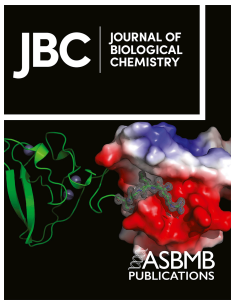
Journal Pre-proof



DNA methyltransferase DNMT3A forms interaction networks with the CpG site and flanking sequence elements for efficient methylation

Michael Dukatz, Marianna Dittrich, Elias Stahl, Sabrina Adam, Alex de Mendoza, Pavel Bashtrykov, Albert Jeltsch

PII: S0021-9258(22)00905-X

DOI: https://doi.org/10.1016/j.jbc.2022.102462

Reference: JBC 102462

To appear in: *Journal of Biological Chemistry*

Received Date: 4 August 2022

Revised Date: 30 August 2022

Accepted Date: 1 September 2022

Please cite this article as: Dukatz M, Dittrich M, Stahl E, Adam S, de Mendoza A, Bashtrykov P, Jeltsch A, DNA methyltransferase DNMT3A forms interaction networks with the CpG site and flanking sequence elements for efficient methylation, *Journal of Biological Chemistry* (2022), doi: https://doi.org/10.1016/j.jbc.2022.102462.

DNA methyltransferase DNMT3A forms interaction networks with the CpG site and flanking sequence elements for efficient methylation

Michael Dukatz[1], Marianna Dittrich[1], Elias Stahl[1], Sabrina Adam[1], Alex de Mendoza[2], Pavel Bashtrykov[1], & Albert Jeltsch[1,*]

[1] Institute of Biochemistry and Technical Biochemistry, University of Stuttgart, Allmandring 31, 70569 Stuttgart, Germany

[2] School of Biological and Behavioural Sciences, Queen Mary University of London, Mile End Road, E1 4NS London, United Kingdom

* Corresponding author
Prof. Dr. Albert Jeltsch
Phone: +49 711 685 64390
Fax: +49 711 685 64392
E-Mail: albert.jeltsch@ibtb.uni-stuttgart.de

Abstract

Specific DNA methylation at CpG and non-CpG sites is essential for chromatin regulation. The DNA methyltransferase DNMT3A interacts with target sites surrounded by variable DNA sequences with its TRD and RD loops, but the functional necessity of these interactions is unclear. We investigated CpG and non-CpG methylation in randomized sequence context using wildtype DNMT3A and several DNMT3A variants containing mutations at DNA-interacting residues. Our data revealed the flanking sequence of target sites between the -2 and up to the +8 position modulates methylation rates >100-fold. Non-CpG methylation flanking preferences were even stronger and favor C(+1). R836 and N838 in concert mediate recognition of the CpG guanine. R836 changes its conformation in a flanking sequence-dependent manner and either contacts the CpG guanine or the +1/+2 flank, thereby coupling the interaction with both sequence elements. R836 suppresses activity at CNT sites, but supports methylation of CAC substrates, the preferred target for non-CpG methylation of DNMT3A in cells. N838 helps to balance this effect and prevent the preference for C(+1) from becoming too strong . Surprisingly, we found L883 reduces DNMT3A activity despite being highly conserved in evolution. However, mutations at L883 disrupt the DNMT3A-specific DNA-interactions of the RD loop, leading to altered flanking sequence preferences. Similar effects occur after the R882H mutation in cancer cells. Our data reveal that DNMT3A forms flexible and interdependent interaction networks with the CpG guanine and flanking residues that ensures recognition of the CpG and efficient methylation of the cytosine in contexts of variable flanking sequences.

Key words

DNMT3A; DNA methylation; enzyme mechanism; enzyme specificity; protein-DNA interaction

Introduction

In human cells, DNA is methylated at the cytosine-C5 position in about 70-80% of all CpG but also at non-CpG sites (1-3). DNA methylation has important roles in the regulation of protein-DNA interactions and it is involved in gene expression, X-chromosome inactivation, genomic stability, cell differentiation and mammalian development (1,4,5). The DNA methyltransferases DNMT3A and DNMT3B are so-called de novo DNA methyltransferases (6) which set up DNA methylation patterns during gametogenesis and post-implantation development (5,7). The DNMT3A enzyme is essential in mammalian development, but it also has important roles in carcinogenesis (8,9) and in the brain (10,11). The catalytically inactive DNMT3-like protein (DNMT3L) has an important regulatory role in this gametogenesis by acting as a stimulator of DNMT3A (6).

The catalytically active C-terminal domain of DNMT3A (DNMT3AC) forms a linear heterotetrameric complex with the C-terminal domain of DNMT3L (DNMT3LC) in a 3L-3A-3A-3L arrangement (Figure 1A) (12). Biochemical studies showed that the DNMT3L subunits in the DNMT3A/3L heterotetramer can be replaced by two additional subunits of DNMT3AC (13,14) yielding a DNMT3A homotetramer as the smallest catalytically active form of DNMT3A. The central DNMT3AC subunits of such homo- or heterotetramers interact via the so-called RD interface and form the DNA binding site of the complex (12). Structures of the DNMT3A/3L complex bound to DNA showed that these subunits interact with two CpG sites in 12 base pair distance and methylate them in opposite DNA strands (15,16) (Figure 1A), but biochemical data showed that other arrangements of co-methylation of CpG sites are also possible (17).

Structural studies showed that the DNA interaction of DNMT3A and DNMT3B is mediated by three protein regions (15,18): 1) a loop directly following the active center (residues G707–K721 in DNMT3A), 2) a loop from the target recognition domain (TRD-loop, residues R831–F848), and 3) the RD interface (RD-loop, residues S881-R887), which together create a continuous DNA-binding surface. Among them, residues in the TRD and RD loops form an interconnected network of contacts with the DNA which mediates recognition of the CpG guanine and of the flanking sequence in the 3' direction (Figure 1A). Both enzymes interact with the CpG guanine with residues from the TRD loop and they prefer methylation of DNA within CpG sites, but they also introduce methylation at lower level at non-CpG sites (19). Non-CpG methylation has been connected with gene regulatory and chromatin modulating functions and to X-chromosome inactivation mainly in the nervous system (2,20). Previous Deep Enzymology studies showed that DNMT3A and DNMT3B methylate CpG and non-CpG sites in strong dependence on their flanking sequences (19). Differences in flanking sequence preferences of DNMT3A and DNMT3B could be connected to sequence and structural differences in the RD loops of both enzymes (18). In addition, the somatic R882H mutation in the RD loop of DNMT3A occurs in a large fraction of AML tumors (8,9) and it was shown to alter the flanking sequence preferences of DNMT3A strongly, making them more similar to those of DNMT3B (21-24). Moreover, R882H was shown to increase the stability of the RD interface which explains its dominant cellular phenotype (24).

Structural studies have documented the involvement of several amino acid residues from the TRD and RD loops in the DNA interaction of DNMT3A (Figure 1B). In a CGT structure (15), The side-chain of R836 from the TRD loop forms an H-bond with the O6 of the CpG guanine plus one additional water-mediated contact to the guanine N7. However, in a CGA structure (16), the R836 side chain moved up and it formed a side-chain H-bond to the N6 atom of the adenine in the +1 flanking position, A(+1), and a second H-bond to the +2 flank site indicating flexible and combined readout of

CpG specificity and the +1/+2 flank. In the CGA structure, the side chain of N838 contacts the O6 of the CpG guanine, while in the CGT structure this residue is flipped up and it forms an H-bond to the phosphodiester group connecting the +2 and +3 flanking site and a direct contact to the +2' base (where ' refers to the non-target DNA strand). Interestingly, the CpG recognition pattern in the CGA structure of DNMT3A resembles the structure of DNMT3B, where N779 (corresponding to N838 of DNMT3A) mediates the CpG guanine recognition. In both DNMT3A structures, the N7 atom of the CpG guanine is bound with a water-mediated H-bond to T834, this contact is also observed in DNMT3B. Moreover, S837 from the TRD loop forms a side chain H-bond to the N7 of G(+3') in both DNMT3A structures, thereby directly contacting the +3 flank. The RD loop residues of DNMT3A are engaged in several contacts to the flanking residues at the 3' side of the CpG (Figure 1C). S881 and R887 form side chain H-bonds to the phosphodiester group between the flanking bases at +5' and +6'. R882 contacts the phosphodiester group between the flanking bases at +3' and +4'. L883 is sandwiched between S837 and the DNA backbone at the +4' and +5' flanking bases with multiple van-der-Waals contacts and a backbone H-bond to the phosphodiester group at this site. The TRD and RD loops are directly connected by the positioning of R836, which pushes R882 away from the DNA in its uplifted conformation (Figure 1C). Moreover, the R882 side chain contacts the side chain of S837 in both structures.

While structural studies document the dependence of the dynamic DNA interaction networks of DNMT3A on the flanking sequences of the CpG site, the functional relevance of this context dependent DNA recognition has not been investigated. It was the aim of this study to unravel the functional consequences of the interconnected recognition of the CpG guanine, the +1 to +3 flanking sequences and flanking sequences further away from the CpG site and determine their effects in CpG and non-CpG methylation. To this end, several mutants of critical residues in DNMT3A were prepared and their activity, CpG specificity and flanking preference in CpG and non-CpG context was determined in Deep Enzymology experiments. Our data provide a comprehensive view on the dynamic interaction of DNMT3A with the CpG sites and the 3' flanking sequences, which defines the roles of individual amino acids residues in this contact network and their interdependence.

Results

Selection of residues for mutagenesis, mutant generation and purification and initial activity assessment

To investigate the interaction of the catalytic domain of DNMT3A with the target CpG guanine and the 3' flanking DNA, seven DNMT3A residues in the TRD and RD loop were selected from the available structures (15,16) and mutated to alanine, viz. R836, S837, N838, S881, R882, L883 and R887. The cancer mutant R882H was investigated as well, because of its medical relevance. A multiple sequence alignment of DNMT3A and DNMT3B enzymes from several vertebrate species and DNMT3 from amphioxus and sea urchin as representatives of non-vertebrates (Figure 1D) showed that all these residues are fully conserved in DNMT3A enzymes, which is matching their putative highly important mechanistic role. However, there are strong differences in roles of these residues between DNMT3A and DNMT3B. In the TRD loop, R836 is replaced by K777 (using human DNMT3B numbering) in most of the DNMT3B sequences and the RD loop is one of the most divergent regions of DNMT3A and DNMT3B, where the DNMT3A residues S881-R882-L883-…-R887 are replaced by G822-R823-G824 and K828 in DNMT3B. DNMT3AC mutations were generated by site-directed

mutagenesis, overexpressed and purified as MBP-tagged proteins. All mutants were obtained at comparable purity (Supplemental Figure 1). Protein concentrations were determined by OD^{280nm} and validated on Coomassie BB stained polyacrylamide gels. Methylation activity was initially determined using radioactively labelled AdoMet and a biotinylated double-stranded 30-mer oligonucleotide substrate with a single CpG site that has been employed as reference substrate in several previous studies (24,25). Initial reaction rates were calculated from time courses of the methylation reactions and averaged over several (3-6) independent experimental repeats (blue bars in Figure 1E). The results revealed a striking 5-fold enhancement of activity for L883A, slight reductions in activity of R836A and R882H and a stronger reduction of activity in the case of R882A. The catalytic activities of N838A, S837A, S881A and R887A were similar to the WT activity.

Analysis of the flanking preferences of DNMT3A mutants in CpG context

To explore the flanking sequence preferences of CpG methylation by DNMT3A WT and the mutants, we next conducted Deep Enzymology experiments (19). We first investigated the methylation of a pool of DNA substrates, in which a target CpG site was flanked by 10 random nucleotides on either side. The substrate pool was methylated by the enzymes, followed by hairpin ligation, bisulfite conversion, PCR amplification and NGS analysis as described previously (18,22,26). In each case, two independent experiments were conducted as detailed in Supplemental Table 1. R882H CpG methylation data were used from our previous publication (22). To exclude effects of the fixed sequences outside of the randomized region, the following analyses were restricted to the -8 to +8 region which then still have two random nucleotides on both sides. Using these data sets, we determined the average methylation levels of all substrates containing a particular base at one of the -8 to +8 flank sites to identify bases favorable or unfavorable for methylation activity. The data were expressed in observed/expected values of the corresponding base in the methylated DNA strands. We first compared the results of the experimental repeats conducted with each mutant, which always showed high correlation of the derived profiles, which was also the case for the WT data determined here and previously (18) (Supplemental Table 2). In general, the average methylation activities determined in the Deep Enzymology experiments were similar to the radioactive methylation rates of the respective mutant (orange bars in Figure 1E). Most strikingly, the rate enhancement with L883A seen in the radioactive kinetics was observed in the NGS assay as well.

Comparison of the effects of flanking sites on CpG activity

Detailed analysis of the flanking sequence preferences revealed strong position and mutant specific effects. DNMT3A WT flanking effects were detectable between the -2 and (mainly) +6 positions (Figure 2A). At the -2 site, T was strongly preferred while G and A were disfavored. C and A were preferred at -1 where T was most disfavored. At +1 to +3, C was strongly preferred (in addition to A at +3) and G was disfavored. Weaker effects were detectable at +4 (preference for C) and +6 (preference for T and disfavor for G). Due to the limited number of reads, average methylation levels could only be determined for all NNCGNNN flanks. The numbers varied between 95% for TCCGCCC and no detectable methylation for 11 NNCGNNN sequence motifs that were all very G-rich, illustrating the strong effect of the flanking sequences on DNMT3A activity in agreement with previous data (18,22,24).

To identify the flanking base pairs most relevant for the catalytic activity, the RMSD value for the deviation of the occurrence of each base from the expected value (obs/exp = 1.0) was determined for each position. The RMSD values of WT and mutant profiles showed effects between the -2 and +7

flanking sites (Figure 2B). Strong overall flanking sequence preferences were observed with R836A, N838A, R882A, and R882H, while effects at S837A, S881A, L883A and R887A were smaller. Largest effects were observed at the -2 and +1 sites, the largest difference between DNMT3A and DNMT3B profiles (18) were observed at -1 and +1 sites. Detailed analyses of the mutant preferences were conducted for the -2 to +5 positions and compared with WT (Figure 2C). R836A showed a strong change of preferences at the +1 site, where the preferences changed from C>T>A>G (WT) to T>C>A>>G indicating that the disfavor for G was even more enhanced and the preferences for T and C swapped. In addition, the disfavor for G(+2) and G(+3) was reduced for R836A. In case of S837A, no big effects were observed, but at the +1 and +2 sites the preferences of WT were enhanced and the preference for A(+3) was reduced. N838A led to strong changes in the flanking sequence preferences at the +1 position, where the preferences changed from C>T>A>G (WT) to C>T>>G~A. In addition, the preference for C(+3) was reduced. S881A did not show big effects. R882A showed a strong elevation of WT effects at -1 and a strong shift of preferences at +1 towards G>A~C>>T, indicating that the preference for G increased while T and C dropped. Moreover, the preference for G(+2) was increased as well. A similar change in preferences at the +1 and +2 sites as R882A was observed for R882H. In addition, the preference for A(+3) dropped. L883A also caused the same change in preferences at +1 as R882A and R882H and the drop of the A(+3) preference as observed for R882H. For R887A, a mild change of preferences at +4 was observed, where G and A were more preferred that by WT.

In summary, these data reveal strong >100-fold flanking sequence preferences of DNMT3A, that were heavily influenced by R836, N838, R882 and L883. To further compile the data and compare the deviations of the preferences of the DNMT3A mutants from the preference profiles of DNMT3A and DNMT3B (taken from (18)), the R-values of the correlation of -2 to +5 flanking profiles of mutants with DNMT3A and DNMT3B WT were determined and plotted, revealing several distinct groups of mutants (Figure 2D). The R882A, R882H and L883A mutants caused large and similar changes of the profile making them similar to DNMT3B. In contrast, R836A caused an equally large but different shift of preferences that does move it towards DNMT3B. R887A and N838A also caused distinct changes of flanking sequence preferences. Finally, S837A and S881A caused only weak effects and cluster together with WT DNMT3A.

Combined readout of the +1 to +3 flank sites

Next, we were interested to determine the combined readout of the base pairs at the +1 to +3 flank sites. To this end, the relative preferences of all 16 +2/+3 flank dinucleotides were determined for each given base at the +1 site, revealing several interesting combined effects with WT DNMT3A (Figure 2E). For example: 1) T(+1) is favored in the +2/+3 context of AG, 2) A(+1) is favored in the +2/+3 context GA, GC and GG, but disfavored in combination with CG, 3) C(+1) is favored in the TG and TT context. The corresponding patterns were determined for all mutants (Supplemental Figure 2) and the correlation of mutant dinucleotide preferences with the WT were determined (Figure 2F). This analysis revealed that R882A by far showed the weakest correlation, which underscores the importance of R882 for the +1 to +3 flank interaction. The more conservative R882H exchange led to much smaller changes. Interestingly, the mutations in the TRD loop led to reduction of the correlation of +2/+3 flank preferences in the presence of A(+1), and in the case of R836A also for G(+1). Finally, L883A showed reduced correlation with all bases at the +1 site, except G. Based on this one can conclude that R836 forms an essential interaction with G(+1), the TRD loop mediates interactions with A(+1), and L883A is required for the +1 to +3 flank interaction, if there is no G at the +1 site. Among the detailed results, the strong changes of +2/+3 site effects of G(+1) with R836A, strong preferences for AAT, CAA and CAT with N838A and TCC with R882H, R882A and L883A are the strongest effects (Supplemental Figure 2).

In summary, these data show a combined readout of DNMT3A with the +1 and +2/3 flanking sites that is mediated by R836 and the RD loop residues, mainly R882A.

Effects of the outer flanks

In the previous experiments shown in Figure 2A and 2B, flanking effects on CpG methylation were detectable also at the +4 to +7 positions. Related to this, in a previous study we showed strong effects of these sites on the co-methylation of CpG sites in a distance of 12 base pairs and found that A/T bases were preferred (17). We suspected that effects of these regions, which we now call "outer flanks", might have been partially masked by the stronger effects of the inner -3 to +3 flanks. Hence, we studied outer flank effects in five different substrates with fixed inner flanks, which were chosen to cover the spectrum from highly preferred to strongly disfavored inner flanks. The substrates were prepared with the different inner flanks surrounded by randomized -10 to -4 and +4 to +10 regions, methylated by DNMT3A WT and the mutants in two experimental repeats and analyzed as described (Supplemental Table 3). The individual repeats showed very good correlation in most cases (Supplemental Table 4) and the data were merged for further analysis.

As shown in Figure 3A, clear and distinct outer flank effects were observed with WT mainly at the -4 and +4 to +8 sites. To estimate the overall effect size, average methylation levels were determined for substrates with each combination of N-NNN bases at the 4, +4, +5 and +6 positions showing that these values differed by up to 8-fold. In general, our data indicate that only small sequence preferences and effects on activity were observed in the context of a highly preferred inner flank (GTACGTCA in Figure 3A), but the influence of the outer flanks on activity increased with more and more disfavored inner flanks. In general, the effect of 3' flanks was stronger than of the 5' flanks and often A/T bases were preferred at these places in agreement with our previous data (17). Interestingly, strong 5' outer flank effects were only observed in the TGCCGTTG substrate which is missing the highly favored T(-2) residue. Corresponding outer flank profiles were also determined for the mutants (Supplemental Figure 3). For comparison the RMSD deviation of the base distributions averaged for all positions were compiled for all five substrates (Figure 3B). The data revealed very small outer flank effects on the most preferred substrate for WT and all mutants. Interestingly, on the more disfavored substrates, the WT showed more pronounced flanking effects than the mutants, which is also clearly visible when comparing the individual profiles (Supplemental Figure 3).

Next, the RMSD values of the differences between the outer flanking sequence preferences of the mutants and WT were calculated for each position and averaged for the five substrates (Figure 3C). The data revealed strong differences between WT and mutant preferences at the 3' side (+4 to +8), which is in agreement with the fact that the mutants were selected for their potential interaction with this part of the DNA. Strongest effects were observed at +4, where many mutants showed strong deviations, like a strong disfavor for G(+4) of L883A in the ATT-ATG and TGC-TTG substrates or a strong preference for G(+4) of R836A and S837A in the GTC-CGA substrate. It is noticeable that strong outer flank effects were observed with R836A, R887A, S881A and R882A/H even at the +6 to +8 sites. To compare the data at a combined level, +4 to +8 outer flanking sequence preferences were averaged for all five substrates for WT and all mutants (Figure 3D). The data clearly indicate the preference of WT for A or T at all these sites, which has been lost or strongly reduced with all mutants with the exception that the T(+6) preference is still observed, though reduced with R887 and R882 mutants.

In summary, these data show that outer flank sequences influence the activity of DNMT3A and largest outer flank effects were observed at unfavored inner flanks.

Analysis of the CpG recognition of DNMT3A WT and mutants

To study the activity and flanking sequence preferences of non-CpG methylation by DNMT3A and its mutants, Deep Enzymology experiments were also conducted with a substrate containing a CpN target site in a randomized sequence context and the data were split into CpG, CpA, CpT and CpC methylation. Read counts and methylation levels of the individual experimental repeats are listed in Supplemental Table 5, correlations of the flanking profiles of individual repeats are shown in Supplemental Table 6. In general, correlations of the CpG and CpA data were good, which is related to the fact that all enzymes were most active in these sequence contexts. In addition, the CpG flank preferences determined with the CpN substrate were compared with the preferences determined with the CpG substrate and found to be highly correlated, despite the fact that CpG methylation levels in this CN methylation analysis were intentionally higher, to make non-CpG methylation more detectable (Supplemental Table 6). Correlations of the CpT data were lower due to the low overall methylation activity leading to small numbers of methylated reads in some cases. The CpC correlation was only satisfying in case of the WT and the most active mutants. Hence the CpT and CpC flanking profiles of some mutants could not be analyzed in detail.

First, the average methylation levels of C in each CpN context were determined and the data fitted to an exponential reaction progress curve to determine the relative methylation rates in the four sequence contexts (CpG, CpA, CpT and CpC) (Figure 4A and Supplemental Figure 4). For the most accurate measurement of the relative methylation rates, the time points describing the individual reaction progress in each methylation experiment were included in the fitting. The relative activity of WT DNMT3A at CpA, CpT and CpC sites in random flanking sequence context was found to be 4.9 %, 2.1 % and 0.4 %, respectively, of the CpG methylation rate. The largest change in CpG recognition was observed with R836A showing a more than 40-fold increase in CpC methylation, combined with a 3- to 4-fold increase in CpA and CpT methylation. S837A and S881A showed similar relative CpA and CpT methylation as WT together with a mild increase in relative CpC methylation. N838A, R882A, R882H, L883A and R887A showed reduced relative non-CpG methylation. Our results demonstrate that R836A is a very critical residue for CpG recognition as already observed before with individual test substrates (15). We show that CpC methylation (which is very low with WT DNMT3A) becomes elevated most prominently, followed by CpA and CpT. Actually, methylation activity of R836A at CpC sites was about 4-fold higher than at CpT sites, while it is 5-fold lower in case of WT. These results are in very good agreement with cellular non-CpG methylation levels of WT and R836A where a strong increase in CpC methylation was observed as well (15). Of note, mutation of N838 led to a decrease in non-CpG methylation indicating a higher CpG specificity although this residue forms direct contacts to the CpG guanine. This paradoxical finding will be discussed later.

Analysis of the flanking preferences of DNMT3A WT and mutants in non-CpG context

As described above for the CpG flanking sequence profiles, we determined the average methylation levels of all substrates containing a particular base at one of the -8 to +8 flank sites to identify bases favorable or unfavorable for activity. In Figure 4B the profiles of WT are shown for the -4 to +4 region, revealing strong and CpN specific effects. At the +1 site, which is directly adjacent to the CpN site, most variable and CpN specific preferences were observed. In case of CpG, a trend of +1 preferences was observed with C>T>A>G. For CpA methylation, the preferences for C(+1) and G(+1) were elevated and a C>G>T>A profile was observed, indicating a relative increase in the preference of CAG methylation. For CpT methylation, a very strong C(+1) preference was seen and CTC by far was the most preferred methylation motif (C>>G>A~T). In case of CpC methylation (C>T~A>>G), the

disfavor for G(+1) was elevated, which may be related to the fact that a CCG site generates a new CpG dinucleotide, which represents the preferred methylation target. Hence the second, and not the first cytosine in CCG sites will be methylated preferably. No noticeable changes of preferences for non-CpG methylation were observed at the +2 to +4 flanking positions except the general observation that effects were elevated. At the -1 site, C was slightly preferred over A for CpG methylation. In contrast, for all non-CpG activities, only A is preferred at this position indicating that CCG methylation is particularly favored. At the -2 site, the highly characteristic preference profile T>C>A, G was observed for CpG and non-CpG methylation, though effects were even more pronounced in case of non-CpG activity. No noticeable effects were observed at the -3 and -4 sites.

The -3 to +3 flanking sequence preferences of DNMT3A WT, DNMT3A mutants and DNMT3B WT (taken from Dukatz et al., 2020 (27)) are compiled in Figure 5. Comparison of all data reveals that WT DNMT3A and most mutants recognize the CpN site and +1 flank in a combined manner. This is clearly illustrated by the relative methylation of dinucleotides comprising the NX part of the CNX sequences (Figure 4C). By comparing these CNX preference patterns, several groups of mutants can be distinguished:

1) WT and some mutants showed small changes in preferences (S837A, S881A, and R887A).

2) R836A showed a preference for T(+1) in all methylation contexts, in particular at non-CpG sites.

3) N838A showed low methylation of CGA and CGG.

4) R882A, R882H, and L883A showed very similar patterns that partially overlap with DNMT3B. These profiles are characterized by a disfavor for C(+1) in all sequence contexts and the fact that CAG is strongly preferred.

More complex, CpN and mutant specific preference profiles are visible in the individual +3 to -3 flanking sequence preferences of the mutants shown in Figure 5, but here only the strongest effects will be highlighted. In case of R836A, the preference for C(+1) observed with WT was lost, and replaced by an increased preference for T(+1), in particular in non-CpG methylation. At the same time, the CAG preference observed for WT was lost as well. Moreover, A(+2) was more preferred for CpT methylation. N838A showed a specific drop in CGA and (weaker) CGG methylation, corresponding to a loss of the preference for A(+1) and G(+1) in CpG methylation, but its +1 flanking preferences are similar to WT for non-CpG methylation. With R882A, R882H, and L883A, G(+1) was strongly preferred in CpA methylation. Moreover, the strong C(+1) preference for WT in CpA and CpT methylation has been lost in these mutants and T(+1) is disfavored.

In summary, these data reveal strong flanking sequence preferences of DNMT3A in non-CpG methylation activity with a prominent preference for C(+1), which in RD loop mutants is altered to a G(+1) preference similarly as observed in DNMT3B.

Discussion

The DNA interaction of DNMT3A faces two critical challenges. First, target CpG sites must be identified and methylation should be mainly introduced at them. Of note, non-CpG methylation exists in the human genome and it has been connected with important functions, but it must be introduced in a controlled manner and aberrant methylation at arbitrary non-CpG sites must be prevented. One reason for this is that the C5-methyl group of thymine is a critical specificity determinant in protein-DNA interaction used by sequence specific DNA binding proteins to discriminate thymine from cytidine (28-30). Hence, aberrant cytosine methylation in a CpG or non-

CpG context has the ability to disturb the DNA interaction of transcription factors and MBD proteins at enhancers and promoters leading to altered gene activity patterns. While CpG methylation has been demonstrated to have strong effects on DNA binding of these proteins (31-35), the potential role of non-CpG methylation has not been studied systematically for most DNA binding proteins.

The second challenge for DNMT3A is that it interacts with a very small 2-base pair CpG "recognition sequence" that is embedded in a very variable sequence context in genomic DNA. It is known that the static and dynamic conformational properties of DNA are heavily modulated by the DNA sequence (29,36,37). For structural reasons, DNMTs interact with larger regions of the DNA leading to the challenge to establish an accurate positioning of the target CpG in the active site of the enzyme although the structural properties of different DNA substrate sequences vary. Previous work has shown that different DNMTs and also TET enzyme act on target sites embedded in different sequence contexts with different efficiency and strong preferences for some flanking contexts over other ones were reported (18,26,27,38). Flanking sequence preferences have been documented to affect cellular 5mC and 5hmC patterns and they could be connected with divergent biological roles of DNMT3A and DNMT3B and the pathogenicity of the R882H cancer mutation in DNMT3A (19). In this work, we have focused on the combined recognition and interaction of DNMT3A with the CpG guanine residue, the adjacent -3 to +3 inner flanking region and the -8 to +8 outer flanking sequence. We have mutated several DNA interacting residues and analyzed DNA methylation of WT DNMT3A and all mutants using libraries of substrates containing CpG and non-CpG sites in randomized flanking sequence context using the Deep Enzymology approach (19). Our data illustrate strong (>100 fold) flanking sequence effects and combined readout of the CpG guanine with the +1 to +3 flank sites which is further connected with readout of outer flanks (+4 to +8). In some cases, these effects were mutant specific, allowing us to identify roles of individual amino acid residues in the formation of the contact networks between DNMT3A and its target DNA.

CpG guanine recognition

In structural studies, R836 or N838 were found to contact the CpG guanine in a manner depending on the +1 flanking sequence (15,16). In addition, a water mediated contact to the N7-atom of the guanine is mediated by T834. In contrast, in DNMT3B, only the N779 and T775 mediated contacts were observed (corresponding to DNMT3A T834 and N838), while DNMT3B K777 (corresponding to R836) adopts an orientation pointing away from the CpG guanine (18). Based on our activity measurements in all possible flanking contexts, both enzymes qualitatively exhibit similar preferences for CpN sites (CpG>>CpA>CpT>CpC). The preference for a CpA as second-best target site could be due to the highly conserved water-mediated contact between T834/T775 and the N7 atom of the CpG guanine, that could be equally formed with adenine. This interpretation is supported by the finding that a T775A mutation in DNMT3B caused an almost complete loss of activity (27).

However, the relative non-CpG activity of DNMT3B was stronger than that of DNMT3A, which may be related to the lack of the possibility to form an arginine-guanine contact in DNMT3B. An important role of R836 for CpG guanine recognition by DNMT3A is further supported by our finding that mutation of R836 led to a pronounced reduction of CpG specificity, while mutation of N838 even led to an increase in CpG recognition. Our observation that the DNMT3A R836A mutant retains a considerable CpG specificity suggests that N838 can partially take over the role of R836 in CpG recognition, when R836 is mutated. The finding of an increased CpG specificity of N838A can be explained by a similar mechanism, proposing that R836 may use the space emptied by the N838A mutation, and this might strengthen its contact to the CpG guanine and thereby improve CpG recognition. The reduced non-CpG activities of R882A, R882H, L883A and R887A can be understood,

because weakening of the DNA contacts by the RD loop through these mutations might increase the requirements for a correct positioning of the TRD loop, leading to more stringent readout of the CpG guanine.

Flanking sequence preferences in CpG and non-CpG methylation

Our data show that DNMT3A has pronounced preferences for inner flanking sequences which lead to >100-fold changes in methylation rates, because in the same library several NNCGNNN flanks showed zero methylation when the most preferred sequence context was already 95% methylated. One general result of the non-CpG flanking sequence analyses of DNMT3A WT and mutants was that flanking effects are elevated on disfavored CpN sites, again indicating that difficult substrates require a supportive flanking context for efficient methylation. One interesting observation with WT DNMT3A was that the C(-1) preference was high for CpG methylation, while in case of all other methylation contexts, A(-1) was preferred. This effect cannot be further interpreted as the mutational study conducted here did not cover the amino acid residues putatively involved in -1 flank interaction. Another very strong effect was that CpT methylation had a very pronounced preference for C(+1) meaning that CTC methylation is highly preferred.

DNA interaction of TRD loop residues R836 and N838

R836 is a key residue in CpG guanine and flanking interaction of DNMT3A. Structural studies detected a striking movement of R836, which contacts the CpG guanine in a CGT structure, but the +1 and +2 flank in a CGA structure. Our data revealed several biochemical effects that are connected to R836. 1) R836A showed a strong change of flanking sequence preferences at the +1 side, where the preference for C(+1) is lost, which is highly characteristic for DNMT3A non-CpG methylation and also has been observed in cells (39). DNA methylation in CpA context has been observed in neurons and it has been shown to have very important biological roles (10,11). Hence, R836 may contact C(+1) in particular on non-CpG target sites where the CpG guanine interaction partner for R836 is absent. 2) R836A showed a strong decrease in G(+1) preference of CpA methylation indicating that this residue stimulates methylation of CAG sites. These effect may be related to the contact of R836 to the N7 atom of A(+1) that is observed in the CGA structure and may be similarly possible with CAG. 3) R836A shows a strong preference for T(+1) in particular in non-CpG methylation which is specific for this mutant. This effect might indicate that the mutated A836 could engage in a van der Waals contact with the methyl group of T(+1) in non-CpG sites. 4) In CpG methylation, R836A showed a specific effect on the +2/+3 flank recognition in the context of G(+1), suggesting that R836 adopts a specific conformation on CGG complexes, which may be related to the very low preference for this sequence. This effect may be related to the ability of R836 to contact G(+1) and also the +2 base by hydrogen bonds, thereby affecting the +2 site interaction on CGG sites. 5) The influence of the folded back conformation of R836 on the +2 and +3 flank interaction of the entire RD loop is also visible in the altered CpG +2/+3 dinucleotide recognition pattern of all TRD loop mutants in the context of A(+1). Based on this, one can conclude that the TRD loop mediates interactions with A(+1) that depend on the correct conformation of R836, S837 and N838.

N838 is the second most important residues for CpG recognition and flank interactions of DNMT3A. In the DNMT3A CGA structure, the side chain of N838 occupies the place of R836 and it contacts the CpG guanine, while in the CGT structure, it forms an H-bond to the +2/+3 phosphodiester group and the +2' base. The N838A exchange led to strong changes in the flanking sequence preferences at the +1 position, which were distinct from R836A. In particular, the N383A mutant showed a very strong

and unique drop of A(+1) and G(+1) preference in CpG (but not in non-CpG) methylation. This effect can be explained by the combined contact of N838 to the CpG guanine and the A(+1) N7 atom seen in the CGA structure. This contact could equally be formed with G(+1) suggesting that N838 is necessary for efficient CGA and CGG methylation. Hence, N838 functions to buffer the effect of R836 preventing an "overshooting" of the C(+1) preference. In summary, R836 and N838 interact with the CpG guanine and flanking sites in a flexible manner ensuring high activity and specificity of DNMT3A in diverse flanking sequence contexts.

Roles of RD loop residues L883 and R882

One of the most striking observations of this study was the very strong rate enhancement caused by the L883A exchange. This finding can be explained in the context of the DNMT3A structure, showing that the hydrophobic L883 side chain is placed at an exposed and hydrophilic position pointing towards the DNA. Nevertheless, this residue is conserved in all DNMT3A enzymes, indicating that it must have an important biological role. Our data show that L883 is required to stabilize the DNMT3A specific conformation of the RD loop, because the L883A mutation caused similar changes in flanking preferences as R882A and R882H. The +1 to +3 flanking sequence preferences of R882A, R882H, and L883A in CpG and non-CpG methylation showed similarity to DNMT3B, including the preference for G(+1) and A(+1) in CpG methylation, G(+1) in non-CpG methylation and absence of the DNMT3A-characteristic C(+1) and C(+2) preference in CpG, and even more in non-CpG methylation. These changes convert the DNMT3A type preference profile for the +1 site (C>T>A>G) into a DNMT3B-like profile (G>A>C>T). Our data indicate that the RD loop residues are needed to stabilize its DNMT3A specific conformation, allowing for several DNMT3A-specific DNA contacts to the flanking sequence. This function apparently is so important that a reduction in activity by the residue is tolerated in evolution.

Disruption of the DNMT3A-specific DNA contacts of the RD loop leads to flanking sequence preference profiles that are DNMT3B-like and this process is one of the important carcinogenic mechanisms of R882H (21-23). However, the prominent and specific role of the R882H cancer mutation is also due to two more unrelated effects. 1) H882 in the R882H mutant has been shown to form new interface contacts leading to a preferred formation of mutant/mutant RD interfaces, which finally causes a dominant effect of the R882H mutation (24). 2) The R882H mutation arises from the deamination of a methylated cytosine in a CpG site, leading to a high mutational pressure at this site.

Effects of the outer flanks and DNA bending

We demonstrate here that the outer flank sequences also showed up to 8-fold and inner flank dependent effects on methylation rates indicating that long range effects connect DNA flanking sequence and the active sites of DNMT3A. Overall, DNMT3A showed a preference for A/T bases in the +4 to +8 outer flank region. This finding supports results of a previous work, where we showed that A and T bases in this region are preferable for co-methylation of CpG sites on a two-site substrate (17). It was structurally interpreted in the context of the DNA bending required for the interaction of two CpG sites with the two acitve centers at the RD-interface of the DNMT3A tetramer leading to co-methylation of both CpG sites. Our data with the single CpG site substrates presented here reveal similar preferences which might suggest that these substrates also adopt a bent conformation when being bound to DNMT3A, although they are lacking a second CpG site. Surprisingly, this effect was lost or reduced with almost all mutants. This may be due to the fact that the mutations weaken the DNA interaction, making the combined DNA interaction of both active

subunit with the DNA less favorable. Instead, in case of the mutants, only one subunit of DNMT3A might contact the DNA at any time. This would remove the requirement for DNA bending and hence reduce the need to have A/T bases in the +4 to +8 flanking region. However, there was one exception to this rule, because T(+6) was preferred by all mutants, suggesting a more specific role of this base.

Conclusion and outlook

Our experimental data document the combined and interdependent readout of the CpG site and flanking residues up to 8 base pairs away from it. By this, the enzyme generates networks of interactions that are necessary for the efficient methylation of CpG sites embedded into different flanking sites. In general, we observed that a substrate containing an unfavorable feature (e.g. a non-CpG target site or an unfavored inner flank) shows elevated flanking preferences for the remaining part of the substrate. Mutational analyses connected the CpG guanine and +1 to +8 flank DNA interaction contact networks mainly with R836 and N838 from the TRD loop and R882 and L883 from the RD loop.

DNMT3A and DNMT3B arose in the whole genome duplication that occurs before the origin of vertebrates estimated at 500 myo (40). Appearance of two DNMT3 paralogs apparently has allowed for an improvement of DNA methylation machinery by specialization of the two DNMT3 genes that led to the preservation of both paralogs in further evolution. Specialization of DNMT3A and DNMT3B include their regulation and targeting by interaction with distinct sets of transcription factors (41) and divergence of their chromatin interaction, e.g. providing the PWWP domain of DNMT3A a preference for H3K36me2, while the PWWP domain of DNMT3B prefers H3K36me3 (42,43). Diverging flanking sequence preferences of DNMT3A and DNMT3B represent another factor of evolutionary specialization of DNMT3 enzymes. Different flanking sequence preferences of DNMT3A and DNMT3B have been shown to explain the specific role of DNMT3B for methylation of SATII sequences (18), leading to the connection of DNMT3B with the ICF syndrome (44,45). Based on structural studies, the RD loop of both proteins was proposed to contribute to these effects (18). Here we show experimentally the strong influence of the RD loop conformation on DNMT3A specific flanking sequence preferences. Our observation that DNMT3A mutants with different mutations in the RD loop exhibit flanking sequence preferences similar to DNMT3B illustrates the key role of this structural element for the DNMT3A/DNMT3B divergence and extends earlier observations for R882H (22,23). In addition, our data illustrate an important role of the TRD loop in the flanking sequence interactions as well, in particular of R836 and N838. This provides another source of mechanistic divergence between DNMT3A and DNMT3B, because R836 is replaced by K777 in DNMT3B which has a different contact potential than arginine and, for example, has been shown to be responsible for the DNMT3B specific preference for G(+1) in non-CpG methylation (18,27), while we show here that R836 mediates the C(+1) preference in non-CpG methylation of DNMT3A. One role of N838 is to balance this effect of R836 by supporting CGA and CGG methylation, to avoid a too strong preference for C(+1).

We consider it very plausible that flanking sequence preferences would be observed with other DNA interaction enzymes and DNA binding proteins, once experimental studies tailored to observe such effects had been conducted. Unfortunately, due to the large number of different enzyme-DNA complexes that need to be considered, a structural analysis of the detailed effects discovered here is far beyond the current state of structural analysis or molecular dynamic simulations. Hence, our data illustrate the clear demand for improved methods to allow a mechanistic understanding of the basic processes determining the interaction of DNMT3A with its DNA substrate. Improved methods should allow the parallel investigation of several protein-DNA complexes for example by automated

molecular dynamics simulations using force fields that correctly describe the structural parameters of the DNA and the energetics of its protein interaction. At the same time, simulation times must be long enough to capture the specific conformational changes that accompany complex formation.

Experimental procedures

Mutagenesis and protein expression

The catalytic, C-terminal domain of DNMT3A (residues 612-912 of Q9Y6K1) and its mutants were cloned as MBP-tagged fusion proteins. Protein expression and purification was conducted essentially as described (24,25). Protein overexpression was carried out in *E. coli* BL21 (DE3) Codon Plus RIL cells transformed with the corresponding plasmids. If needed, cleavage of the MBP tag was performed as described (24). Mutagenesis was conducted as described (27). The purity of the preparations was estimated to be >95% from Coomassie stained SDS gels. The concentrations of the proteins were determined by UV spectrophotometry and confirmed by densitometric analysis of Coomassie BB stained SDS–polyacrylamide gels.

Radioactive DNA methylation kinetics

Methylation activities of DNMT3A WT and mutants were determined using an avidin-biotin methylation plate assay using a biotinylated double-stranded 30-mer oligonucleotide with a single CpG site (GAA GCT GGG ACT TCCGGG AGG AGA GTG CAA) basically as described (24,27). Methylation reactions with DNMT3A were conducted at 37 °C with 2 µM enzyme in methylation buffer (20 mM HEPES pH 7.5, 1 mM EDTA, 50 mM KCl, 0.25 mg/mL bovine serum albumin) in the presence of 1 µM of the biotinylated substrate. The reactions were started by adding 0.76 µM radioactively labelled AdoMet (Perkin Elmer). The initial slope of the enzymatic reaction was determined by linear regression.

Flanking sequence preference analysis with a randomized substrate and bioinformatics analysis

The preparation of the substrate with CpN site in a 10 base pair randomized sequence context, methylation, bisulfite conversion and library preparation was conducted as described (18,26). In addition, the effect of the outer flanks was tested using five substrates with fixed -3 to +3 inner flanks and randomized -10 to -4 and +4 to +10 outer flanks, which were selected to cover the entire range of preferred (low ranks) to disfavored inner flanks (high ranks):

- PB955: inner flank GTACGTCA, rank 57
- MD221: inner flank CTACGGCA, rank 112
- MD222: inner flank GTCCGCGA, rank 564
- PB954: inner flank ATTCGATG, rank 3814
- PB956: inner flank TGCCGTTG, rank 3883

Library methylation was conducted at 37°C as described (18,26) using 5 ng/µl of the library in buffer containing 20 mM HEPES pH 7.5, 1 mM EDTA, 50 mM KCl, 100 µg/ml BSA, and 0.8 mM AdoMet (Sigma). Enzyme concentrations (0.5-5 µM) and incubation times (30 -120 min) were as indicated. The methylation reactions were stopped by shock freezing in liquid nitrogen, then treated with proteinase

K (NEB) for 2 hours at 42 °C, and purified by PCR Clean-up kit (Macherey-Nagel). Afterwards, bisulfite conversion and NGS library preparation were conducted as described (18,24,26). Different sets of barcodes were introduced in the PCR steps to distinguish different samples and experiments. NGS data analysis was conducted basically as described (27).

Data Availability

NGS kinetic raw data will be available at DaRUS at https://doi.org/10.18419/darus-2993

Anonymous review access can be found at:

https://darus.uni-stuttgart.de/privateurl.xhtml?token=9329d2ef-4945-4237-b959-67a9b308a81a

Biochemical data are provided with this paper.

Funding

This work has been supported by the Deutsche Forschungsgemeinschaft (JE 252/10 and JE252/36).

Author contributions

AJ and MD devised the study. MD performed the experimental work with support from MD and ES. MD, SA and PB conducted the NGS analyses. MD, SA, PB and AJ were involved in NGS data analysis and bioinformatic analyses. AM prepared the MSA. AJ prepared the figures and manuscript draft. All authors were involved in final data analysis and interpretation and writing of the manuscript. All authors approved the final version of the manuscript.

Competing Interest

The authors declare no competing interests.

References

1. Schubeler, D. (2015) Function and information content of DNA methylation. *Nature* **517**, 321-326
2. He, Y., and Ecker, J. R. (2015) Non-CG Methylation in the Human Genome. *Annu Rev Genomics Hum Genet* **16**, 55-77
3. Jeltsch, A., Broche, J., and Bashtrykov, P. (2019) Molecular Processes Connecting DNA Methylation Patterns with DNA Methyltransferases and Histone Modifications in Mammalian Genomes. *Genes* **10**
4. Bergman, Y., and Cedar, H. (2013) DNA methylation dynamics in health and disease. *Nature structural & molecular biology* **20**, 274-281
5. Zeng, Y., and Chen, T. (2019) DNA Methylation Reprogramming during Mammalian Development. *Genes* **10**
6. Gowher, H., and Jeltsch, A. (2018) Mammalian DNA methyltransferases: new discoveries and open questions. *Biochemical Society transactions* **46**, 1191-1202
7. Chen, Z., and Zhang, Y. (2019) Role of Mammalian DNA Methyltransferases in Development. *Annual review of biochemistry*
8. Yang, L., Rau, R., and Goodell, M. A. (2015) DNMT3A in haematological malignancies. *Nat Rev Cancer* **15**, 152-165
9. Hamidi, T., Singh, A. K., and Chen, T. (2015) Genetic alterations of DNA methylation machinery in human diseases. *Epigenomics* **7**, 247-265
10. Kinde, B., Gabel, H. W., Gilbert, C. S., Griffith, E. C., and Greenberg, M. E. (2015) Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. *Proc Natl Acad Sci U S A* **112**, 6800-6806
11. Christian, D. L., Wu, D. Y., Martin, J. R., Moore, J. R., Liu, Y. R., Clemens, A. W., Nettles, S. A., Kirkland, N. M., Papouin, T., Hill, C. A., Wozniak, D. F., Dougherty, J. D., and Gabel, H. W. (2020) DNMT3A Haploinsufficiency Results in Behavioral Deficits and Global Epigenomic Dysregulation Shared across Neurodevelopmental Disorders. *Cell Rep* **33**, 108416
12. Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A., and Cheng, X. (2007) Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449**, 248-251
13. Jurkowska, R. Z., Anspach, N., Urbanke, C., Jia, D., Reinhardt, R., Nellen, W., Cheng, X., and Jeltsch, A. (2008) Formation of nucleoprotein filaments by mammalian DNA methyltransferase Dnmt3a in complex with regulator Dnmt3L. *Nucleic Acids Res* **36**, 6656-6663
14. Jurkowska, R. Z., Rajavelu, A., Anspach, N., Urbanke, C., Jankevicius, G., Ragozin, S., Nellen, W., and Jeltsch, A. (2011) Oligomerization and binding of the Dnmt3a DNA methyltransferase to parallel DNA molecules: heterochromatic localization and role of Dnmt3L. *J Biol Chem* **286**, 24200-24207
15. Zhang, Z. M., Lu, R., Wang, P., Yu, Y., Chen, D., Gao, L., Liu, S., Ji, D., Rothbart, S. B., Wang, Y., Wang, G. G., and Song, J. (2018) Structural basis for DNMT3A-mediated de novo DNA methylation. *Nature* **554**, 387-391
16. Anteneh, H., Fang, J., and Song, J. (2020) Structural basis for impairment of DNA methylation by the DNMT3A R882H mutation. *Nature communications* **11**, 2294
17. Emperle, M., Bangalore, D. M., Adam, S., Kunert, S., Heil, H. S., Heinze, K. G., Bashtrykov, P., Tessmer, I., and Jeltsch, A. (2021) Structural and biochemical insight into the mechanism of dual CpG site binding and methylation by the DNMT3A DNA methyltransferase. *Nucleic Acids Res* **49**, 8294-8308
18. Gao, L., Emperle, M., Guo, Y., Grimm, S. A., Ren, W., Adam, S., Uryu, H., Zhang, Z. M., Chen, D., Yin, J., Dukatz, M., Anteneh, H., Jurkowska, R. Z., Lu, J., Wang, Y., Bashtrykov, P., Wade, P. A., Wang, G. G., Jeltsch, A., and Song, J. (2020) Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. *Nature communications* **11**, 3355

19. Jeltsch, A., Adam, S., Dukatz, M., Emperle, M., and Bashtrykov, P. (2021) Deep enzymology studies on DNA methyltransferases reveal novel connections between flanking sequences and enzyme activity. *J Mol Biol*, 167186
20. Guo, J. U., Su, Y., Shin, J. H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., Zhu, H., Chang, Q., Gao, Y., Ming, G. L., and Song, H. (2014) Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat Neurosci* **17**, 215-222
21. Emperle, M., Rajavelu, A., Kunert, S., Arimondo, P. B., Reinhardt, R., Jurkowska, R. Z., and Jeltsch, A. (2018) The DNMT3A R882H mutant displays altered flanking sequence preferences. *Nucleic Acids Res* **46**, 3130-3139
22. Emperle, M., Adam, S., Kunert, S., Dukatz, M., Baude, A., Plass, C., Rathert, P., Bashtrykov, P., and Jeltsch, A. (2019) Mutations of R882 change flanking sequence preferences of the DNA methyltransferase DNMT3A and cellular methylation patterns. *Nucleic Acids Res* **47**, 11355-11367
23. Norvil, A. B., AlAbdi, L., Liu, B., Tu, Y. H., Forstoffer, N. E., Michie, A. R., Chen, T., and Gowher, H. (2020) The acute myeloid leukemia variant DNMT3A Arg882His is a DNMT3B-like enzyme. *Nucleic Acids Res* **48**, 3761-3775
24. Mack, A., Emperle, M., Schnee, P., Adam, S., Pleiss, J., Bashtrykov, P., and Jeltsch, A. (2022) Preferential Self-interaction of DNA Methyltransferase DNMT3A Subunits Containing the R882H Cancer Mutation Leads to Dominant Changes of Flanking Sequence Preferences. *J Mol Biol* **434**, 167482
25. Emperle, M., Dukatz, M., Kunert, S., Holzer, K., Rajavelu, A., Jurkowska, R. Z., and Jeltsch, A. (2018) The DNMT3A R882H mutation does not cause dominant negative effects in purified mixed DNMT3A/R882H complexes. *Scientific reports* **8**, 13242
26. Adam, S., Anteneh, H., Hornisch, M., Wagner, V., Lu, J., Radde, N. E., Bashtrykov, P., Song, J., and Jeltsch, A. (2020) DNA sequence-dependent activity and base flipping mechanisms of DNMT1 regulate genome-wide DNA methylation. *Nature communications* **11**, 3723
27. Dukatz, M., Adam, S., Biswal, M., Song, J., Bashtrykov, P., and Jeltsch, A. (2020) Complex DNA sequence readout mechanisms of the DNMT3B DNA methyltransferase. *Nucleic Acids Res* **48**, 11495-11509
28. Garvie, C. W., and Wolberger, C. (2001) Recognition of specific DNA sequences. *Molecular cell* **8**, 937-946
29. Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., and Mann, R. S. (2010) Origins of specificity in protein-DNA recognition. *Annual review of biochemistry* **79**, 233-269
30. Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordan, R., and Rohs, R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences* **39**, 381-399
31. Maurano, M. T., Wang, H., John, S., Shafer, A., Canfield, T., Lee, K., and Stamatoyannopoulos, J. A. (2015) Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Rep* **12**, 1184-1195
32. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schubeler, D., Vinson, C., and Taipale, J. (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**
33. Kribelbauer, J. F., Laptenko, O., Chen, S., Martini, G. D., Freed-Pastor, W. A., Prives, C., Mann, R. S., and Bussemaker, H. J. (2017) Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes. *Cell Rep* **19**, 2383-2395
34. Shimbo, T., and Wade, P. A. (2016) Proteins That Read DNA Methylation. *Adv Exp Med Biol* **945**, 303-320
35. Liu, K., Xu, C., Lei, M., Yang, A., Loppnau, P., Hughes, T. R., and Min, J. (2018) Structural basis for the ability of MBD domains to bind methyl-CG and TG sites in DNA. *J Biol Chem* **293**, 7344-7354
36. Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. (2009) The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248-1253

37. Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H. J., Rohs, R., and Mann, R. S. (2015) Deconvolving the recognition of DNA shape from sequence. *Cell* **161**, 307-318
38. Adam, S., Bracker, J., Klingel, V., Osteresch, B., Radde, N. E., Brockmeyer, J., Bashtrykov, P., and Jeltsch, A. (2022) Flanking sequences influence the activity of TET1 and TET2 methylcytosine dioxygenases and affect genomic 5hmC patterns. *Commun Biol* **5**, 92
39. Lee, J. H., Park, S. J., and Nakai, K. (2017) Differential landscape of non-CpG methylation in embryonic stem cells and neurons caused by DNMT3s. *Scientific reports* **7**, 11295
40. de Mendoza, A., Poppe, D., Buckberry, S., Pflueger, J., Albertin, C. B., Daish, T., Bertrand, S., de la Calle-Mustienes, E., Gomez-Skarmeta, J. L., Nery, J. R., Ecker, J. R., Baer, B., Ragsdale, C. W., Grutzner, F., Escriva, H., Venkatesh, B., Bogdanovic, O., and Lister, R. (2021) The emergence of the brain non-CpG methylation system in vertebrates. *Nat Ecol Evol* **5**, 369-378
41. Jeltsch, A., and Jurkowska, R. Z. (2016) Allosteric control of mammalian DNA methyltransferases - a new regulatory paradigm. *Nucleic Acids Res* **44**, 8556-8575
42. Baubec, T., Colombo, D. F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A. R., Akalin, A., and Schubeler, D. (2015) Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature* **520**, 243-247
43. Weinberg, D. N., Papillon-Cavanagh, S., Chen, H., Yue, Y., Chen, X., Rajagopalan, K. N., Horth, C., McGuire, J. T., Xu, X., Nikbakht, H., Lemiesz, A. E., Marchione, D. M., Marunde, M. R., Meiners, M. J., Cheek, M. A., Keogh, M. C., Bareke, E., Djedid, A., Harutyunyan, A. S., Jabado, N., Garcia, B. A., Li, H., Allis, C. D., Majewski, J., and Lu, C. (2019) The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature* **573**, 281-286
44. Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-257
45. Xu, G. L., Bestor, T. H., Bourc'his, D., Hsieh, C. L., Tommerup, N., Bugge, M., Hulten, M., Qu, X., Russo, J. J., and Viegas-Pequignot, E. (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402**, 187-191

Figure legends

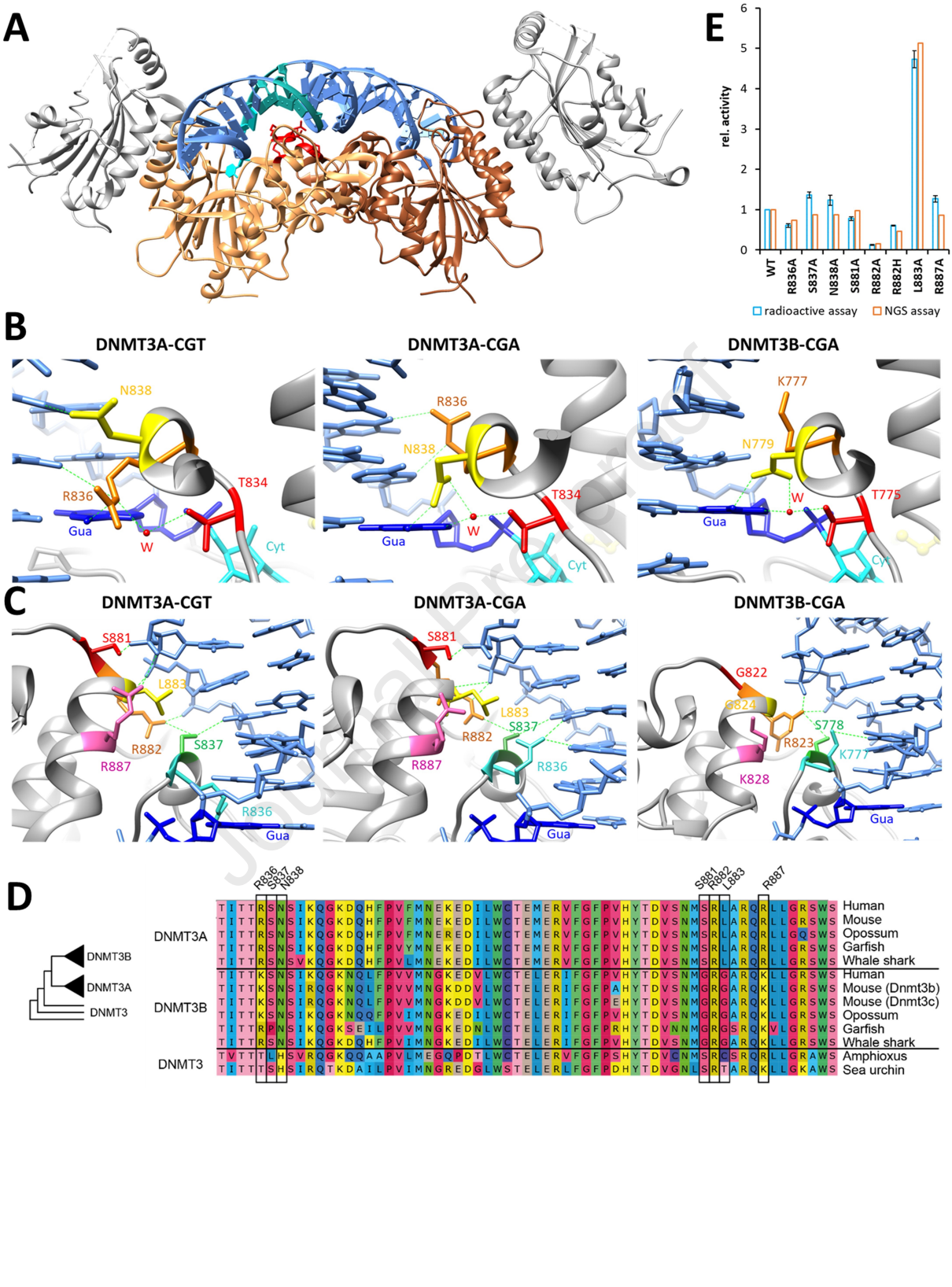
Figure 1. Structure and activity of DNMT3A complexes. A) Structure of the DNMT3A/3L heterotetramer (15). DNMT3L is shown in light and dark grey, DNMT3A in orange and light orange. The DNA is colored blue. The flipped cytosine and the CpG guanine are colored cyan, the +1 to +3 flank base pairs are colored turquoise. The mutated residues are indicated in red in the light orange DNMT3A subunit. **B)** Structural snapshots of the TRD loop regions in DNMT3A (CGT complex 6BRR (15) and CGA complex 6W8B (16)) and DNMT3B (CGT complex 6U8W (18)). The flipped cytosine is shown in cyan and the CpG guanine in dark blue. W, water. **C)** Structural snapshots of the RD loop regions in DNMT3A and DNMT3B. The flipped cytosine is shown in cyan and the CpG guanine in dark blue. **D)** Multiple sequence alignment of DNMT3A and DNMT3B in representative vertebrate species and Amphioxus as an example of a closely related non-vertebrate. **E)** Activity analysis of the mutants investigated in this study with radioactive assays or the Deep Enzymology NGS assay. NGS activity of R882A is based on the MBP-cleaved data and WT refers to the activity of MBP-cleaved WT taken from Mack et al., 2022 (24). NGS activity of R882H is based on the data from Emperle et al. (2019) (22).

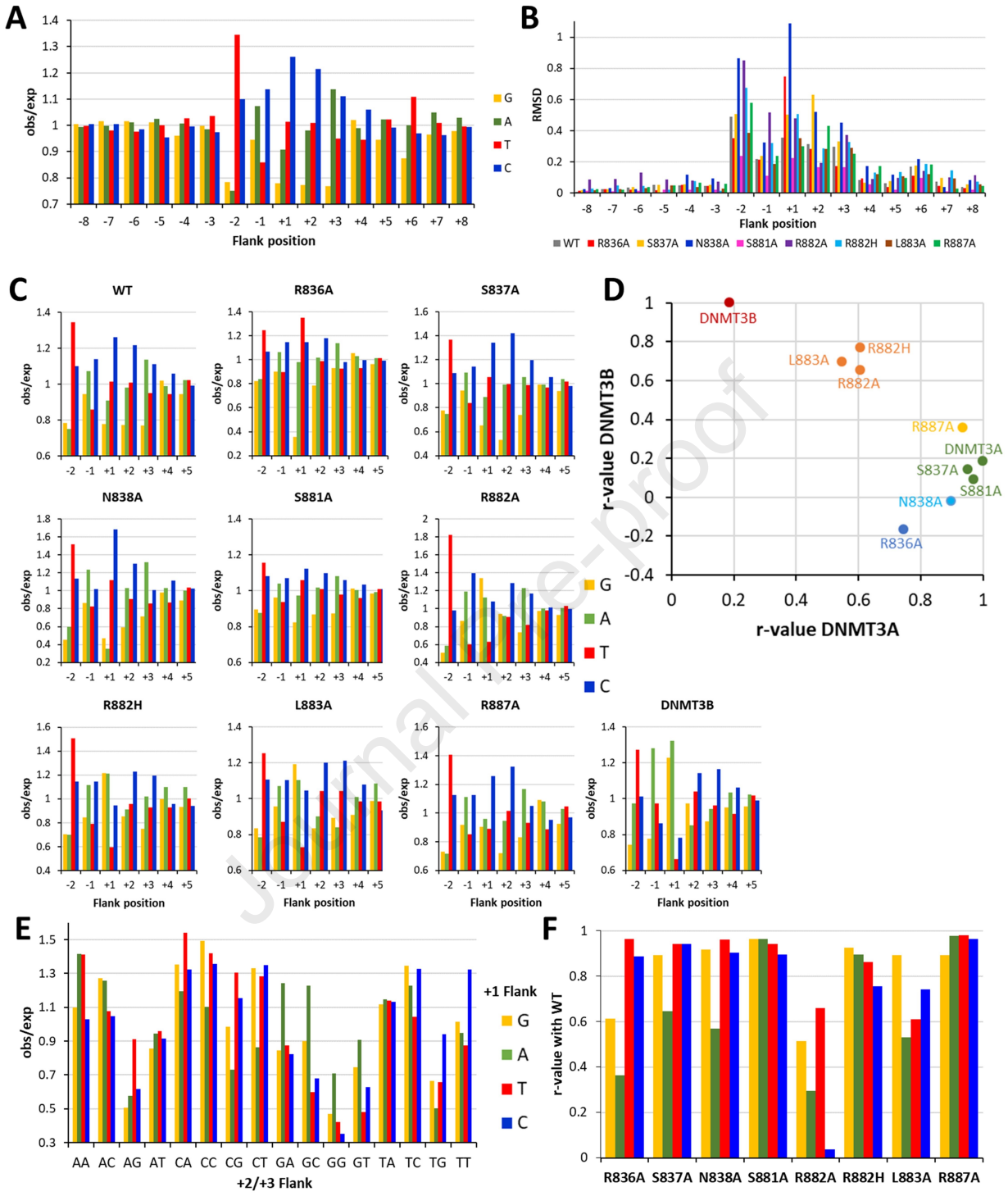
Figure 2. Analysis of flanking sequence effects in CpG methylation. A) Flanking sequence preferences of WT DNMT3A for the -8 to +8 region. **B)** Compilation of RSMD values of flanking sequence effects of WT DNMT3A and mutants at each flank position. **C)** Flanking sequence preferences of WT and DNMT3A mutants in the -2 to +3 flank region. **D)** Compilation of the r-values of the correlation of mutant flanking sequence preferences with WT DNMT3A and DNMT3B. **E)** +2/+3 flanking preferences in dependence of the base at the +1 flank site of WT DNMT3A. Mutant profiles are shown in Supplemental Figure 4. **F)** Correlation of mutant +2/+3 flank preferences with WT DNMT3A for different bases at the +1 position.

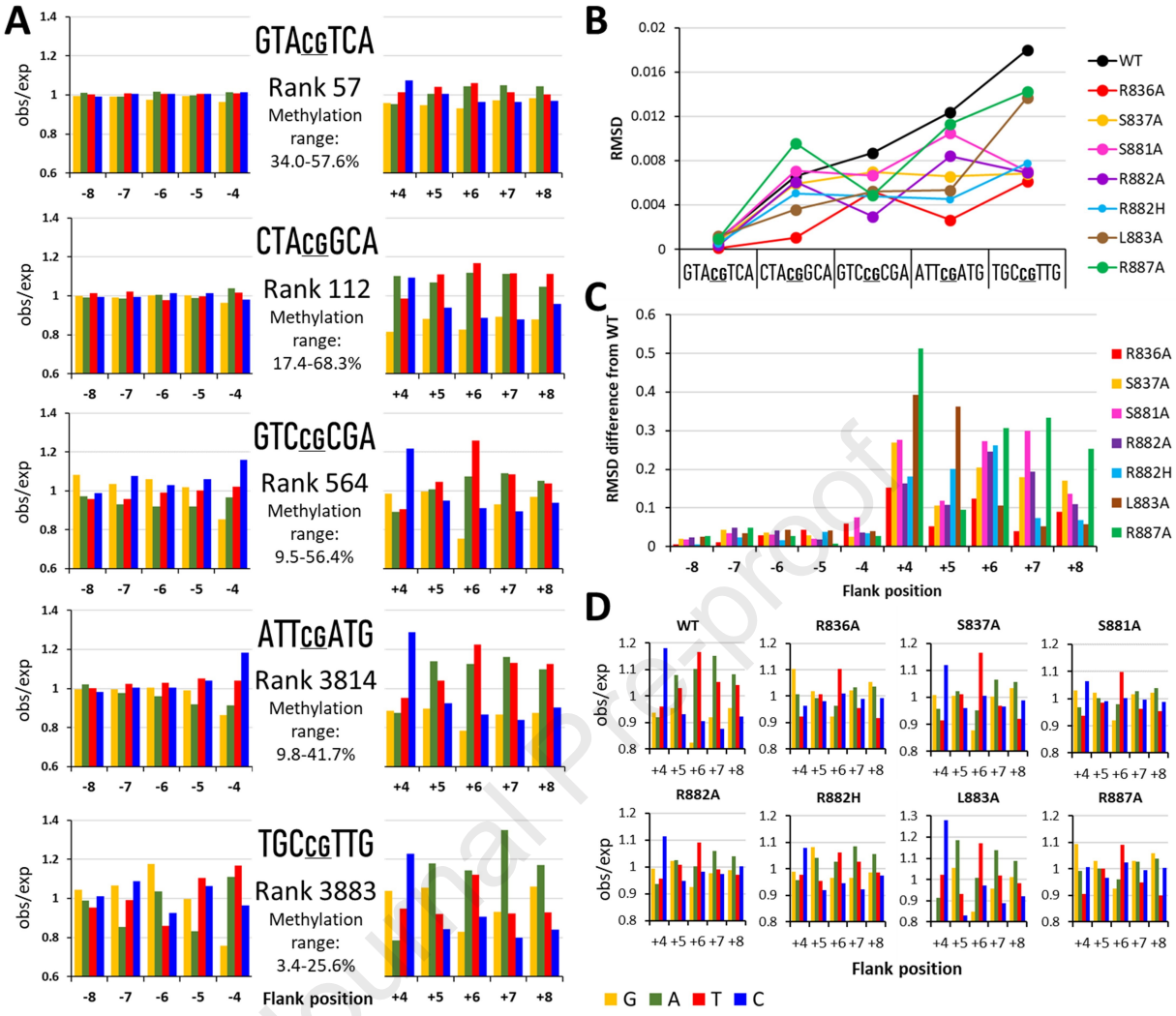
Figure 3. Outer flank effect on DNMT3A activity. A) Effects of the -8 to +8 flanks on the methylation of five substrates with fixed -3 to +3 inner flank with DNMT3A WT. Rank refers to the preference of the corresponding inner flank where small numbers indicate high preference. Methylation range refers to the combined sequences of the -4, +4, +5 and +6 flank sites. **B)** RMSD values of the deviation of the flanking profiles averaged for all -8 to -4 and +4 to +8 sites. **C)** RMSD values of the deviation of the flanking profiles of WT and mutants averaged for all five substrates at the different flanking sites. **D)** +4 to +8 outer flank sequence preferences for WT and mutants averaged for all five substrates.

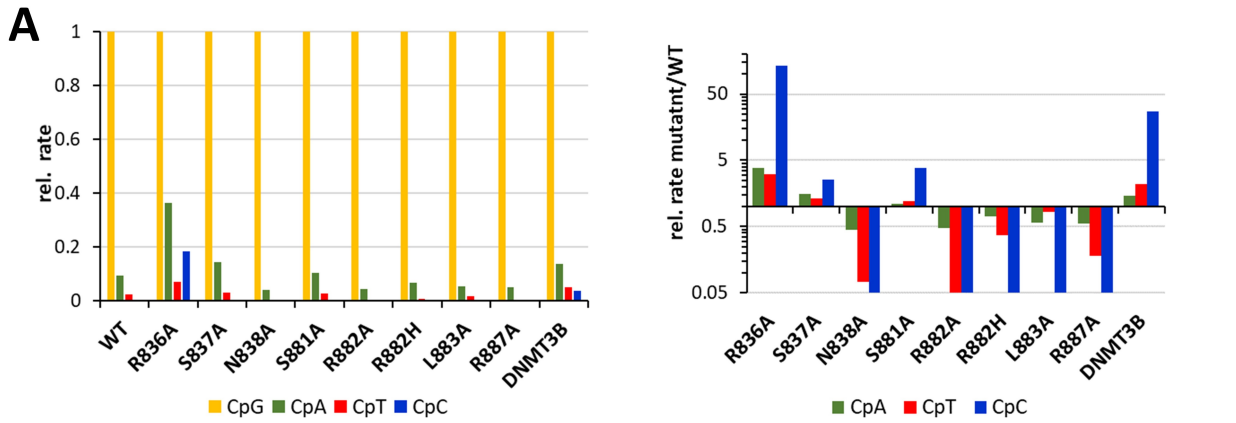
Figure 4. Non-CpG methylation of WT and mutants. A) Relative non-CpG methylation rates of DNMT3A WT and mutants. **B)** Flanking sequence preferences of WT DNMT3A for non-CpG methylation. Relative methylation rates of 0.05 correspond to the limit of detection. **C)** Combined readout of the CpN base and flank position +1. Grey shaded experiments could not be analyzed due to the low methylation levels.

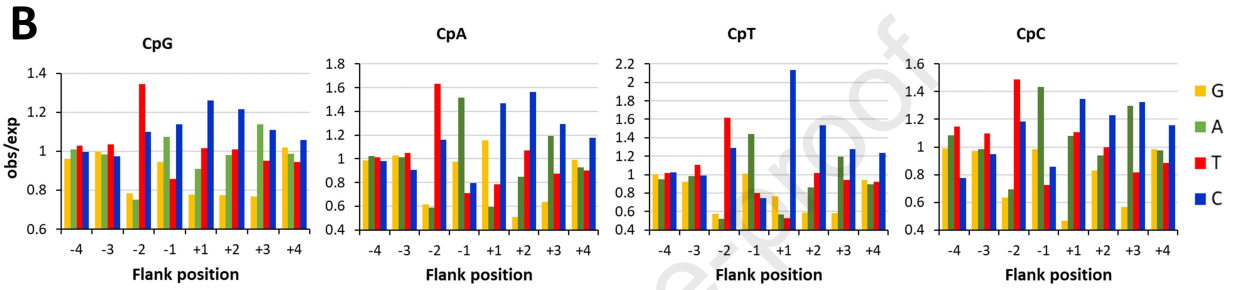
Figure 5. Flanking sequence preferences of non-CpG methylation of WT and mutants. The -3 to +3 flanking preference profiles are shown. Some experiments could not be analyzed due to the low methylation levels (n.d.).


A
E
rel. activity
WT
R836A
S837A
N838A
S881A
R882A
R882H
L883A
R887A
radioactive assay
NGS assay
B
DNMT3A-CGT
DNMT3A-CGA
DNMT3B-CGA
N838
R836
T834
Gua
W
Cyt
K777
N779
T775
C
S881
L883
R882
S837
R887
G822
G824
S778
R823
K777
K828
D
R836
S837
N838
S881
R882
L883
R887
DNMT3B
DNMT3A
DNMT3
Human
Mouse
Opossum
Garfish
Whale shark
Mouse (Dnmt3b)
Mouse (Dnmt3c)
Amphioxus
Sea urchin









C

WT

	G	A	T	C
G	0.942	0.965	1.006	1.087
A	1.153	0.594	0.785	1.468
T	0.769	0.57	0.527	2.134
C	0.467	1.081	1.106	1.346

R836A

	G	A	T	C
G	0.724	1.028	1.16	1.088
A	0.543	0.695	1.727	1.034
T	0.479	0.812	1.964	0.745
C	0.046	0.358	2.849	0.746

S837A

	G	A	T	C
G	0.921	0.965	1.021	1.093
A	0.986	0.634	0.885	1.495
T	0.625	0.47	0.487	2.417
C	0.51	0.76	0.947	1.783

N838A

	G	A	T	C
G	0.532	0.399	1.144	1.579
A	1.143	0.609	0.874	1.336
T	0.865	0.815	0.809	1.547
C				

S881A

	G	A	T	C
G	0.944	0.99	1.025	1.04
A	0.9	0.568	0.902	1.63
T	0.55	0.419	0.628	2.402
C	0.702	0.787	0.976	1.535

R882A

	G	A	T	C
G	1.232	1.102	0.704	0.962
A	1.661	0.854	0.558	0.926
T				
C				

R882H

	G	A	T	C
G	1.263	1.16	0.698	0.878
A	1.726	0.887	0.581	0.805
T	1.269	1.129	0.645	0.958
C				

L883A

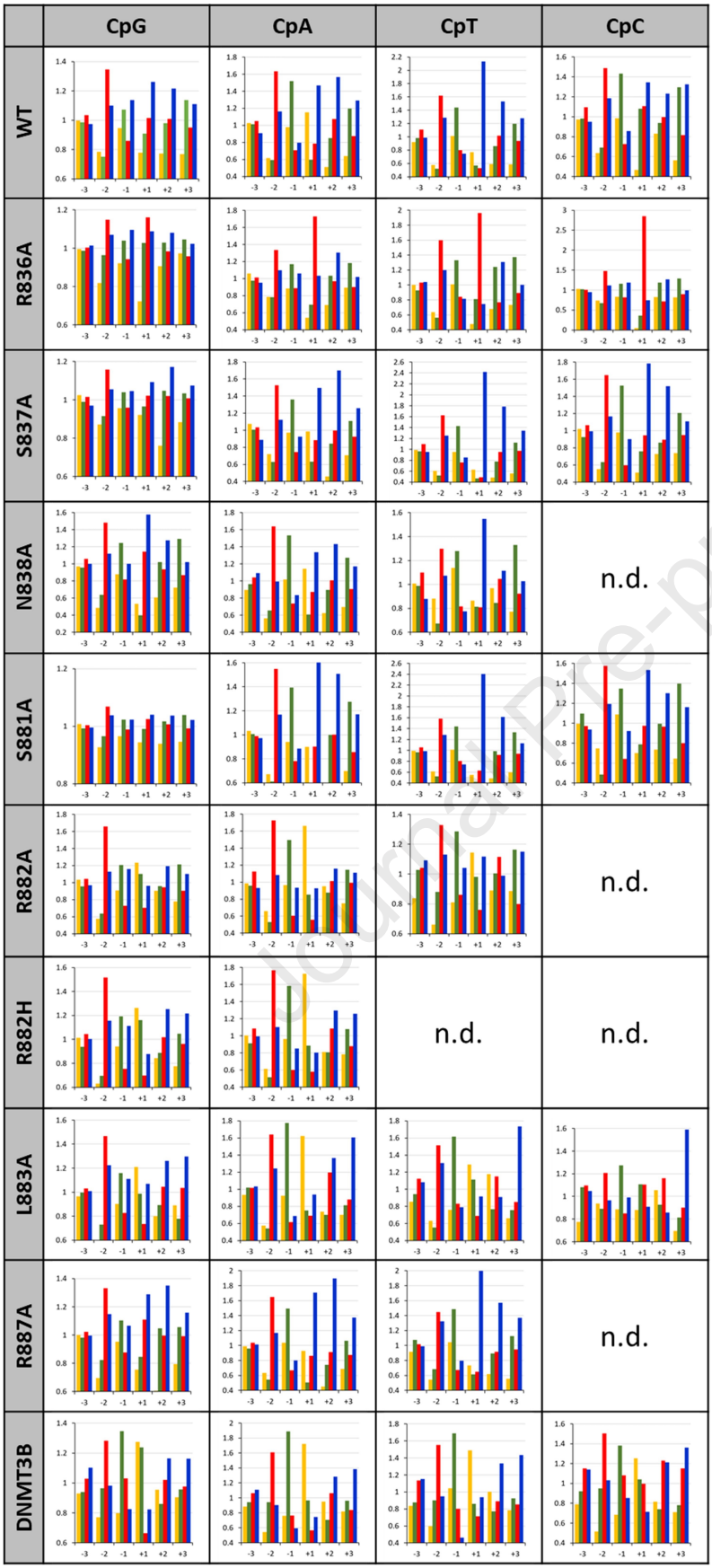
	G	A	T	C
G	1.21	0.987	0.734	1.068
A	1.624	0.751	0.689	0.937
T	1.288	1.111	0.686	0.914
C	0.88	1.107	1.104	0.909

R887A

	G	A	T	C
G	0.755	0.847	1.11	1.288
A	0.928	0.504	0.861	1.706
T	0.736	0.615	0.65	1.999
C				

DNMT3B

	G	A	T	C
G	1.274	1.237	0.664	0.824
A	1.722	0.967	0.566	0.745
T	1.488	0.861	0.713	0.938
C	1.252	1.039	0.996	0.713



CRediT author statement

Michael Dukatz: Conceptualization, Formal analysis, Investigation, Visualization, Writing - Original Draft, Writing - Review & Editing

Marianna Dittrich: Formal analysis, Investigation, Writing - Review & Editing

Elias Stahl: Formal analysis, Investigation, Writing - Review & Editing

Sabrina Adam: Methodology, Investigation, Writing - Review & Editing

Alex de Mendoza: Investigation, Visualization, Writing - Review & Editing

Pavel Bashtrykov: Software, Methodology, Investigation, Writing - Review & Editing

Albert Jeltsch: Conceptualization, Formal analysis, Writing - Original Draft, Visualization, Writing - Review & Editing, Project administration, Funding acquisition