may be due to random variation among cell lines or other technical issues. In any case, the cytosine in this sequence context is likely to be hydroxymethylated at some developmental stage and therefore the modified cytosine is in a biologically relevant sequence context.

We [84, 85] and others [112, 123] have envisaged a role for 5hmC in transcriptional regulation. Several groups have demonstrated that 5hmC has a role in stem cell pluripotency [89] and in the oxidative demethylation of 5-methylcytosine [87, 124, 125]. We do not believe that these studies conflict with the presence of an endonuclease that preferentially cleaves 5hmC-modified DNA. Indeed, cytosine methylation has been demonstrated to have several cellular functions, likewise cytosine hydroxymethylation potentially has multiple functions.

Previous reports [117, 126, 127] demonstrate that EndoG creates double stranded breaks at CG rich regions. However, these studies were performed before the identification of 5hmC in the genome and cytosine hydroxymethylation has not been implicated as a requirement for this cleavage. While we observe EndoG-mediated cleavage at CG rich regions, EndoG cleaves 5hmC-modified DNA much more efficiently. EndoG resides primarily in the mitochondrial inner membrane space [128] and in the nucleus at a lower concentration [129]. Importantly, the liver nuclear extracts used for this study were free of any detectable mitochondrial contaminants. EndoG is thought to induce apoptosis in a caspase independent manner [126, 130,

131]. It has also been shown that EndoG can generate primers necessary for mitochondrial DNA replication [132]. Nevertheless, an Endog-deficient (Endog-/-) mouse did not show apoptotic defects and did not show any deficiencies consistent with ineffective or inefficient mitochondrial DNA replication [133, 134]. The absence of these defects in the Endog-/- mice suggests an alternative role for EndoG in nucleic acid metabolism. Indeed, several studies have shown that EndoG is necessary for the initiation of recombination by the creation of double-stranded DNA breaks [117, 135]. The data in our study shows that EndoG prefers to cleave 5hmC-modified DNA, can promote the formation of strand breaks *in vivo* and that this break can be used to initiate recombination. Taken together, our data suggests that EndoG via its 5hmC-mediated cleavage activity can initiate recombination at hydroxymethylated cytosine residues.

Future perspectives

Our finding that 5hmC within the gene body has a limited effect on transcription will have a profound effect on future studies. Indeed, most studies have focused on the relation of 5hmC to transcription. While we believe that this relation is likely correct, it opens the door to studies that question the role of 5hmC in nucleic acid processes outside transcriptional regulation. We expect that these future studies will help solve important biological problems.

In our second study we identified a protein and a mechanism that results in conservative recombination. We expect that our finding is only the beginning of research aimed at identifying roles of 5hmC in other nucleic acid transactions.

References

1. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-3.
2. McCarthy, B.J. and J.J. Holland, *Denatured DNA as a direct template for in vitro protein synthesis.* Proc Natl Acad Sci U S A, 1965. **54**(3): p. 880-6.
3. Sonenberg, N. and A.G. Hinnebusch, *Regulation of translation initiation in eukaryotes: mechanisms and biological targets.* Cell, 2009. **136**(4): p. 731-45.
4. Todeschini, A.L., A. Georges, and R.A. Veitia, *Transcription factors: specific DNA binding and specific gene regulation.* Trends Genet, 2014.
5. Berger, S.L., *The complex language of chromatin regulation during transcription.* Nature, 2007. **447**(7143): p. 407-12.
6. Wilkie, G.S., K.S. Dickson, and N.K. Gray, *Regulation of mRNA translation by 5'- and 3'-UTR-binding factors.* Trends Biochem Sci, 2003. **28**(4): p. 182-8.
7. Arnold, R.J. and J.P. Reilly, *Observation of Escherichia coli ribosomal proteins and their posttranslational modifications by mass spectrometry.* Anal Biochem, 1999. **269**(1): p. 105-12.
8. Gelderblom, H.R., *Structure and Classification of Viruses*, in *Medical Microbiology*, S. Baron, Editor. 1996: Galveston (TX).
9. Chargaff, E., R. Lipshitz, and C. Green, *Composition of the desoxypentose nucleic acids of four genera of sea-urchin.* J Biol Chem, 1952. **195**(1): p. 155-60.
10. Elson, D. and E. Chargaff, *On the desoxyribonucleic acid content of sea urchin gametes.* Experientia, 1952. **8**(4): p. 143-5.
11. Brown, T.A., in *Genomes*. 2002: Oxford.
12. Malats, N. and F. Calafell, *Basic glossary on genetic epidemiology.* J Epidemiol Community Health, 2003. **57**(7): p. 480-2.
13. Yip, S.P., *Sequence variation at the human ABO locus.* Ann Hum Genet, 2002. **66**(Pt 1): p. 1-27.
14. Holzfeind, P. and B. Redl, *Structural organization of the gene encoding the human lipocalin tear prealbumin and synthesis of the recombinant protein in Escherichia coli.* Gene, 1994. **139**(2): p. 177-83.
15. Sabatini, L.M., Y.Z. He, and E.A. Azen, *Structure and sequence determination of the gene encoding human salivary statherin.* Gene, 1990. **89**(2): p. 245-51.
16. Valenzuela, P., et al., *Structure of yeast phenylalanine-tRNA genes: an intervening DNA segment within the region coding for the tRNA.* Proc Natl Acad Sci U S A, 1978. **75**(1): p. 190-4.
17. Russell, J. and J.C. Zomerdijk, *The RNA polymerase I transcription machinery.* Biochem Soc Symp, 2006(73): p. 203-16.
18. Engel, C., et al., *RNA polymerase I structure and transcription regulation.* Nature, 2013. **502**(7473): p. 650-5.
19. Acker, J., et al., *Interactions between the human RNA polymerase II subunits.* J Biol Chem, 1997. **272**(27): p. 16815-21.
20. Lee, Y., et al., *MicroRNA genes are transcribed by RNA polymerase II.* EMBO J, 2004. **23**(20): p. 4051-60.
21. Ishiguro, A., G.A. Kassavetis, and E.P. Geiduschek, *Essential roles of Bdp1, a subunit of RNA polymerase III initiation factor TFIIIB, in transcription and tRNA processing.* Mol Cell Biol, 2002. **22**(10): p. 3264-75.

22. Pascali, C. and M. Teichmann, *RNA polymerase III transcription - regulated by chromatin structure and regulator of nuclear chromatin organization.* Subcell Biochem, 2013. **61**: p. 261-87.

23. Cvitkovic, I. and M.S. Jurica, *Spliceosome database: a tool for tracking components of the spliceosome.* Nucleic Acids Res, 2013. **41**(Database issue): p. D132-41.

24. Will, C.L. and R. Luhrmann, *Spliceosome structure and function.* Cold Spring Harb Perspect Biol, 2011. **3**(7).

25. Bellofatto, V. and J. Wilusz, *Transcription and mRNA stability: parental guidance suggested.* Cell, 2011. **147**(7): p. 1438-9.

26. Desrosiers, R., K. Friderici, and F. Rottman, *Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells.* Proc Natl Acad Sci U S A, 1974. **71**(10): p. 3971-5.

27. Fu, Y., et al., *Gene expression regulation mediated through reversible m(6)A RNA methylation.* Nat Rev Genet. **15**(5): p. 293-306.

28. Meyer, K.D., et al., *Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons.* Cell, 2012. **149**(7): p. 1635-46.

29. Wei, C.M., A. Gershowitz, and B. Moss, *5'-Terminal and internal methylated nucleotide sequences in HeLa cell mRNA.* Biochemistry, 1976. **15**(2): p. 397-401.

30. Fu, Y., et al., *Gene expression regulation mediated through reversible m(6)A RNA methylation.* Nat Rev Genet, 2014. **15**(5): p. 293-306.

31. Edwalds-Gilbert., G., *Regulation of mRNA Splicing by Signal Transduction.* Nature Education, 2010. **3(9)**(43).

32. Jenner, L., et al., *Crystal structure of the 80S yeast ribosome.* Curr Opin Struct Biol, 2012. **22**(6): p. 759-67.

33. Rabl, J., et al., *Crystal structure of the eukaryotic 40S ribosomal subunit in complex with initiation factor 1.* Science, 2011. **331**(6018): p. 730-6.

34. Kanner, E.M., M. Friedlander, and S.M. Simon, *Co-translational targeting and translocation of the amino terminus of opsin across the endoplasmic membrane requires GTP but not ATP.* J Biol Chem, 2003. **278**(10): p. 7920-6.

35. Hortsch, M., D. Avossa, and D.I. Meyer, *Characterization of secretory protein translocation: ribosome-membrane interaction in endoplasmic reticulum.* J Cell Biol, 1986. **103**(1): p. 241-53.

36. Rapaport, F., et al., *Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.* Genome Biol, 2013. **14**(9): p. R95.

37. Remenyi, A., H.R. Scholer, and M. Wilmanns, *Combinatorial control of gene expression.* Nat Struct Mol Biol, 2004. **11**(9): p. 812-5.

38. Sproul, D., N. Gilbert, and W.A. Bickmore, *The role of chromatin structure in regulating the expression of clustered genes.* Nat Rev Genet, 2005. **6**(10): p. 775-81.

39. Katan-Khaykovich, Y. and K. Struhl, *Heterochromatin formation involves changes in histone modifications over multiple cell generations.* Embo J, 2005. **24**(12): p. 2138-49.

40. Verzijlbergen, K.F., et al., *Multiple histone modifications in euchromatin promote heterochromatin formation by redundant mechanisms in Saccharomyces cerevisiae.* BMC Mol Biol, 2009. **10**: p. 76.

41. Kimura, H., *Histone modifications for human epigenome analysis.* J Hum Genet, 2013. **58**(7): p. 439-45.

42. Bednar, J., et al., *Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin.* Proc Natl Acad Sci U S A, 1998. **95**(24): p. 14173-8.

43. Boulard, M., et al., *Histone variant nucleosomes: structure, function and implication in disease.* Subcell Biochem, 2007. **41**: p. 71-89.

44. Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications.* Nature, 2000. **403**(6765): p. 41-5.

45. Jenuwein, T. and C.D. Allis, *Translating the histone code.* Science, 2001. **293**(5532): p. 1074-80.

46. Latham, J.A. and S.Y. Dent, *Cross-regulation of histone modifications.* Nat Struct Mol Biol, 2007. **14**(11): p. 1017-24.

47. Kreutzer, D.A. and J.M. Essigmann, *Oxidized, deaminated cytosines are a source of C --> T transitions in vivo.* Proc Natl Acad Sci U S A, 1998. **95**(7): p. 3578-82.

48. Barsky, D., et al., *New insights into the structure of abasic DNA from molecular dynamics simulations.* Nucleic Acids Res, 2000. **28**(13): p. 2613-26.

49. Moe, A., et al., *Incision at hypoxanthine residues in DNA by a mammalian homologue of the Escherichia coli antimutator enzyme endonuclease V.* Nucleic Acids Res, 2003. **31**(14): p. 3893-900.

50. Branzei, D. and M. Foiani, *Interplay of replication checkpoints and repair proteins at stalled replication forks.* DNA Repair (Amst), 2007. **6**(7): p. 994-1003.

51. Branzei, D. and M. Foiani, *Maintaining genome stability at the replication fork.* Nat Rev Mol Cell Biol, 2010. **11**(3): p. 208-19.

52. Hoeijmakers, J.H., *Genome maintenance mechanisms for preventing cancer.* Nature, 2001. **411**(6835): p. 366-74.

53. Fiers, W., et al., *Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene.* Nature, 1976. **260**(5551): p. 500-7.

54. Morris, P., et al., *Genomic characterization of mycobacteriophage Giles: evidence for phage acquisition of host DNA by illegitimate recombination.* J Bacteriol, 2008. **190**(6): p. 2172-82.

55. Wilson, G.G., *Organization of restriction-modification systems.* Nucleic Acids Res, 1991. **19**(10): p. 2539-66.

56. Wilson, G.G. and N.E. Murray, *Restriction and modification systems.* Annu Rev Genet, 1991. **25**: p. 585-627.

57. Warren, R.A., *Modified bases in bacteriophage DNAs.* Annu Rev Microbiol, 1980. **34**: p. 137-58.

58. Wyatt, G.R. and S.S. Cohen, *A new pyrimidine base from bacteriophage nucleic acids.* Nature, 1952. **170**(4338): p. 1072-3.

59. Morera, S., et al., *High resolution crystal structures of T4 phage beta-glucosyltransferase: induced fit and effect of substrate and metal binding.* J Mol Biol, 2001. **311**(3): p. 569-77.

60. Lunt, M.R. and E.A. Newton, *Glucosylated Nucleotide Sequences from T-Even Bacteriophage Deoxyribonucleic Acids.* Biochem J, 1965. **95**: p. 717-23.

61. Marinus, M.G., *DNA methylation and mutator genes in Escherichia coli K-12.* Mutat Res, 2010. **705**(2): p. 71-6.

62. Marinus, M.G. and N.R. Morris, *Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12.* J Bacteriol, 1973. **114**(3): p. 1143-50.

63. Geier, G.E. and P. Modrich, *Recognition sequence of the dam methylase of Escherichia coli K12 and mode of cleavage of Dpn I endonuclease.* J Biol Chem, 1979. **254**(4): p. 1408-13.

64. Lobner-Olesen, A., O. Skovgaard, and M.G. Marinus, *Dam methylation: coordinating cellular processes.* Curr Opin Microbiol, 2005. **8**(2): p. 154-60.

65. Gomez-Eichelmann, M.C. and J. Ramirez-Santos, *Methylated cytosine at Dcm (CCATGG) sites in Escherichia coli: possible function and evolutionary implications.* J Mol Evol, 1993. **37**(1): p. 11-24.

66. Dianov, G. and T. Lindahl, *Reconstitution of the DNA base excision-repair pathway.* Curr Biol, 1994. **4**(12): p. 1069-76.

67. Robertson, A.B. and S.W. Matson, *Reconstitution of the very short patch repair pathway from Escherichia coli.* J Biol Chem, 2012. **287**(39): p. 32953-66.

68. Bell, D.C. and C.G. Cupples, *Very-short-patch repair in Escherichia coli requires the dam adenine methylase.* J Bacteriol, 2001. **183**(12): p. 3631-5.

69. Palmer, B.R. and M.G. Marinus, *The dam and dcm strains of Escherichia coli--a review.* Gene, 1994. **143**(1): p. 1-12.

70. Yu, Z., et al., *The protein that binds to DNA base J in trypanosomatids has features of a thymidine hydroxylase.* Nucleic Acids Res, 2007. **35**(7): p. 2107-15.

71. Borst, P. and R. Sabatini, *Base J: discovery, biosynthesis, and possible functions.* Annu Rev Microbiol, 2008. **62**: p. 235-51.

72. Razin, A. and A.D. Riggs, *DNA methylation and gene function.* Science, 1980. **210**(4470): p. 604-10.

73. Ehrlich, M. and R.Y. Wang, *5-Methylcytosine in eukaryotic DNA.* Science, 1981. **212**(4501): p. 1350-7.

74. Berger, J. and A. Bird, *Role of MBD2 in gene regulation and tumorigenesis.* Biochem Soc Trans, 2005. **33**(Pt 6): p. 1537-40.

75. Jones, P.L., et al., *Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription.* Nat Genet, 1998. **19**(2): p. 187-91.

76. Nan, X., et al., *Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex.* Nature, 1998. **393**(6683): p. 386-9.

77. Kriaucionis, S. and N. Heintz, *The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain.* Science, 2009. **324**(5929): p. 929-30.

78. Tahiliani, M., et al., *Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.* Science, 2009. **324**(5929): p. 930-5.

79. Branco, M.R., G. Ficz, and W. Reik, *Uncovering the role of 5-hydroxymethylcytosine in the epigenome.* Nat Rev Genet, 2012. **13**(1): p. 7-13.

80. Cannon, S.V., A. Cummings, and G.W. Teebor, *5-Hydroxymethylcytosine DNA glycosylase activity in mammalian tissue.* Biochem Biophys Res Commun, 1988. **151**(3): p. 1173-9.

81. Xu, Y., et al., *Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells.* Mol Cell, 2011. **42**(4): p. 451-64.

82. Wossidlo, M., et al., *5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming.* Nat Commun, 2011. **2**: p. 241.

83. Robertson, A.B., et al., *Pull-down of 5-hydroxymethylcytosine DNA using JBP1-coated magnetic beads*. Nat Protoc, 2012. **7**(2): p. 340-50.

84. Robertson, A.B., et al., *A novel method for the efficient and selective identification of 5-hydroxymethylcytosine in genomic DNA*. Nucleic Acids Res, 2011. **39**(8): p. e55.

85. Robertson, J., A.B. Robertson, and A. Klungland, *The presence of 5-hydroxymethylcytosine at the gene promoter and not in the gene body negatively regulates gene expression*. Biochem Biophys Res Commun, 2011. **411**(1): p. 40-3.

86. Ito, S., et al., *Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification*. Nature, 2010. **466**(7310): p. 1129-33.

87. Guo, J.U., et al., *Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain*. Cell, 2011. **145**(3): p. 423-34.

88. Schiesser, S., et al., *Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing*. Angew Chem Int Ed Engl, 2012. **51**(26): p. 6516-20.

89. Ficz, G., et al., *Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation*. Nature, 2011. **473**(7347): p. 398-402.

90. Dawlaty, M.M., et al., *Combined deficiency of Tet1 and Tet2 causes epigenetic abnormalities but is compatible with postnatal development*. Dev Cell, 2013. **24**(3): p. 310-23.

91. Gao, Y., et al., *Replacement of Oct4 by Tet1 during iPSC induction reveals an important role of DNA methylation and hydroxymethylation in reprogramming*. Cell Stem Cell, 2013. **12**(4): p. 453-69.

92. Costa, Y., et al., *NANOG-dependent function of TET1 and TET2 in establishment of pluripotency*. Nature, 2013. **495**(7441): p. 370-4.

93. Gu, T.P., et al., *The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes*. Nature, 2011. **477**(7366): p. 606-10.

94. Gama-Sosa, M.A., et al., *The 5-methylcytosine content of DNA from human tumors*. Nucleic Acids Res, 1983. **11**(19): p. 6883-94.

95. Goelz, S.E., et al., *Hypomethylation of DNA from benign and malignant human colon neoplasms*. Science, 1985. **228**(4696): p. 187-90.

96. Fearon, E.R. and B. Vogelstein, *A genetic model for colorectal tumorigenesis*. Cell, 1990. **61**(5): p. 759-67.

97. Diala, E.S., et al., *Extent of DNA methylation in human tumor cells*. J Natl Cancer Inst, 1983. **71**(4): p. 755-64.

98. Ohtani-Fujita, N., et al., *CpG methylation inactivates the promoter activity of the human retinoblastoma tumor-suppressor gene*. Oncogene, 1993. **8**(4): p. 1063-7.

99. Merlo, A., et al., *5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers*. Nat Med, 1995. **1**(7): p. 686-92.

100. Melki, J.R., P.C. Vincent, and S.J. Clark, *Concurrent DNA hypermethylation of multiple genes in acute myeloid leukemia*. Cancer Res, 1999. **59**(15): p. 3730-40.

101. Aaby, P., et al., *Spacing, crowding, and child mortality in Guinea-Bissau*. Lancet, 1983. **2**(8342): p. 161.

102. Ko, M., et al., *Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2*. Nature, 2010. **468**(7325): p. 839-43.

103. Langemeijer, S.M., M.G. Aslanyan, and J.H. Jansen, *TET proteins in malignant hematopoiesis.* Cell Cycle, 2009. **8**(24): p. 4044-8.

104. Langemeijer, S.M., et al., *Acquired mutations in TET2 are common in myelodysplastic syndromes.* Nat Genet, 2009. **41**(7): p. 838-42.

105. Blanco, M.A., et al., *Identification of staphylococcal nuclease domain-containing 1 (SND1) as a Metadherin-interacting protein with metastasis-promoting functions.* J Biol Chem, 2011. **286**(22): p. 19982-92.

106. Lian, C.G., et al., *Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma.* Cell, 2012. **150**(6): p. 1135-46.

107. Chen, Q., et al., *TET2 promotes histone O-GlcNAcylation during gene transcription.* Nature, 2013. **493**(7433): p. 561-4.

108. Wu, H., et al., *Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells.* Genes Dev. **25**(7): p. 679-84.

109. Kinney, S.M., et al., *Tissue-specific distribution and dynamic changes of 5-hydroxymethylcytosine in mammalian genomes.* J Biol Chem, 2011. **286**(28): p. 24685-93.

110. Stroud, H., et al., *5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells.* Genome Biol, 2011. **12**(6): p. R54.

111. Williams, K., J. Christensen, and K. Helin, *DNA methylation: TET proteins-guardians of CpG islands?* EMBO Rep, 2012. **13**(1): p. 28-35.

112. Wu, H., et al., *Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells.* Genes Dev, 2011. **25**(7): p. 679-84.

113. Wu, H. and Y. Zhang, *Tet1 and 5-hydroxymethylation: a genome-wide view in mouse embryonic stem cells.* Cell Cycle, 2011. **10**(15): p. 2428-36.

114. Gadji, M., et al., *Nuclear remodeling as a mechanism for genomic instability in cancer.* Adv Cancer Res, 2011. **112**: p. 77-126.

115. Pornthanakasem, W., et al., *LINE-1 methylation status of endogenous DNA double-strand breaks.* Nucleic Acids Res, 2008. **36**(11): p. 3667-75.

116. Nestor, C., et al., *Enzymatic approaches and bisulfite sequencing cannot distinguish between 5-methylcytosine and 5-hydroxymethylcytosine in DNA.* Biotechniques, 2010. **48**(4): p. 317-9.

117. Huang, K.J., C.C. Ku, and I.R. Lehman, *Endonuclease G: a role for the enzyme in recombination and cellular proliferation.* Proc Natl Acad Sci U S A, 2006. **103**(24): p. 8995-9000.

118. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.

119. Su, Z., L. Han, and Z. Zhao, *Conservation and divergence of DNA methylation in eukaryotes: new insights from single base-resolution DNA methylomes.* Epigenetics, 2011. **6**(2): p. 134-40.

120. Booth, M.J., et al., *Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution.* Science, 2012. **336**(6083): p. 934-7.

121. Sun, Z., et al., *High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells.* Cell Rep, 2013. **3**(2): p. 567-76.

122. Yu, M., et al., *Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome.* Cell, 2012. **149**(6): p. 1368-80.

123.   Williams, K., et al., *TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity.* Nature, 2011. **473**(7347): p. 343-8.

124.   Cortellino, S., et al., *Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair.* Cell, 2011. **146**(1): p. 67-79.

125.   Ito, S., et al., *Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine.* Science, 2011. **333**(6047): p. 1300-3.

126.   Parrish, J., et al., *Mitochondrial endonuclease G is important for apoptosis in C. elegans.* Nature, 2001. **412**(6842): p. 90-4.

127.   Cote, J., J. Renaud, and A. Ruiz-Carrillo, *Recognition of (dG)n.(dC)n sequences by endonuclease G. Characterization of the calf thymus nuclease.* J Biol Chem, 1989. **264**(6): p. 3301-10.

128.   Ohsato, T., et al., *Mammalian mitochondrial endonuclease G. Digestion of R-loops and localization in intermembrane space.* Eur J Biochem, 2002. **269**(23): p. 5765-70.

129.   Zan, H., et al., *Endonuclease G plays a role in immunoglobulin class switch DNA recombination by introducing double-strand breaks in switch regions.* Mol Immunol, 2011. **48**(4): p. 610-22.

130.   Li, L.Y., X. Luo, and X. Wang, *Endonuclease G is an apoptotic DNase when released from mitochondria.* Nature, 2001. **412**(6842): p. 95-9.

131.   van Loo, G., et al., *Endonuclease G: a mitochondrial protein released in apoptosis and involved in caspase-independent DNA degradation.* Cell Death Differ, 2001. **8**(12): p. 1136-42.

132.   Cote, J. and A. Ruiz-Carrillo, *Primers for mitochondrial DNA replication generated by endonuclease G.* Science, 1993. **261**(5122): p. 765-9.

133.   David, K.K., et al., *EndoG is dispensable in embryogenesis and apoptosis.* Cell Death Differ, 2006. **13**(7): p. 1147-55.

134.   Irvine, R.A., et al., *Generation and characterization of endonuclease G null mice.* Mol Cell Biol, 2005. **25**(1): p. 294-302.

135.   Huang, K.J., B.V. Zemelman, and I.R. Lehman, *Endonuclease G, a candidate human enzyme for the initiation of genomic inversion in herpes simplex type 1 virus.* J Biol Chem, 2002. **277**(23): p. 21071-9.

List of Appendices

Appendix I

Robertson J.*, Robertson A.B.*, Klungland A. "The presence of 5-hydroxymethylcytosine at the gene promoter and not in the gene body negatively regulates gene expression" (* contributed equally)

Biochem Biophys Res Commun. 2011 Jul 22;411(1):40-3.

Appendix II

Robertson A.B.*, **Robertson J.***, Fusser M., Klungland A. "Endonuclease G Preferentially Cleaves 5-Hydroxymethylcytosine-Modified DNA creating a substrate for recombination" (* contributed equally)

Nucleic Acids Res., in press (2014).