The Biochemistry of DNA Oxidation- and Repair-Mediated Active DNA Demethylation

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Alain Weber

aus

Zürich, Schweiz

Basel, 2015

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Primo Schär (Fakultätsverantwortlicher und Dissertationsleiter)
Prof. Dr. Orlando Schärer (Korreferent)

Basel, den 21.04.2015

<div align="right">

Prof. Dr. Jörg Schibler

Dekan der Philosophisch-Naturwissenschaftlichen Fakultät

</div>

Acknowledgements

Table of Contents

Abbreviations

5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5hmU	5-hydroxymethyluracil
5mC	5-methylcytosine
α-KG	α-ketoglutarate
A	Adenine
AID	Activation-induced deaminase
AP	Apurinic/apyrimidinic
APE1	AP endonuclease 1
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide
BER	Base excision repair
bp	Base pair
C	Cytosine
CpG	C - G dinucleotide
CGI	CpG island
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
dRP	5'-deoxyribose-phosphate
DSB	Double-strand break
EMSA	Electrophoretic mobility shift assay
ESC	Embryonic stem cell
G	Guanine
Gadd45	Growth arrest and DNA-damage-inducible protein 45
GST	Glutathione S-transferase
H3K4/9/27	Histone 3 lysine 4/9/27
HCP	High CpG density promoter
ICP	Intermediate CpG density promoter
ICR	Imprint control region
IP	Immunoprecipitation
LC/MS/MS	Liquid chromatography - tandem mass spectrometry
LCP	Low CpG density promoter
LIG3	DNA ligase 3
LINE	Long interspersed nuclear element
LTR	Long terminal repeat
MBD4	Methyl-CpG-binding domain protein 4
MUG	Mismatch-directed uracil-DNA glycosylase
NER	Nucleotide excision repair
(Ni-)NTA	(Nickel -) Nitrilotriacetic acid

PGC	Primordial germ cells
POLβ	DNA polymerase β
RNA	Ribonucleic acid
SAM	S-adenosyl-L-methionine
SIM	SUMO interaction motif
SINE	Short interspersed nuclear element
SMUG1	Single-strand selective monofunctional uracil DNA glycosylase 1
SUMO	Small ubiquitin-like modifier
ss	Single-strand
T	Thymine
TDG	Thymine DNA glycosylase
TET1-3	Ten-eleven-translocation family of proteins 1-3
THase	Thymine-7-hydroxylase
TSS	Transcription start site
U	Uracil
UNG2	Uracil-DNA glycosylase 2
wt	Wild-type
XPG	Xeroderma pigmentosum group G
XRCC1	X-ray repair complementing defective repair in chinese hamster cells 1
Y2H	Yeast two-hybrid
Δcat	Catalytically dead mutant

1 Summary

Cells of multicellular organisms, no matter how specialized they are, share the same genetic information, stored in their deoxyribonucleic acid (DNA) sequence. They obtain their identity during lineage commitment and differentiation, where specific gene expression patterns are established and subsequently maintained. This process does not involve the alteration of the DNA sequence itself; instead, it is achieved through mechanisms that modulate the accessibility of the DNA to the transcription machinery and thus control how the genetic code is read and applied. Faithful development and survival of complex multicellular organisms is thus not only depending on the genetic code but is also controlled by an additional layer of information called the epigenetic code. In mammals, the epigenetic information is stored mainly in two forms, posttranslational histone tail modifications and DNA methylation. DNA methylation of the fifth carbon of cytosines (C) yielding 5-methylcytosine (5mC) is predominantly found in palindromic CpG dinucleotides affecting roughly 60 - 80% of them (Bird 2002). Epigenetic memory that comprises both layers of epigenetic information is generally maintained during cell division and in particular DNA methylation poses a fundamental and heritable barrier that prevents regression into an undifferentiated state and loss of cellular identity (Messerschmidt et al. 2014; Seisenberger et al. 2013). DNA methylation is established by DNA methyltransferases (DNMTs) that, in our current understanding, either focus on the (*de novo*) establishment or the maintenance of DNA methylation across cell generations (Jurkowska et al. 2011; Law and Jacobsen 2010). Despite its crucial role, DNA methylation patterns are not only statically maintained but are also subject to dynamic regulation through active as well as passive mechanisms. DNA demethylation events have been observed locus-specifically in differentiated cells (Kangaspeska et al. 2008; M. S. Kim et al. 2009b; Metivier et al. 2008) as well as on a global scale during early development (Oswald et al. 2000; Seisenberger et al. 2012; Smith et al. 2014). Although global erasure of DNA methylation can be obtained efficiently through passive dilution by inhibiting the methylation maintenance machinery, a major caveat of this process is the dependence on repeated DNA replication, reducing the dynamic flexibility. In contrast, DNA demethylation also occurs in an active manner, involving enzymatic activities that can process 5mC and revert it back to unmodified C. While the catalytic mechanism of DNA methylation is well understood and established, the process of removing DNA methylation has puzzled researchers for a long time and a variety of mechanisms have been proposed (Ooi and Bestor 2008; S. C. Wu and Zhang 2010). Many of these pathways, however, have failed to find sufficient support, most often due to a lack of reproducibility or convincing biochemical as well as biological evidence.

In recent years, major advances have been made in the understanding of DNA demethylation and some promising candidate mechanisms have emerged (H. Wu and Zhang 2014). Compelling biochemical as well as biological evidence points towards an involvement of the ten eleven-translocation (TET) family of dioxygenases in the removal of DNA methylation (Pastor et al. 2013). The family consists of 3 members, TET1-3, that share a conserved catalytic core domain enabling iterative oxidation of 5mC to generate 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), which could serve as intermediates in active or passive DNA demethylation

processes (He et al. 2011; Huang et al. 2014; Inoue et al. 2011; Ito et al. 2011). The thymine DNA glycosylase (TDG), originally identified as biochemical activity excising thymine (T) and uracil (U) when mispaired with guanine (G), is able to recognize and excise the TET-mediated 5mC oxidation products 5fC and 5caC (He et al. 2011; Maiti and Drohat 2011). A role for TDG in epigenetic programming and DNA demethylation has also been implicated by gene inactivation studies in animals (Cortazar et al. 2011; Cortellino et al. 2011; Saito et al. 2011) as well as in ES cells (Raiber et al. 2012; L. Shen et al. 2013; C. X. Song et al. 2013a). Together, these findings gave rise to a novel concept of active DNA demethylation (Kohli and Zhang 2013; H. Wu and Zhang 2014) involving TET-mediated 5mC oxidation followed by TDG-initiated DNA repair to release the oxidized 5mC derivative and re-establish the unmethylated state. Despite the fact that this mechanism is plausible and has been widely accepted, there is in fact little evidence supporting a direct link of TET with TDG and DNA repair and mechanistic details, coordination, regulation and targeting of this process remain to be clarified.

In order to gain further insight into TET and TDG-mediated active DNA demethylation, I set out to address some of the imminent mechanistic questions by *in vitro* reconstitution of oxidative DNA demethylation along the TET-TDG axis in combination with base excision repair (BER). I first showed that TDG by itself has no detectable enzymatic activity on 5mC and 5hmC but efficiently recognizes and processes 5caC, a modification that does not affect regular Watson-Crick basepairing (Supplementary results 4.4.1) (He et al. 2011; Maiti and Drohat 2011). The proposed oxidative DNA demethylation mechanism implies a coupled action of TET and TDG to facilitate an efficient but coordinated removal of 5mC. To investigate the coupling mechanism, I tested a potential physical interaction of the two enzymes, through multiple experimental approaches. I could demonstrate that TET1 and TDG physically interact through domains located in the N-terminus as well as the catalytic domain (TET1$_{CD}$) of TET1. Recombinant TET1$_{CD}$/TDG complex, purified from *Escherichia coli (E.coli)* cells co-expressing both proteins, turned out to act as 'demethylase' by combining both enzymatic activities to remove 5mC and 5hmC from synthetic DNA oligonucleotides. After successful reconstitution of 5mC base release with purified recombinant proteins, I combined this activity with the BER machinery and showed complete reconstitution of active DNA demethylation via oxidized intermediates *in vitro*, providing the first experimental evidence that this process is functional in the proposed manner. Moreover, investigation of the process operating at symmetrically modified CpGs suggested that symmetric DNA demethylation is obtained through a processive mechanism that is highly coordinated and acts sequentially on both strands to protect the DNA from the formation of DNA double strand breaks (DSBs). However, the sequential and coordinated repair of two nearby substrates on opposite strands, beneficial in terms of avoiding the formation of DNA DSBs, could have an impact on mutagenesis of CpG dinucleotides. I could show that at fully methylated CpG sites, where spontaneous hydrolytic deamination may occur coincident with oxidative DNA demethylation, the repair of the resulting G/T mismatch is highly disfavored in presence of a G·5caC base pair. The preferential repair of 5caC can then occasionally create a C to T mutation and, hence, lead to the loss of the CpG dinucleotide **(Appendix I)**. Additional experiments revealed that neither TET nor TDG activity is restricted to double-stranded DNA or a CpG context, suggesting that TET-TDG-mediated

DNA demethylation might also occur in other biologically relevant contexts including non-CpG methylation, single-stranded DNA or R-loops (Supplementary results 4.4.1).

The process of active DNA demethylation by TET-TDG-BER has to occur in a tightly regulated and highly coordinated manner to ensure accuracy and genome integrity. TDG was previously described to be regulated by posttranslational modification and non-covalent interaction with the small ubiquitin-like modifiers (SUMO), SUMO1 and SUMO2/3 (Hardeland et al. 2002; Mohan et al. 2007; Steinacher and Schar 2005) and other BER factors are also amongst the increasing number of reported SUMO targets (Cremona et al. 2012; Weber et al. 2014). The biochemical investigation of the functional consequences of SUMO modification, however, has been lagging behind due to the difficulty to generate appreciable amounts of recombinant SUMOylated proteins. Therefore, I, in collaboration with David Schürmann, established a recombinant SUMOylation system, coupling efficient SUMO-conjugation with affinity purification of modified target proteins, and present tools and strategies to generate SUMOylated proteins using versatile binary expression vector systems in protease-deficient *E.coli.* We successfully modified the BER factors TDG and XRCC1 and could show that purified SUMO-modified TDG had retained the expected biochemical properties **(Appendix II)**. I was then also able to modify the N-terminus of TET1 with SUMO1 as well as SUMO3 using the recombinant SUMOylation system and identified SUMO interaction motifs (SIMs) in the TET1 sequence by *in silico* prediction. This indicates that SUMOylation might also be prominently involved in the coordination and regulation of TET-TDG mediated DNA demethylation processes (Supplementary results 4.4.2).

We reasoned that other factors previously proposed to contribute to DNA demethylation might be involved in the concerted action of the TET and TDG enzymatic activities, exerting a regulatory or structural function. The growth arrest and DNA-damage-inducible protein 45 (Gadd45) family of proteins has previously been implicated in active DNA demethylation through Xeroderma pigmentosum group G (XPG)-dependent DNA repair (Barreto et al. 2007; Schmitz et al. 2009) or BER of activation-induced deaminase (AID)-based deamination products (Cortellino et al. 2011; Rai et al. 2008). In collaboration with Zheng Li and Guoliang Xu at the Chinese Academy of Science in Shanghai, I set out to re-investigate a potential role of Gadd45a in the context of oxidative DNA demethylation and provide several lines of evidence that Gadd45a serves as a regulator in the TET-TDG-mediated DNA demethylation pathway. Together, we showed that Gadd45a synergizes with TET and TDG to activate a methylated reporter gene in transfected cells. Moreover, Gadd45a physically interacted with TDG and potentiated TDG glycosylase activity to remove 5fC and 5caC from genomic DNA of transfected HEK293T cells. Finally, deletion of Gadd45a/b in mouse ES cells led to hypermethylation at specific genomic loci, which also gained increased DNA methylation levels and are enriched in 5fC in TDG-deficient cells. Despite the diverse molecular functions that have been attributed to Gadd45a, we were able to connect Gadd45 proteins with DNA demethylation along the TET-TDG axis and propose a regulatory function. My specific contribution to this work was the biochemical examination of Gadd45a on the enzymatic function of TET and TDG **(Appendix III)**.

Taken together, the work presented in my PhD thesis advances our understanding of TET-TDG-mediated active DNA demethylation and the underlying mechanisms. I was able to show that TET

associates with BER by physically interacting with TDG and to provide proof of concept that DNA demethylation can be achieved through the coordinated action of an intricate network of enzymes consisting of TET, TDG and the core components of the DNA BER system. Without question, additional factors and regulatory mechanisms, like Gadd45a and SUMO modification, for which I was able to provide preliminary evidence, will turn out to contribute to coordination, targeting and regulation of this active DNA demethylation process. Additional findings that 5mC oxidation and repair by TET1 and TDG is neither limited to CpG dinucleotides nor to double-stranded DNA suggest that this pathway of DNA demethylation could operate in as yet unidentified biological contexts.

2 Introduction

DNA is the carrier of the genetic information, stored in every cell and needed for proper development and functioning of all known living organisms. The genetic information is encoded by the sequential assembly of four building blocks, the nucleotides. The nucleotides are composed of a nucleobase, the purines adenine (A) and guanine (G) or the pyrimidines cytosine (C) and thymine (T), that are attached to a monosaccharide sugar called deoxyribose and a phosphate group. The DNA consists of two anti-parallel complementary single-strands, coiled around each other to form a double-stranded α-helix. Complementarity is established through the Watson-Crick base pairing, where A pairs with T via two and G pairs with C via three hydrogen bonds.

2.1 Genome Maintenance and DNA Repair

DNA is a fragile chemical structure, susceptible to damage by reactive chemicals, by high energetic radiation and by spontaneous decay. DNA damaging agents can originate from exogenous sources like ionizing radiation, but are predominantly endogenously produced by cellular metabolic processes. DNA damage most frequently affects bases and includes various chemical modifications like methylation, oxidation, alkylation, deamination or even the hydrolytic loss of entire bases (Lindahl et al. 1997). Repairing these often mutagenic or cytotoxic DNA alterations is important for cell survival and genome integrity. Unrepaired base damage can lead to genetic mutations or interfere with replication fork progression and replication associated chromatid breakage. To minimize the deleterious effects of DNA damage and maintain genome integrity, nature has evolved various repair systems, each of which tackling a specific class of DNA lesions (Dalhus et al. 2009; Germann et al. 2012; Scharer 2003).

I will now provide a very general overview on DNA damage and briefly introduce the most common DNA repair pathways, a more detailed introduction on the relevant repair mechanism for this thesis will follow in chapter 2.2.1. DNA DSBs represent the most severe form of DNA damage and are repaired either by non-homologous end joining (NHEJ) or by homologous recombination (HR). As the name implies, NHEJ repairs DSBs without using the information of homologous sequences and is therefore often error prone. After the DSB ends are processed to restore a ligatable configuration, they are simply re-ligated. Unlike in NHEJ, repair by HR uses sequence homology from the sister chromatid as a template for the repair of a broken chromatid. Frequent base modifications caused by alkylating agents can be repaired by direct damage reversal, employing a mechanism where the alkyl group is either transferred to a reactive cysteine group of the repair protein or oxidized and subsequently cleaved off. The majority of base lesions, DNA polymerase errors arising during DNA replication, UV-light and chemical induced bulky DNA adducts and pyrimidine dimers are repaired by numerous excision repair pathways. Errors in DNA replication (misincorporation of nucleotides) are restored by the mismatch repair system (MMR) that recognizes and excises the mismatched base from the newly synthesized DNA strand, which is then re-synthesized at the original parental DNA strand. Lesions that cause more serious helix distortion but only affect one strand are dealt with by the nucleotide

excision repair pathway (NER). After the lesion is recognized and verified, the helix is unwound and an oligonucleotide of 24-32 nucleotides in length containing the damage is released. Finally, gap synthesis and strand re-ligation completes successful repair. Modifications of single bases are generally recognized and repaired by the BER machinery which can excise and replace single erroneous bases (chapter 2.2.1).

Besides its role in maintaining genome integrity, DNA repair, in particular BER, has also been implicated a role in the maintenance of epigenetic DNA modifications (Schar and Fritsch 2010). The epigenome is referred to as the total of functionally relevant chemical modifications to DNA and histone tails, which together guide chromatin structuring and gene expression programs of individual cells at distinct developmental stages (Mohn and Schubeler 2009) (chapter 2.3). Potential repair-mediated mechanisms that are involved in the shaping of the epigenome will be discussed in chapter 2.5.

2.2 DNA Base Modifications and Repair

Nucleobases frequently engage in chemical reactions due to presence of DNA damaging agents, resulting in a variety of products of hydrolysis, deamination, alkylation and oxidation, which can lead to the loss of entire bases or the alteration of coding properties, both of which can be mutagenic or cytotoxic. Base damages are generally recognized and processed by the BER machinery. Besides these mutagenic or toxic base modifications, there are, however, also biologically relevant ones that are placed intentionally as part of the epigenetic code to instruct important biological processes, i.e. DNA methylation of the fifth carbon of C yielding 5mC (chapter 2.3.2). 5mC, which accounts for roughly 1% of total DNA bases is considered the 5th base and is itself a target for further modification. For example, through stepwise oxidation of the methyl group, generating 3 more chemically distinct C derivatives that are present in vertebrate DNA in comparably low amounts; 5hmC, 5fC and 5caC (Delatte et al. 2014). The biological function of these bases is only starting to be unveiled and the evidence available suggests that they have epigenetic functionality. To what extent they represent programmed modifications or oxidative lesions of 5mC, however, is currently unclear. Notably, the epigenetically relevant DNA base modifications, i.e. the methylation of C, is, from a structural point of view, not much different from base alkylations representing DNA damage. It is therefore not surprising that DNA repair mechanisms, and particularly DNA glycosylase-initiated BER, specialized in safeguarding DNA base integrity have been adopted for the control of epigenetic DNA methylation (see chapter 2.5). As this PhD thesis mainly focused on the role of said BER pathway and TDG in epigenetic maintenance, particularly in active DNA demethylation, this DNA repair pathway is introduced in more detail in the following section.

2.2.1   DNA Base Excision Repair (BER)

Substrate Recognition: DNA base modifications or damages are usually processed by the BER pathway (Baute and Depicker 2008; G. L. Dianov and Hubscher 2013; A. B. Robertson et al. 2009). BER is typically targeted and initiated by modification-specific DNA glycosylases that recognize and excise substrate bases from DNA. Most DNA glycosylases recognize base modifications that potentially compromise genome integrity like deaminated Cs or are known to be cytotoxic like oxidation products of all four bases and some types of alkylation damage. Still little is known about the mechanisms how modified bases are found in the vast quantity of undamaged bases in the genome. It is proposed that DNA glycosylases have different strategies of damage pre-selection by establishing loose base contacts through conserved residues close to the catalytic site without the need of fully inserting every base into the active pocket (Jacobs and Schar 2011). This scanning mode allows fast coverage of the genome with a minimized effort. Additionally, damage recognition by DNA glycosylases might also benefit from the fact that most lesions show reduced base pairing stability and helix distortion because of the mismatched context (Yang 2006). These sterical features, but often also specific interactions with the opposing base and strand must be important for the recognition of genuine DNA bases like T when mispaired with G but not when paired with A. For the final damage verification, the putative lesion is flipped out from the interior of the DNA helix and is inserted in the catalytic pocket of the DNA glycosylase, providing more room for molecular interactions. Specific molecular interactions in the catalytic pocket generally lead to a restriction of substrate tolerance but at the same time increases specificity, conferring high damage-specificity to DNA glycosylases. In contrast to DNA damages that predominantly occur in a stochastic way and are randomly distributed across the genome, epigenetic base modifications, i.e. DNA cytosine methylation is most often placed in a targeted way. For the recognition of these modifications, DNA glycosylases might need targeting to the respective loci and don't need to scan the genome themselves.

BER Mechanism: The mechanism of BER (Fig. 2.1) has initially been described and reconstituted with purified enzyme many years ago (G. Dianov and Lindahl 1994; G. L. Dianov and Hubscher 2013; Kubota et al. 1996). Core BER is initiated by a damage-specific DNA glycosylase which hydrolyzes the N-glycosidic bond that links the damaged base with the deoxyribose moiety of the nucleoside without cleaving the phosphodiester bond adjacent to the damaged base. Thereby, base removal from a DNA strand generates a repair intermediate, the so-called apurinic/apyrimidinic site (AP-site) or abasic site. AP-sites can give rise to DNA strand-breaks and impede the progression of RNA or DNA polymerases, resulting in cell toxicity or dysfunction (Wilson and Barsky 2001). Protection and coordinated processing of AP-sites is therefore crucial for cell survival and mutation avoidance and the initial hydrolysis of an irregular base must be tightly coupled to the subsequent incision and repair of the AP-site.

Strand incision at the abasic site is exerted by the AP-endonuclease (APE1) which catalyzes the hydrolytic cleavage of the phosphodiester bond 5' of the abasic site, generating a single-strand break (SSB) with a 5'-deoxyribose phosphate (5'-dRP) residue and a 3'-hydroxyl (3'-OH) end. The single-strand break is recognized by Poly [ADP-ribose] polymerase 1 (PARP1), which then recruits X-ray repair cross-complementing protein 1 (XRCC1), DNA ligase 3 (LIG3) and DNA polymerase β (POLβ)

to insert a new nucleotide and seal the nick. The faithful repair of a SSB requires a 3'-OH and a 5'-phosphate end that allows the downstream action of a DNA polymerase and DNA ligase to seal the nick. Removal of the 5'-dRP moiety and incorporation of a new nucleotide is performed simultaneously by the POLβ. Subsequently, strands are re-ligated by a heterodimer composed of the scaffold protein XRCC1 and LIG3.

SSBs with unligatable ends can also arise spontaneously, by hydrolysis of AP-sites or by the action of bifunctional DNA glycosylases. In contrast to monofunctional DNA glycosylases, these enzymes have the capacity to convert an AP-site into a SSB by β- or β,δ-elimination through an associated DNA lyase activity, generating a 5'-phosphate (5'-P) and a 3'-polyunsaturated aldehyde (3'-PUA) or a 3'-phosphate (3'-P) end, respectively. However, an AP-endonuclease with its intrinsic 3' diesterase activity is then still required to remove the 3' blocking α, β-unsaturated aldehyde, generated by AP lyases. After the conventional 3' end has been restored, gap filling and re-ligation can continue. BER that only incorporates one nucleotide and subsequent re-ligation of the 3'-OH and 5'-P ends is referred to as short-patch BER (SP-BER).
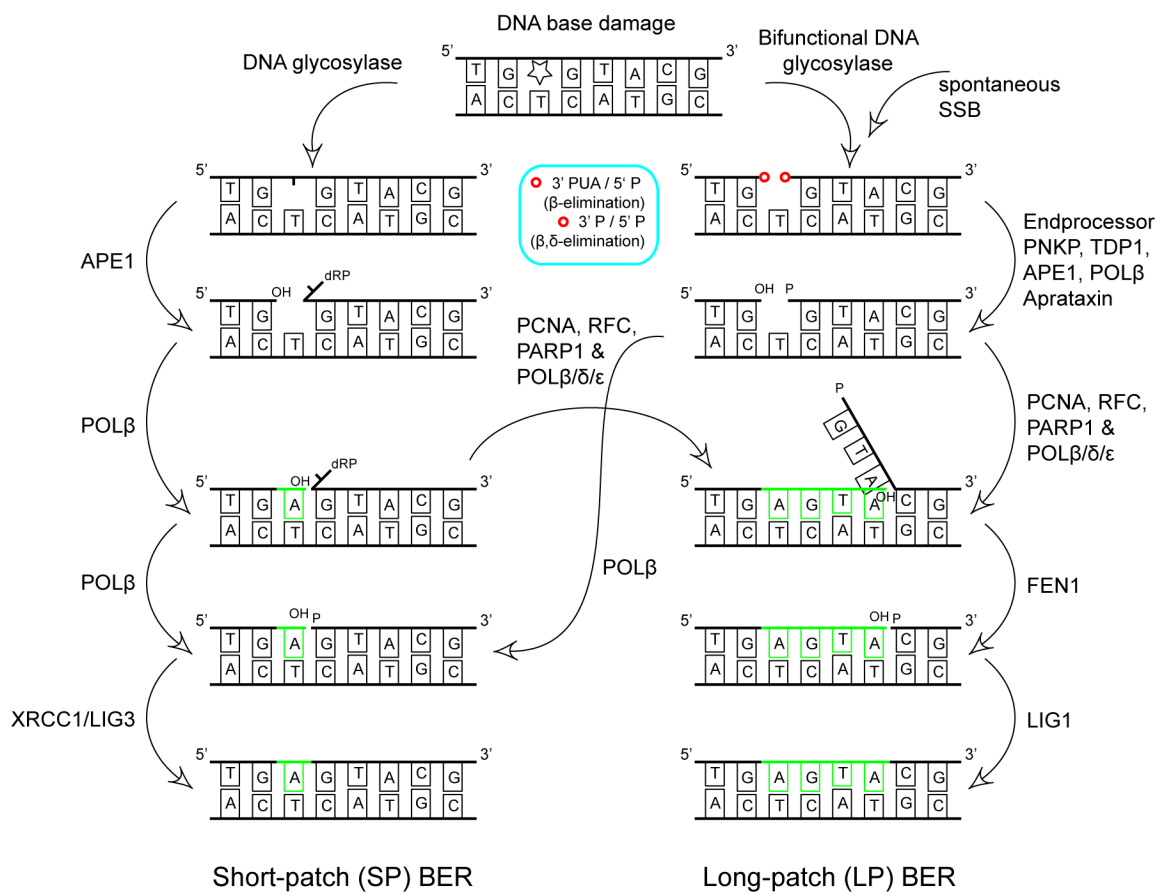
Fig. 2.1: Schematic short- and long patch BER pathways. BER is initiated by a DNA glycosylase to excise the damaged base. The DNA backbone is subsequently cleaved at the AP-site by APE1 or the lyase activity of a bifunctional DNA glycosylase. In the short-patch (SP) pathway, POLβ processes the free DNA ends and incorporates a new nucleotide. Finally the remaining nick is ligated by the XRCC1/LIG3 complex. The long-patch (LP) pathway is employed when the strand break produces ends that are refractory to processing by POLβ. Depending on the DNA ends one of the available endprocessors is employed. POLβ and/or POLδ/ε accomplish strand displacement by incorporation of multiple nucleotides. The DNA flap structure is cleaved by FEN1 followed by strand re-ligation by LIG1. Inspired by (Baute and Depicker 2008; Y. J. Kim and Wilson 2012; A. B. Robertson et al. 2009).

In case the 5'-end is blocked and cannot be processed by SP-BER, an alternative long-patch pathway (LP-BER) is engaged. During mammalian LP-BER, POLβ also incorporates the first nucleotide, elongation and strand displacement are then carried out by the replicative DNA polymerases δ or ε in presence of proliferating cell nuclear antigen (PCNA), replication factor C (RFC) and PARP1. The resulting 5' DNA "flap" structure of 2-12 nucleotides is then removed by the flap endonuclease 1 (FEN1) and subsequently the nick is sealed by DNA ligase I (LIG1).

Besides APE1 and POLβ other SSB end-processors were described, which may be needed downstream of the strand nicking reaction (G. L. Dianov and Hubscher 2013). Polynucleotide Kinase Phosphatase (PNKP) dephosphorylates 3'-ends and phosphorylates 5'-hydroxyl ends; Aprataxin processes 5'-termini blocked by abortive ligation reactions and tyrosyl DNA phosphodiesterase (TDP1) repairs SSBs generated by abortive DNA topoisomerase reactions.

It is expected that the BER process occurs tightly regulated and orchestrated in cells, likely requiring regulatory processes such as posttranslational protein modifications, but it remains to be clarified how the individual steps are activated in a coordinated manner. Two popular models for the coordination of BER have been discussed in the literature, a stepwise 'passing the baton' model based on transient protein-protein interactions and a model suggesting a continuous process performed completely by preassembled DNA repair complexes (G. L. Dianov and Hubscher 2013). However, an emerging concept of how such complex processes can be coordinated is regulation of protein-protein interactions and protein conformations by SUMOylation that could feed into either of the two mentioned models (Jacobs and Schar 2011).

2.2.2 Thymine DNA Glycosylase (TDG)

Discovery and Classification: TDG was initially discovered in mammalian cells when looking for an activity capable of recognizing and processing a T when mispaired with G (Brown and Jiricny 1988; Wiebauer and Jiricny 1989), a situation that can occur through spontaneous hydrolytic deamination of 5mC. Purification of the activity and subsequent molecular cloning of the gene led to the description of human TDG, which was the first reported mismatch-specific DNA glycosylase at that time (Neddermann and Jiricny 1993; Neddermann et al. 1996). TDG became the founding member of a protein family called mismatch-directed uracil-DNA glycosylases (MUG), a subgroup of monofunctional uracil-DNA glycosylases (UDGs) that share a common and characteristic α/β-fold structure (Aravind and Koonin 2000). All MUGs have a comparable and rather simple architecture; a conserved core domain containing the active site is flanked by less conserved N- and C-terminal domains of variable lengths. Within the core domain, the MUG orthologs share between 37-52% sequence identity at the amino acid level (Cortazar et al. 2007). A common structural feature of the MUG proteins is their large catalytic cavity which allows a rather broad spectrum of substrate bases for excision (Barrett et al. 1999). A certain degree of substrate specificity is however granted due to specific interactions between residues in the active site and the G on the complementary strand, giving an evident preference for bases paired with G (Barrett et al. 1998). The human TDG is composed of 410 amino

acids, the mouse variant was found to occur in two splice variants, TDGa and TDGb, with a sequence length of 421 and 397 amino acids, respectively. TDGb is missing 25 amino acids from the N-terminus, which doesn't affect the catalytic domain, and it is currently not clear whether the two isoforms have distinct biological roles.

Recognition and Repair of Base Modifications: Insight into the mechanism of base processing by MUG proteins came from structural studies of the *E.coli* Mug (eMug) (Barrett et al. 1999). The structure suggests an intercalation/nucleotide flipping mechanism; a conserved insertion loop intercalates into the DNA and flips out the substrate base into the catalytic pocket where it is processed. In this scenario the residues from the insertion loop mimic Watson-Crick base pairing with the opposite G through specific contacts to stabilize the enzyme/substrate complex (Maiti et al. 2008; Maiti et al. 2009). In addition, specific contacts to the G positioned 3' to the target nucleotide provide a structural basis for a potential lesions in a CpG context (Maiti et al. 2008). Owing to its name, TDG is best known for its ability to process T in a G/T mismatch, but like other members of the MUG family, it turned out to have a rather broad substrate spectrum including ethenoadducts, deaminated purines, oxidized pyrimidines and derivatives of U when paired with G, with U being the most common physiological one (Borys-Brzywczy et al. 2005; Cortazar et al. 2007; Hardeland et al. 2003). TDG is believed to associate with DNA through its flexible N-terminal domain that switches from an open to a clamp like conformation upon binding to DNA, thereby stabilizing the glycosylase on the DNA (Hardeland et al. 2002; Steinacher and Schar 2005). This clamp structure may enable sliding along the DNA in order to detect substrate bases that are then flipped into the active site pocket upon encounter. A highly conserved asparagine residue (N151 in mouse, N140 in human TDG) in the active site promotes the hydrolysis of the N-glycosidic bond between the base and the deoxyribose-phosphate backbone through an activated water molecule (Barrett et al. 1998; Hardeland et al. 2000). TDG binds product AP-sites with high affinity and fails to dissociate (Hardeland et al. 2000; Waters et al. 1999); it is thus considered to be fully product inhibited and protects the hazardous repair intermediate from forming spontaneous SSBs or DSBs. Biochemical studies have shown that dissociation could be accomplished by either posttranslational modification of TDG with SUMO proteins (Hardeland et al. 2002; Steinacher and Schar 2005) (see below) or in presence of an excess of downstream BER factors, which stimulate enzymatic turnover (Fitzgerald and Drohat 2008; Waters et al. 1999).

SUMOylation of TDG: Posttranslational protein modifications, e.g. phosphorylation, acetylation, attachment of small polypeptides, are a rapid and energetically inexpensive means to reversibly modulate protein function. Enzymatic activity, subcellular localization, stability and also interaction with other proteins can be regulated by such modifications (Barry and Lock 2011). TDG was shown to be modified by SUMO1 and SUMO2/3 that are attached to an acceptor lysine (K341 in mouse TDGa, K330 in human TDG) within a SUMOylation consensus motif (VKEE) (Hardeland et al. 2002; Weber et al. 2014) (Appendix II). SUMO is conjugated to its substrates by an enzymatic cascade consisting of an activating enzyme (E1), a conjugating enzyme (E2) and, sometimes, a SUMO ligase (E3). In contrast to ubiquitylation, SUMOylation utilizes only a single conjugation enzyme, Ubc9, a single heterodimeric activating enzyme consisting of SAE1 and SAE2 (Aos1 and Uba2 in humans) and a limited number of E3-ligases (Weber et al. 2014) (Appendix II). SUMO1, conjugated to the C-terminus

of TDG, interacts functionally with the N-terminus and neutralizes its DNA binding capacity facilitating enzymatic turnover (Hardeland et al. 2002; Weber et al. 2014) (Appendix II). In addition to the SUMOylation site, TDG also contains two SIMs that mediate non-covalent SUMO-interactions with either SUMO alone or other SUMOylated proteins to induce complex formation (Mohan et al. 2007). In this regard, SUMOylation might also exert a regulatory function in the BER pathway, where XRCC1 and APE1 have also been described as SUMO targets (Weber et al. 2014) (Appendix II), to ensure correct orchestration of the enzymatic cascade. The details of this potential regulation are, however, not yet fully understood. Besides SUMOylation, TDG was also reported to be modified by ubiquitylation (Hardeland et al. 2007), phosphorylation and acetylation (Mohan et al. 2010), which could play important roles in the regulation of TDG abundance, localization or enzymatic activity.

Biological Roles: With its ability to process T from G/T mismatches, TDG was believed to predominantly counteract C to T transitions arising from hydrolytic deamination of 5mC at CpG sites. An increasing body of biochemical and genetic evidence, however, suggested that TDG could also be acting outside classical repair and play essential roles in various other biological processes such as embryonic development, regulation of gene expression and DNA demethylation (Cortazar et al. 2007; Sjolund et al. 2013). Over the years, numerous interactions with transcription factors, chromatin modifiers and DNA methyltransferases (DNMTs) have been described, proposing a functional role of TDG in gene regulation either as transcriptional co-factor or by modulating DNA methylation states itself. For example, TDG was shown to interact with the nuclear receptors retinoic acid receptor (RAR) and retinoid X receptor (RXR) (Um et al. 1998) stimulating receptor binding and activating of reporter genes. A similar effect was observed with estrogen receptor α (ERα) that was also shown to physically interact with TDG. There, TDG acted as a transcriptional co-activator for ERα-regulated genes (D. Chen et al. 2003a). Most convincing biological evidence for a role of TDG outside of DNA repair was then described in TDG knockout studies in mice, where it was found to be required for embryonic development (Cortazar et al. 2011; Cortellino et al. 2011; Saito et al. 2011). TDG knockout was shown to confer embryonic lethality and molecular studies pointed towards epigenetic and transcriptional dysregulation. Interestingly, TDG knockout mouse embryonic stem cells (ESCs) show largely normal gene expression patterns. Only upon differentiation into neuronal progenitor cells and mouse embryonic fibroblasts (MEFs), gene expression differences between TDG proficient and deficient cells arise and these are accompanied by epigenetic aberrations i.e. increased *de novo* DNA methylation of promoter CpG islands (CGIs) (Cortazar et al. 2011). This phenotype, not observed for any other DNA glycosylase so far, is most likely not arising from a DNA repair defect alone but is linked to a potential role of TDG in establishing and maintaining proper DNA methylation patterns as well as gene regulation during differentiation. Recent studies further substantiated an involvement of TDG in genome-wide methylation dynamics (L. Shen et al. 2013; C. X. Song et al. 2013a) and mechanistic details of the role of TDG are starting to unravel.

2.3 Epigenetic Memory and DNA Methylation

Every cell of an organism, no matter how specialized, contains an entire copy of the genome, representing an excess of unused genetic information for the majority of cells. Differentiation into and maintenance of specific cell types is not obtained by altering the genome itself but rather by controlling the readout and application of the genetic code. The mechanisms that have evolved to facilitate differential readouts of the genome are thought to modulate DNA accessibility for the transcription machinery. Thus, the information required for the growth, development and survival of a multicellular organism is not only stored genetically in the DNA sequence but also epigenetically as epigenetic memory in the superimposed code that determines gene expression patterns. Epigenetic memory is established by DNA cytosine methylation and a plethora of post-translational histone tail modifications (Fig. 2.2) and relies on faithful inheritance of these marks (Lee et al. 2014). These epigenetic modifications act in a concerted way to determine chromatin structure and regulate transcriptional activity at specific loci. As this thesis focuses mainly on DNA methylation I will only provide a very superficial view on histone modifications and their epigenetic function here. Detailed discussions about the expanding number of histone tail modifications and their implicated functions can be found in (Bannister and Kouzarides 2011; Kouzarides 2007). So-called repressive histone marks can induce a compaction of chromatin (heterochromatin), whereas activating histone marks decondense chromatin, resulting in a more open, accessible form (euchromatin) (Bannister and Kouzarides 2011). Heterochromatin is largely inaccessible for protein complexes involved in gene transcription and thus contains mostly transcriptionally inactive genes and repetitive sequences (Hubner et al. 2013). DNA methylation also contributes to chromatin compaction and can modulate transcriptional activity through other mechanisms, e.g. by recruiting repressive complexes (see chapter 2.3.2). These epigenetic instructions are needed to guide and control important biological processes including cell differentiation and cell lineage maintenance, cellular reprogramming and gametogenesis (Smith and Meissner 2013). A simplified model allows an allocation of biological processes to be regulated by either short- or long-term epigenetic memory. However, this is a very general view mainly applicable to vertebrates and there are likely to be exceptions to the rule.

2.3.1 Short- and Long-term Epigenetic Memory

Different mechanisms have evolved in cells to ensure long-term epigenetic stability and heritability but also to maintain plasticity in transcriptional programs, depending on the requirements of the cell (Mohn and Schubeler 2009; Reik 2007). Current knowledge suggests that the short-term epigenetic memory is mainly established by histone modifications and used to transiently repress or activate specific loci (Fig. 2.2 A). Dynamically regulated processes allow a rather fast adaptation of gene expression patterns in response to internal or external stimuli for example during cell differentiation. In pluripotent ESCs, for instance, genes that are required during development and differentiation are often held in a so-called bivalent chromatin state, which is characterized by both activating (histone 3 lysine 4 methylation, H3K4me) as well as repressive (histone 3 lysine 27 methylation, H3K27me) marks around the transcription start site (TSS), whereas the DNA remains methylation free (Bernstein et al.
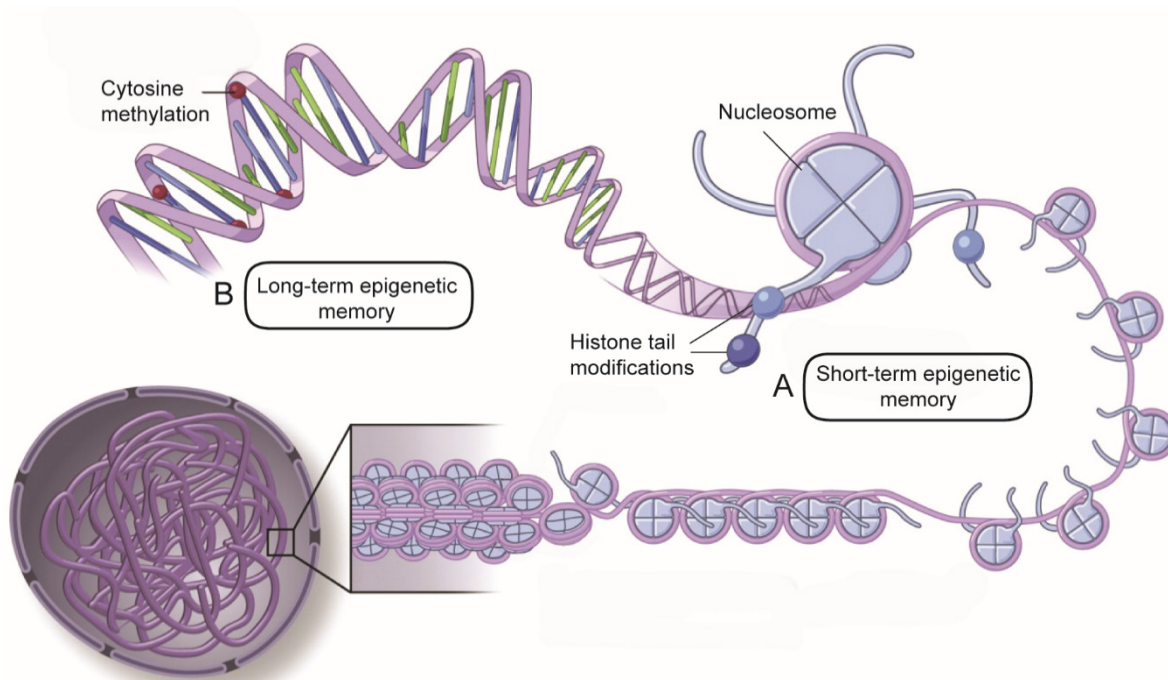
Fig. 2.2: The two main layers of epigenetic memory. (A) Short-term epigenetic memory mainly established by histone tail modifications. (B) Long-term epigenetic memory through DNA methylation in combination with histone tail modifications. Adapted from (Yan et al. 2010).

2006). Removal of the repressive or activating histone mark at the start of differentiation then leads to rapid gene activation or repression, respectively. How the epigenetic information encoded by histone modifications is maintained during DNA replication is not yet completely understood and subject to ongoing research. Putative models are discussed in (Probst et al. 2009).

Long-term epigenetic memory, on the other hand, is sustained throughout the development and lifespan of an organism and is achieved by repressive histone marks in combination with DNA methylation (Fig. 2.2 B). The acquisition of these repressive marks leads to compaction of the genome, which in turn helps to keep it in a stably silenced state. The genomic DNA methylation patterns are established early in development and are generally statically maintained by the DNA methylation machinery during the lifespan of a cell (see chapter 2.3.2). These regions include imprinted genes, repeat regions, transposons and the inactivated X chromosome, where stable transcriptional silencing is crucial for proper development of an organism and genome integrity (Jones 2012; K. D. Robertson 2005; Smith and Meissner 2013). Long-term silencing is also required for pluripotency and developmental genes that need to be shut down during cell lineage commitment to protect the cell from dedifferentiation and loss of identity.

2.3.2   DNA Cytosine Methylation

In mammals, enzymatic DNA methylation occurs almost entirely at the fifth carbon of Cs resulting in 5mC predominantly in CpG dinucleotides (Bird 2002). These CpG dinucleotides are not evenly distributed across the genome and the CpG content is inversely correlated with the level of DNA methylation. Isolated DNA islands with a higher than average content of CpGs, hence called CGIs, are

generally hypomethylated, the rest of the genome is mostly depleted of CpGs and hypermethylated (Cohen et al. 2011; Deaton and Bird 2011). The number of CGIs in mammalian genomes is estimated to be around 30'000; they frequently span gene regulatory elements and thus seem to be involved in gene regulation (Cohen et al. 2011; Illingworth and Bird 2009). DNA methylation at CpG rich gene regulatory elements is generally associated with gene repression. Mechanistically, transcriptional activity is repressed by directly inhibiting association of DNA binding factors or by recruiting methyl-CpG binding proteins that administrate a repressive function through co-repressor molecules (Ballestar and Wolffe 2001; Bird 2002). Lately, most of the research on DNA methylation has focused on CGIs at promoter regions, however, 5mC also appears in gene bodies, enhancers and in non-CpG contexts (Lister et al. 2009). Non-CpG methylation has been described in plants and mammalian stem cells, but recent evidence indicated that it also occurs in somatic cells and could be involved in the regulation of gene expression, but its function in this context is only starting to be unraveled (Pinney 2014).

Methylation of Cs affects an organism on several levels, ranging from altering biochemical and chemical properties down to consequences on the whole organism (Fig. 2.3) (Franchini et al. 2012). Starting from altering van der Waals radii around C5 of cytosine, base stacking, DNA structure and protein association by promoting or inhibiting protein-DNA interactions, DNA methylation also stimulates the compaction of chromatin and the subnuclear localization of gene loci, represses transposons, alters DNA replication efficacy or influences the rate of transcription and RNA processing. Altogether, DNA methylation substantially influences and regulates many important biological processes such as cellular proliferation and differentiation, pluripotency, genetic imprinting, and oncogenesis, hence ensuring proper development and long-term survival of an organism (Jones 2012). DNA methylation in mammals is catalyzed by a group of proteins called the DNA methyltransferases (DNMTs) that classically either focus on the establishment or the maintenance of DNA methylation patterns. In the classical model the methylation system involving a maintenance methyltransferase DNMT1 and the *de novo* methyltransferases DNMT3A/B provides methylation fidelity at two levels: (1) *De novo* methylation, potentially guided by chromatin modifications, establishes and maintains DNA methylation states across genomic regions, which are then (2) maintained by faithful copying at hemimethylated CpGs of newly synthesized DNA to preserve previously established DNA methylation patterns. Notably, this traditional assignment of roles to DNMTs is not universally applicable and emerging concepts have pointed towards an involvement of the *de novo* DNMTs in DNA methylation maintenance at specific loci (Jones and Liang 2009).

De Novo DNA Methylation

The relatively well-established catalysis of *de novo* DNA methylation is carried out by the so-called *de novo* methyltransferases DNMT3A and DNMT3B in the presence of S-Adenosyl-L-methionine (SAM) (Hermann et al. 2004). A third member of the DNA methyltransferase family DNMT3L has no catalytic activity itself but acts as a regulatory factor in the process of DNA methylation and was shown to stimulate DNMT3A and DNMT3B both *in vitro* and *in vivo* (Ooi et al. 2007; Suetake et al. 2004). DNA
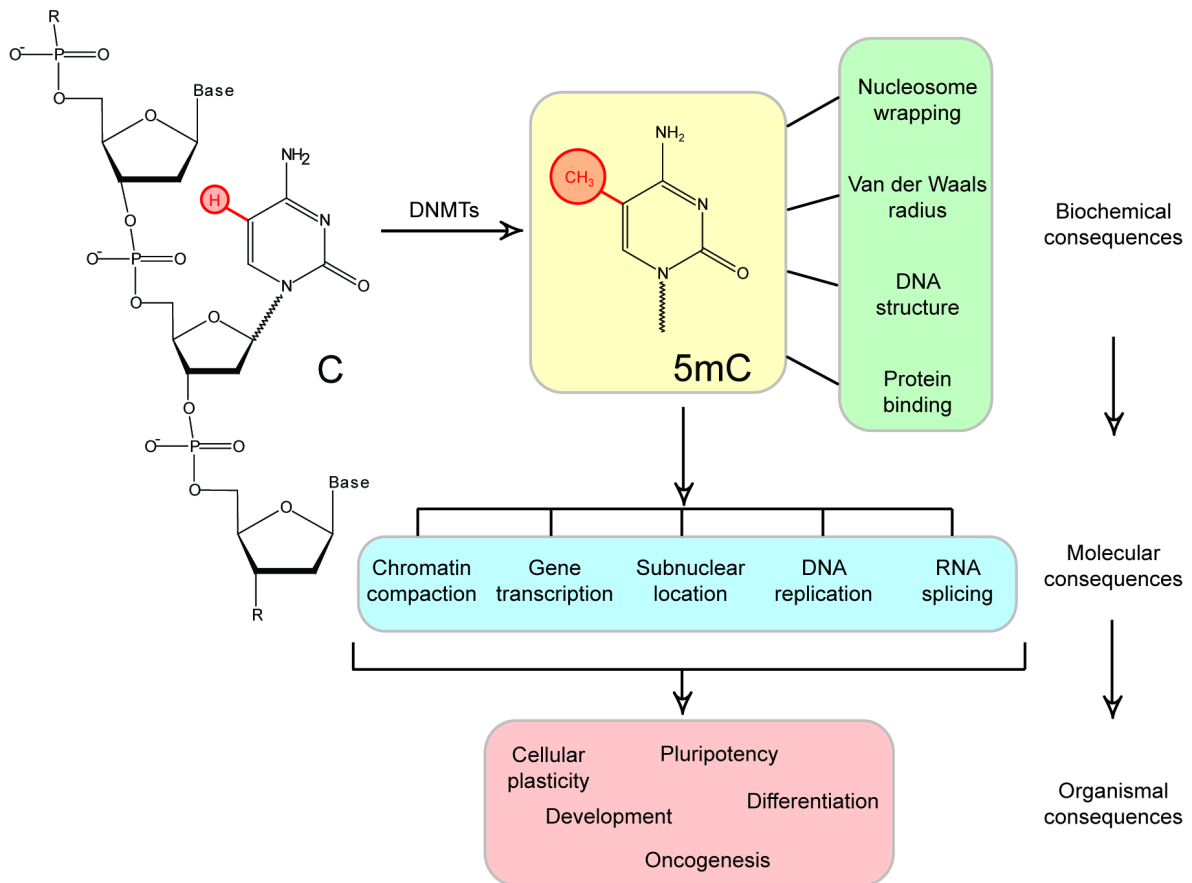
Fig. 2.3: The biological roles and consequences of DNA cytosine methylation. The methyl group on cytosine can induce direct as well as indirect biochemical changes to DNA that serve as molecular signals for biological functions via various means and determines development, physiology, and pathology of an organism. Inspired by (Franchini et al. 2012).

methylation patterns are mainly established during early embryogenesis, at around the time of implantation, through the activity of DNMT3A and DNMT3B (Law and Jacobsen 2010). The importance of this process was demonstrated when it was shown that a lack of *de novo* DNA methylation is incompatible with normal development and causes early embryonic lethality (Okano et al. 1999). Mechanistically, it is not yet fully understood how the methyltransferases are targeted and regulated to generate cell-type specific DNA methylation patterns during embryogenesis and germ cell development. It appears, however, that *de novo* DNA methylation can be directed by chromatin context and influenced by nucleosome positioning, histone modifications as well as chromatin-associated repressor proteins (Denis et al. 2011). The hierarchical relationship between DNA methylation and histone modifications, however, is complex and still needs to be resolved. A current view is that transient downregulation of genes through histone modifications precedes induction of DNA methylation and, thus, that inactive genes are more susceptible to *de novo* DNA methylation than active ones (Smith and Meissner 2013). This idea is supported by studies showing that DNMT3L interacted with unmethylated H3K4 to enable *de novo* DNA methylation, but the interaction was strongly inhibited by the presence of the transcriptional activation mark H3K4 methylation (Ooi et al. 2007; Y. Zhang et al. 2010b). On the other hand, the repressive chromatin mark H3K9 methylation, did not negatively affect *de novo* DNA methylation, indicating that the methylation machinery can discriminate between different chromatin states thus keeping transcriptionally active regions free of methylation. An induction of DNA methylation would then require coordinated action of a H3K4me3

demethylase or nucleosome remodeler to provide a nucleosome at previously depleted regions (e.g. lymphoid-specific helicase (LSH) (H. Zhu et al. 2006)) and subsequent action of a H3K9 methyltransferase (e.g. G9A) to initiate heterochromatin formation and recruitment of DNMT3A and DNMT3B to ensure long-term silencing (Dong et al. 2008).

Maintenance DNA Methylation

DNA methylation maintenance is a process that assures stable reproduction of DNA methylation patterns between cell generations. This is achieved by DNMT1 and its partner UHRF1, which copy the DNA methylation signature from the parental onto the daughter strand during DNA replication and repair. Consistently, it was found that the DNA methylation maintenance machinery prefers hemimethylated over unmethylated DNA (Bostick et al. 2007; Hermann et al. 2004). Intriguingly however, DNMT1 seems unable to maintain DNA methylation on its own as a gradual loss of methylation was observed in ESCs that lack DNMT3A and DNMT3B (T. Chen et al. 2003b). Furthermore, maintenance by DNMT1 alone might be challenging at regions with high frequency of 5mC, possibly requiring a more complex maintenance system. In a newly emerging concept, the bulk of DNA methylation is still maintained by DNMT1 but at sites with higher CpG density it involves specific targeting and cooperation between the DNMTs. DNA methylation patterns at CGIs are proposed to be maintained by a template process involving DNMT1 and a stochastic process involving *de novo* methylation by DNMT3 methyltransferases (Jones and Liang 2009). Unlike DNMT1, which is regarded as a 'reader' of DNA sequences that places methyl groups at hemimethylated CpGs of newly synthesized DNA regardless of chromatin state, DNMT3 enzymes are anchored to nucleosomes and do not 'read' DNA but instead have a constitutive activity to methylate sites missed by DNMT1 in highly methylated regions (Jeong et al. 2009). The importance of proper establishment and maintenance of DNA methylation in embryonic development was demonstrated by studies showing that mice lacking either DNMT1 or DNMT3B are not viable and die during embryonic development and mice without DNMT3A die within a few weeks after birth (E. Li et al. 1992; Okano et al. 1999).

2.3.3 Biological Functions of DNA Methylation

Transcription control: CGIs are often associated with gene regulatory regions but nearly half of the identified CGIs are "orphans" that are not associated with annotated sequences (Illingworth et al. 2010). These "orphans" behave like promoter CGIs but their function is not completely understood. The hypomethylated state of CGIs is believed to be mediated by transcription factor binding (Macleod et al. 1994). Furthermore, unmethylated CpG-rich regions are bound by CXXC finger protein 1 (CFP1), which recruits histone H3K4 methyltransferases to maintain these sites in an unmethylated and transcriptionally active state (Thomson et al. 2010). By contrast, DNA methylation at promoter regions has been linked to transcriptional repression. This is however not always the case and depends on the CpG density at gene regulatory elements. Promoters have thus been classified into three categories based on the CpG content; high CpG density promoters (HCPs), intermediate CpG density promoters (ICPs) and low CpG density promoters (LCPs) (Meissner et al. 2008). Whereas HCPs and ICPs are

downregulated upon methylation, LCPs remain transcriptionally active regardless of their methylation status (Meissner et al. 2008). Similarly to promoter regions, enhancers also have characteristic DNA methylation patterns and hypomethylation is associated with transcription factor binding and active gene expression (Stadler et al. 2011). However, as described above, dynamic gene expression patterns are often mediated by the short-term epigenetic memory and histone modifications. DNA methylation serves more to manifest and stabilize the transcriptionally silent state.

Imprints: Genomic imprinting is a developmental process where parent-of-origin-specific gene expression is established. Manifestation of imprints, e.g. allele-specific DNA methylation, is coordinated by differential DNA methylation at cis-regulatory elements called imprinting control regions (ICRs). Methylation of ICRs guides downstream processes to maintain allele-specific gene clusters either repressed or active. Methylation at imprints is introduced during germ cell differentiation by the *de novo* DNMTs and relies on the enzymatically inactive DNMT3L (Kaneda et al. 2004). These established patterns are maintained throughout life; even in the zygote, where global erasure of DNA methylation is observed (see chapter 2.4.2), imprinted genes are protected and escape this epigenetic reprogramming process. Proper establishment of epigenetic imprinting patterns is crucial for embryogenesis and DNMT3L-deficient mice lacking maternal DNA methylation at ICRs die by embryonic day 10.5 (Henckel et al. 2009).

Transposable elements: Roughly 40% of the mammalian genome is comprised of transposable elements of various classes (Lander et al. 2001). The three major classes are; long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and long terminal repeat- (LTRs) containing elements. LINE as well as LTR elements encode strong promoters that must be kept in a hypermethylated and thus repressed state because active transposable elements are highly mutagenic and recombinogenic; loss of DNA methylation may cause transcriptional activation and (retro)transposition. Furthermore, activation of retrotransposons can also deregulate the expression of neighboring genes by acting as enhancers or promoters (Girard and Freeling 1999). Mechanistically, the H3K9 methyltransferase SETDB1 appears to act prior to DNMT recruitment and DNA methylation is then established to reinforce and stabilize the repressed state (Karimi et al. 2011).

Pericentromeric repeats: These elements are non-coding but show latent transcriptional activity, similar to transposable elements. Minor and major satellite elements extend from the centromere in thousands to tens of thousands of tandem copies (Lehnertz et al. 2003). Transcriptional repression and heterochromatization of these regions is essential for proper chromosome alignment and segregation during mitosis. Continued transcription of these repeats causes rearrangements in the vicinity of the centromeres likely due to chromosome misalignment during mitosis (Gopalakrishnan et al. 2009). Also this process appears to be orchestrated in a sequential manner, where H3K9 methylation is deposited by SUV39H1, which in turn recruits DNMT3B to stably silence these regions (Lehnertz et al. 2003).

X chromosome inactivation: Gene dosage control in females is obtained by random inactivation of one of the two X chromosomes. Early evidence suggested that DNA methylation plays a role in this process as treatment with the DNA demethylating agent 5-azadeoxycytidine resulted in reactivation of

several X-linked genes (Mohandas et al. 1981). It is, however, not entirely clear whether silencing precedes DNA methylation. Experimental evidence, however, indicated that *de novo* DNA methylation is again not the initiator of silencing but rather the long-term fixation of X inactivation (Lock et al. 1987). The current model suggests that inactivation is induced by the expression of a cis-acting non-coding RNA, Xist, which coats the X chromosome, followed by chromatin changes (repressive H3K9 and H3K27 methylation) and ultimately DNA methylation at promoter CGIs (Wutz 2011).

2.4 Dynamics of DNA Methylation

2.4.1 DNA Methylation Stability/Fidelity

DNA methylation is an important component in the proper development of an organism and guides the differentiation of developmentally potent cells into any specific cell type. As development proceeds and cells progress in differentiation, they are directed towards their future lineages through the establishment of DNA methylation landscapes, which finally need to be stably maintained in the terminally differentiated cells. In our current understanding, bulk genomic DNA methylation is stably maintained within cell lineages, tissues and throughout life and includes regions that could cause genomic instability like repetitive elements and retrotransposons. Besides methylation of these bulk genomic regions, cells also need to dynamically establish specific gene expression patterns mediated in part by DNA methylation at gene regulatory elements. With increasing commitment, cells lose their plasticity and the pluripotency network is downregulated concomitant with an upregulation of cell line-specific genes (Borgel et al. 2010). In this context, DNA methylation provides a framework and epigenetic barrier, which guides and restricts differentiation and prevents regression into an undifferentiated state (Messerschmidt et al. 2014; Seisenberger et al. 2013). Stability and heritability of these cell type-specific DNA methylation landscapes and gene expression patterns at given developmental stages is crucial for cell fate and errors could lead to aberrant cell function and a loss of cell identity. Regarding the extent of DNA methylation, it is likely that the fidelity of the DNA methylation patterns varies among different genomic regions, depending on their function reflected in CpG density and chromatin state. Bulk genomic regions might not need the same degree of fidelity and are more tolerant to stochastic errors than gene regulatory elements where even small errors could have a much larger impact (Schar and Fritsch 2010). In agreement with that, promoter-associated CGIs were found to be more effectively protected from *de novo* DNA methylation than CGIs outside promoter regions (Ushijima et al. 2003). It appears that a constant methylation pressure sets and maintains the genome in a highly methylated state, at promoter-associated CGIs and other gene regulatory elements, however, supplementary mechanisms have evolved to keep them methylation free and in an active or transcription-poised state. Fidelity of DNA methylation patterns thus seems to be generally higher in regions that are more dynamically regulated and important for progression of development.

Loss of methylation fidelity in somatic tissues leads to epigenetic instability, which is associated with disease development. Unscheduled epigenetic changes correlate with cellular ageing and alterations

in the distribution and levels of 5mC can be found in most, if not all, cancers. Cancer epigenomes are often characterized by genome-wide DNA hypomethylation accompanied by hypermethylation of CGI containing gene promoters (Jones and Baylin 2002). Extensive DNA hypomethylation leads to a loss of cell identity and can be associated with an increased developmental potency (see chapter 2.4.2), which is in part reflected in the renewing and growth potential of cancer cells and tumors. The range of diseases associated with epigenetic alterations and particularly with dysregulated DNA methylation expands beyond cancer as exemplified by imprinting disorders (e.g. loss of imprinting (LOI), Prader-Willi syndrome (PWS)), repeat-instability diseases (e.g. fragile X syndrome (FRAXA), facioscapulohumeral muscular dystrophy) and others (K. D. Robertson 2005).

2.4.2 Resetting DNA Methylation Patterns

Despite the robust mechanisms of establishment and maintenance of DNA methylation patterns, DNA methylation does not under all circumstances represent a static epigenetic modification but is also subject to dynamic regulation. Dynamic regulation, however, implicates the existence of mechanisms of DNA demethylation. To what extent and under what conditions such mechanisms operate to dynamically regulate DNA methylation patterns is not yet fully understood. DNA demethylation has been observed to occur at specific loci in differentiated cells as well as on a global scale during early embryonic development (Fig. 2.4). It is believed that epigenetic reprogramming, including large scale erasure of DNA methylation, is closely associated with cell fate transition and restoration of developmental potency (Seisenberger et al. 2013). In the mouse model, widespread erasure of DNA methylation has been detected in the zygote immediately after fertilization and again in the establishment of the primordial germ cells (PGCs) which are the direct progenitors of sperm and oocytes (Messerschmidt et al. 2014; Seisenberger et al. 2013; Smith and Meissner 2013).

DNA demethylation in PGCs: PGCs are derived from the epiblast, which is already directed towards somatic lineages. At early stages of mouse embryogenesis (E6.5), these progenitors show characteristics in DNA methylation, chromatin modification and gene expression profiles that are very similar to their somatic origin (i.e. pluripotency and germline-specific genes are tightly repressed by DNA methylation). Until E12.5-E13.5, however, DNA methylation virtually disappears in both parental genomes, imprints are erased and the X chromosome is reactivated in the female progenitors. Most recent studies have revealed that this global loss of DNA methylation occurs in two distinct phases involving both passive and active mechanisms (Seisenberger et al. 2012; Yamaguchi et al. 2013). In the first phase, beginning at around E8.0, bulk DNA methylation is lost indiscriminately, most likely through passive dilution due to a downregulation of the DNA methylation machinery (Kagiwada et al. 2012; Seisenberger et al. 2012). Regions that escape this first wave of passive DNA demethylation including imprints, germline-specific genes and CGIs on the inactive X chromosome are then affected in the second phase of DNA demethylation. Demethylation at these sites appears to require active triggering by modifying 5mC (Hackett et al. 2013; Vincent et al. 2013; Yamaguchi et al. 2013) (putative mechanisms are discussed in detail in chapter 2.5). After the extensive loss of DNA methylation, both genomes undergo *de novo* DNA methylation. Interestingly, the extent of re-methylation seems to be
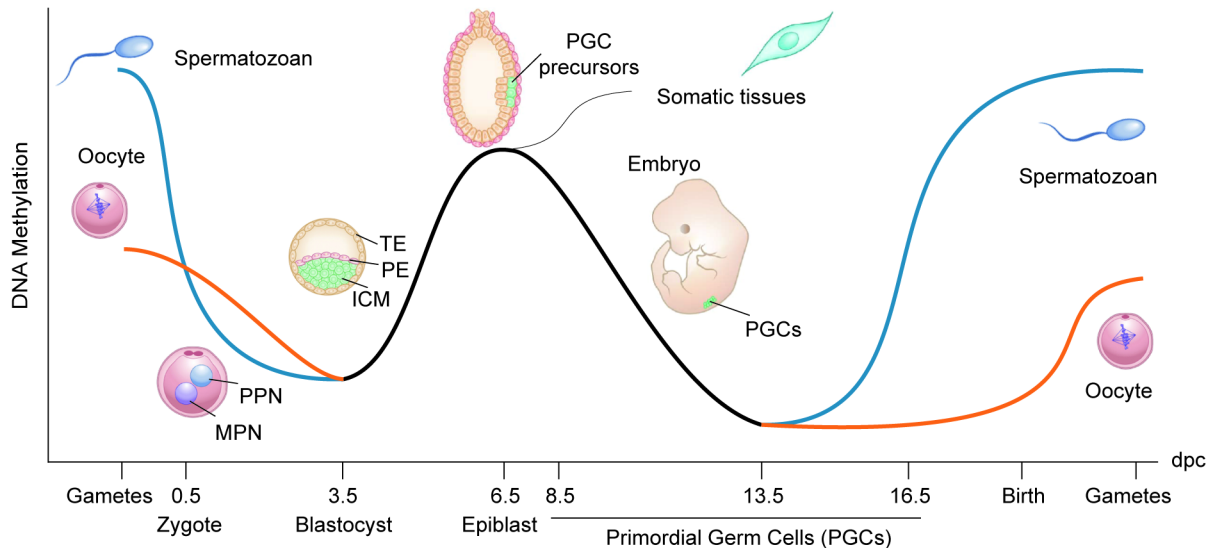
Fig. 2.4: DNA methylation dynamics in the mammalian lifecycle. Two major waves of DNA demethylation are observed in the mouse germline. The first, in the early embryo until the blastocyst stage (E3.5), affects both the paternal (blue) and maternal (orange) genome. Methylation is lost again in the primordial germ cells (PGCs) between E6.5 and E13.5 as they emerge from the epiblast. dpc, days post coitum; PPN, paternal pronucleus; MPN, maternal pronucleus, TE, trophoectoderm; PE, primitive endoderm; ICM, inner cell mass; PGC, primordial germ cell. Inspired by (Lee et al. 2014; Seisenberger et al. 2013) with illustrations from (Cantone and Fisher 2013).

different for male and female germ cells, with sperm showing a highly methylated genome (~85% CpG methylation), while oocytes are only moderately methylated (~ 30% CpG methylation) (Seisenberger et al. 2013; Smallwood et al. 2011; Smith et al. 2012).

DNA demethylation in early embryos: The newly established DNA methylation patterns in sperm and oocyte are reprogrammed again in the zygote shortly after fertilization. This second wave of global DNA demethylation is, however, not as widespread as in PGCs and at some regions, including imprinted loci, some retrotransposons and centromeres, DNA methylation patterns are retained. Here DNA demethylation includes genomic regions important for maintaining pluripotency, which are silenced in germ cells but needed in the early stages of development. Paternal and maternal genomes follow different DNA demethylation kinetics; whereas in the paternal pronucleus the methylation signal is lost rapidly after fertilization, the maternal genome undergoes a more gradual loss of DNA methylation (Santos et al. 2002; Santos and Dean 2004; Wossidlo et al. 2010) (mechanisms discussed in chapter 2.5). New methylation marks are acquired during and following implantation of the embryo when there is extensive *de novo* DNA methylation covering up to roughly 70% of all CpGs (Lepikhov et al. 2010; Morgan et al. 2005).

For either of the DNA demethylation events, the biological reasons remain incompletely understood but for both situations the reprogramming assures that epigenetic traits from previous generations are not carried over onto newly developing organisms. An erasure of epigenetic memory may also increase epigenetic plasticity to facilitate the major changes in transcriptional programs that are associated with the reestablishment of developmental potency (Seisenberger et al. 2013). The extensive reprogramming could also serve as an opportunity to correct accumulated epimutations at the generational boundary and hence ensure that future phenotypes are not affected by events in the past. In the context of these reprogramming processes, recent evidence suggested that both passive

20

and active DNA demethylation mechanisms are involved and different pathways may operate sequentially or in parallel in order to obtain such a widespread but controlled loss of DNA methylation (Guibert et al. 2012; Kawasaki et al. 2014; Seisenberger et al. 2012; Smith et al. 2012; Smith et al. 2014) (see chapter 2.5).

DNA demethylation in somatic cells: Dynamic epigenetic changes can also be observed post developmentally in somatic cells at specific loci, mostly upon nuclear- or hormone-receptor activation of gene promoters (Jost et al. 1991; Kangaspeska et al. 2008; M. S. Kim et al. 2009b; Kress et al. 2006; Lucarelli et al. 2001; Metivier et al. 2008; Saluz et al. 1986; Thomassin et al. 2001; Wilks et al. 1982). In contrast to global epigenetic reprogramming that occurs during early development, these rapid site-specific DNA demethylation events do not occur to erase parental marks but rather function in the dynamic regulation of gene transcription to allow rapid response to specific stimuli. Here, the DNA demethylation process occurs independent of cell division and is believed to be an active process. Considering this, the CpG methylation levels at such sites appear to reflect rather a steady-state than a stable condition, a dynamic equilibrium that can shift depending on the requirements of the cell. The underlying mechanisms of active DNA demethylation are still a matter of debate but major advances have been made in this regard recently and will be discussed in chapter 2.5.

2.4.3   Breaking DNA Methylation Patterns

The epigenetic marks that are acquired during development and differentiation restrict the cell's lineage potential and they become increasingly committed. This process is thought to be unidirectional in mammals, meaning that differentiated cells generally do not switch fates and epigenetic marks are stably passed on during mitosis, restricting cellular plasticity of the following cell generation. However, the remarkable epigenetic stability of differentiated cells *in vivo* can be circumvented and reversed *in vitro* by nuclear transfer (SCNT), cell fusion and induced pluripotent stem cell (iPSC) technology (Jullien et al. 2011; J. Xu et al. 2015; Yamanaka and Blau 2010). These approaches are capable of breaking the stable epigenetic state of differentiated cells by simulating the physiological processes of early development, leading to altered patterns of gene expression and to an induction of pluripotency. Apparently, it is sufficient to provide 4 key pluripotency factors Oct4, Sox2, Klf4 and Myc (OSKM) to reprogram a cell towards an induced pluripotent stem cell (iPSC) (Takahashi et al. 2007). Since the initial description of iPSCs, many cell types have been reprogrammed using a variety of reprogramming factors (J. Xu et al. 2015). This indicates that cells, no matter how committed, retain a certain degree of plasticity and established epigenetic barriers in terminally differentiated cells can be overcome. The fact that epigenetic reprogramming can be induced principally in any cell type indicates that potent processes capable of actively modulating the epigenome exist, particularly in early development and are strongly associated with pluripotency. These advances in epigenetic reprogramming provide a valuable opportunity to generate pluripotent stem cells from adult cells of the same individual that can be suitable for various applications.

So far the understanding of the molecular mechanisms underlying nuclear reprogramming are incomplete and it remains unclear to what degree it is a deterministic or stochastic process (Jullien et al. 2011). Although some parts of the epigenome can be deterministically and efficiently reprogrammed under the proposed conditions, other regions seem resistant to these programs and the epigenetic stability is sustained throughout the process. To break such a blockage and allow full reprogramming, stochastic processes may be required in addition that depend on random access of the involved factors to their targets spread throughout the genome. If and how DNA demethylation of target CGIs in promoter regions is mediated to reactivate gene expression is also still unclear, but evidence points towards an active DNA demethylation pathway (Simonsson and Gurdon 2004) and DNA demethylation seems to be a limiting step in reprogramming towards iPS cells (Mikkelsen et al. 2008).

2.5   DNA De-Methylation

The most recent advances in DNA methylation dynamics suggest that different mechanisms operate sequentially or in parallel to achieve genome-wide as well as locus-specific DNA demethylation. In both genome-wide reprogramming events it is widely accepted that the bulk of the genomic DNA methylation is erased passively (Kohli and Zhang 2013; Lee et al. 2014; H. Wu and Zhang 2014). Active mechanisms however may still operate as rapid inducers of DNA demethylation at specific loci and the current evidence points towards a prominent involvement of DNA repair pathways like NER or the BER pathway in these processes. Several lines of experimentation revealed a generation and repair of DNA strand breaks in the course of DNA demethylation (Barreto et al. 2007; Hajkova et al. 2010; Kress et al. 2006; Le May et al. 2012; Wossidlo et al. 2010), indicating a strand incision process potentially initiated by a DNA glycosylase or a repair associated endonuclease as discussed in detail in chapter 2.2.1. It is thus not surprising that proposed DNA demethylation pathways often, but not always, involve DNA repair. In the following section an overview of putative DNA demethylation scenarios will be given, ranging from very early observations to the most recent and promising pathways still being discussed today.

2.5.1   Scenarios of Active DNA Demethylation

Direct Reversal of DNA Methylation

Probably the most direct way to achieve DNA demethylation is to directly remove the methyl group by an enzymatic activity capable of breaking the rather strong C-C bond (Fig. 2.5 A). This energetically unfavorable reaction involving no cofactors other than water was proposed to be carried out by MBD2, a methyl CpG-binding protein (Bhattacharya et al. 1999). So far, however, this activity could not be reproduced and the results have been contested (Bird 2002). Furthermore, MBD2 knockout mice are viable and fertile and show no abnormalities in the pattern of genomic DNA methylation (Hendrich et al. 2001). Another enzymatic activity potentially capable of directly reverting DNA methylation is the

repair machinery for alkylation damages. Reversal of alkylation damage (1-methyladenine, 1mA and 3-methylcytosine, 3mC) in mammals is achieved by the AlkB homologs ABH2 and ABH3 employing an oxidative mechanism, where the methyl group is oxidized and subsequently released from the base as free formaldehyde (Duncan et al. 2002) (Fig. 2.5 B). Although the C-C bond in 5mC is energetically more difficult to break than the C-N bond in 1mA and 3mC, it still remains possible that a novel type of oxidase could directly demethylate 5mC by an oxidative mechanism. Recent studies identified structurally related dioxygenases, the TET proteins, which can directly act on 5mC. They however do not revert the DNA methylation directly but oxidize the methyl group to generate new 5mC derivatives (see chapter 2.5.2).

Direct 5mC Removal by DNA Glycosylases

More than 20 years ago, an enzymatic activity was discovered in extract of chicken embryos that was capable of excising 5mC from a DNA substrate (Jost 1993) (Fig. 2.5 C). At around the same time, the enzymatic removal of 5mC in HeLa cell nuclear extracts was reported and proposed to be mediated by a DNA glycosylase (Vairapandi and Duker 1993). A 5mC demethylation activity was also found in differentiating mouse myoblasts between the 3rd and 5th day of differentiation (Jost and Jost 1994) and an active, DNA glycosylase-initiated, demethylation mechanism was suggested (Jost et al. 1995). The 5mC DNA glycosylase activity in developing chicken embryos was then attributed to TDG and recombinant protein showed 5mC glycosylase activity *in vitro* (B. Zhu et al. 2000b). The same activity was also shown to be present in recombinant human MBD4 glycosylase (B. Zhu et al. 2000a). However, recombinant TDG and MBD4 show 30-40 fold lower activity on G·5mC than on G/T mismatch substrates *in vitro* (B. Zhu et al. 2000b; B. Zhu et al. 2000a) and this activity still remains to be reproduced independently. It is thus questionable whether TDG and MBD4 are reliable and efficient 5mC DNA glycosylases. It could recently, however, be shown that a fusion of TDG with a sequence-specific DNA binding domain can induce DNA demethylation and transcription of targeted genes (Gregory et al. 2012), though it is not clear if TDG was directly responsible for 5mC removal or if other factors were involved as well. Still, it is an interesting indication that targeting of DNA glycosylases to the right places might facilitate DNA demethylation. Knockout of the two DNA glycosylases MBD4 and TDG gave somewhat of a variable outcome. MBD4 knockout mice are viable and fertile and do not show aberrant DNA methylation but rather increased mutability (Millar et al. 2002). In agreement with that, global DNA demethylation of the paternal pronucleus in zygotes does not seem to rely specifically on MBD4 (Santos and Dean 2004). In contrast to MBD4, a TDG knockout in mouse causes embryonic lethality around E11.5, likely owing to developmental defects caused by epigenetic aberrations as discussed in chapter 2.2.2. In addition, there is indication that TDG plays a role in demethylation of imprinted genes in PGCs (Cortellino et al. 2011). Apparently, the catalytic functionality of TDG is required for its role in DNA demethylation as the knock-in of the catalytic mutant (N151A) has the same severe phenotype (Cortellino et al. 2011). Despite this strong experimental evidence, it remains doubtful that the role of TDG is to directly act as a DNA demethylase due to a lack of biochemical evidence supporting such a glycosylase activity. The existence of a DNA glycosylase initiated DNA demethylation mechanism is, however, well accepted and supported by strong genetic and

biochemical evidence in plants. In *Arabidopsis thaliana* active DNA demethylation is achieved by BER of 5mC initiated by the DNA glycosylases ROS1 and DEMETER (Choi et al. 2002; Gong et al. 2002). In mammals, the mechanism appears to be more sophisticated and relies on additional 5mC modifiers that render the base more susceptible to DNA damage repair and thereby triggering DNA demethylation.

Deamination-coupled Removal of 5mC by Base Excision Repair

An alternative mechanism to direct 5mC excision by DNA glycosylases is through deamination of 5mC to T, which generates a G/T mismatch (Fig. 2.5 D). In this case, TDG or MBD4 might again contribute to the DNA demethylation process. Both, cytidine deaminases and DNMTs have been proposed to catalyze the deamination step in this process. AID and apolipoprotein B mRNA editing enzyme, catalytic polypeptide 1 (APOBEC1) were shown to have weak 5mC deaminase activity *in vitro*, resulting in G/T mismatches, though these enzymes were shown to have a strong preference for single-stranded DNA (Morgan et al. 2004). More recent evidence, however, contested the deamination capacity of AID and APOBEC on 5mC and both enzymes showed a clear preference for unmodified C *in vitro* (Larijani et al. 2005; Nabel et al. 2012; Rangam et al. 2012). Still, studies in Zebrafish embryos suggested a potential coupling of a deaminase with a DNA glycosylase in the context of DNA demethylation (Rai et al. 2008); overexpression of AID and MBD4 together triggered demethylation of the bulk genome and injected methylated DNA fragments whereas overexpression of either protein alone had no effect. Thus, AID and MBD4 were suggested to be functionally coupled and their overexpression can demethylate DNA via a detectable G/T intermediate (Rai et al. 2008).

It also appears noteworthy that the genes encoding AID and APOBEC1 are both located within a cluster of pluripotency genes including NANOG and STELLA and are expressed in mouse oocytes, embryonic germ cells and ESCs, in line with a function associated with a pluripotent cell state (Morgan et al. 2004). Consistently, AID was shown to contribute to genome-wide DNA demethylation in mouse PGCs, but DNA methylation levels are also drastically reduced to below 20% by E13.5 in the absence of AID (Popp et al. 2010) implicating a prominent involvement of other pathways as well. This is supported by studies showing that AID knockout mice do not show major developmental defects and are viable and fertile (Muramatsu et al. 2000); APOBEC1 knockout mice are also viable and fertile (Morrison et al. 1996). The deaminases might however have a functional overlap and only a double knockout could further elucidate the importance of a cytidine deaminase in active DNA demethylation and epigenetic reprogramming. In addition, deamination of a symmetrical CpG dinucleotide (methylated or not) could give rise to a double mismatch and biochemical evidence that either of TDG or MBD4 possesses such a processing activity is lacking so far. Most recent findings suggested that global DNA demethylation in PGCs occurs in two distinct phases (Seisenberger et al. 2012), where the first and more extensive one is believed to occur mostly passive by inactivation of the DNA methylation machinery. Specific sequences escaped this process until the beginning of the second wave of DNA demethylation that appeared to occur more targeted, and AID could possibly have a role in this second phase of active DNA demethylation. Additional evidence for a putative involvement of

AID in active DNA demethylation came from reprogramming studies, where somatic cells were fused to ESCs to induce nuclear reprogramming and restore pluripotency (Bhutani et al. 2009). In these non-dividing heterokaryons, AID contributed to locus-specific active DNA demethylation of the OCT4 and NANOG promoter and was required for epigenetic reprogramming of somatic cell nuclei (Bhutani et al. 2009). AID also appeared to stabilize reprogramming of somatic cells using the iPSC approach by aiding DNA demethylation of pluripotency genes (Kumar et al. 2013). Interestingly, it is now also hypothesized that AID-mediated DNA demethylation is not necessarily induced by the deamination of 5mC but rather deamination of C to U in methylated regions, followed by subsequent LP-BER initiated by the Uracil-DNA glycosylase UNG2 instead of TDG (Franchini et al. 2014; Santos et al. 2013; D. Wu et al. 2014). This pathway could allow removal of several methylated CpG sites in close proximity using only a single repair event and thereby also reduce the genotoxic potential. Another non-enzymatic factor, Gadd45a, was suggested to play a role and serve as a scaffold to couple AID and MBD4 (Rai et al. 2008), AID and TDG (Cortellino et al. 2011) or assist in a NER (see below) dependent DNA demethylation pathway (Barreto et al. 2007). However, a functional role of Gadd45a in DNA demethylation could not be substantiated upon overexpression in human cells (Jin et al. 2008) and global as well as locus-specific DNA hypermethylation was not observed in Gadd45a-deficient mice (Engel et al. 2009).

Another DNA demethylation process coupling deamination with base excision has been proposed for DNMT3A/B and TDG. Initial evidence that methyltransferases could have deamination activity was found in bacteria, where they were shown to actively deaminate C and 5mC under suboptimal reaction conditions, i.e. when the concentration of their essential cofactor SAM is too low (J. C. Shen et al. 1992; Yebra and Bhagwat 1995). The mammalian methyltransferases DNMT3A/B, which commonly catalyze *de novo* DNA methylation, have also been shown to deaminate C and 5mC to U and T, respectively, in the absence of SAM (Metivier et al. 2008). Consistent with a functional cooperation, DNMT3A or DNMT3B were also shown to interact physically with TDG and stimulate its enzymatic activity (Boland and Christman 2008; Y. Q. Li et al. 2007). Furthermore, the recruitment of DNMT3A/B together with TDG and other BER factors to transcriptionally stimulated gene promoters coincided with cyclical, replication-independent, DNA methylation and demethylation processes at these sites (e.g. pS2 gene) (Kangaspeska et al. 2008; Metivier et al. 2008). This indicates a potential functional liaison of the deamination activity of DNMTs and the G/T, G/U processing activity of TDG to achieve locus-specific DNA demethylation. However, this mechanism would require that SAM levels also rapidly cycle *in vivo* to achieve 5mC deamination by DNMT3A/B. It is difficult to imagine that this is possible without biological consequences, given that SAM is also involved in many other essential biochemical processes. Alternatively, the TDG-DNMT3A/B interaction may simply ensure recruitment and targeting of the glycosylase to methylated DNA, where it exhibits a protective function by preventing C to T mutations caused by spontaneous deamination of 5mC (Boland and Christman 2008; Y. Q. Li et al. 2007). The incorporated C could then quickly be re-methylated by the DNMT3s to maintain the methylated state. This would also enable and facilitate a dynamic regulation of methylation states, where DNA demethylation and re-methylation occur in a cyclical manner.

Nucleotide Excision Repair initiated 5mC Removal

Another DNA repair pathway implicated in active DNA demethylation is NER that could erase 5mC by repairing short genomic regions containing methylated Cs (Fig. 2.5 E). In HeLa cells, NER factors have been shown to sequentially assemble at gene promoters upon gene activation even in absence of exogenous genotoxic attack (Le May et al. 2010). A differential role of the NER machinery present at gene promoters and at more distal gene regions has been proposed, where the former is functionally coupled to transcription and proper DNA demethylation. Consistently, it could be shown that transcriptional activation of RARβ2 leads to XPG-induced DNA breaks and DNA demethylation at the promoter region, suggesting an involvement of NER (Le May et al. 2012). NER was previously associated with active DNA demethylation in combination with Gadd45a as mentioned before. Gadd45a was identified in a screen for cDNAs, which activate expression of a methylation-silenced reporter gene, and proposed to directly interact with the NER endonuclease XPG (Barreto et al. 2007). Transfection of Gadd45a in cultured cells led to DNA demethylation by recruiting the DNA repair machinery to specific sites and a reduction in global DNA methylation, whereas knockdown of Gadd45a as well as XPG led to a global increase of 5mC (Barreto et al. 2007). These results were contested and a functional role of Gadd45a in DNA demethylation could not be substantiated upon overexpression in human cells as mentioned above (Jin et al. 2008). However, it has been shown that
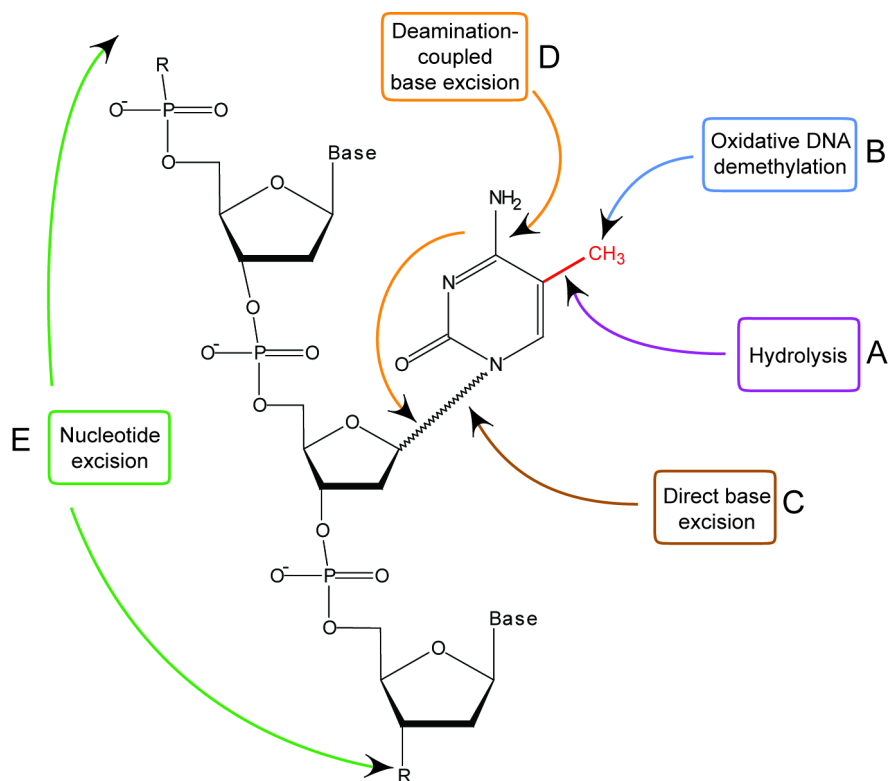
Fig. 2.5: Illustration of possible active DNA demethylation scenarios. (A) Hydrolytic release of the methyl group as methanol. (B) Direct reversal of methylation using an oxidative mechanism. (C) Release of 5mC by a DNA glycosylase, the predominant mechanism in plants. (D) 5mC deamination followed by G/T mismatch repair. (E) Excision of multiple nucleotides using NER. Details of these mechanisms can be found in the main text. Inspired by (J. K. Zhu 2009)

Gadd45a is associated with active rDNA repeats and knockdown of Gadd45a or components of the NER machinery led to increased promoter methylation (Schmitz et al. 2009).

2.5.2 TET-initiated DNA Demethylation

The search for a biological activity that could potentially trigger DNA demethylation by modifying 5mC was motivated by two pathways involved in oxidative modification of T. The first involves biosynthesis of the „Base J" in the genome of the parasite *Trypanosoma brucei* and the other the conversion of free T to U in the pyrimidine salvage pathway. „Base J", the trivial name for β-D- glucosyl-hydroxymethyluracil, is a modified T, functions in transcriptional regulation and is generated in a two-step process. A T in the context of genomic DNA is first oxidized to 5hmU by the J-binding proteins (JBP) 1 or 2; the addition of a glucose to 5hmU by a glycosyltransferase then completes biosynthesis of Base J (Borst and Sabatini 2008). JBP1/2 are members of Fe(II)/α-ketoglutarate (α-KG)-dependent dioxygenase family of enzymes. Another member of this family, Thymine-7-hydroxylase (THase), also acts on T to generate 5hmU but rather on the free base and not in the context of DNA. 5hmU is then iteratively further oxidized to 5-formyluracil (5fU) and 5-carboxyluracil (5caU) (Liu et al. 1973). The isoorotate decarboxylase finally completes the T to U conversion through decarboxylation of 5caU (Smiley et al. 2005).

A large bioinformatics screen for mammalian paralogues of JBP1/2 finally led to the discovery of the TET family of proteins (Tahiliani et al. 2009). The family consists of 3 members, TET1-3, sharing a conserved catalytic core comprised of a cysteine-rich region followed by a double-stranded β helix (DSβH) region (Tahiliani et al. 2009). TET1 and TET3 contain an N-terminal CXXC domain, enabling DNA binding and specific recognition of CpG dinucleotides. TET2 was separated from its CXXC domain during evolution (Ko et al. 2013). The IDAX protein that was originally encoded within an ancestral TET2 gene was then shown to physically and functionally interact with TET2 (Ko et al. 2013). The three TET proteins display different expression patterns during development and across tissues (Ito et al. 2010). TET1 and TET2 are highly expressed in mouse ESCs and both are present in PGCs albeit not at the same level (Hackett et al. 2013; Wossidlo et al. 2011; Yamaguchi et al. 2012). TET3 appears neither in ESCs nor PGCs but is the only TET enzyme present in mouse oocytes and one-cell zygotes (Iqbal et al. 2011; Wossidlo et al. 2011). Moreover, TET3 but also TET2 are more broadly expressed in various tissues compared to TET1 (Ito et al. 2010). In contrast to the above described Fe(II)/α-KG-dependent dioxygenase family members, the TET proteins were suggested to act on 5mC rather than T bases in DNA. Indeed, it was found that overexpression of TET1 led to a reduction of 5mC in genomic DNA and purified TET enzymes converted 5mC in oligonucleotides to 5hmC through oxidation (Ito et al. 2010; Tahiliani et al. 2009). The similarity between oxidation of T by the THase and oxidation of 5mC has raised the question if the TET proteins could also mediate iterative oxidation of 5mC beyond 5hmC. Although a first attempt to verify the presence of these 5mC oxidative derivatives in genomic DNA was unsuccessful (Globisch et al. 2011), they could eventually be detected in ESCs under physiological conditions (He et al. 2011; Ito et al. 2011; Pfaffeneder et al. 2011) and the generation of 5fC and 5caC by TET was shown *in vitro* (He et al. 2011; Ito et al. 2011).

Recent structural studies provided first insight into the putative mechanism of 5mC oxidation. Seemingly, the enzyme flips the 5mC out of the DNA double helix into its catalytic pocket where it is stabilized by hydrogen bonds. Importantly, the cysteine-rich DNA binding domain of the human TET2 catalytic domain also played a role in stabilizing the DSβH, forming a compact structure (Hashimoto et al. 2014; Hu et al. 2013). It is, however, still unclear if the TET enzymes operate in a continuous processive or a distributive way for the three successive oxidation steps. Notably, 5hmC, 5fC and 5caC are chemically distinct cytosine modifications that pair normally with G, have no mutagenic potential and are found at different levels in the mammalian genome (Inoue et al. 2011; Ito et al. 2011; Pfaffeneder et al. 2011). Although little is known about the biological implications of these modifications, these oxidized 5mC derivatives extend the previously accepted bimodal state of cytosine modifications to five distinct states in the mammalian genome (H. Wu and Zhang 2014). Oxidation of the C5-substituent leads to altered steric as well as electron density properties, opening up new potential ways of TET-initiated DNA demethylation. Three main possibilities for DNA demethylation following 5mC oxidation were suggested and will be discussed in detail below; (i) passive dilution of 5mC oxidized derivatives, (ii) direct removal of the oxidized C5-substituent and (iii) excision of the modified nucleotide mediated by DNA repair (Fig. 2.4).

Regulation of TET-initiated DNA demethylation could be provided on various levels, that remain incompletely understood. For example, little is known about the substrate preference of the TET proteins; interestingly, it appears that the rate of initial oxidation from 5mC to 5hmC is significantly higher than that of oxidizing 5hmC or 5fC (Hashimoto et al. 2014; Ito et al. 2011), indicating some discrimination between oxidative states. Notably, the methyl group of 5mC is not involved in the TET2-DNA interaction and it remains to be clarified if TET proteins have distinct affinity for any of the C derivatives (Hu et al. 2013). For the catalysis of 5mC oxidation, TET enzymes require several supplements including oxygen, Fe(II) and α-KG and regulation of TET activity might also occur at the level of metabolites and cofactors. α-KG is produced primarily by the isocitrate dehydrogenases (IDHs) in the tricarboxylic acid (TCA) cycle. Mutations in IDH can result in accumulation of 2-hydroxyglutarate (2-HG) at the expense of α-KG, which competitively inhibits the TET enzymes and leads to DNA hypermethylation (Figueroa et al. 2010; Kaelin and McKnight 2013). Furthermore, ascorbic acid was shown to stimulate the catalytic activity of TET proteins by directly interacting with the catalytic domain and facilitating conformational changes or by recycling the cofactor Fe(II) (Yin et al. 2013). In mouse ESCs, providing ascorbic acid greatly enhanced TET activity leading to significantly higher levels of 5mC oxidation and a blastocyst like methylome (Blaschke et al. 2013). Genomic targeting, probably mediated by the N-terminal domain (including the CXXC domain) of the protein, might also play an essential role in the regulation of TET activity. Genome-wide analysis revealed that TET1 preferentially localized to unmethylated CpG rich regions in ESCs (K. Williams et al. 2011; H. Wu et al. 2011) and also IDAX (and TET2 alongside) was shown to be enriched at unmethylated CpG sequences (Ko et al. 2013). The CXXC domains of the TET proteins, however, do not restrict the protein to localize to unmethylated regions (Y. Xu et al. 2011; Y. Xu et al. 2012; H. Zhang et al. 2010a) and may thus have increased flexibility in targeting the TET proteins to genomic loci to be regulated. Given the differential expression patterns of the TET proteins during development and across tissues, TET expression might also be regulated posttranscriptionally by microRNAs.

Indeed, the microRNA miR-22 was recently shown to negatively regulate TET protein levels in breast cancer development and hematopoietic stem cell transformation (S. J. Song et al. 2013b; S. J. Song et al. 2013c). Another microRNA miR-29 also appeared to directly regulate TET1-3 mRNA levels, thus indirectly affecting DNA methylation levels (P. Zhang et al. 2013).

TET-mediated DNA Demethylation Pathways

Passive dilution: Symmetrically modified CpG dinucleotides are generally maintained in the fully methylated state by the maintenance DNA methylation system (DNMT1/UHRF1), thus preserving DNA methylation patterns through cell divisions as discussed in chapter 2.3.2. A process similar to the passive dilution of 5mC might also be employed or even facilitated by TET-mediated oxidative modification. *In vitro* evidence supported such a scenario and UHRF1 was shown to bind hemi-5hmC ten-fold less efficiently than hemi-5mC modified DNA (Hashimoto et al. 2012b). In addition, the activity of recombinant DNMT1 was up to 50-fold reduced at these hemi-modified sites (Hashimoto et al. 2012b). Whether hemi-modified CpGs carrying a 5fC or 5caC also impair DNMT1 activity is currently not known, but the low abundance of these modifications indicates that they are rather short-lived and passive dilution might only play a minor role in the removal of these modifications. Interestingly, *in vitro* methylase activity of DNMT3A/B on hemi-modified 5mC and hemi-modified 5hmC DNA was comparable (Hashimoto et al. 2012b) and DNMT3A/B could potentially take over a maintenance function in cells with high TET activity to prevent passive dilution of oxidized 5mC derivatives. In contrast to non-enzymatically induced passive loss of 5mC, preceding oxidation of 5mC followed by passive dilution might not be regarded as genuinely passive but rather active-passive (Kohli and Zhang 2013). This kind of passive dilution may be effective even in the presence of maintenance DNA methylation and provides an alternative way to reduce DNA methylation levels.

Direct removal of C5-substituents: Whereas for 5mC the direct removal of the methyl group is energetically highly unfavorable, oxidation of the C5-substituent may render such a reaction more feasible. In bacteria for example, 5hmC is directly converted to C by bacterial methyltransferases (M.HpaII, M.SssI, M.AluI) (Liutkeviciute et al. 2009). Such an activity was also attributed to all three mammalian DNMTs, observed to directly convert 5hmC to unmodified C *in vitro* (C. C. Chen et al. 2012). Apparently the function of DNMT3A can be regulated by the redox status; while reducing conditions favor the methyltransferase activity, the dehydroxymethylation is active under oxidizing conditions (C. C. Chen et al. 2012). The validity and physiological relevance of such a mechanism is currently not clear but it is an interesting observation. Analogous to the thymidine salvage pathway, the unmodified state might be obtained through deformylation of 5fC and decarboxylation of 5caC. There is in fact experimental evidence for a decarboxylase activity in ESC lysates (Schiesser et al. 2012) but the protein responsible for 5caC decarboxylation needs yet to be identified.
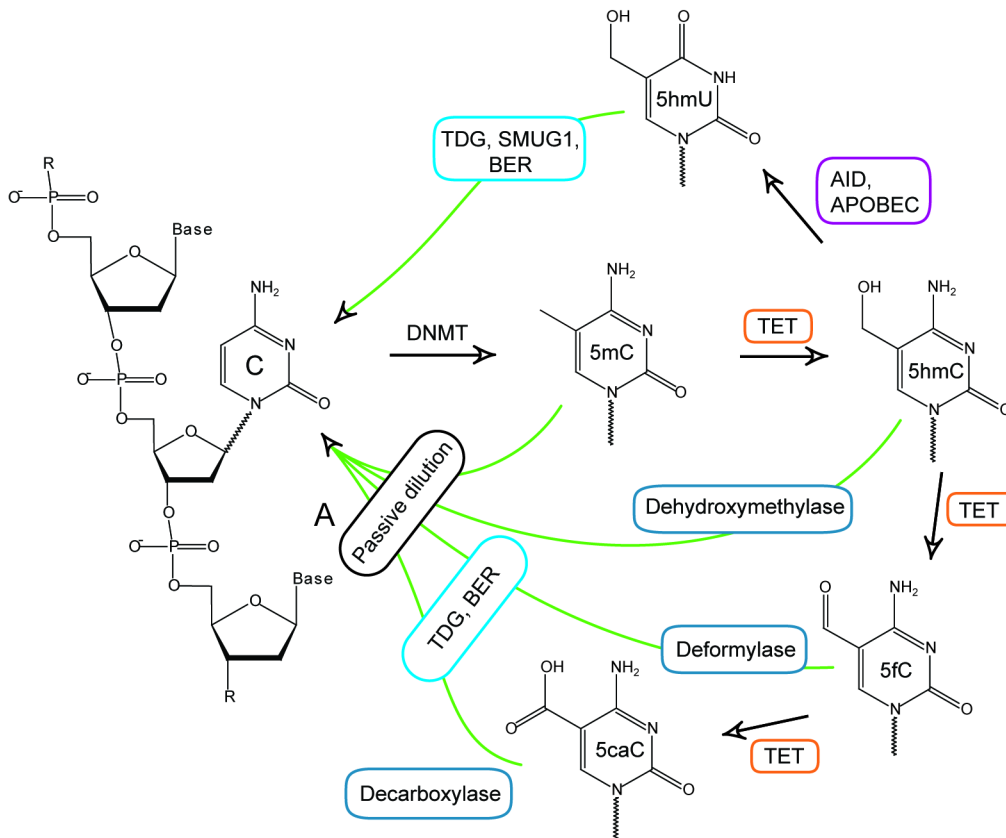
Fig. 2.6: TET-initiated DNA demethylation. Three main possibilities for DNA demethylation following 5mC oxidation were suggested; (i) passive dilution of 5mC oxidized derivatives (ii) direct removal of the oxidized C5-substituent by either dehydroxylation, deformylation or decarboxylation and (iii) excision of the modified nucleotide mediated by DNA repair. Another possibility is 5mC oxidation followed by deamination and DNA repair. Details in the main text.

Excision of the modified nucleotide: A likely and widely accepted scenario is the engagement of DNA repair downstream of 5mC oxidation to revert the base modification. An interesting side note here is that the plant 5mC glycosylases; DME and ROS1, also have a significant 5hmC excision activity *in vitro*, although 5hmC does not appear to be present in *Arabidopsis* (Jang et al. 2014). Early studies showing *in vitro* evidence of a mammalian 5hmC glycosylase activity in calf thymus extract (Cannon et al. 1988) were never reproduced substantiating that DNA glycosylase mediated DNA demethylation pathways differ between plants and animals. After the discovery of TET proteins and the oxidized C derivatives, the role of TDG in DNA demethylation was revisited and it could indeed be shown that 5fC as well as 5caC are specifically recognized and excised by TDG but not by UNG2, SMUG1 or MBD4 *in vivo* and *in vitro* (He et al. 2011; Maiti and Drohat 2011; L. Zhang et al. 2012). The molecular basis for this unusual activity of TDG, which otherwise prefers mismatched bases, lies in the stability of the N-glycosidic bond, which was previously shown to have a major effect on the catalytic efficiency of TDG (Bennett et al. 2006; Hardeland et al. 2000). Oxidation of 5mC weakens the N-glycosidic bond and stabilizes the transition state of the glycosylase reaction (R. T. Williams and Wang 2012) thus enabling efficient base excision. In addition, it could be shown that residues in the active site of TDG engaged in additional molecular interactions with 5fC and particularly 5caC, mediating efficient recognition of these base modifications (Hashimoto et al. 2012a; L. Zhang et al. 2012). Seemingly, 5fC can also alter the structure of the DNA helix, an attribute that is usually related to faulty base modifications and thereby facilitate initial base recognition and initiation of BER (Raiber et al. 2015).

Although there is yet no biochemical evidence that another DNA glycosylase acts on the TET-oxidized 5mC derivatives, a recent study showed that in the absence of TDG the NEIL glycosylases could partially reactivate reporter gene expression in ES cells (Muller et al. 2014). They hypothesized but did not show that the NEIL glycosylases can also initiate DNA demethylation via BER after TET-mediated C oxidation by direct excision of 5fC and 5caC.

5hmC could potentially also be deaminated to generate 5hmU, a base modification that is readily detected and excised by TDG as well as SMUG1 (Jacobs and Schar 2011), opening up the possibility of a coupled oxidation-deamination pathway with 5hmC as a key intermediate to remove 5mC. It could indeed be shown that AID/APOBEC deaminases specifically promoted 5hmC removal and did not have a significant effect on 5mC (Guo et al. 2011) and that expression of TET1 resulted in a significant amount of 5hmU in HEK293 cells (Guo et al. 2011). 5hmU, however, does not necessarily originate from deamination of 5hmC but might also be generated through oxidation of T by TET1 itself (Pfaffeneder et al. 2014) and could thus be an artefact or byproduct of TET1 overexpression. Furthermore, biochemical studies that evaluated the feasibility of AID/APOBEC mediated 5hmC deamination showed that this activity, if at all detectable, is extremely weak in several *in vitro* assays and in cells (Nabel et al. 2012; Rangam et al. 2012), arguing against a prominent role of such a pathway. Yet, 5mC deamination is still a feasible scenario for DNA demethylation as discussed above, but it more likely operates in parallel rather than in combination with the TET-mediated oxidation.

With the discovery of the TET proteins came along a re-evaluation of the role of Gadd45a in DNA demethylation (chapter 2.5.1). Gadd45a was shown to interact with TDG and overexpression together with TET led to DNA demethylation and activation of a reporter gene in HEK293T (Appendix III). In addition, knockout of Gadd45a/b led to hypermethylation of specific loci that also display 5fC enrichment in the absence of TDG (Appendix III). Another study suggested a direct interaction of Gadd45a with TDG in long non-coding RNA (lncRNA)-mediated DNA demethylation, also requiring the TET but not the NER proteins (Arab et al. 2014).

Finally, evidence starts to emerge for a direct interaction of TET proteins with TDG and components of the BER machinery and a DNA demethylation mechanism triggered by 5mC oxidation and subsequent processing by TDG and BER is now the first biologically and biochemically validated active DNA demethylation mechanism (Appendix I and III). This major breakthrough in DNA demethylation research has led to reevaluation of many biological processes that rely on DNA demethylation and it becomes more and more evident that TET but also TDG-initiated DNA repair contribute to active DNA demethylation (Kohli and Zhang 2013; Pastor et al. 2013; H. Wu and Zhang 2014).

3 Aim of the Thesis

TDG was initially discovered as a DNA glycosylase that excises T when mispaired with G. Recent compelling evidence, however, suggests that the essential role of this DNA repair enzyme is the control of DNA methylation (Cortazar et al. 2011; Cortellino et al. 2011; Raiber et al. 2012; L. Shen et al. 2013; C. X. Song et al. 2013a). Current models supported by both biochemical as well as genetic evidence implicate TDG-initiated BER in active DNA demethylation following the step-wise oxidation of 5mC to 5hmC, 5fC and 5caC by TET proteins (Inoue et al. 2011; Ito et al. 2011). 5fC and 5caC were shown to be efficiently excised from DNA by TDG (He et al. 2011; Maiti and Drohat 2011) and the phenotype of TDG deficient mice and cells is consistent with a dysregulation of DNA methylation patterning (Cortazar et al. 2011; Cortellino et al. 2011; Raiber et al. 2012; L. Shen et al. 2013; C. X. Song et al. 2013a). Although DNA demethylation through this pathway is a plausible and therefore widely accepted scenario, the actual evidence proofing the functionality of such a mechanism is scarce. A molecular and functional interaction between TET, TDG and BER has not been established and key mechanistic features and regulatory aspects of the complex multistep process have not been resolved.

The overall aim of my PhD thesis was to provide a proof of functionality of a TET, TDG and BER based DNA demethylation system. By reconstituting an operational TET, TDG and BER system with purified proteins *in vitro*, I aimed to biochemically investigate the process of active DNA demethylation. Specifically, the interplay of TET and TDG and functional interactions with additional factors such as Gadd45a and the BER proteins was to be elucidated and characterized. Aside, I aimed to establish novel tools and biochemical assays, facilitating the functional investigation of implicated regulatory factors such as posttranslational SUMO modification or RNA interactions.

A major part of my thesis project, thus, comprised the production and purification of highly active recombinant proteins. With the purified proteins it was then possible to study their enzymatic activities individually or in combination using biochemical assays to unravel the process of DNA demethylation. My work also comprised the establishment and successful application of protein expression systems facilitating the production of recombinant SUMO-modified proteins for mechanistic studies as well as of biochemical assays with RNA containing DNA demethylation substrates.

To this end, I – in collaboration with others – was able to extend previous biochemical findings and provide a biochemically validated concept of TET-initiated and DNA repair-mediated active DNA demethylation involving TDG, BER and in part Gadd45a. I fully reconstituted the DNA demethylation process for the first time and my results revealed a highly coordinated and sequential mechanism to achieve DNA demethylation at symmetrically modified CpGs.

4 Results

This chapter provides an overview and summary of the results obtained during my PhD thesis, which are attached as manuscripts in the appendix. Additional important results not contained in the manuscripts are presented here in chapter 4.4.

4.1 Biochemical Reconstitution of TET1-TDG-BER Dependent Active DNA Demethylation Reveals a Highly Coordinated Mechanism (Appendix I)

DNA methylation at the C5 position of Cs (5-methylcytosine, 5mC) is one of the best described epigenetic modifications and is found predominantly within CpG dinucleotides, affecting 60-90% of such sites (Bird 2002). Originally considered a static developmentally established and maintained epigenetic mark, recent advances have indicated that DNA methylation is subject to dynamic regulation through passive as well as active enzymatic processes. Mechanisms of DNA demethylation have long been under debate but recent findings have advanced our understanding of this process and some candidate pathways have received a lot of experimental support (H. Wu and Zhang 2014). A widely accepted model suggests an involvement of the TET family of dioxygenases and the DNA glycosylase TDG in the removal of DNA methylation (Kohli and Zhang 2013; Pastor et al. 2013). TET proteins catalyze the oxidation of 5mC to 5hmC, 5fC and 5caC (He et al. 2011; Inoue et al. 2011; Ito et al. 2011; Tahiliani et al. 2009). TDG, originally identified as a DNA glycosylase for the excision of T and U mispaired with guanine G, was then shown to recognize and excise the TET-mediated 5mC oxidation products 5fC and 5caC. These findings implicated an engagement of the BER pathway in the exchange of 5mC for an unmethylated C (He et al. 2011; Maiti and Drohat 2011). However, although plausible as a mechanism, there is yet only little evidence supporting a direct link of TET with BER and key features relating to the regulation, coordination and progression of oxidative DNA demethylation remain to be resolved. In this study, we set out to reconstitute the process of oxidative DNA demethylation with purified proteins *in vitro* and to address the biochemical features of this pathway.

The mechanism of TET- and TDG-dependent DNA demethylation implies a coupled action of both enzymes and we reasoned that they might physically interact to facilitate an efficient but coordinated removal of 5mC. To this end, we provide evidence of a direct interaction between the two proteins following three complementary approaches (Appendix I, Fig. 1). Results obtained from gel filtration (Appendix I, Fig. 1 A), yeast two-hybrid (Y2H) (Appendix I, Fig. 1 B) and co-affinity purification (Appendix I, Fig. 1 C) supported a direct TET-TDG interaction and led us to conclude that TET1 and TDG physically interact through domains located in the TET1 N-terminus ($TET1_N$) as well as the TET1 catalytic domain ($TET1_{CD}$). When measuring the catalytic activity of the complex formed between $TET1_{CD}$ and TDG by means of a standardized nicking assay (Hardeland et al. 2000), I detected efficient base release from oligonucleotide substrates containing 5mC or 5hmC modifications (Appendix I, Fig. 2 A). As both catalytic activities were required for the base release in these assays,

the results strongly indicate a concerted action of TET and TDG and, hence, the functionality of the TET and TDG complex. Evaluation of the enzymatic activity of the purified catalytic domain (TET1$_{CD}$) by itself (Appendix I, Fig. 2 B) and reconstitution with purified TET1$_{CD}$ and TDG (Appendix I, Fig. 2 C) then provided proof for a successful cooperation of both enzymes in the oxidation and excision of 5mC and 5hmC from DNA to generate an abasic-site DNA repair intermediate.

As TET1 and TDG were shown to physically interact, we reasoned that this interaction might also have functional consequences. I thus tested potential stimulatory effects on the partner's enzymatic activity using different approaches (Appendix I, Fig. 3). To measure the effect of TDG on the ability of TET1$_{CD}$ to produce 5mC derivatives, I treated DNA substrates containing 5mC with TET1$_{CD}$ by itself or in complex with a catalytically inactive TDG (TDGΔcat). The abundance of 5mC derivatives in the DNA was then measured by the base release assay using active TDG (Appendix I, Fig. 3 A) or by LC/MS/MS analysis (Appendix I, Fig. 3 B). Neither analysis resulted in a significant difference between TET1$_{CD}$ by itself and in complex with TDGΔcat, indicating the TDG interaction did not stimulate TET1 activity on either 5mC or 5hmC. Interestingly, the presence of TET1$_{CD}$ had a stabilizing effect on TDG activity and increased the efficiency of 5caC processing (Appendix I, Fig. 3 C). From these results, we conclude that physical interaction of TET1 and TDG does not affect their catalytic efficiency but has a stabilizing effect, particularly on TDG.

The initial steps of DNA demethylation, i.e. the release of the 5mC base, are suggested to be followed by BER to restore the DNA sequence with an unmodified C. To formally proof the validity of such a mechanism and to provide a tool to investigate its biochemical mode of action, I reconstituted the full DNA demethylation pathway with purified TET1$_{CD}$, TDG and BER proteins. Analysis of individual steps of the process (Appendix I, Fig. 4 A) as well as of the end product (Appendix I, Fig. 4 B) showed not only that it is possible to release 5mC in a TET-TDG dependent manner but also that an unmodified C can be correctly incorporated *in vitro*, thus reestablishing the unmethylated state via the oxidative DNA demethylation pathway.

It is known that DNA methylation is most often present at palindromic CpG dinucleotides, involving methylation of both strands. We therefore asked how the DNA demethylation process is coordinated in a context of symmetrical DNA methylation to avoid a deleterious formation of DSBs. The analysis of fully methylated CpGs processed by the TET1$_{CD}$/TDG complex showed that the 5mCs in both strands are equal targets for demethylation, indicating an absence of DNA strand preference or a randomly distributed but coordinated action of the TET1$_{CD}$/TDG complex (Appendix I, Fig. 5 A). To further examine the events at symmetrically modified CpGs, I separated the TDG from the TET1 activity and measured the kinetics of 5caC processing strand-specifically. As for the activity of the complex, there was no detectable strand specificity, but the processing of one strand apparently blocked the access to the other, indicating that TDG did not dissociate easily from the abasic site (Appendix I, Fig. 5 C). Complete reconstitution of DNA repair on a symmetrically modified substrate then showed that the unmethylated state was correctly reestablished on both strands (Appendix I, Fig. 5 D). Importantly, the repair process of symmetrically modified CpGs, that ultimately requires breaking of both strands, did not induce a detectable level of DSBs. Hence, full DNA demethylation and repair of the two strands

occurred in a processive but sequential manner, restoring the unmodified state and maintaining genome integrity during the DNA demethylation process.

Having evidence that repair at symmetrically modified CpGs occurs sequentially, I set out to investigate potential interferences between oxidation-induced DNA demethylation and repair of mismatched DNA. Spontaneous or induced deamination of 5mC on one strand while the other strand is being actively demethylated could generate a G·5caC modification next to a G/T mismatch within the same CpG dinucleotide. Interestingly, when offered both substrates at the same time, TDG almost exclusively processed 5caC with largely unaffected rate, leaving the G/T mismatch completely unrepaired (Appendix I, Fig. 6 B). Sequential repair in this situation is of biological relevance, as it occasionally promoted C to T mutations when BER of the 5caC·G incorporated two (or more) instead of just one nucleotide, and thereby generated an A opposite T in the opposite strand (Appendix I, Fig. 6 C). Such events may contribute to the depletion of CpG dinucleotides in the genome.

In conclusion, we provide a proof of concept for the mechanism of repair-mediated oxidative DNA demethylation (Appendix I, Fig. 6 D). We showed *in vitro* that TET1 does indeed associate with TDG to release 5mC from DNA and together with the BER machinery productively demethylates DNA via oxidative intermediates. With the gained knowledge, some pending mechanistic questions could be addressed. In particular, my results provide evidence for a highly coordinated and sequential mechanism to achieve DNA demethylation even at symmetrically methylated CpGs (Appendix I, Fig. 6 D middle path), which is important to avoid the formation of DNA DSBs and, hence, gross genomic instability (Appendix I, Fig. 6 D right path). The sequential repair of both lesions, however, turns into a disadvantage in situations of coincident deamination and oxidation at fully methylated CpGs. My results clearly showed that a strong preference for 5caC masks the nearby G/T mismatch to fix a C to T mutation and hence a loss of the CpG dinucleotide (Appendix I, Fig. 6 D left path).

Contribution: I designed and cloned the bacterial TET1 expression constructs as well as the TDG-GST construct. I performed the gel filtration and the co-affinity purifications that were used to study the interaction between TET1 and TDG. I purified TET1$_{CD}$, TDG as well as the TET1$_{CD}$/TDG complex that were used in the biochemical assays. I performed and established (where necessary) the biochemical assays (base release, oxidation, *in vitro* reconstitution). I wrote the manuscript and designed all the figures based on my own results.

4.2 Versatile Recombinant SUMOylation System for the Production of SUMO-Modified Protein (Appendix II)

Posttranslational modification by ubiquitin-like polypeptides, so-called UBLs, has been implicated in the regulation and coordination of a variety of vital biological processes, including DNA repair, cellular response to DNA replication stress, regulation of gene expression and epigenetic DNA and histone modifications (Cubenas-Potts and Matunis 2013; Flotho and Melchior 2013; Jackson and Durocher 2013; Zhao 2007). Also, in the context of DNA repair-mediated DNA demethylation, SUMOylation is likely to have regulatory functions, given its well-documented ability to modulated TDG function. The

SUMO peptides represent a prominent subfamily of the UBLs and exist in four different isoforms (SUMO1, SUMO2, SUMO3 and SUMO4) in mammalian cells, all of them showing high 3D-structural resemblance (Dohmen 2004; Gareau and Lima 2010; Geiss-Friedlander and Melchior 2007; Jackson and Durocher 2013). The mechanism of SUMO conjugation appears to be the same for all SUMO isoforms; they are covalently attached to their target proteins by a cascade of enzymatic reactions resembling ubiquitin conjugation (Hay 2013; Schwartz and Hochstrasser 2003). These involve a heterodimeric activating enzyme E1 (SAE1/SAE2), a single conjugating enzyme E2 (Ubc9) and, in some cases, an E3 protein ligase (Appendix I, Fig. 1 A). The number of proteins identified as SUMO targets has been increasing over recent years, mechanistic insight into the impact of SUMO interactions and modifications on the respective molecular processes, however, is only slowly emerging. A main reason for this is the difficulty to generate appreciable amounts of recombinant SUMOylated proteins for biochemical and structural investigation.

In this part of my PhD thesis, we set out to establish a versatile and potent SUMOylation system in *E.coli*, to facilitate the production of homogeneously SUMOylated recombinant proteins for subsequent biochemical and structural studies. We designed and experimentally validated two alternative two-component vector systems for simultaneous expression of the complete SUMOylation machinery alongside a target protein of interest. The first, termed pSUMO-based system, expresses all essential components for SUMO modification from a single plasmid (pSUMO1-3) (Appendix II, Fig. 1 B) and can be combined with commonly used bacterial expression vectors (Appendix II, Fig. 1 C). The second was named SUMO-E2-fusion system and is composed of a SUMO-activating vector (pSA1-3), encoding the SUMO polypeptide and SUMO-E1 (Appendix II, Fig. 3 A), and a SUMO-conjugating vector (pSC) (Appendix II, Fig. 3 C and E), containing a TARGET-Ubc9-GST fusion expression cassette. Both systems are suitable for in-cell as well as in-extract SUMOylation (Appendix II, Fig. 6), thus providing versatility in respect to the experimental setup and requirements of a target protein. In this publication, we evaluated the newly developed vector systems with a well-established and a putative SUMO target, the DNA BER proteins TDG and XRCC1 (Bruderer et al. 2011; Gocke et al. 2005; Jacobs and Schar 2011).

Our systems differ from previously established ones (Lens et al. 2011; Mencia and de Lorenzo 2004; O'Brien and DeLisa 2012; Uchimura et al. 2004b), in that they consist of human components only, provide two different tags for affinity purification and avoid a fusion of the two subunits of SUMO-activating enzyme E1 (SAE1 and SAE2), previously shown to reduce SUMO-E1 activity (Uchimura et al. 2004a)(Appendix II, Fig. 1 B). Evaluation of our pSUMO-based system consistently led to a faster appearance and higher yields of in-cell SUMOylated mouse TDG (mTDG) and human XRCC1 (hXRCC1) protein when compared to the system by Saitoh and colleagues (Uchimura et al. 2004a) (Appendix II, Fig. 1 D and E).

SUMOylation efficiency and specificity can be enhanced by SUMO-E3 ligases (O'Brien and DeLisa 2012) that promote structural proximity between the substrate and the SUMO-conjugating enzyme Ubc9. To avoid co-expression of respective SUMO-E3 ligases, their function can be mimicked by fusing Ubc9 directly to SUMO1 or target proteins of interest (Jakobs et al. 2007; E. T. Kim et al.

2009a). To provide such opportunity, we used a TARGET-Ubc9 fusion approach in our SUMO-E2-fusion system. The generation of SUMO-conjugated protein was efficiently obtained in-cell using this fusion approach with human (hTDG) as well as mouse TDG (mTDG) (Appendix II, Fig. 3 D, F, G) and appeared to be affected by the configuration of the TARGET-Ubc9 construct that was designed in two variants (Appendix II, Fig. 3 D compared to F).

In contrast to TDG, in-cell modification of the hXRCC1 SUMO-E2-fusion was not satisfactory due to inefficient protein expression. To circumvent such barriers, we developed a strategy where SUMO-modification can be obtained in crude *E.coli* extracts without prior enrichment of the SUMOylation and target factors (Appendix II, Fig. 4 A). Using this approach, fast and efficient SUMOylation of hXRCC1 was detected (Appendix II, Fig. 4 B and E) that turned out to completely rely on the Ubc9 fusion (Appendix II, Fig. 4 C). This indicated that the TARGET-Ubc9 fusion efficiently mimicked SUMO-E3 ligase functionality and enhanced SUMOylation of hXRCC1 in mixed lysates (Appendix II, Fig. 4 D). This in-extract approach can be employed using lysates from pSA or pSUMO expressing cells (Appendix II, Fig. 4 A); particularly pSUMO offers high combinatorial flexibility and can be combined with lysates from virtually any potential SUMO target (Appendix II, Fig. 6).

The generation of homogeneously SUMO-modified target proteins was finally achieved by purification and separation from unmodified isoforms by sequential GST and Ni-NTA affinity chromatography or vice versa. We could show that purification following both routes work in principle and result in a clear enrichment of SUMO-modified target protein (Appendix II, Fig. 2). Treatment of purified mTDG-SUMO1 with recombinant SUMO protease SenP2 then generated the unmodified isoform demonstrating that the detected high molecular bands are indeed SUMO1-modified mTDG protein (Appendix II, Fig. 5 A and B compare lanes 3 and 4). Moreover, we showed that the purified in-cell SUMOylated mTDG displays previously described modification-dependent enzymatic features (Hardeland et al. 2002), when we evaluated its capacity to release uracil from G/U mismatched DNA substrates in a base release assay (Appendix II, Fig. 5 C and D).

In conclusion, we developed tools and strategies to produce SUMO-modified proteins using versatile binary expression vector systems in protease-deficient *E.coli* (Appendix II, Fig. 6). They are designed to be applied for biochemical and structural studies of proteins and biological pathways potentially regulated by SUMOylation, i.e. the coordination of active repair-mediated DNA demethylation. Using this system I could already successfully SUMOylate TET1 (see chapter 4.4.1) corroborating a potential involvement of this regulatory pathway in oxidative DNA demethylation.

Contribution: I cloned cDNAs of target proteins into the pSC constructs that were assembled by David Schürmann. Together with David Schürmann, I designed and conducted in-cell as well as in-extract SUMOylation of TDG and XRCC1 and performed immunoblot analysis (Appendix II, Fig. 1 – 4). I purified in-cell SUMOylated mTDG and made the biochemical evaluation (Appendix II, Fig. 5). I prepared and designed all the figures and wrote the manuscript.

4.3 Gadd45a promotes DNA demethylation through TDG (Appendix III)

In vertebrates, cytosine methylation in CpG-rich regulatory gene promoters and enhancers inversely correlates with transcriptional activity of associated genes by mechanisms involving recruitment of transcriptional repressors and chromatin condensation. Reversal of DNA methylation is subject of ongoing research and a widely accepted model suggests that active DNA demethylation can be obtained by a concerted action of TET proteins and TDG-initiated BER (Kohli and Zhang 2013). Results obtained during my PhD project have provided a proof of the mechanistic concept of this pathway (Appendix I), it is however likely that auxiliary factors also contribute to DNA demethylation by modulating TET and TDG activities and coordinating these with BER processes. Gadd45 family proteins are multi-faceted nuclear factors that have been implicated in active DNA demethylation, potentially through XPG-dependent DNA repair (Barreto et al. 2007; Schmitz et al. 2009) or BER of AID-based deamination products (Cortellino et al. 2011; Rai et al. 2008). However, despite the connection of Gadd45 proteins with DNA demethylation, their role in the process has remained unsolved and controversial (Jin et al. 2008; Niehrs and Schafer 2012). This collaborative study initiated by Li Zheng and Guoliang Xu at the Chinese Academy of Science in Shanghai, aimed to investigate the role of Gadd45a in TET-TDG-dependent oxidative DNA demethylation and activation of methylation-silenced genes.

Candidate factors (Appendix III, Table S2) were first tested in combination with TET and TDG in a HEK293T cell-based firefly luciferase reporter assay for their capabilities to promote reactivation of a methylated reporter gene. Seemingly, the presence of Gadd45a stimulated the re-expression of the silenced reporter and co-expression of Gadd45a with Tet2CD and TDG increased the luciferase activity by 35-fold (Appendix III, Fig. 1 A). The effect was strongly dependent on the enzymatic activities of both Tet2CD and TDG and gene activation was lost upon co-expression with the catalytically inactive mutants (Appendix III, Fig. 1 A), suggesting that Gadd45a cooperated with TET and TDG in the activation of a methylation-silenced reporter gene.

To further evaluate the role of Gadd45a in TET and TDG-mediated DNA demethylation, the occurrence of genomic 5caC was monitored in HEK293T cells. Interestingly, a substantial reduction of the genomic 5caC level (30% reduction) was detected upon co-transfection of Gadd45a with TET2 (Appendix III, Fig. 2 A) and selective removal of 5fC and 5caC but not 5mC or 5hmC was measured (Appendix III, Fig. 2 B), suggesting that Gadd45a potentially cooperated with endogenous TDG in the removal of genomic 5fC and 5caC. Consistently, in TDG knockout HEK293T cells (TDG KO) (Appendix III, Fig. 4 A and B and Fig. S3), overexpression of Gadd45a had no effect on the TET2-generated genomic 5caC levels (Appendix III, Fig. 4 C). It thus appears that the role of Gadd45a in oxidative DNA demethylation relies on the presence of TDG and its enzymatic function.

To further elucidate a potential interplay between Gadd45a and TDG in DNA demethylation, their potential to engage in physical interaction with each other was tested. Findings from co-affinity purification (Appendix III, Fig. 3 A), co-immunoprecipitation (Co-IP) (Appendix III, Fig. 3 B) as well as Y2H (Appendix III, Fig. 3 C) assays supported the existence of physical interactions between the two. Interaction domain mapping then corroborated that the interaction is mediated by the N-terminal

domain of Gadd45a (amino acids 1-137) and the N- (amino acids 1-132) as well as C-terminus (amino acids 178-397) of TDG (Appendix III, Figure 3 D).

The effect of Gadd45a on TDG-mediated removal of 5fC/5caC was then evaluated by co-transfecting an *in vitro* generated 5fC/5caC-containing reporter plasmid (Appendix III, Fig. 5 A) with the Gadd45a expression vector into HEK293T cells. M.SssI-assisted bisulfite sequencing (MAB-seq) of the recovered plasmid revealed a substantially reduced level of 5fC/5caC (58% vs. 89%) (Appendix III, Fig. 5 B), whereas 5fC/5caC was largely retained without Gadd45a (84% vs. 89%) (Appendix III, Fig. 5 B). Consistent with the idea of TDG activity mediating the Gadd45a effect, there was no significant reduction of 5fC/5caC in TDG KO cells upon expression of Gadd45a (89.5% vs. 89.2%) (Appendix III, Fig. 5 B).

We next evaluated the biological relevance of Gadd45a in TET-TDG-mediated modulation of epigenetic states in ES cells. Interestingly, in cells lacking both Gadd45a and Gadd45b (*Gadd45a/Gadd45b* double knockout, DKO), no obvious difference in global methylation was observed at the genomic scale (Appendix III, Fig. S6). However, locus-specific hypermethylation but not hypomethylation was detectable (Appendix III, Table S3) as exemplified by the two genomic loci, the *Plagl1* and *Cilp2* genes (Appendix III, Fig. 6 A). Overall, the hypermethylated regions in Gadd45 deficient cells significantly overlapped with regions showing enrichment of 5hmC and 5fC in TDG deficient mouse ES cells (C. X. Song et al. 2013a; Yu et al. 2012) (Appendix III, Fig. 6 C). In these regions epigenetic states are considered to be dynamically maintained involving TET and TDG and the results presented here support the hypothesis that Gadd45 proteins functionally cooperate with the TET and TDG enzymes in the dynamic regulation of genomic targets in mouse ES cells.

It remains yet unclear how Gadd45a exactly regulates TDG activity in cells. From the direct interaction and the stimulation of TDG-dependent DNA demethylation by Gadd45a in cells, we reasoned that Gadd45a might impact on the enzymatic activity of TDG. I contributed to this study by examining this hypothesis *in vitro* using purified proteins. Evaluation of 5caC base release under single turnover conditions showed that Gadd45a did not stimulate the processing rate of TDG (Appendix III, Fig. S8 C and D). In presence of a large excess of substrate, however, Gadd45a increased the overall processing, albeit this effect was not specific compared to a BSA control (Appendix III, Fig. S8 A and B). Finally, EMSAs indicated that Gadd45a increased TDG DNA binding, thus exerting a stabilizing function, but again this effect turned out to be largely unspecific (data not included in the publication).

In conclusion, several lines of evidence provided in this study suggest that Gadd45a serves as a regulator in the TET-TDG-mediated DNA demethylation pathway: (i) Gadd45a enhanced TET and TDG-dependent reactivation of a methylated reporter gene in transfected cells (Appendix III, Fig. 1), (ii) Gadd45a interacted physically with TDG both *in vitro* and *in vivo* (Appendix III, Fig. 3), (iii) overexpression of Gadd45a enhanced TDG-dependent removal of 5fC and 5caC from genomic or transfected plasmid DNA in HEK293T cells (Appendix III, Fig. 2, 4, 5), (iv) deletion of Gadd45a/b in mouse ES cells led to locus-specific hypermethylation (Appendix III, Fig. 6). As Gadd45a did not stimulate TDG activity when examined biochemically, we propose a model where Gadd45a (and

Gadd45b) acts as a scaffold protein and promote TET-TDG-mediated DNA demethylation by coordinating the assembly of demethylation complexes at specific loci.

Contribution: My contribution started at a later stage of the project and consistent in the evaluation of the stimulatory effect of Gadd45a on TDG activity *in vitro*. I purified Gadd45a and TDG protein used in the base release assays. I performed the base release assays shown in Figure S8 as well as a series electrophoretic mobility shift assays (EMSA), results of which were not included in the final version of the manuscript. I wrote the relevant parts in the methods section and integrated my results in the discussion.

4.4 Supplementary Results

4.4.1 Biochemistry of TDG and TET1 suggests function in RNA-containing structures

The discovery of the TET proteins and the products of TET-mediated 5mC oxidation has greatly advanced our understanding of active DNA demethylation and led to reevaluation of previous mechanistic hypotheses regarding globally as well as locally observed DNA demethylation events (Kohli and Zhang 2013; Lee et al. 2014; Seisenberger et al. 2013; H. Wu and Zhang 2014). Also the DNA glycosylase TDG has received a great deal of attention and a role in TET-mediated DNA demethylation was soon suggested, particularly after the TET-mediated 5mC oxidation products 5fC and 5caC were identified to be readily recognized and excised by TDG (Hashimoto et al. 2012a; He et al. 2011; Maiti and Drohat 2011; L. Zhang et al. 2012) (chapter 4.1 and 4.3, Appendix I and III).

TDG has previously been implicated in active DNA demethylation in different contexts: in analogy to the glycosylase-initiated DNA demethylation mechanism in plants, it has formerly been suggested to be a 5mC-specific glycosylase by JP Jost (Jost 1993; Jost et al. 1995; B. Zhu et al. 2000b). It was also JP Jost and colleagues who first suggested that the glycosylase requires an RNA moiety for efficient DNA demethylation and that the RNA might exert a targeting function (Fremont et al. 1997; Jost et al. 1997). Careful biochemical evaluation of TDG interactions with RNA in our lab has confirmed that TDG does have affinity for RNA and that RNA can stimulate the TDG catalytic activity in certain contexts, not to the extent however that the enzyme would become proficient in 5mC excision (Christophe Kunz, unpublished results). Preliminary analysis of TDG cross-linking immunoprecipitation (CLIP) data as well as RNA immunoprecipitation (RIP) performed with mouse ES cells confirmed that TDG interacts with RNA *in vivo* (Christophe Kunz, unpublished results). To date, the biological role of this interaction is not entirely clear and the mechanistic implications of such interactions have not been addressed. Inside cells RNAs are present in multiple forms, they can be single- and double-stranded, highly structured, and present in RNA/DNA hybrids and R-loops. R-loops are generally formed during gene transcription at the transcription bubble where they have been suggested to play a role in transcriptional regulation as well as in other processes, e.g. recombination (Aguilera and Garcia-Muse 2012). TDG was also attributed a role in transcriptional regulation (Cortazar et al. 2007; Dalton and Bellacosa 2012; Sjolund et al. 2013) where the association with RNA in hybrids or R-loops might be biologically relevant.

Here, I wanted to combine the latest insight in oxidative DNA demethylation with previous observations on the biochemical properties of TDG. First, I evaluated TDG's affinity for 5mC and TET-mediated DNA demethylation intermediates as well as its activity on these bases in regular DNA duplex contexts. I then verified if oxidative DNA demethylation could occur in other substrate contexts, i.e. RNA/DNA hybrids or R-loops by testing the enzymatic activities of TDG and TET in newly generated assays with such hybrid substrates *in vitro*.

RESULTS

TDG binds to and processes 5caC but not 5mC or 5hmC

Following the discovery of the TET-mediated 5mC oxidation, I first set out to evaluate the affinity and activity of TDG for 5mC and its oxidized DNA demethylation intermediates. Binding affinity of TDG was measured by a standardized electrophoretic mobility shift assay (EMSA) (Fig. 4.1 A). TDG was incubated with labeled synthetic 60 bp DNA substrates containing one modified C in a CpG context (5mC, 5hmC, 5caC or T), alone as well as in the presence of unlabeled unspecific (60 bp homoduplex) or specific (60 bp G·5caC containing) competitor DNA. In the absence of competitor DNA, all the substrates were bound with high affinity using a two-fold molar excess of enzyme over substrate (100 nM DNA, 200 nM enzyme) (Fig. 4.1 A). In the presence of 10- and 20-fold molar excess of unspecific competitor DNA (1 µM and 2 µM), 5mC and 5hmC binding was reduced 10- to 20-fold, suggesting that

A

B
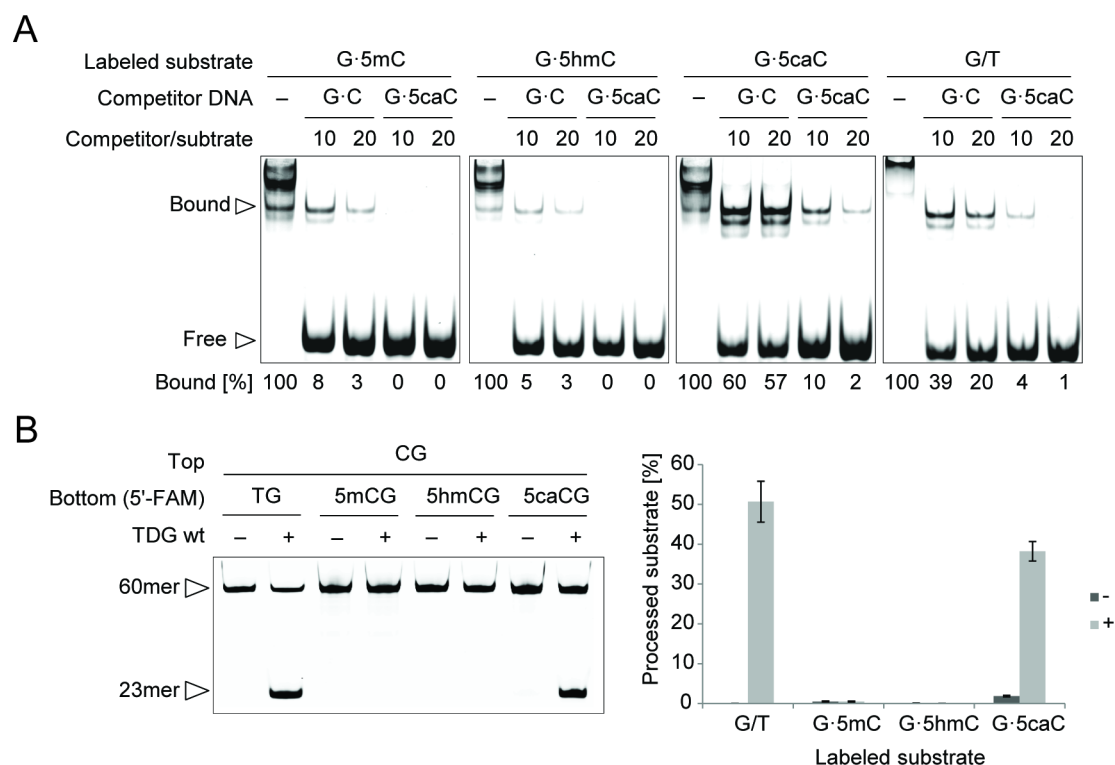
Fig. 4.1: TDG binds and processes 5caC specifically but not 5mC and 5hmC. (A) 1 pmol of 5' fluorescein labeled DNA substrate containing one modified C (5mC, 5hmC or 5caC) or a G/T mismatch were incubated with 2 pmol of TDG and 10 or 20 pmol unlabeled competitor DNA as indicated at 37°C for 15 min. The formation of DNA/protein complexes was examined by EMSA. (B) The ability to generate alkaline-sensitive abasic sites on the indicated substrates was assayed with a base release assay using equimolar enzyme and substrate concentrations at 37°C for 15 min. Products were separated by denaturing gel electrophoresis, visualized with fluorescent scanning and quantified. The 60mer substrate DNA and the 23mer product fragment are indicated by arrow heads. (n = 3; mean values with standard deviations).

a high degree of unspecific DNA binding was occurring under the assay conditions applied (Fig. 4.1 A, panel 1 and 2). In contrast, 57% of G·5caC and 20% of G/T containing DNA remained specifically bound in presence of a 20-fold molar excess of competitor DNA, indicating a strong preference and high affinity for these substrates (Fig. 4.1 A, panel 3 and 4). This result is interesting because unlike G/T, which represents a mismatch, 5caC forms regular Watson-Crick base pairs and cannot be recognized by mismatching characteristics. To directly compare the affinities of TDG for these two substrates, an unlabeled 5caC-containing DNA was used as competitor. Presence of the specific competitor DNA completely abolished binding to 5mC and 5hmC containing substrates (Fig. 4.1 A, panel 1 and 2) but also binding to a G/T mismatch was almost completely lost (Fig. 4.1 A, panel 4). As in this experimental setup TDG most likely generated AP-sites and DNA binding thus also reflected binding to these, the assay was repeated with a catalytically inactive TDG (TDGΔcat N151A). The result was virtually the same using TDGΔcat (data not shown) indicating that TDG has a higher affinity for the normally paired 5caC than for G/T mismatches. I then tested the ability of TDG to release the modified bases in a standardized base release assay (Hardeland et al. 2000). Incubation of equimolar concentrations (25 nM) of 5mC and 5hmC containing substrates with TDG at 37°C for 15 minutes did not produce any detectable base release products (Fig. 4.1 B). In contrast, 5caC and a G/T control substrate were well processed (38% and 50% respectively) and despite the lower binding efficiency for G/T as compared to 5caC (~1/3 in presence of a 20-fold molar excess of competitor DNA) (Fig. 4.1 A), processing of the mismatch was more efficient. Taken together, I could confirm more quantitatively previous findings that TDG has neither a specific binding affinity nor activity towards 5mC- and 5hmC-containing substrates. I could furthermore show the activity of TDG on 5caC, a TET-mediated 5mC oxidation product, and that TDG binds 5caC with higher affinity than a T mispaired with G.

TDG processes 5caC in RNA/DNA hybrids as well as in R-loops

Since TDG associated with RNAs *in vivo*, we hypothesized that its engagement in active DNA demethylation may rely on the presence of an RNA. I therefore evaluated the ability of TDG to process 5caC in RNA/DNA hybrid- as well as in R-loop substrates. I first tested TDG activity on various substrate bases in RNA/DNA hybrids (Fig. 4.2 A, schematic view) under single turnover conditions (10-fold molar excess of enzyme over substrate) (Fig. 4.2 A). The results suggested that TDG is active on U and to a lower extent on T in these hybrids but not on 5mC and 5hmC substrates (Fig. 4.2 A, lanes 2 - 5). Interestingly, in comparison to processing in dsDNA, the excision was more efficient for 5caC than for T in this context (Fig. 4.2 A, compare lanes 2 and 6). To compare the TDG activity on G·5caC and G/T in the hybrid substrates, we followed base release over time using equimolar enzyme and substrate concentrations (Fig. 4.2 B). Quantification of the reaction products revealed that processing of G·5caC in the context of a RNA/DNA hybrid is not much different than processing in dsDNA (Appendix I, Fig. 5 B). However, excision of T from a G/T mismatch was drastically reduced in this context (Fig. 4.2 B, right panel). Using the same experimental setup, base release was then monitored in the context of R-loop substrates (Fig. 4.2 C). Here, the difference appeared even more evident than in the hybrid substrates - albeit 5caC processing was slightly reduced when compared to excision in RNA/DNA hybrids, removal of T from a G/T mismatch was barely detectable in R-loops (Fig. 4.2 C). Here, the difference appeared even more evident than in the hybrid substrates - albeit
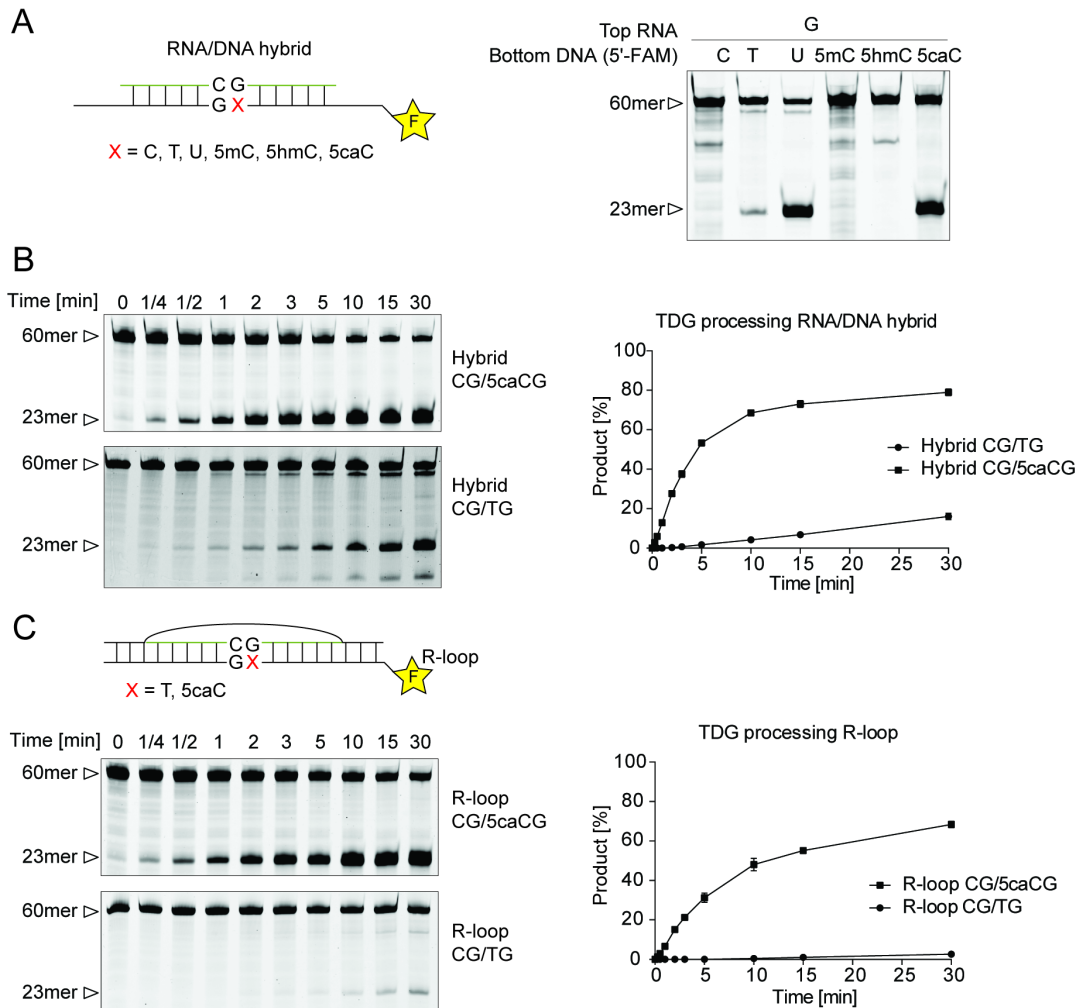
Fig. 4.2: TDG processes 5caC in RNA/DNA hybrids and in R-loops. (A) Schematic illustration of RNA/DNA hybrid substrate used in the base release assay, an unlabeled 25 nt RNA strand was annealed to complementary labeled 60 nt DNA strand. Shown are the results obtained with substrates containing G·C, G/T, G/U, G·5mC, G·5hmC or G·5caC base pairs. Products were separated by denaturing gel electrophoresis, visualized and quantified. Positions of the 60mer substrate DNA and product fragments are indicated. (B) Time-course base release assay on an RNA/DNA hybrid substrate containing G·5caC or G/T. Reactions were stopped after indicated timepoints by the addition of NaOH and separated by denaturing gel electrophoresis. Strands were then visualized by fluorescent scanning and quantified. (n = 3, mean values with standard deviations). (C) Schematic illustration of an R-loop substrate used in the base release assay, an unlabeled 25 nt RNA strand was annealed within the bubble structure of an unlabeled and a labeled 60 nt DNA strand. Release of 5caC or T was again monitored over time in a time-course base release assay. (n = 3, mean values with standard deviations).

5caC processing was slightly reduced when compared to excision in RNA/DNA hybrids, removal of T from a G/T mismatch was barely detectable in R-loops (Fig. 4.2 C).

Altogether, these experiments establish that purified TDG is active on RNA/DNA hybrids excising U as well as T in the DNA strand when mispaired with G in the RNA strand. Testing of 5mC and the DNA demethylation intermediates 5hmC and 5caC revealed no activity towards 5mC and 5hmC, whereas 5caC was well recognized and excised in these hybrid substrates. Comparing TDG kinetics on hybrid substrates containing either G·5caC or G/T, excision of 5caC occurred faster and with a higher overall efficiency than for T. Similar findings apply to processing of these bases in an R-loop context, where removal of T was almost not detectable while 5caC was readily recognized and released.

TET1$_{CD}$ is active on single-stranded DNA, RNA/DNA hybrids and in a non-CpG context

I next evaluated if the TET proteins are also enzymatically active in RNA/DNA hybrids as well as in R-loop substrates. Furthermore, I tested if 5mC can be released in a non-CpG context by the activity of TET1$_{CD}$ and TDG in a 60 bp substrate containing two methylated Cs in a CpA context. Substrates were first incubated with TET1$_{CD}$ alone at 37°C for 30 min before adding TDG to excise the TET-mediated 5mC oxidation products. The reactions were carried out using a 10-fold (ss, hybrid, R-loop) or five-fold (non-CpG) molar excess of TET1$_{CD}$ and TDG over substrate and equimolar end concentrations of TET1$_{CD}$ and TDG. Since TDG was highly active in ssDNA (Appendix I, Fig. 5 B), RNA/DNA hybrid and R-loop context (see Fig. 4.2), we reasoned that TDG activity was not limiting in these reactions and the efficiency of base release largely reflected the activity of the TET protein. The results showed that TET1$_{CD}$ was highly active on ssDNA (Fig. 4.3 A, lanes 2 and 6). In the presence of TDG both 5mC and 5hmC were released via an oxidized 5mC derivative at the same efficiency (Fig. 4.3 A, lanes 2 and 6). The activity was then found to be greatly reduced on hybrids and even more pronounced on R-loop substrates. Whereas substantial base release was detected from 5mC and 5hmC containing RNA/DNA hybrid substrates (Fig. 4.3 A, lanes 3 and 7), the amounts of products formed in R-loop substrates was strongly reduced (Fig. 4.3 A, lanes 4 and 8). In the non-CpG methylated substrate both 5mCs were targeted by TET1$_{CD}$ and TDG and subsequently released roughly to the same amounts (Fig. 4.3 B, lane 3). Taken together, the results suggested that the activity of TET1$_{CD}$ is neither restricted to double-stranded DNA nor to a CpG context since a substantial amount of oxidation followed by TDG-dependent base release could also be detected in single-stranded DNA as well as in RNA/DNA hybrids. In contrast to TDG, TET1$_{CD}$ did not appear to be active on R-loop substrates.

CONCLUSION & OUTLOOK

TDG has been associated with DNA demethylation and transcriptional regulation through multiple
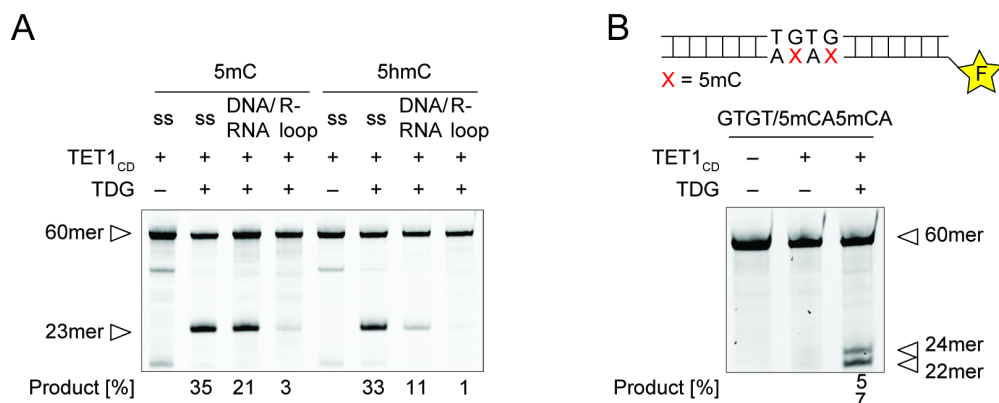
Fig. 4.3: TET1$_{CD}$ can convert 5mC in single-stranded DNA and in a non-CpG context. (A) The activity of TET1$_{CD}$ was analyzed by base release assays in combination with TDG. After incubating substrates at 37°C for 30 min with TET1$_{CD}$, an equimolar amount of TDG was added and incubated for another 30 min. Shown are the results obtained with substrates containing either 5mC or 5hmC in single-stranded DNA (ss), RNA/DNA hybrids or R-loops. Products were separated by denaturing gel electrophoresis, visualized and quantified. Positions of the 60mer substrate DNA and product fragments are indicated by arrow heads. (B) Combined TET1$_{CD}$ and TDG base release assay on duplex DNA containing two 5mC modifications in a CpA context. Arrows indicate the 60mer substrate and the two product fragments.

pathways (Dalton and Bellacosa 2012). But the most convincing biological as well as biochemical evidence has now pointed towards a mechanism where 5mC is first modified by the TET dioxygenases followed by TDG-mediated excision of the oxidized 5mC derivatives and BER to restore the unmodified state (He et al. 2011; Inoue et al. 2011; Ito et al. 2011; Maiti and Drohat 2011) (Appendix I). Shortly after the discovery of the TET proteins and its oxidation products, I could show that TDG binds and processes 5caC with high specificity but not 5mC or 5hmC. These results furthermore revealed that TDG binds 5caC with even higher affinity than a G/T mismatch, despite the fact that 5caC forms a normal base pair with G. Structural studies support these results and suggested that of all TDG substrates analyzed so far, 5caC established the most specific and strong hydrogen bonding interactions within the TDG active site, making it a high affinity and specificity substrate for TDG (Hashimoto et al. 2012a; L. Zhang et al. 2012). Despite the higher affinity for 5caC, the excision rate appeared to be lower than for G/T, indicating that the rate limiting step for 5caC processing was the cleavage of the N-glycosidic bond of the base. Cleavage efficiency of TDG was previously shown to depend on the stability of the N-glycosidic bond (Bennett et al. 2006; Hardeland et al. 2000), which is weakened for 5caC as well as T. However, 5caC excision may require a slightly higher activation energy than T excision (R. T. Williams and Wang 2012).

TDG has previously been shown to associate with RNA and it was hypothesized that the RNA might target or stimulate the DNA demethylation activity (Fremont et al. 1997; Jost et al. 1997). Results obtained in our lab also showed a stimulating effect of a complementary RNA on the catalysis of TDG; processing of G/T and G/U mismatches was enhanced but also excision of T and U from single-stranded DNA could be detected if the RNA induced a mismatch. Direct processing of 5mC by TDG, however, was not detected neither in single- nor double-stranded DNA in presence of the RNA (Christophe Kunz, unpublished results). I now evaluated processing of RNA/DNA hybrids (25mer RNA paired with complementary 60mer DNA) containing the TET-mediated oxidation products 5hmC and 5caC in the DNA moiety to see if oxidative DNA demethylation could occur in such a context. RNA/DNA hybrids are transient structures that can occur during transcription, when a newly synthesized RNA strand anneals to its complementary DNA strand. Not surprisingly, direct excision of 5mC and 5hmC by TDG alone could not be detected in such substrates but 5caC was found to be well processed and, unlike in duplex DNA, excised more efficiently than a T from a G/T mismatch. This difference in processing efficiency is even more pronounced in R-loop substrates, where the excision of T was almost completely abolished but 5caC was still well processed. Hence, 5caC processing of TDG in RNA/DNA hybrids and R-loops is biochemically evident but the biological relevance of this activity remains yet speculative. The activity on RNA-containing structures might be ascribed to the role of TDG in transcriptional regulation that has previously been suggested to occur either through interactions with transcription factors or by direct modulation of epigenetic states (D. Chen et al. 2003a; Dalton and Bellacosa 2012; Sjolund et al. 2013; Um et al. 1998). Furthermore, R-loops, typically present at or shortly behind transcription bubbles, were suggested to exert, besides others, DNA methylation-related functions (Aguilera and Garcia-Muse 2012) such as the protection of CGIs against DNA hypermethylation (Ginno et al. 2012). In this regard, R-loops may function as a signal to recruit DNA demethylation factors, including TDG, to keep CGI promoters in a hypomethylated state. Data from TDG chromatin immunoprecipitation sequencing (ChIPseq) indicated that TDG indeed

associated with gene regulatory elements, but was not restricted to these regions (Annika Wirz, unpublished results). The induction of DNA nicks through DNA demethylation at gene regulatory elements, triggered by short (jump-start) transcripts forming hybrid and R-loop structures, could also induce chromatin reorganization around TSSs that can ultimately lead to gene activation at silenced or transcriptionally poised loci. The potential biological roles of TDG together with RNA are manifold but convincing evidence if and how TDG might be involved in such processes is yet scarce and the subject still needs further investigation. Arguing against a model of oxidative DNA demethylation directly on hybrid and R-loop structures is the observation that $TET1_{CD}$ had reduced activity on 5mC and 5hmC in RNA/DNA hybrids and almost no detectable activity on 5mC and 5hmC in R-loops compared to single-strand context. It might however still be possible that TET and TDG act in an uncoupled way to obtain DNA demethylation at certain loci. Seemingly, in contrast to TDG, RNA had an inhibitory effect on the $TET1_{CD}$ catalytic activity in this experimental setup. However, since there was substantial processing in ssDNA, it can be hypothesized that $TET1_{CD}$ might still be recruited to R-loops but acts on the looped DNA single-strand rather than the one paired with the RNA. Together with TDG, DNA methylation could be actively removed, but also here more data is needed to support such a concept. Moreover, the role of the TET1 N-terminus is largely unknown but might be necessary for specific targeting and/or stability on particular substrates, i.e. hybrids or R-loops, thus also affecting its enzymatic efficiency. Interestingly, methylation of C does not exclusively occur in the context of DNA but can also be detected in various RNAs as 5-methyl-ribo-cytosine (5mrC). Yet, not much is known about its biological function (Motorin et al. 2010). Recently, it has been shown that $TET1_{CD}$ actively converts 5mrC to 5-hydroxymethyl-ribo-cytosine (5hmrC) in single-stranded RNA *in vitro* and 5hmrC could also be detected in total RNA of ES cells (Fu et al. 2014). Binding of TDG to RNA could be connected to the ability of TET to oxidize 5mrC in RNA and oxidative DNA demethylation might also be occurring in RNA or other RNA structures. However, it still needs to be clarified if TET is capable of performing more than the initial oxidation step, thus, if potential TDG substrates can be generated. In addition, TDG activity and downstream BER on an RNA substrate could not yet be shown and as there is yet no described biological function of RNA methylation it needs to be clarified if reversal of RNA methylation is biologically relevant at all.

Another noteworthy observation is that oxidative DNA demethylation is not only restricted to a CpG context and base release can also be detected in a non-CpG methylated substrate (Fig. 4.3 B). Although CpG dinucleotides remain the primary site for DNA methylation in mammals, there is good evidence that DNA methylation also occurs in non-CpG contexts (CpA, CpT and CpC) in mammalian cells (Lister et al. 2009). The function of non-CpG methylation is, however, just starting to be defined (Pinney 2014).The efficient release of a 5mC in a CpA context indicates that both $TET1_{CD}$ and TDG do not require a G 3' to the 5mC for specific recognition and processing. Altogether, the activities of $TET1_{CD}$ on 5mC and TDG on the TET-mediated 5mC oxidation product 5caC are robust. Interestingly, neither a double-stranded context nor a CpG sequence is required for efficient processing indicating that there is still plenty to be explored to relate these activities to biologically relevant processes.

MATERIALS AND METHODS

TDG and TET1$_{CD}$ expression constructs

For expression of an N-terminally His6-tagged mouse TDG, the murine TDGa cDNA was cloned into pET28c (Novagen-Merck). The TDG catalytic mutant (N151A) was generated by site-directed mutagenesis on the pET28-mTDGa vector using the QuickChange Site-directed Mutagenesis Kit (Stratagene) with primers mmTdg-N151A-s (5'-TATTGTGATCATTGGCATTGCCCCGGGATTAATGGCTGC-3') and mmTdg-N151A-as (5'-GCAGCCATTAATCCCGGGGCAATGCCAATGATCACAATA-3') according to the manufacturer's instructions. For the N-terminally His6-tagged TET1 catalytic domain (TET1$_{CD}$ aa 1367 - 2057) expression construct the cDNA of the mouse TET1 protein was amplified by adaptor PCR from an expression vector kindly provided by H. Leonhardt and introduced into the NotI and PacI sites of the pCDFDuet-1 vector (Novagen).

Recombinant protein expression and purification

The expression vectors were introduced into *E.coli* BL21(DE3) cells by electroporation. Overnight pre-cultures were diluted with fresh pre-warmed LB medium and grown at 30°C to an OD600 level of 0.6 - 0.8. Cultures were grown under selective pressure using either 50 mg/L of Kanamycin (TDG) or 50 mg/L of Streptomycin (TET1$_{CD}$). Protein expression was induced by the addition of the isopropyl-β-D-thiogalactose (IPTG) to a final concentration of 250 μM and cultures were further incubated at 25°C for 3 h (TET1$_{CD}$) or at 15°C for 16 h (TDG).

TDG was then purified as described previously (Kunz et al. 2009). Briefly, mTDG wt and Δcat were expressed from vectors pET28c-mTDGa.0 and pET28c-mTDGa.1, respectively, as described above. Cell lysis was carried out in Ni-NTA lysis buffer (50 mM Na-phosphate pH 7.5, 500 mM NaCl, 20% glycerol, 0.1% Tween-20, 20 mM imidazole, 20 mM β-mercaptoethanol, 0.1 mM PMSF) by sonication followed by extract clarification. The cleared supernatant was loaded onto a 5 mL HisTrap FF crude column (GE Healthcare) and bound protein was eluted with 400 mM imidazole. Eluted protein was dialyzed against Heparin buffer (25 mM Na-phosphate pH 7.0, 250 mM NaCl, 20% glycerol, 20 mM β-mercaptoethanol, 0.1 mM PMSF), loaded onto a 5 mL HiTrap Heparin HP column (GE Healthcare) and bound protein was eluted with a linear gradient of 250 mM – 1.5 M NaCl. For ion exchange, relevant fractions were pooled, dialyzed against AIEX buffer (50mM Bicine pH 8.8, 25 mM NaCl, 20% glycerol, 20 mM β-mercaptoethanol, 0.1 mM PMSF) and loaded onto a 1 mL Resource Q column (GE Healthcare). Bound protein was eluted with a linear salt gradient of 25 mM – 1 M NaCl and the purest fractions were finally dialyzed against storage buffer (50 mM Tris-HCl pH 8.0, 50 mM NaCl, 10% glycerol, 1 mM DTT), frozen on dry-ice and stored at -80 °C.

TET1$_{CD}$ was purified as described in Appendix I. The cleared lysate was loaded onto a 1 mL HisTrap FF crude column (GE Healthcare), bound protein was eluted with 400 mM imidazole and relevant fractions dialyzed against CIEX buffer (50 mM HEPES pH 7.2, 25 mM NaCl, 20% glycerol, 5 mM DTT and 0.1 mM PMSF). Dialyzed fractions were then loaded onto a 1 mL Resource S column (GE Healthcare) and bound protein was eluted with a linear salt gradient of 25 mM – 1 M NaCl and the

purest fractions were finally dialyzed against storage buffer (50 mM HEPES pH 7.2, 100 mM NaCl, 20% glycerol, 5 mM DTT), frozen on dry-ice and stored at -80 °C.

Oligonucleotides and substrate annealing

The oligonucleotides were synthesized by Microsynth or supplied from Adam Robertson (5mC, 5hmC and 5caC containing oligonucleotides). Typically a 5' fluorescein labeled upper strand was paired with its complementary lower strand. RNA/DNA hybrids were generated in annealing buffer (10 mM Tris-Cl pH 8, 50 mM NaCl) containing an unlabeled 25mer RNA (Subs25uG_RNA) and a complementary 5'-fluorescein labeled 60mer DNA (Subs60lx-F) by heating at 95°C for 5 min and ramping to 4°C at a rate of 0.02°C/s. The R-loop substrates were generated in a two-step process in annealing buffer (10 mM Tris-Cl pH 7.5, 2 mM $MgCl_2$, 1 mM DTT). First, an unlabeled 25mer RNA (Subs25uG-RNA) and a 5'-fluorescein labeled DNA (Subs60lx-F) were heated at 90°C for 1 min and the reaction was ramped to 71°C at a rate of 0.02°C/s before the unlabeled 60mer DNA (Subs60u_R-loop) was added and the reaction was ramped to 4°C.

Name	Sequence	Description
RNA Subs25uG	5'-GAUCCGAUGUCGACCUCAAACCUAG-3'	Unlabeled 25mer RNA, upper strand
Subs60uG	5'-TAGACATTGCCCTCGAGGTACCATGGATCCGATGTCGACCTCAAACCTAGACGAATTCCG-3'	Unlabeled 60mer, upper strand
Subs60lx-F	5'-CGGAATTCGTCTAGGTTTGAGGTxGACATCGGATCCATGGTACCTCGAGGGCAATGTCTA-3'	5'-fluorescein labeled 60mer, lower strand, x: T, U, 5mC, 5hmC, 5caC
Subs60u R-loop	5'-TAGACATTGCCCTCGAGGTACCATATCGAATCCTGATCAACACCCAACCGACGAATTCCG-3'	Unlabeled 60mer R-loop, upper strand
Subs60uGT	5'-TAGACATTGCCCTCGAGGTACCTATGTATATGTGTGTGTATGTGGCTTAGACGAATTCCG-3'	Unlabeled 60mer non-CG methylation, upper strand
Subs60lCA-F	5'-CGGAATTCGTCTAAGCCACATAxAxACACATATACATAGGTACCTCGAGGGCAATGTCTA-3'	5'-fluorescein labeled 60mer, non-CG methylation, lower strand, x: 5mC

Table 4.1: List of oligonucleotides used to generate substrates for *in vitro* assays.

Electrophoretic mobility shift assay

EMSAs were performed using the double-stranded oligonucleotide substrates described above. Standard EMSAs were carried out in a total reaction volume of 10 µl containing 2 pmol of recombinant protein and 1 pmol of labeled DNA substrate with varying amounts of unlabeled competitor DNA in 1x reaction buffer (50 mM Tris-HCl pH 8.0, 1 mM DTT, 5% glycerol, and 1 mM EDTA). After incubation at

37 °C for 15 min, the reactions were loaded immediately onto 6% (19:1) native polyacrylamide gels and separated in 0.5x TBE for 50 min at 12 V/cm at room temperature. The fluorescein-labeled DNA was also visualized with a Typhoon 9400 (GE Healthcare) and quantified using the ImageQuant TL software (GE Healthcare).

Base release assay

The catalytic activity of the TDG and TET1$_{CD}$ in combination with TDG was monitored by means of a standardized nicking assay (Hardeland et al. 2000). TDG reactions were carried out in a reaction volume of 20 µL containing nicking buffer (50 mM Tris-HCl pH 8.0, 1 mM EDTA, 1 mM DTT, 1 mg/ml BSA), 0.5 pmol of substrate and 5 pmol purified TDG at 37°C for 15 min. The reactions for combined TET1$_{CD}$ and TDG were carried out in 40 µL reaction volume containing TET reaction buffer (50 mM HEPES pH 8, 50 mM NaCl, 1 mM Di-Sodium-Ketoglutarate, 2 mM Ascorbic acid, 75 µM Fe(II), 1 mM ATP), 0.5 pmol of substrate and 5 pmol purified TET1$_{CD}$ at 37°C for 30 min before addition of 5 pmol purified TDG and incubation for another 30 min. Reactions were stopped by addition of 1 M NaOH to a final concentration of 100 mM and heating at 99°C for 10 min. After EtOH precipitation at -20°C overnight, the products were separated in a 15% denaturing polyacrylamide gel and labeled DNA was detected using the red or blue fluorescence mode of the Typhoon 9400 (GE Healthcare) and analyzed quantitatively by ImageQuant TL software (v7.0, GE Healthcare).

TDG time-course reactions were carried out in 200 µL reaction volume containing nicking buffer (50 mM Tris-HCl pH 8, 1 mM DTT, 0.1 mg/mL BSA, 1 mM EDTA), 5 pmol of labeled substrate DNA and 5 pmol of purified TDG. After the indicated times of incubation at 37°C, 20 µL aliquots were withdrawn and the reactions were stopped by the addition of 1 M NaOH to an end concentration of 100 mM and heating at 99°C for 10 min. Reaction products were analyzed by denaturing polyacrylamide gel electrophoresis and analyzed as described above.

4.4.2 TET proteins as potential SUMO targets

Posttranslational protein modifications, such as phosphorylation, acetylation, attachment of small polypeptides, are fast and energetically inexpensive ways to reversibly modulate stability, enzymatic activity and interaction dynamics of proteins. Modification by the SUMO proteins has turned out to be a widely applied regulatory mechanism; the number of SUMO substrates is still expanding alongside the number of mechanisms/pathways that are known to be regulated by SUMO. Molecular pathways affected by SUMO-modified proteins include most fundamental cellular transactions such as signal transduction, transcriptional regulation and DNA repair (Zhao 2007), and genome maintenance in general (Gill 2004). Notably SUMOylation plays a crucial role in developmental processes; depletion of non-redundant components of the SUMO pathway result in embryonic lethality (Lomeli and Vazquez 2011).

Previous biochemical studies have shown that SUMOylation of TDG provides an important regulatory function in the process of BER, facilitating the controlled dissociation of the glycosylase from the AP-

site and passing the repair intermediate onto APE1 for further processing (Hardeland et al. 2002). SUMOylation might not only be important for modulating DNA affinity of TDG but could also play an essential role in the coordination and regulation of the entire repair process by orchestrating the assembly of proper function of downstream BER factors to the damaged site (Roland Steinacher, unpublished results). Given that a SUMOylation deficient variant of TDG causes a cell differentiation defect in mouse ESCs similar to that of a catalytic deficient TDG (Roland Steinacher, unpuslished data) we wondered whether SUMOylation could have a regulatory function in DNA demethylation and coordinate processes upstream of TDG. As a beginning, I therefore tested if TET1 or any of the other TET proteins are themselves targets of SUMO modification, making use of the newly established SUMOylation tools (see chapter 4.2, Appendix II).

RESULTS

I first conducted an *in silico* screen for potential SUMO conjugation consensus motifs (ψKxE, where ψ is a large hydrophobic residue, generally isoleucine, leucine or valine and x can be any residue (Sampson et al. 2001) in the TET1 sequence using various available SUMOylation prediction tools (SUMOsp 2.0, SUMOplot, SUMOFI). Owing to its large size, multiple SUMO sites were predicted and the number varies slightly depending on the used tool (Fig. 4.4 A). The number of potential sites was reduced by overlapping the results with the highest score and examination of the conservation of the

A

TET1 predicted SUMO sites

SUMOsp 2.0			**SUMOplot**			**SUMOFI**	
Position	Peptide	Score	Position	Context	Score	Position	Hit Sequence
1208	EK **VKVE** P	4.341	K1783	STSHL **VKDE** STDFC	0.93	1207 - 1210	**VKvE**
1127	DQ **KKQE** Q	3.868	K1208	LAKEK **VKVE** PSDSL	0.93	1220 - 1223	**FKtE**
1322	VT **EKRE** A	3.5	K639	KKPKV **LKTD** FNNKP	0.91	1782 - 1785	**VKdE**
1783	HL **VKDE** S	1.877	K1221	LPTCQ **FKTE** SGGQT	0.85		
1221	CQ **FKTE** S	1.261	K802	GECDH **LKGP** RNTLL	0.8		

B

SUMO site conservation

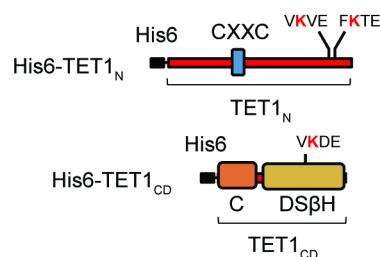| TET1	Position	Context
mouse	K1208	LAKEK **VKVE** PSDSL
human	K1256	SAEEK **VKVE** PLDSL
mouse	K1221	LPTCQ **FKTE** SGGQT
human	K1269	LSLFH **LKTE** SNGKA
mouse	K1783	STSHL **VKDE** STDFC
human	K1813	TVQPE **VKSE** TEPHF

C

Fig. 4.4: *In silico* screening for SUMO sites in mouse TET1. (A) The protein sequence of TET1 was analyzed for potential SUMO modification consensus motifs using three different prediction tools (SUMOsp 2.0, SUMOplot and SUMOFI, see Materials and Methods). Displayed are outputs with respective sequence context and score as received from the analysis tool, target lysine residues are highlighted in red. (B) The list of predicted TET1 SUMOylation sites in mouse was aligned with a list containing predicted sites for human TET1 (not shown). Conserved sites between mouse and human TET proteins are displayed in red. (C) Scheme of the expression constructs used for the validation of the predicted SUMO sites. The positions of the predicted SUMO acceptor sites are K1208, K1221 and K1783. His6-tagged TET1$_N$ (aa 301 - 1366) containing two and His6-tagged TET1$_{CD}$ (aa 1367 - 2057) containing one predicted SUMO site, respectively, were used in combination with pSUMO1 and pSUOM3. Abbreviations; CXXC, CXXC zinc finger domain (aa 568 – 607); Cat, TET1 catalytic domain (aa 1367 – 2057); C, Cys-rich region; DSβH, double-stranded beta helix.

predicted sites between mouse and human TET1, which then gave a final list of candidate sites (Fig. 4.4 B). Two of the resulting candidate SUMOylation sites of mouse TET1 are located outside the catalytic domain in close proximity to each other (K1208 and K1221) and one site is within the catalytic domain (K1783) (Fig. 4.4 C).

To evaluate the potential SUMOylation of TET1, I used the newly established in-cell and in-extract SUMOylation system for protein modification in *E.coli* (Weber et al. 2014) (Appendix II). TET1 was split into two expression constructs either containing the N-terminally His6-tagged TET1 N-terminal fragment (TET1$_N$ aa 301 - 1366) or the N-terminally His6-tagged TET1 catalytic domain (TET1$_{CD}$ aa 1367 - 2057) (Fig. 4.4 C). SUMOylation was then induced either in-cell by co-expressing the TET1 target constructs with pSUMO1 or pSUMO3 or in-extract by expressing the target and SUMOylation vectors separately followed by mixing of the cell lysates (for detailed methods see below and Appendix II).

Following the in-extract approach first, I thus prepared crude lysates from *E.coli* cultures expressing separately the pSUMO1 and pSUMO3 components (250 μM IPTG at 30°C, 3 hours induction) and TET1$_N$ (250 μM IPTG at 25°C, 3 hours induction) as well as TET1$_{CD}$ (250 μM IPTG at 25°C, 3 hours induction) mixed the extracts at different volume ratios and incubated the combined lysates at 30°C in the presence of ATP for 1 hour. As a positive control, I used bacterial lysates of cells expressing TDG, which was previously shown to be efficiently modified in this procedure. Under these conditions, I could not detect TET1$_N$ by immunoblotting, indicating an insufficient production of soluble protein due
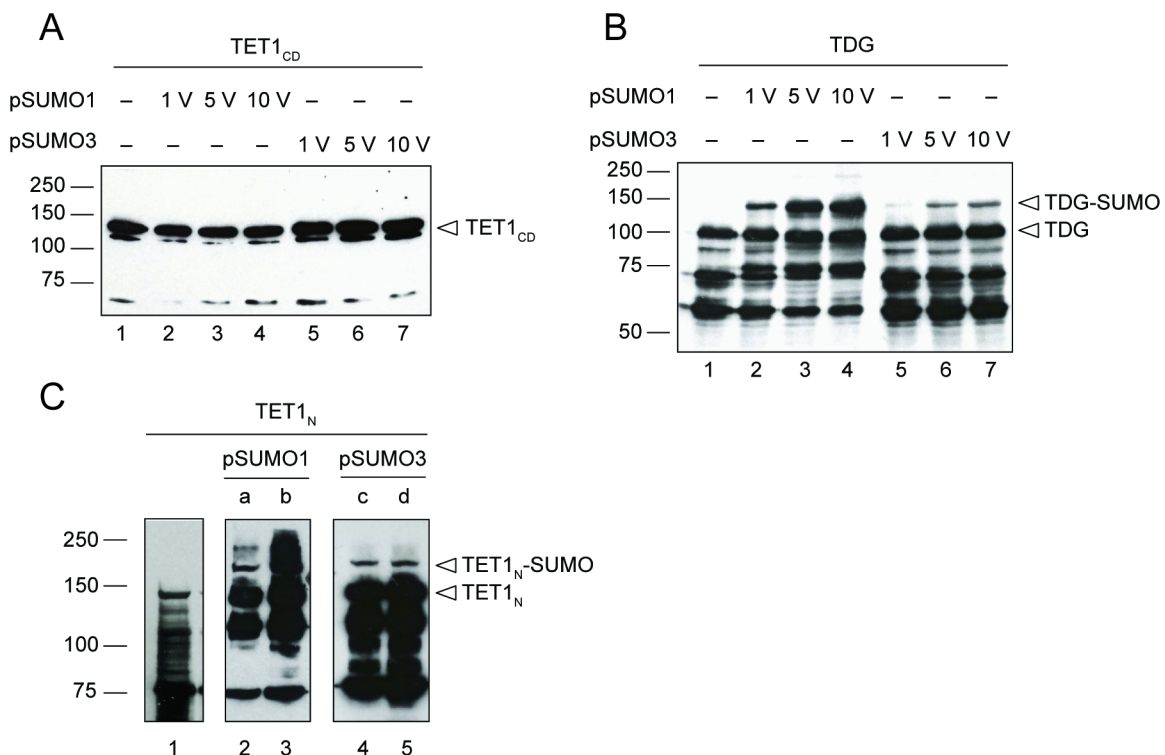

Fig. 4.5: TET1$_N$ but not TET1$_{CD}$ is SUMO-modified. Extracts from *E.coli* BL21(DE3) cells expressing the SUMO1 or the SUMO3 system (250 μM IPTG at 30°C for 3 h) were mixed with extracts from cells expressing TET1$_{CD}$ (A) or TDG (B) (250 μM IPTG at 25°C for 3 h) with the indicated volume (V) ratios and analyzed using immunoblot with respective antibodies. (C) Immunoblot analysis of TET1$_N$ SUMOylation in *E.coli* cells, expressing the TET1$_N$ from pACYC-mTET1-N and the SUMO system from either the pSUMO1 or the pSUMO3 plasmid (250 μM IPTG at 25°C for 3 h). a, b, c, d; different *E.coli* clones expressing the respective plasmids.

to codon-usage or folding issues (data not shown). However, TET1$_{CD}$ was readily detectable but did not seem to be SUMO-modified, neither by SUMO1 nor by SUMO3 (Fig. 4.5 A, lanes 2-7). Even the presence of a large excess of SUMOylation-potent extract did not produce any detectable SUMO-modified TET1$_{CD}$ (Fig. 4.5 A, lanes 3, 4, 6 and 7). By contrast, a substantial amount of SUMO1- as well as SUMO3-modified TDG was produced using the same procedures (Fig. 4.5 B, lanes 2-7), confirming the activity of the SUMO extracts in the TET experiment.

Next, I evaluated SUMOylation of the TET1$_N$ protein (Fig. 4.4 C) by co-expression with the recombinant SUMO system (pSUMO1 and pSUMO3) in *E.coli*. Two co-expressing *E.coli* clones each (SUMO1: a, b; SUMO3: c, d; Fig. 4.5 C) were analyzed for SUMOylation of TET1$_N$ under the applied expression conditions (250 µM IPTG at 25°C, 3 hours induction). I observed an additional prominent band in TET1 immunoblots migrating around 170 kDa, possibly corresponding to SUMO-modified protein, as compared to the negative control (TET1$_N$ expression alone) (Fig. 4.5 C). Seemingly, in-cell SUMOylation of TET1$_N$ could be achieved when it was co-expressed with the SUMOylation machinery from the pSUMO1 as well as pSUMO3 vector under these experimental conditions, albeit the efficiency was lower using pSUMO3.

CONCLUSION AND OUTLOOK

To investigate potential regulatory mechanisms in the process of DNA demethylation, I tested TET1 as a potential target for SUMO modification. Profound *in silico* analysis revealed that TET1 has several predicted SUMO conjugation consensus motifs, three of which are conserved between mouse and men, suggesting that these could represent genuine sites of modification. To establish a proof of concept I examined modification of mouse TET1, using the newly established SUMOylation system (Weber et al. 2014) (Appendix II). This revealed that only the N-terminus but not the catalytic domain of TET1 can be modified by SUMO1 as well as SUMO3. Notably, these results were obtained from a single experiment, but as the outcome was clearly positive, they can be taken as a strong indication for that TET1 being a SUMO-modified protein. From the observed pattern by immunoblot analysis, it appears that TET1$_N$ is modified at a single site. It can, however, not be concluded which of the two sites in the N-terminus is the target and further mapping studies, including site-directed mutagenesis of the predicted SUMO motifs, are required to ultimately identify the SUMO site. It will then be interesting to investigate the functional as well as structural consequences of the SUMO modification, particularly in the process of DNA demethylation in cooperation with TDG and BER. TDG has been described as SUMO target before affecting its structure and DNA binding properties to facilitate enzymatic turnover (Hardeland et al. 2002). Besides being covalently SUMO-modified, TDG also contains two SIMs, one in the N- and one in the C-terminal domain (Mohan et al. 2007), that could promote an interaction with TET or downstream BER factors. *In silico* analysis also predicted two conserved SIMs in the TET1 sequence (data not shown), further substantiating a potential involvement of SUMO in the coordination of TET-mediated processes. There are also multiple predicted SUMOylation sites for the other TET proteins TET2 and TET3 that are conserved between species (Fig. 4.6 A and B). As more of the DNA demethylation factors are being identified as SUMO targets and interactors, it will be of great interest to explore the role of SUMO interactions, modifications, and transfer in the orchestration of the DNA repair mediated DNA demethylation.

A

TET2 predicted SUMO sites

SUMOsp 2.0

Position	Peptide	Score
927	RP **IKTE** P	4.559
906	LL **QKQE** Q	4.397
953	SR **IKQE** I	4.071
111	KI **LKLD** Q	3.588
303	PF **QKAE** H	3.426
732	GQ **IKVE** E	2.474

SUMOplot

Position	Context	Score
K953	NMSSR **IKQE** ISSPS	0.94
K927	QMLRP **IKTE** PVSKP	0.94
K732	SCFGQ **IKVE** ESFCV	0.94
K111	HPNKI **LKLD** QKAKG	0.91
K1000	KSQKH **VKVP** TDIQA	0.82
K528	PGWIE **LKAP** NLHEA	0.8

SUMOFI

Position	Hit Sequence
731 - 734	**IKvE**
926 - 929	**IKtE**
952 - 955	**IKqE**

SUMO site conservation

TET2	Position	Context
mouse	K927	QMLRP **IKTE** PVSKP
human	K982	QMHRP **IKVE** PGCKP

B

TET3 predicted SUMO sites

SUMOsp 2.0

Position	Peptide	Score
1270	RS **IKQE** P	6.341
1061	EK **IKQE** A	5.431
631	KK **IKIE** S	3.991
355	PK **VKVE** A	3.782
1092	PS **LKVE** P	3.493
1496	GV **VKEE** K	2.422
1434	NP **VKVE** E	1.91
1470	GA **LKSE** E	1.469

SUMOplot

Position	Context	Score
K1270	CYNRS **IKQE** PIDPL	0.94
K1061	STPEK **IKQE** ALELA	0.94
K631	RSPKK **IKIE** SSGAV	0.94
K1496	PSKGV **VKEE** KSGPT	0.93
K1434	KLWNP **VKVE** EGRIP	0.93
K355	PTKPK **VKVE** APSSS	0.93
K1470	KLFGA **LKSE** EKLWD	0.91
K1092	SLKPS **LKVE** PQNHF	0.91
K1552	HATTP **LKKP** NRCHP	0.8

SUMOFI

Position	Hit Sequence
354 - 357	**VKvE**
630 - 633	**IKiE**
1060 - 1063	**IKqE**
1091 - 1094	**LKvE**
1269 - 1272	**IKqE**
1433 - 1436	**VKvE**
1469 - 1472	**LKsE**
1495 - 1498	**VKeE**

SUMO site conservation

TET3	Position	Context		TET3	Position	Context
mouse	K355	PTKPK **VKVE** APSSS		mouse	K1270	CYNRS **IKQE** PIDPL
human	K356	PTKPK **VKVE** APSSS		human	K1262	CYNRS **IKQE** PVDPL
mouse	K631	RSPKK **IKIE** SSGAV		mouse	K1434	KLWNP **VKVE** EGRIP
human	K623	RSPKQ **IKIE** SSGAV		human	K1426	KLWNP **MKGE** EGRIP
mouse	K1061	STPEK **IKQE** ALELA		mouse	K1470	KLFGA **LKSE** EKLWD
human	K1053	STPEK **IKQE** ALELA		human	K1462	KLFGA **LKSE** EKLWD
mouse	K1092	SLKPS **LKVE** PQNHF		mouse	K1496	PSKGV **VKEE** KSGPT
human	K1084	GLKPS **LKVE** PQNHF		human	K1488	PSKGA **VKEE** KGGGG

Fig. 4.6: *In silico* **screening for SUMO sites in mouse TET2 and TET3.** The protein sequence of (A) TET2 and (B) TET3 was analyzed for potential SUMOylation motifs (SUMOsp 2.0, SUMOplot and SUMOFI see Materials and Methods for details) which were then overlapped with a list containing predicted sites for human TET1 (not shown). Displayed are output sites with respective sequence context and score as received from the analysis tool (for SUMOsp only sites with a score > 1, for SUMOplot > 0.8 are shown), target lysine residues are highlighted in red.

MATERIALS AND METHODS

In silico analysis for the prediction of SUMOylation sites

For prediction of SUMOylation sites, online tools were used that are freely available on the internet. SUMOsp can be downloaded or used in an online version under http://sumosp.biocuckoo.org/. The SUMOplot analysis program is also an online tool that can be reached under http://www.abgent.com/sumoplot/. The SUmo MOtif FInder (SUMOFI) tool can be found under http://cbg.garvan.unsw.edu.au/sumofi/form.do.

Vector construction

The plasmids with the complete SUMO system (pSUMO1 and pSUMO3) were assembled by standard cloning methods as described in Appendix II and the plasmids, vector sequences and maps are available from Addgene (http://www.addgene.org/), plasmid ID 52258 and 52260. The cDNA of the mouse TET1 protein was amplified by adaptor PCR from an expression vector kindly provided by H. Leonhardt. For $TET1_N$ (aa 301 – 1366) the amplified fragment was introduced into the BamHI and PacI sites of pACYC-Duet1 vector (Novagen) and for $TET1_{CD}$ (aa 1367 – 2057) the fragment was introduced into the NotI and PacI sites of the pCDF-Duet1 vector (Novagen).

Recombinant protein expression, in-cell and in-extract SUMOylation

Detailed methods for in-cell and in-extract SUMOylation can be found in Appendix II. Briefly, expression vectors were introduced into *E.coli* BL21(DE3) cells by electroporation. Overnight pre-cultures were diluted 20 times, grown at 30°C to an OD600 of ~ 0.8 and protein expression was induced by the addition of the isopropyl-β-D-thiogalactose (IPTG) and incubated at indicated temperatures for the indicated times. Cells were harvested by centrifugation and soluble protein fractions were extracted by sonication in lysis buffer for in-cell modification (50 mM Na-phosphate buffer pH 7.5, 300 mM NaCl, 10% glycerol, 0.1% Tween-20, 1 mM DTT, 1 mM PMSF) or in SUMOylation buffer for in-extract modification (50 mM Tris-HCl pH 8, 50 mM NaCl, 10% glycerol, 0.5 mM DTT, 5 mM $MgCl_2$, 5 mM ATP). Crude lysates were then cleared by centrifugation with >20'000 g at 4°C for 30 min. For in-extract SUMOylation the lysates were mixed at indicated volume ratios, SUMOylation was triggered by the addition of ATP to a final concentration of 5 mM and reaction mixtures were incubated at 30°C for 1 h.

Analytical gel electrophoresis, western blotting and protein detection

Protein fractions were analyzed by standard SDS-polyacrylamide gel electrophoresis (SDS-PAGE) followed by immunoblotting using chemiluminescence (WesternBright ECL, Advansta) according to the provider's protocol. Antibodies were diluted in 5% non-fat dry milk TBS (100 mM Tris-HCl pH 8, 150 mM NaCl) supplemented with 0.2% Tween-20: TDG, rabbit polyclonal ab 141, 1:20'000; $TET1_{CD}$, rabbit polyclonal α-TET1 ab (Millipore), 1:5'000; $TET1_N$, rabbit polyclonal α-TET1 ab (Genetex), 1:10'000.

5 Concluding Discussion and Outlook

DNA cytosine methylation represents a key DNA modification throughout many phylogenies of life, which gained an important epigenetic function in multicellular organisms. Originally regarded as a rather static and rigid DNA mark, recent research has revealed unexpected dynamic features; the reversibility of DNA methylation was shown to be at the center of programming the genome function in mammalian development and life. Experimental evidence implicates the co-operation of both active and passive DNA demethylation processes to erase epigenetic marks. The search for candidate factors and mechanisms capable of actively reversing DNA methylation has been accompanied by controversy and uncertainty and led to the description of a variety of potential pathways (Ooi and Bestor 2008; S. C. Wu and Zhang 2010). Although still debated, the currently most plausible mechanisms involve DNA repair and are triggered by enzymatic activities that modify 5mC in order to make it visible for the repair machinery. Deamination of 5mC is one of them and would introduce a mismatch lesion to initiate DNA repair; several potential deaminases have been proposed to carry out the initial step in such a pathway (Fritz and Papavasiliou 2010). DNA demethylation triggered by oxidation of 5mC has also emerged as a plausible mechanism. Here, proteins of the TET family of DNA dioxygenases were suggested as candidates, owing to their ability to oxidize the methyl group of 5mC (Inoue et al. 2011; Ito et al. 2011). Already the initial conversion of 5mC to 5hmC might facilitate passive DNA demethylation through replicative dilution, but further oxidation to 5fC and 5caC generates substrates for the DNA glycosylase TDG (He et al. 2011; Maiti and Drohat 2011), thereby initiating an active process to remove 5mC and restore an unmodified C. TDG was previously implicated in the control of DNA methylation (Dalton and Bellacosa 2012) but a clear mechanistic concept supported by both biological as well as biochemical evidence has not been established until now.

During my PhD thesis, major advances were made in the field of active DNA demethylation and compelling evidence supporting an involvement of TDG in DNA demethylation has accumulated. In the proposed mechanisms TDG acts downstream of either 5mC oxidation by the TET dioxygenases or 5mC deamination by the AID deaminases, but many mechanistic details remain unresolved. My work focused on the establishment of a reconstituted *in vitro* DNA demethylation system consisting of TET, TDG, BER and auxiliary factors to proof the functionality of a DNA oxidation and repair mediated demethylation. This system can then be employed to study mechanisms of coordination and regulation of the process and to address pertinent questions regarding its impact on genome stability. According to the generally accepted model of oxidative DNA demethylation, 5mC is iteratively oxidized to 5hmC, 5fC and 5caC by the TET proteins followed by TDG-initiated DNA repair. This mechanism postulates a coupled action of TET and TDG to facilitate an efficient but coordinated removal of 5mC. The result of my work provides conclusive evidence for a direct physical interaction of TDG with TET1, mediated by domains located in the N- as well as C-terminus, linking TET1 to DNA repair. These results are also supported by recent findings in ES cells that showed co-localization of the two factors at sites undergoing dynamic epigenetic changes (L. Shen et al. 2013; C. X. Song et al. 2013a) (Annika Wirz, unpublished results). Moreover, the purified recombinant protein complex turned out to be highly active and capable of removing 5mC and 5hmC from a synthetic DNA substrate. Subsequent

biochemical experiments established a coupled action of both enzymes and showed for the first time that 5mC can be released via an oxidative intermediate using the combined activities of TET1$_{CD}$ and TDG. In the additional presence of the purified components of the core BER pathway, a methylated piece of DNA could for the first time be fully demethylated through a cascade of defined and coordinated enzymatic reactions. Hence, I was able to establish biochemical proof of functionality of TET-TDG-BER mediated active DNA demethylation.

To what extent, where and under what conditions oxidative DNA demethylation occurs in cells still needs more clarification. DNA repair-mediated demethylation processes harbor a certain genotoxic potential, raising concerns about how genome stability is maintained at developmental stages requiring a great number of such repair events. I used the *in vitro* reconstitution system to study the repair process at fully methylated CpG sites, where complete DNA demethylation by BER inevitably involves nicking of both strands and thereby generates an evident risk of DNA DSB formation. Therefore, tight regulation and coordination of the repair process in this delicate situation is crucial to avoid gross genomic instability. Indeed, the omission of DNA POLβ and XRCC1/LIG3 from the reconstituted DNA demethylation assay gave rise to the formation of substantial amount of detectable DSBs whereas faithful symmetrical 5caC repair could be reconstituted in the presence of all the required repair factors. My data suggests that the repair processes at both DNA strands occurs in an arbitrary but sequential manner, where the initially attacked strand is completely repaired prior to processing of the opposite strand. I propose that the unusual high affinity of TDG to its product (Hardeland et al. 2000; Waters et al. 1999) provides for a highly coordinated progression through the BER steps, altogether preventing an unscheduled processing of 5caC on the opposite strand and facilitating efficient replacement of symmetrical 5mC with unmodified C. Such a mode of action would resolve a main conceptual caveat associated with excision repair-mediated active DNA demethylation, namely the destabilization of the genome through excessive DNA repair. 5mC is prone to deaminate with a higher rate than C, yielding a T (J. C. Shen et al. 1994), and this was used to explain the underrepresentation of CpG dinucleotides in the mammalian genome. As cells have evolved efficient repair mechanisms to counter the mutability of 5mC deamination, however, this concept always appeared to be somewhat incomplete. We reasoned that 5mC deamination could coincide with oxidative DNA demethylation and generated a substrate reflecting such a situation by placing a 5caC next to a G/T mismatch within a CpG dinucleotide. Here, base release by TDG did not occur in a random manner but with high preference at the 5caC, leaving the G/T mismatch unrepaired. Given this, the observed sequential action of the TDG-BER system at symmetrically modified CpGs, which avoids the induction of deleterious DNA DSBs, could have mutagenic consequences. My results show that in case the repair synthesis associated with the replacement of 5caC incorporates two (or more) instead of only one nucleotide, the G/T mismatch is resolved to an A·T base pair and, hence, the CpG dinucleotide mutated into a TpG. My data thus provide strong evidence that C to T mutations at methylated CpGs can arise when deamination clashes with oxidative DNA demethylation. In this case the repair system itself prevents G/T correction and favors mutation giving a possible explanation for the observed underrepresentation of CpG dinucleotides in the genome.

Interestingly, the initial oxidation step by TET appeared to be much more efficient than the oxidation of 5hmC and 5fC in our *in vitro* experiments, resulting in higher amounts of 5hmC compared to 5fC and 5caC, which correlates with their levels in cells (Ito et al. 2011; Pfaffeneder et al. 2011). 5hmC is considered to have a distinct biological function as it attracts a specific set of "reader" proteins (Spruijt et al. 2013). This raised the question how TET activity could be regulated in cells to ensure that 5hmC is sustained and further oxidation is only induced in presence of the right signals. To this end, I tested whether the presence of TDG or Gadd45a, another factor implicated in DNA demethylation, could stimulate $TET1_{CD}$ enzymatic activity and thus exert a regulatory function in the demethylation process. However, neither TDG nor Gadd45a did modulate any of the $TET1_{CD}$-mediated oxidation steps. In contrast, presence of $TET1_{CD}$ stabilized the enzymatic activity of TDG, substantiating a functional interaction of them. Yet, little is known about the biological function and relevance of 5hmC and further studies regarding its impact on epigenetically regulated processes will contribute to the understanding of its catalysis and/or removal. 5fC and particularly 5caC occur at levels comparable to regular oxidative damage and might thus also be regarded as such, the role of TDG could then again be considered as classical repair enzyme. The low abundance of 5fC and 5caC might, however, simply reflect the high excision rate of these modifications and high dynamics of the demethylation process. This idea is supported by the observation of an accumulation of these demethylation intermediates in the absence of TDG (L. Shen et al. 2013; C. X. Song et al. 2013a)(Annika Wirz, Angelika Jacobs, unpublished results).

Notably, I could establish some additional biochemical features of $TET1_{CD}$ and TDG. The enzymatic activity of both proteins is neither restricted to CpGs nor to double stranded DNA, suggesting that oxidative DNA demethylation might also occur in other contexts. TDG has previously been associated with RNA and was shown to process RNA/DNA hybrid structures (Christophe Kunz, unpublished results). I evaluated this activity in the light of oxidative DNA demethylation and demonstrated 5caC processing in RNA/DNA hybrids as well as in R-loops. Yet the biological relevance of this activity remains speculative. R-loops are typically found at the transcription bubble and suggested to exert various specific functions (Aguilera and Garcia-Muse 2012), e.g. the protection of CpG islands from hypermethylation and, thereby, the regulation of genes (Ginno et al. 2012). In this regard, R-loops might function as signal to recruit DNA demethylation factors including TDG to keep CGI promoters in a hypomethylated state. Indeed, data from our group showed that TDG is recruited to CpG-rich gene regulatory elements together with TET proteins (Annika Wirz, unpublished results). Whether this recruitment is targeted by R-loops or other RNA/DNA hybrids, however is not yet known. Although TDG was previously described to be involved in the regulation of gene expression (Cortazar et al. 2011; Dalton and Bellacosa 2012; Sjolund et al. 2013), the underlying mechanism are not completely understood. Implicating a role of TET-TDG and BER in gene regulation, one could speculate that base release and subsequent nicking of the DNA might facilitate chromatin transactions at transcription start sites, which could contribute to gene activation.

The biochemically investigated and here described sophisticated cascade of enzymatic events leading to DNA demethylation needs to be tightly regulated and orchestrated in cells, and is likely to involve regulatory processes such as posttranslational protein modifications. Given that SUMO-modification

and interaction is prominent in the regulation of TDG, SUMOylation and SUMO-binding might have coordinative and regulatory functions at specific stages of active DNA demethylation as well. Biochemical studies have previously implicated SUMOylation in modulating TDG activity (Hardeland et al. 2002; Steinacher and Schar 2005) and also in the coordination of the downstream BER process by mediating protein-protein interactions (Roland Steinacher, unpublished results). To investigate a function of SUMOylation in DNA demethylation, we developed a versatile SUMOylation system to generate recombinant SUMOylated proteins in *E.coli*. Using this system, we efficiently modified and purified the BER factors TDG and XRCC1. In addition, preliminary results indicated that TET1 is also a SUMOylation target and is modified in its N-terminal domain. This and the presence of SIMs in both TDG and TET1 might hint to a potential contribution of SUMO in the process of DNA demethylation. The inclusion of a dynamic SUMOylation system in the reconstituted DNA demethylation assay could facilitate the investigation of the regulation of individual steps in this process. Future studies will provide additional insight into the structural as well as functional implications of SUMO-modification on individual targets but also more globally on the coordination of the DNA demethylation process.

In collaboration with Li Zheng and Guoliang Xu at the Chinese Academy of Science in Shanghai, we also explored other putative regulators of the DNA demethylation process by the TET-TDG cascade including those previously implicated. Among them is the multifunctional Gadd45a that was proposed to promote demethylation through XPG-dependent DNA repair (Barreto et al. 2007; Schmitz et al. 2009) or BER of AID-based deamination products (Cortellino et al. 2011; Rai et al. 2008). Notably, Gadd45a has no described enzymatic activity itself but based on our results several lines of evidence point towards a regulatory role in the TET-TDG oxidative DNA demethylation pathway. We could show that Gadd45a enhanced the activation of a methylated reporter gene in HEK293T cells when co-expressed with catalytically active TDG and TET. Furthermore, Gadd45a physically interacted with TDG and overexpression of Gadd45a increased the TDG-dependent removal of 5fC and 5caC from genomic and transfected plasmid DNA. Gadd45a and Gadd45b double knockout mouse ES cells were found to be hypermethylated at specific genomic loci that were described to be targets of TDG and show 5fC enrichment in TDG-deficient cells. Despite the diverse molecular functions that have been attributed to Gadd45a (Niehrs and Schafer 2012), our data indicate that DNA demethylation and activation of repressed genes is promoted specifically through TDG. Interestingly, Gadd45a does not appear to stimulate TDG in substrate binding, processing, or turnover in a specific manner when evaluated biochemically. We therefore favor a model whereby Gadd45a promotes TET-TDG mediated DNA demethylation by coordinating the recruitment of TDG into a demethylation complex. Gadd45a (and Gadd45b) could act as a scaffold protein which interacts with other epigenetic regulators and directs TDG to the target loci to efficiently excise 5fC and 5caC.

In conclusion, my mechanistic investigations provided novel insight into individual steps of the active DNA demethylation process and thereby extend our current understanding of this aspect of epigenetic reorganization occurring in cells undergoing differentiation. The most conclusive active DNA demethylation scenarios comprise an initial triggering event by modifying 5mC either through deamination or oxidation to make it susceptible for DNA repair. My work has focused on the oxidative DNA demethylation and I provide functional evidence that DNA demethylation along the TET-TDG

axis is biochemically valid in various contexts. I, in collaboration with others, also suggest additional regulatory mechanisms and factors that could feed into this pathway, i.e. SUMOylation and Gadd45a, but the exact regulation and targeting of these processes *in vivo* still needs to be resolved. In contrast to the TET-initiated pathway, the deamination-triggered process is in many cases disfavored by the biochemical properties of the suggested deaminases that mainly function in epigenetically unrelated processes. In addition, deamination of bases in the genome, even programmed ones, harbors a certain mutagenic potential that needs to be counteracted immediately. It is possible that AID-mediated DNA demethylation occurs under certain conditions at specific loci but other mechanisms, like oxidative DNA demethylation, are likely to operate in parallel and play a more prominent role. Biochemical as well as biological evidence has been accumulating, corroborating that the TET-TDG-mediated DNA demethylation pathway is a frequently implicated epigenetic mechanism at different stages of the mammalian lifecycle (Kohli and Zhang 2013; Pastor et al. 2013; H. Wu and Zhang 2014). My data on the interaction of TET1 and TDG, their combined activity on 5mC as well as the reconstitution of the DNA demethylation process including downstream BER further supports this pathway, rendering it biochemically valid throughout. The fact that both proteins are also able to act outside CpG and double stranded DNA contexts makes it all the more versatile and likely to be involved in so far unexplored situations. In contrast to deamination-induced DNA demethylation, the oxidative pathway does not generate mutagenic intermediates. Instead, the TET-mediated C modifications appear to be truly epigenetic marks in the sense that they do not alter the DNA sequence but may be recognized by specific reader proteins. Through oxidation of 5mC, the genome can retain the necessary flexibility without directly affecting its integrity and coding information. Still, as I have shown, oxidative DNA demethylation may harbor a mutagenic potential under conditions where active DNA demethylation clashes with deamination of 5mC within a CpG dinucleotide. Nevertheless, active DNA demethylation events seemingly only make up a minor part in global methylation dynamics and TET proteins can also contribute to DNA demethylation in a targeted and repair-independent way by masking the methylation mark for passive dilution through oxidation without the need to excise the base.

One could now ask whether active repair-mediated DNA demethylation is required at all. The answer is likely yes, but further studies are needed to evaluate where and under what circumstances DNA methylation states need to be maintained by or targeted for active DNA demethylation. Considering the iterative modification of C-5mC-5hmC-5fC-5caC-C as an ongoing cycle at certain loci under certain conditions, TET, TDG, BER and the DNMTs might operate as mediators of a dynamic equilibrium of DNA methylation and demethylation, establishing a steady-state rather than a stable methylation state at individual CpGs. Other factors like Gadd45a potentially also feed into this cycle and the purpose of the DNA repair mediated cycling may not be to regulate methylation in the first place but to maintain certain loci in a respective state of epigenetic plasticity. Most likely, this process of active epigenetic maintenance does not occur genome-wide but rather targeted and might be particularly important during phases where epigenetic patterns and transcriptional programs need to be correctly established, i.e. during cell differentiation or reprogramming of developmental potency. Disrupting this DNA methylation-demethylation cycle at any point, e.g. by depletion of TDG, would then lead to a locus-specific manifestations of aberrant epigenetic states.

With the results presented in my PhD thesis we are able to advance our understanding of the mechanism of active DNA demethylation. I, in collaboration with others, provide a biochemically validated concept for the coordinated and productive demethylation of DNA and additionally suggest potential regulatory mechanisms. Still much is to learn about the regulation, targeting and the sequence of events of this dynamic process but we are closer than ever to finally settle the debate about how DNA methylation can be reversed.

6 References

Aguilera, A. and Garcia-Muse, T. (2012), 'R loops: from transcription byproducts to threats to genome stability', *Molecular cell,* 46 (2), 115-24.

Arab, K., et al. (2014), 'Long noncoding RNA TARID directs demethylation and activation of the tumor suppressor TCF21 via GADD45A', *Molecular cell,* 55 (4), 604-14.

Aravind, L. and Koonin, E. V. (2000), 'The alpha/beta fold uracil DNA glycosylases: a common origin with diverse fates', *Genome Biol,* 1 (4), RESEARCH0007.

Ballestar, E. and Wolffe, A. P. (2001), 'Methyl-CpG-binding proteins. Targeting specific gene repression', *Eur J Biochem,* 268 (1), 1-6.

Bannister, A. J. and Kouzarides, T. (2011), 'Regulation of chromatin by histone modifications', *Cell Res,* 21 (3), 381-95.

Barreto, G., et al. (2007), 'Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation', *Nature,* 445 (7128), 671-5.

Barrett, T. E., et al. (1998), 'Crystal structure of a G:T/U mismatch-specific DNA glycosylase: mismatch recognition by complementary-strand interactions', *Cell,* 92 (1), 117-29.

Barrett, T. E., et al. (1999), 'Crystal structure of a thwarted mismatch glycosylase DNA repair complex', *The EMBO journal,* 18 (23), 6599-609.

Barry, J. and Lock, R. B. (2011), 'Small ubiquitin-related modifier-1: Wrestling with protein regulation', *The international journal of biochemistry & cell biology,* 43 (1), 37-40.

Baute, J. and Depicker, A. (2008), 'Base excision repair and its role in maintaining genome stability', *Crit Rev Biochem Mol Biol,* 43 (4), 239-76.

Bennett, M. T., et al. (2006), 'Specificity of human thymine DNA glycosylase depends on N-glycosidic bond stability', *J Am Chem Soc,* 128 (38), 12510-9.

Bernstein, B. E., et al. (2006), 'A bivalent chromatin structure marks key developmental genes in embryonic stem cells', *Cell,* 125 (2), 315-26.

Bhattacharya, S. K., et al. (1999), 'A mammalian protein with specific demethylase activity for mCpG DNA', *Nature,* 397 (6720), 579-83.

Bhutani, N., et al. (2009), 'Reprogramming towards pluripotency requires AID-dependent DNA demethylation', *Nature,* 463 (7284), 1042-7.

Bird, A. (2002), 'DNA methylation patterns and epigenetic memory', *Genes Dev,* 16 (1), 6-21.

Blaschke, K., et al. (2013), 'Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells', *Nature,* 500 (7461), 222-6.

Boland, M. J. and Christman, J. K. (2008), 'Characterization of Dnmt3b:thymine-DNA glycosylase interaction and stimulation of thymine glycosylase-mediated repair by DNA methyltransferase(s) and RNA', *J Mol Biol,* 379 (3), 492-504.

Borgel, J., et al. (2010), 'Targets and dynamics of promoter DNA methylation during early mouse development', *Nature genetics,* 42 (12), 1093-100.

Borst, P. and Sabatini, R. (2008), 'Base J: discovery, biosynthesis, and possible functions', *Annual review of microbiology,* 62, 235-51.

Borys-Brzywczy, E., et al. (2005), 'Mismatch dependent uracil/thymine-DNA glycosylases excise exocyclic hydroxyethano and hydroxypropano cytosine adducts', *Acta biochimica Polonica,* 52 (1), 149-65.

Bostick, M., et al. (2007), 'UHRF1 plays a role in maintaining DNA methylation in mammalian cells', *Science,* 317 (5845), 1760-4.

Brown, T. C. and Jiricny, J. (1988), 'Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells', *Cell,* 54 (5), 705-11.

Bruderer, R., et al. (2011), 'Purification and identification of endogenous polySUMO conjugates', *EMBO reports,* 12 (2), 142-8.

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

Cantone, I. and Fisher, A. G. (2013), 'Epigenetic programming and reprogramming during development', *Nature structural & molecular biology,* 20 (3), 282-9.

Chen, C. C., Wang, K. Y., and Shen, C. K. (2012), 'The Mammalian de novo DNA Methyltransferases Dnmt3a and Dnmt3b Are Also DNA 5-Hydroxymethyl Cytosine Dehydroxymethylases', *J Biol Chem*.

Chen, D., et al. (2003a), 'T:G mismatch-specific thymine-DNA glycosylase potentiates transcription of estrogen-regulated genes through direct interaction with estrogen receptor alpha', *J Biol Chem,* 278 (40), 38586-92.

Chen, T., et al. (2003b), 'Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b', *Mol Cell Biol,* 23 (16), 5594-605.

Choi, Y., et al. (2002), 'DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in arabidopsis', *Cell,* 110 (1), 33-42.

Cohen, N. M., Kenigsberg, E., and Tanay, A. (2011), 'Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection', *Cell,* 145 (5), 773-86.

Cortazar, D., et al. (2007), 'The enigmatic thymine DNA glycosylase', *DNA Repair (Amst),* 6 (4), 489-504.

Cortazar, D., et al. (2011), 'Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability', *Nature,* 470 (7334), 419-23.

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

Cremona, C. A., et al. (2012), 'Extensive DNA damage-induced sumoylation contributes to replication and repair and acts in addition to the mec1 checkpoint', *Mol Cell,* 45 (3), 422-32.

Cubenas-Potts, C. and Matunis, M. J. (2013), 'SUMO: a multifaceted modifier of chromatin structure and function', *Developmental cell,* 24 (1), 1-12.

Dalhus, B., et al. (2009), 'DNA base repair--recognition and initiation of catalysis', *FEMS Microbiol Rev,* 33 (6), 1044-78.

Dalton, S. R. and Bellacosa, A. (2012), 'DNA demethylation by TDG', *Epigenomics,* 4 (4), 459-67.

Deaton, A. M. and Bird, A. (2011), 'CpG islands and the regulation of transcription', *Genes Dev,* 25 (10), 1010-22.

Delatte, B., Deplus, R., and Fuks, F. (2014), 'Playing TETris with DNA modifications', *The EMBO journal,* 33 (11), 1198-211.

Denis, H., Ndlovu, M. N., and Fuks, F. (2011), 'Regulation of mammalian DNA methyltransferases: a route to new mechanisms', *EMBO reports,* 12 (7), 647-56.

Dianov, G. and Lindahl, T. (1994), 'Reconstitution of the DNA base excision-repair pathway', *Current biology : CB,* 4 (12), 1069-76.

Dianov, G. L. and Hubscher, U. (2013), 'Mammalian base excision repair: the forgotten archangel', *Nucleic Acids Res,* 41 (6), 3483-90.

Dohmen, R. J. (2004), 'SUMO protein modification', *Biochim Biophys Acta,* 1695 (1-3), 113-31.

Dong, K. B., et al. (2008), 'DNA methylation in ES cells requires the lysine methyltransferase G9a but not its catalytic activity', *The EMBO journal,* 27 (20), 2691-701.

Duncan, T., et al. (2002), 'Reversal of DNA alkylation damage by two human dioxygenases', *Proceedings of the National Academy of Sciences of the United States of America,* 99 (26), 16660-5.

Engel, N., et al. (2009), 'Conserved DNA methylation in Gadd45a(-/-) mice', *Epigenetics,* 4 (2), 98-9.

Figueroa, M. E., et al. (2010), 'Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation', *Cancer Cell,* 18 (6), 553-67.

Fitzgerald, M. E. and Drohat, A. C. (2008), 'Coordinating the initial steps of base excision repair. Apurinic/apyrimidinic endonuclease 1 actively stimulates thymine DNA glycosylase by disrupting the product complex', *J Biol Chem,* 283 (47), 32680-90.

Flotho, A. and Melchior, F. (2013), 'Sumoylation: a regulatory protein modification in health and disease', *Annual review of biochemistry,* 82, 357-85.

Franchini, D. M., Schmitz, K. M., and Petersen-Mahrt, S. K. (2012), '5-Methylcytosine DNA demethylation: more than losing a methyl group', *Annual review of genetics,* 46, 419-41.

Franchini, D. M., et al. (2014), 'Processive DNA demethylation via DNA deaminase-induced lesion resolution', *PLoS One,* 9 (7), e97754.

Fremont, M., et al. (1997), 'Demethylation of DNA by purified chick embryo 5-methylcytosine-DNA glycosylase requires both protein and RNA', *Nucleic Acids Res,* 25 (12), 2375-80.

Fritz, E. L. and Papavasiliou, F. N. (2010), 'Cytidine deaminases: AIDing DNA demethylation?', *Genes Dev,* 24 (19), 2107-14.

Fu, L., et al. (2014), 'Tet-mediated formation of 5-hydroxymethylcytosine in RNA', *Journal of the American Chemical Society,* 136 (33), 11582-5.

Gareau, J. R. and Lima, C. D. (2010), 'The SUMO pathway: emerging mechanisms that shape specificity, conjugation and recognition', *Nature reviews. Molecular cell biology,* 11 (12), 861-71.

Geiss-Friedlander, R. and Melchior, F. (2007), 'Concepts in sumoylation: a decade on', *Nat Rev Mol Cell Biol,* 8 (12), 947-56.

Germann, M. W., Johnson, C. N., and Spring, A. M. (2012), 'Recognition of damaged DNA: structure and dynamic markers', *Medicinal research reviews,* 32 (3), 659-83.

Gill, G. (2004), 'SUMO and ubiquitin in the nucleus: different functions, similar mechanisms?', *Genes & development,* 18 (17), 2046-59.

Ginno, P. A., et al. (2012), 'R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters', *Molecular cell,* 45 (6), 814-25.

Girard, L. and Freeling, M. (1999), 'Regulatory changes as a consequence of transposon insertion', *Developmental genetics,* 25 (4), 291-6.

Globisch, D., et al. (2011), 'Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates', *PLoS One,* 5 (12), e15367.

Gocke, C. B., Yu, H., and Kang, J. (2005), 'Systematic identification and analysis of mammalian small ubiquitin-like modifier substrates', *J Biol Chem,* 280 (6), 5004-12.

Gong, Z., et al. (2002), 'ROS1, a repressor of transcriptional gene silencing in Arabidopsis, encodes a DNA glycosylase/lyase', *Cell,* 111 (6), 803-14.

Gopalakrishnan, S., et al. (2009), 'DNMT3B interacts with constitutive centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions', *Human molecular genetics,* 18 (17), 3178-93.

Gregory, D. J., Mikhaylova, L., and Fedulov, A. V. (2012), 'Selective DNA demethylation by fusion of TDG with a sequence-specific DNA-binding domain', *Epigenetics,* 7 (4), 344-9.

Guibert, S., Forne, T., and Weber, M. (2012), 'Global profiling of DNA methylation erasure in mouse primordial germ cells', *Genome Res,* 22 (4), 633-41.

Guo, J. U., et al. (2011), 'Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain', *Cell*.

Hackett, J. A., et al. (2013), 'Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine', *Science,* 339 (6118), 448-52.

Hajkova, P., et al. (2010), 'Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway', *Science,* 329 (5987), 78-82.

Hardeland, U., et al. (2000), 'Separating substrate recognition from base hydrolysis in human thymine DNA glycosylase by mutational analysis', *J Biol Chem,* 275 (43), 33449-56.

Hardeland, U., et al. (2002), 'Modification of the human thymine-DNA glycosylase by ubiquitin-like proteins facilitates enzymatic turnover', *EMBO J,* 21 (6), 1456-64.

Hardeland, U., et al. (2003), 'The versatile thymine DNA-glycosylase: a comparative characterization of the human, Drosophila and fission yeast orthologs', *Nucleic Acids Res,* 31 (9), 2261-71.

Hardeland, U., et al. (2007), 'Cell cycle regulation as a mechanism for functional separation of the apparently redundant uracil DNA glycosylases TDG and UNG2', *Nucleic Acids Res,* 35 (11), 3859-67.

Hashimoto, H., et al. (2012a), '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 (20), 10203-14.

Hashimoto, H., et al. (2012b), 'Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation', *Nucleic acids research,* 40 (11), 4841-9.

Hashimoto, H., et al. (2014), 'Structure of a Naegleria Tet-like dioxygenase in complex with 5-methylcytosine DNA', *Nature,* 506 (7488), 391-5.

Hay, R. T. (2013), 'Decoding the SUMO signal', *Biochemical Society transactions,* 41 (2), 463-73.

He, Y. F., et al. (2011), 'Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA', *Science,* 333 (6047), 1303-7.

Henckel, A., et al. (2009), 'Histone methylation is mechanistically linked to DNA methylation at imprinting control regions in mammals', *Human molecular genetics,* 18 (18), 3375-83.

Hendrich, B., et al. (2001), 'Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development', *Genes Dev,* 15 (6), 710-23.

Hermann, A., Gowher, H., and Jeltsch, A. (2004), 'Biochemistry and biology of mammalian DNA methyltransferases', *Cell Mol Life Sci,* 61 (19-20), 2571-87.

Hu, L., et al. (2013), 'Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation', *Cell,* 155 (7), 1545-55.

Huang, Y., et al. (2014), 'Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells', *Proceedings of the National Academy of Sciences of the United States of America,* 111 (4), 1361-6.

Hubner, M. R., Eckersley-Maslin, M. A., and Spector, D. L. (2013), 'Chromatin organization and transcriptional regulation', *Current opinion in genetics & development,* 23 (2), 89-95.

Illingworth, R. S. and Bird, A. P. (2009), 'CpG islands--'a rough guide'', *FEBS Lett,* 583 (11), 1713-20.

Illingworth, R. S., et al. (2010), 'Orphan CpG islands identify numerous conserved promoters in the mammalian genome', *PLoS genetics,* 6 (9), e1001134.

Inoue, A., et al. (2011), 'Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development', *Cell Res,* 21 (12), 1670-6.

Iqbal, K., et al. (2011), 'Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine', *Proc Natl Acad Sci U S A,* 108 (9), 3642-7.

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

Ito, S., et al. (2011), 'Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine', *Science.*

Jackson, S. P. and Durocher, D. (2013), 'Regulation of DNA Damage Responses by Ubiquitin and SUMO', *Mol Cell.*

Jacobs, A. L. and Schar, P. (2011), 'DNA glycosylases: in DNA repair and beyond', *Chromosoma.*

Jakobs, A., et al. (2007), 'Ubc9 fusion-directed SUMOylation (UFDS): a method to analyze function of protein SUMOylation', *Nat Methods,* 4 (3), 245-50.

Jang, H., et al. (2014), 'Excision of 5-hydroxymethylcytosine by DEMETER family DNA glycosylases', *Biochemical and biophysical research communications,* 446 (4), 1067-72.

Jeong, S., et al. (2009), 'Selective anchoring of DNA methyltransferases 3A and 3B to nucleosomes containing methylated DNA', *Mol Cell Biol,* 29 (19), 5366-76.

Jin, S. G., Guo, C., and Pfeifer, G. P. (2008), 'GADD45A does not promote DNA demethylation', *PLoS Genet,* 4 (3), e1000013.

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

Jones, P. A. and Baylin, S. B. (2002), 'The fundamental role of epigenetic events in cancer', *Nat Rev Genet,* 3 (6), 415-28.

Jones, P. A. and Liang, G. (2009), 'Rethinking how DNA methylation patterns are maintained', *Nat Rev Genet,* 10 (11), 805-11.

Jost, J. P. (1993), 'Nuclear extracts of chicken embryos promote an active demethylation of DNA by excision repair of 5-methyldeoxycytidine', *Proc Natl Acad Sci U S A,* 90 (10), 4684-8.

Jost, J. P. and Jost, Y. C. (1994), 'Transient DNA demethylation in differentiating mouse myoblasts correlates with higher activity of 5-methyldeoxycytidine excision repair', *J Biol Chem,* 269 (13), 10040-3.

Jost, J. P., Saluz, H. P., and Pawlak, A. (1991), 'Estradiol down regulates the binding activity of an avian vitellogenin gene repressor (MDBP-2) and triggers a gradual demethylation of the mCpG pair of its DNA binding site', *Nucleic Acids Res,* 19 (20), 5771-5.

Jost, J. P., et al. (1995), 'Mechanisms of DNA demethylation in chicken embryos. Purification and properties of a 5-methylcytosine-DNA glycosylase', *J Biol Chem,* 270 (17), 9734-9.

Jost, J. P., et al. (1997), 'The RNA moiety of chick embryo 5-methylcytosine- DNA glycosylase targets DNA demethylation', *Nucleic Acids Res,* 25 (22), 4545-50.

Jullien, J., et al. (2011), 'Mechanisms of nuclear reprogramming by eggs and oocytes: a deterministic process?', *Nat Rev Mol Cell Biol,* 12 (7), 453-9.

Jurkowska, R. Z., Jurkowski, T. P., and Jeltsch, A. (2011), 'Structure and function of mammalian DNA methyltransferases', *Chembiochem : a European journal of chemical biology,* 12 (2), 206-22.

Kaelin, W. G., Jr. and McKnight, S. L. (2013), 'Influence of metabolism on epigenetics and disease', *Cell,* 153 (1), 56-69.

Kagiwada, S., et al. (2012), 'Replication-coupled passive DNA demethylation for the erasure of genome imprints in mice', *EMBO J,* 32 (3), 340-53.

Kaneda, M., et al. (2004), 'Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting', *Nature,* 429 (6994), 900-3.

Kangaspeska, S., et al. (2008), 'Transient cyclical methylation of promoter DNA', *Nature,* 452 (7183), 112-5.

Karimi, M. M., et al. (2011), 'DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs', *Cell Stem Cell,* 8 (6), 676-87.

Kawasaki, Y., et al. (2014), 'Active DNA demethylation is required for complete imprint erasure in primordial germ cells', *Scientific reports,* 4, 3658.

Kim, E. T., et al. (2009a), 'Enhanced SUMOylation of proteins containing a SUMO-interacting motif by SUMO-Ubc9 fusion', *Biochem Biophys Res Commun,* 388 (1), 41-5.

Kim, M. S., et al. (2009b), 'DNA demethylation in hormone-induced transcriptional derepression', *Nature,* 461 (7266), 1007-12.

Kim, Y. J. and Wilson, D. M., 3rd (2012), 'Overview of base excision repair biochemistry', *Current molecular pharmacology,* 5 (1), 3-13.

Ko, M., et al. (2013), 'Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX', *Nature*.

Kohli, R. M. and Zhang, Y. (2013), 'TET enzymes, TDG and the dynamics of DNA demethylation', *Nature,* 502 (7472), 472-9.

Kouzarides, T. (2007), 'Chromatin modifications and their function', *Cell,* 128 (4), 693-705.

Kress, C., Thomassin, H., and Grange, T. (2006), 'Active cytosine demethylation triggered by a nuclear receptor involves DNA strand breaks', *Proc Natl Acad Sci U S A,* 103 (30), 11112-7.

Kubota, Y., et al. (1996), 'Reconstitution of DNA base excision-repair with purified human proteins: interaction between DNA polymerase beta and the XRCC1 protein', *EMBO J,* 15 (23), 6662-70.

Kumar, R., et al. (2013), 'AID stabilizes stem-cell phenotype by removing epigenetic memory of pluripotency genes', *Nature*.

Kunz, C., et al. (2009), 'Base excision by thymine DNA glycosylase mediates DNA-directed cytotoxicity of 5-fluorouracil', *PLoS Biol,* 7 (4), e91.

Lander, E. S., et al. (2001), 'Initial sequencing and analysis of the human genome', *Nature,* 409 (6822), 860-921.

Larijani, M., et al. (2005), 'Methylation protects cytidines from AID-mediated deamination', *Mol Immunol,* 42 (5), 599-604.

Law, J. A. and Jacobsen, S. E. (2010), 'Establishing, maintaining and modifying DNA methylation patterns in plants and animals', *Nat Rev Genet,* 11 (3), 204-20.

Le May, N., et al. (2012), 'XPG and XPF endonucleases trigger chromatin looping and DNA demethylation for accurate expression of activated genes', *Molecular cell,* 47 (4), 622-32.

Le May, N., et al. (2010), 'NER factors are recruited to active promoters and facilitate chromatin modification for transcription in the absence of exogenous genotoxic attack', *Mol Cell,* 38 (1), 54-66.

Lee, H. J., Hore, T. A., and Reik, W. (2014), 'Reprogramming the methylome: erasing memory and creating diversity', *Cell Stem Cell,* 14 (6), 710-9.

Lehnertz, B., et al. (2003), 'Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin', *Current biology : CB,* 13 (14), 1192-200.

Lens, Z., et al. (2011), 'Purification of SUMO-1 modified IkappaBalpha and complex formation with NF-kappaB', *Protein expression and purification,* 80 (2), 211-6.

Lepikhov, K., et al. (2010), 'DNA methylation reprogramming and DNA repair in the mouse zygote', *Int J Dev Biol,* 54 (11-12), 1565-74.

Li, E., Bestor, T. H., and Jaenisch, R. (1992), 'Targeted mutation of the DNA methyltransferase gene results in embryonic lethality', *Cell,* 69 (6), 915-26.

Li, Y. Q., et al. (2007), 'Association of Dnmt3a and thymine DNA glycosylase links DNA methylation with base-excision repair', *Nucleic Acids Res,* 35 (2), 390-400.

Lindahl, T., Karran, P., and Wood, R. D. (1997), 'DNA excision repair pathways', *Curr Opin Genet Dev,* 7 (2), 158-69.

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

Liu, C. K., Hsu, C. A., and Abbott, M. T. (1973), 'Catalysis of three sequential dioxygenase reactions by thymine 7-hydroxylase', *Archives of biochemistry and biophysics,* 159 (1), 180-7.

Liutkeviciute, Z., et al. (2009), 'Cytosine-5-methyltransferases add aldehydes to DNA', *Nat Chem Biol,* 5 (6), 400-2.

Lock, L. F., Takagi, N., and Martin, G. R. (1987), 'Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation', *Cell,* 48 (1), 39-46.

Lomeli, H. and Vazquez, M. (2011), 'Emerging roles of the SUMO pathway in development', *Cellular and molecular life sciences : CMLS,* 68 (24), 4045-64.

Lucarelli, M., et al. (2001), 'The dynamics of myogenin site-specific demethylation is strongly correlated with its expression and with muscle differentiation', *J Biol Chem,* 276 (10), 7500-6.

Macleod, D., et al. (1994), 'Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island', *Genes & development,* 8 (19), 2282-92.

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

Maiti, A., Morgan, M. T., and Drohat, A. C. (2009), 'Role of two strictly conserved residues in nucleotide flipping and N-glycosylic bond cleavage by human thymine DNA glycosylase', *The Journal of biological chemistry,* 284 (52), 36680-8.

Maiti, A., et al. (2008), 'Crystal structure of human thymine DNA glycosylase bound to DNA elucidates sequence-specific mismatch recognition', *Proceedings of the National Academy of Sciences of the United States of America,* 105 (26), 8890-5.

Meissner, A., et al. (2008), 'Genome-scale DNA methylation maps of pluripotent and differentiated cells', *Nature,* 454 (7205), 766-70.

Mencia, M. and de Lorenzo, V. (2004), 'Functional transplantation of the sumoylation machinery into Escherichia coli', *Protein expression and purification,* 37 (2), 409-18.

Messerschmidt, D. M., Knowles, B. B., and Solter, D. (2014), 'DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos', *Genes & development,* 28 (8), 812-28.

Metivier, R., et al. (2008), 'Cyclical DNA methylation of a transcriptionally active promoter', *Nature,* 452 (7183), 45-50.

Mikkelsen, T. S., et al. (2008), 'Dissecting direct reprogramming through integrative genomic analysis', *Nature,* 454 (7200), 49-55.

Millar, C. B., et al. (2002), 'Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice', *Science,* 297 (5580), 403-5.

Mohan, R. D., et al. (2007), 'SUMO-1-dependent allosteric regulation of thymine DNA glycosylase alters subnuclear localization and CBP/p300 recruitment', *Molecular and cellular biology,* 27 (1), 229-43.

Mohan, R. D., et al. (2010), 'Opposing regulatory roles of phosphorylation and acetylation in DNA mispair processing by thymine DNA glycosylase', *Nucleic acids research,* 38 (4), 1135-48.

Mohandas, T., Sparkes, R. S., and Shapiro, L. J. (1981), 'Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation', *Science,* 211 (4480), 393-6.

Mohn, F. and Schubeler, D. (2009), 'Genetics and epigenetics: stability and plasticity during cellular differentiation', *Trends Genet,* 25 (3), 129-36.

Morgan, H. D., et al. (2004), 'Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming', *J Biol Chem,* 279 (50), 52353-60.

Morgan, H. D., et al. (2005), 'Epigenetic reprogramming in mammals', *Hum Mol Genet,* 14 Spec No 1, R47-58.

Morrison, J. R., et al. (1996), 'Apolipoprotein B RNA editing enzyme-deficient mice are viable despite alterations in lipoprotein metabolism', *Proc Natl Acad Sci U S A,* 93 (14), 7154-9.

Motorin, Y., Lyko, F., and Helm, M. (2010), '5-methylcytosine in RNA: detection, enzymatic formation and biological functions', *Nucleic acids research,* 38 (5), 1415-30.

Muller, U., et al. (2014), 'TET-mediated oxidation of methylcytosine causes TDG or NEIL glycosylase dependent gene reactivation', *Nucleic acids research,* 42 (13), 8592-604.

Muramatsu, M., et al. (2000), 'Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme', *Cell,* 102 (5), 553-63.

Nabel, C. S., et al. (2012), 'AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation', *Nat Chem Biol,* 8 (9), 751-8.

Neddermann, P. and Jiricny, J. (1993), 'The purification of a mismatch-specific thymine-DNA glycosylase from HeLa cells', *The Journal of biological chemistry,* 268 (28), 21218-24.

Neddermann, P., et al. (1996), 'Cloning and expression of human G/T mismatch-specific thymine-DNA glycosylase', *The Journal of biological chemistry,* 271 (22), 12767-74.

Niehrs, C. and Schafer, A. (2012), 'Active DNA demethylation by Gadd45 and DNA repair', *Trends in cell biology,* 22 (4), 220-7.

O'Brien, S. P. and DeLisa, M. P. (2012), 'Functional reconstitution of a tunable E3-dependent sumoylation pathway in Escherichia coli', *PLoS One,* 7 (6), e38671.

Okano, M., et al. (1999), 'DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development', *Cell,* 99 (3), 247-57.

Ooi, S. K. and Bestor, T. H. (2008), 'The colorful history of active DNA demethylation', *Cell,* 133 (7), 1145-8.

Ooi, S. K., et al. (2007), 'DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA', *Nature,* 448 (7154), 714-7.

Oswald, J., et al. (2000), 'Active demethylation of the paternal genome in the mouse zygote', *Curr Biol,* 10 (8), 475-8.

Pastor, W. A., Aravind, L., and Rao, A. (2013), 'TETonic shift: biological roles of TET proteins in DNA demethylation and transcription', *Nature reviews. Molecular cell biology,* 14 (6), 341-56.

Pfaffeneder, T., et al. (2011), 'The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA', *Angew Chem Int Ed Engl,* 50 (31), 7008-12.

Pfaffeneder, T., et al. (2014), 'Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA', *Nature chemical biology,* 10 (7), 574-81.

Pinney, S. E. (2014), 'Mammalian Non-CpG Methylation: Stem Cells and Beyond', *Biology,* 3 (4), 739-51.

Popp, C., et al. (2010), 'Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency', *Nature,* 463 (7284), 1101-5.

Probst, A. V., Dunleavy, E., and Almouzni, G. (2009), 'Epigenetic inheritance during the cell cycle', *Nature reviews. Molecular cell biology,* 10 (3), 192-206.

Rai, K., et al. (2008), 'DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45', *Cell,* 135 (7), 1201-12.

Raiber, E. A., et al. (2015), '5-Formylcytosine alters the structure of the DNA double helix', *Nature structural & molecular biology,* 22 (1), 44-9.

Raiber, E. A., et al. (2012), 'Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase', *Genome Biol,* 13 (8), R69.

Rainsford, K. D. (2013), 'Ibuprofen: from invention to an OTC therapeutic mainstay', *International journal of clinical practice. Supplement,* (178), 9-20.

Rangam, G., et al. (2012), 'AID enzymatic activity is inversely proportional to the size of cytosine C5 orbital cloud', *PLoS One,* 7 (8), e43279.

Reik, W. (2007), 'Stability and flexibility of epigenetic gene regulation in mammalian development', *Nature,* 447 (7143), 425-32.

Robertson, A. B., et al. (2009), 'DNA repair in mammalian cells: Base excision repair: the long and short of it', *Cell Mol Life Sci,* 66 (6), 981-93.

Robertson, K. D. (2005), 'DNA methylation and human disease', *Nat Rev Genet,* 6 (8), 597-610.

Saito, Y., et al. (2011), 'Embryonic lethality in mice lacking mismatch-specific thymine DNA glycosylase is partially prevented by DOPS, a precursor of noradrenaline', *Tohoku J Exp Med,* 226 (1), 75-83.

Saluz, H. P., Jiricny, J., and Jost, J. P. (1986), 'Genomic sequencing reveals a positive correlation between the kinetics of strand-specific DNA demethylation of the overlapping estradiol/glucocorticoid-receptor binding sites and the rate of avian vitellogenin mRNA synthesis', *Proc Natl Acad Sci U S A,* 83 (19), 7167-71.

Sampson, D. A., Wang, M., and Matunis, M. J. (2001), 'The small ubiquitin-like modifier-1 (SUMO-1) consensus sequence mediates Ubc9 binding and is essential for SUMO-1 modification', *J Biol Chem,* 276 (24), 21664-9.

Santos, F. and Dean, W. (2004), 'Epigenetic reprogramming during early development in mammals', *Reproduction,* 127 (6), 643-51.

Santos, F., et al. (2002), 'Dynamic reprogramming of DNA methylation in the early mouse embryo', *Dev Biol,* 241 (1), 172-82.

Santos, F., et al. (2013), 'Active demethylation in mouse zygotes involves cytosine deamination and base excision repair', *Epigenetics & chromatin,* 6 (1), 39.

Schar, P. and Fritsch, O. (2010), 'DNA repair and the control of DNA methylation', *Prog Drug Res,* 67, 51-68.

Scharer, O. D. (2003), 'Chemistry and biology of DNA repair', *Angew Chem Int Ed Engl,* 42 (26), 2946-74.

Schiesser, S., et al. (2012), 'Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing', *Angewandte Chemie,* 51 (26), 6516-20.

Schmitz, K. M., et al. (2009), 'TAF12 recruits Gadd45a and the nucleotide excision repair complex to the promoter of rRNA genes leading to active DNA demethylation', *Mol Cell,* 33 (3), 344-53.

Schwartz, D. C. and Hochstrasser, M. (2003), 'A superfamily of protein tags: ubiquitin, SUMO and related modifiers', *Trends Biochem Sci,* 28 (6), 321-8.

Seisenberger, S., et al. (2013), 'Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences,* 368 (1609), 20110330.

Seisenberger, S., et al. (2012), 'The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells', *Mol Cell,* 48 (6), 849-62.

Shen, J. C., Rideout, W. M., 3rd, and Jones, P. A. (1992), 'High frequency mutagenesis by a DNA methyltransferase', *Cell,* 71 (7), 1073-80.

--- (1994), 'The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA', *Nucleic acids research,* 22 (6), 972-6.

Shen, L., et al. (2013), 'Genome-wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics', *Cell*.

Simonsson, S. and Gurdon, J. (2004), 'DNA demethylation is necessary for the epigenetic reprogramming of somatic cell nuclei', *Nat Cell Biol,* 6 (10), 984-90.

Sjolund, A. B., Senejani, A. G., and Sweasy, J. B. (2013), 'MBD4 and TDG: multifaceted DNA glycosylases with ever expanding biological roles', *Mutation research,* 743-744, 12-25.

Smallwood, S. A., et al. (2011), 'Dynamic CpG island methylation landscape in oocytes and preimplantation embryos', *Nat Genet,* 43 (8), 811-4.

Smiley, J. A., et al. (2005), 'Genes of the thymidine salvage pathway: thymine-7-hydroxylase from a Rhodotorula glutinis cDNA library and iso-orotate decarboxylase from Neurospora crassa', *Biochim Biophys Acta,* 1723 (1-3), 256-64.

Smith, Z. D. and Meissner, A. (2013), 'DNA methylation: roles in mammalian development', *Nat Rev Genet,* 14 (3), 204-20.

Smith, Z. D., et al. (2012), 'A unique regulatory phase of DNA methylation in the early mammalian embryo', *Nature,* 484 (7394), 339-44.

Smith, Z. D., et al. (2014), 'DNA methylation dynamics of the human preimplantation embryo', *Nature,* 511 (7511), 611-5.

Sneader, W. (2000), 'The discovery of aspirin: a reappraisal', *BMJ,* 321 (7276), 1591-4.

Song, C. X., et al. (2013a), 'Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming', *Cell*.

Song, S. J., et al. (2013b), 'MicroRNA-antagonism regulates breast cancer stemness and metastasis via TET-family-dependent chromatin remodeling', *Cell,* 154 (2), 311-24.

Song, S. J., et al. (2013c), 'The oncogenic microRNA miR-22 targets the TET2 tumor suppressor to promote hematopoietic stem cell self-renewal and transformation', *Cell Stem Cell,* 13 (1), 87-101.

Spruijt, C. G., et al. (2013), 'Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives', *Cell*.

Stadler, M. B., et al. (2011), 'DNA-binding factors shape the mouse methylome at distal regulatory regions', *Nature*.

Steinacher, R. and Schar, P. (2005), 'Functionality of human thymine DNA glycosylase requires SUMO-regulated changes in protein conformation', *Curr Biol,* 15 (7), 616-23.

Suetake, I., et al. (2004), 'DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction', *J Biol Chem,* 279 (26), 27816-23.

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

Takahashi, K., et al. (2007), 'Induction of pluripotent stem cells from adult human fibroblasts by defined factors', *Cell,* 131 (5), 861-72.

Thomassin, H., et al. (2001), 'Glucocorticoid-induced DNA demethylation and gene memory during development', *EMBO J,* 20 (8), 1974-83.

Thomson, J. P., et al. (2010), 'CpG islands influence chromatin structure via the CpG-binding protein Cfp1', *Nature,* 464 (7291), 1082-6.

Uchimura, Y., Nakao, M., and Saitoh, H. (2004a), 'Generation of SUMO-1 modified proteins in E. coli: towards understanding the biochemistry/structural biology of the SUMO-1 pathway', *FEBS Lett,* 564 (1-2), 85-90.

Uchimura, Y., et al. (2004b), 'Overproduction of eukaryotic SUMO-1- and SUMO-2-conjugated proteins in Escherichia coli', *Anal Biochem,* 331 (1), 204-6.

Um, S., et al. (1998), 'Retinoic acid receptors interact physically and functionally with the T:G mismatch-specific thymine-DNA glycosylase', *J Biol Chem,* 273 (33), 20728-36.

Ushijima, T., et al. (2003), 'Fidelity of the methylation pattern and its variation in the genome', *Genome Res,* 13 (5), 868-74.

Vairapandi, M. and Duker, N. J. (1993), 'Enzymic removal of 5-methylcytosine from DNA by a human DNA-glycosylase', *Nucleic Acids Res,* 21 (23), 5323-7.

Vincent, J. J., et al. (2013), 'Stage-Specific Roles for Tet1 and Tet2 in DNA Demethylation in Primordial Germ Cells', *Cell Stem Cell*.

Waters, T. R., et al. (1999), 'Human thymine DNA glycosylase binds to apurinic sites in DNA but is displaced by human apurinic endonuclease 1', *The Journal of biological chemistry,* 274 (1), 67-74.

Weber, A. R., Schuermann, D., and Schar, P. (2014), 'Versatile recombinant SUMOylation system for the production of SUMO-modified protein', *PLoS One,* 9 (7), e102157.

Wiebauer, K. and Jiricny, J. (1989), 'In vitro correction of G.T mispairs to G.C pairs in nuclear extracts from human cells', *Nature,* 339 (6221), 234-6.

Wilks, A. F., et al. (1982), 'Estrogen induces a demethylation at the 5' end region of the chicken vitellogenin gene', *Proc Natl Acad Sci U S A,* 79 (14), 4252-5.

Williams, K., et al. (2011), 'TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity', *Nature*.

Williams, R. T. and Wang, Y. (2012), 'A density functional theory study on the kinetics and thermodynamics of N-glycosidic bond cleavage in 5-substituted 2'-deoxycytidines', *Biochemistry,* 51 (32), 6458-62.

Wilson, D. M., 3rd and Barsky, D. (2001), 'The major human abasic endonuclease: formation, consequences and repair of abasic lesions in DNA', *Mutat Res,* 485 (4), 283-307.

Wossidlo, M., et al. (2010), 'Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes', *EMBO J,* 29 (11), 1877-88.

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

Wu, D., et al. (2014), 'Uracil-DNA glycosylase is involved in DNA demethylation and required for embryonic development in the zebrafish embryo', *The Journal of biological chemistry,* 289 (22), 15463-73.

Wu, H. and Zhang, Y. (2014), 'Reversing DNA methylation: mechanisms, genomics, and biological functions', *Cell,* 156 (1-2), 45-68.

Wu, H., et al. (2011), 'Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells', *Nature.*

Wu, S. C. and Zhang, Y. (2010), 'Active DNA demethylation: many roads lead to Rome', *Nat Rev Mol Cell Biol,* 11 (9), 607-20.

Wutz, A. (2011), 'Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation', *Nature reviews. Genetics,* 12 (8), 542-53.

Xu, J., Du, Y., and Deng, H. (2015), 'Direct Lineage Reprogramming: Strategies, Mechanisms, and Applications', *Cell Stem Cell,* 16 (2), 119-34.

Xu, Y., et al. (2011), 'Genome-wide Regulation of 5hmC, 5mC, and Gene Expression by Tet1 Hydroxylase in Mouse Embryonic Stem Cells', *Mol Cell.*

Xu, Y., et al. (2012), 'Tet3 CXXC domain and dioxygenase activity cooperatively regulate key genes for Xenopus eye and neural development', *Cell,* 151 (6), 1200-13.

Yamaguchi, S., et al. (2013), 'Dynamics of 5-methylcytosine and 5-hydroxymethylcytosine during germ cell reprogramming', *Cell Res,* 23 (3), 329-39.

Yamaguchi, S., et al. (2012), 'Tet1 controls meiosis by regulating meiotic gene expression', *Nature,* 492 (7429), 443-7.

Yamanaka, S. and Blau, H. M. (2010), 'Nuclear reprogramming to a pluripotent state by three approaches', *Nature,* 465 (7299), 704-12.

Yan, M. S., Matouk, C. C., and Marsden, P. A. (2010), 'Epigenetics of the vascular endothelium', *Journal of applied physiology,* 109 (3), 916-26.

Yang, W. (2006), 'Poor base stacking at DNA lesions may initiate recognition by many repair proteins', *DNA Repair (Amst),* 5 (6), 654-66.

Yebra, M. J. and Bhagwat, A. S. (1995), 'A cytosine methyltransferase converts 5-methylcytosine in DNA to thymine', *Biochemistry,* 34 (45), 14752-7.

Yin, R., et al. (2013), 'Ascorbic acid enhances Tet-mediated 5-methylcytosine oxidation and promotes DNA demethylation in mammals', *Journal of the American Chemical Society,* 135 (28), 10396-403.

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

Zhang, H., et al. (2010a), 'TET1 is a DNA-binding protein that modulates DNA methylation and gene transcription via hydroxylation of 5-methylcytosine', *Cell research,* 20 (12), 1390-3.

Zhang, L., et al. (2012), 'Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA', *Nat Chem Biol,* 8 (4), 328-30.

Zhang, P., et al. (2013), 'Ten-eleven translocation (Tet) and thymine DNA glycosylase (TDG), components of the demethylation pathway, are direct targets of miRNA-29a', *Biochemical and biophysical research communications,* 437 (3), 368-73.

Zhang, Y., et al. (2010b), 'Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail', *Nucleic acids research,* 38 (13), 4246-53.

Zhao, J. (2007), 'Sumoylation regulates diverse biological processes', *Cell Mol Life Sci,* 64 (23), 3017-33.

Zhu, B., et al. (2000a), '5-Methylcytosine DNA glycosylase activity is also present in the human MBD4 (G/T mismatch glycosylase) and in a related avian sequence', *Nucleic Acids Res,* 28 (21), 4157-65.

Zhu, B., et al. (2000b), '5-methylcytosine-DNA glycosylase activity is present in a cloned G/T mismatch DNA glycosylase associated with the chicken embryo DNA demethylation complex', *Proc Natl Acad Sci U S A,* 97 (10), 5135-9.

Zhu, H., et al. (2006), 'Lsh is involved in de novo methylation of DNA', *The EMBO journal,* 25 (2), 335-45.

Zhu, J. K. (2009), 'Active DNA demethylation mediated by DNA glycosylases', *Annu Rev Genet,* 43, 143-66.

Appendix

Biochemical Reconstitution of TET1-TDG-BER Dependent Active DNA Demethylation Reveals a Highly Coordinated Mechanism

Alain R. Weber[1], Claudia Krawczyk[1], Adam B. Robertson[2], Anna Kusnierczyk[3], Cathrine B. Vågbø[3], David Schuermann[1], Arne Klungland[2] and Primo Schär[1]*

[1] Department of Biomedicine, University of Basel, Basel, Switzerland

[2] Institute of Medical Microbiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway

[3] Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

* To whom correspondence should be addressed. Tel: +41 61 267 0767; Fax: +41 61 267 3566; Email: primo.schaer@unibas.ch

ABSTRACT

Cytosine methylation in CpG dinucleotides is an epigenetic DNA modification established and maintained by DNA methyltransferases. Recent research has substantiated the coexistence of active DNA demethylation processes, facilitating dynamic regulation of methylation states. One postulated pathway of active demethylation operates through TET hydroxylases that oxidize 5-methylcytosine (5mC) to generate substrates for TDG dependent base excision repair (BER), which then replaces 5mC with C. Although mechanistically plausible, direct evidence for a functional and productive coupling of TET with BER has never been presented. Here, we show that TET1 and TDG interact to constitute a 5mC demethylase and proof by biochemical reconstitution that the TET/TDG-BER system is capable of productive DNA demethylation. We also show that intrinsic features of this mechanism assure a sequential demethylation of symmetrically methylated CpGs, thereby avoiding DNA double-strand break formation, but can mutate CpGs if 5mC oxidation in one strand coincides with 5mC deamination in the other.

DNA methylation in mammals occurs almost entirely at the C5 position of cytosines (5-methylcytosine, 5mC) and is found predominantly within CpG dinucleotides, affecting 60-90% of such sites[1]. Modulating chromatin states and thereby transcriptional activity and genome stability, DNA methylation plays an important epigenetic role in various biological processes[2]. It was long viewed as a largely static DNA base modification but recent research has shown that, under specific circumstances, DNA methylation undergoes dynamic regulation. This is best illustrated by its genome-wide erasure during early embryonic development[3-5] or in maturing primordial germ cells[6]. Locus directed DNA demethylation has also been observed in somatic cells upon triggering transcriptional activation in various ways[7-9]. Both passive and active pathways of DNA demethylation were proposed to operate in these contexts but the mechanisms underlying active demethylation, in particular, have remained controversial[10].

Recent evidence strengthens the case for an involvement of the ten eleven-translocation (TET) family of dioxygenases[11]. TET proteins oxidize 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), all of which have been implicated as intermediates of DNA demethylation[12-15]. Considering active mechanisms, 5fC and 5caC appear of particular relevance as these bases are substrates for the thymine DNA glycosylase (TDG), a DNA repair protein previously characterized as an enzyme excising a variety of cytosine and 5mC base derivatives from DNA[12,16-18]. TET and TDG together thus constitute catalytic activities capable of oxidation and removal of 5mC in DNA. This biochemical reasoning is supported by the phenotype of TET[19-21] and TDG knockout mice[22-24] as well as embryonic stem cells (ESCs), all showing aberrations in DNA methylation[22,25-28].

The evident mechanistic view is thus that TET and TDG cooperate to initiate active DNA demethylation by sequential oxidation and excision of 5mC in DNA. The resulting baseless site (AP-site) is then repaired by the DNA base excision repair (BER) system to restore the original unmethylated DNA sequence. An engagement of the core BER pathway would imply that the AP-site is first recognized and incised by an AP endonuclease (i.e. APE1), which generates a DNA single-strand break (SSB) that attracts and activates Poly [ADP-ribose] polymerase 1 (PARP1). PARP1 then facilitates the engagement of X-ray repair cross-complementing protein 1 (XRCC1), DNA ligase 3 (LIG3) and DNA polymerase β (POLβ) for insertion of a new nucleotide and ligation of the DNA nick[29]. Although this mechanism is plausible and widely anticipated, there is in fact little evidence supporting a direct link between TET and BER; a productive action of TET with the BER system on a 5mC

substrate has not been shown, nor have the basic mechanistic features of such a process been addressed.

The aim of this study was therefore to reconstitute the full DNA demethylation system *in vitro* and to address specific properties of the DNA transactions involved. We investigated physical and functional interactions between TET1 and TDG and tested the hypothesis that methylated DNA substrates can be fully converted to unmethylated DNA through oxidation and BER of 5mC. With the *in vitro* DNA demethylation system established, we addressed the strand-specificity of the reaction, whether symmetrically modified CpGs can be demethylated without DNA fragmentation, and how complex lesions such as the simultaneous deamination of a symmetrically opposite 5mC within a CpG dinucleotide do affect the demethylation outcome. The data proof full functionality of a TET1-TDG-BER based DNA demethylation system on hemi- and fully methylated DNA and show that the molecular transactions involved are coordinated in a manner that avoids DNA DSB formation but can generate mutations if deamination and demethylation events coincide within a CpG.

RESULTS

TET1 and TDG interact physically

The model of TET-TDG mediated oxidative DNA demethylation anticipates a coupled action of TET and TDG to facilitate an efficient but coordinated removal of 5mC. To address the mode of cooperation of TET and TDG, we examined a potential physical interaction of the two enzymes. We first co-expressed a full length C-terminally 6His-tagged TET1 (TET1-His6) with a C-terminally GST-tagged TDG (TDG-GST). Although co-expression with TDG positively affected full-length TET1 expression, enrichment of TET1-His6 via Ni-NTA chromatography yielded prominent fragments migrating at sizes corresponding to 140 – 150 kDa, 90 kDa and 60 – 70 kDa. Size-fractionation of the eluate by gel filtration (Fig. 1a) showed that specific fragments of TET1-His6 (140 – 150 kDa, 60 – 70 kDa) co-eluted with full-length TDG-GST in high molecular weight fractions (200 – 600 kDa). This occurred at high ionic strength (500 mM NaCl), indicating that the two proteins form a stable complex. Given the enrichment on the Ni-NTA resin, we concluded that the TET1 fragments isolated correspond to the C-terminus of TET1 comprising the catalytic domain (Fig. 1a). A prominent 90 kDa fragment of TET1-

His6, however, eluted in lower molecular weight fractions (90 – 200 kDa) showing only a partial overlap with TDG-GST, suggesting a weak interaction with TDG.

To confirm and further characterize the TET1-TDG interaction, we used the yeast two-hybrid system. The TET1 protein was split into 4 fragments spanning the entire protein sequence (Fig. 1b) and fused to the Gal4 binding domain (bait). The TET1 constructs were then co-expressed separately with TDG fused to the Gal4 activation domain (prey). Growth on selective medium, allowing colony formation only if the reporter gene is activated by direct interaction of bait and prey proteins, indicated an interaction between TDG and the TET1 fragments 2 and 4, the former harboring the CXXC domain in the N-terminus, the latter spanning the C-terminal catalytic domain. These results indicate that the interaction between TET1 and TDG is mediated by domains located in the TET1 N-terminus (amino acids (aa) 397 – 931) as well as the catalytic domain (aa 1367 – 2057) (Fig. 1b). We next performed co-precipitations from lysates of *E.coli* cells co-expressing TDG-GST with either a His6-labeled TET1 N-terminal fragment (TET1$_N$; aa 301 – 1366) or the TET1 catalytic domain (TET1$_{CD}$; aa 1367 – 2057) (Fig. 1c). After enrichment of His6-TET1$_N$ and His6-TET1$_{CD}$ using a Ni-NTA resin, TDG-GST was clearly co-eluting in the bound fraction of both TET1 fragments (Fig. 1c Ni-NTA). The outcome was the same when we enriched for TDG-GST, both TET1 fragments could be detected in the bound fraction after GST affinity purification (Fig. 1c, GST). The results of all protein interaction assays are consistent and lead us to conclude that TET1 and TDG physically interact through domains located in the TET1 N-terminus (aa 397 – 931) as well as the TET1 catalytic domain (aa 1367 – 2057).

A TET1$_{CD}$/TDG protein complex releases 5mC and 5hmC via oxidized intermediates

To examine the catalytic activity of the TET/TDG complex, we decided to work with TET1$_{CD}$, which forms a complex with TDG as described above. We thus co-expressed the catalytic domain of TET1 (His6-TET1$_{CD}$) and TDG-GST as well as combinations of the respective catalytic-dead variants (His6-TET1$_{CD}$Δcat (H1652Y; D1654A) with TDG-GST; His6-TET1$_{CD}$ with TDGΔcat-GST (N151A)) in *E.coli* and enriched the complexes by Ni-NTA chromatography (Supplementary Fig. S1). The catalytic activities were then measured by means of a standardized base release assay[30] with 2 fluorescein labeled synthetic 60 bp substrates containing a single 5mC or 5hmC. Incubation of the enriched His6-TET1$_{CD}$/TDG-GST with both these DNA substrates (1 h at 37°C) generated a substantial amount of

DNA incisions at the position of the modified cytosines (Fig. 2a). This 5mC/5hmC excision activity was not detectable when either of the two proteins was mutated at its catalytic site. Hence, the excision of 5mC and 5hmC from DNA was dependent on the catalytic activities of both TET1 and TDG.

To confirm that the base released in this assay was indeed oxidized 5mC or 5hmC, we examined the 5mC oxidation products generated by His6-TET1$_{CD}$ purified in the absence of TDG. Purification of His6-TET1$_{CD}$ by Ni-NTA and ion exchange chromatography yielded two prominent protein fragments, both corresponding to TET1$_{CD}$ (Supplementary Fig. S2). Mass spectrometric analysis identified the smaller fragment as an N-terminal truncation lacking ~240 amino acids comprising the highly conserved Cys-rich domain, which was shown to be essential for the catalytic activity[31]. We tested the catalytic activity of this His6-TET1$_{CD}$ preparation in a plasmid based DNA oxidation assay (Fig. 2b). To this end, we *in vitro* methylated highly pure plasmid DNA using the M.SssI CpG methyltransferase, creating fully CpG methylated plasmids (200 pmol mCpG sites / µg DNA) and then reacted the methylated DNA (200 ng; 40 pmol mCpGs) with purified His6-TET1$_{CD}$ (500 ng; 6 pmol) at 37°C for 1 hour. Immunoblot analyses with specific antibodies against 5mC, 5hmC, 5fC, and 5caC then showed that all detectable 5mC was fully oxidized to 5hmC, 5fC, and 5caC under these conditions. Hence, the His6-TET1$_{CD}$ efficiently carried out the predicted 5mC oxidation steps *in vitro* with the conversion of 5mC to 5hmC representing the most efficient step (Fig. 2b).

We next used separately purified TET and TDG proteins to reconstitute the 5mC release. To allow for preformation of the TET/TDG complex, we mixed His6-TET1$_{CD}$ with His6-TDG (Supplementary Fig. S2) at a molar ration of 2:1 (most active ratio as evaluated by titration) before addition of DNA substrates. A two-fold molar excess of TET/TDG (50 nM) over substrate DNA (25 nM) and incubation of 60 min at 37°C resulted in efficient release of both 5mC and 5hmC (Fig. 2c). Notably, 5mC was nearly as efficiently excised as 5hmC, indicating that TDG turnover[32] rather than 5mC oxidation was rate limiting in this setup. As expected, 5caC was processed most efficiently. While it is fair to assume that TET1 was not required for 5caC release, it is noteworthy that its presence did not appear to interfere with TDG activity.

Together, these results establish that TET1$_{CD}$ and TDG activities cooperate to efficiently excise 5mC from DNA, thereby generating alkaline labile AP-sites in DNA.

TET stabilizes TDG activity

To address whether TET1 and TDG cooperate at the level of their catalytic activities, we examined the effect of TDG on the efficiency of 5mC oxidation by $TET1_{CD}$. In this setup, we used purified, catalysis deficient TDG to limit excision of TET-generated 5fC and 5caC. We thus pre-incubated His6-$TET1_{CD}$ (50 nM) with or without His6-TDGΔcat (25 nM), then added 5mC substrate (25 nM), stopped the reactions at different time points, and monitored the presence of 5fC and 5caC in the recovered product DNA by digestion with purified catalytic proficient TDG (250 nM). This showed that the presence of TDGΔcat had a minor but not statistically significant effect on 5mC oxidation by His6-TET_{CD} (Fig. 3a). Further evaluation of a potential effect of TDG on TET1 in a plasmid oxidation assay followed by LC/MS/MS analysis essentially supported this finding. We did, however, observe a slightly reduced conversion of 5hmC to 5fC and 5caC in presence of TDGΔcat under these multiple turnover conditions (Fig. 3b). This may indicate that TDG is rate limiting in the turnover of the $TET1_{CD}$/TDG complex. Notably, the lack of detectable 5caC in the reactions with TDGΔcat is explained by the residual activity of the TDG catalytic mutant towards optimally suited substrates such as 5caC (see Fig. 2c)[30]. From these results, we conclude that, under single turnover conditions, purified His6-$TET1_{CD}$ carries out all three 5mC oxidation steps equally efficiently irrespective of whether it is in a complex with TDG or not, while its efficiency in 5hmC oxidation is slightly reduced under multiple turnover conditions. Vice versa, the presence of a two-fold molar excess of $TET1_{CD}$ had a positive effect on TDG activity when compared to BSA added to the same molarity. The overall efficiency of base excision from 5caC•G (equimolar enzyme/substrate) and T•G substrates (five-fold excess of substrate) was enhanced, whereas the initial rate of excision by TDG was not affected (Fig. 3c). Hence, the interaction of $TET1_{CD}$ and TDG does not significantly impact on the $TET1_{CD}$ catalysis, but the presence of the $TET1_{CD}$ stabilized the activity of TDG.

In vitro reconstitution of TET/TDG-BER mediated active DNA demethylation

The current model TET1/TDG mediated active DNA demethylation postulates the engagement BER in the accurate restoration of the unmethylated DNA sequence following 5fC or 5caC excision. To formally proof the functionality of such a pathway and to provide a tool to investigate mechanistic features, we aimed to reconstitute the entire process of active DNA demethylation with defined

components. In addition to TET1$_{CD}$ and TDG, we thus purified to near homogeneity the enzymes of the core BER pathway[33], APE1, POLβ and XRCC1/LIG3 (Supplementary Fig. S2). Using a 60 bp substrate containing a single 5mC, we first performed demethylation in step-by-step reactions to monitor the DNA intermediates generated. The combined action of His6-TET1$_{CD}$ and His6-TDG generated an AP-site cleavable either chemically by NaOH or enzymatically by APE1 to generate a 23 nt oligonucleotide with or without a 3'-phosphate, respectively (Fig. 4a, lanes 3, 4). Following strand incision by APE1, POLβ was able to insert a dCMP, thus generating a 24 nt oligonucleotide as the main product (Fig. 4a, lane 5). Addition of either T4 ligase or an XRCC1/LIG3 complex then efficiently ligated the nicked intermediate, restoring a continuous 60mer DNA fragment (Fig. 4a, lanes 6 and 7). The nearly complete re-ligation confirmed the efficient removal of the 5'-dRP remains of the cleaved AP-site by POLβ. The results establish that TET1 and TDG convert 5mC to DNA repair intermediates amenable to processing by the core BER system.

To test the accuracy of the reconstituted DNA demethylation process, we performed the reaction with a 5'-fluorescein labeled 59 bp DNA substrate presenting a hemi-methylated CpG dinucleotide within a recognition site for the *Hpa*II endonuclease (CCGG) (Fig. 4b). Due to its methylation sensitivity, *Hpa*II will not be able to cleave this substrate unless it undergoes successful and complete demethylation. We thus subjected the hemi-methylated substrate to demethylation by the reconstituted TET/TDG-BER system and examined the generation of a cleavable restriction site by digestion with *Hpa*II (Fig. 4b). As expected, the asymmetrically methylated substrate was fully resistant to *Hpa*II cleavage (Fig. 4b, lane 2). Incubation in the presence of the reconstituted DNA demethylation system, however, generated *Hpa*II digestible DNA products, indicating that the methylated DNA fragment was converted into an intact unmethylated fragment (Fig. 4b, lane 3). Together, these results proof that TET1/TDG mediated oxidation and excision of 5mC generates intermediates for BER, which then acts to efficiently restore the original DNA sequence in an unmethylated configuration.

Coordinated operation of TET/TDG-BER prevents DNA DSB formation during demethylation of symmetrically methylated CpGs

CpGs in mammalian DNA are mostly symmetrically methylated, generating a potential conflict for excision-repair mediated DNA demethylation; i.e. once started, a DNA demethylation event in one

DNA strand would have to be completed before another event starts at the symmetrically opposite

5mC, which would otherwise lead to the formation of a DNA DSB. We therefore asked if in a

symmetrically methylated CpG dinucleotide, demethylation events would generate DSBs or be

confined to one strand at a time. For this purpose, we generated three 60 bp DNA substrates with

either a fluorescein labeled bottom strand containing a single 5mC, a TexasRed labeled top strand

containing a single 5mC, or both strands labeled and presenting a symmetrically methylated CpG (Fig.

5a). Incubation of all these substrates with a two-fold molar excess of TET1$_{CD}$/TDG produced a solid

5mC release from both the bottom and the top strands, irrespective of whether the CpG was hemi- or

symmetrically methylated (Fig. 5a). Activities on top and bottom strands in hemi-methylated substrates

were similar, indicating the absence of sequence context effects in this setup (Fig. 5a, lanes 2, 4).

Notably, the same reaction conditions applied to the substrate with 5mC modifications on both strands

produced approximately half the amount of incised product on each DNA strand with the total activity

remaining constant (Fig. 5a, lane 6). These results show that TET1$_{CD}$/TDG can act on both strands on

a substrate containing a symmetrically methylated CpG.

To further investigate the demethylation events at symmetrically modified CpGs, we separated TDG

from TET1 activities and measured the kinetics of 5caC processing in the context of potentially arising

DNA demethylation intermediates. Using equimolar substrate and enzyme concentrations[34], we

evaluated substrates containing a 5caC on the labeled DNA strand opposite an unmodified C, a 5mC

or a 5hmC within the same CpG (Fig. 5b). Under the conditions applied, both initial rate and overall

5caC excision by TDG was not notably affected by the modification status of the symmetrically

opposite C (Fig. 5b). 5caC was processed with appreciable efficiency even in single-stranded DNA,

corroborating the high affinity of TDG for this substrate. The situation when 5caC arises in both

strands simultaneously is of particular interest as it raises the possibility that TDG initiated BER will

induce DNA DSBs. We thus evaluated the behavior of TDG in such a context, monitoring the release

of 5caC from both strands in a time course base release assay with a substrate carrying labels on both

strands. Similar to the combined activity of TET1/TDG on 5mC, the activity of TDG alone was

distributed evenly to both strands carrying the 5caC (Fig. 5c), amounting to approximately 50%

processing for each strand. The plateauing of single strand incision at 50% indicated that the

processing of one DNA strand by TDG inhibited base release from the other strand. This is a likely

consequence of TDGs tight interaction with AP-sites[30,35], the coordinated dissociation of which[32,36]

may favor completion of the repair process to initiation of an additional repair event at the opposite

strand. To test whether repair of a symmetrical demethylation intermediate is at all possible and can be achieved without the generation of DSBs, we used a substrate with a symmetrically 5caC modified *Hpa*II site in a reconstituted TDG-BER assay and analyzed the generation of cleavable *Hpa*II sites. Controls showed that both hemi- and symmetrically 5caC modified substrates were fully resistant to *Hpa*II cleavage (Fig. 5d, lane 4, 6). Incubation of the symmetrically 5caC modified DNA with the TDG-BER system, however, generated an appreciable amount of *Hpa*II cleavable product. Thus, the 5caCs in both DNA strands were promptly replaced with unmodified Cs (Fig. 5d, lane 8). Notably, symmetrical repair, which essentially requires the breaking of both strands in the process, was not associated with the generation of detectable levels of DSBs (Fig. 5d, lane 7), indicating that repair events at both strands proceeded in a sequential manner.

From these results, we conclude that DNA demethylation *in vitro* has no apparent strand and, hence sequence context preference. TET1/TDG is capable of initiating active DNA demethylation in both strands of a fully methylated CpG but the repair of symmetrically positioned 5caC modifications does not generate DNA DSBs. Once initiated in one strand, BER is completed before it restarts on the other strand, indicating that demethylation of symmetrically methylated CpGs occurs in a sequential manner protecting genome integrity.

Active DNA demethylation inhibits G•T repair at methylated CpGs

Another issue of BER mediated demethylation at symmetrically methylated CpGs is the potential collision with 5mC deamination. 5mC in genomic DNA is susceptible to spontaneous hydrolytic deamination[37] generating in a thymine paired with a guanine. Such G•T mismatches are recognized and excised also by TDG. Enzymatic deamination coupled to BER has also been considered as a mechanism of active DNA demethylation[7,23,38]; it would replace a 5mC with an unmodified C through a mutagenic intermediate. To investigate potential interferences between deamination and oxidation-induced DNA demethylation pathways, we evaluated G•T and G•5caC processing efficiencies in kinetic base release assays, using equimolar substrate/enzyme (His6-TDG) concentrations (25 nM). When provided on separate DNA molecules, TDG processed the G•T mismatch more efficiently than the 5caC substrate (Fig. 6a), showing that the mismatch is a preferred substrate as reported previously[16]. In a substrate where the G•5caC modification is next to a G•T mismatch within the same

CpG dinucleotide, reflecting a spontaneous deamination event on one strand while the other is being actively demethylated, however, TDG processes almost exclusively the 5caC, leaving the G•T mismatch untouched (Fig. 6b). Furthermore, the processing rate of 5caC was largely unaffected by the presence of the G•T mismatch, indicating that in this configuration 5caC is clearly the preferred substrate. The result was essentially the same when the modifications were inversed within the same double-stranded substrate (Supplementary Fig. S5), thus excluding DNA strand- or sequence context effects as an explanation for the preference for 5caC.

This strong preference of TDG for the non-mutagenic 5caC next to a pre-mutagenic G•T mismatch implies that TET/TDG mediated active DNA demethylation has a potential to mutate CpG dinucleotides if it coincides with a deamination event. To test this possibility, we used our fully reconstituted BER setup on a 59 bp substrate containing a G•5caC next to a G•T mismatch within an *Msc*I recognition site and analyzed the generation of mutant demethylation products by endonuclease digestion. 5caC directed sequential BER of this substrate would generate C to T mutations and thus create an *Msc*I restriction site if two or more nucleotides were incorporated during the DNA re-synthesis step (Fig. 6c). In the absence of the TDG-BER machinery, no *Msc*I cleavage products were detectable (Fig. 6c, lane 3). Full reconstitution of TDG-BER, however, generated a cleavable product, indicating that the 5caC was correctly replaced with a C but an A was incorporated opposite of T, thus manifesting the C to T transition and a loss of a CpG dinucleotide (Fig. 6c, lane 4).

DISCUSSION

Recent biochemical as well as biological evidence points towards an active DNA demethylation mechanism involving TET proteins as well as the DNA glycosylase TDG[12-14,16,17]. The current model suggests that DNA demethylation through this pathway occurs in a stepwise manner via TET catalyzed oxidation of 5mC to 5fC and 5caC that are then excised by TDG dependent BER to restore an unmethylated sequence. Despite the general acceptance of this pathway, experimental evidence that directly links TET activity to TDG and BER is missing and fundamental mechanistic questions have not been addressed. The data presented here provide proof of a successful coupling of 5mC oxidation and TDG-initiated BER in a cascade of enzymatic reactions that productively demethylates DNA. *In vitro* reconstitution of the active demethylation of symmetrically modified CpGs revealed a

mechanism that is intrinsically coordinated to operate sequentially on both strands. While this prevents the formation of DNA DSBs, and hence genomic instability, the process can be mutagenic if 5mC deamination and oxidative demethylation events coincide on opposite strands in a CpG dinucleotide.

In line with co-localization evidence reported previously[39] our biochemical work provides strong evidence for a direct physical interaction of TET1 with TDG, linking 5mC oxidation to base excision. This interaction, which involves structures within the N-terminus and the catalytic domain of TET1, allowed us to enrich a functional TET1/TDG complex from *E.coli* lysates that was highly active and capable of removing 5mC from a synthetic DNA substrate. In contrast to previous studies, showing 5mC conversion by TET and base excision by TDG in separate assays[12,13,16], our data demonstrate the functionality of a complex of both enzymes in couple 5mC oxidation and excision.

The relative high abundance of 5hmC in cells compared to 5fC and 5caC[13,40] suggests that 5mC oxidation by TET enzymes is tightly regulated. One way of explanation is that the rate of the initial oxidation of 5mC to 5hmC by TET enzymes is higher than that of the subsequent oxidations of 5hmC or 5fC. The latter oxidation steps may depend on or be stimulated by the presence of additional factors, such as the TDG[13,41] and/or Gadd45[42,43]. We examined this possibility but did not measure a stimulatory effect of TDG on $TET1_{CD}$ catalysis at any step of oxidation, neither did we observe such an effect for Gadd45a added to the $TET1_{CD}$/TDG complex (Supplementary Fig. S3). These experiments were done with $TET1_{CD}$, however, leaving the possibility that the missing N-terminus with its zinc finger CXXC domain may play an allosteric regulatory role. Additional work is needed to address the important question of TET1 regulation; i.e. to identify the factors and mechanisms determining the patterning of genomic 5hmC, 5fC and 5caC generation. The reconstituted demethylation assay presented here will be instrumental in this endeavor.

The engagement of a DNA glycosylase in active DNA demethylation inevitably generates a need for repair of AP-sites. Evidence supporting an involvement of the BER pathway has been reported for PGCs, where an increase of DNA SSBs and BER activity was linked to active global DNA demethylation[44], and in a recent study, demonstrating that various BER proteins co-precipitate with TET1, when overexpressed in HEK293T cells[39]. With the successful *in vitro* reconstitution of TET1/TDG-BER-mediated DNA demethylation, we provide the first direct evidence for a physical and functional coupling of these factors in the oxidation and excision of 5mC and the re-synthesis of an unmethylated CpG. While, from a mechanistic point of view, BER-mediated DNA demethylation seems

a straightforward process, it raises concerns regarding potential adverse effects on genome stability, particularly where the density of CpGs undergoing demethylation is high and excessive formation of DNA strand-breaks might occur. It is therefore fair to assume that active demethylation in cells is a highly orchestrated process, controlled through regulatory mechanisms also involving posttranslational modifications[18,32,36]. Our *in vitro* DNA demethylation system does not recapitulate regulatory actions of this kind but it does inform on intrinsic features of the mechanism regarding the potential of DNA DSB formation and the handling of complex substrates.

A distributive mode of action of TET proteins, for instance, would produce a variety of demethylation intermediates with 5caC placed opposite from 5mC, 5hmC or an unmodified C within CpG dinucleotides, the precise configuration of which may then determine the efficiency of initiation of BER. This, however, seems an unlikely regulatory concept as TDG processed 5caC with high efficiency irrespective of the opposite C modification. Yet, our experiments also indicate that while the TET/TDG demethylase is capable of acting on both strands at symmetrically modified CpGs, it does so in a sequential manner without producing DNA DSBs. Even with substrates containing the efficiently processed 5caC in both strands, TDG-mediated BER did not generate detectable DNA DSBs, and this was not due to a preferential action of TDG on one strand only. Owing to its high affinity to AP-sites, TDG will not easily turnover in these reactions[30,35]. In the specific case of the occurrence of symmetrical lesions within CpG dinucleotides, such as during symmetrical DNA demethylation, this feature of TDG constitutes an important protective mechanism; not only will it provide an opportunity to coordinate AP-site repair but also protect the substrate in the opposite strand from being repaired at the same time. Biochemical evidence suggests that TDG will dissociate only in the presence of downstream BER factors, which then allows for efficient repair of the AP-site, restoration of the DNA strand with an unmodified C, and ultimately initiation of repair at the opposite strand. The importance of this coupling of 5caC excision to the BER process in this delicate situation is highlighted by the observation that in absence of POLβ and XRCC1/LIG3, TDG and APE1 generated an appreciable amount of DSBs in symmetrically modified substrates (Supplementary Fig. S4). We therefore argue that BER in the context of active DNA demethylation occurs in a processive manner, where the initially attacked strand is fully repaired prior to processing of the opposite strand (Fig. 6d). This may explain how the replacement of symmetrical 5mC with unmodified C can occur without destabilizing the genome.

Another complex situation that may arise is the coincident deamination and oxidation of symmetrically opposed 5mCs in CpG dinucleotides. Methylated CpGs are well known for their increased mutability, which is to a large extent due to the higher rate of hydrolytic deamination of methylated cytosines compared unmethylated cytosines[45]. Such deamination events occur with an appreciable frequency in the genome and will generate G•T mispairs within methylated CpG dinucleotides[37]. This observation alone, however, does not explain the relatively high C to T mutation rates as cells have efficient mechanisms in place to repair the pre-mutagenic G•T mismatch, e.g. TDG or MBD4 mediated BER[18]. Our data on G•T versus 5caC repair in CpG dinucleotides provide a plausible biochemical explanation of how G•T mismatches might occasionally escape correction and turn into mutations. While, consistent with previous observations[16], TDG processed the G•T mismatch with higher efficiency than 5caC when the two lesions were analyzed separately, 5caC was processed with a striking preference when both substrates were present within the same CpG, reflecting a situation where spontaneous deamination occurs at a site that is being actively demethylated (Fig. 6d). Evidently, TDG has a high affinity for 5caC, even when base-paired with a G, consistent with the uniquely specific active site contacts it establishes with this substrate[46,47]. The sequential repair of both lesions, which, as argued above, helps to avoid the formation of DNA DSBs, turns into a disadvantage in this particular situation. The initial 5caC processing in the context of DNA demethylation would mask a nearby G•T mismatch for repair and fix the C to T mutation within the CpG dinucleotide whenever the re-synthesis step of BER incorporates more than just one nucleotide (Fig. 6d).

In conclusion, our data provide proof of functionality of an active DNA demethylation pathway based on the coupled oxidation and excision repair of 5mC, they provide insight into how intrinsic features of the mechanism allow demethylation of symmetrically methylated CpGs without the formation of DNA DSBs and how it may contribute to C to T mutagenesis within methylated CpG dinucleotides. Having a fully reconstituted DNA demethylation process established will allow future investigations into the detailed mechanism of the process, including the important aspect of regulation.

MATERIALS AND METHODS

Bacterial expression vectors: The plasmids for the expression of TDG (GI: 37589917) (pTG-mTDGa.0, pET28-mTDGa.0, pET28-mTDGa.1), TET1 (GI: 568968019) (pCDF-mTET1), TET1 catalytic domain

(aa 1367 – 2057) (pCDF-His-mTET1CD, pCDF-His-mTET1CDΔcat), TET1 N-terminus (aa 301 - 1366) (pACYC-mTET1-N), APE1 (GI:18375501) (pPRS125, pEThis-APE1.0), POLβ (GI:4505931) (pPRS112, pQE30-6HIS-Polβ), XRCC1 (GI:190684675) (pET-XRCC1) and LIG3 (GI:73747829) (pGEX4T-Lig3) were assembled by standard cloning methods based on PCR amplification with adaptor-oligonucleotides providing suitable restriction sites.

Antibodies: TDG, rabbit polyclonal ab 141, 1:20'000; TET1$_{CD}$, rabbit polyclonal α-TET1 ab (Millipore), 1:5'000; TET1-N, rabbit polyclonal α-TET1 ab (Genetex), 1:10'000; 5mC, mouse monoclonal α-5mC ab (Diagenode), 1:250; 5hmC, rabbit polyclonal α-5hmC (Active motif), 1:20'000, 5fC rabbit polyclonal α-5fC (Active motif), 1:2'500; 5caC, rabbit polyclonal α-5caC (Active motif), 1:2'000.

5-carboxyethyl-N4-benzoyl-dC CE phosphoramidite: The 5caC phosphoramidite (5-carboxyethyl-N4-benzoyl-dC CE) was synthesized in collaboration with Glen Research (USA).

Oligonucleotides: 60mer (Substrate 1 and 3) or 59mer (Substrate 2) double-stranded oligonucleotide substrates containing different modifications were prepared by annealing of two complementary oligonucleotides synthesized by Adam Robertson or Microsynth (Switzerland). The upper strand was either unlabeled or carried a 5'-TexasRed label, the lower strand was unlabeled or carried a 5'-fluorescein label. Substrate 1 (standard) upper strand 5'-TAGACATTGCCCTCGAGGTACCATGGATCCGATGTXGACCTCAAACCTAGACGAATTCCG -3' where X = C, T, 5mC, 5hmC, 5caC. Substrate 1 lower strand strand 5'-CGGAATTCGTCTAGGTTTGAGGTXGACATCGGATCCATGGTACCTCGAGGGCAATGTCTA-3', where X = T, 5mC, 5hmC or 5caC. Substrate 2 upper strand 5'-TAGACATTGCCCTCGACGACCCGCCGCCGCGCXGGCCACCCGCACCTAGACGAATTCCG -3' where X = C, T, 5mC, 5hmC, 5caC. Substrate 2 lower strand 5'-CGGAATTCGTCTAGGTGCGGGTGGCXGGCGCGGCGGCGGGTCGTCGAGGGCAATGTCTA -3' where X = 5mC, 5hmC, 5caC. Substrate 3 upper strand 5'-TAGACATTGCCCTCGACGGTGCCCTCXGGGCCGCGCGTCGCGCTCCCTAGACGAATTCCG -3' where X = C. Substrate 3 lower strand 5'-CGGAATTCGTCTAGGGAGCGCGACGCGCGGCCXGGAGGGCACCGTCGAGGGCAATGTCTA -3' where X = 5mC, 5hmC.

Recombinant protein expression

The expression vectors were introduced into *E.coli* BL21(DE3) cells by electroporation. Overnight pre-cultures were diluted with fresh pre-warmed LB medium and grown at 30°C to an OD_{600} level of 0.6 - 0.8. Cultures were grown under selective pressure using respective antibiotics at concentrations of either 100 mg/L of (Ampicillin) or 50 mg/L (Kanamycin, Streptomycin) for single plasmid expressions and half the concentration of each antibiotic when co-expressing two plasmids. Protein expression was induced using following conditions: TET1 (250 μM IPTG, 25°C for 3 h), TET1/TDG (250 μM IPTG, 25°C for 3 h), TDG (250 μM IPTG, 15°C for 16 h), APE1 (500 μM IPTG, 25°C for 6 h), POLβ (500 μM IPTG, 25°C for 3 h), LIG3 and XRCC1 were co-expressed (250 μM IPTG, 25°C for 4 h). Finally, cells were harvested by centrifugation and soluble protein fractions were extracted by sonication (Bioruptor, Diagenode) or homogenization (Emulsiflex C-3, Avestin) in lysis buffer (50 mM Na-phosphate buffer pH 7.5, 300 mM NaCl, 20% glycerol, 0.1% Tween-20, 1 mM DTT, 1 mM PMSF), if not stated otherwise. Crude lysates were then cleared by centrifugation with >30'000 g at 4°C for 60 min.

Protein purification

For $TET1_{CD}$ purification, the cleared lysate was loaded onto a 1 mL HisTrap FF crude column (GE Healthcare), bound protein was eluted with 400 mM imidazole and relevant fractions dialyzed against CIEX buffer (50 mM HEPES pH 7.2, 25 mM NaCl, 20% glycerol, 5 mM DTT and 0.1 mM PMSF). Dialyzed fractions were then loaded onto a 1 mL Resource S column (GE Healthcare) and bound protein was eluted with a linear salt gradient of 25 mM – 1 M NaCl and purest fractions finally dialyzed against storage buffer (50 mM HEPES pH 7.2, 100 mM NaCl, 20% glycerol, 5 mM DTT), frozen on dry-ice and stored at -80 °C.

The BER proteins were purified as followed; in brief, APE1 and POLβ were purified with Ni-NTA affinity and ion exchange chromatography as described before[34,48]. TDG was purified by Ni-NTA affinity, heparin affinity and ion exchange chromatography as described before[49]. LIG3 and XRCC1 were purified as complex by Ni-NTA, GST and Ni-NTA affinity chromatography according to the manufacturer's instructions. Highly pure fractions were dialyzed against storage buffer (10 mM Tris-HCl pH8, 50 mM NaCl, 10% glycerol), snap frozen and stored at -80°C.

Gel filtration: Gel filtration was performed using a Superdex 200 10/300GL column (GE Healthcare) and an ÄKTA Purifier 10 (GE Healthcare) according to the manufacturer's instructions. Ni-NTA enriched fractions were prepared as described above. Ni-NTA elution fractions were pooled, concentrated to 8 mg/mL using Amicon Ultra Centrifugal Filters (Millipore) and buffer was changed to gel filtration running buffer (50 mM Na-phosphate pH 7.5, 500 mM NaCl, 20% glycerol). 4 mg of the enriched fraction was then loaded onto the gel filtration column. Column washing, loading and sampling of the fractions was done according to the manufacturer's instructions. 0.5 mL fractions were collected and 20 µL of each fraction was used for SDS-PAGE and Western blot analysis.

Batch Ni-NTA and GST affinity purification: To study the interaction of TDG and TET1, Ni-NTA and GST pull-down assays were performed. TDG-GST was co-expressed with a TET1 N-terminal fragment (His6-TET1-N aa 301 - 1366) or the TET1 catalytic domain (His6-TET1$_{CD}$ aa 1367 - 2057) in *E.coli* as described above. 5 mg of cleared *E.coli* lysate was then incubated with 25 µL of Glutathione Magnetic Beads (Thermo Scientific) or Ni-NTA Sepharose beads (Roche) in binding buffer (50 mM Na-phosphate pH 7.5, 300 mM NaCl, 20% glycerol, 0.1% Tween-20, 1 mM DTT, 1 mM PMSF) in a total volume of 1 mL at room temperature for 2 h. The beads were rinsed 3 times with 500 µL binding buffer and bound proteins were analyzed by SDS-PAGE and Western blotting.

Partial purification of the TET1$_{CD}$/TDG complex for activity assays was done via Ni-NTA affinity purification as described above. As catalytic mutants, His6-TET1$_{CD}$Δcat (H1652Y; D1654A) and TDGΔcat-GST (N151A) were used.

Analytical gel electrophoresis and western blotting

Protein fractions were analyzed by standard SDS-polyacrylamide gel electrophoresis (SDS-PAGE) followed by Coomassie blue staining or by immunoblotting using chemiluminescence (WesternBright ECL, Advansta) according to the provider's protocol. Antibodies were diluted in 5% non-fat dry milk TBS (100 mM Tris-HCl pH 8, 150 mM NaCl) supplemented with 0.2% Tween-20.

Yeast Two Hybrid assay

To confirm the interaction between TET1 and TDG yeast two hybrid was performed using he Matchmaker yeast-two hybrid system (Clontech). TET1 was divided into 4 overlapping fragments (TET1-1 aa 1-491; TET1-2 aa 397-931; TET1-3 aa 870-1403; TET1-4 aa 1367-2057) that were cloned into the BD (pAS2.1 BD FLAG) of the Gal4 protein, TDG was cloned into the AD (pACT2 AD) of the Gal4 protein. The *Saccharomyces cerevisiae* strain AH109 was co-transformed with 50-500 ng of bait and prey plasmids according to the Clontech manual. Interactions were assessed by spotting serial dilutions of cells on selective medium (SC-LEU-TRP-ADE-HIS) supplemented with 2.5 mM 3AT (3-Amino-1,2,4-triazole), a competitive inhibitor of the HIS3 gene product. Cells were incubated at 30°C for 6 to 7 days.

Base release assay

The catalytic activity of the TET1/TDG complex was monitored by means of a standardized nicking assay [30]. The reactions were carried out in a reaction volume of either 40 µL when using partially purified TET1/TDG complex from Ni-NTA affinity purification fractions or 20 µL when using purified recombinant protein containing TET reaction buffer (50 mM HEPES pH 8, 50 mM NaCl, 1 mM Di-Sodium-Ketoglutarate, 2 mM Ascorbic acid, 75 µM Fe(II), 1 mM ATP), 0.5 pmol of substrate and 10 uL of Ni-NTA pulldown or 2 pmol purified $TET1_{CD}$ and 1 pmol purified TDG (preincubated together on ice for 5') respectively. Reactions were incubated at 37°C for 1 h and stopped by addition of 1 M NaOH to a final concentration of 100 mM and heating for 10 min at 99°C. After EtOH precipitation at -20°C overnight, the products were separated in a 15% denaturing polyacrylamide gel and labeled DNA was detected using the red or blue fluorescence mode of the Typhoon 9400 (GE Healthcare) and analyzed quantitatively by ImageQuant TL software (v7.0, GE Healthcare).

TDG time course reactions were carried out in 200 µL reaction volume containing nicking buffer (50 mM Tris-HCl pH 8, 1 mM DTT, 0.1 mg/mL BSA, 1 mM EDTA), 5 pmol of labeled substrate DNA and 5 pmol of purified TDG. After the indicated times of incubation at 37°C, 20 µL aliquots were withdrawn and the reactions were stopped by the addition of 1 M NaOH to an end concentration of 100 mM and heating for 10 min at 99°C. Reaction products were analyzed by denaturing polyacrylamide gel electrophoresis and analyzed as described above.

In vitro methylation and oxidation of plasmid DNA and slot blot analysis

In vitro methylation of pUC19 plasmid DNA was performed using M.SssI CpG methyltransferase (New England Biolabs) according to the manufacturer's instructions.

For the _in vitro_ oxidation 200 ng of methylated plasmid was incubated with 500 ng purified His6-TET1$_{CD}$ from _E.coli_ (see above). The reaction was carried out in TET reaction buffer (50 mM HEPES pH 8, 50 mM NaCl, 1 mM Di-Sodium-Ketoglutarate, 2 mM Ascorbic acid, 75 µM Fe(II), 1 mM ATP) and incubated at 37°C for 1 h. Reaction was stopped with the addition of NaOH and EDTA to a final concentration of 400 mM and 10 mM respectively and heating at 99°C for 10 min. The denatured DNA was blotted using the Bio-Rad slot blot system according to the manufacturer's instruction. Hybond-N+ nylon membranes (Amersham) were UV-crosslinked, blocked with 5% milk and immunostaining against 5mC, 5hmC, 5fC and 5caC was performed using chemiluminescence (WesternBright ECL, Advansta) according to the provider's protocol. Antibodies were diluted in 5% non-fat dry milk TBS (100 mM Tris-HCl pH 8, 150 mM NaCl) supplemented with 0.2% Tween-20.

LC/MS/MS analysis

Plasmid DNA samples were enzymatically hydrolyzed to deoxyribonucleosides by incubation at 45ºC for 40 min in 10 mM ammonium acetate buffer pH 5.3 containing 5 mM magnesium chloride and 0.2 U nuclease P1 from Penicillium citrinum (Sigma, N8630). The samples were then buffered in ammonium bicarbonate to a final concentration of 100 mM and incubated at 37ºC for 30 min with 0.0002 U phosphodiesterase I from Crotalus adamanteus venom (Sigma, P3243) and 0.3 U alkaline phosphatase from _E.coli_ (Sigma, P5931). The reactions were stopped and contaminants, that could potentially clog the HPLC column, were precipitated by adding three volume equivalents of ice-cold acetonitrile and centrifugation at 16,000 g for 30 min. The supernatants were collected in new tubes and vacuum centrifuged at room temperature until dry. Salt residues, originating from buffers, were partially evaporated by re-dissolving the samples in 100 µL of water and vacuum-drying one more time. The standards for 5-me(dC), 5-hm(dC), 5-ca(dC) and 5-f(dC) were prepared to contain the same amount of salts as the samples and followed the same desalting procedure. The samples were then finally dissolved in 50 µL of water for LC/MS/MS analysis of 5-me(dC), 5-hm(dC), 5-ca(dC) and 5-f(dC). For quantification of unmodified nucleosides (dA, dC, dG and dT) samples were diluted 1:10

19

with water. For some of the samples 1:10 dilution was also used during quantification of 5-me(dC).

Quantification was performed with the use of an LC-20AD HPLC system (Shimadzu) coupled to an

API 5000 triple quadrupole (ABSciex) operating in positive electrospray ionization mode. The

chromatographic separation was performed at 40ºC with the use of an Ascentis Express C18 2.7 µm

150 x 2.1 mm i.d. column protected with an Ascentis Express Cartridge Guard Column (Supelco

Analytical) with an Exp Titanium Hybrid Ferrule (Optimize Technologies Inc.). The mobile phase

consisted of A (water, 0.1% formic acid) and B (methanol, 0.1% formic acid) solutions. Following

conditions were employed during chromatography: for unmodified nucleosides – 0.13 mL/min flow,

starting at 10% B for 0.1 min, ramping to 60% B over 2.4 min and re-equilibrating with 10% B for 4.5

min; for 5-me(dC), 5-hm(dC), 5-ca(dC), and 5-f(dC) - 0.14 mL/min flow, starting at 5% B for 0.5 min,

ramping to 45% B over 8 min and re-equilibrating with 5% B for 5 min. For mass spectrometry

detection the multiple reaction monitoring (MRM) was implemented using the following mass

transitions: 252.2/136.1 (dA), 228.2/112.1 (dC), 268.2/152.1 (dG), 243.2/127.0 (dT), 242.1/126.0 [5-
me(dC)], 258.1/142.0 [5-hm(dC)], 256.1/140.0 [5-f(dC)], 272.1/156.0 [5-ca(dC)].

BER reconstitution

The BER reconstitution reaction was carried out stepwise to analyze individual stages of the process.

The reaction mixture containing 1 pmol labeled 60 bp or 59 bp DNA, 5 pmol His6-TET1$_{CD}$ and 2 pmol

His6-TDG were incubated at 37°C for 30 min in TET reaction buffer (50 mM HEPES pH 8, 50 mM

NaCl, 1 mM Di-Sodium-Ketoglutarate, 2 mM Ascorbic acid, 75 µM Fe(II), 1 mM ATP) to generate an

AP-site. The reaction mixture was then supplemented with 70 mM KCl, 7 mM MgCl$_2$, 200 µM dCTP or

dNTP, 2 mM ATP, 500 µg/mL BSA, 1 mM DTT and 10 pmol APE1 and incubated at 37°C for 5 min.

0.5 pmol DNA polymerase β was then added and the reaction mixture incubated for a further 5 min.

Finally 2 pmol XRCC1/DNA ligase III complex was added for a 10 min incubation. Reactions were

terminated by the addition of stop buffer (50 mM Tris-Cl pH 8, 0.5% SDS, 100 mM NaBH$_4$) and

incubation on ice for 20 min. The reaction products were analyzed by denaturing polyacrylamide gel

electrophoresis and analyzed as described above.

For the analysis of the endproduct with *Hpa*II or *Msc*I endonuclease digest, the reconstitution reaction

was carried out by adding all the factors at the same time and incubation at 37°C for 1 h followed by

EtOH precipitation of the labeled DNA at -20°C overnight. The recovered DNA was then treated with a

total of 5 U *Hpa*II or *Msc*I endonuclease (New England Biolabs) at 37°C for 60 min, fragments were

separated in 8% native polyacrylamide gels and detected as described above.

REFERENCES

1.   Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**, 6-21 (2002).
2.   Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* (2012).
3.   Guo, H. *et al.* The DNA methylation landscape of human early embryos. *Nature* **511**, 606-10 (2014).
4.   Smith, Z.D. *et al.* DNA methylation dynamics of the human preimplantation embryo. *Nature* **511**, 611-5 (2014).
5.   Oswald, J. *et al.* Active demethylation of the paternal genome in the mouse zygote. *Curr Biol* **10**, 475-8 (2000).
6.   Seisenberger, S. *et al.* The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol Cell* **48**, 849-62 (2012).
7.   Metivier, R. *et al.* Cyclical DNA methylation of a transcriptionally active promoter. *Nature* **452**, 45-50 (2008).
8.   Kangaspeska, S. *et al.* Transient cyclical methylation of promoter DNA. *Nature* **452**, 112-5 (2008).
9.   Kim, M.S. *et al.* DNA demethylation in hormone-induced transcriptional derepression. *Nature* **461**, 1007-12 (2009).
10. Schar, P. & Fritsch, O. DNA repair and the control of DNA methylation. *Prog Drug Res* **67**, 51-68 (2010).
11. Pastor, W.A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature reviews. Molecular cell biology* **14**, 341-56 (2013).
12. He, Y.F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-7 (2011).
13. Ito, S. *et al.* Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science* (2011).
14. Inoue, A., Shen, L., Dai, Q., He, C. & Zhang, Y. Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. *Cell Res* **21**, 1670-6 (2011).
15. Huang, Y. *et al.* Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 1361-6 (2014).
16. 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* (2011).
17. Kohli, R.M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472-9 (2013).
18. Jacobs, A.L. & Schar, P. DNA glycosylases: in DNA repair and beyond. *Chromosoma* (2011).
19. Dawlaty, M.M. *et al.* Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. *Cell Stem Cell* **9**, 166-75 (2011).
20. Dawlaty, M.M. *et al.* Combined deficiency of tet1 and tet2 causes epigenetic abnormalities but is compatible with postnatal development. *Dev Cell* **24**, 310-23 (2013).
21. Gu, T.P. *et al.* The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* (2011).
22. Cortazar, D. *et al.* Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature* **470**, 419-23 (2011).
23. Cortellino, S. *et al.* Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell* **146**, 67-79 (2011).
24. Saito, Y. *et al.* Embryonic lethality in mice lacking mismatch-specific thymine DNA glycosylase is partially prevented by DOPS, a precursor of noradrenaline. *Tohoku J Exp Med* **226**, 75-83 (2011).

25. Dawlaty, M.M. *et al.* Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Developmental cell* **29**, 102-11 (2014).

26. Shen, L. *et al.* Genome-wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics. *Cell* (2013).

27. Song, C.X. *et al.* Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. *Cell* (2013).

28. Raiber, E.A. *et al.* Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol* **13**, R69 (2012).

29. Dianov, G.L. & Hubscher, U. Mammalian base excision repair: the forgotten archangel. *Nucleic Acids Res* **41**, 3483-90 (2013).

30. Hardeland, U., Bentele, M., Jiricny, J. & Schar, P. Separating substrate recognition from base hydrolysis in human thymine DNA glycosylase by mutational analysis. *J Biol Chem* **275**, 33449-56 (2000).

31. Hu, L. *et al.* Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell* **155**, 1545-55 (2013).

32. Steinacher, R. & Schar, P. Functionality of human thymine DNA glycosylase requires SUMO-regulated changes in protein conformation. *Curr Biol* **15**, 616-23 (2005).

33. Kubota, Y. *et al.* Reconstitution of DNA base excision-repair with purified human proteins: interaction between DNA polymerase beta and the XRCC1 protein. *EMBO J* **15**, 6662-70 (1996).

34. Hardeland, U., Bentele, M., Jiricny, J. & Schar, P. The versatile thymine DNA-glycosylase: a comparative characterization of the human, Drosophila and fission yeast orthologs. *Nucleic Acids Res* **31**, 2261-71 (2003).

35. Waters, T.R., Gallinari, P., Jiricny, J. & Swann, P.F. Human thymine DNA glycosylase binds to apurinic sites in DNA but is displaced by human apurinic endonuclease 1. *The Journal of biological chemistry* **274**, 67-74 (1999).

36. Hardeland, U., Steinacher, R., Jiricny, J. & Schar, P. Modification of the human thymine-DNA glycosylase by ubiquitin-like proteins facilitates enzymatic turnover. *EMBO J* **21**, 1456-64 (2002).

37. Shen, J.C., Rideout, W.M., 3rd & Jones, P.A. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic acids research* **22**, 972-6 (1994).

38. Rai, K. *et al.* DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. *Cell* **135**, 1201-12 (2008).

39. Muller, U., Bauer, C., Siegl, M., Rottach, A. & Leonhardt, H. TET-mediated oxidation of methylcytosine causes TDG or NEIL glycosylase dependent gene reactivation. *Nucleic acids research* **42**, 8592-604 (2014).

40. Pfaffeneder, T. *et al.* The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. *Angew Chem Int Ed Engl* **50**, 7008-7012 (2011).

41. Hashimoto, H. *et al.* Structure of a Naegleria Tet-like dioxygenase in complex with 5-methylcytosine DNA. *Nature* **506**, 391-5 (2014).

42. Li, Z. *et al.* Gadd45a promotes DNA demethylation through TDG. *Nucleic acids research* **43**, 3986-97 (2015).

43. Niehrs, C. & Schafer, A. Active DNA demethylation by Gadd45 and DNA repair. *Trends in cell biology* **22**, 220-7 (2012).

44. Hajkova, P. *et al.* Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. *Science* **329**, 78-82 (2010).

45. Cohen, N.M., Kenigsberg, E. & Tanay, A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* **145**, 773-86 (2011).

46. Hashimoto, H., Zhang, X. & Cheng, X. Activity and crystal structure of human thymine DNA glycosylase mutant N140A with 5-carboxylcytosine DNA at low pH. *DNA repair* **12**, 535-40 (2013).

47. Zhang, L. *et al.* Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nat Chem Biol* **8**, 328-30 (2012).

48. El-Andaloussi, N. *et al.* Arginine methylation regulates DNA polymerase beta. *Molecular cell* **22**, 51-62 (2006).

49. Kunz, C. *et al.* Base excision by thymine DNA glycosylase mediates DNA-directed cytotoxicity of 5-fluorouracil. *PLoS Biol* **7**, e91 (2009).

ACKNOWLEDGMENTS

FIGURE LEGENDS

Figure 1: TET1 physically interacts with TDG. (**a**) Size fractionation by gel filtration at high ionic strength (500 mM NaCl) of Ni-NTA enriched lysates of *E.coli* cells co-expressing TET1-His6 and TDG-GST from constructs indicated. Fractions were analyzed by SDS-PAGE and TET1 and TDG detected by immunoblotting. Shown is the SDS-PAGE stained with Coomassie Blue (left panel) and an immunoblot (right panel); molecular weights of gel filtration standards are indicated. (**b**) Yeast two hybrid analysis of the TET1 TDG interaction. TET1 domains cloned into the GAL4 activation domain (AD), TET1-1 (aa 1 - 491), TET1-2 (aa 397 - 931), TET1-3 (aa 870 - 1403), TET1-4 (aa 1367 - 2057) are indicated at the top. Shown is the growth reactions of serial dilutions of strains co-expressing TET1 domains fused to the AD and the TDG fused to the GAL4 binding domain (BD) and respective negative controls (TET1 domains or TDG co-expressed with the vector control (V)) on permissive and selective media as indicated. The large T antigen (lTAg) and p53 fused to the AD and BD, respectively, served as a positive control. (**c**) Immunoblot analysis of fractions obtained from Ni-NTA and GST purifications using *E.coli* extracts co-expressing His6-TET1$_N$ and TDG-GST (left panel), or co-expressing His6-TET1$_{CD}$ and TDG-GST (right panel). Expression constructs used are indicated; TET1$_N$ (aa 301-1366), TET1$_{CD}$ (aa 1367-2057).

Figure 2: Combined TET1 and TDG activity releases 5mC through oxidized intermediates. (**a**) Base excision activity of a Ni-NTA enriched His6-TET1$_{CD}$/TDG-GST complex on synthetic DNA substrates as indicated. The ability to generate alkaline sensitive AP-sites in two different substrates containing either single G•5mC or G•5hmC base pairs was assayed with enriched His6-TET1$_{CD}$/TDG-GST complexes consisting of wildtype proteins or respective mutant variants (His6-TET1$_{CD}$Δcat/TDG-GST, His6-TET1$_{CD}$/TDGΔcat-GST). Products were separated by denaturing gel electrophoresis, visualized with fluorescent scanning and quantified; positions of the 60mer substrate DNA and product fragment are indicated. (**b**) Slot blot analysis of plasmid oxidation by purified His6-TET1$_{CD}$. *In vitro* methylated pUC19 plasmid DNA (800 nM) was treated with His6-TET1$_{CD}$ (125 nM) and cytosine modifications

were detected by immunblotting with specific antibodies against 5mC, 5hmC, 5fC and 5caC. (**c**) Reconstitution of 5mC/5hmC base release with purified His6-TET1$_{CD}$ and His6-TDG proteins. DNA substrates (25 nM) containing either G•5mC, G•5hmC or G•5caC base pairs were reacted with a preassembled His6-TET1$_{CD}$/His6-TDG complex (50 nM), reaction products separated by denaturing gel electrophoresis, visualized and quantified. Positions of the 60mer substrate DNA and product fragments are indicated. Shown are mean values with standard deviations (n = 3).

Figure 3: TET1$_{CD}$ stabilizes TDG activity. (**a**) Stimulatory effect of His6-TDGΔcat on His6-TET1$_{CD}$ analyzed by base release assay. His6-TET1$_{CD}$ (50 nM) or His6-TET1$_{CD}$/His6-TDGΔcat (25 nM) complex were incubated with DNA substrate (25 nM) containing a G•5mC base pair for the indicated times. Recovered DNA was then assayed by a base release assay using His6-TDG (250 nM) to monitor presence of oxidized 5mC species. Shown is a typical gel-scan (top panel) with mean values and standard deviations of 3 independent experiments (bottom panel). (**b**) LC/MS/MS analysis of plasmid oxidation assays using His6-TET1$_{CD}$ or TET1$_{CD}$/His6-TDGΔcat. *In vitro* methylated pUC19 plasmid DNA (660 nM) was treated with either TET1$_{CD}$ (100 nM) or a preassembled TET1$_{CD}$/TDGΔcat complex for the indicated time to determine the stimulatory effect of TDG on TET1 catalysis. DNA was analyzed by LC/MS/MS; shown are mean values (modif. nts/10^6 unmodif. nts) with standard deviations of 3 independent experiments. (**c**) The effect of His6-TET1$_{CD}$ on His6-TDGΔcat catalysis assessed in base release assays. The time-dependent generation of AP-sites was measured after reaction of a 60mer substrate containing either a single G•5caC or G•T base pair with a preassembled His6-TDG/His6-TET1$_{CD}$ complex or His6-TDG/BSA. Equimolar concentrations of G•5caC substrate (25 nM) and TDG protein (25 nM) (upper pannel) or a 5-fold molar excess of G•T substrate (125 nM) over TDG protein (25 nM) (lower panel) were incubated at 37°C and reactions stopped by the addition of NaOH after the indicated times. Product formation was monitored and quantified using denaturing gel electrophoresis and fluorescent scanning. Shown are mean values with standard deviations (n = 1 (G•T), n = 3 (G•5caC).

Figure 4: Full reconstitution of TET-TDG-BER mediated DNA demethylation. (**a**) Intermediate steps of the oxidative DNA demethylation reaction were reconstituted and visualized by denaturing gel electrophoresis. Labeled 60mer substrate DNA containing one G•5mC base pair was incubated sequentially with TET/TDG-BER enzymes at concentrations indicated. Reaction products were separated by denaturing gel electrophoresis and visualized by fluorescent scanning; sizes of the
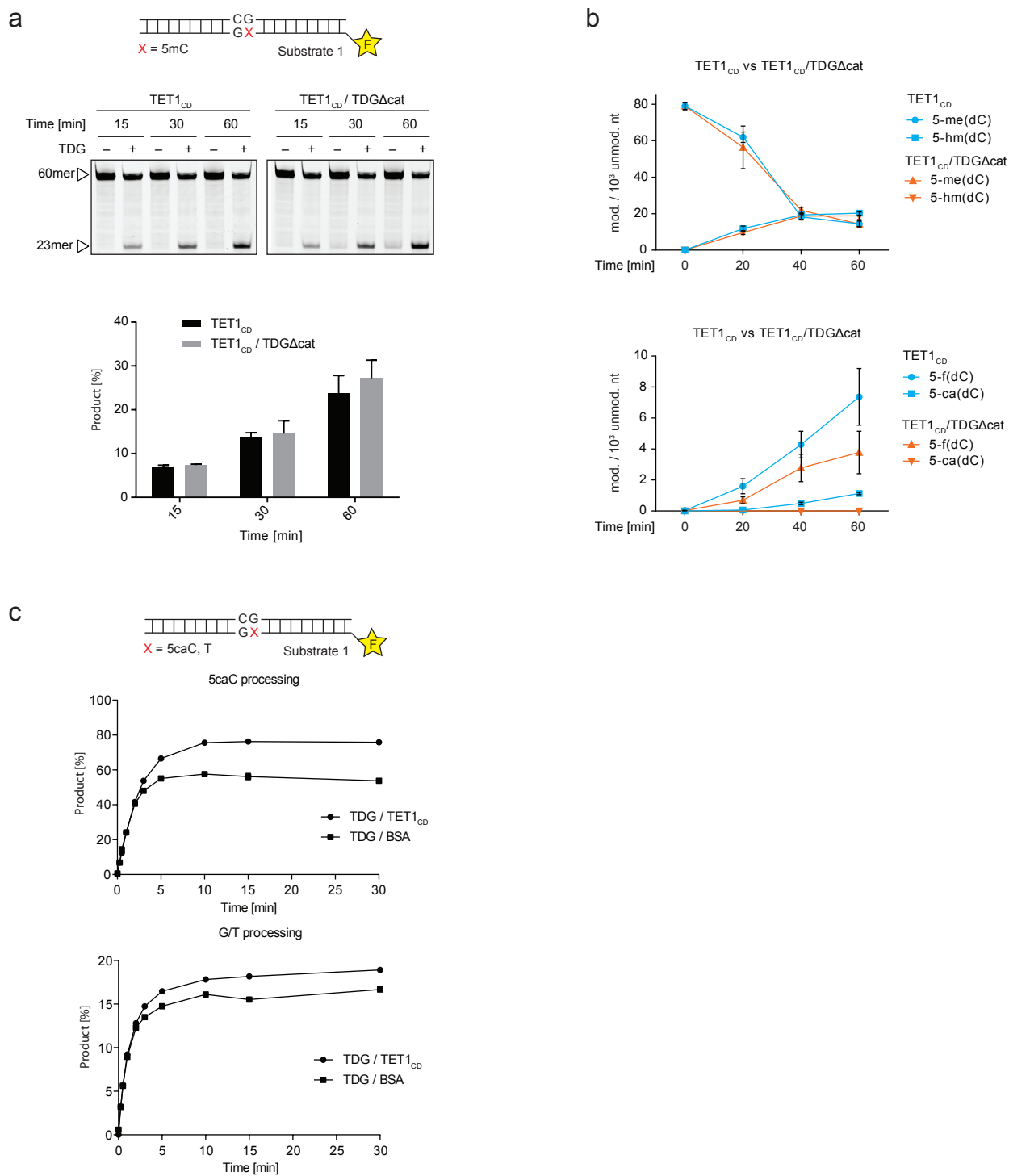
60mer substrate DNA and reaction products are indicated. (**b**) Complete DNA demethylation by the reconstituted TET/TDG-BER system analyzed by the generation of a *Hpa*II sensitive restriction site. Reconstituted DNA demethylation was done with a 5'-labeled 59 bp substrate containing one G•5mC base pair within a *Hpa*II recognition site (CCGG). Recovered DNA was digested with methylation sensitive *Hpa*II endonuclease and analyzed by native polyacrylamide gel electrophoresis; positions of the 59 bp substrate DNA and product fragment are indicated.

Figure 5: Processing of differentially modified CpGs by TET1$_{CD}$/TDG or TDG. (**a**) Base release from fully methylated CpGs by His6-TET1$_{CD}$/His6-TDG. His6-TET1$_{CD}$/His6-TDG complex preformed from purified proteins was incubated with labeled 60mer substrates containing a single 5mC modification on either the fluorescent labeled top (5'-TexasRed, TR) or bottom strand (5'-fluorescein, FAM) or containing a fully methylated CpG with labels on both strands as illustrated. Product formation was monitored and quantified using denaturing gel electrophoresis and fluorescent scanning (Texas Red, R-channel; fluorescein, F-channel); positions of the 60mer substrate DNA and the resulting base incision products of both strands are indicated. (*) unlabeled DNA strand. (**b**) Release of 5caC from differentially modified CpGs by His6-TDG monitored over time. 60mer DNA substrates containing 5caC opposite C, 5mC or 5hmC in a CpG dinucleotide or 5caC in single-stranded (ss) DNA were incubated with His6-TDG for the indicated times. Products were separated by denaturing gel electrophoresis, visualized by fluorescent scanning and quantified. Shown are mean percentages of product formation with standard deviations (n = 3). (**c**) 5caC release from a symmetrically modified CpG dinucleotide by His6-TDG. A substrate containing 5caC on both strands within a CpG dinucleotide and labels of both strand was incubated with His6-TDG for indicated times. Reactions were analyzed by denaturing gel electrophoresis and both strands visualized by fluorescent scanning adjusted to detect either of the two labels. Shown are mean percentages of product formation with standard deviation (n = 3). (**d**) Full reconstitution of TDG-BER on a symmetrically modified 5caC substrate. A labeled 59 bp substrate containing a symmetrically 5caC modified base pair within a *Hpa*II recognition site (CCGG) was incubated with TDG and BER factors. Recovered DNA was digested with *Hpa*II endonuclease and analyzed by native polyacrylamide gel electrophoresis, visualized and quantified by fluorescent scanning. Unmodified (CG/CG) and hemi-modified (CG/5caCG) substrates were used as controls for the *Hpa*II digest. Positions of the 59 bp substrate DNA and product fragments are indicated; ssDNA, free single-stranded DNA.
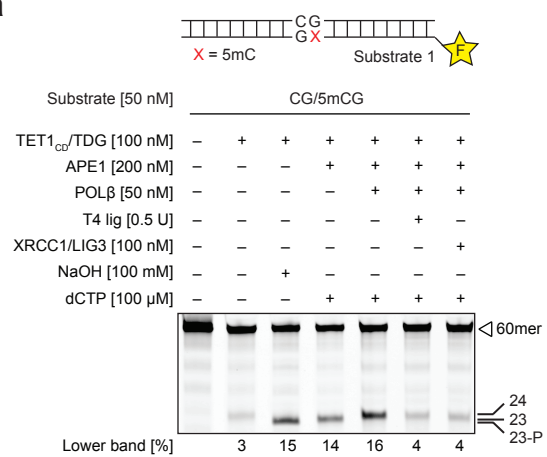
Figure 6: DNA demethylation blocks G•T repair and can induce mutations. (**a**) Enzymatic activity of

TDG on G•5caC and G•T containing substrates. Release of 5caC and T by His6-TDG (25 nM) was

monitored over time on 5'-labeled 60 bp substrates (25 nM) containing either a G•5caC or G•T base

pair. Reactions were stopped at indicated times, separated by denaturing gel electrophoresis,

visualized with fluorescent scanning and quantified. Shown are mean percentages of product

formation with standard deviations (n = 3) (**b**) Base release from a substrate containing a G•5caC next

to a G•T mismatch. Substrate preference of TDG (25 nM) was evaluated on a 59 bp DNA fragment

(25 nM) containing 5caC on the labeled top strand (5', TexasRed) and T on the labeled bottom strand

(5', fluorescein) within the same CpG context as illustrated. Reactions were stopped after indicated

times, separated by denaturing gel electrophoresis, and both strands visualized by fluorescent

scanning and quantified. Shown are mean percentages of product formation with standard deviations

(n = 3). (**c**) Full reconstitution of TDG-BER on a G•5caC/G•T containing substrate. A labeled 59 bp

substrate containing a G•5caC next to a G•T mismatch was incubated with His6-TDG and BER

factors. Correct repair of the 5caC as well as the introduction of an A opposite of T was monitored by

*Msc*I digestion and analyzed by native polyacrylamide gel electrophoresis and fluorescent scanning.

Unmodified (CG/CG) substrate DNA digested with *Hpa*II was used as size marker; positions of the

substrate DNA and product fragments are indicated. ssDNA indicates free single-stranded DNA. (**d**)

Mechanistic model of TET/TDG-BER-mediated DNA demethylation. In presence of all the necessary

factors, DNA demethylation at fully methylated CpGs occurs in a coordinated and sequential manner

to correctly reestablish the unmodified state (regular BER). Lack of coordination, e.g. in the absence of

downstream BER factors, repair-mediated DNA demethylation can lead to the induction of DNA DSBs

(uncoupled BER). Coincident oxidation and hydrolytic deamination at fully methylated CpG sites can

lead to increased C to T transitions caused by the sequential repair mechanism (coincident
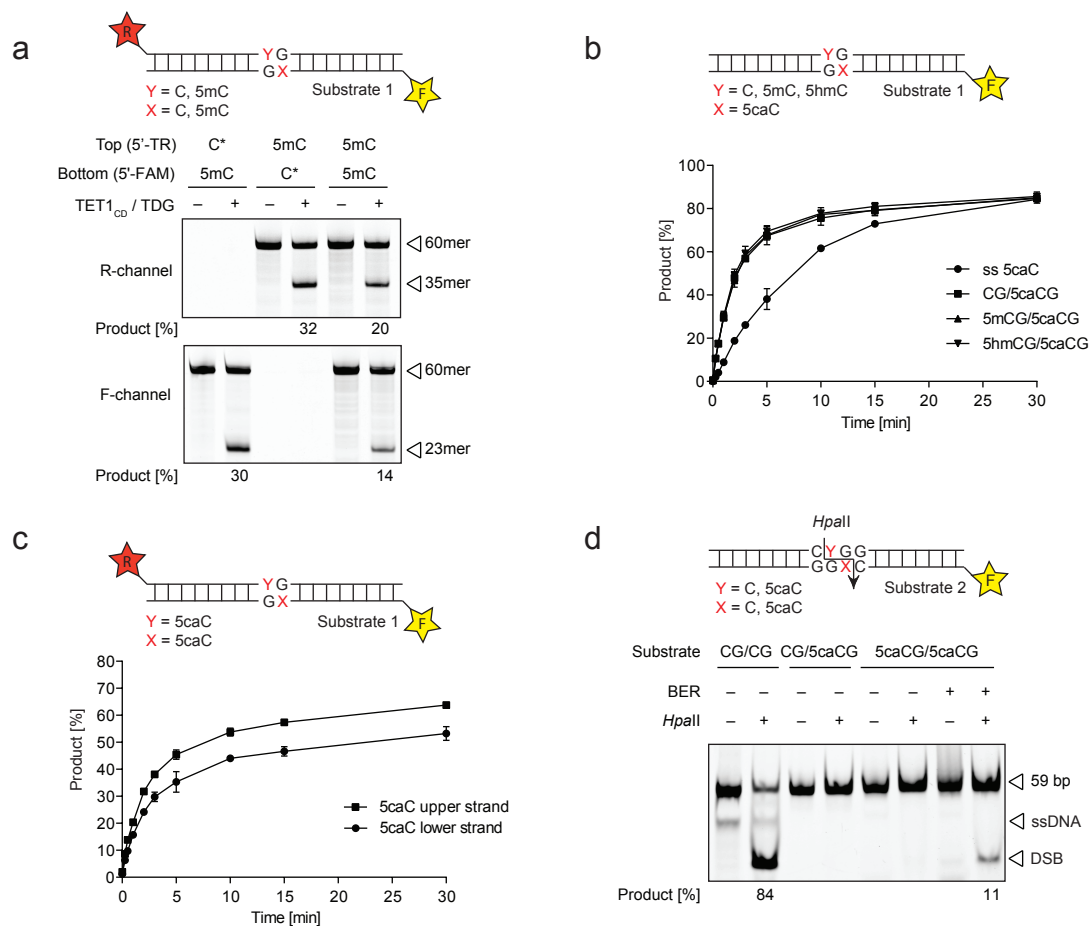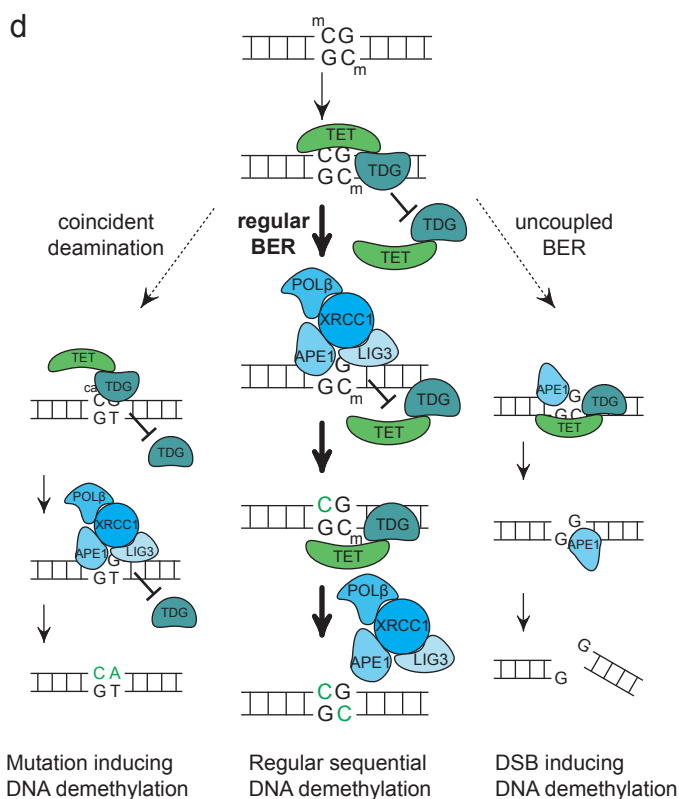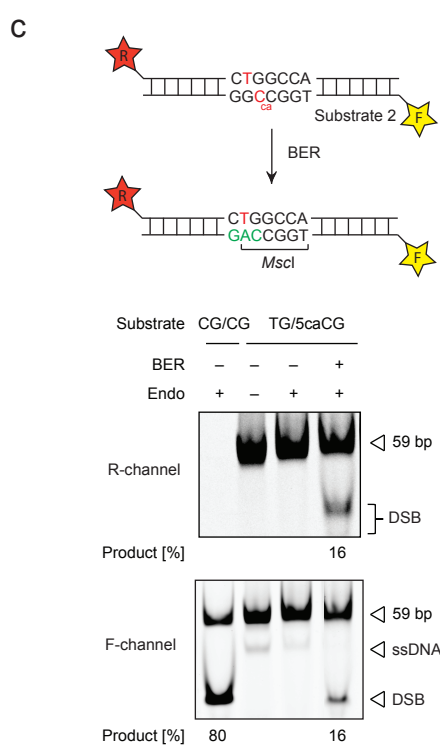
deamination).

a

b

c

a

Substrate [50 nM]	CG/5mCG						
TET1$_{CD}$/TDG [100 nM]	–	+	+	+	+	+	+
APE1 [200 nM]	–	–	–	+	+	+	+
POLβ [50 nM]	–	–	–	–	+	+	+
T4 lig [0.5 U]	–	–	–	–	–	+	–
XRCC1/LIG3 [100 nM]	–	–	–	–	–	–	+
NaOH [100 mM]	–	–	+	–	–	–	–
dCTP [100 µM]	–	–	–	+	+	+	+
Lower band [%]		3	15	14	16	4	4

60mer

24
23
23-P

b

Substrate	CG/CG	CG/5mCG	
TET1$_{CD}$/BER	–	–	+
Product [%]	85		18

59 bp
ssDNA
DSB

a

b

c

d

coincident deamination

regular BER

uncoupled BER

Mutation inducing DNA demethylation

Regular sequential DNA demethylation

DSB inducing DNA demethylation
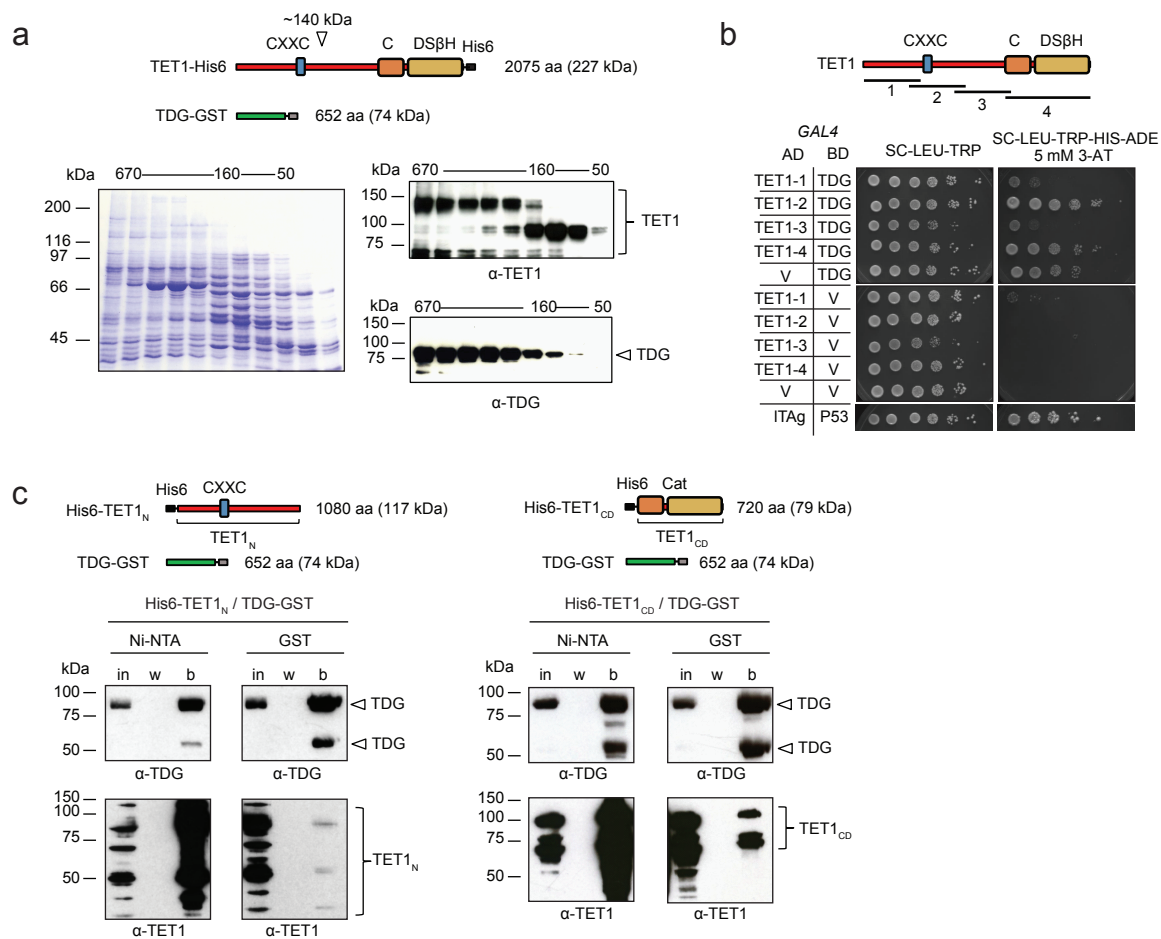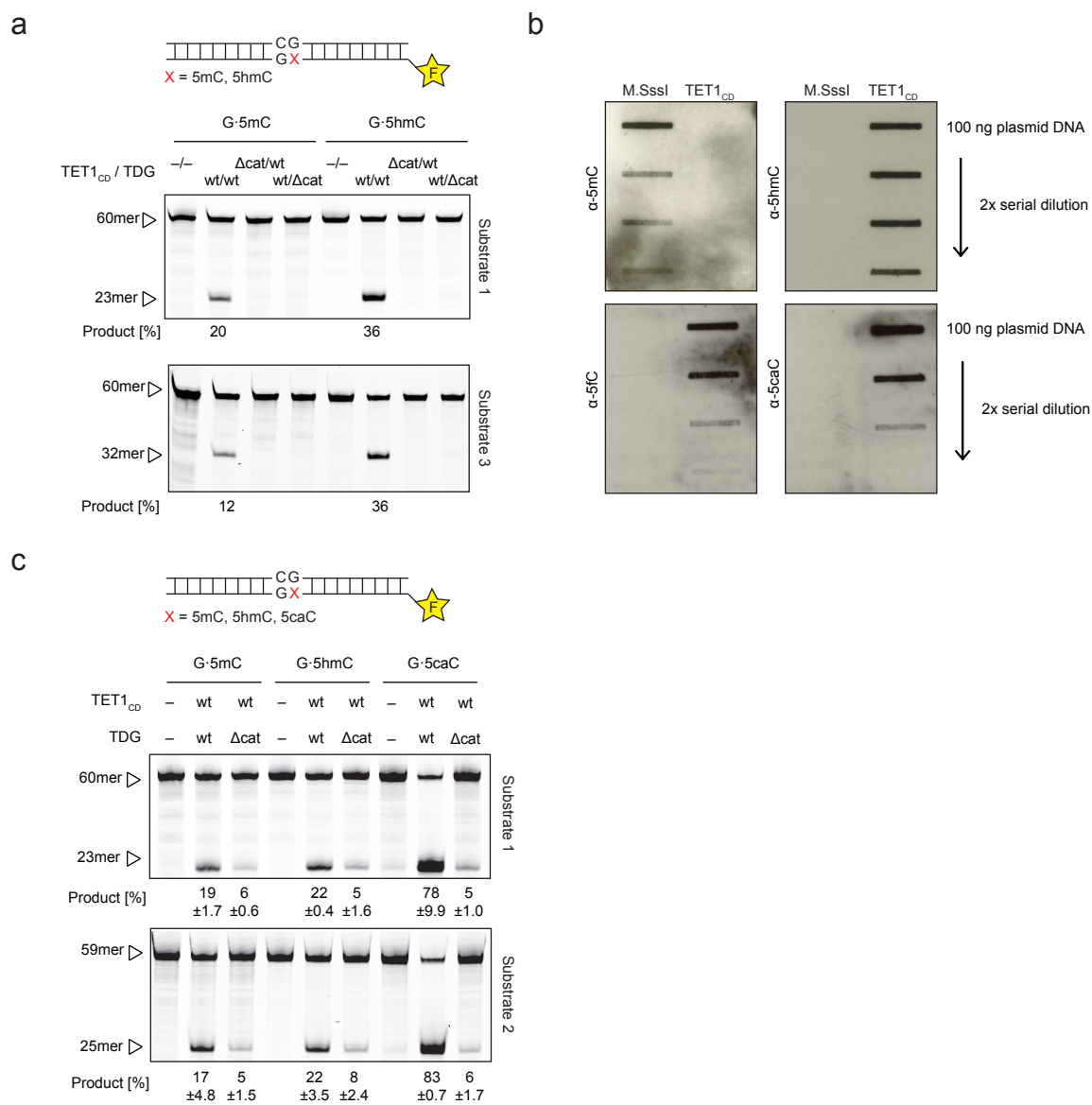
Supplementary Information

Biochemical Reconstitution of TET1-TDG-BER Dependent Active DNA Demethylation Reveals a Highly Coordinated Mechanism

Alain R. Weber[1], Claudia Krawczyk[1], Adam B. Robertson[2], Anna Kusnierczyk[3], Cathrine B. Vågbø[3], David Schuermann[1], Arne Klungland[2] and Primo Schär[1]*

[1] Department of Biomedicine, University of Basel, Basel, Switzerland

[2] Institute of Medical Microbiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway

[3] Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

* To whom correspondence should be addressed. Tel: +41 61 695 3060; Fax: +41 61 267 3566; Email: primo.schaer@unibas.ch

SUPPLEMENTARY DATA

METHODS

Double-strand break assay

Double-strand break assays were carried out in 20 μL reaction volumes containing incision buffer (50 mM HEPES pH 8, 70 mM KCl, 7 mM MgCl$_2$, 500 μg/mL BSA, 1 mM DTT), 0.5 pmol of labeled substrate, 1 pmol APE1 and 1 pmol TDG. After incubation at 37°C for 30 min proteinase K was added to a final concentration of 100 μg/mL and the reaction was incubated at 37°C for another 30 min. Samples were then separated in 8% native polyacrylamide gels and detected and quantified as described above.

FIGURE LEGENDS

Figure S1: Ni-NTA affinity purification of *E.coli* lysates co-expressing His6-TET1$_{CD}$ and TDG-GST. After co-expression of His6-TET1$_{CD}$ and TDG-GST (250 μM IPTG, 25°C for 3 h) cells were lysed and Ni-NTA affinity purification was carried out using Ni-NTA Sepharose beads according to the manufacturer's instructions. Fractions were analyzed by SDS-PAGE followed by (**A**) Coomassie staining or (**B**) immunoblotting using anti TET1 antibody (Millipore). An immunoblot of TDG from Ni-NTA co-affinity purification can be found in Fig. 1. in = input, f = flow, w = wash, e = elution.

Figure S2: Purified His6-TET1$_{CD}$ and BER proteins used in biochemical assays. Purified His6-TET1$_{CD}$, His6-TDG, APE1-His6, His6-POLβ, LIG3/XRCC1-His6 used in the biochemical assays were analyzed by SDS-PAGE followed by Coomassie staining. Purification procedures are described in materials and methods section. Except for LIG3 the affinity tags were not cleaved off. * catalytically inactive C-terminal TET1$_{CD}$ fragment.

Figure S3: Gadd45a has no stimulatory effect on TET1$_{CD}$ when in complex with TDGΔcat. *In vitro* methylated pUC19 plasmid was treated with a preassembled His6-TET1$_{CD}$/His6-TDGΔcat or His6-TET1$_{CD}$/His6-TDGΔcat/His6-Gadd45a complex and resulting cytosine modifications were analyzed by LC/MS/MS according to materials and methods. (n = 3, mean values with standard deviations).

Figure S4: TDG/APE1 induce DNA DSBs at symmetrically modified CpGs. Base release assay using TDG and APE1 on a labeled 59 bp substrate containing either a single 5caC or a symmetrically 5caC modified base pair within a *Hpa*II recognition site (CCGG). Reactions were directly analyzed with native polyacrylamide gel electrophoresis to monitor the appearance of DSBs. The positions of the 59 bp substrate DNA and product fragment are indicated.

Figure S5: Processing of differentially modified CpGs by TDG. Base release assay with His6-TDG on labeled 60 bp substrates containing T on the labeled top strand (5' TexasRed, TR) and 5caC or C on the labeled bottom strand (5' fluorescein, FAM) within the CpG context, illustrated at the top. Reactions were carried out at 37°C for 15 min with the indicated enzyme/substrate ratio, stopped by the addition of NaOH and separated by denaturing gel electrophoresis. Both strands were then visualized by fluorescent scanning and quantified.

a

b

α-TET1

TET1$_{CD}$/TDGΔcat vs TET1$_{CD}$/TDGΔcat/Gadd45a

TET1$_{CD}$/TDGΔcat
▲ 5-me(dC)
▼ 5-hm(dC)
TET1$_{CD}$/TDGΔcat/Gadd45a
● 5-me(dC)
■ 5-hm(dC)

TET1$_{CD}$/TDGΔcat vs TET1$_{CD}$/TDGΔcat/Gadd45a

TET1$_{CD}$/TDGΔcat
▲ 5-f(dC)
▼ 5-ca(dC)
TET1$_{CD}$/TDGΔcat/Gadd45a
● 5-f(dC)
■ 5-ca(dC)
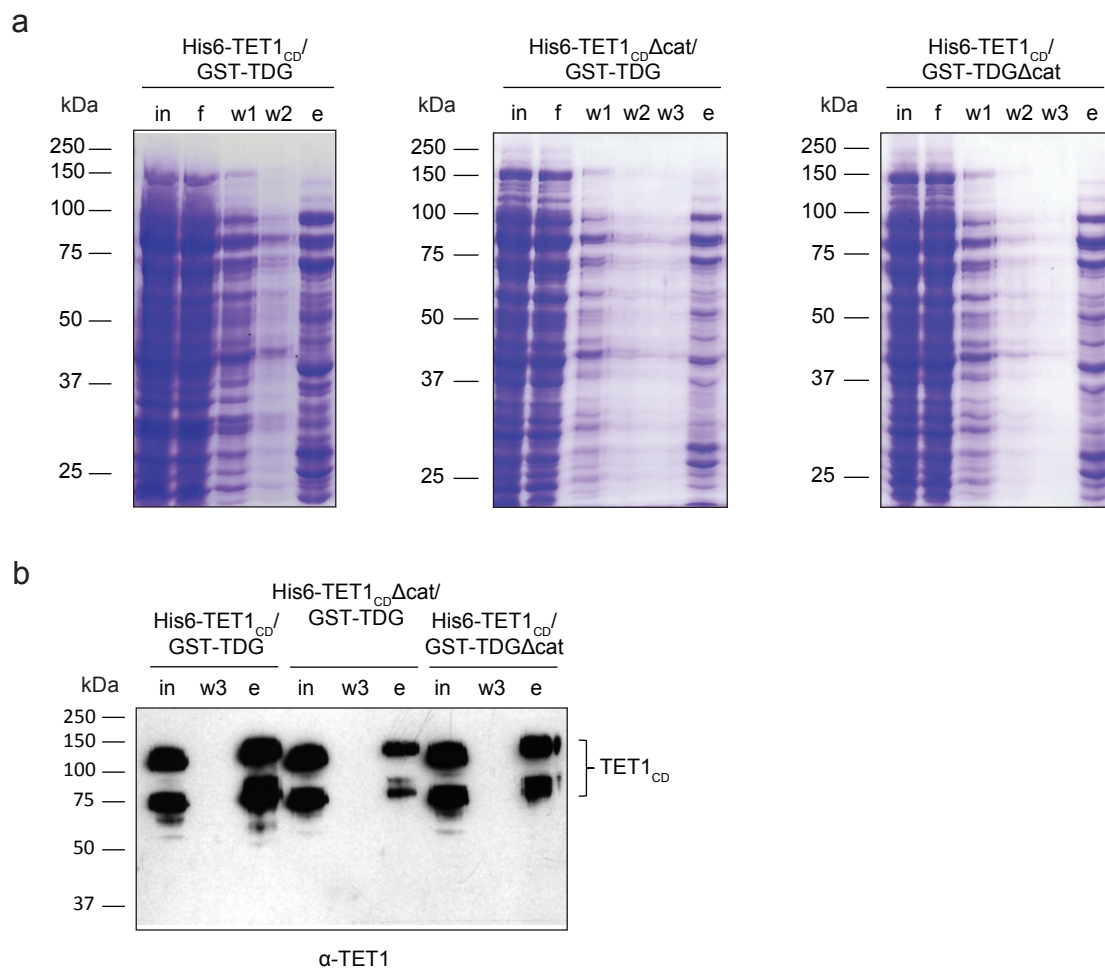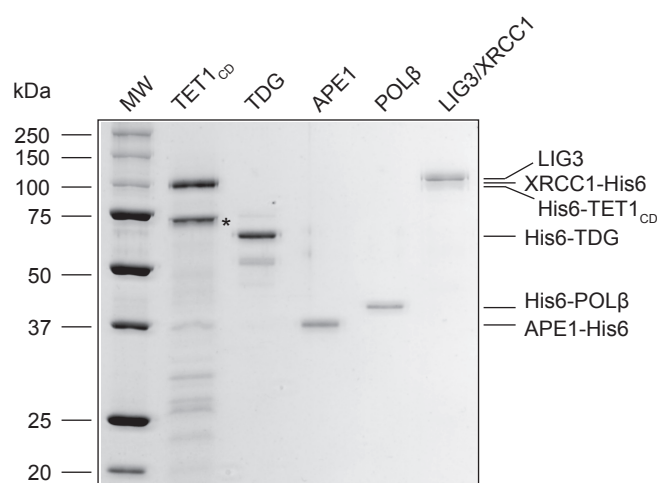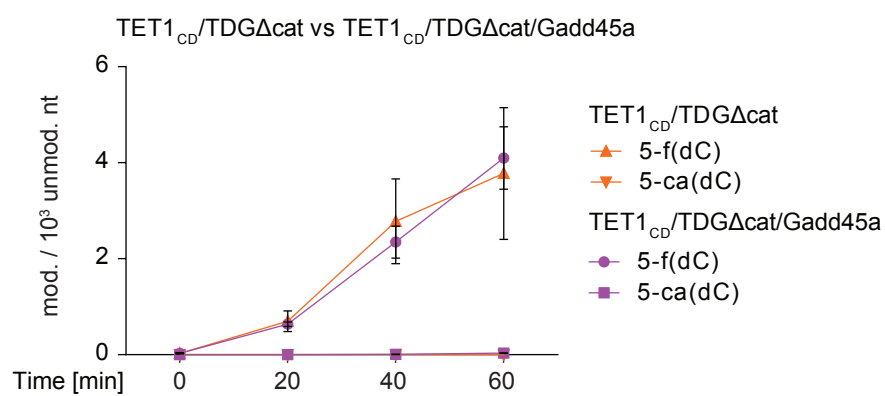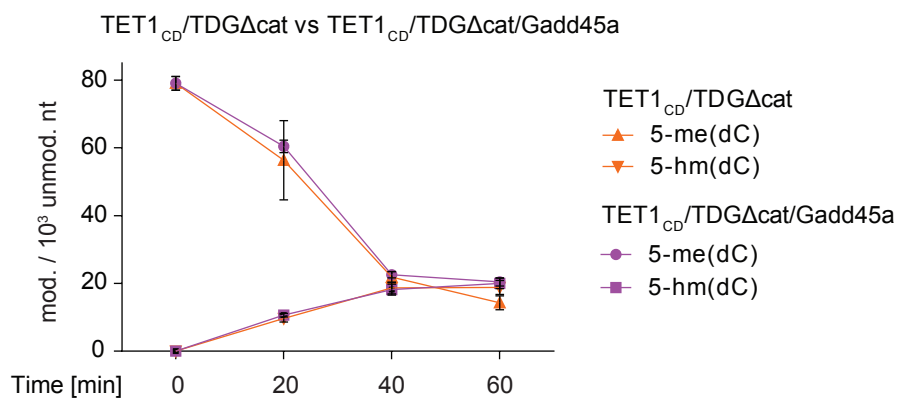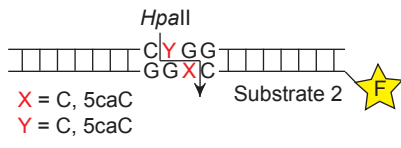
Substrate [25 nM]: CG/CG, CG/5caCG, 5caCG/5caCG

TDG [50 nM]: − − − + − + − + − +
APE1 [50 nM]: − − − − + + − − + +
HpaII: − + − − − − − − − −
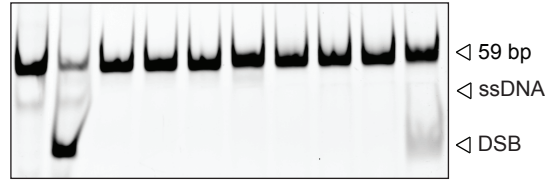
59 bp
ssDNA
DSB

X = C, 5caC
Y = C, 5caC

HpaII

Substrate 2

Product [%]: 83, 15

Y = T
X = C, 5caC

Substrate 1

Top (5'-TR)	T			T		
Bottom (5'-FAM)	C			5caC		
TDG:substrate	−	1	4	−	1	4

R-channel
◁ 60mer
◁ 35mer

Product [%]	4	52	78	5	5	7

F-channel
◁ 60mer
◁ 23mer

Product [%]				1	39	77

PLOS ONE
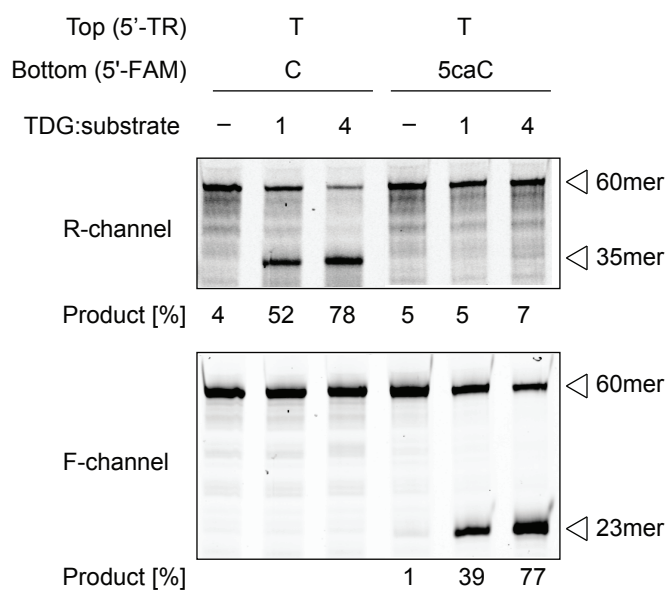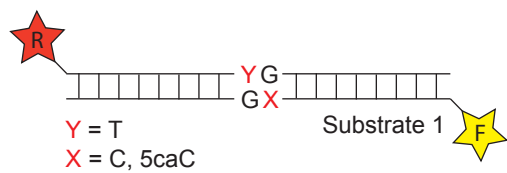
Versatile Recombinant SUMOylation System for the Production of SUMO-Modified Protein

Alain R. Weber, David Schuermann*, Primo Schär*

Department of Biomedicine, University of Basel, Basel, Switzerland

Abstract

Posttranslational modification by small ubiquitin-like modifiers (SUMO) is being associated with a growing number of regulatory functions in diverse cellular processes. The biochemical investigation into the underlying molecular mechanisms, however, has been lagging behind due to the difficulty to generate sufficient amounts of recombinant SUMOylated proteins. Here, we present two newly designed two-component vector systems for the expression and purification of SUMO-modified target proteins in *Escherichia coli*. One system consists of a vector for SUMO conjugation, expressing human SUMO-activating (SAE1/SAE2) and conjugating (Ubc9) enzymes together with His$_6$-tagged SUMO1, 2 or 3, that can be combined with commonly used expression constructs for any gene of interest. To facilitate SUMOylation of targets normally requiring a SUMO-E3 ligase for efficient modification, a second system is designed to express the target protein as a fusion with the human SUMO-conjugating enzyme Ubc9, thus compensating the absence of a potential SUMO ligase. We demonstrate the proficiency of these systems by SUMOylation of two DNA repair proteins, the thymine DNA glycosylase (TDG) and XRCC1, and describe purification schemes for SUMOylated proteins in native and active form. This SUMO toolbox facilitates ''in-cell'' and ''in-extract'' production and purification of recombinant SUMO-modified target proteins for functional and structural analysis.

Introduction

Posttranslational modification by ubiquitin-like polypeptides, so-called UBLs, affects a large number of proteins, thereby regulating a variety of cellular processes [1,2]. The SUMO (small ubiquitin-like modifier) peptides represent a prominent subfamily of the UBLs and exist in four different isoforms (SUMO1, SUMO2, SUMO3 and SUMO4) in mammalian cells, each encoded by a different gene. Although these SUMOs differ to some extent in their amino acid sequences – SUMO2 and SUMO3 share sequence identity of 97% with each other and about 50% with SUMO1 – they all show high 3D-structural resemblance [3–6]. SUMOs, as all UBLs, are attached to their target proteins by a sequence of enzymatic reactions resembling those of ubiquitin conjugation [7,8], involving a heterodimeric activating enzyme E1 (SAE1/SAE2), a single conjugating enzyme E2 (Ubc9) and, in some cases, an E3 protein ligase (Figure 1A). SUMO itself is first synthesized as a precursor peptide that is then trimmed by a SUMO-specific isopeptidase (sentrin-specific proteases; SenPs) to expose an internal glycine-glycine (GG) motif at the C-terminus. The carboxyl group of this mature SUMO peptide is then linked via a thioester to a cysteine residue in SAE2 in an ATP-dependent manner [9]. Subsequently, the activated SUMO is transferred to a cysteine residue of the SUMO-conjugating enzyme Ubc9 [10]. Ubc9 can recognize substrate proteins directly [11] and catalyze the formation of a peptide bond involving the C-terminal carboxyl group of SUMO and an ε-amino group of a target lysine within the SUMOylation consensus motif ΨKxE (Ψ, hydrophobic residue; x, any residue) of the substrate protein [12]. Often, however, SUMO conjugation is additionally promoted by SUMO-E3 ligases, which act as substrate-specific adapters (Figure 1A).

SUMO modification concerns a wide spectrum of target proteins, implicating functions in a variety of vital biological processes such as the cellular response to DNA replication stress, the repair of DNA damage, the regulation of gene expression and epigenetic DNA and histone modifications [2,5,13]. Although research into protein SUMOylation has identified a large number of targets over recent years, there is limited insight into the functional consequences of the modification. Investigations into the immediate biochemical and structural impact of SUMO modification have been a challenge due to difficulties in producing SUMO-modified proteins in sufficient amounts and homogeneity. Enrichment of native endogenous SUMOylated proteins by cell fractionation is generally limited by the low abundance of such proteins and the action of efficient SUMO proteases [14]. *In vitro* modification of enriched target proteins with recombinant SUMOylation enzymes is a more promising approach, typically yielding mixtures of modified and unmodified target proteins contaminated with the SUMOylation enzymes, hence requiring subsequent purification steps [15,16]. Also, co-expression of
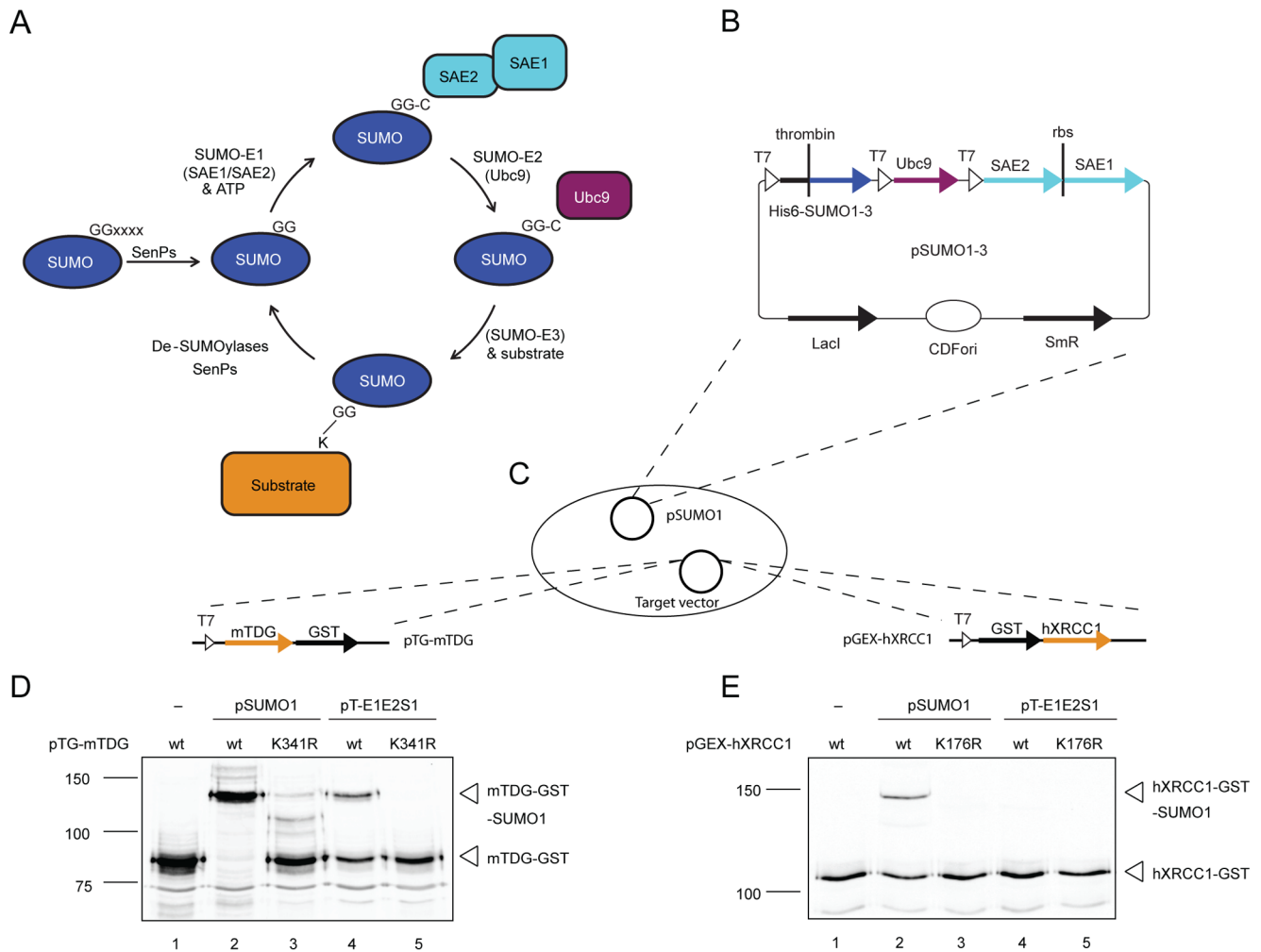
Figure 1. The pSUMO-based SUMOylation system modifies proteins in *E.coli*. (**A**) Scheme of the *in vivo* SUMO maturation, SUMO conjugation and deconjugation process (for detailed description, see ''Introduction''). (**B**) pSUMO vectors containing the humanized SUMOylation system consisting of N-terminally His$_6$-tagged SUMO1, 2 or 3, the SUMO-conjugating enzyme E2 (Ubc9) and both subunits of the SUMO-activating enzyme (SAE1 and SAE2) as a cistronic expression unit with an internal ribosomal binding site (rbs). Expression of the respective cDNAs is under the control of a lac-repressor (LacI) regulated T7 promoter. (**C**) Scheme of the experimental setup of the pSUMO-based in-cell SUMO conjugation. *E.coli* BL21 cells were used containing pSUMO1 in combination with pTG-mTDG or pGEX-hXRCC1 plasmids were used for the co-expression of the complete SUMO system with C- and N-terminally GST-tagged mTDG and hXRCC1, respectively. Immunoblot analyses of mTDG (**D**) and hXRCC1 (**E**) SUMOylation in *E.coli* cells, expressing the SUMO target and the SUMO system from either the pSUMO1 or the pT-E1E2S1 plasmid (250 μM IPTG at 25°C for 2 h). Co-expression of target proteins mutated in the SUMO acceptor sites of mTDG (K341R) and hXRCC1 (K176R) were included to assess the specificity of the SUMOylation system.
doi:10.1371/journal.pone.0102157.g001

SUMO targets with SUMO1 and SUMO-activating and -conjugating enzymes of different origin (human, mouse, *Xenopus laevis*) in *E.coli* was shown to produce SUMO-modified protein [17–19], but in our experience, the yields of specifically modified proteins were often poor, impeding efficient purification and subsequent biochemical analysis.

To streamline the production of homogeneously SUMOylated recombinant proteins for biochemical and structural studies, we set out to establish an optimized and versatile SUMOylation system, coupling efficient SUMO-conjugation with affinity purification of modified target proteins. We designed and experimentally validated two alternative two-component vector systems for simultaneous expression of mature SUMO1, 2 or 3 polypeptides, SUMO-E1, SUMO-E2 and a target protein of interest in *E.coli*. In contrast to previous approaches [18–20], we used SUMOylation enzymes of human origin only, physically separated the SAE1 and

SAE2 E1 subunits and added a protease-cleavable His$_6$-tag to the SUMOs to facilitate purification of modified protein. To overcome a possible rate limitation by the absence of an appropriate SUMO-E3 ligase in the *E.coli* system, one system is designed to express the target protein as fusion to the SUMO-conjugating enzyme Ubc9, a strategy that was successfully applied for SUMOylation of ectopically expressed p53 and STAT1 in HEK293, HeLa, COS-7 and CHO cells [21,22]. We evaluated these newly developed vector systems with the DNA base excision repair (BER) enzymes thymine DNA glycosylase (TDG) and XRCC1. TDG is a well-studied SUMO target [23]; SUMO modification of TDG was shown to effect conformational changes that promote enzymatic turnover [24] and may also regulate the subcellular localization [25]. XRCC1 acts as central scaffold factor in BER [26] and was identified as a putative SUMO target in an *in vitro* screening approach. It was found SUMOylated in HeLa cells

following heat shock treatment [27,28], but the function of this modification remains to be elucidated. We assessed the efficiency and specificity of SUMO modification of these proteins in our recombinant systems as well as the proficiency of the purification procedure to generate biologically active SUMO proteins. Finally, we provide guidance to optimize the experimental setup and conditions for the SUMOylation of any target protein in bacteria, discussing the use of the different SUMOylation vectors and expression strategies for "in-cell" or "in-extract" SUMOylation.

Materials and Methods

Vector construction

The plasmids with the complete SUMO system (pSUMO1-3), the SUMO-activating (pSA1-3) and the SUMO-conjugating (pSC-PreE2/IntE2) plasmids were assembled by standard cloning methods based on PCR amplification with adaptor oligonucleotides providing suitable restriction sites. The plasmid DNA, vector sequences and maps are available from Addgene (http://www.addgene.org), plasmid ID 52258-52284. The cDNAs of the human SUMOylation components SAE1, SAE2, Ubc9 and SUMO1-3 were amplified from pGEX-based bacterial expression vectors kindly provided by R. Hay and M. Hottiger. pSUMO1-3, pSA1-3 and pSC-PreE2/IntE2 vectors are based on pCDFDuet-1 (Novagen) and pTXB3 (New England BioLabs), respectively. The cDNAs of the SUMO target proteins were amplified by adaptor PCR and introduced into the NcoI and EcoRI site (hXRCC1, hTDG) or NcoI and XhoI (mTDG) of the pSC-IntE2/PreE2 vectors. The consensus SUMOylation motif (VKEE) was deleted by site-directed mutagenesis in hTDG (K330A), mTDG (K341R), hXRCC1 (K176R).

Recombinant protein expression, in-cell SUMOylation and cell lysis

The expression vectors were introduced into *E.coli* BL21(DE3) cells by electroporation. Overnight pre-cultures were diluted with fresh pre-warmed LB medium and grown at 30°C to OD_{600} levels as indicated. Cultures were grown under selective pressure using either 100 mg/L of Ampicillin or 50 mg/L of Streptomycin for single plasmid expressions and half the concentration of each antibiotic when co-expressing two plasmids. Protein expression was induced by the addition of the isopropyl-β-D-thiogalactose (IPTG) to the final concentrations as specified in the results and cultures were further incubated as indicated. Finally, cells were harvested by centrifugation and soluble protein fractions were extracted by sonication in lysis buffer (50 mM Na-phosphate buffer pH 7.5, 300 mM NaCl, 10% glycerol, 0.1% Tween-20, 1 mM DTT, 1 mM PMSF), if not stated otherwise. Crude lysates were then cleared by centrifugation with >20'000 g at 4°C for 30 min.

Purification of SUMOylated protein

Small-scale protein preparations were performed with Glutathione sepharose HP (GE Healthcare) or cOmplete His-tag purification (Roche Applied Science) resins. Cleared lysates were incubated with 100 μL resin in lysis buffer at 4°C for 3 h, prior to loading onto gravity flow columns (BioRad). Unbound proteins were washed out with 20 and 10 bed volumes of lysis buffer with 0.3 and 1 M NaCl, respectively. After a final wash step with 10 bed volumes of lysis buffer, bound proteins were eluted by the addition of 250 mM imidazole or 10 mM reduced glutathione to the lysis buffer. Large-scale protein purification was carried out on an ÄKTA purifier 10 system using pre-packed columns (GE Healthcare). To enrich for SUMOylated proteins, cell lysates

were loaded onto a 5 mL HisTrap crude FF column (GE Healthcare), washed with an imidazole gradient from 0 to 40 mM over 2 column volumes (CV) and bound proteins were eluted with 10 CV lysis buffer containing 400 mM imidazole. Peak fractions were pooled and dialyzed 3 times for 30 min against 300 mL GST loading buffer (50 mM Na-phosphate buffer pH 7.5, 500 mM NaCl, 15% glycerol, 0.1% Tween-20, 1 mM DTT, 0.5 mM PMSF) and loaded on a 1 mL GSTrap HP column (GE Healthcare). Unbound protein was washed out by a NaCl gradient from 0.5 to 1 M over 10 CV. Subsequently, bound target protein was released from the GST-Ubc9 moiety either by induced self-splicing at 4°C in cleavage buffer (50 mM Tris-HCl pH 8, 500 mM NaCl, 15% glycerol, 0.1% Tween-20, 0.1 mM PMSF, 50 mM DTT) for 16 h or by the application of 80 U of PreScission protease (GE Healthcare) according to the manufacturer's instructions. Cleaved protein was eluted, dialyzed 3 times for 30 min against 300 mL storage buffer (20 mM Tris-HCl pH 8, 50 mM NaCl, 10% glycerol, 1 mM DTT), snap-frozen and stored at −80°C.

Analytical gel electrophoresis, western blotting and protein detection

Protein fractions were analyzed by standard SDS-polyacrylamide gel electrophoresis (SDS-PAGE) followed by Coomassie blue staining or by immunoblotting using the Odyssey imaging system (LI-COR Biosciences) or chemiluminescence (WesternBright ECL, Advansta) according to the provider's protocol. Antibodies were diluted in 5% non-fat dry milk TBS (100 mM Tris-HCl pH 8, 150 mM NaCl) supplemented with 0.2% Tween-20: human TDG, rabbit polyclonal ab 141 (raised against recombinant full-length hTDG), 1:5'000; mouse TDG, rabbit polyclonal ab L58 (raised against recombinant full-length mTDG), 1:5'000; XRCC1, rabbit polyclonal ab (Sigma-Aldrich X0629), 1:2'000 and mouse monoclonal ab (33-2-5; Abcam ab1838), 1:1'000; SUMO1, mouse monoclonal α-GMP1 ab (21C7; Life Technologies 33-2400), 1:1'000 and rabbit polyclonal α-SUMO1 ab (Sigma-Aldrich S8070) 1:1'000.

Base release assay

A 60 bp heteroduplex DNA containing a G·U mismatch was prepared by annealing an unlabeled oligonucleotide (5′-TAGA-CATTGC CCTCGAGGTA CCATGGATCC GATGTCGACC TCAAACCTAG ACGAATTCCG-3′) to a 5′-fluorescein-labeled uracil-containing oligonucleotide (5′-ATCTGTAACG GGAGCTCCAT GGTACCTAGG CTACAGUTGG AGTTTGGATC TGCTTAAGGC-3′) by heating to 95°C for 5 min and gradual cooling to 25°C with a rate of 0.02°C/s.

Reactions were carried out in 20 μL nicking buffer (50 mM Tris-HCl pH 8, 1 mM EDTA, 1 mM DTT, 0.1 mg/mL BSA) containing 5 pmol of substrate DNA and 0.5 pmol of TDG protein, unless stated otherwise, at 37°C for the indicated time periods. AP-sites were then chemically cleaved by the addition of 2 μL of 1 M NaOH and boiling at 99°C for 10 min. DNA was precipitated overnight at −20°C after adding 2.2 μL of 3 M Na acetate pH 5.2, 0.5 μL yeast tRNA (10 mg/mL) and 67.5 μL ethanol. Subsequently, DNA was pelleted by centrifugation, washed with 70% ethanol, air-dried for 10 min, resuspended in 10 μL gel loading buffer (1×TBE, 90% formamide), heated at 99°C for 5 min and loaded on a 15% denaturing polyacrylamide gel (8 M urea, 1× TBE) for analysis. Gels were run at 13 V/cm for 30 min and labelled DNA was detected using the blue fluorescence mode of the Typhoon 9400 (GE Healthcare) and analyzed quantitatively by ImageQuant TL software (v7.0, GE Healthcare).

SUMOylation and de-SUMOylation assays

In vitro SUMOylation with purified recombinant protein was carried out in 50 µL SUMOylation buffer (50 mM Tris-HCl pH 8, 50 mM NaCl, 10% glycerol, 0.5 mM DTT, 5 mM MgCl$_2$, 5 mM ATP), containing 80 pmol SUMO1, 16 pmol Ubc9, 4 pmol SAE1/SAE2 and 16 pmol target protein. Reactions were incubated at 30°C for 30 min. De-SUMOylation was carried out in SenP2 buffer (25 mM Tris-HCl pH 8, 150 mM NaCl, 2 mM DTT, 0.1% Tween-20) using an excess amount of SenP2 protease. The reaction mixture was incubated at RT for 30 min.

For in-extract SUMOylation, cells expressing pSUMO1, pSA1 and target vectors were lysed in SUMOylation buffer without ATP and mixed at indicated volume ratios. SUMOylation was triggered by the addition of ATP to a final concentration of 5 mM and reaction mixtures were incubated at 30°C for 1 h.

Results

Recombinant human SUMOylation system modifies target proteins in *E.coli*

To provide a humanized SUMO-E1/E2 conjugation system for modification of target proteins in a recombinant bacterial expression setup, we constructed a series of CDFori-based vectors (pSUMO1-3), expressing the SUMO-E1 and -E2 enzymes as well as the mature SUMO1, 2 or 3 polypeptides (C-terminal GG motif) under the control of the phage T7 promoter (Figure 1B). Unique to this system is that it bases on human proteins only, expresses the heterodimeric SAE1-SAE2 complex (SUMO-E1) from a bi-cistronic unit and provides SUMO polypeptides with an N-terminal His$_6$-tag separated by a thrombin cleavage site, facilitating the enrichment of modified proteins by affinity chromatography and the removal of the affinity tag.

We validated the functionality of the system by co-expression with an established and a postulated SUMOylation target of the BER pathway, TDG and XRCC1. To ensure a stable maintenance of the pSUMO1 and the target expression constructs in *E.coli*, we chose pMB1ori-based plasmids for the expression of the mouse TDG (mTDG) and the human XRCC1 (hXRCC1) (Figure 1C). First, we compared our pSUMO1 with the previously published pT-E1E2S1 vector [18] for the efficiency to produce SUMO1-conjugated C-terminally GST-tagged mTDG (pTG-mTDG) and N-terminally GST-tagged hXRCC1 (pGEX-hXRCC1) when co-expressed with the recombinant SUMO system in *E.coli* BL21(DE3) cells. Under the applied experimental conditions, we observed a significantly higher efficiency of SUMO1 modification with the newly designed pSUMO1 vector for both substrates, yielding nearly 100% SUMOylated mTDG and about 20% SUMOylated hXRCC1 (Figure 1D and E, compare lanes 2 and 4). The highly efficient modification in the presence of pSUMO1 may generate some unspecific SUMO conjugation as evidenced by the low amount of mis-targeted modification notable with an mTDG mutated in the major SUMO acceptor lysine (mTDG-K341R) (Figure 1D, lane 3). hXRCC1 mutated in its predicted SUMO acceptor lysine (hXRCC1-K176R; unpublished information kindly provided by Roland Steinacher), however, showed no detectable SUMOylation (Figure 1E, lane 3), thus demonstrating the selectivity of pSUMO1-mediated SUMO modification.

The SUMO conjugation system presented here generates SUMOylated products with GST- and His$_6$-tags fused to the target protein and the SUMO polypeptide, respectively. Purification of the modified target can thus be achieved through successive GST and Ni-NTA affinity chromatography steps (Figure 2A) as shown here for mTDG and hXRCC1. Modification of the target

proteins was obtained by induced in-cell SUMOylation in *E.coli* at 25°C for 2 hours, co-transformed with pSUMO1 and pTG-mTDG or pGEX-hXRCC1. Yields of SUMO-modified mTDG and hXRCC1 proteins were estimated by stained analytic SDS-PAGE and found to be around 5 and 1.5 mg per liter bacterial culture, respectively. Thus, substantial amounts of recombinant proteins are expected to be purifiable from the in-cell SUMOylation system.

SUMOylated proteins were enriched by fractionation of the *E.coli* lysates using either sequential GST and Ni-NTA affinity chromatography (work flow 1) or vice versa (work flow 2) (Figure 2A). Following purification work flow 1, both, unmodified and modified mTDG and hXRCC1 were enriched in the elution fractions of the GST affinity column as detected by SDS-PAGE analysis followed by Coomassie staining (Figure 2B and F, GST lanes e1 and e2) and immunoblotting with anti-mTDG and anti-hXRCC1 antibodies (Figure 2E and G). Applying the pooled GST elutions to a Ni-NTA affinity column led to a further enrichment of the SUMOylated protein fractions. mTDG eluted from the column as homogeneously SUMOylated protein fraction (Figure 2B and E, Ni-NTA lanes e1 and e2). A prominent protein, however, migrating at about 20 kDa co-eluted in the main fraction and turned out to be free SUMO1 (Figure 2C). The Ni-NTA step also enriched the proportion of SUMO1-modified hXRCC1 but did not separate it entirely from unmodified hXRCC1 (Figure 2F and G, Ni-NTA lanes e1 and e2). This may be due to the propensity of hXRCC1 to dimerize through its BRCT domain under purification conditions [29], SUMO-mediated protein-protein interactions or a possible dimerization of the GST-tag, thus forming hXRCC1-SUMO1/hXRCC1 heterodimers. Following work flow 2, we observed an efficient enrichment of SUMO1-modified mTDG but also of free SUMO1 and probably some *E.coli* proteins on the Ni-NTA column (Figure 2D, Ni-NTA lanes e1 and e2). These impurities could then be separated from the SUMO1-modified target protein by GST affinity purification, which yielded homogeneously SUMOylated mTDG (Figure 2D, GST lanes e1 and e2).

A SUMO-E2-fusion system to facilitate SUMOylation of suboptimal targets

Having confirmed the functionality of the humanized SUMOylation system in *E.coli*, we aimed to optimize its robustness for targets, requiring a SUMO-E3 ligase for efficient modification, by expressing the target protein as a fusion with the SUMO-conjugating enzyme [21]. To this end, we split the expression units for SUMOylation into two compatible vectors, one for SUMO activation (SA) and the other for SUMO conjugation (SC). The SUMO-activating vector (pSA1, pSA2, or pSA3) contains expression cassettes encoding mature human SUMO proteins (SUMO1-3), N-terminally fused to a His$_6$-tag separated by a thrombin cleavage site, as well as both subunits of the human SUMO-activating enzyme E1 (SAE1 and SAE2) as a bi-cistronic unit (Figure 3A). The SUMO-conjugating vectors pSC-PreE2 (Figure 3E) and pSC-IntE2 (Figure 3C) were designed to express the target protein with a C-terminal fusion to the SUMO-E2 enzyme Ubc9 and the GST-tag. The inclusion of the PreScission cleavage site in pSC-PreE2 facilitates the specific release of the target proteins from the Ubc9-GST fusion by a protease digestion. The linker in pSC-IntE2 separates the SUMO target from the SUMO-E2 portion through the *Mycobacterium xenopi* GyrA intein sequence (Figure 3C) and facilitates the release of the modified target protein by self-splicing in a reducing environment [30], i.e. without protease treatment. The two constructs also provide alternative TARGET-Ubc9 configurations, should one or the
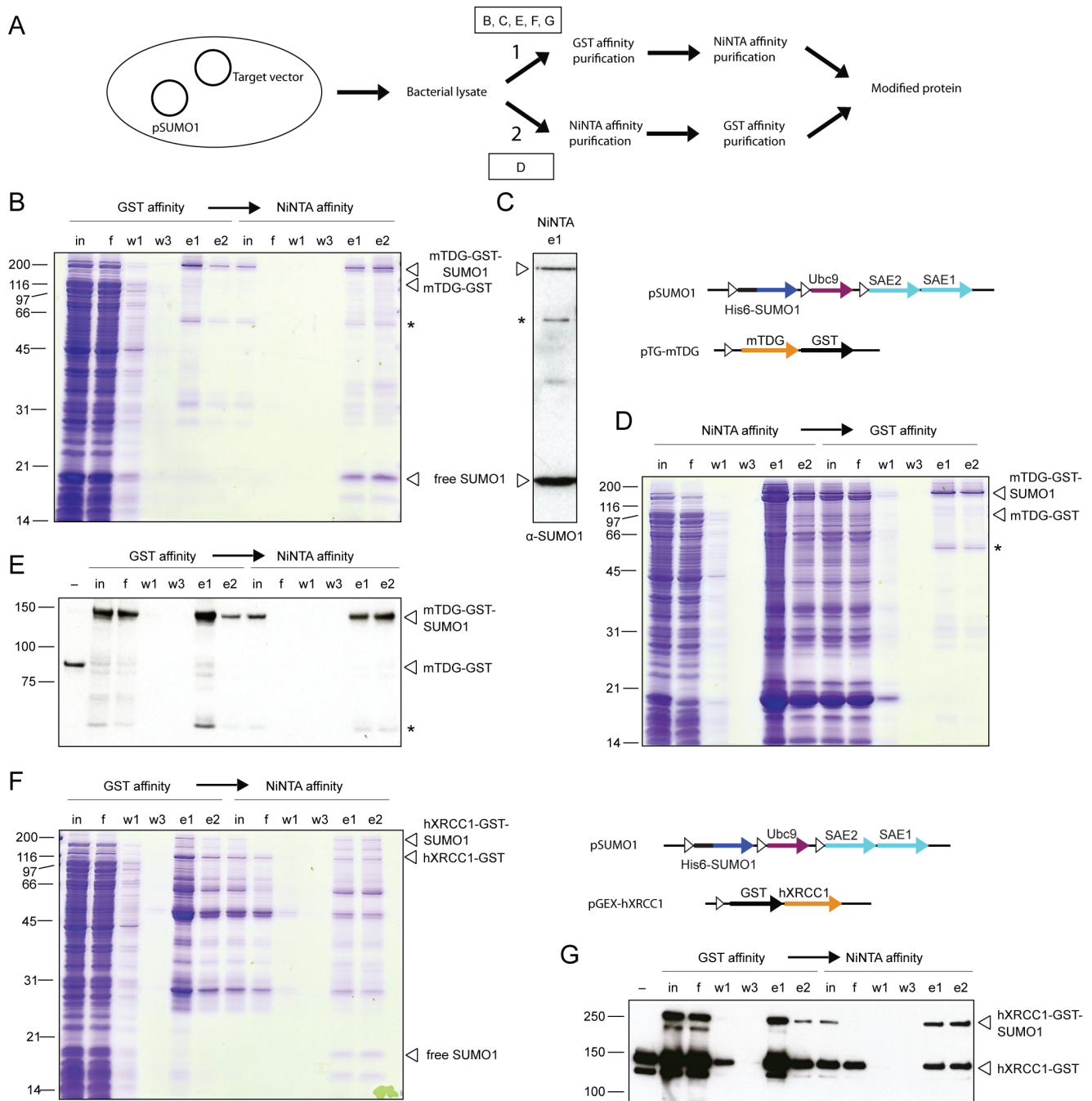
Figure 2. Purification of SUMO1-modified mTDG and hXRCC1 produced by in-cell SUMOylation. (**A**) Purification scheme for in-cell SUMO-modified protein. Cell lysates are subjected to subsequent GST- and Ni-NTA-affinity purification (work flow 1) or vice versa (work flow 2). Boxed letters indicate the corresponding sub-panels. Fractions of the purification of SUMO1-modified mTDG from purification work flow 1 (**B**) and 2 (**D**) and hXRCC1 from work flow 1 (**F**) were analyzed by SDS-PAGE and subsequent Coomassie blue staining and immunoblotting using monoclonal anti-GMP1 (**C**), polyclonal anti-mTDG (**E**) and anti-hXRCC1 (**G**) antibodies, respectively. in, input (cleared lysate or dialyzed elution fractions); f, flow through; w, wash steps; e, elution fractions; *, SUMO1-modified truncated mTDG.
doi:10.1371/journal.pone.0102157.g002

other fusion cause structural constraints that compromise SUMOylation efficiency and specificity.

We first tested the functionality of the SUMO-E2-fusion system in in-cell SUMO conjugation (Figure 3B), using as targets the human TDG (hTDG), which SUMOylates *in vitro* with intermediate efficiency, its mouse ortholog mTDG as a control for high efficient SUMOylation, and hXRCC1 as an inefficiently modified

substrate. We thus introduced the SUMO1-activating (pSA1) and respective SUMO TARGET-E2 vectors (pSC-hTDG-IntE2, pSC-hTDG-PreE2, pSC-mTDG-IntE2, pSC-hXRCC1-IntE2, pSC-hXRCC1-PreE2) into *E.coli* BL21(DE3). Protein expression was then induced with 250 μM and 1 mM IPTG at 15°C for 3 and 6 hours and 37°C for 1 hour, respectively. In the control reactions without SUMO activation, the non-modified full-length
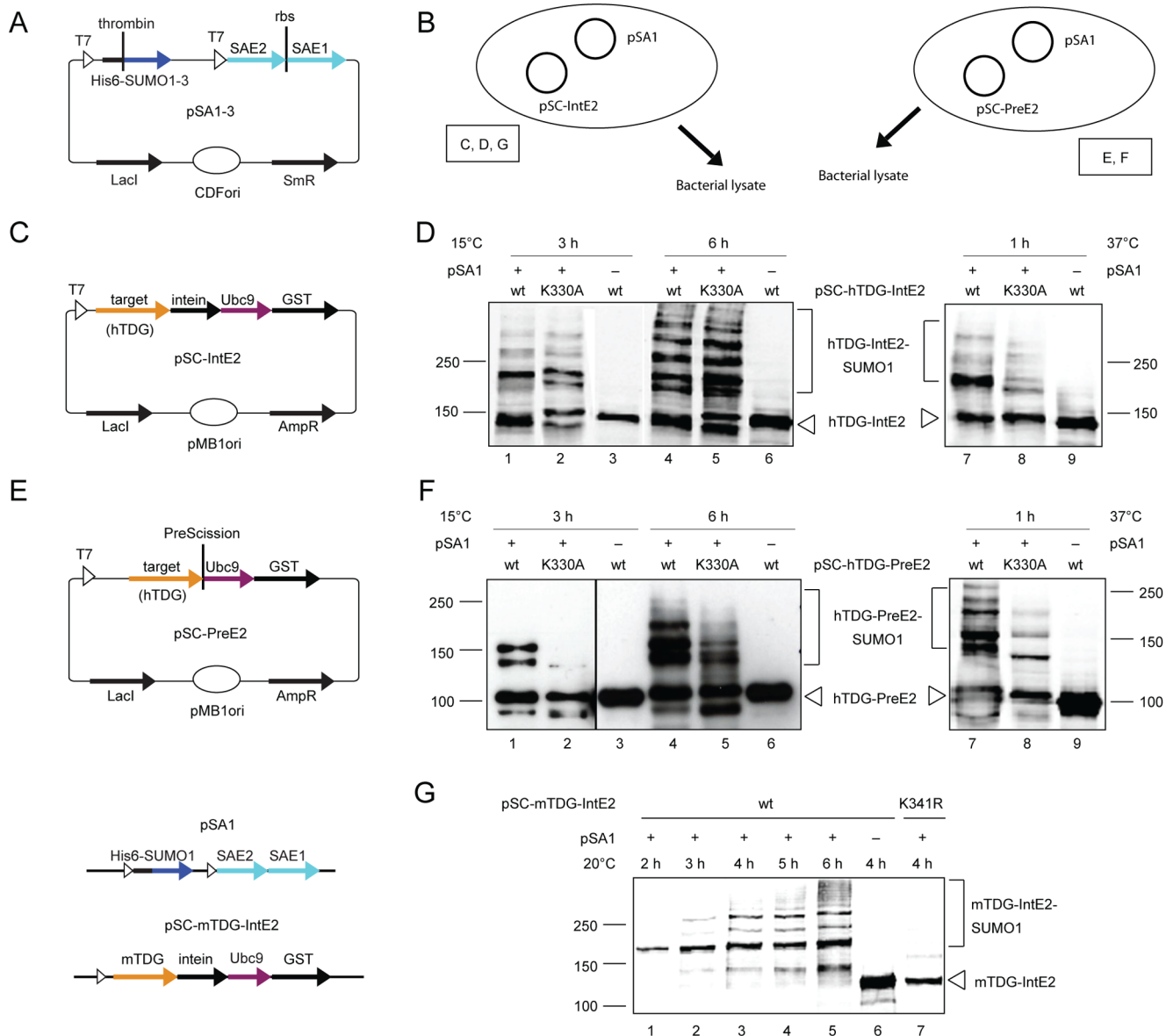
Figure 3. In-cell SUMOylation of TDG with the SUMO-E2-fusion system. Scheme of the SUMO-activating vectors pSA1-3, which are identical to the pSUMO1-3 vectors but lack the Ubc9 expression unit (**A**) and the Ubc9-fusion SUMO-conjugating vectors pSC-PreE2 (**E**) and pSC-IntE2 (**C**) for expression of target protein fused to a GST-tagged Ubc9 under the control of the T7 promoter. A PreScission protease cleavage site or a *Mycobacterium xenopi* GyrA intein sequence in the linker region allows for the release of the modified target from the Ubc9-GST fusion. (**B**) Experimental setup of in-cell SUMO conjugation with the SUMO-E2-fusion system. pSA1 is co-expressed with target proteins either from pSC-IntE2 or pSC-PreE2 vectors in *E.coli* BL21. Boxed letters indicate the corresponding sub-panels. Immunoblot analysis of lysates of *E.coli* cells expressing the SUMO-activating proteins (pSA1) and wild-type (wt) or SUMO acceptor site-mutated (K330A) human TDG (hTDG) from the SUMO-conjugating vectors pSC-IntE2 (**D**) or pSC-PreE2 (**F**). Expression was induced with 250 μM IPTG at 15°C for 3 and 6 h or with 1 mM IPTG at 37°C for 1 h. (**G**) SUMOylation of mouse TDG (mTDG) expressed from the pSC-IntE2 vector was followed over time by immunoblot analysis. Expression was induced with 500 μM IPTG and cells were incubated at 20°C.
doi:10.1371/journal.pone.0102157.g003

hTDG-IntE2 and hTDG-PreE2 fusion proteins appeared as a prominent bands migrating just below 150 kDa and around 100 kDa, respectively (Figure 3D and F, lanes 3, 6 and 9). After 3 hours of co-expression at 15°C, prominent slower migrating polypeptides appeared in cells expressing hTDG-IntE2 or hTDG-PreE2, corresponding to the SUMO1-modified hTDG fusion proteins (Figure 3D and F, lane 1). Similarly efficient and specific modification of hTDG-IntE2 or -PreE2 occurred when co-expression was done for 1 hour at 37°C under strong IPTG induction (Figure 3D and F, lanes 7 and 8). After prolonged

expression for 6 hours at 15°C, additional high molecular weight SUMO modification products became apparent (Figure 3D and F, lane 4). We interpret these to represent hTDG isoforms with multiple SUMO chains or poly-SUMO chains attached, most of which do not form at the major acceptor site and are not normally seen with endogenous hTDG. The majority of these SUMO conjugates also appeared for the hTDG-K330A variant, which is mutated in the major SUMO-acceptor site (Figure 3D and F, lanes 5), hence reflecting mis-targeted modifications at either the hTDG or the fused (intein-)Ubc9-GST polypeptides, which are known

and predicted SUMOylation targets. The generation of SUMO conjugates also appears to be influenced by the configuration of the TARGET-Ubc9 construct as, generally, the hTDG-IntE2 fusion generated more heterogeneity in modification products than the hTDG-PreE2 fusion, most likely because of various modifications in the intein-Ubc9-GST moiety. As with the pSUMO system (Figure 1), in-cell modification of mTDG expressed from the pSC-IntE2 vector was highly efficient, with close to 100% yields of SUMOylated mTDG (Figure 3G, lanes 1–5). At early time-points of induced protein expression (500 μM IPTG, 20°C), this setup produced a single prominent SUMOylated form of mTDG-IntE2. Only upon prolonged expression, we observe slower migrating multi-SUMOylated mTDG. As only little non-specific SUMOylation occurred with an mTDG mutated in the SUMO acceptor site (K341R) (Figure 3G, lane 7), we interpret the slower migrating mTDG-SUMO1 products to predominantly represent SUMO chain formation.

In-extract SUMOylation increases efficiency and flexibility of the SUMO-E2-fusion system

While in-cell SUMOylation with the SUMO-E2-fusion system was highly efficient for both TDGs, the outcome was not satisfactory for hXRCC1 (data not shown), due to inefficient expression of the fusion protein. To work around such constraints, we resorted to a strategy of expressing pSA1 and pSC-IntE2/-PreE2 vectors separately in E.coli cells and performing SUMO modification of target proteins in crude E.coli extracts without prior purification of the necessary SUMOylation factors (Figure 4A).

To assess the potential of in-extract SUMOylation, we expressed hXRCC1 from pSC-PreE2 or pSC-IntE2 and the SUMO activating factors from pSA1 in separate E.coli cultures at 25°C and 30°C, respectively, for 3 hours. The crude lysates of these cultures were then mixed at variable volume ratios and incubated at 30°C in the presence of 5 mM ATP for 1 hour. We then compared the efficiency of in-extract SUMOylation of hXRCC1 by the SUMO-E2-fusion system with that of extracts from cells expressing GST-tagged hXRCC1 (pGEX-hXRCC1) and the pSUMO1 system (Figure 4B and C). The SUMO-E2-fusion system produced a substantial amount of SUMO1-modified hXRCC1-PreE2 protein, which was fully dependent on an intact SUMO acceptor site (Figure 4B). By contrast, when hXRCC1 was provided without the Ubc9 fusion, in-extract SUMOylation was not detectable, even in the presence of an excess of extract providing the complete SUMOylation components (Figure 4C). Although this can be explained partly by a reduced SUMOylation capacity of the pSUMO1 lysate compared to the pSA1 lysate (Figure 4D, compare lane 2 and 3), these results show that the SUMO-E2-fusion system facilitates efficient in-extract SUMOylation of a suboptimal SUMO target like hXRCC1 (Figure 4D, compare lane 1 and 2).

To assess the impact of the configuration of the TARGET-Ubc9 fusion, we also performed in-extract SUMOylation with hXRCC1 and hTDG expressed from the pSC-IntE2 plasmid. We thus prepared crude lysates from E.coli expressing the pSA1 components (30°C, 3 hours induction) and the TARGET-IntE2 fusion (25°C, 3 hours induction) and incubated mixed extracts at different volume ratios at 30°C for 1 hour. This produced an appreciable amount of SUMO1-modified hXRCC1-IntE2 fusion protein (Figure 4E), largely in an ATP-dependent manner. The residual SUMO conjugation, notable without addition of ATP, at a 5-fold excess of activating over conjugating lysates (Figure 4E, lane 5) most likely reflects the pre-existence of a small amount of activated SUMO1 in the extracts. Compared to the hXRCC1-

PreE2 fusion, however, the in-extract SUMOylation of hXRCC1-IntE2 fusion appeared to be less efficient; the maximum yield of SUMOylated product was generally lower for hXRCC1-IntE2 than for the hXRCC1-PreE2 fusion. Consistent with the observations from in-cell modification, in-extract SUMOylation of the TARGET-IntE2 fusions may also be less specific than that of the TARGET-PreE2 fusions, as some residual SUMO modification of hXRCC1 and hTDG mutated in the main SUMO acceptor sites appeared in the presence of an excess of SUMO-activating lysate (Figure 4E, lane 6; and Figure 4F, lane 5). These results indicate that a fusion of the SUMO-E2 enzyme to the target protein can substantially enhance in-extract SUMOylation efficiency and may be useful to compensate rate limitations due to the lack of a proper SUMO-E3 ligase in the recombinant system. Notably, the stimulatory effect of the Ubc9 fusion was less pronounced for hTDG (Figure 4F and G); in-extract SUMOylation with either IntE2-fused or non-fused hTDG generated approximately 50% modified hTDG protein with some tendency to mis-targeted modification. This is consistent with TDG's high propensity of SUMO modification in the absence of an E3 ligase and may reflect a high affinity of hTDG for SUMO1-loaded Ubc9.

Altogether, these results show that efficient SUMOylation of hTDG and hXRCC1 can be achieved with the pSA- and pSC-based SUMO-E2-fusion system. In-cell modification experiments resulted in high SUMOylation efficiency for either of the hTDG-Ubc9 fusions, but also generated considerable amounts of mis-targeted modification, either in the target protein itself or in the SUMO-E2-fusion-tag. For this particular target, stronger induction of expression at higher temperature for shorter times markedly improved the SUMOylation specificity without affecting overall protein levels. Hence, induction conditions have a strong influence on in-cell SUMOylation efficiency and specificity and thus provide opportunities for target-specific fine-tuning of the system. Overall, the PreE2 fusions seem to SUMOylate more efficiently than the IntE2 fusions. Attempts to do in-cell SUMOylation of hXRCC1 indicated that the co-expression of larger Ubc9 fusions with all SUMO components in E.coli may not yield satisfactory results. In such cases, SUMO modification in mixed extracts provides a valuable alternative.

In-cell modified mTDG is biochemically active

To demonstrate that the SUMO-E2-fusion system produces authentically modified target protein, we performed in-cell modification with mTDG expressed from the pSC-IntE2 plasmid and purified the modified protein through consecutive enrichment over GST and Ni-NTA affinity columns. To test the suitability of the intein linker sequence for on-column release of the Ubc9-GST moiety, we eluted the SUMO-conjugated mTDG from the GST matrix by induced intein self-splicing in presence of 50 mM DTT. We then compared the eluted mTDG-SUMO1 with purified recombinant mTDG, either unmodified or SUMO1-modified in a defined in vitro SUMOylation system by immunoblot analysis using anti-mTDG (Figure 5A) and anti-SUMO1 antibodies (Figure 5B). Both antibodies detected in-cell modified mTDG as a prominent protein band migrating at around 80 kDa (Figure 5A and B, lane 3 and 5). A few higher molecular weight mTDG-SUMO1 species were also apparent, as expected for the very efficiently modified mTDG, while unmodified mTDG was hardly detectable (Figure 5A, lane 3). Upon treatment with the recombinant SUMO protease SenP2, SUMO1 was cleaved from the in-cell as well as from the in vitro SUMOylated mTDG (Figure 5A and B, lane 4 and 6) to generate the unmodified isoform, indicating that the
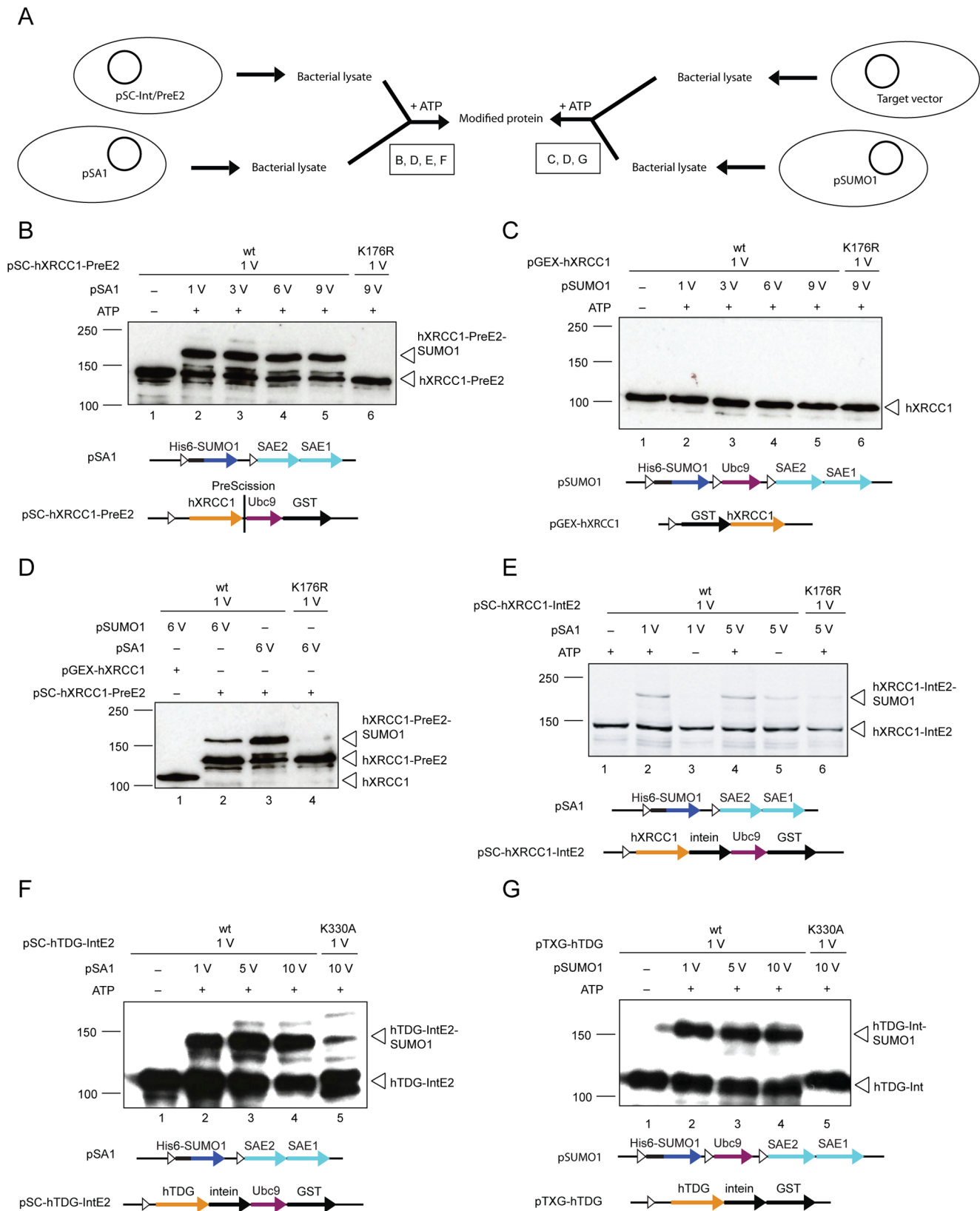
Figure 4. The SUMO-E2-fusion system allows SUMOylation of targets in crude cell extracts. (**A**) In-extract SUMOylation procedure using lysate from pSUMO1- or pSA1-expressing bacteria. Boxed letters indicate the corresponding sub-panels. In-extract SUMOylation efficiency with or without the addition ATP to a final concentration of 5 mM (30°C for 1 h) was assessed by immunoblot analysis. Extracts from *E.coli* BL21(DE3) cells expressing the SA1 (**B**) or the SUMO1 (**C**) system (250 μM IPTG at 30°C for 3 h) were mixed with extracts from cells expressing the fusion of Ubc9 to wild-type hXRCC1 (wt) or hXRCC1-K176R from the pSC-PreE2 plasmid (250 μM IPTG at 25°C for 3 h) with the indicated volume (V) ratio. (**D**) Direct

comparison of the SUMOylation efficiency of hXRCC1-PreE2 and hXRCC1 not fused to Ubc9 with either the SA1 or SUMO1 extracts. (**E**) Crude *E.coli* BL21(DE3) cell extracts expressing wild-type (wt) or SUMO acceptor site-mutated hXRCC1 (K176R) (250 µM IPTG at 25°C for 3 h) from the pSC-IntE2 plasmids were mixed with extracts with the SA system (250 µM IPTG at 30°C for 3 h) at the indicated volume (V) ratio. Applying the same experimental conditions as above, the SUMOylation of wild-type (wt) and the SUMOylation-deficient (K330A) hTDG mutant was analyzed comparing co-incubation of extracts from *E.coli* BL21(DE3) cells expressing pSA1 and the TARGET-IntE2-fusion (**F**) or pSUMO1 and the non-Ubc9 fusion (**G**).
doi:10.1371/journal.pone.0102157.g004

detected high molecular bands are indeed SUMO1-modified mTDG protein.

To confirm that the enriched SUMO1-modified mTDG shows previously described modification-induced enzymatic features [31], we analyzed purified recombinant mTDG (Figure 5A and
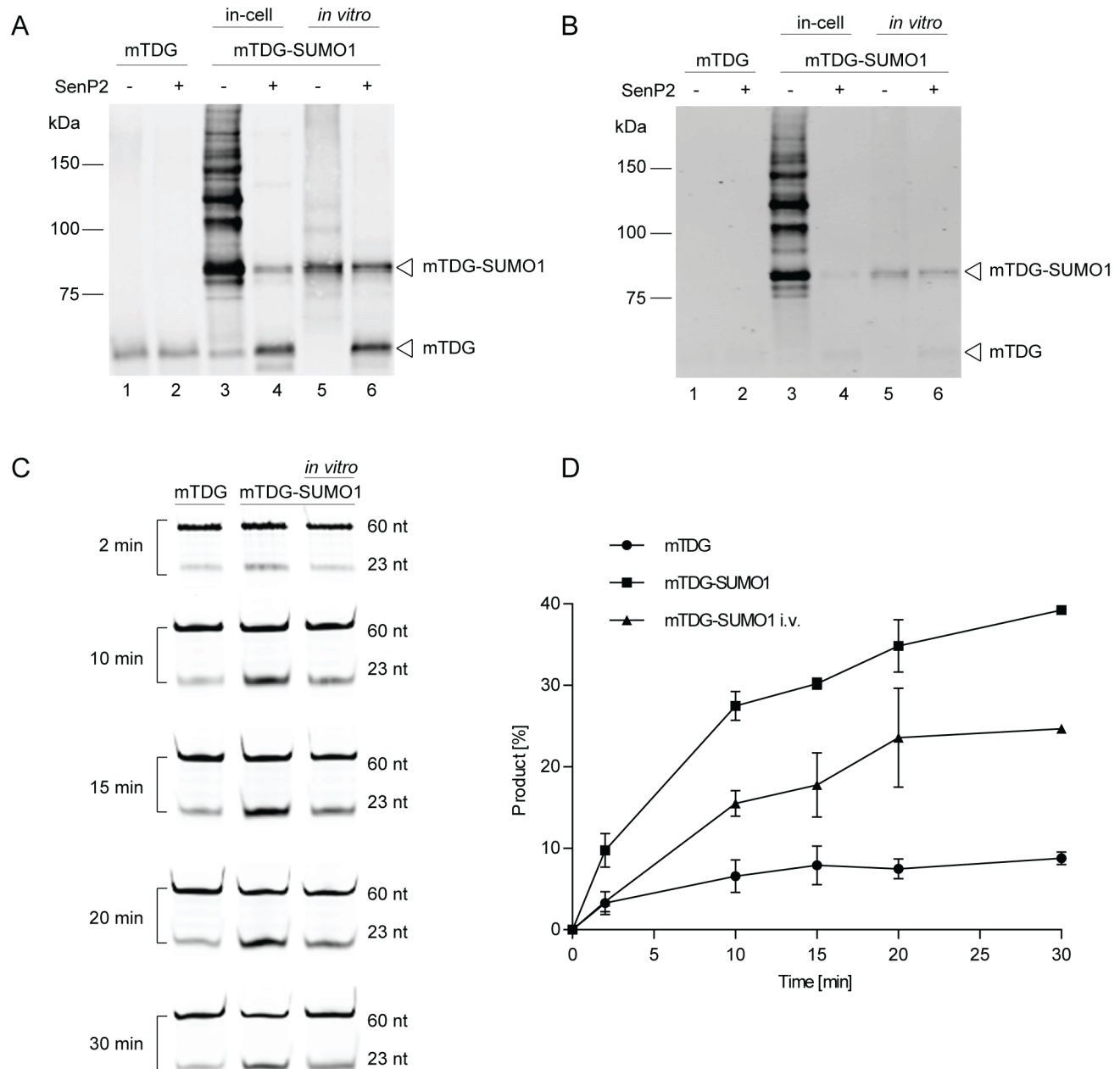

Figure 5. In-cell SUMOylated TDG is active and shows enzymatic turnover. 200 ng of purified mouse TDG (mTDG), purified in-cell SUMOylated mTDG and *in vitro* SUMOylated samples were analyzed by immunoblot analysis using anti-mTDG (**A**) as well as anti-SUMO1 (**B**) antibodies. Conjugated SUMO1 was cleaved with the recombinant SUMO protease SenP2 at RT for 30 min (lanes 2, 4, 6) and compared to untreated samples (lanes 1, 3, 5). (**C**) Enzymatic activity and turnover of unmodified TDG (mTDG), *in vitro* (mTDG-SUMO1 i.v.) and purified in-cell SUMOylated TDG (mTDG-SUMO1) were assessed by the base release assay with a 10-fold molar excess of G·U mismatched oligonucleotides over enzyme. Samples were taken at the indicated time-points and the relative amounts of processed 23 nucleotide product (23 nt) versus unprocessed 60 nucleotide substrate (60 nt) was quantified and depicted in (**D**). Error bars, SEM of 2 experiments.
doi:10.1371/journal.pone.0102157.g005

B, lane 1), purified in-cell SUMOylated mTDG (Figure 5A and B, lane 3) and *in vitro* SUMOylated mTDG (Figure 5A and B, lane 5) for their capacity to release uracil from G·U mismatched DNA substrates in a base release assay (Figure 5C and D). A 10-fold excess of DNA substrate over enzyme was provided in all reactions to allow an assessment of enzymatic turnover induced by mTDG SUMOylation. For unmodified mTDG, substrate processing plateaued at a product-enzyme ratio of about 2, in agreement with previous reports, showing that unmodified TDG has a strong affinity to the product AP-site and, therefore, shows the kinetics of product inhibition [24,32]. By contrast, in-cell SUMOylated mTDG processed up to an 8-fold excess of substrate without reaching a plateau after 30 min, even more efficiently than *in vitro* SUMOylated mTDG. These findings are in line with the previously established effect of mTDG SUMOylation on the turnover rate of mTDG in a base release assay with of G·U mismatched DNA substrate. We therefore conclude that the purification of SUMO1-modified mTDG generated from the enhanced SUMO-E2-fusion system, using vectors pSA1 and pSC-mTDG-IntE2, yields functionally intact protein with the expected biochemical properties.

Discussion

Over the past decade, posttranslational protein modification by SUMO polypeptides has emerged as a key regulatory mechanism of important cellular processes [2,5,8,33,34]. The number of known and suspected SUMO targets is increasing rapidly, many of which being identified by large-scale proteomic or by bioinformatic approaches. However, insight into mechanisms underlying SUMO-regulated biological transactions is lagging behind due to the difficulty to produce recombinant SUMO-modified proteins of sufficient homogeneity and quality for biochemical analyses. To this end, we developed humanized SUMOylation systems and strategies for simple purification of modified proteins from *E.coli* (Figure 6). The pSUMO-based system expresses all essential components for SUMO modification from a single plasmid (pSUMO1-3), which is compatible with the commonly used bacterial expression vectors carrying a target protein of interest with a suitable affinity tag for purification. The SUMO-E2-fusion system is composed of a SUMO-activating vector (pSA1-3) and a SUMO-conjugating vector (pSC), containing a TARGET-Ubc9-GST fusion expression cassette. This system is suitable for both, in-cell and in-extract SUMOylation, depending on the requirements of a target protein and, the latter offering useful combinatorial options.

Unlike previously introduced SUMOylation systems [17–19,35], the ones presented here consist of human components only and combine two different affinity-tags on the SUMO polypeptide and the target protein to facilitate separation of modified and unmodified target protein. The expression of the SUMO-activating enzyme E1 (SAE1 and SAE2) as separate subunits from a cistronic expression cassette avoids the SAE1-SAE2 fusion, which was shown to reduce E1 activity [20]. Consistently, SUMOylated mTDG and hXRCC1 protein appeared faster and at higher levels with our *E.coli* SUMOylation setup when compared to the system by Saitoh and colleagues [20].

SUMO-E3 ligases strongly enhance the SUMOylation efficiency and specificity in the *in vivo* situation by promoting proximity between the substrate and the SUMO-loaded Ubc9. Accordingly, more efficient and specific SUMOylation of target proteins were reported for an *E.coli* SUMOylation system when the respective E3 ligases were co-expressed [35]. However, SUMO-E3 ligases for many proteins are not known and the additional expression of

such a component is likely to reduce the overall production of recombinant proteins, and is thus not suitable as a universal strategy. Fusion of Ubc9 with SUMO1 or target proteins provides an alternative [36] and has been successfully applied in mammalian cells to compensate rate limiting SUMO-E3 activity [21]. The engineered proximity makes the SUMOylation independent of the affinity between target protein and activated SUMO-E2, which is a determinant for modification efficiency [25,37]. This TARGET-Ubc9 fusion approach is part of our *E.coli* SUMO-E2-fusion system for in-cell and in-extract modification. Using this system, we observed a fast and efficient SUMOylation of recalcitrant targets such as hXRCC1, which was only inefficiently modified without the Ubc9 fusion. Notably, canonical SUMOylation of Ubc9 itself was suggested to have a regulatory function in target discrimination, in particular for target proteins with a high affinity to SUMO [38,39], and to stimulate the formation of SUMO chains [40]. This might explain the multiple SUMOylation events observed with the SUMO-E2-fusion system, particularly pronounced upon prolonged time of induction. The Ubc9-fusion might not only facilitate the modification of targets but could also enhance SUMOylation of the fusion-tag itself, especially when the SUMOylation site in the target is not available, i.e. mutated. However, the specificity and efficiency of target protein SUMOylation appears to be little affected by these unscheduled SUMOylation events and they will be eliminated in the course of protein purification by cleaving off the target from the fusion-tag. As authentic regulatory mechanisms are lacking and, hence, target site selection may be biased to some extent in recombinant *E.coli* SUMOylation systems, caution should be applied if the SUMO-E2-fusion is used for the validation of SUMO acceptor sites (Figure 6). In any case, the conditions for in-cell and in-extract SUMOylation with our vector systems need to be carefully evaluated and controlled so that off-target SUMOylation or non-canonical SUMO-chain formation not observed in the authentic host system can be avoided.

Based on our experience with the SUMO targets TDG and XRCC1, we can provide some basic guidelines for how to purify SUMOylated targets, although optimal conditions can vary and, thus, have to be evaluated individually. We recommend starting with the pSUMO-based system, i.e. the co-expression of a gene of interest fused to a suitable affinity tag with the SUMOylation factors provided by plasmids pSUMO1-3. The efficiency as well as the specificity of target SUMOylation is difficult to predict and likely to depend on the abundance of soluble recombinant proteins and on the affinity of the target protein with the SUMO components. For instance, SUMO modification of minor acceptor sites or the formation of poly-SUMO chains as observed with heterologous SUMOylation systems [17–19] may be favored when the cellular concentration of activated Ubc9 is high. It is therefore crucial to initially determine optimal induction conditions that ensure an optimal balance between expression and specificity, i.e. produce predominantly mono-SUMOylated protein, carrying the SUMO at the authentic acceptor site. In the case of TDG, which has a well-defined SUMOylation site, either a strong and fast (1 mM IPTG) induction at a high temperature (37°C) or a mild induction (250 µM IPTG) at a lower temperature (15°C) for 3 to 4 hours gave the best results with respect to SUMOylation efficiency and specificity. Applying prolonged induction times yielded more protein but also produced unwanted multi-SUMOylation at unspecific sites.

If the pSUMO-based system does not yield satisfactory SUMOylation, the SUMO-E2-fusion system, expressing the target protein from either the pSC-IntE2 or pSC-PreE2 vector, provides an alternative strategy (Figure 6). Although the intein self-cleavage
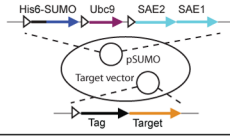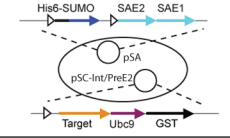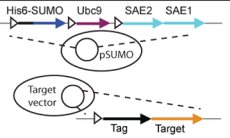
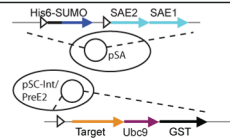|			Pros	Cons	Application	
In-vivo	Host	[cell diagram]	- Specific target modification - Regulation mechanisms present	- Low yields	- Mapping of SUMOylation sites - Biological function studies	
In-cell	Recombinant	His6-SUMO Ubc9 SAE2 SAE1 / pSUMO / Target vector / Tag Target	- Any compatible target vector - Specific target modification	- No SUMO-E3 ligase function - Target needs activated SUMO-E2 affinity	- Confirmation of SUMO site mapping	- Generation of large amounts of modified protein for structural and biochemical evaluation
In-cell	Recombinant	His6-SUMO SAE2 SAE1 / pSA / pSC-Int/PreE2 / Target Ubc9 GST	- Enhanced efficiency of target modification	- Reduced expression levels - Reduced specificity (?)		
In-extract	Recombinant	His6-SUMO Ubc9 SAE2 SAE1 / pSUMO / Target vector / Tag Target	- Combinatorial flexibility - Optimize induction conditions for target and SUMO system individually - Specific target modification	- No SUMO-E3 ligase function - Target needs activated SUMO-E2 affinity - Mixing ratios (stochiometry) have to be evaluated	- Confirmation of SUMO site mapping	
In-extract	Recombinant	His6-SUMO SAE2 SAE1 / pSA / pSC-Int/PreE2 / Target Ubc9 GST	- Enhanced efficiency of target modification - Specific target modification	- Size of fusion protein - Mixing ratios (stochiometry) have to be evaluated		
In-vitro	Recombinant	SAE2 SAE1 / Ubc9 / SUMO / Substrate	- Stochiometry can be adapted - Specificity	- Laborious to purify all components - Lower efficiency (?)	- Mapping of SUMOylation sites - Mechanistic evaluations	

Figure 6. Summarizing table. Overview on our newly introduced SUMOylation systems indicating advantages, disadvantages and putative applications in comparison to host *in vivo* and purely *in vitro* systems.
doi:10.1371/journal.pone.0102157.g006

mechanism used in the pSC-IntE2 vector has the advantage that no additional protease is required to remove the Ubc9 and the affinity-tag, the rather large fusion product may negatively affect the protein yields and the complex intein structure might also have an adverse effect on SUMOylation specificity. If so, the pSC-PreE2 vector with a conventional protease cleavage site offers a valuable alternative. In our experiments, we usually observed higher expression levels, increased modification efficiencies and better site specificity with targets expressed from pSC-PreE2. We recommend applying the SUMO-E2-fusion system for in-cell SUMOylation of small and easy to express proteins that SUMOylate poorly with the pSUMO-based system because of limited affinity to the activated Ubc9 or the requirement of a SUMO-E3 ligase.

For SUMO targets that are difficult to be produced and modified by co-expression, we evaluated an alternative procedure to obtain large quantities of SUMOylated proteins by in-extract conjugation, which does not require the prior purification of the SUMOylation factors. We independently expressed the TAR-GET-Ubc9 fusion protein and the SA system in *E.coli* BL(21) cells and prepared cleared lysate under *in vitro* SUMOylation-proficient buffer conditions. Co-incubation of these lysates resulted in a satisfactory SUMOylation of hXRCC1. Also, our data show that the Ubc9 fusion enhances SUMOylation of hXRCC1 in mixed lysates, suggesting a successful mimicking of a SUMO-E3 ligase function missing in this context. A significant advantage of this in-extract procedure is that expression conditions of the SA components and the TARGET-Ubc9 fusion constructs can be fine-tuned individually. In the case of hXRCC1, for instance, the

SA system expressed most efficiently following induction with 250 μM IPTG, 30°C for 3 hours, whereas the optimal conditions for hXRCC-Ubc9 were induction with 250 μM IPTG, 25°C for 3 hours. Hence, the possibility to fine-tune these conditions will impact on the yields and quality of SUMO-modified protein. In-extract modification may also produce good yields with the pSUMO-based SUMOylation system, although in this case a prerequisite is that the target protein has a high intrinsic affinity to SUMO1-loaded Ubc9, such as TDG. This strategy offers even higher combinatorial flexibility as existing expression vectors for potential SUMO targets can be used without having to consider plasmid replication incompatibility or selection markers (Figure 6).

Purification of GST- and His6-tagged protein can be carried out by applying the cell extracts onto GST and Ni-NTA affinity columns irrespective of whether SUMOylation was performed in-cell or in-extract. The sequence of affinity column purification has to be evaluated individually but should only have a minor effect on the final purity. First enriching the targets on the GST affinity column gives the advantage that protein can directly be eluted from the column via release from the GST-fusion-tag. Doing so, possible modifications of the fusion-tag are eliminated and the separation of modified and unmodified target by the subsequent Ni-NTA column is not affected by a possible GST dimerization. Yet, our experience from purifying TDG and XRCC1 was that yields were better when extracts were first fractionated on the Ni-NTA column and then on the GST column. In addition, this work flow resulted in less co-purification of free SUMO1 polypeptides; an issue that will apply to all targets harboring an intrinsic SUMO interaction domain. Also, non-covalent interaction with free

SUMO but also SUMO-conjugated target protein results in the co-purification of unmodified target and may require further purification steps such as size exclusion or ion-exchange chromatography or adaptation of buffer conditions (salt concentration, detergents).

In conclusion, we present tools and strategies to generate SUMOylated proteins using versatile binary expression vector systems in protease-deficient *E.coli*. They are designed to be applied for SUMOylation-related experimentation, complementary to classical investigation in the native host or *in vitro* (see Summary Figure 6). We provide purification work flows to enrich for SUMOylated protein that retains the expected biochemical properties. Owing to its high SUMOylation efficiency, the system will be suitable for screening and testing of predicted SUMOylation targets, but also for large scale purifications of modified proteins as required for biochemical and structural studies. Depending on the target, some degree of fine-tuning of expression

and modification conditions will be needed to limit non-specific SUMO-conjugation. As the vector systems are available for SUMO1, SUMO2 and SUMO3, modified proteins with the variant forms can be produced to analyze and compare SUMO-specific functional properties and consequences.

Acknowledgments

Author Contributions

Conceived and designed the experiments: DS ARW. Performed the experiments: ARW DS. Analyzed the data: ARW DS PS. Contributed to the writing of the manuscript: ARW DS PS.

References

1. Zhao J (2007) Sumoylation regulates diverse biological processes. Cell Mol Life Sci 64: 3017–3033.
2. Flotho A, Melchior F (2013) Sumoylation: a regulatory protein modification in health and disease. Annual review of biochemistry 82: 357–385.
3. Dohmen RJ (2004) SUMO protein modification. Biochim Biophys Acta 1695: 113–131.
4. Geiss-Friedlander R, Melchior F (2007) Concepts in sumoylation: a decade on. Nat Rev Mol Cell Biol 8: 947–956.
5. Jackson SP, Durocher D (2013) Regulation of DNA damage responses by ubiquitin and SUMO. Molecular cell 49: 795–807.
6. Gareau JR, Lima CD (2010) The SUMO pathway: emerging mechanisms that shape specificity, conjugation and recognition. Nature reviews Molecular cell biology 11: 861–871.
7. Schwartz DC, Hochstrasser M (2003) A superfamily of protein tags: ubiquitin, SUMO and related modifiers. Trends Biochem Sci 28: 321–328.
8. Hay RT (2013) Decoding the SUMO signal. Biochemical Society transactions 41: 463–473.
9. Johnson ES (2004) Protein modification by SUMO. Annu Rev Biochem 73: 355–382.
10. Desterro JM, Thomson J, Hay RT (1997) Ubch9 conjugates SUMO but not ubiquitin. FEBS Lett 417: 297–300.
11. Bernier-Villamor V, Sampson DA, Matunis MJ, Lima CD (2002) Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. Cell 108: 345–356.
12. Sampson DA, Wang M, Matunis MJ (2001) The small ubiquitin-like modifier-1 (SUMO-1) consensus sequence mediates Ubc9 binding and is essential for SUMO-1 modification. J Biol Chem 276: 21664–21669.
13. Cubenas-Potts C, Matunis MJ (2013) SUMO: a multifaceted modifier of chromatin structure and function. Developmental cell 24: 1–12.
14. Mikolajczyk J, Drag M, Bekes M, Cao JT, Ronai Z, et al. (2007) Small ubiquitin-related modifier (SUMO)-specific proteases: profiling the specificities and activities of human SENPs. The Journal of biological chemistry 282: 26217–26224.
15. Flotho A, Werner A, Winter T, Frank AS, Ehret H, et al. (2012) Recombinant reconstitution of sumoylation reactions in vitro. Methods in molecular biology 832: 93–110.
16. Werner A, Moutty MC, Moller U, Melchior F (2009) Performing in vitro sumoylation reactions using recombinant enzymes. Methods in molecular biology 497: 187–199.
17. Lens Z, Dewitte F, Van Lint C, de Launoit Y, Villeret V, et al. (2011) Purification of SUMO-1 modified IkappaBalpha and complex formation with NF-kappaB. Protein expression and purification 80: 211–216.
18. Uchimura Y, Nakamura M, Sugasawa K, Nakao M, Saitoh H (2004) Overproduction of eukaryotic SUMO-1- and SUMO-2-conjugated proteins in Escherichia coli. Anal Biochem 331: 204–206.
19. Mencia M, de Lorenzo V (2004) Functional transplantation of the sumoylation machinery into Escherichia coli. Protein expression and purification 37: 409–418.
20. Uchimura Y, Nakao M, Saitoh H (2004) Generation of SUMO-1 modified proteins in E. coli: towards understanding the biochemistry/structural biology of the SUMO-1 pathway. FEBS Lett 564: 85–90.
21. Jakobs A, Koehnke J, Himstedt F, Funk M, Korn B, et al. (2007) Ubc9 fusion-directed SUMOylation (UFDS): a method to analyze function of protein SUMOylation. Nat Methods 4: 245–250.
22. Niedenthal R (2009) Enhanced detection of in vivo SUMO conjugation by Ubc9 fusion-dependent sumoylation (UFDS). Methods in molecular biology 497: 63–79.
23. Jacobs AL, Schar P (2012) DNA glycosylases: in DNA repair and beyond. Chromosoma 121: 1–20.
24. Hardeland U, Steinacher R, Jiricny J, Schär P (2002) Modification of the human thymine-DNA glycosylase by ubiquitin-like proteins facilitates enzymatic turnover. The EMBO journal 21: 1456–1464.
25. Mohan RD, Rao A, Gagliardi J, Tini M (2007) SUMO-1-dependent allosteric regulation of thymine DNA glycosylase alters subnuclear localization and CBP/p300 recruitment. Molecular and cellular biology 27: 229–243.
26. Caldecott KW (2003) XRCC1 and DNA strand break repair. DNA Repair (Amst) 2: 955–969.
27. Gocke CB, Yu H, Kang J (2005) Systematic identification and analysis of mammalian small ubiquitin-like modifier substrates. J Biol Chem 280: 5004–5012.
28. Bruderer R, Tatham MH, Plechanovova A, Matic I, Garg AK, et al. (2011) Purification and identification of endogenous polySUMO conjugates. EMBO reports 12: 142–148.
29. Cuneo MJ, Gabel SA, Krahn JM, Ricker MA, London RE (2011) The structural basis for partitioning of the XRCC1/DNA ligase III-alpha BRCT-mediated dimer complexes. Nucleic acids research 39: 7816–7827.
30. Chong S, Mersha FB, Comb DG, Scott ME, Landry D, et al. (1997) Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. Gene 192: 271–281.
31. Hardeland U, Steinacher R, Jiricny J, Schar P (2002) Modification of the human thymine-DNA glycosylase by ubiquitin-like proteins facilitates enzymatic turnover. EMBO J 21: 1456–1464.
32. Fitzgerald ME, Drohat AC (2008) Coordinating the initial steps of base excision repair. Apurinic/apyrimidinic endonuclease 1 actively stimulates thymine DNA glycosylase by disrupting the product complex. The Journal of biological chemistry 283: 32680–32690.
33. Bologna S, Ferrari S (2013) It takes two to tango: Ubiquitin and SUMO in the DNA damage response. Frontiers in genetics 4: 106.
34. Ulrich HD (2012) Ubiquitin and SUMO in DNA repair at a glance. Journal of cell science 125: 249–254.
35. O'Brien SP, DeLisa MP (2012) Functional reconstitution of a tunable E3-dependent sumoylation pathway in Escherichia coli. PLoS One 7: e38671.
36. Kim ET, Kim KK, Matunis MJ, Ahn JH (2009) Enhanced SUMOylation of proteins containing a SUMO-interacting motif by SUMO-Ubc9 fusion. Biochem Biophys Res Commun 388: 41–45.
37. Takahashi H, Hatakeyama S, Saitoh H, Nakayama KI (2005) Noncovalent SUMO-1 binding activity of thymine DNA glycosylase (TDG) is required for its SUMO-1 modification and colocalization with the promyelocytic leukemia protein. The Journal of biological chemistry 280: 5611–5621.
38. Knipscheer P, Flotho A, Klug H, Olsen JV, van Dijk WJ, et al. (2008) Ubc9 sumoylation regulates SUMO target discrimination. Mol Cell 31: 371–382.
39. Duda DM, van Waardenburg RC, Borg LA, McGarity S, Nourse A, et al. (2007) Structure of a SUMO-binding-motif mimic bound to Smt3p-Ubc9p: conservation of a non-covalent ubiquitin-like protein-E2 complex as a platform for selective interactions within a SUMO pathway. Journal of molecular biology 369: 619–630.
40. Klug H, Xaver M, Chaugule VK, Koidl S, Mittler G, et al. (2013) Ubc9 Sumoylation Controls SUMO Chain Formation and Meiotic Synapsis in Saccharomyces cerevisiae. Molecular cell 50: 625–636.

Published online 6 April 2015

Gadd45a promotes DNA demethylation through TDG

Zheng Li[1], Tian-Peng Gu[1], Alain R. Weber[2], Jia-Zhen Shen[1], Bin-Zhong Li[1], Zhi-Guo Xie[1], Ruichuan Yin[3], Fan Guo[4], Xiaomeng Liu[4], Fuchou Tang[4,5], Hailin Wang[3], Primo Schär[2] and Guo-Liang Xu[1,6,*]

[1]Group of DNA Metabolism, The State Key Laboratory of Molecular Biology, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China, [2]Department of Biomedicine, University of Basel, Basel 4048, Switzerland, [3]The State Key Laboratory of Environmental Chemistry and Ecotoxicology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China, [4]Biodynamic Optical Imaging Center, College of Life Sciences, Peking University, Beijing 100871, China, [5]Ministry of Education Key Laboratory of Cell Proliferation and Differentiation, Peking University, Beijing 100871, China and [6]School of Life Science and Technology, ShanghaiTech University, 319 Yue Yang Road, Shanghai 200031, China

ABSTRACT

Growth arrest and DNA-damage-inducible protein 45 (Gadd45) family members have been implicated in DNA demethylation in vertebrates. However, it remained unclear how they contribute to the demethylation process. Here, we demonstrate that Gadd45a promotes active DNA demethylation through thymine DNA glycosylase (TDG) which has recently been shown to excise 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) generated in Ten-eleven-translocation (Tet)—initiated oxidative demethylation. The connection of Gadd45a with oxidative demethylation is evidenced by the enhanced activation of a methylated reporter gene in HEK293T cells expressing Gadd45a in combination with catalytically active TDG and Tet. Gadd45a interacts with TDG physically and increases the removal of 5fC and 5caC from genomic and transfected plasmid DNA by TDG. Knockout of both Gadd45a and Gadd45b from mouse ES cells leads to hypermethylation of specific genomic loci most of which are also targets of TDG and show 5fC enrichment in TDG-deficient cells. These observations indicate that the demethylation effect of Gadd45a is mediated by TDG activity. This finding thus unites Gadd45a with the recently defined Tet-initiated demethylation pathway.

INTRODUCTION

Methylation at position 5 of cytosine (5-methylcytosine, 5mC) in DNA is a major epigenetic modification that regulates gene transcription and other functions of the genome (1,2). Cytosine methylation in CpG-rich regulatory gene promoters and enhancers inversely correlates with transcriptional activity of associated genes as it causes chromatin condensation and thus gene silencing. Since patterns of 5mC are subject to mitotic inheritance through maintenance methylation during DNA replication, active demethylation is required for a rapid and efficient erasure of 5mC (3). Both locus-specific and genome-wide demethylation have been documented (4). For instance, the promoter of the estrogen receptor target gene *pS2* undergoes active demethylation in the cyclic activation of transcription (5). Genome-wide demethylation in primordial germ cells is believed to be important for erasing the parental methylation patterns (6). Demethylation of the zygotic genome is associated with remodeling of the parental epigenomes, presumably to establish developmental competence for the early embryo (7–10). Multiple mechanisms have been proposed to achieve active demethylation, which include direct removal of the exocyclic methyl group from the cytosine via C–C bond cleavage, replacement of the methylated cytosine base and nucleotide respectively through DNA base excision repair (BER) and nucleotide excision repair pathways (11). However, most of the proposed mechanisms have not been validated biochemically and genetically (12,13).

Compelling biochemical and genetic evidence has suggested that members of the Ten-eleven-translocation (Tet) family of DNA dioxygenases function to reverse DNA methylation (4,14). Tet enzymes catalyze the oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (15–17). Thymine DNA glycosylase (TDG), originally identified as a DNA glycosylase for the excision of thymine and uracil mispaired with guanine, is able to recognize and excise the Tet-generated oxidation products 5fC and 5caC, leading to the incorporation of unmethylated cytosine via the BER pathway (15,18–20). The functional relevance of TDG in

the regulation of DNA methylation is well-established by gene inactivation experiments at both animal (21,22) and ES cell levels (15,23,24). Given the importance of DNA methylation in stem cell biology and cancer, the study of Tet/TDG-mediated demethylation has become a major focus over the recent years. However, it is still unclear how the oxidative demethylation process is regulated.

Gadd45 (growth arrest and DNA-damage-inducible protein 45) family proteins are multi-faceted nuclear factors implicated in active DNA demethylation, apart from maintenance of genomic stability, DNA repair and suppression of cell growth (25,26). Overexpression of Gadd45a activates methylation-silenced reporter genes and promotes global DNA demethylation (27). However, despite the connection of Gadd45 proteins with DNA demethylation in several contexts, including neuronal activity-induced demethylation in the mouse brain (28) and deaminase-related demethylation in *Xenopus laevis* embryos (29), if and how they exactly promote DNA demethylation has remained unresolved and controversial (26,30).

In this study, we investigate the role of Gadd45a in active demethylation and activation of silenced genes. We find that Gadd45a interacts physically and functionally with TDG and contributes to DNA demethylation and gene activation in a TDG-dependent manner. In mouse ES cells, inactivation of *Gadd45a/b* leads to hypermethylation at loci most of which overlap with those depending on TDG for demethylation. These findings connect Gadd45 proteins with the Tet-TDG axis, functionally integrating the seemingly diverse demethylation pathways.

MATERIALS AND METHODS

Materials

Primary antibodies used for western blotting assays were as follows: anti-Flag (Sigma, F7425), anti-HA (Sigma, H6908), anti-GAPDH (Sigma, 9545). Anti-Tet2 and anti-TDG antibodies were as described previously (31). The *Tdg* knockout ES cell line was described (32).

Luciferase reporter assay

The reporter plasmid pCpGL-CMV-firefly luciferase was generated by subcloning the Cytomegalovirus (CMV) promoter from pcDNA3.1 (Invitrogen) into the CpG-free pCpGL-basic vector (33). By replacing the firefly luciferase gene with the renilla luciferase gene, an analogous control reporter plasmid pCpGL-CMV-renilla luciferase was constructed. The firefly plasmid was *in vitro* methylated with CpG methylase M.SssI (NEB) and the complete methylation was verified by digestion with methylation sensitive enzymes TaiI (Fermentas). HEK293T cells were transiently transfected in 12-well plate with 500 ng expression constructs (Tet2/TDG/Gadd45a) each, 20 ng methylated firefly luciferase reporter and 0.2 ng unmethylated renilla luciferase reporter as an internal control for normalization. Forty-six hours after transfection, luciferase activities were measured using the dual-luciferase reporter assay system (Promega) according to the manufacturer's instructions. Each experiment was repeated at least three times.

HPLC analysis of nucleosides

The nucleosides were analyzed by HPLC according to He *et al.* (15). Briefly, 200 µg of genomic DNA extracted from HEK293T cells were heat-denatured and hydrolyzed into mononucleotides with 0.5 U of nuclease P1 (Sigma) at 37°C overnight (the reaction buffer containing 20 mM NaAc, pH 5.3, 0.2 mM $ZnSO_4$). The nucleotides were then dephosphorized by incubation with calf intestinal alkaline phosphatase (CIAP, TaKaRa) for at least 2 h at 37°C. The reactions were then concentrated into 35 µl and analyzed on an Agilent 1260 HPLC machine with an AQ-C18 column of 5-µm particle size, 25 cm × 4.6 mm. The mobile phase was 10 mM KH_2PO_4, pH 3.7, running at 0.6 ml per min and the detector was set at 280 nm. 5hmC and 5caC nucleoside standards were prepared by dephosphorylation of 5-Hydroxymethyl-dCTP and 5-Carboxy-dCTP (TriLink), and 2′-deoxycytidine (C) and 5-methyl 2′-deoxycytidine (5mC) were bought from Sigma.

UHPLC- MS/MS analysis of mononucleosides

UHPLC-MS/MS analysis of modified mononucleosides was carried out according to Yin *et al.* (34). It was performed on an Agilent 1290 UHPLC system coupled with a G6410B triple quadrupole mass spectrometer (Agilent Technologies, Palo Alto, CA, USA). An isocratic elution with 5.0% methanol, 95% water and 0.1% formic acid running at 0.25 ml/min was used for UHPLC separation of mononucleosides. The eluate from the column was injected into electrospray ionization-triple quadrupole mass spectrometry. Positive multiple reaction monitoring modes were used: m/z 242→126 for 5mC (collision energy, 5eV); m/z 258→142 for 5hmC (5eV); m/z 256→140 for 5fC (5eV); m/z 272→156 for 5caC (5eV). 2′-deoxyguanosine (dG) in all samples was also measured to normalize the amount of DNA on column. Each sample was analyzed at least three times. The frequency of modified mononucleosides was calculated by corresponding standard curves.

Establishment of TDG KO HEK293T cell lines

Targeting of *TDG* in HEK293T cells was carried out by using TALEN mediated homologous recombination. TALENs against *TDG* were designed by online software TAL Effector Nucleotide Targeter 2.0 (35) to target the exon 2 of *TDG*. The TALENs were composed of domains targeting to the left arm sequence: 5′-TCAGCTATTCCCTTCAGCA-3′ and the right arm sequence: 5′-TCAGTTGTTGAAATGGAAA-3′. TALEN constructs were assembled using FastTALE TALEN Assembly Kit (SiDanSai). The TALENs plasmids and targeting vectors (PGK-hygromycin and PGK-puromycin) were transfected into HEK293T cells using Lipofectamine 2000. Selective medium with 200 µg/ml hygromycin and 1.5 µg/ml puromycin was applied 48 h after transfection and replaced every two days. Individual positive colonies were picked and expanded, which were further characterized by genotyping polymerase chain reaction (PCR) and western blotting using antibody against TDG.

Generation of *Gadd45a/Gadd45b* double-knockout mouse ES cell lines

Gene targeting of *Gadd45a/b* was accomplished by using CRISPR/Cas9 system according to Ran *et al.* (36). The target sequences were designed on the website (http://crispr.genome-engineering.org/) to disrupt the first exon containing the start codon of the *Gadd45a* or *Gadd45b* genes. Expression vector px330 containing the Cas9 and mCherry genes was digested with BbsI and the linearized plasmid purified using QIAquick PCR Purification Kit (Qiagen). Paired oligonucleotides for sgRNA at each targeted site (Supplementary Table S1) were annealed and ligated to the linearized vector. The plasmids were confirmed by sequencing.

V6.5 ES cells were cultured on feeder cells with standard ES cell culture conditions. Cells were transfected with px330-sgRNA plasmids using Lipofectamine 2000. After 48 h, mCherry-positive cells were sorted by flow cytometry using BD FACS Aria II (BD) and cultured on feeder layers. After recovering for 4–6 days, individual colonies were picked and genotyped by PCR genotyping and sequencing.

Reduced representation bisulfite sequencing (RRBS)

Reduced representation bisulfite sequencing (RRBS) was performed as described in Guo *et al.* (32) with slight modification. Briefly, genomic DNA was digested by MspI (Fermentas) and the 200–500 bp DNA fragments were selected to undergo bisulfite conversion. Analysis of RRBS data was performed according to Guo *et al.* (32), the adaptor-trimmed reads were mapped to the mouse genome (mm9) using Bismark (v. 0.76, http://www.bioinformatics.babraham.ac.uk/projects/bismark/). The methylation level of each single CpG site was calculated using the number of RRBS-measured C (methylated) divided by the sum of measured C (methylated) and T (unmethylated). CpG sites with at least five uniquely mapped reads were chosen for further analysis.

The differentially methylated CpG sites were identified if the absolute methylation level difference was >25% between wild-type (WT) and Gadd45 DKO ES cells (one-tailed Fisher's exact test with $P < 0.05$, with Benjamini–Hochberg false discovery rate <0.05). The neighboring hypermethylated (or hypomethylated) CpG sites were merged to form a differentially methylated region if there was at most one CpG site between these two hypermethylated (or hypomethylated) CpG sites and the distance between them was <100 bp.

Preparation of the 5fC/5caC-containing plasmid (oxi-5mC plasmid)

For the preparation of oxi-5mC plasmid DNA, 100 ng of M.SssI-methylated pCpGL-CMV-firefly plasmids were oxidized by using 2 μg of the TET2 protein (37) in the reaction buffer (50 mM HEPES, pH 8.0, 50 mM NaCl, 2 mM ascorbic acid, 1mM 2-oxoglutarate, 100 μM Fe(NH$_4$)$_2$(SO$_4$)$_2$, 1 mM adenosine triphosphate (ATP) and 1 mM Dithiothreitol (DTT)) for 1 h at 37°C. The reaction was then added with 1 μg of the TET2 protein for further incubation of 1 h. The plasmid DNA was recovered by phenol–chloroform
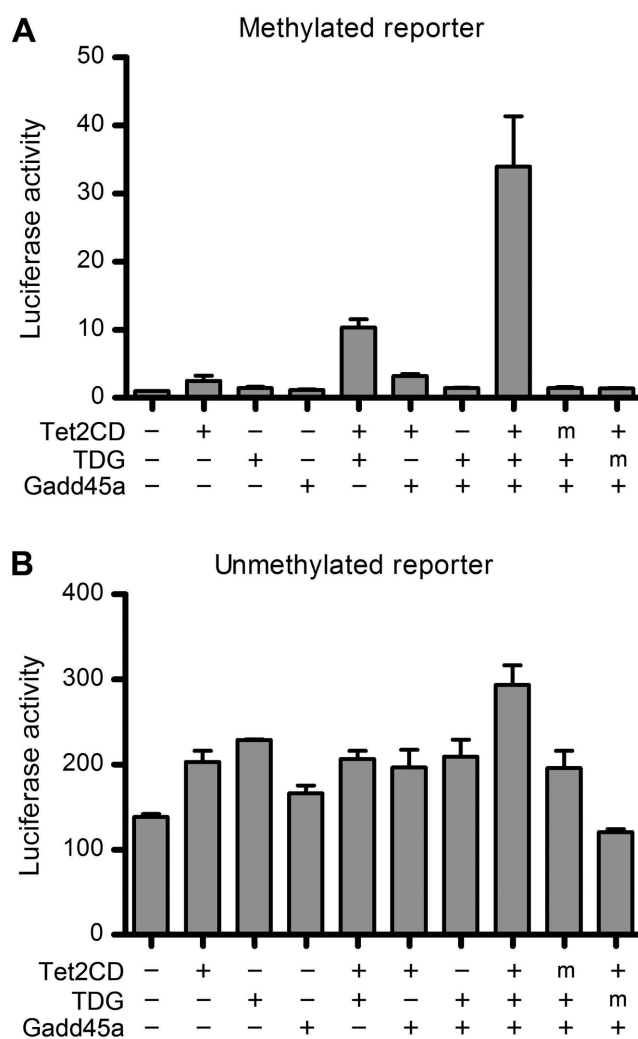

Figure 1. Gadd45a synergizes with Tet and TDG in activation of a methylated reporter gene in a dual luciferase reporter assay. A methylated (**A**) or unmethylated (**B**) pCpGL-CMV-firefly luciferase reporter was co-transfected into HEK293T cells with Gadd45a, Tet2CD and TDG in various combinations indicated. Shown is the firefly luciferase activity normalized to the level of control pCpGL-CMV-renilla luciferase relative to that of cells without Tet, TDG and Gadd45 expression, which is set to 1. 'm' denotes catalytically inactive mutants of Tet2CD and TDG. Data are represented as the mean ± SEM of three independent experiments.

extraction and the oxidative level of the DNA was tested by M.SssI-assisted bisulfite sequencing (MAB-seq) (32).

Isolation of transfected plasmid

Forty-eight hours after transfection, nuclei were extracted from the transfected HEK293T cells using the Wizard Genomic DNA Purification Kit (Promega) according to the manufacturer's instructions. The plasmid DNA was then isolated using phenol–chloroform extraction.

Bisulfite sequencing

A total of 200 ng of genomic DNA were treated with the EZ DNA Methylation-Direct Kit (Zymo Research). Specific genomic regions were PCR-amplified using Taq HS en-
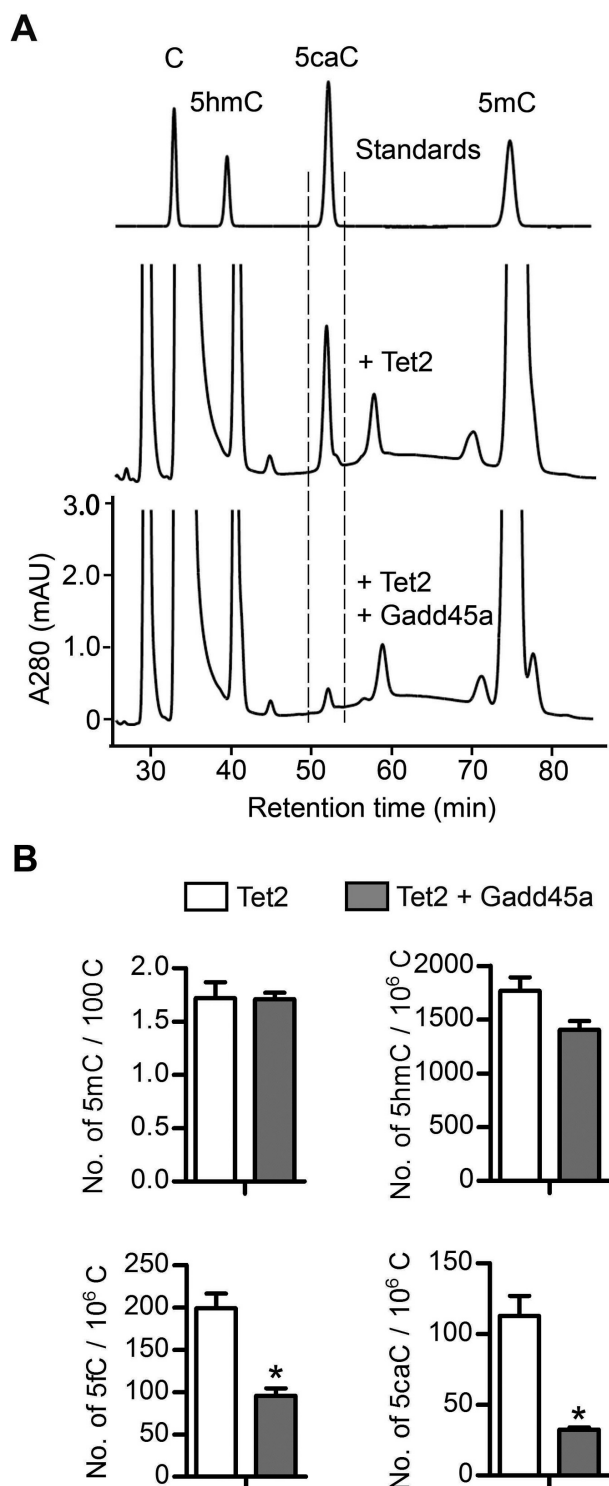
Figure 2. Ectopic expression of Gadd45a reduces genomic 5fC and 5caC generated by Tet2 in transfected HEK293T cells. (**A**) HPLC analysis of 5caC in the genomic DNA of cells overexpressing full-length Tet2 with and without Gadd45a. The y-axis indicates OD280nm of nucleoside preparations from genomic DNA. Defined nucleosides were used as standards (C, 5mC, 5hmC and 5caC). Dashed lines indicate the elution position of 5caC. (**B**) Mass spectrometric quantification of modified cytosines (5mC, 5hmC, 5fC and 5caC) in the genomic DNA of cells ectopically expressing Tet2 alone or with Gadd45a. Data are represented as the mean ± SEM from two independent experiments, and are analyzed by two-tailed *t*-test, $P < 0.05$ (*).

zyme (Takara). The PCR products were gel-purified by using Gel Extraction Kit (Qiagen) and cloned into pMD 19-T (Takara) for sequencing. Data were analyzed by an online called BISMA (38) (http://services.ibc.uni-stuttgart.de/BDPC/BISMA/) and duplicated clones were deleted by this software.

M.SssI-assisted bisulfite sequencing (MAB-seq)

MAB-seq was performed to map sequence-specific 5fC and 5caC distribution as described (31,32). A total of 100 ng of genomic DNA were methylated by M.SssI (NEB) following the vendor's instructions. The methylated DNA sample was then purified by phenol–chloroform extraction. Bisulfite conversion, sequencing and data analysis were performed as described above.

Recombinant protein expression and purification

For purification of Flag-tagged proteins, coding sequences were cloned into a modified pcDNA4 (Invitrogen) vector. The expression plasmids were transfected into HEK293T cells using polyethylenimine (PEI, Sigma). After culturing for 48 h, the cells were harvested and lysed in lysis buffer (50 mM Tris, pH 7.5, 500 mM NaCl, 1% NP-40, 1× protease inhibitors without ethylenediaminetetraacetic acid (EDTA)). The recombinant proteins were purified by using the FLAG M2 affinity gel (Sigma) according to the manufacturer's instructions. The eluted proteins were dialyzed against storage buffer (20 mM HEPES, pH 7.4, 50 mM NaCl and 50% glycerol) for 4 h at 4°C and then stored at −80°C.

For purification of Glutathione S-transferase (GST) fusion proteins, GST-tagged proteins were overexpressed in *Escherichia coli* strain BL21 (DE3) Codon-Plus-RIL (Stratagene). Cells were grown at 37°C in LB medium to an optical density of 0.6–0.8 at 600 nm and then induced with 0.2 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) at 20°C overnight. Purification was performed using Glutathione Sepharose 4B (GE Healthcare) according to the manufacturer's instructions.

Purification of TDG used in the *in vitro* glycosylase assay was described before (39).

GST-pull down assay

To confirm the interaction of TDG and Gadd45a, GST pull-down assay was performed essentially as described (40). A total of 4 μg of the GST–TDG fusion protein purified from *E. coli* were incubated with 20 μl Glutathione Sepharose 4B beads in the binding buffer [20 mM Tris–HCl (pH 7.9), 0.1 M NaCl, 1 mM EDTA, 5 mM MgCl₂, 0.1% NP-40, 1 mM DTT, 0.2 mM phenylmethanesulfonyl fluoride (PMSF)] in a total volume of 200 μl at 4°C for 2 h. Then 1 μg of Flag-Gadd45a was added to the slurry and incubated at 4°C for another 2 h. The Sepharose beads were washed three times with the washing buffer [20 mM Tris–HCl (pH 7.9), 0.15 M NaCl, 1 mM EDTA, 5 mM MgCl₂, 0.1% NP-40, 1 mM DTT, 0.2 mM PMSF] and bound proteins were analyzed by sodium dodecyl sulphate-polyacrylamide gel electrophoresis and visualized by western blotting.

The interaction of Flag-TDG and GST-Gadd45a was determined using the same protocol.

Yeast two-hybrid assay

To confirm the interaction between Gadd45a and TDG, Gadd45a was cloned in-frame with GAL4 DNA binding domain (GAL4DBD) in the bait vector pGBKT7 (Clontech) and TDG fragments were fused with the GAL4 activation domain (GAL4AD) in the prey vector pGADT7. Yeast was co-transformed with pGBKT7-Gadd45a and pGADT7-TDG. Colonies were selected on the SD medium lacking His, Leu, Trp and adenine according to the recommended protocol (Clontech).

Base release assay

The 60-mer double-stranded oligonucleotide substrates containing different modifications were prepared by annealing of an unlabeled upper strand oligonucleotide 5′-TAGACATTGCCCTCGAGGTACCATGGATCC GATGTCGACCTCAAACCTAGACGAATTCCG-3′ to a 5′-fluorescein-labeled lower oligonucleotide strand 5′-F-CGGAATTCGTCTAGGTTTGAGGT$\underline{X}$GACATC GGATCCATGGTACCTCGAGGGCAATG$\overline{\underline{T}}$CTA-3′, where $X =$ T or 5caC.

Multiple turnover base release assays were carried out in a total volume of 200 μl containing 25 nM of the labeled DNA substrate (G•T), 100 nM of unlabeled DNA substrate, 25 nM of recombinant TDG and 1 μM Gadd45a (or BSA) in 1× reaction buffer [50 mM Tris–HCl (pH 8.0), 1 mM EDTA, 1 mM DTT] at 37°C and 20 μl samples were taken at indicated time points. Generated AP-sites were cleaved by the addition of NaOH to a final concentration of 100 mM and heating to 99°C for 10 min. Subsequently, DNA was ethanol precipitated overnight at −20°C in 0.3 M Na-acetate (pH 5.2) and in the presence of 0.4 mg/ml carrier tRNA. The DNA was collected by centrifugation (20 min, 20 000 *g*, 4°C) and washed in 70% ethanol. Air-dried pellets were resuspended in loading buffer (1× Tris-Borate-EDTA (TBE) buffer, 90% formamide), heated at 99°C for 5 min and then immediately chilled on ice. Reaction products were separated on 15% denaturing polyacrylamide gels in 1× TBE. The fluorescein-labeled DNA was visualized with a Typhoon 9400 (GE Healthcare) and quantified using the ImageQuant TL software (GE Healthcare).

Single turnover base release assays were carried out as described above in a total volume of 200 μl containing 25 nM of the labeled DNA substrate (G•5caC), 250 nM recombinant TDG and 1 μM Gadd45a in 1× reaction buffer [50 mM Tris–HCl (pH 8.0), 1 mM EDTA, 1 mM DTT] at 30°C and 20 μl samples were taken at indicated time points.

RESULTS

Gadd45a activates the expression of a methylated reporter gene in cooperation with Tet and TDG

Active DNA demethylation can be achieved by a concerted action of Tet and TDG (41). We reasoned that other factors previously proposed to contribute to demethylation might exert their effects by modulating the Tet-TDG axis. To test this possibility, we used a cell-based firefly luciferase reporter assay to test anti-gene silencing effects of factors of interest (Supplementary Table S2). The luciferase-reporter plasmids were free of CpGs except for the 0.6-kb CMV promoter (33) regions, which contained 37 CpGs that we fully methylated *in vitro* by reaction with the CpG-specific bacterial methyltransferase M.SssI prior to transfection. Methylation of the CMV promoter conferred 100-fold repression of the firefly luciferase activity but co-transfection of the methylated reporter plasmid with Tet2 catalytic domain (Tet2CD) together with TDG, increased the activity by 13-fold, thus partially alleviating the repression (Figure 1A). Gadd45a, which we tested as a candidate for a modulatory factor, was able to stimulate the silenced reporter expression by 35-fold, but only when co-transfected with Tet2CD and TDG (Figure 1A). Expression of Gadd45a did not affect the protein levels of Tet2CD and TDG in transfected cells (Supplementary Figure S1). Moreover, its effect was dependent on the enzymatic activities of Tet and TDG because no stimulation effect was observed in co-transfection with the catalytically inactive mutants. By contrast, Gadd45a in combination with Tet and TDG had only a two-fold stimulatory effect on expression of the reporter gene on an unmethylated plasmid (Figure 1B). These observations suggest that Gadd45a cooperates with Tet and TDG in the activation of a methylation-silenced reporter gene.

Reduction of 5fC and 5caC by Gadd45a in HEK293T cells overexpressing Tet2

Since the effect of Gadd45a on methylated reporter activation depends on catalytically active Tet and TDG, we next investigated whether Gadd45a might impact on the functions of Tet and TDG in transfected cells. The occurrence of 5caC in the genomic DNA of HEK293T cells transfected with a Tet enzyme could be detected by HPLC analysis (15). We took advantage of this system to examine whether Gadd45a can regulate the level of 5caC generated in transfected cells. As reported previously (15), 5caC could be detected in Tet2-transfected cells (Figure 2A). However, the 5caC level was reduced to 30% by co-transfection of Gadd45a with Tet2, while the 5mC and 5hmC levels were unchanged. Gadd45a expression did not influence the protein expression of transfected Tet2 and endogenous TDG (Supplementary Figure S2A and B). Drastic reduction in 5fC and 5caC levels was confirmed by triple quadruple mass spectrometry quantification (Figure 2B). The selective removal of 5fC and 5caC but not 5mC and 5hmC could suggest that Gadd45a might not affect the oxidation function of Tet2 but promote the function of the endogenous TDG which is likely rate-limiting in removing genomic 5fC and 5caC from HEK293T cells.

Gadd45a interacts with TDG

Having demonstrated the functional relationship between Gadd45a and TDG in reporter gene reactivation and the regulatory effect of Gadd45a on the 5caC and 5fC levels in transfected cells, we wondered whether there is a direct protein–protein interaction between them. As shown in Figure 3A, purified Flag-TDG from HEK293T cells could
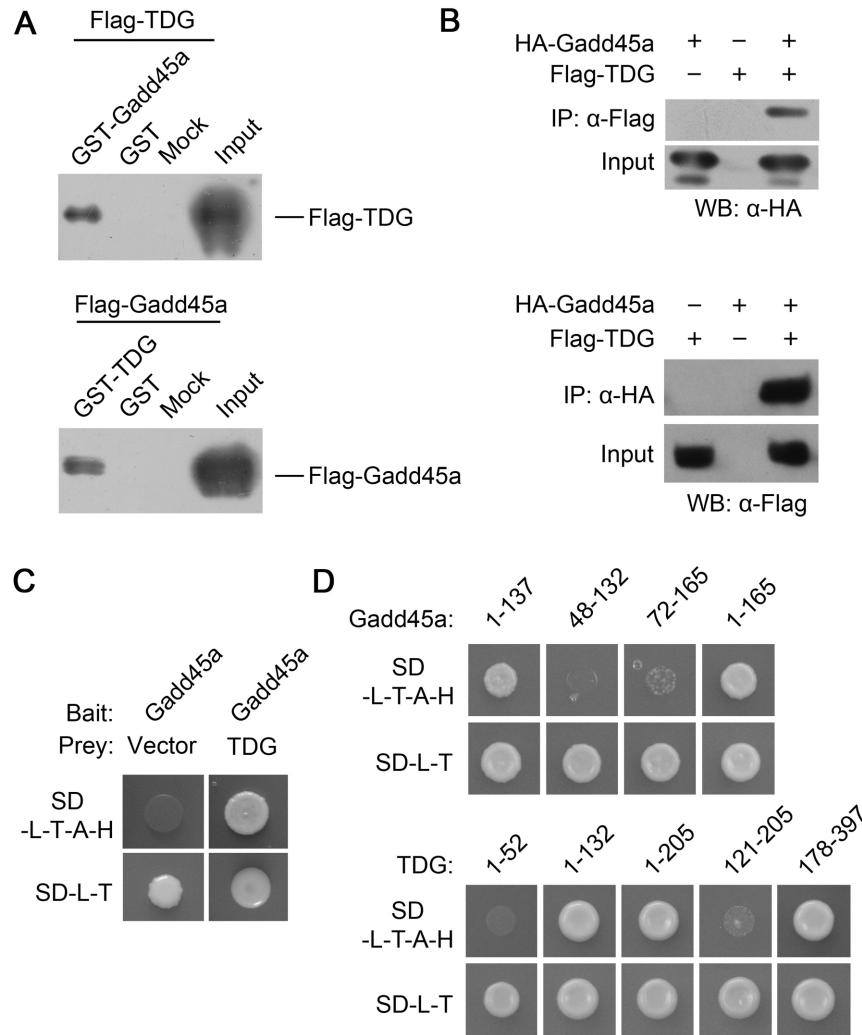
Figure 3. Gadd45a interacts with TDG *in vitro*. (**A**) GST and Flag pull-down assay. Western blot analysis of Flag-TDG in fractions obtained from a GST pull-down assay using GST-Gadd45a (lane 1) or GST (lane 2) (upper panel) and *viceversa* (lower panel). Mock is the control without GST or its fusion. Ten percent of the input was loaded. (**B**) Co-immunoprecipitation (Co-IP) assay. Immunoprecipitates of lysates of HEK293T cells transfected with HA-Gadd45a and Flag-TDG alone or in combination were resolved by SDS-PAGE and detected by western blotting with the indicated antibodies. (**C**) Yeast two-hybrid assay. Appearance of colonies in the SD-L-T-A-H selective medium indicates that the reporter genes have been activated due to the interaction between the bait (GAL4DBD-Gadd45a) and prey (TDG-GAL4AD) hybrid proteins. Yeast strains harboring the Gadd45a bait construct and an empty prey vector were used as negative controls. (**D**) Mapping of Gadd45a-TDG interaction domains. The protein regions tested in yeast two-hybrid assays are indicated above.

bind with bacterial recombinant protein GST-Gadd45a. The protein association could also be shown with co-immunoprecipitation (Co-IP) assay using extracts from HEK293T cells transfected with epitope-tagged TDG and Gadd45a (Figure 3B). To further confirm and characterize the interaction, we performed yeast two-hybrid assay. As expected, yeasts harboring both a Gadd45a bait and a TDG prey construct could grow on the selective medium lacking His, Leu, Trp and adenine because of the activation of reporter genes due to the interaction between Gadd45a and TDG (Figure 3C). To map the interacting domains, deletion analysis was performed with the bait and prey constructs. We found that the N-terminal domain of Gadd45a (amino acids1–137) mediates the interaction with TDG via a N (amino acids 1–132) and a C-terminal (amino acids 178–397) domain of TDG (Figure 3D). These data provide

a confirmation of the physical interaction between Gadd45a and TDG.

The effect of Gadd45a on 5caC removal is exclusively mediated by TDG

TDG is the only known enzyme capable of specifically excise 5fC and 5caC from DNA (15). To determine whether Gadd45a affects the reduction of 5caC in genomic DNA through the endogenous TDG, we generated HEK293T cell lines deficient in TDG. The *TDG* gene was disrupted by using TALEN technology (Supplementary Figure S3). Exon 2 was replaced in the two alleles, by the hygromycin-resistance gene in one gene copy and by the puromycin-resistance gene in the other. Deletion of this exon would lead to a frameshift behind the first seven N-terminal codons. Genomic PCR confirmed that the exon 2 sequences had been deleted,
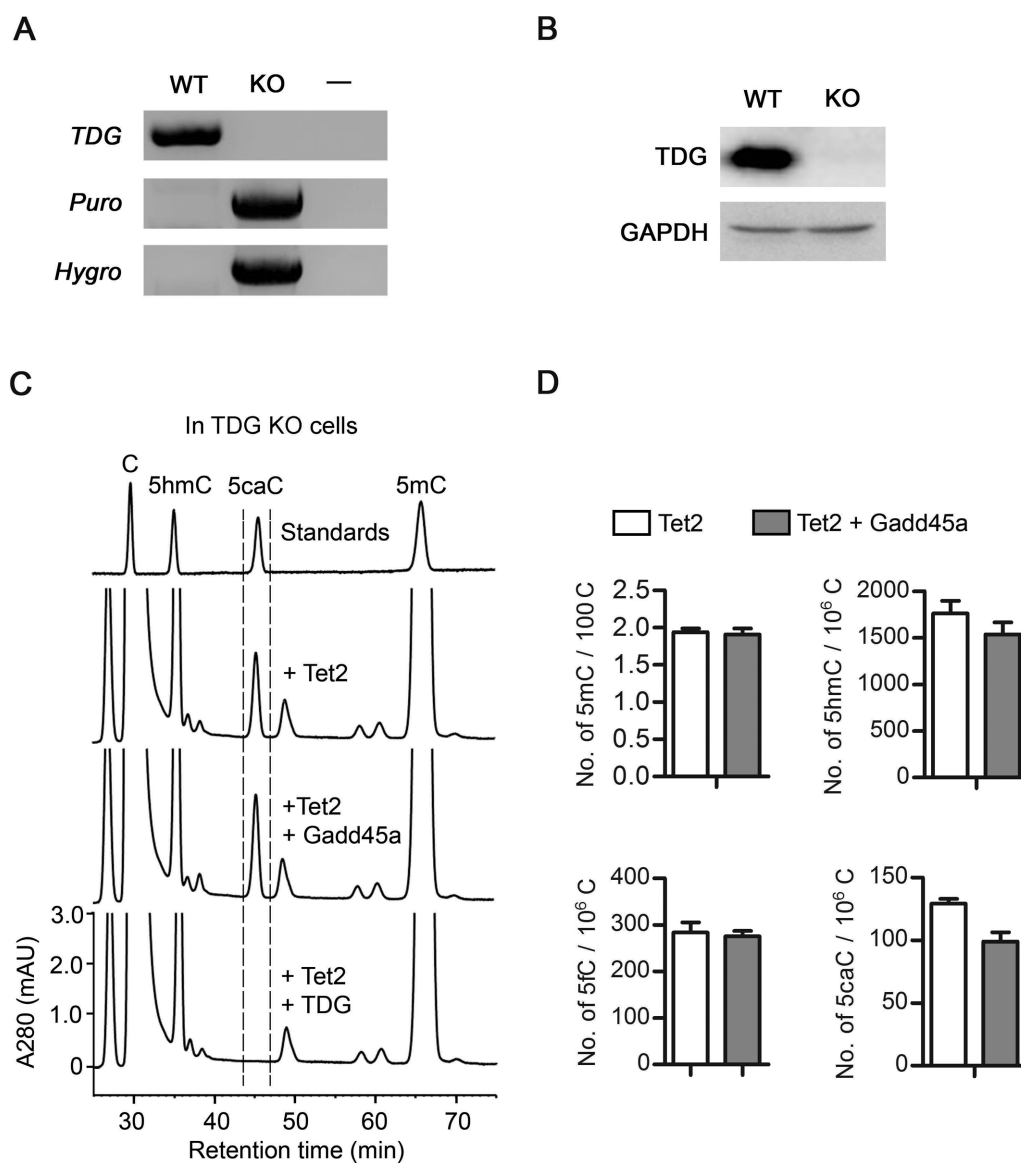
Figure 4. Reduction of genomic 5caC by Gadd45a depends on TDG in transfected HEK293T cells. (**A**) Confirmation of the *TDG* gene knockout (KO) in HEK293T cells by PCR genotyping of genomic DNA. The two *TDG* alleles present in wild-type (WT) cells were replaced by puromycin (Puro) and hygromycin (Hygro) drug selection markers in KO cells (Supplementary Figure S3). (**B**) Loss of the endogenous TDG protein in the established knockout cell line confirmed by western analysis using anti-TDG antibody. Detection with anti-GAPDH served as a loading control. (**C**) HPLC analysis of genomic 5caC in TDG-knockout cells expressing ectopic full-length Tet2 alone and together with Gadd45a. Transfection of the TDG control eliminates 5caC generated by Tet2, as previously reported (15). (**D**) Mass spectrometry quantification of modified cytosines in the genomic DNA of TDG knockout cells ectopically expressing Tet2 alone or Tet2 together with and Gadd45a. Data are represented as the mean ± SEM from two independent experiments.

with the acquisition of drug selection markers in the established knockout cell line (Figure 4A). Loss of TDG protein expression was confirmed by western blot analysis using an anti-TDG antibody (Figure 4B). As shown in the HPLC analysis of genomic DNA from transfected cells, expression of Gadd45a in TDG-deficient cells did not change the level of 5caC generated by Tet2 (Figure 4C). Mass spectrometry analyses of the genomic DNA samples confirmed that levels of 5fC and 5caC were not changed by Gadd45a expression in TDG-knockout cells (Figure 4D). The possibility that Gadd45a may increase the expression level of Tet2 thus masking its effect by generating higher level of 5caC was ruled out (Supplementary Figure S2C). Additionally,

the complete removal of 5caC upon re-expression of ectopic TDG in the knockout cells validated the cellular competence for TDG function. These results indicate that the function of Gadd45a in DNA demethylation is realized exclusively through its effect on TDG.

Gadd45a promotes conversion of 5fC and 5caC to unmodified cytosine in a TDG-dependent manner

As shown above, Gadd45a reduces the buildup of genomic 5fC and 5caC formed by ectopic Tet2 in transfected cells. To directly demonstrate the role of Gadd45a, we prepared *in vitro* a 5fC/5caC-containing reporter plasmid and trans-
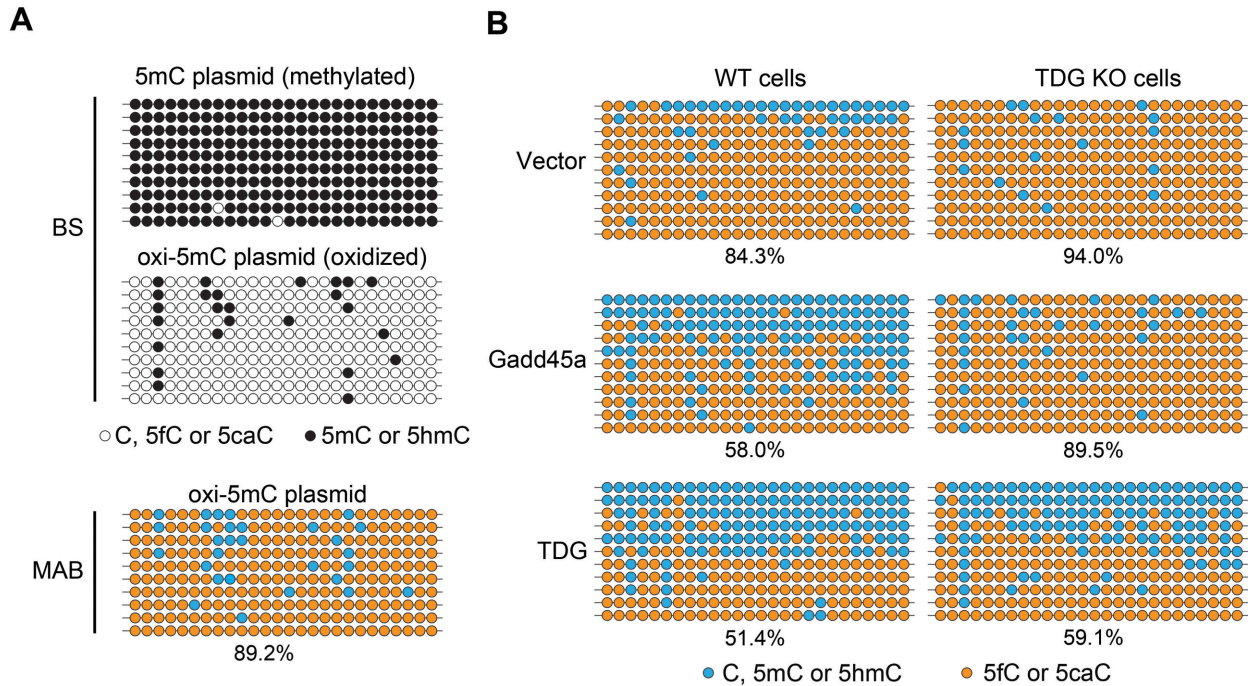
Figure 5. Gadd45a promotes conversion of 5fC and 5caC into unmodified cytosine in a TDG-dependent manner in transfected cells. (**A**) Characterization of oxidized pCpGL-CMV reporter plasmid DNA (oxi-5mC plasmid). Plasmid DNA was methylated *in vitro* with CpG methyltransferase M.SssI and then oxidized using human TET2 enzyme to obtain oxi-5mC DNA. BS-seq and MAB-seq profiles show occurrence of expected modifications at CpGs in the CMV promoter region. There was about 10% 5mC or 5hmC remaining in the oxi-5mC DNA due to incomplete oxidization with TET2. The white circles denote unmodified C or higher oxidized forms of 5mC (5fC and 5caC) and the black circles denote 5mC or 5hmC in BS analysis. The blue circles in the MAB-seq profile represent C, 5mC or 5hmC. The orange circles represent higher oxidized forms (5fC or 5caC). (**B**) MAB-seq analysis of oxidized reporter plasmid DNA recovered from transfected HEK293T cells. The oxi-5mC plasmid DNA was transfected into WT or TDG knockout cells with Gadd45a, TDG or vector control. Percentages of the oxi-5mC bases (5fC and 5caC) are indicated.

fected the plasmid with the Gadd45a expression plasmid into HEK293T cells. The recovered reporter plasmid DNA was then analyzed by M.SssI-assisted bisulfite sequencing (MAB-seq) which allows to distinguish unmodified cytosines from 5fC/5caC in a specific sequence at single-base resolution (31,32). We first confirmed the efficient methylation (5mC 99.2%) and oxidation (5fC + 5caC; 89.2%) of the CpG sites in the CMV promoter region of the prepared plasmid (oxi-5mC) using bisulfite-seq and MAB-seq (Figure 5A; Supplementary Figure S4). Interestingly, in the recovered plasmid upon co-transfection with Gadd45a, the combined level of 5fC and 5caC was reduced to 58% while the DNA without Gadd45a co-transfection retained the majority of 5fC/5caC (84.3 versus 89.2%, Figure 5B). Consistent with the idea of TDG activity mediating the Gadd45a effect, overexpression of TDG also reduced the 5fC/5caC level to 51%. Moreover, Gadd45a had no effect on the 5fC/5caC level in TDG KO cells (89.5 versus 89.2%, Figure 5B), indicating that the endogenous TDG is essential to mediate the function of Gadd45a in HEK293T cells. Taken together, Gadd45a promotes demethylation by positively impacting TDG for the conversion of 5fC and 5caC into C.

Gadd45a/b knockout causes locus-specific hypermethylation in mouse ES cells

If Gadd45a positively regulates Tet/TDG-mediated DNA demethylation in ES cells, Gadd45a deficiency would lead to hypermethylation in genomic regions known to be the target loci of active demethylation. To eliminate the effect of potential functional redundancy with Gadd45b that has 55% protein sequence identity and has a similar mRNA expression level with Gadd45a in ES cells, we generated *Gadd45a/Gadd45b* double-knockout (DKO) mouse ES cell lines using the CRISPR/Cas9 system. Nucleotide insertions and deletions at the targeted sites were identified which would lead to frame shift mutation or early termination, thus gene inactivation (Supplementary Figure S5). To compare genomic methylation between DKO and the WT ES cells, we performed genome-scale single-base-resolution analysis using RRBS. In the DKO cells, no obvious methylation difference was observed at the genome scale (Supplementary Figure S6). Both the mutant and WT ES cells have a CpG methylation frequency of around 33% in the RRBS-covered genome part and a similar overall gene distribution pattern. However, 68 specific regions were found hypermethylated which contained more than 4 CpG sites of significantly increased methylation (Supplementary Table S3), as exemplified by the genomic loci from the *Plagl1* and *Cilp2* genes (Figure 6A). In sharp contrast, no hypomethylated regions could be identified in these DKO ES cells. Primer-based conventional bisulfite sequencing confirmed significantly elevated methylation in the corresponding regions of *Plagl1* and *Cilp2* (Figure 6B). Interestingly, these two regions also gained hypermethylation in control ES cells deficient in TDG (Figure 6B). Since 5mC and 5hmC
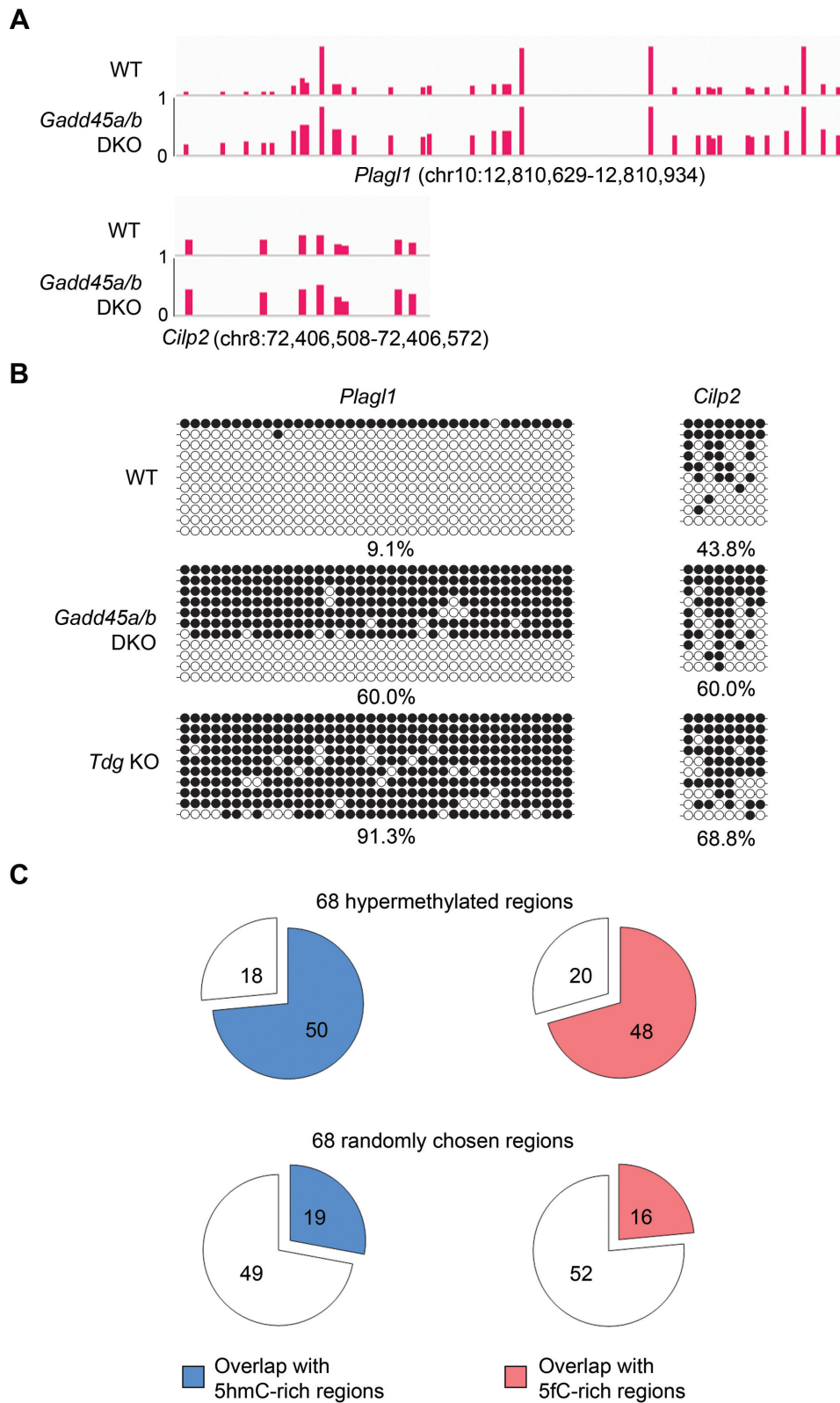
Figure 6. Locus-specific hypermethylation in *Gadd45a/b* DKO mouse ES cells. (**A**) Representative loci of hypermethylation in *Gadd45a/b* DKO ES cells mapped by RRBS analysis. Vertical bars indicate the methylation level (0–1) at individual CpGs in the genomic regions indicated. (**B**) Confirmation of increased methylation at the selected loci by bisulfite sequencing in WT, *Gadd45a/b* DKO and *Tdg* KO ES cells. Black and white circles denote methylated and unmethylated CpG sites respectively. Percentages of the CpGs sites resistant to bisulfite conversion are indicated below. (**C**) Comparison of the 68 hypermethylated regions identified in *Gadd45a/b* DKO ES cells (top pie charts) or 68 randomly chosen RRBS covered regions (bottom) with the pool of 5hmC-rich regions (24,42) (left) and the pool of 5fC-rich regions (24) (right) previously identified in *Tdg* KO ES cells.

are not distinguishable in conventional bisulfite sequencing, we performed Tet-assisted bisulfite sequencing (TAB-seq) to determine 5hmC in the *Plagl1* and *Cilp2* regions. We found 5hmC very rare and no increase in both cell lines deficient in Gadd45 or TDG (Supplementary Figure S7). This indicates that the hypermethylation was actually composed of 5mC rather than 5hmC. Further inspection of the Gadd45 deficiency–caused hypermethylated regions revealed that most of them were overlapped with 5hmC and 5fC-enriched regions previously mapped in ES cells (24,42). Among the 68 hypermethylated regions, 50 are overlapped with 174 655 regions reported to be enriched in 5hmC and 48 are overlapped with 120 052 regions enriched in 5fC (Figure 6C). By comparison, much fewer regions randomly chosen from the RRBS-covered sequences are overlapped with 5hmC or 5fC regions. Overall, these results support the conclusion that Gadd45 proteins functionally cooperate with the Tet and TDG enzymes in demethylation of genomic targets in mouse ES cells.

DISCUSSION

Mechanisms of DNA demethylation have been intensively studied but remain elusive. There exist a wide variety of factors that belong to different pathways but appear to regulate DNA methylation negatively (13). As biochemical evidence revealing their participation in the demethylation process is lacking in most cases, it is plausible that the role of these factors in DNA methylation control may be indirect. In the presented work, we have examined the role of Gadd45a in the context of the recently proposed Tet-TDG mediated demethylation pathway (41). Our data indicate that Gadd45a is involved in oxidative demethylation by regulating the activity of TDG in cells.

The Tet family of proteins, or DNA dioxygenases by nature, have emerged as the most plausible enzymes for initiation of active demethylation, owing to their ability to oxidize the methyl group of 5mC (14,43). While the simple oxidation of 5mC to 5hmC will demethylate DNA passively through replicative dilution, further oxidation to 5fC and 5caC generates substrates for TDG, which will remove these bases, triggering their replacement by unmethylated cytosine, potentially via a conventional BER process (19,41). As demethylation is a highly regulated process, we reasoned that auxiliary factors including those previously implicated in demethylation may function to modulate the activity of the Tet-TDG system. Among such factors is the multifunctional Gadd45a initially identified to antagonize methylation by its activation effect on methylated reporter genes (27). Gadd45a was proposed to promote demethylation potentially through XPG-dependent DNA repair (27,44) or BER of AID-based deamination products (22,29). Our finding that Gadd45a serves as a regulator in the Tet-TDG pathway is supported by several lines of evidence. First, Gadd45a synergizes with Tet and TDG to activate methylated reporter gene in transfected cells (Figure 1), and overexpression of Gadd45a alone cannot reactive the methylated reporter gene, consistent with the previous report by Jin *et al.* (30). Second, Gadd45a interacts with TDG physically *in vitro* (Figure 3). Third, Gadd45a potentiates TDG glycosylase to remove 5fC and 5caC from genomic DNA of transfected HEK293T cells (Figure 2). Forth, endogenous TDG is required to mediate the effect of ectopic Gadd45a in decreasing Tet2-generated 5caC (Figures 4 and 5). Lastly, deletion of *Gadd45a/b* in mouse ES cells leads to hypermethylation at specific genomic loci which also gain increased methylation in *Tdg*-deficient cells and are enriched in 5fC (Figure 6). Since the molecular functions of Gadd45a appear to be diverse (26), it is noteworthy that Gadd45a promotes demethylation specifically through TDG in our studies. However, since the Gadd45 deficiency-caused hypermethylated regions are much fewer than the previously mapped 5hmC and 5fC enriched regions in ES cells (24,42) (68 versus 174 655 and 120 052), Gadd45 may only contribute to the control of a small fraction of Tet-TDG regulated genomic targets.

It is curious how Gadd45a exactly regulates TDG activity in cells. In light of the direct interaction between the two proteins and the stimulation of TDG-dependent DNA demethylation by Gadd45a in cells, we examined the possibility that Gadd45a might impact on the enzymatic activity of TDG *in vitro* using purified proteins. Under multiple-turnover conditions, addition of an excess of Gadd45a stimulated initial processing by as well as overall efficiency of TDG but it didn't improve its rate limiting turnover (Supplementary Figure S8A and B). The stimulation by Gadd45a, however, was in a similar range as observed when a molar equivalent of BSA was added to the reaction, indicating a non-specific effect of the protein on the structural integrity of TDG. Likewise, under single-turnover conditions, the addition of Gadd45a had no measurable effect on TDG activity, suggesting that Gadd45a does not modulate TDG substrate recognition potential (Supplementary Figure S8C and D). Taken together, Gadd45a does not appear to stimulate TDG in substrate binding, processing or turnover in a specific manner. We therefore favor a model whereby Gadd45a promotes Tet-TDG-mediated demethylation by coordinating the recruitment of TDG into demethylation complexes. Gadd45a (and Gadd45b) could act as a scaffold protein which interacts with other epigenetic regulators (26) and directs TDG to the target loci in the context of chromatin to efficiently excise 5fC and 5caC. This is consistent with the recently proposed model in which recruitment of both Gadd45a and TDG is implicated in demethylation of the tumor suppressor gene *TCF21* targeted by a long non-coding RNA (45). Further investigation will be needed to elucidate the exact molecular mechanism by which Gadd45 proteins regulate the function of the TDG enzyme in specific chromatin contexts.

ACCESSION NUMBER

The Gene Expression Omnibus accession number for the RRBS data reported in this manuscript is GSE62496.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Jaenisch,R. and Bird,A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, **33**(Suppl.), 245–254.
2. Goll,M.G. and Bestor,T.H. (2005) Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.*, **74**, 481–514.
3. Simonsson,S. and Gurdon,J. (2004) DNA demethylation is necessary for the epigenetic reprogramming of somatic cell nuclei. *Nat. Cell Biol.*, **6**, 984–990.
4. Seisenberger,S., Peat,J.R., Hore,T.A., Santos,F., Dean,W. and Reik,W. (2013) Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **368**, 20110330.
5. Kangaspeska,S., Stride,B., Metivier,R., Polycarpou-Schwarz,M., Ibberson,D., Carmouche,R.P., Benes,V., Gannon,F. and Reid,G. (2008) Transient cyclical methylation of promoter DNA. *Nature*, **452**, 112–115.
6. Hajkova,P., Erhardt,S., Lane,N., Haaf,T., El-Maarri,O., Reik,W., Walter,J. and Surani,M.A. (2002) Epigenetic reprogramming in mouse primordial germ cells. *Mech. Dev.*, **117**, 15–23.
7. Mayer,W., Niveleau,A., Walter,J., Fundele,R. and Haaf,T. (2000) Demethylation of the zygotic paternal genome. *Nature*, **403**, 501–502.
8. Oswald,J., Engemann,S., Lane,N., Mayer,W., Olek,A., Fundele,R., Dean,W., Reik,W. and Walter,J. (2000) Active demethylation of the paternal genome in the mouse zygote. *Curr. Biol.*, **10**, 475–478.
9. Gu,T.P., Guo,F., Yang,H., Wu,H.P., Xu,G.F., Liu,W., Xie,Z.G., Shi,L., He,X., Jin,S.G. *et al.* (2011) The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature*, **477**, 606–610.
10. Wossidlo,M., Nakamura,T., Lepikhov,K., Marques,C.J., Zakhartchenko,V., Boiani,M., Arand,J., Nakano,T., Reik,W. and Walter,J. (2011) 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat. Commun.*, **2**, 241.
11. Walsh,C.P. and Xu,G.L. (2006) Cytosine methylation and DNA repair. *Curr. Top. Microbiol. Immunol.*, **301**, 283–315.
12. Ooi,S.K. and Bestor,T.H. (2008) The colorful history of active DNA demethylation. *Cell*, **133**, 1145–1148.
13. Wu,H. and Zhang,Y. (2014) Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell*, **156**, 45–68.
14. Pastor,W.A., Aravind,L. and Rao,A. (2013) TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat. Rev. Mol. Cell Biol.*, **14**, 341–356.
15. He,Y.F., Li,B.Z., Li,Z., Liu,P., Wang,Y., Tang,Q., Ding,J., Jia,Y., Chen,Z., Li,L. *et al.* (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, **333**, 1303–1307.
16. Ito,S., Shen,L., Dai,Q., Wu,S.C., Collins,L.B., Swenberg,J.A., He,C. and Zhang,Y. (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, **333**, 1300–1303.
17. Tahiliani,M., Koh,K.P., Shen,Y., Pastor,W.A., Bandukwala,H., Brudno,Y., Agarwal,S., Iyer,L.M., Liu,D.R., Aravind,L. *et al.* (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930–935.
18. Maiti,A. and Drohat,A.C. (2011) Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J. Biol. Chem.*, **286**, 35334–35338.
19. Kohli,R.M. and Zhang,Y. (2013) TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, **502**, 472–479.
20. Jacobs,A.L. and Schar,P. (2012) DNA glycosylases: in DNA repair and beyond. *Chromosoma*, **121**, 1–20.
21. Cortazar,D., Kunz,C., Selfridge,J., Lettieri,T., Saito,Y., MacDougall,E., Wirz,A., Schuermann,D., Jacobs,A.L., Siegrist,F. *et al.* (2011) Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature*, **470**, 419–423.
22. Cortellino,S., Xu,J., Sannai,M., Moore,R., Caretti,E., Cigliano,A., Le Coz,M., Devarajan,K., Wessels,A., Soprano,D. *et al.* (2011) Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell*, **146**, 67–79.
23. Shen,L., Wu,H., Diep,D., Yamaguchi,S., D'Alessio,A.C., Fung,H.L., Zhang,K. and Zhang,Y. (2013) Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell*, **153**, 692–706.
24. Song,C.X., Szulwach,K.E., Dai,Q., Fu,Y., Mao,S.Q., Lin,L., Street,C., Li,Y., Poidevin,M., Wu,H. *et al.* (2013) Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*, **153**, 678–691.
25. Fornace,A.J Jr, Alamo,I. Jr and Hollander,M.C. (1988) DNA damage-inducible transcripts in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 8800–8804.
26. Niehrs,C. and Schafer,A. (2012) Active DNA demethylation by Gadd45 and DNA repair. *Trends Cell Biol.*, **22**, 220–227.
27. Barreto,G., Schafer,A., Marhold,J., Stach,D., Swaminathan,S.K., Handa,V., Doderlein,G., Maltry,N., Wu,W., Lyko,F. *et al.* (2007) Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation. *Nature*, **445**, 671–675.
28. Ma,D.K., Jang,M.H., Guo,J.U., Kitabatake,Y., Chang,M.L., Pow-Anpongkul,N., Flavell,R.A., Lu,B., Ming,G.L. and Song,H. (2009) Neuronal activity-induced Gadd45b promotes epigenetic DNA demethylation and adult neurogenesis. *Science*, **323**, 1074–1077.
29. Rai,K., Huggins,I.J., James,S.R., Karpf,A.R., Jones,D.A. and Cairns,B.R. (2008) DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. *Cell*, **135**, 1201–1212.
30. Jin,S.G., Guo,C. and Pfeifer,G.P. (2008) GADD45A does not promote DNA demethylation. *PLoS Genet.*, **4**, e1000013.
31. Hu,X., Zhang,L., Mao,S.Q., Li,Z., Chen,J., Zhang,R.R., Wu,H.P., Gao,J., Guo,F., Liu,W. *et al.* (2014) Tet and TDG mediate DNA demethylation essential for mesenchymal-to-epithelial transition in somatic cell reprogramming. *Cell Stem Cell*, **14**, 512–522.
32. Guo,F., Li,X., Liang,D., Li,T., Zhu,P., Guo,H., Wu,X., Wen,L., Gu,T.P., Hu,B. *et al.* (2014) Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell*, **15**, 447–458.
33. Klug,M. and Rehli,M. (2006) Functional analysis of promoter CpG methylation using a CpG-free luciferase reporter vector. *Epigenetics*, **1**, 127–130.
34. Yin,R., Mao,S.Q., Zhao,B., Chong,Z., Yang,Y., Zhao,C., Zhang,D., Huang,H., Gao,J., Li,Z. *et al.* (2013) Ascorbic acid enhances Tet-mediated 5-methylcytosine oxidation and promotes DNA demethylation in mammals. *J. Am. Chem. Soc.*, **135**, 10396–10403.
35. Doyle,E.L., Booher,N.J., Standage,D.S., Voytas,D.F., Brendel,V.P., Vandyk,J.K. and Bogdanove,A.J. (2012) TAL effector-nucleotide targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Res.*, **40**, W117–W122.
36. Ran,F.A., Hsu,P.D., Wright,J., Agarwala,V., Scott,D.A. and Zhang,F. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.*, **8**, 2281–2308.

37. Hu,L., Li,Z., Cheng,J., Rao,Q., Gong,W., Liu,M., Shi,Y.G., Zhu,J., Wang,P. and Xu,Y. (2013) Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation. *Cell*, **155**, 1545–1555.

38. Rohde,C., Zhang,Y., Reinhardt,R. and Jeltsch,A. (2010) BISMA–fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. *BMC Bioinformatics*, **11**, 230.

39. Kunz,C., Focke,F., Saito,Y., Schuermann,D., Lettieri,T., Selfridge,J. and Schar,P. (2009) Base excision by thymine DNA glycosylase mediates DNA-directed cytotoxicity of 5-fluorouracil. *PLoS Biol.*, **7**, e91.

40. Nguyen,T.N. and Goodrich,J.A. (2006) Protein-protein interaction assays: eliminating false positive interactions. *Nat. Methods*, **3**, 135–139.

41. Xu,G.L. and Walsh,C.P. (2014) Enzymatic DNA oxidation: mechanisms and biological significance. *BMB Rep.*, **47**, 609–618.

42. Yu,M., Hon,G.C., Szulwach,K.E., Song,C.X., Zhang,L., Kim,A., Li,X., Dai,Q., Shen,Y., Park,B. *et al.* (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, **149**, 1368–1380.

43. Tan,L. and Shi,Y.G. (2012) Tet family proteins and 5-hydroxymethylcytosine in development and disease. *Development*, **139**, 1895–1902.

44. Schmitz,K.M., Schmitt,N., Hoffmann-Rohrer,U., Schafer,A., Grummt,I. and Mayer,C. (2009) TAF12 recruits Gadd45a and the nucleotide excision repair complex to the promoter of rRNA genes leading to active DNA demethylation. *Mol. Cell*, **33**, 344–353.

45. Arab,K., Park,Y.J., Lindroth,A.M., Schafer,A., Oakes,C., Weichenhan,D., Lukanova,A., Lundin,E., Risch,A., Meister,M. *et al.* (2014) Long noncoding RNA TARID directs demethylation and activation of the tumor suppressor TCF21 via GADD45A. *Mol. Cell*, **55**, 604–614.

Figure S1. Related to Figure 1
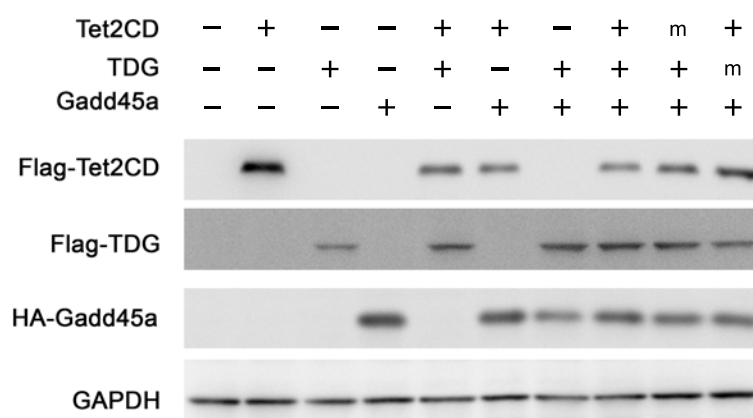
Figure S2. Related to Figure 2 and Figure 4

A

Tet2 +　　／　　Gadd45a

Flag-Tet2

GAPDH

WT HEK293T cells

B

Vector　Tet2　Tet2 + Gadd45a

TDG

GAPDH

WT HEK293T cells

C

Tet2 +　　／　　TDG　　Gadd45a

Flag-Tet2

GAPDH

TDG KO HEK293T cells
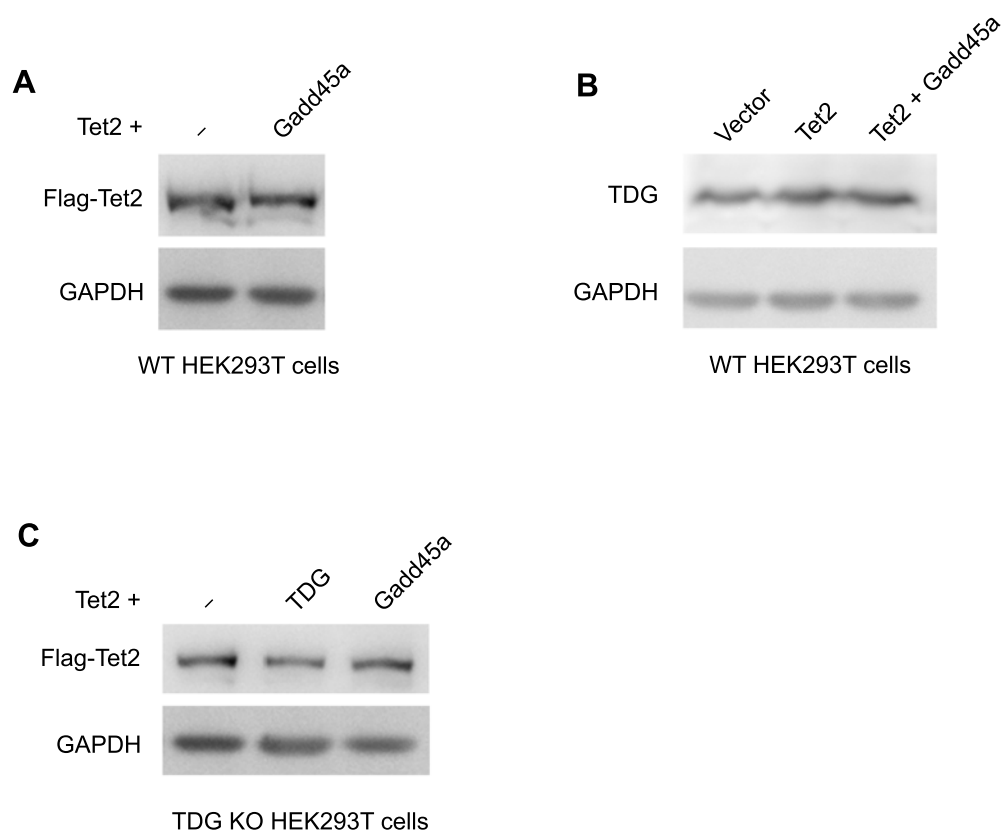
Figure S3. Related to Figure 4

1 kb

ATG

Exon 1

TALEN cleavage

Exon 2

Exon 3

Targeting vector (hygro)
PGK-hygro

WT allele I

WT allele II

Targeting vector (puro)
PGK-puro

Hygromycin + Puromycin

PGK-hygro — Targeted allele I

PGK-puro — Targeted allele II

— HDR left arm (1.44 kb)   — HDR right arm (1.02 kb)
— TALEN left arm   — TALEN right arm
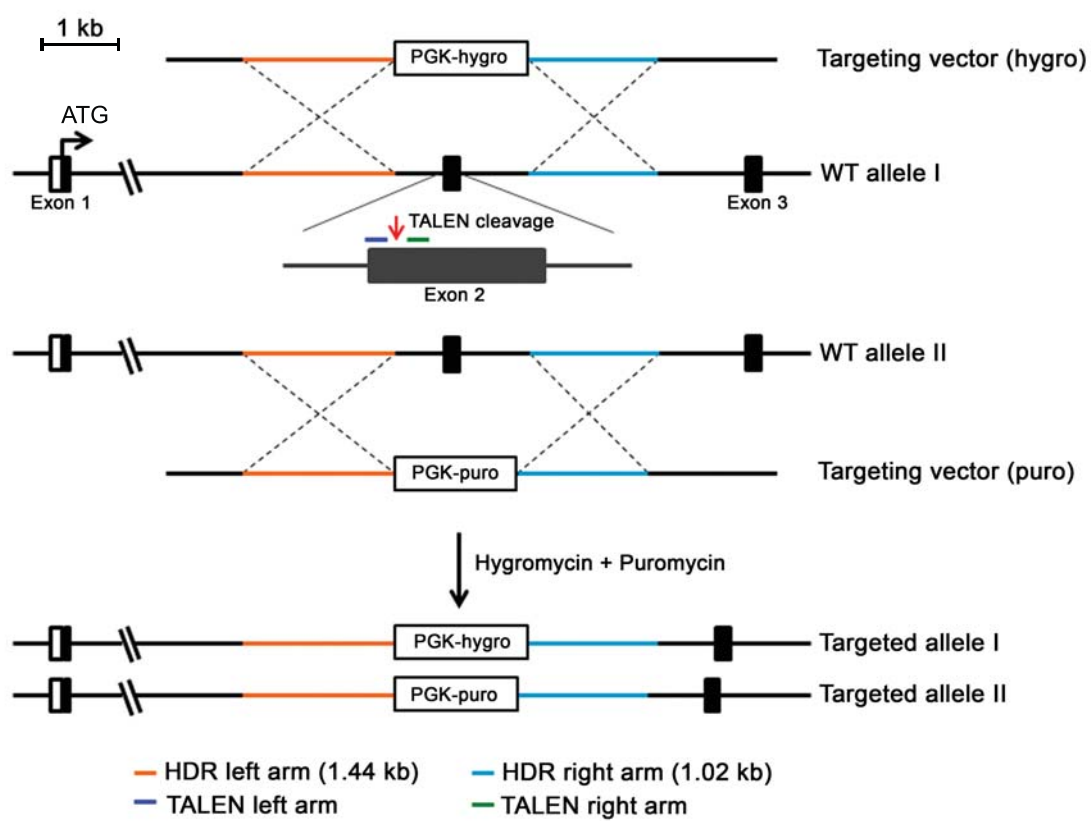
Figure S4. Related to Figure 5

A
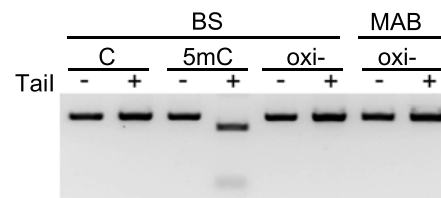
B
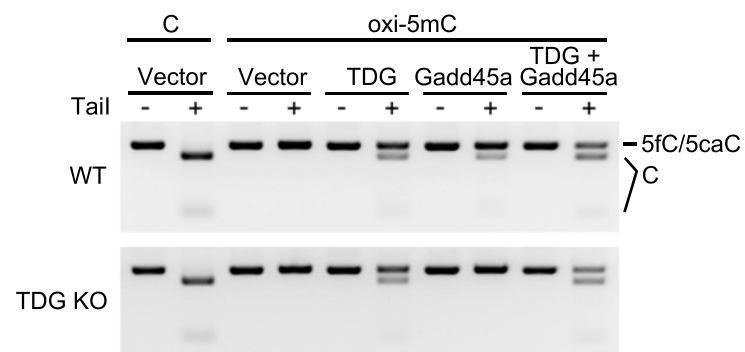
Figure S5. Related to Figure 6

A

50 bp

Exon 1   2

Gadd45a

EcoRI
cacttgcaatatgactttggag**GAATTC**cggctgcagagc

Exon 1   2

Gadd45b

BstUI
tggaagagctggtggcgagcgacaa**CGCG**gttcagaa

B

Gadd45a

Gadd45b

C

Double mutant clone #2

GGACTCGCACTTGCAATATGACTTTGGAGGAATTCTCGGCTGCAGAGCAGAAG
Gadd45a GGACTCGCACTTGCAATATGACTTTG--------CTCGGCTGCAGAGCAGAAG -8 bp
GGACTCGCACTTGCAATATGACTTTG--------CTCGGCTGCAGAGCAGAAG

GCATCATGACCCTGGAAGAGCTGGTGGCGAGCGACAACGCGGTTCAGAAGTG
Gadd45b GCATCATGACCCTGGAAGAGCTGGTGGCGAG--------GCGGTTCAGAAGTG -8 bp
GCATCATGACCCTGGAAGAGCTGGTGGCGAG--------GCGGTTCAGAAGTG

Double mutant clone #6

GGACTCGCACTTGCAATATGACTTTGGAGGAATTCTCGGCTGCAGAGCAGAAG
Gadd45a GGACTCGCACTTGCAATATGACTTTGGAGGAATTCTCGGCTGCAGAGCAGAAG +1 bp
                                    T
GGACTCGCACTTGCAATATGACTTTGGAGGAATTCTCGGCTGCAGAGCAGAAG
                                    T

GCATCATGACCCTGGAAGAGCTGGTGGCGAGCGACAACGCGGTTCAGAAGTG
Gadd45b GCATCATGACCCTGGAAGAGCTGGTGGCGAGCG--------GTTCAGAAGTG -8 bp
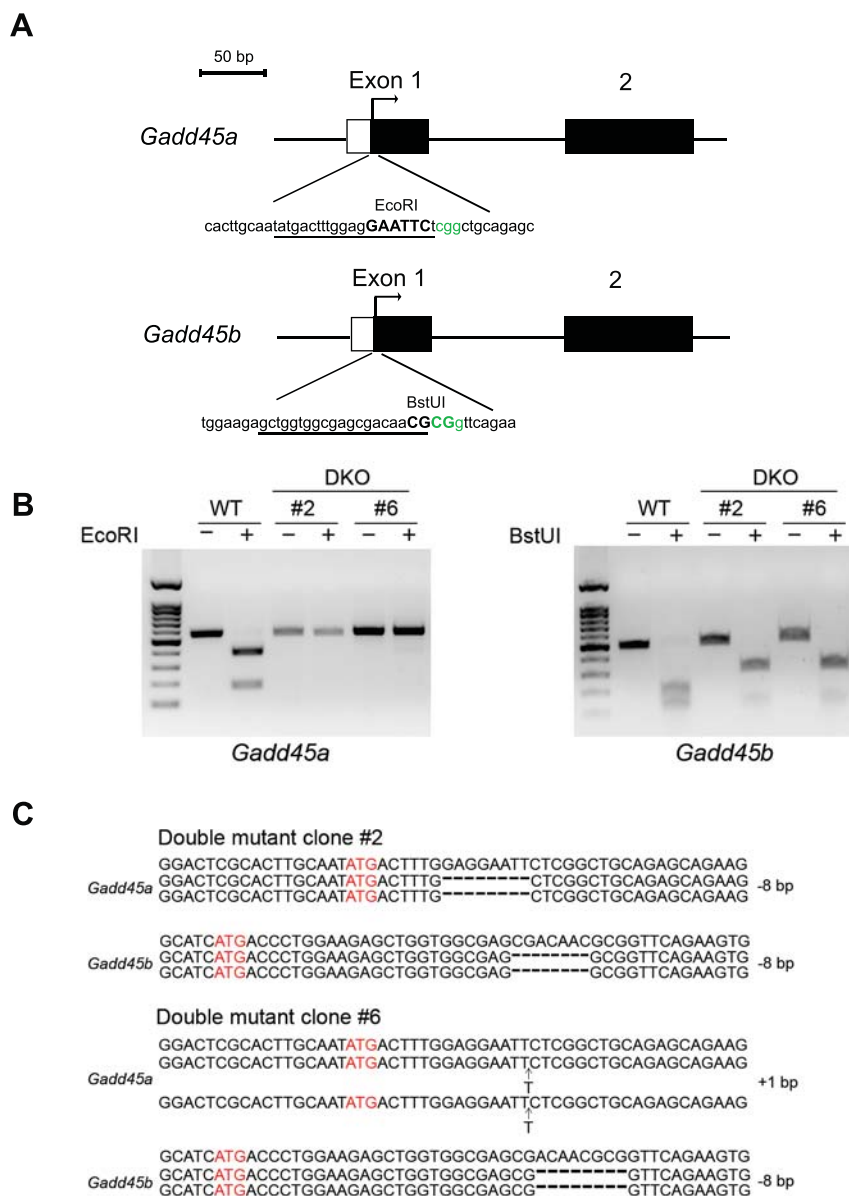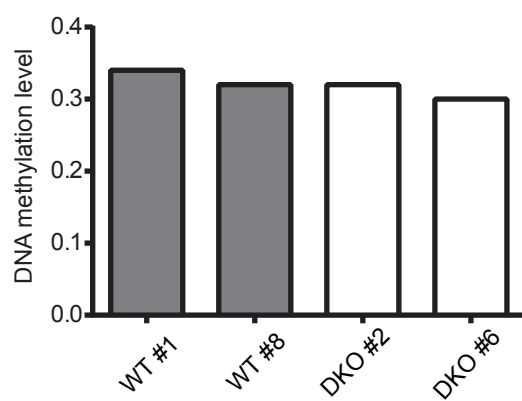GCATCATGACCCTGGAAGAGCTGGTGGCGAGCG--------GTTCAGAAGTG

Figure S6. Related to Figure 6

A

B
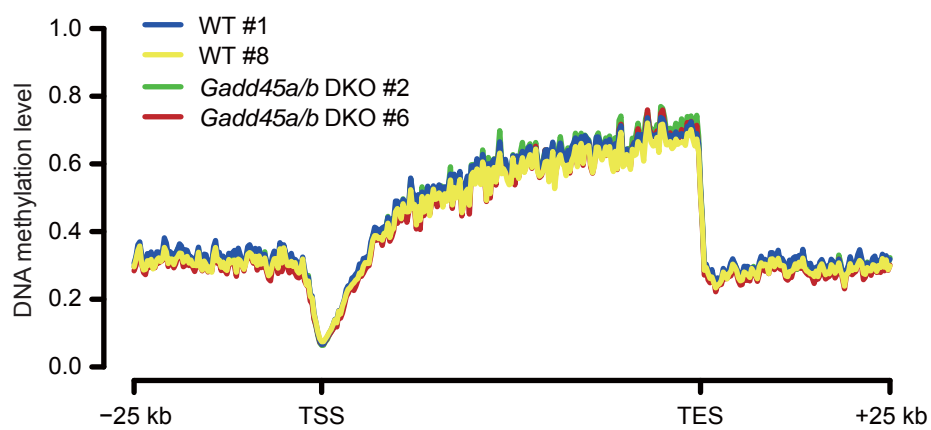
Figure S7. Related to Figure 6

Plagl1

Cilp2

WT

Gadd45a/b
DKO

Tdg KO

● 5hmC

Figure S8.

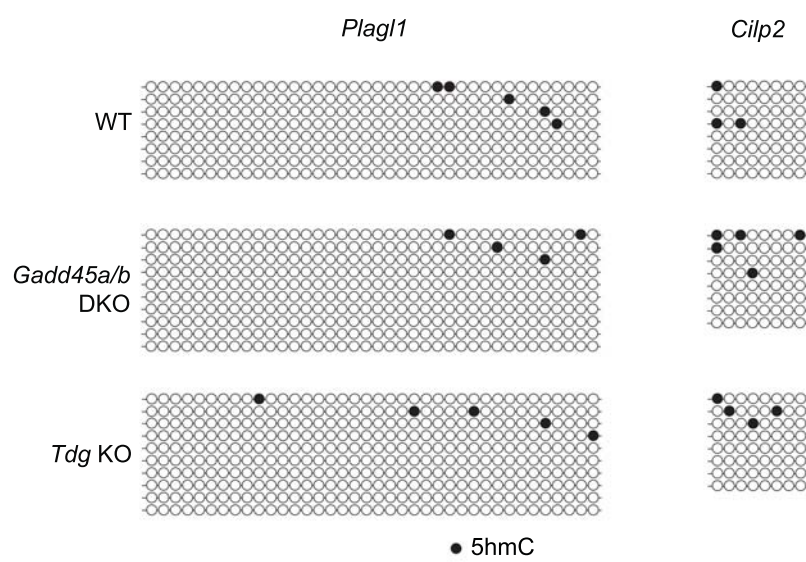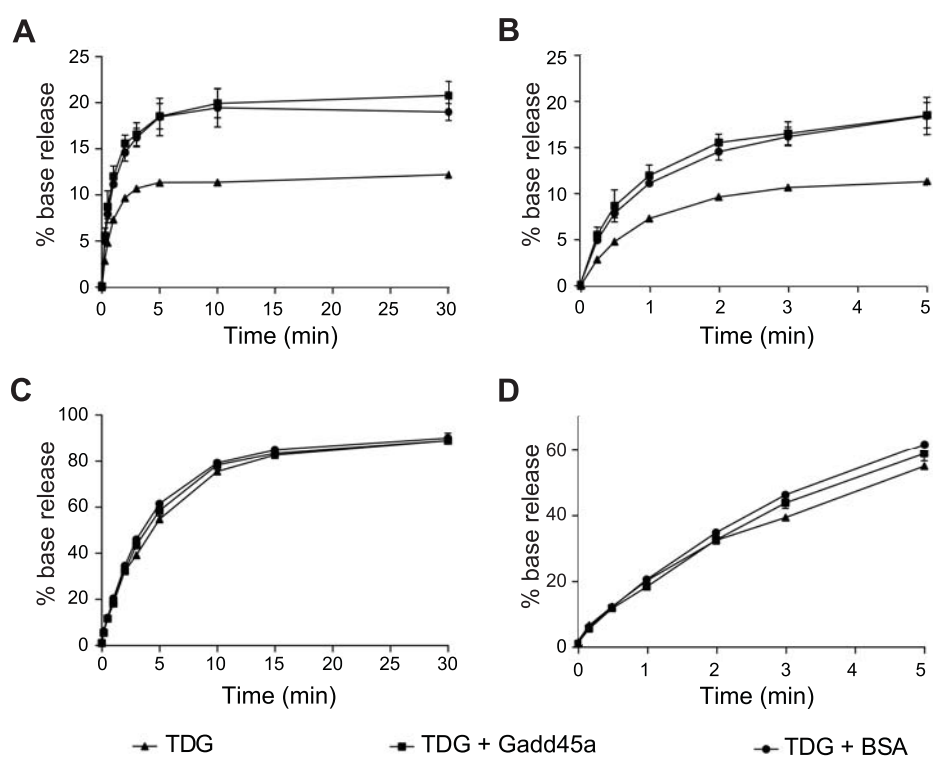**A**

B

C

D

TDG   TDG + Gadd45a   TDG + BSA

Supplementary Figure Legends

Figure S1. Western blotting assay to confirm the expression of proteins co-transfected with the luciferase reporter plasmids. Anti-Tet2, -Flag and -HA antibodies were used to detect the respective proteins expressed in transfected HEK293T cells in the luciferase assay (Figure 1). Detection with anti-GAPDH served as a loading control.

Figure S2. Western blotting assays to confirm the unaltered expression of ectopic Flag-Tet2 and endogenous TDG in transfected HEK293T cells.

(A) Flag-Tet2 expression was not altered in wild-type cells when co-transfected with Gadd45a. Western blotting analysis was performed with anti-Flag. The GAPDH detection served as a loading control of the protein extracts.

(B) Endogenous TDG level was not affected by transfection with Tet2 alone or with Gadd45a. Western analysis was performed with anti-TDG.

(C) Flag-Tet2 expression was not altered when co-transfected with Gadd45a or TDG in TDG knockout HEK293T cells. Western analysis was performed with anti-Flag.

Figure S3. Gene targeting strategy of *TDG* in HEK293T cells. The two alleles of exon 2 were replaced by the puromycin or hygromycin resistance genes using TALEN technology.

Figure S4. Characterization of the oxi-pCpGL-CMV luciferase reporter plasmid and analysis of recovered oxi-plasmid after transfection with other expression constructs.

(A) COBRA analysis of C, 5mC, oxi-, or M.SssI-incubated oxi- plasmids prepared *in vitro*. The 5mC plasmid was generated from C plasmid by M.SssI methylation, and oxi-plasmid from the 5mC plasmid by Tet2 oxidation.

(B) Restriction assay for 5fC and 5caC analysis of oxi-plasmid recovered from nuclei of HEK293T cells transfected with empty vector, Gadd45a, TDG or Gadd45a and TDG expression constructs. The recovered plasmid DNA was treated with M.SssI and bisulfite conversion. PCR products amplified from the CMV region were digested with TaiI. The appearance of shorter fragments from TaiI digestion indicates 5fC/5caC removal and cytosine

restoration at the TaiI CpG sites in the CMV promoter. Unmodified plasmid (C-CMV) was used as a control to monitor the *in vitro* M.SssI methylation efficiency.

Figure S5. Targeting of the *Gadd45a* and *Gadd45b* genes in mouse ES cells.

(A) Schematic of the Cas9/sgRNA-targeting sites in *Gadd45a* and *Gadd45b*. The sgRNA-targeting sequence is underlined, and the protospacer adjacent motif (PAM) sequence is labeled in green. The nucleotides at the target regions used for restriction analysis are bold.

(B) Genotyping of DKO ES cell lines #2 and #6 by restriction analysis. Gadd45a PCR products amplified from the exon 1 region were digested with EcoRI. Expected fragment sizes: WT = 180 + 369 bp, Mutant = 549 bp. Gadd45b PCR products were digested with BstUI. Expected fragment sizes: WT = 182 + 118 + 113 + 56 + 24 bp, Mutant = 300 + 113 + 56 + 24 bp.

(C) The sequences of mutated alleles in double-knockout ES cell lines #2 and #6. The start codon is labeled in red.

Figure S6. Global DNA methylation was unaltered in *Gadd45a/b* knockout mouse ES cells.

(A) DNA methylation levels of wild-type (WT) and *Gadd45a/b* DKO (DKO) ES cells. The y-axis indicates the proportion of methylated CpG sites in the RRBS-covered genome part, calculated based on the RRBS data. # denotes independent ES cell lines.

(B) Averaged DNA methylation levels along the gene bodies and 25-kb regions upstream and downstream of the gene body. TSS, transcription start site. TES, transcription end site.

Figure S7. 5hmC profiles of the selected two regions (*Plagl1* and *Cilp*) in *Gadd45a/b* double knockout (DKO), *Tdg* knockout (KO) and wild-type (WT) ES cells. 5hmC was determined by Tet-assisted bisulfite sequencing (TAB-seq).

Figure S8. Gadd45 does not stimulate the activity of TDG *in vitro*.

(A) Quantitation of base release activities of TDG with or without Gadd45a (or BSA) under multiple-turnover conditions. A 60-bp double-strand oligonucleotide duplex (125 nM) containing a G•T mismatch end-labeled with fluorescein was incubated with purified TDG (25 nM) and Gadd45a or BSA (1 µM) at 37°C. The 60-mer substrate band and a 23-mer product

band separated on a denaturing gel were quantified using the ImageQuant TL software. The y

axis indicates the percentage of 23-mer product bands to the initial 60-mer substrate. The x

axis indicates different reaction time points.

(B) Quantitation of base release activities of TDG in the initial phase of the reaction (A). Data

are presented as means ± SD from three independent experiments.

(C) Quantitation of base release activities of TDG with BSA or Gadd45a under single-turnover

condition. G•5caC-DNA (25 nM) was used as substrate and incubated with 250 nM TDG and 1

μM Gadd45a or BSA.

(D) shows the activity of TDG in the initial phase of (C).

Supplementary Materials and Methods

Tet-assisted bisulfite sequencing (TAB-seq)

Genomic DNA extracted from mouse ES cells were treated as described (1). Briefly, 200 ng of

genomic DNA were sonicated into fragments in average size of 500 bp and then glucosylated

with T4 Phage β-glucosyltransferase (T4-BGT) (NEB) according to the manufacturer's

instructions. The glucosylated DNA was recovered by phenol-chloroform extraction and

oxidized with recombinant TET2 protein twice, followed by bisulfite treatment. PCR primers

were the same as those used in standard bisulfite sequencing.

Supplemental References

1.	Yu, M., Hon, G.C., Szulwach, K.E., Song, C.X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B.
	et al. (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome.
	Cell, **149**, 1368-1380.

Table S1. List of oligonucleotides and PCR primers used in the this study.

TDG genotyping PCR primers (5'-3'):

hTDG-exon2-F2	GGCTGACTTGACAGGACTGA
hTDG-exon2-R2	GGCTGGAACTTCTTCTGGCA
hTDG geno.puro-F	CGACCACCAGGGCAAGG
hTDG geno.hygro-F	GTCCGAGGGCAAAGGAATGA
hTDG geno.R	TACCCAACAGGAAGCCCTGA

Gadd45 genotyping PCR primers (5'-3'):

G45a-geno-F	GAGGCAGCAGTGCAGAGTTC
G45a-geno-R	CAGAGCCACATCCCGGTCGTC
G45b-geno-F	TTGGGAAGGGAAGTCCCACCG
G45b-geno-R	CCGAGAGCGCTGGGAAAGTCC

Gadd45 sgRNA oligos used for preparing sgRNA expression constructs (5'-3'):

G45a-sgRNA1-F	CACCGTATGACTTTGGAGGAATTCT
G45a-sgRNA1-R	AAACAGAATTCCTCCAAAGTCATAC
G45b-sgRNA1-F	CACCGGCTGGTGGCGAGCGACAACG
G45b-sgRNA1-R	AAACCGTTGTCGCTCGCCACCAGCC

Bisulfite primers (5'-3'):

Plagl1-BS-F	ATTTGTTATTTAGTTTGGGTTGG
Plagl1-BS-R	ACCCAAATTCAAAATTTATCAC
Cilp2-BS-F	GTTAGGGTGGTGGAAGGTTT
Cilp2-BS-R	ACTTCACCTCCTACCAAAA
CMV-BS-F	GAATTTTTGTAGGATTAGTGGATTT
CMV-BS-R	CACTTAATATACTACCAAATAAACAA

Table S2. Luciferase assay in 293T cells expressing ectopic candidate genes in the presence of Tet2 or Tet2 and TDG. Renilla luciferase activity was used for the normalization of firefly luciferase activity. Luciferase signals were normalized to the signal (arbitrarily set at 1) from cells transfected with empty vector and methylated reporter gene.

	Tet2	Tet2 + TDG
empty vector	2.9	10.5
AID	5.3	8.8
Gadd45a	4.0	30.6
Gadd45b	6.6	50.7
Gadd45g	7.6	25.9
MBD4	6.1	23.6
RNF4	11.4	75.2
Elp3	3.4	7.1
MBD2	3.2	8.2

Table S3. Hypermethylated loci identified in *Gadd45a/b* DKO ES cells.

#Chr	CpG site start at	CpG site end at	No. of hypermethylated CpG sites (methylation level DKO-WT >= 25%)
chr10	12810629	12811034	25
chr17	29989346	29989629	12
chr8	124497875	124498055	11
chr7	114353723	114354069	10
chr11	11926195	11926386	9
chr11	116464994	116465161	9
chr2	156846689	156846869	9
chr1	63247019	63247224	8
chr10	28916824	28917000	8
chr17	39985424	39985567	8
chr6	47974857	47974999	8
chr10	94778414	94778594	7
chrX	160374376	160374599	7
chr1	173939752	173939942	6
chr12	11444962	11445130	6
chr13	96374092	96374300	6
chr14	52360659	52360837	6
chr14	55262846	55263070	6
chr14	80061327	80061490	6
chr19	61302514	61302645	6
chr6	117821502	117821665	6
chr7	30578032	30578263	6
chr7	117146779	117146919	6
chr8	29640647	29640801	6
chr1	108016012	108016241	5
chr15	51859452	51859578	5
chr18	61225434	61225598	5
chr2	156846988	156847212	5
chr2	165611147	165611298	5
chr2	174110204	174110339	5
chr5	120904875	120905074	5
chr5	143006075	143006216	5
chr7	4959731	4959983	5
chr7	19478584	19478731	5
chr7	50931969	50932121	5
chr7	52632486	52632699	5
chr8	72406508	72406672	5
chr8	86192449	86192618	5
chr9	43915985	43916108	5
chr10	10675791	10675963	4
chr11	11926800	11926959	4
chr11	96573894	96574066	4
chr11	114333878	114334019	4
chr15	79447772	79447944	4
chr15	102009414	102009586	4
chr17	32497199	32497327	4
chr17	48088489	48088609	4
chr17	88349236	88349388	4
chr18	20760061	20760222	4
chr18	86377421	86377559	4

chr2	154162024	154162158	4
chr3	90077809	90077940	4
chr4	145427159	145427298	4
chr5	110497981	110498132	4
chr6	100432791	100432929	4
chr7	4960086	4960227	4
chr7	80664446	80664603	4
chr7	125633325	125633465	4
chr7	134379244	134379386	4
chr8	3994244	3994402	4
chr8	74225409	74225552	4
chr8	87396661	87396811	4
chr8	97852302	97852476	4
chr8	107898587	107898755	4
chr9	43916469	43916643	4
chr9	57606552	57606685	4
chrX	91225968	91226103	4
chrX	133464204	133464373	4