Published in final edited form as:

J Mol Biol. 2014 November 11; 426(22): 3703–3712. doi:10.1016/j.jmb.2014.09.010.

The carboxy-terminal domain of ROS1 is essential for 5-methylcytosine DNA glycosylase activity

Samuel Hong^{1,2}, Hideharu Hashimoto¹, Yoke Wah Kow³, Xing Zhang¹, and Xiaodong Cheng^{1,*}

Samuel Hong: shong27@emory.edu; Hideharu Hashimoto: hhashi3@emory.edu; Yoke Wah Kow: ykow@emory.edu; Xing Zhang: xzhan02@emory.edu; Xiaodong Cheng: xcheng@emory.edu

¹Department of Biochemistry, Emory University School of Medicine, 1510 Clifton Road, Atlanta, GA 30322

²Molecular and Systems Pharmacology graduate program, Emory University School of Medicine, 1510 Clifton Road, Atlanta, GA 30322

³Department of Radiation Oncology, Emory University School of Medicine, 1510 Clifton Road, Atlanta, GA 30322

Abstract

Arabidopsis thaliana Repressor of silencing 1 (ROS1) is a multi-domain bifunctional DNA glycosylase/lyase, which excises 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) as well as thymine and 5-hydroxymethyluracil (i.e., the deamination products of 5mC and 5hmC) when paired with a guanine, leaving an apyrimidinic (AP) site that is subsequently incised by the lyase activity. ROS1 is slow in base excision and fast in AP lyase activity, indicating that the recognition of pyrimidine modifications might be a rate-limiting step. In the C-terminal half, the enzyme harbors a Helix-hairpin-Helix DNA glycosylase domain followed by a unique C-terminal domain. We show that the isolated glycosylase domain is inactive for base excision, but retains partial AP lyase activity. Addition of the C-terminal domain restores the base excision activity and increases the AP lyase activity as well. Furthermore, the two domains remain tightly associated and can be co-purified by chromatography. We suggest that the C-terminal domain of ROS1 is indispensable for the 5mC DNA glycosylase activity of ROS1.

Keywords

5-methylcytosine; DNA 5mC glycosylase; epigenetic regulation; DNA demethylation; Repressor of silencing

© 2014 Elsevier Ltd. All rights reserved.

*To whom correspondence should be addressed: Tel: +1 404 727 8491; Fax: +1 404 727 3746; xcheng@emory.edu.

Author Contributions SH performed all experiments; HH provided purified TDG and MBD4 enzymes; YWK assisted in data analysis; XZ and XC organized and designed the scope of the study; and all were involved in analyzing data and preparing the manuscript.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

In eukaryotic genomes, DNA methyltransferases convert a proportion of cytosines into 5-methylcytosine (5mC)¹. Mammalian ten-eleven-translocation (Tet) dioxygenases then convert a fraction of these to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) in consecutive oxidation reactions^{2,3,4}. Mammalian TDG, named after thymine DNA glycosylase, excises the mismatched base from G:X mismatches, where X is uracil, thymine or 5-hydroxymethyluracil (5hmU). These are, respectively, the deamination products of cytosine, 5mC and 5hmC. In addition, TDG excises the Tet enzyme products 5fC and 5caC but not 5mC and 5hmC, when paired with a guanine^{4,5,6,7}. The resulting apurinic/apyrimidinic (AP) site is enzymatically converted to normal cytosine through the base excision repair pathway, altering DNA methylation patterns utilized for epigenetic control. Mammalian DNA glycosylase(s) that excise 5mC or 5hmC have not been identified but such activities have been reported^{8,9,10,11}.

In *Arabidopsis thaliana*, a family of 5mC DNA glycosylases has been identified: REPRESSOR OF SILENCING 1 (ROS1)¹², DEMETER (DME)¹³, DME-like 2 (DML2) and DME-like 3 (DML3)¹⁴. ROS1 is a 1393-residue, multi-domain protein (Fig. 1a): the N-terminal domain containing a lysine-rich stretch involved in non-specific DNA binding and sliding along DNA^{15,16}, followed by the central Helix-hairpin-Helix (HhH) DNA glycosylase domain containing an iron-sulfur (4Fe-4S) cluster¹², and a unique uncharacterized domain at the C-terminus. The central glycosylase domain (GD) has an atypical insertion of ~230 residues – whose sequence and length vary among the ROS1 family members – which is not found in other characterized HhH DNA glycosylases¹⁷. Like ROS1, mammalian Tet proteins have an atypical insertion into their catalytic domains, and the insertion is not required for the *in vitro* catalytic activity¹⁸. ROS1, and its family members, is a bifunctional DNA glycosylase/lyase whose glycosylase activity excises a 5mC base from the DNA backbone and then its lyase activity cleaves the DNA backbone at the AP site^{19,20,21}.

The amino acids sequences within the C-terminal domain (CTD) are conserved among the ROS1 family members, but no homologous sequence has been found in other phyla. Introduction of random point mutations or deletions in the corresponding domain in DME resulted in abrogation of the 5mC excision activity²². Here we show that the isolated glycosylase domain of ROS1 does not possess the 5mC excision activity, but retains partially the AP lyase activity. The addition of the CTD restores the 5mC excision activity. The two domains remain tightly associated and could be co-purified by chromatography.

Results

ROS1 glycosylase domain and the C-terminal domain

First, we constructed a deletion variant of ROS1, deleting the N-terminal 509 residues and replacing the internal insertion (residues 628-855) with a 5-residue linker, which we refer to as ROS1^N (Fig. 1a). We measured the base excision and the AP lyase activities of the purified ROS1 full-length (FL), ROS1^N and its catalytic mutant D971N (see below), using various 32-base pair (bp) DNA oligonucleotides (oligos), each containing a single variable

base opposite a guanine (G:X pair), where X is C, 5mC, 5hmC, 5fC, 5caC, U, T, 5hmU, or AP. These substrates bear either the 'natural' base pairs, mismatches of the deamination products, or the product of base excision. Both FL and N deletion excised 5mC and 5hmC but not C, 5fC or 5caC (Fig. 1b–c), and 5hmC excision was weaker (by a factor of ~1.6) than 5mC excision for both ROS1 and ROS1 N (Fig. 1b–c). The *in vitro* excision activity on 5hmC has recently been reported for ROS1 and its family members ($k_{\text{cat}} = 0.3\text{--}1\text{ h}^{-1}$ under single turnover conditions)^{23,24}. However, the significance of this activity is unclear, because no homologs of Tet dioxygenases have been identified in *Arabidopsis* and data on the existence of 5hmC in *Arabidopsis* are conflicting: one study detected no 5hmC²³, whereas another study found low levels of 5hmC in the DNA of leaves and flowers²⁵.

In addition, ROS1 N is also active on G:T and G:5hmU mismatches, but no activity was observed on G:U mismatch (Fig. 1d). The activity on G:T is comparable with that on G:5mC. This observation indicates that ROS1 is sensitive to pyrimidine modifications at the ring-5 position.

In the structurally characterized HhH DNA glycosylases, a conserved aspartate, Asp138 of *Escherichia coli* endonuclease III (Endo III)²⁶, Asp138 of *E. coli* MutY²⁷, Asp238 of *E. coli* AlkA²⁸, Asp268 of human OGG1²⁹, and Asp534 of mouse MBD4^{30,31}, has been suggested to activate a catalytic nucleophile (such as a water molecule or a nearby lysine residue) for the attack on the deoxyribose C1' carbon atom of the target nucleotide. The equivalent residue in ROS1 is Asp971¹⁷, and the mutation of Asp971 to asparagine (D971N) abolished the base excision activity but not the AP lyase activity (Fig. 1b). One interesting observation is that the AP lyase activity of ROS1 is substantially faster than the base excision activity (comparing Fig. 1e to Fig. 1c–d). Both ROS1 FL and ROS1 N showed ~90% cleavage of AP sites in 15 min compared to ~80% excision of 5mC over 20 h under the same conditions. ROS1 is known for slow turnover kinetics²¹, and our observation of the fast AP lyase activity of ROS1 suggests that an initial stage of 5mC excision reaction, or probably the recognition of pyrimidine modifications, is a rate-limiting step.

ROS1 glycosylase domain and the C-terminal domain associate tightly

Most structurally characterized HhH DNA glycosylases, like EndoIII²⁶, hOGG1²⁹, AlkA²⁸, and MBD4^{30,31} exist as or have an isolated glycosylase domain active on its own *in vitro*. We asked whether ROS1 glycosylase domain could function on its own and thus purified the isolated glycosylase domain (GD) and the C-terminal domain (CTD) individually. We note that the isolated domains, particularly the CTD, were somewhat problematic during expression and/or purification with low yield, more impurity, and tendency to aggregate (see Materials and Methods). Nevertheless, as shown in Figure 2a, the isolated glycosylase domain is inactive on 5mC and 5hmC excisions, but retains residual AP lyase activity (lane 2). However, addition of CTD (with estimated 3:1 molar ratio of CTD:GD) restored partial activity of base excision on 5mC and 5hmC and increased the AP lyase activity as well (Fig. 2a, lane 4).

We reasoned that, in order to restore the base excision activity of ROS1, the C-terminal domain must interact with the glycosylase domain, either directly or through DNA. To test this notion, and to overcome the problems of the isolated GD and CTD domains, we

engineered a new construct, termed as ROS1 N:P, in which the PreScission protease recognition sequence (LEVLFQG/P) was inserted in the linker between GD and CTD (Fig. 3a). The 8-residue insertion did not affect the 5mC and 5hmC excision activities (Fig. 3b, lanes 2–3) and AP lyase activity (Fig. 2a, bottom panel). Analytical size exclusion chromatography measurements revealed that the two cleaved fragments of ROS1 associated together in the presence of 500 mM NaCl (Fig. 3c, 3d, and 3e), suggesting the interactions between the two domains are hydrophobic, a plausible reason that the isolated domains tend to aggregate in aqueous solution. Approximately the same base excision and AP lyase activities were observed with and without the protease cleavage (Fig. 3b, lanes 3–4 and Fig. 2a, bottom panel).

Mouse MYH does not possess 5mC DNA glycosylase activity

We were intrigued by the observation that adding ROS1 CTD could restore the base excision activity by ROS1 GD. We asked whether ROS1 CTD could also allow other glycosylases to be active on 5mC and 5hmC for four reasons: (1) ROS1 shares some common substrates (such as 5-hydroxyuracil) with several DNA glycosylases (Nth1, Neil1)²¹, (2) some of them are oxidized pyrimidine-specific DNA glycosylases that have been characterized in mammalian cells (Nth1, Neil1/2, TDG)³², (3) several mouse DNA glycosylases (Neil1/2, Nth1 and Ogg1) were identified to bind 5mC- or 5hmC-containing oligos in a DNA pull-down experiment combined with quantitative mass spectrometry³³, and (4) mouse MutY homolog (mMYH) has recently been suggested to possess 5mC DNA glycosylase activity³⁴. However, none of the mammalian enzymes we examined, including mMYH (Fig. 2d–e), showed 5mC or 5hmC DNA glycosylase activity with and without the addition of ROS1 CTD, whereas they were active on their respective substrates (Fig. 2b–d). Furthermore, the addition of ROS1 CTD had no effect on the activities of mMBD4 and mMYH on their cognate substrates (bottom panels of Fig. 2c–d). We were unable to observe the suggested 5mC activity for mMYH using either the 32-bp oligos (Fig. 2d) or the 71-bp sequence from the mouse IL-2 promoter used by Wu and Zheng³⁴ (Fig. 2e).

Among the HhH DNA glycosylases, mammalian MYH (MutY homolog) shares a similar domain organization as that of ROS1 N (Fig. 3f). MutY cleaves the adenine opposite 8-oxoguanine (oxoG), which arises from unrepaired oxoG after DNA replication. In the structure of *bacillus stearothermophilus* MutY in complex with DNA, the C-terminal domain recognizes 8oxoG and the opposite Ade flips out into the active site of the glycosylase domain where the excision occurs³⁵. Mouse MYH (mMYH) is also known to excise Ade opposite Gua with comparable efficiencies as that of A:oxoG³⁶. For comparison with ROS1, we generated the analogous mMYH:P construct, in which the PreScission protease recognition sequence was inserted in the linker connecting the mMYH glycosylase domain and its C-terminal domain (Fig. 3f). Unlike ROS1, the protease cleaved mMYH fragments eluted as two separate and delayed peaks in the size exclusion chromatography (Fig. 3h), clearly showing that the two domains dissociated after cleavage (Fig. 3i–3j). The separated glycosylase domain of mMYH has much reduced activity on adenine excision (Fig. 3g, lane 4), similar to *E. coli* MutY where the glycosylase domain alone has reduced activity^{27,37}.

We attempted to test whether the ROS1 C-terminal domain could allow mMYH glycosylase domain to be active on 5mC by generating a hybrid enzyme (Fig. 3k). The fusion enzyme has reduced activity on G:A mismatch (Fig. 3k, lane 6), similar to that of cleaved mMYH glycosylase domain (Fig. 3g, lane 4), but is not active on a G:5mC pair (Fig. 3k, lane 3).

Discussion

Extensive studies on DNA repair glycosylase enzymes, such as human uracil DNA glycosylase (hUNG) and 8-oxoguanine DNA glycosylases (hOGG and bacterial MutM), showed that they recognize damaged bases through a multi-step interrogation process [reviewed in ^{38,39}]. Allowing only a true substrate to reach the active site, these enzymes distort DNA by bending it followed by intrahelical interrogation to detect a lesion, by flipping of potential substrate nucleotides to varying degrees and rejecting non-substrate nucleotide back to DNA helices. The two-domain structure of MutY in complex with DNA containing an 8oxoG:A mismatch³⁵ revealed that the C-terminal domain contributes specific contacts to the intra-helical stacked 8oxoG lesion that are functionally important for lesion recognition and thus enzymatic excision of extra-helical adenine opposite 8oxoG by the catalytic glycosylase domain. In other words, the two domains of MutY are primarily responsible, respectively, for essential interaction with the bases on opposite DNA strands; as changing oxoG:A to C:A significantly reduces the activity of MYH from calf thymus⁴⁰. In the case of ROS1, the two domains, the glycosylase domain and the C-terminal domain, strongly associate with each other and seem to be insensitive to the base identity paired with the modified cytosine, as ROS1 N is active on all four pairs of 5mC:X or 5hmC:X (X=G, A, T, or C) (Fig. 3l–3m). Yet, somehow the target 5mC or 5hmC must be recognized intrahelically, flipped out and delivered to the active site of the glycosylase domain to allow excision. The precise way in which the two interacting domains of ROS1 mediate specific DNA recognition and excision awaits the solution of a protein-DNA complex structure.

One possibility is that the CTD stabilizes the glycosylase domain and stimulates its intrinsic excision and lyase activities. A precedent is mammalian DNA methyltransferase 3a (Dnmt3a) and its interaction with Dnmt3-like protein (Dnmt3L). Dnmt3a has low activity on its own but forms oligomers. Interaction with Dnmt3L disrupts the Dnmt3a oligomer, forming a Dnm3L-3a tetramer via the catalytic domain of Dnmt3a, and stimulates the Dnmt3a activity^{41,42,43,44}. Previous published work by Ariza and colleagues showed, based on structurally homology modeling and site-directed mutagenesis, that the ROS1 glycosylase domain interacts with both strands of DNA^{17,45}. It was suggested that residues Phe589 and Tyr1028 in the glycosylase domain are involved in recognition of flipped-out 5mC in the cleavage center¹⁷ and residues Arg903 and Met905 interact with the orphan guanine in the complementary strand⁴⁵. However, these suggested interactions would be post-base flipping, and would not account for the step(s) of the intrahelical modification-interrogation that precedes specific extrahelical base recognition. It is conceivable that CTD could have nonspecific DNA binding activity and thus stimulates the modification interrogation process by the glycosylase domain. Alternatively, we speculate that the CTD is a novel DNA substrate recognition domain responsive to pyrimidine modifications at the ring-C5 position. The CTD might function in the early steps of intrahelical interrogation to

detect the C5 modification and in initiation of base flipping that precedes specific binding in the active site of the glycosylase domain.

Arabidopsis ROS1 and mammalian TDG are the two DNA glycosylases currently implicated in so-called DNA demethylation pathways by removing modified cytosine base⁴⁶: ROS1 excises 5mC and 5hmC but not 5fC and 5caC, whereas TDG removes 5fC and 5caC but not 5mC and 5hmC^{5,7}. It is worthy to note that ROS1 is inactive on G:U^{20,21}, whereas TDG, even named after thymine DNA glycosylase, has much faster activity on G:U mismatch^{47,48}. The four chemically modified forms of cytosine might not be genetically equivalent in terms of base pairing. A strong intramolecular hydrogen bond has been observed between the exocyclic *N*4 amino group (NH₂) and the carbonyl oxygen (O=C) at ring carbon-5 position of 5fC, in the free nucleoside form^{49,50}, and the carboxyl moiety (COO⁻) of 5caC in the protein bound form⁵¹. It was hypothesized that the existence of such an intra-base hydrogen bond would shift the amino-imino equilibrium^{48,52,53}, which would enable 5fC and 5caC to form two, instead of three, hydrogen bonds with an opposite guanine, equivalent to a G:T or G:U 'wobble' pair. Previously observed mutagenic potential of 5fC and 5caC *in vivo* and *in vitro*^{49,52,53,54,55} suggested the possible existence of the imino tautomeric form. TDG might take advantage of the tendency of G:5fC and G:5caC to form a mismatch-like wobble hydrogen bonding pattern and turn them into substrates, whereas ROS1 is insensitive to mismatches.

Besides ROS1, which recognizes and excises 5mC from the 'natural' G:5mC base pair, another enzyme, PabI in *Pyrococcus abyssi*, initially identified as a restriction enzyme, actually is a sequence-specific adenine DNA glycosylase⁵⁶. The dimeric PabI recognizes a palindromic 5'-GTAC-3', hydrolyses the *N*-glycosidic bond between the adenine base and the sugar and produces two opposing AP sites, which are subsequently cleaved by AP endonucleases to introduce a double-strand break. Thus, not every DNA glycosylase is involved in DNA repair, and some may actually generate damage.

Materials and Methods

Protein Expression and Purification

The full-length (FL) ROS1 (pXC1135) was expressed in *Escherichia coli* dcm⁻ BL21-CodonPlus(DE3)-RIL (Stratagene) as a 6xHis fusion in a pET28a vector (Novagen). ROS1 N (pXC1256), ROS1 N:D971N (pXC1273), ROS1 N:P (pXC1327), ROS1:GD (pXC1278), ROS1:CTD (pXC1297), mMYH (pXC1321), mMYH:P (pXC1332), mMYH-ROS1 hybrid (pXC1338) were ligated into a modified pET28b vector (Novagen) as a N-terminal 6xHis-SUMO fusion and expressed in *Escherichia coli* dcm⁻ BL21-CodonPlus(DE3)-RIL (Stratagene). Bacterial cells were cultured in Lysogeny Broth (LB) at 37 °C, and protein expression was induced at 16 °C overnight or at 23°C for 2 h (ROS1:FL). Cells were harvested and stored in -80 °C. Cell pellet was thawed and lysed by sonication in lysis buffer (20 mM Tris-HCl pH 7.5, 500 mM NaCl, 5% glycerol and 1 mM dithiothreitol). Lysate was clarified by centrifugation at 18,000 rpm for 1 h. The fusion protein was isolated on a Nickel-charged HisTrap affinity column (GE Healthcare).

For ROS1:FL, eluted fractions from the nickel column were further purified on tandem HiTrap Q and HiTrap SP ion exchange columns (GE Healthcare). For ROS1 N, ROS1 N:D971N, ROS1 N:P, and mMYH-ROS1 hybrid, eluted fractions from the nickel column were purified on Heparin affinity column (GE Healthcare) followed by cleavage of the 6xHis-SUMO tag via yeast ubiquitin-like-specific protease 1 (Ulp1; purified in-house), and then purified on a HiTrap Q column. For ROS1 N, 6xHis-SUMO tag-cleaved sample was passed through Glutathione Sepharose 4B (GE Healthcare) pre-bound with GST-tagged Uracil glycosylase inhibitor protein (GST-UGI; purified in-house)⁵⁷ to remove any residual *E. coli* Uracil DNA glycosylase activity. ROS1 N, ROS1 N:D971N, and ROS1 N:P were further purified by Superdex 200 16/60 size exclusion column (GE Healthcare).

For ROS1:GD, ROS1:CTD, mMYH, and mMYH:P, eluted fractions from the nickel column were cleaved of their 6xHis-SUMO tag and further purified on an Heparin column. For ROS1:GD, mMYH, and mMYH:P, eluted fractions from the Heparin column were further purified on Superdex 200 (16/60 or 10/300 GL) size exclusion column (GE healthcare). Final protein concentrations were estimated by absorbance at 280 nm for ROS1 N (absorbance coefficient $\epsilon=1.084$), ROS1 N:P ($\epsilon=1.071$), ROS1:GD ($\epsilon=1.450$), mMYH ($\epsilon=1.472$), mMYH:P ($\epsilon=1.456$) or, for ROS1:FL and CTD, by Coomassie Blue staining using Bovine Serum Albumin as a standard. Compared with ROS1 N, the isolated ROS1 GD domain had lower yield and a broader peak on size exclusion column with some of the fractions overlapping with a major *E. coli* contaminant. The CTD domains had lower yield and higher impurity. In addition, the CTD was aggregated as it eluted in void volume in size exclusion chromatography (data now shown).

DNA glycosylase activity assay

Activities of ROS1 and its variants, and other DNA glycosylases, on various DNA oligos labeled with 6-carboxy-fluorescein (FAM) were performed in reaction buffer: 50 mM Tris-HCl pH 8.0, 1 mM EDTA, 1 mM DTT, and 0.1% bovine serum albumin) for the indicated time in room temperature (~23 °C or otherwise indicated). For reactions with ROS1 N:P and mMYH:P after being cleaved by the Prescission protease (purified in-house), the protease ($0.1 \mu\text{g } \mu\text{l}^{-1}$) were present in the reaction mixture. Reactions were stopped by adding 2 μl Proteinase K (1 mg ml^{-1}) and incubating at 50 °C for 15 min. For substrates except G:AP, reactions were stopped by adding 0.1 N NaOH and incubating at 95°C on a heat block for 10 min. An aliquot (20 μl) of loading buffer (98% formamide, 1 mM EDTA, and trace amount of bromophenol blue and xylene cyanole) was added, and samples were heated in 95°C on a heat block for 10 min. Samples were then immediately transferred to ice water to cool and loaded on a $10 \times 10 \text{ cm}^2$ denaturing PAGE gel containing 15% acrylamide, 7 M urea and 24% formamide in 1x Tris-Borate-EDTA (TBE). For G:AP substrates, reactions were stopped by adding 20 μl of loading buffer and samples were loaded on the gel without heating. The gel was run in 1x TBE at 200 V for 75 min. Typhoon Trio+ (GE Healthcare) was used to visualize the intensities from FAM-labeled DNA. The image-processing program ImageJ was used to quantify the intensities and data points were fit to a curve using Prism 6.0 (GraphPad).

Various 32-bp oligos labeled with FAM (synthesized by New England Biolabs) were used as substrates: (FAM)-5'-TCG GAT GTT GTG GGT CAG XGC ATG ATA GTG TA-3' (where X=C, 5mC, 5hmC, 5fC, 5caC, U, T, or 5hmU) and its complementary strand 5'-TAC ACT ATC ATG CYC TGA CCC ACA ACA TCC GA-3' (where Y=G, A, T, or C). Oligo containing G:AP was generated by incubating G:U oligo with 1 Unit of *E. coli* uracil DNA glycosylase (UDG; New England Biolabs, catalog #M0280) for 30 min in room temperature (~23 °C). In addition, 71-bp oligos (synthesized by Sigma) were used for testing mMYH activity: (FAM)-5'-CAT GAG TTA CTT TTG TGT CTC CAC CCC AAA GAG GAA AAT TTG TTT CAT ACA GAA GGX GTT CAT TGT ATG AA-3' (where X=A or 5mC) and its complementary strand 5'-TTC ATA CAA TGA ACG CCT TCT GTA TGA AAC AAA TTT TCC TCT TTG GGG TGG AGA CAC AAA AGT AAC TCA TG-3'.

Acknowledgments

We thank J.K. Zhu (Purdue University) for providing the ROS1 FL construct, J.I. Cohen (NIH) for the GST-UGI construct, S. Mitra (University of Texas Medical Branch at Galveston) for purified hNth1, hNeil1, and hNeil2, and J.R. Horton for comments. National Institutes of Health grant GM049245-21 supported this work to X.C. (who is a Georgia Research Alliance Eminent Scholar).

Abbreviations used

ROS1	Repressor of silencing 1
HhH	Helix-hairpin-Helix
TDG	thymine DNA glycosylase
5mC	5-methylcytosine
5hmC	5-hydroxymethylcytosine
5fC	5-formylcytosine
5caC	5-carboxylcytosine

References

1. Goll MG, Bestor TH. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem.* 2005; 74:481–514. [PubMed: 15952895]
2. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science.* 2009; 324:930–935. [PubMed: 19372391]
3. Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science.* 2011; 333:1300–1303. [PubMed: 21778364]
4. He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science.* 2011; 333:1303–1307. [PubMed: 21817016]
5. Maiti A, Drohat AC. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem.* 2011; 286:35334–35338. [PubMed: 21862836]
6. Zhang L, Lu X, Lu J, Liang H, Dai Q, Xu GL, et al. Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nat Chem Biol.* 2012; 8:328–330. [PubMed: 22327402]

7. Hashimoto H, Hong S, Bhagwat AS, Zhang X, Cheng X. Excision of 5-hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase domain: its structural basis and implications for active DNA demethylation. *Nucleic Acids Res.* 2012; 40:10203–10214. [PubMed: 22962365]
8. Vairapandi M, Duker NJ. Enzymic removal of 5-methylcytosine from DNA by a human DNA-glycosylase. *Nucleic Acids Res.* 1993; 21:5323–5327. [PubMed: 8265344]
9. Vairapandi M, Liebermann DA, Hoffman B, Duker NJ. Human DNA-demethylating activity: a glycosylase associated with RNA and PCNA. *J Cell Biochem.* 2000; 79:249–260. [PubMed: 10967552]
10. Vairapandi M. Characterization of DNA demethylation in normal and cancerous cell lines and the regulatory role of cell cycle proteins in human DNA demethylase activity. *J Cell Biochem.* 2004; 91:572–583. [PubMed: 14755686]
11. Cannon SV, Cummings A, Teebor GW. 5-Hydroxymethylcytosine DNA glycosylase activity in mammalian tissue. *Biochem Biophys Res Commun.* 1988; 151:1173–1179. [PubMed: 3355548]
12. Gong Z, Morales-Ruiz T, Ariza RR, Roldan-Arjona T, David L, Zhu JK. ROS1, a repressor of transcriptional gene silencing in Arabidopsis, encodes a DNA glycosylase/lyase. *Cell.* 2002; 111:803–814. [PubMed: 12526807]
13. Gehring M, Huh JH, Hsieh TF, Penterman J, Choi Y, Harada JJ, et al. DEMETER DNA glycosylase establishes MEDEA polycomb gene self-imprinting by allele-specific demethylation. *Cell.* 2006; 124:495–506. [PubMed: 16469697]
14. Ortega-Galisteo AP, Morales-Ruiz T, Ariza RR, Roldan-Arjona T. Arabidopsis DEMETER-LIKE proteins DML2 and DML3 are required for appropriate distribution of DNA methylation marks. *Plant Mol Biol.* 2008; 67:671–781. [PubMed: 18493721]
15. Ponferrada-Marin MI, Martinez-Macias MI, Morales-Ruiz T, Roldan-Arjona T, Ariza RR. Methylation-independent DNA binding modulates specificity of Repressor of Silencing 1 (ROS1) and facilitates demethylation in long substrates. *J Biol Chem.* 2010; 285:23032–23039. [PubMed: 20489198]
16. Ponferrada-Marin MI, Roldan-Arjona T, Ariza RR. Demethylation initiated by ROS1 glycosylase involves random sliding along DNA. *Nucleic Acids Res.* 2012; 40:11554–11562. [PubMed: 23034804]
17. Ponferrada-Marin MI, Parrilla-Doblas JT, Roldan-Arjona T, Ariza RR. A discontinuous DNA glycosylase domain in a family of enzymes that excise 5-methylcytosine. *Nucleic Acids Res.* 2011; 39:1473–1484. [PubMed: 21036872]
18. Hashimoto H, Pais JE, Zhang X, Saleh L, Fu ZQ, Dai N, et al. Structure of a Naegleria Tet-like dioxygenase in complex with 5-methylcytosine DNA. *Nature.* 2014; 506:391–395. [PubMed: 24390346]
19. Agius F, Kapoor A, Zhu JK. Role of the Arabidopsis DNA glycosylase/lyase ROS1 in active DNA demethylation. *Proc Natl Acad Sci U S A.* 2006; 103:11796–11801. [PubMed: 16864782]
20. Morales-Ruiz T, Ortega-Galisteo AP, Ponferrada-Marin MI, Martinez-Macias MI, Ariza RR, Roldan-Arjona T. DEMETER and REPRESSOR OF SILENCING 1 encode 5-methylcytosine DNA glycosylases. *Proc Natl Acad Sci U S A.* 2006; 103:6853–6858. [PubMed: 16624880]
21. Ponferrada-Marin MI, Roldan-Arjona T, Ariza RR. ROS1 5-methylcytosine DNA glycosylase is a slow-turnover catalyst that initiates DNA demethylation in a distributive fashion. *Nucleic Acids Res.* 2009; 37:4264–4274. [PubMed: 19443451]
22. Mok YG, Uzawa R, Lee J, Weiner GM, Eichman BF, Fischer RL, Huh JH. Domain structure of the DEMETER 5-methylcytosine DNA glycosylase. *Proc Natl Acad Sci U S A.* 2010; 107:19225–19230. [PubMed: 20974931]
23. Jang H, Shin H, Eichman BF, Huh JH. Excision of 5-hydroxymethylcytosine by DEMETER family DNA glycosylases. *Biochem Biophys Res Commun.* 2014; 446:1067–1072. [PubMed: 24661881]
24. Brooks SC, Fischer RL, Huh JH, Eichman BF. 5-methylcytosine recognition by Arabidopsis thaliana DNA glycosylases DEMETER and DML3. *Biochemistry.* 2014; 53:2525–2532. [PubMed: 24678721]
25. Yao Q, Song CX, He C, Kumaran D, Dunn JJ. Heterologous expression and purification of Arabidopsis thaliana VIM1 protein: in vitro evidence for its inability to recognize

- hydroxymethylcytosine, a rare base in Arabidopsis DNA. *Protein Expr Purif.* 2012; 83:104–111. [PubMed: 22459921]
26. Thayer MM, Ahern H, Xing D, Cunningham RP, Tainer JA. Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. *Embo J.* 1995; 14:4108–4120. [PubMed: 7664751]
 27. Guan Y, Manuel RC, Arvai AS, Parikh SS, Mol CD, Miller JH, et al. MutY catalytic core, mutant and bound adenine structures define specificity for DNA repair enzyme superfamily. *Nat Struct Biol.* 1998; 5:1058–1064. [PubMed: 9846876]
 28. Hollis T, Ichikawa Y, Ellenberger T. DNA bending and a flip-out mechanism for base excision by the helix-hairpin-helix DNA glycosylase, Escherichia coli AlkA. *Embo J.* 2000; 19:758–766. [PubMed: 10675345]
 29. Bruner SD, Norman DP, Verdine GL. Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. *Nature.* 2000; 403:859–866. [PubMed: 10706276]
 30. Wu P, Qiu C, Sohail A, Zhang X, Bhagwat AS, Cheng X. Mismatch repair in methylated DNA. Structure and activity of the mismatch-specific thymine glycosylase domain of methyl-CpG-binding protein MBD4. *J Biol Chem.* 2003; 278:5285–5291. [PubMed: 12456671]
 31. Hashimoto H, Zhang X, Cheng X. Excision of thymine and 5-hydroxymethyluracil by the MBD4 DNA glycosylase domain: structural basis and implications for active DNA demethylation. *Nucleic Acids Res.* 2012; 40:8276–8284. [PubMed: 22740654]
 32. Brooks SC, Adhikary S, Robinson EH, Eichman BF. Recent advances in the structural mechanisms of DNA glycosylases. *Biochim Biophys Acta.* 2013; 1834:247–271. [PubMed: 23076011]
 33. Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T, Jansen PW, Bauer C, et al. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell.* 2013; 152:1146–1159. [PubMed: 23434322]
 34. Wu L, Zheng Q. Active demethylation of the IL-2 Promoter in CD4(+) T cells is mediated by an inducible DNA glycosylase, Myh. *Mol Immunol.* 2014; 58:38–49. [PubMed: 24291244]
 35. Fromme JC, Banerjee A, Huang SJ, Verdine GL. Structural basis for removal of adenine mispaired with 8-oxoguanine by MutY adenine DNA glycosylase. *Nature.* 2004; 427:652–656. [PubMed: 14961129]
 36. Ushijima Y, Tominaga Y, Miura T, Tsuchimoto D, Sakumi K, Nakabeppu Y. A functional analysis of the DNA glycosylase activity of mouse MUTYH protein excising 2-hydroxyadenine opposite guanine in DNA. *Nucleic Acids Res.* 2005; 33:672–682. [PubMed: 15681617]
 37. Chmiel NH, Golinelli MP, Francis AW, David SS. Efficient recognition of substrates and substrate analogs by the adenine glycosylase MutY requires the C-terminal domain. *Nucleic Acids Res.* 2001; 29:553–564. [PubMed: 11139626]
 38. Parikh SS, Putnam CD, Tainer JA. Lessons learned from structural results on uracil-DNA glycosylase. *Mutat Res.* 2000; 460:183–199. [PubMed: 10946228]
 39. David SS, O'Shea VL, Kundu S. Base-excision repair of oxidative DNA damage. *Nature.* 2007; 447:941–950. [PubMed: 17581577]
 40. McGoldrick JP, Yeh YC, Solomon M, Essigmann JM, Lu AL. Characterization of a mammalian homolog of the Escherichia coli MutY mismatch repair protein. *Mol Cell Biol.* 1995; 15:989–996. [PubMed: 7823963]
 41. Chedin F, Lieber MR, Hsieh CL. The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proc Natl Acad Sci USA.* 2002; 99:16916–16921. [PubMed: 12481029]
 42. Suetake I, Shinozaki F, Miyagawa J, Takeshima H, Tajima S. DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. *J Biol Chem.* 2004; 279:27816–27823. [PubMed: 15105426]
 43. Jia D, Jurkowska RZ, Zhang X, Jeltsch A, Cheng X. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature.* 2007; 449:248–251. [PubMed: 17713477]
 44. Jurkowska RZ, Anspach N, Urbanke C, Jia D, Reinhardt R, Nellen W, et al. Formation of nucleoprotein filaments by mammalian DNA methyltransferase Dnmt3a in complex with regulator Dnmt3L. *Nucleic Acids Res.* 2008; 36:6656–6663. [PubMed: 18945701]

45. Parrilla-Doblas JT, Ponferrada-Marin MI, Roldan-Arjona T, Ariza RR. Early steps of active DNA demethylation initiated by ROS1 glycosylase require three putative helix-invading residues. *Nucleic Acids Res.* 2013; 41:8654–8664. [PubMed: 23868090]
46. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature.* 2013; 502:472–479. [PubMed: 24153300]
47. Bennett MT, Rodgers MT, Hebert AS, Ruslander LE, Eisele L, Drohat AC. Specificity of human thymine DNA glycosylase depends on N-glycosidic bond stability. *J Am Chem Soc.* 2006; 128:12510–12519. [PubMed: 16984202]
48. Hashimoto H, Zhang X, Cheng X. Selective excision of 5-carboxylcytosine by a thymine DNA glycosylase mutant. *J Mol Biol.* 2013; 425:971–976. [PubMed: 23337108]
49. Munzel M, Lischke U, Stathis D, Pfaffeneder T, Gnerlich FA, Deiml CA, et al. Improved synthesis and mutagenicity of oligonucleotides containing 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine. *Chemistry.* 2011; 17:13782–13788. [PubMed: 22069110]
50. Burdzy A, Noyes KT, Valinluck V, Sowers LC. Synthesis of stable-isotope enriched 5-methylpyrimidines and their use as probes of base reactivity in DNA. *Nucleic Acids Res.* 2002; 30:4068–4074. [PubMed: 12235391]
51. Liu Y, Olanrewaju YO, Zhang X, Cheng X. DNA recognition of 5-carboxylcytosine by a zfp57 mutant at an atomic resolution of 0.97 Å. *Biochemistry.* 2013; 52:9310–9317. [PubMed: 24236546]
52. Kamiya H, Murata-Kamiya N, Karino N, Ueno Y, Matsuda A, Kasai H. Mutagenicity of 5-formyluracil in mammalian cells. *Nucleic Acids Symp Ser.* 2000:81–82. [PubMed: 12903278]
53. Kamiya H, Tsuchiya H, Karino N, Ueno Y, Matsuda A, Harashima H. Mutagenicity of 5-formylcytosine, an oxidation product of 5-methylcytosine, in DNA in mammalian cells. *J Biochem.* 2002; 132:551–555. [PubMed: 12359069]
54. Karino N, Ueno Y, Matsuda A. Synthesis and properties of oligonucleotides containing 5-formyl-2'-deoxycytidine: in vitro DNA polymerase reactions on DNA templates containing 5-formyl-2'-deoxycytidine. *Nucleic Acids Res.* 2001; 29:2456–2463. [PubMed: 11410651]
55. Kellinger MW, Song CX, Chong J, Lu XY, He C, Wang D. 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nat Struct Mol Biol.* 2012; 19:831–833. [PubMed: 22820989]
56. Miyazono K, Furuta Y, Watanabe-Matsui M, Miyakawa T, Ito T, Kobayashi I, et al. A sequence-specific DNA glycosylase mediates restriction-modification in *Pyrococcus abyssi*. *Nat Commun.* 2014; 5:3178. [PubMed: 24458096]
57. Reddy SM, Williams M, Cohen JI. Expression of a uracil DNA glycosylase (UNG) inhibitor in mammalian cells: varicella-zoster virus can replicate in vitro in the absence of detectable UNG activity. *Virology.* 1998; 251:393–401. [PubMed: 9837803]
58. Ikeda S, Biswas T, Roy R, Izumi T, Boldogh I, Kurosky A, et al. Purification and characterization of human NTH1, a homolog of *Escherichia coli* endonuclease III. Direct identification of Lys-212 as the active nucleophilic residue. *J Biol Chem.* 1998; 273:21585–21593. [PubMed: 9705289]
59. Hazra TK, Izumi T, Boldogh I, Imhoff B, Kow YW, Jaruga P, et al. Identification and characterization of a human DNA glycosylase for repair of modified bases in oxidatively damaged DNA. *Proc Natl Acad Sci U S A.* 2002; 99:3523–3528. [PubMed: 11904416]
60. Hazra TK, Kow YW, Hatahet Z, Imhoff B, Boldogh I, Mookapati SK, et al. Identification and characterization of a novel human DNA glycosylase for repair of cytosine-derived lesions. *J Biol Chem.* 2002; 277:30417–30420. [PubMed: 12097317]

Research Highlights

ROS1 is slow in base excision and fast in lyase activity

The isolated glycosylase domain is inactive for base excision but retains lyase activity

Addition of the C-terminal domain restores the base excision activity

The glycosylase domain and the C-terminal domain remain tightly associated

Mouse MutY homolog (MYH) does not possess 5-methylcytosine DNA glycosylase activity

Figure 1. ROS1 glycosylase domain (GD) and the C-terminal domain (CTD)

(a) Domain organizations of ROS1 full-length (FL) and ROS1 N.

(b) Activities of ROS1 FL (top panel), ROS1 N (middle panel), and ROS1 N D971N (bottom panel) on 32-bp oligos for indicated time under the single-turnover condition ($[S_{DNA}] = 50$ nM and $[E_{FL}] = 100$ nM or $[E_N] = 100$ nM or $[E_{D971N}] = 500$ nM). Labels S for substrate and P for product.

(c) The time course (0–24 h) of ROS1 N reactions ($[E_N] = 500$ nM) on five oligos ($[S_{DNA}] = 50$ nM) with various modifications under the single-turnover condition. Data ($\pm$ error bars) were averaged from three independent experiments ($n=3$).

(d) The time course (0–24 h) of ROS1 N reactions ($[E_N] = 500$ nM) on three oligos with G:X mismatches ($[S_{DNA}] = 50$ nM). Data ($\pm$ error bars) were averaged from three independent experiments ($n=3$).

(e) The time course (0–15 min) of ROS1 N AP lyase reactions (under three enzyme concentrations) on oligo ($[S_{DNA}] = 50$ nM) with an AP site.

Figure 2. Effect of C-terminal domain (CTD) on ROS1 glycosylase domain (GD) activity

(a) Activities of ROS1 N ($[E_N] = 0.5 \mu\text{M}$), the glycosylase domain ($[E_{GD}] \approx 0.5 \mu\text{M}$), and the C-terminal domain ($[E_{CTD}] \approx 1.5 \mu\text{M}$) on 32-bp oligos ($[S_{DNA}] = 50 \text{ nM}$) at 20 h (G:5mC and G:5hmC) or 15 min (G:AP) reactions. Bottom panel: The time course of AP lyase activities of ROS1 N, ROS1 N:P (with and without the protease cleavage), and GD.

(b) Bifunctional DNA glycosylases, with and without ROS1 CTD ($[E_{CTD}] \approx 1.5 \mu\text{M}$), on G:5mC and G:5hmC 32-bp oligos ($[S_{DNA}] = 50 \text{ nM}$) at 20 h reactions. AP lyase activities (15 min; 1 h for hOgg1) under the same condition are shown as positive controls. The enzyme concentrations used were $0.1 \mu\text{g} \mu\text{L}^{-1}$ of hNth1⁵⁸, 1.6 U of hOGG1 (New England Biolabs, catalog #M0241), and $50 \text{ ng} \mu\text{L}^{-1}$ of hNeil1⁵⁹ and hNeil2⁶⁰.

(c) The glycosylase domains of hTDG⁷ and mMBD4³¹ ($[E] = 500 \text{ nM}$), with and without ROS1 CTD ($[E_{CTD}] \approx 1.5 \mu\text{M}$), on G:5mC and G:5hmC 32-bp oligos ($[S_{DNA}] = 50 \text{ nM}$) at 20 h reactions. Activities on G:T mismatch (30 min) under the same condition are shown as positive controls. Bottom panel: the activity of mMBD4 on G:T substrate is unaffected by the addition of ROS1 CTD.

(d) mMYH ($[E] = 500 \text{ nM}$), with and without ROS1 CTD ($[E_{CTD}] \approx 1.5 \mu\text{M}$), on G:5mC and G:5hmC 32-bp oligos ($[S_{DNA}] = 50 \text{ nM}$) at 20 h reactions. Activities on G:A mismatch (2 h) under the same condition is shown as a positive control. Bottom panel: the activity of mMYH on G:A substrate is unaffected by the addition of ROS1 CTD.

(e) mMYH or ROS1 N ($[E] = 500 \text{ nM}$) on 71-bp oligos from the mouse IL-2 promoter³⁴ ($[S_{DNA}] = 50 \text{ nM}$) at 20 h reactions. mMYH is only active on G:A mismatch (lane 2), while ROS1 N is active on G:5mC (lane 6) under the same condition.

Figure 3. ROS1 glycosylase domain and the C-terminal domain associate tightly

(a) The Prescission protease recognition sequence (LEVLFQGP) was inserted between ROS1 GD and CTD.

(b) Activities of ROS1 N, before and after the protease cleavage, on 32-bp oligos (G:5mC) at 20 h reaction in room temperature ($[E]=500$ nM and $[S_{DNA}]=50$ nM).

(c) Elution profiles of ROS1 N:P in two consecutive runs on a Superdex 200 (10/300 GL) column (GE Healthcare) before and after the protease cleavage, in 20 mM Tris-HCl pH 8.0, 5% glycerol, 1 mM dithiothreitol, and 500 mM NaCl. Peak heights reflected relative OD280 absorbance and the retention volume shown as fractions.

(d-e) SDS-PAGE (15%) analyses of S200 fractions containing ROS1 N:P, before and after the protease cleavage.

(f) The Prescission protease recognition sequence (LEVLFQGP) was inserted between mMYH GD and its CTD.

(g) Activities of mMYH, before and after the protease cleavage, on 32-bp oligos (G:A) at 30 min reaction in 37 °C ($[E]=500$ nM and $[S_{DNA}]=50$ nM).

(h) Elution profiles of mMYH:P in two consecutive runs on a Superdex 200 (10/300 GL) column, before and after the protease cleavage, under the same conditions as shown in panel c.

(i-j) SDS-PAGE (15%) analyses of S200 fractions containing mMYH:P, before and after the protease cleavage.

(k) The hybrid enzyme generated by fusing the N-terminal mMYH glycosylase domain to the ROS1 C-terminal domain (top) and its 20 h reaction ($[E]=500$ nM and $[S_{DNA}]=50$ nM) on G:5mC and G:A 32-bp oligos (bottom).

(l-m) The effect of the opposing base on 5mC and 5hmC excision by ROS1 N:P before and after the protease digestion. Reactions were performed with $[E]=500$ nM and $[S]=50$ nM for 4 h.