Supporting Information

for

Direct Sequencing of DNA 5-Methylcytosine by Engineered Dioxygenase NTET-assisted eNAPS

Shan Zhang,[1,†] Neng-Bin Xie,[1,†] Li Zeng,[1] Fang-Yin Gang,[1] Yao-Hua Gu,[1] Min Wang,[1] Xia Guo,[1] Tong-Tong Ji,[1] Jun Xiong,[1,*] Bi-Feng Yuan[1,2,*]

[1] Department of Occupational and Environmental Health, School of Public Health, Department of Radiation and Medical Oncology, Zhongnan Hospital of Wuhan University, State Key Laboratory of Metabolism and Regulation in Complex Organisms, Wuhan University, Wuhan 430071, China.

[2] Hubei Provincial Center for Disease Control and Prevention & NHC Specialty Laboratory of Food Safety Risk Assessment and Standard Development, Wuhan 430079, China.

[†] These authors contributed equally to this work.

* Corresponding authors:

Bi-Feng Yuan. E-mail: bfyuan@whu.edu.cn

Jun Xiong. Email: jxiong@whu.edu.cn

Table of Contents

Supplementary methods

Enzymatic digestion of DNA

Nucleoside analysis was performed on a Shimadzu LC-MS/MS system comprising an LC-30AD UPLC coupled to an 8045 triple quadrupole mass spectrometer (Shimadzu, Kyoto, Japan) equipped with an electrospray ionization (Turbo Ionspray source). Chromatographic separation was achieved using a Shim-pack GIST C18 column (2.1 × 100 mm, 2.0 μm; Shimadzu) maintained at 40°C with a flow rate of 0.3 mL/min. 2 mM NH_4HCO_3 (solvent A, pH 8.6) and methanol (solvent B) were employed as mobile phases. A gradient of 0–3 min of 5% B, 3–10 min of 5–80% B, 10–12 min of 80% B, 12–13 min of 80–5% B, and 13–20 min of 5% B was used for the separation of digested nucleosides. Mass spectrometric detection was conducted in positive ion mode with the following optimized parameters: interface temperature: 300°C; collision-induced dissociation (CID) gas pressure: 230 kPa; capillary voltage: 4.0 kV. Nucleosides were quantified via multiple reaction monitoring (MRM) using the following mass transitions (precursor ions → product ions): A (252.1 → 136.1), G (268.1 → 152.1), C (228.1 → 112.1), T (243.1 → 127.1), 5mC (242.1 → 126.1), 5hmC (258.1 →142.1), 5fC (256.2 →140.2) and 5caC (272.1 →156.1). The MRM parameters of the analytes were optimized to achieve maximal detection sensitivity.

Table S1. Sequences of unmodified dsDNA substrates.

Name	Sequence (5′ to 3′)
DNA-CG	CACTGGCAGCAGCCACTGGTAACGGGATTAGCAGAGCGAGGTATGTAG GCGGTGATCGGTATCATTACCCCCATGAACAGAAATCCCCCTTACACGG AGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTAACATGGCCCGCTT TATCAGAAGCCAGACATTAACGCTTCTGGAGAAACTCAACGAGCTGGAC GCGGATGAACAGGCAGACATCTGTGAATCGCTTCACGACCACGCTGATG AGCTTTACCGCAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAAGTA ATTGTTAGTGGAATGT
DNA-CC	CACTGGCAGCAGCCACTGGTAACCGGATTAGCAGAGCGAGGTATGTAG GCGGTGATCGGTATCATTACCCCCATGAACAGAAATCCCCCTTACACGG AGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTAACATGGCCCGCTT TATCAGAAGCCAGACATTAACGCTTCTGGAGAAACTCAACGAGCTGGAC GCGGATGAACAGGCAGACATCTGTGAATCGCTTCACGACCACGCTGATG AGCTTTACCGCAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAAGTA ATTGTTAGTGGAATGT
DNA-CT	CACTGGCAGCAGCCACTGGTAACTGGATTAGCAGAGCGAGGTATGTAG GCGGTGATCGGTATCATTACCCCCATGAACAGAAATCCCCCTTACACGG AGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTAACATGGCCCGCTT TATCAGAAGCCAGACATTAACGCTTCTGGAGAAACTCAACGAGCTGGAC GCGGATGAACAGGCAGACATCTGTGAATCGCTTCACGACCACGCTGATG AGCTTTACCGCAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAAGTA ATTGTTAGTGGAATGT
DNA-CA	CACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAG GCGGTGATCGGTATCATTACCCCCATGAACAGAAATCCCCCTTACACGG AGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTAACATGGCCCGCTT TATCAGAAGCCAGACATTAACGCTTCTGGAGAAACTCAACGAGCTGGAC GCGGATGAACAGGCAGACATCTGTGAATCGCTTCACGACCACGCTGATG AGCTTTACCGCAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAAGTA ATTGTTAGTGGAATGT

Table S2. Sequences of 5mC- and 5hmC-modified dsDNA substrates.

Name	Sequence (5′ to 3′)
DNA-5mCG	CACTGGCAGCAGCCACTGGTAA5mCGGGATTAGCAGAGCGAGGTA TGTAGGCGGTGATCGGTATCATTACCCCCATGAACAGAAATCCCC CTTACACGGAGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTA ACATGGCCCGCTTTATCAGAAGCCAGACATTAACGCTTCTGGAGA AACTCAACGAGCTGGACGCGGATGAACAGGCAGACATCTGTGAAT CGCTTCACGACCACGCTGATGAGCTTTACCGCAGCTGCCTCGCGCG TTTCGGTGATGACGGTGAAAAGTAATTGTTAGTGGAATGT
DNA-5mCC	CACTGGCAGCAGCCACTGGTAA5mCCGGATTAGCAGAGCGAGGTA TGTAGGCGGTGATCGGTATCATTACCCCCATGAACAGAAATCCCC CTTACACGGAGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTA ACATGGCCCGCTTTATCAGAAGCCAGACATTAACGCTTCTGGAGA AACTCAACGAGCTGGACGCGGATGAACAGGCAGACATCTGTGAAT CGCTTCACGACCACGCTGATGAGCTTTACCGCAGCTGCCTCGCGCG TTTCGGTGATGACGGTGAAAAGTAATTGTTAGTGGAATGT
DNA-5mCT	CACTGGCAGCAGCCACTGGTAA5mCTGGATTAGCAGAGCGAGGTA TGTAGGCGGTGATCGGTATCATTACCCCCATGAACAGAAATCCCC CTTACACGGAGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTA ACATGGCCCGCTTTATCAGAAGCCAGACATTAACGCTTCTGGAGA AACTCAACGAGCTGGACGCGGATGAACAGGCAGACATCTGTGAAT CGCTTCACGACCACGCTGATGAGCTTTACCGCAGCTGCCTCGCGCG TTTCGGTGATGACGGTGAAAAGTAATTGTTAGTGGAATGT
DNA-5mCA	CACTGGCAGCAGCCACTGGTAA5mCAGGATTAGCAGAGCGAGGTA TGTAGGCGGTGATCGGTATCATTACCCCCATGAACAGAAATCCCC CTTACACGGAGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTA ACATGGCCCGCTTTATCAGAAGCCAGACATTAACGCTTCTGGAGA AACTCAACGAGCTGGACGCGGATGAACAGGCAGACATCTGTGAAT CGCTTCACGACCACGCTGATGAGCTTTACCGCAGCTGCCTCGCGCG TTTCGGTGATGACGGTGAAAAGTAATTGTTAGTGGAATGT
DNA-5hmC	CACTGGCAGCAGCCACTGGTAA5hmCAGGATTAGCAGAGCGAGGT ATGTAGGCGGTGATCGGTATCATTACCCCCATGAACAGAAATCCC CCTTACACGGAGGCATCAGTGACCAAACAGGAAAAAACCGCCCTT AACATGGCCCGCTTTATCAGAAGCCAGACATTAACGCTTCTGGAG AAACTCAACGAGCTGGACGCGGATGAACAGGCAGACATCTGTGA ATCGCTTCACGACCACGCTGATGAGCTTTACCGCAGCTGCCTCGCG CGTTTCGGTGATGACGGTGAAAAGTAATTGTTAGTGGAATGT

Table S3. Sequences of PCR primers for dsDNA preparation and amplification of eNAPS-treated DNA.

Name	Sequence (5′ to 3′)
5mCG-F	CACTGGCAGCAGCCACTGGTAA5mCGGGATTAGCAGAGCGAGGTAT
5mCG-R	ACATTCCACTAACAATTACTTTTCACCGTCATC
5mCC-F	CACTGGCAGCAGCCACTGGTAA5mCCGGATTAGCAGAGCGAGGTAT
5mCC-R	ACATTCCACTAACAATTACTTTTCACCGTCATC
5mCT-F	CACTGGCAGCAGCCACTGGTAA5mCTGGATTAGCAGAGCGAGGTAT
5mCT-R	ACATTCCACTAACAATTACTTTTCACCGTCATC
5mCA-F	CACTGGCAGCAGCCACTGGTAA5mCAGGATTAGCAGAGCGAGGTAT
5mCA-R	ACATTCCACTAACAATTACTTTTCACCGTCATC
5hmC-F	CACTGGCAGCAGCCACTGGTAA5hmCAGGATTAGCAGAGCGAGGTAT
5hmC-R	ATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTG
PS-E	TGGTGCTCGAGTGCGGCCGCAGGCTTAACGACATTCCACTAACAATTAC
PS-F	CAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCAC
PS-R	TGGTGCTCGAGTGCGGCCGCAG

Table S4. Amino acid sequences of eNTET, wtNTET, and TDG.

Name	Amino acid sequence (N-terminal to C-terminal)
eNTET	MSKSNEPGKATGEGKPVNNKWLNNAGKDLGSPVPDRIANKLRDKEFESFDD FRETFWEEVSKDPELSKQFSRNNNDRMKVGKAPKTRTQDVSGKRTSFELNH QKPIEQNGGVYDMDNISVVTPKRNIDIEGMTTFKQQTIKEKETKRKYCIKGTT ANLTQTHPNGPVCVNRGEEVANTTTLLDSGGGINKKSLLQNLLSKCKTTFQQ SFTNANITLKDEKWLKNVRTAYFVCDHDGSVELAYLPNVLPKELVEEFTEKF ESIQTGRKKDTGYSGILDNSMPFNYVTADLSQELGQYLSEIVNPQINYYISKL LTCVSSRTINYLVSLNDSYYALNNCLYPSTAFNSLKPSNDGHRIRKPHKDNLD ITPSSAFYFGNFQNTEGYLELTDKNCKVFVQPGDVLFFKGNEYKHVVANITS GWRIGLVYFAHKGSKTKPYYEDTQKNSLKIHKETK
wtNTET	MTTFKQQTIKEKETKRKYCIKGTTANLTQTHPNGPVCVNRGEEVANTTTLLD SGGGINKKSLLQNLLSKCKTTFQQSFTNANITLKDEKWLKNVRTAYFVCDH DGSVELAYLPNVLPKELVEEFTEKFESIQTGRKKDTGYSGILDNSMPFNYVTA DLSQELGQYLSEIVNPQINYYISKLLTCVSSRTINYLVSLNDSYYALNNCLYPS TAFNSLKPSNDGHRIRKPHKDNLDITPSSLFYFGNFQNTEGYLELTDKNCKVF VQPGDVLFFKGNEYKHVVANITSGWRIGLVYFAHKGSKTKPYYEDTQKNSL KIHKETK
TDG	SKKSGKSAKSKEKQEKITDTFKVKRKVDRFNGVSEAELLTKTLPDILTFNLDI VIIGINPGLMAAYKGHHYPGPGNHFWKCLFMSGLSEVQLNHMDDHTLPGKY GIGFTNMVERTTPGSKDLSSKEFREGGRILVQKLQKYQPRIAVFNGKCIYEIFS KEVFGVKVKNLEFGLQPHKIPDTETLCYVMPSSSARCAQFPRAQDKVHYYIK LKDLRDQLKGIERNMDV

Table S5. Sequences of FAM-labeled DNA substrates.

Name	Sequence (5′ to 3′)
FAM-5mC	FAM-CACTGGCAGCAGCCACTGGTAA5mCAGGATTAGCAGAGCGAGGTAT
5mC-R	ATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTG
FAM-5hmC	FAM-CACTGGCAGCAGCCACTGGTAA5hmCAGGATTAGCAGAGCGAGGTAT
5hmC-R	ATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTG

Table S6. Sequences of primers used for detection of 5mC in the gene body of *EGFR* of human lung tissues by eNAPS.

Name	Sequence (5′ to 3′)
EGFR (chr7:55109538 and chr7:55109595)	Extension： GCCACGACCGGCGCCGAGCTCGGAATGATAGAATTTG TCC Forward: TGATTGTAATGTTTGTTGTTTATTGGTTAG Reverse: AATTTATACCTAAACCAAAAAAAAATATCCC

Table S7. Sequences of primers used for detection of 5mC in the gene body of EGFR of human lung tissues by BS-seq.

Name	Sequence (5′ to 3′)
EGFR (chr7:55109538)	Forward: TGATTGTAATGTTTGTTGTTTATTGGTTAG Reverse: AATTTATACCTAAACCAAAAAAAAATATCCC
EGFR (chr7:55109595)	Forward: ATTGTAATGTTTGTTGTTTATTGGTTAGGGTAGTTTTT Reverse: GCCAAAATAATAAAATTTATCCTATAAATCAAAAATTT

Table S8. Information of the individual 5mC sites detected in genome of human lung tissue.

Genome location	Gene	Gene element
chr7: 55109538, CRCh38	EGFR	Gene body
chr7: 55109595, CRCh38	EGFR	Gene body

Figure S1. Purification of eNTET protein. (A) Size-exclusion chromatography (SEC) purification of eNTET using a Superdex-200 10/300 column (AKTA purifier). (B) SDS-PAGE analysis of the fractions of peak containing purified eNTET protein.

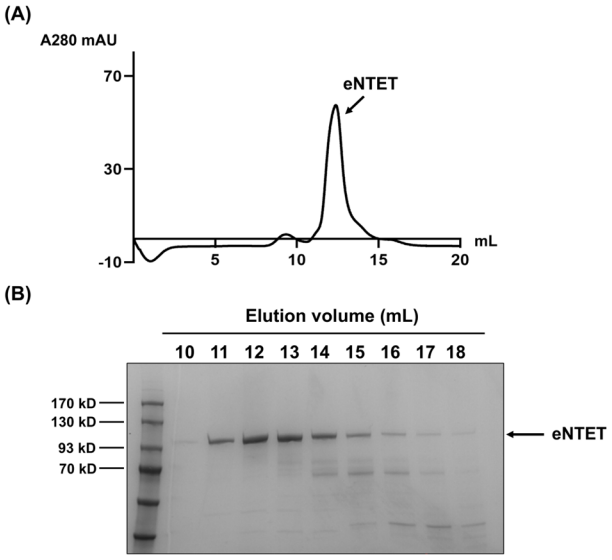

Figure S2. Expression and purification of TDG protein. (A) The schematic illustration of plasmid for the expression of TDG protein. (B) SDS-PAGE analysis of the purified TDG protein.

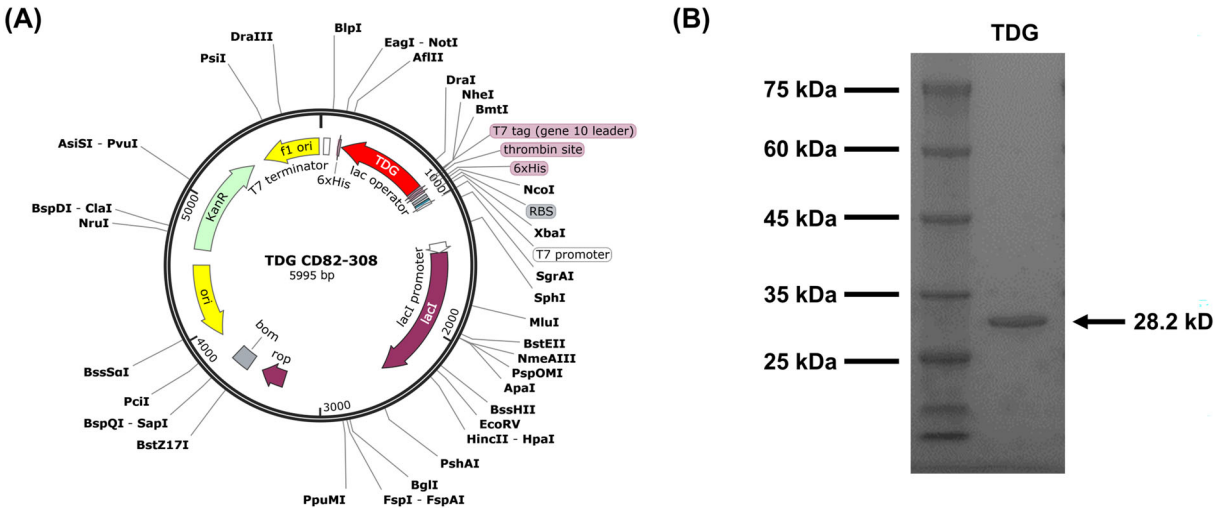

Figure S3. Expression and purification of wtNTET protein. (A) The schematic illustration of plasmid for the expression of wtNTET protein. (B) SDS-PAGE analysis of the purified wtNTET protein.

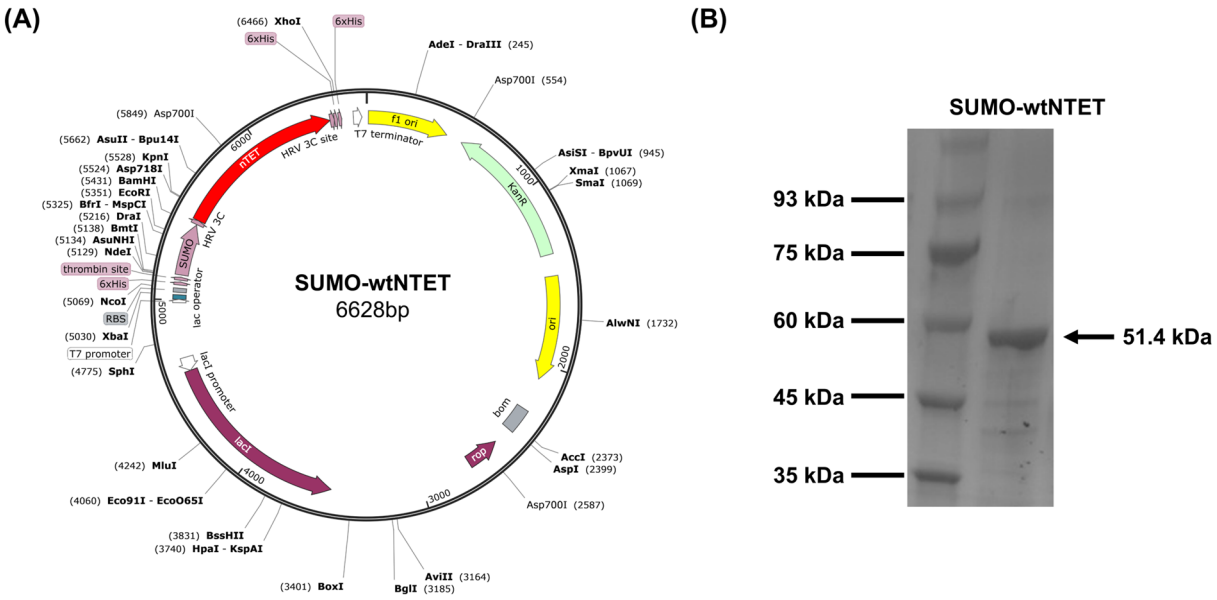

Figure S4. The oxidation activities of eNTET and wtNTET proteins toward 5mC.

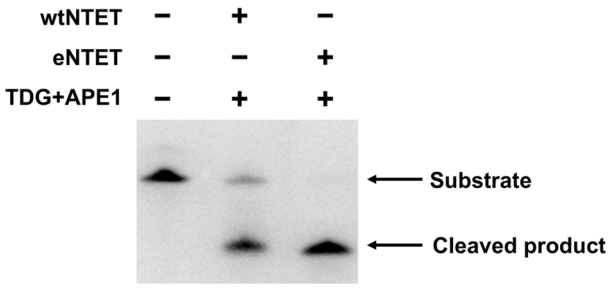

Figure S5. Extracted ion chromatograms of 5mC, 5hmC, 5fC, and 5caC after wtNTET treatment using LC-MS/MS analysis.

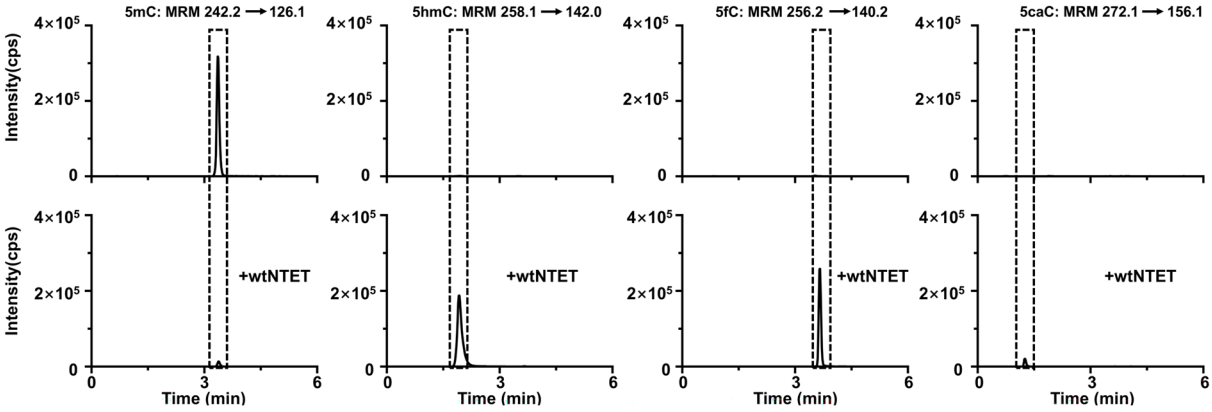

Figure S6. Sanger sequencing of 5hmC by eNAPS without β-GT. (A) Schematic illustration of the readout of 5hmC by eNAPS without β-GT. (B) Sanger sequencing result of 5hmC by eNAPS without β-GT.

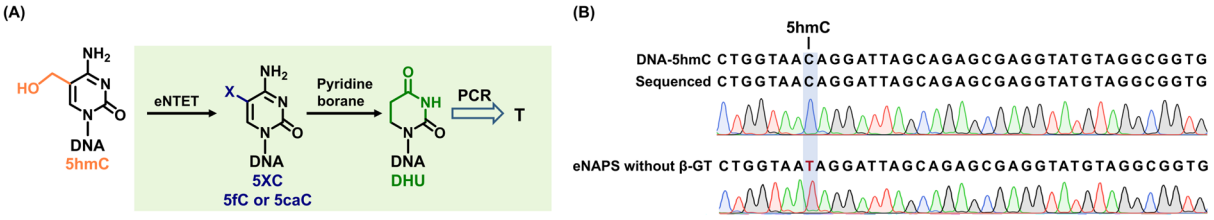

Figure S7. Evaluation of the performance of eNAPS by colony sequencing. Fifty clones were selected for the evaluation. Red, read as T.