How to cite: *Angew. Chem. Int. Ed.* **2024**, *63*, e202318837
doi.org/10.1002/anie.202318837

Epigenetics

Evolved Readers of 5-Carboxylcytosine CpG Dyads Reveal a High Versatility of the Methyl-CpG-Binding Domain for Recognition of Noncanonical Epigenetic Marks

Brinja Kosel, Katrin Bigler, Benjamin C. Buchmuller, Suchandra R. Acharyya, Rasmus Linser, and Daniel Summerer**

Abstract: Mammalian genomes are regulated by epigenetic cytosine (C) modifications in palindromic CpG dyads. Including canonical cytosine 5-methylation (mC), a total of four different 5-modifications can theoretically co-exist in the two strands of a CpG, giving rise to a complex array of combinatorial marks with unique regulatory potentials. While tailored readers for individual marks could serve as versatile tools to study their functions, it has been unclear whether a natural protein scaffold would allow selective recognition of marks that vastly differ from canonical, symmetrically methylated CpGs. We conduct directed evolution experiments to generate readers of 5-carboxylcytosine (caC) dyads based on the methyl-CpG-binding domain (MBD), the widely conserved natural reader of mC. Despite the stark steric and chemical differences to mC, we discover highly selective, low nanomolar binders of symmetric and asymmetric caC-dyads. Together with mutational and modelling studies, our findings reveal a striking evolutionary flexibility of the MBD scaffold, allowing it to completely abandon its conserved mC recognition mode in favour of noncanonical dyad recognition, highlighting its potential for epigenetic reader design.

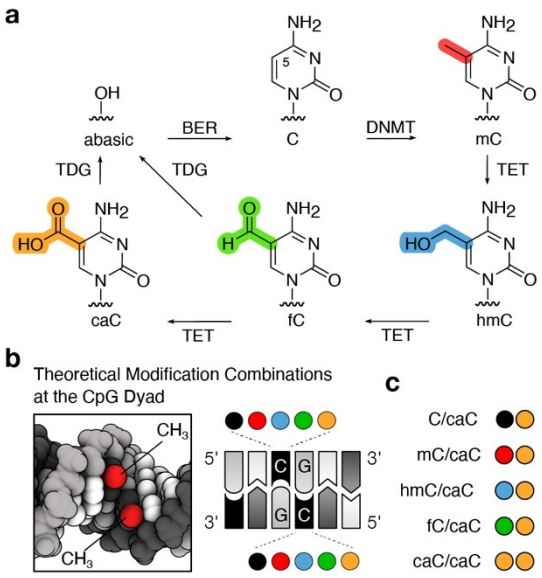

Figure 1. Mammalian cytosine modifications in CpG dyads. (a) DNMT and TET enzymes convert C to mC, hmC, fC, and caC. Thymine DNA glycosylase (TDG) and base excision repair (BER) can actively restore C. C can also be restored by inhibited maintenance methylation in oxidized CpG ("active modification-passive dilution").[2] (b) Combinations of modifications in the two CpG strands (left: cartoon of DNA duplex containing an mC/mC CpG; PDB 329d.[12] (c) CpG dyads containing caC used in this study.

Mammalian genome functions are dynamically regulated by the enzymatic introduction and erasure of substituents at the 5-carbon atom of cytosine (C, Figure 1a).[1] Cs in palindromic CpG dyads can be transformed to 5-methylcytosine (mC) by DNA methyltransferases (DNMTs). mC can be maintained in a strand-symmetric state (i.e., in both strands of the CpG) after replication, and the resulting "mC/mC" sites play essential roles for transcription regulation, differentiation and development. mC can also be oxidized to 5-hydroxymethyl- (hmC), 5-formyl- (fC) and 5-carboxycytosine (caC) by ten-eleven-translocation dioxygenases (TETs) in a non-processive manner, followed by potential active or passive de-modification processes (Figure 1a). This gives rise to a complex landscape of combinatorial CpG marks with unique physicochemical properties in the major groove of genomic DNA (Figure 1b–c).[2] Indeed, oxidized mCs have generally been shown to influence DNA interactions of MBDs,[3] nucleosomes,[4] transcription factors,[5] RNA polymerase II,[6] and other proteins.[7] Importantly, profiling studies have thereby shown that MBDs and many transcription factors are capable of combinatorial read-out, i.e. that they differentially interact with individual combinations of oxidized mCs in CpG dyads.[3b,5]

Protein probes for cytosine modifications are widely used for studying their biological functions, e.g., by affinity

[*] M. Sc. B. Kosel, M. Sc. K. Bigler, Dr. B. C. Buchmuller, M. Sc. S. R. Acharyya, Prof. Dr. R. Linser, Prof. Dr. D. Summerer
Faculty of Chemistry and Chemical Biology
TU Dortmund University
Otto-Hahn-Str. 4a, 44227 Dortmund (Germany)
E-mail: daniel.summerer@tu-dortmund.de
rasmus.linser@tu-dortmund.de

enrichment/sequencing/mapping, by imaging, and by functional assays.[8] However, probes that selectively recognize user-defined CpG duplex marks involving oxidized mCs are elusive.[7b] Candidate scaffolds for engineering are MBD proteins, the natural readers of mC/mC dyads[9] that are widely used as probes for affinity enrichment.[8,10] These share a small, highly conserved MBD domain (Figure 2a) that recognizes the CpG guanines by two arginines via the Hoogsteen face, whereas the mC methyl groups are recognized by two different hydrophobic pockets (Figure 2a).[11]

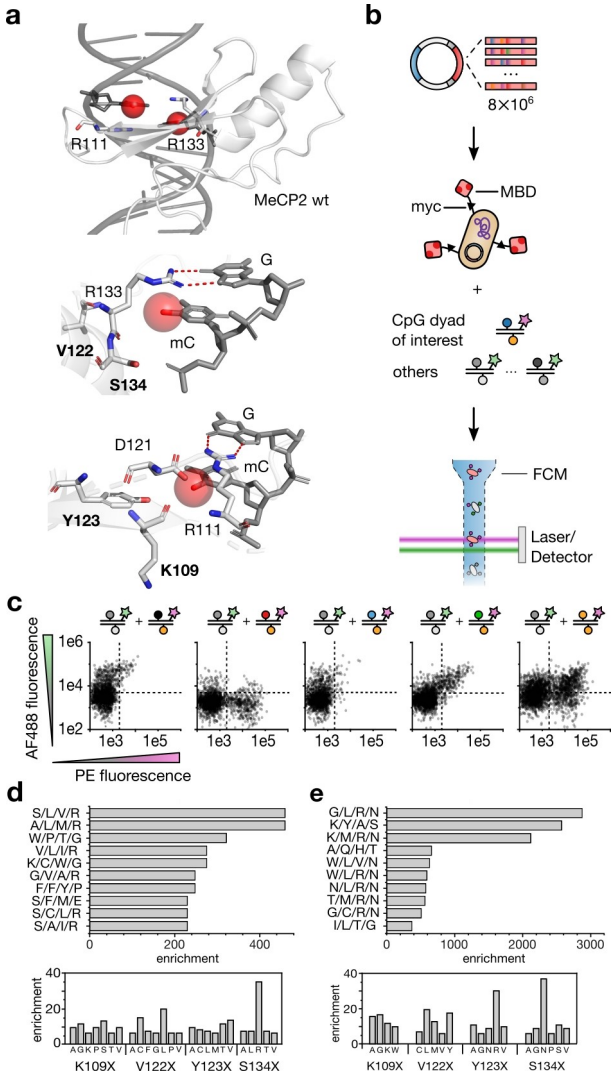

Figure 2. Directed evolution of MBD domains for caC-containing CpG dyads. (a) Structures of the MeCP2 MBD bound to DNA containing an mC/mC CpG (PDB 3c2i)[11] showing conserved interactions and library random sites (bold). 5-Methyl groups of mC are shown as red spheres. (b) Bacterial surface display of MBD mutant library and example DNA probes used for FCM-selections. (c) FCM analysis of the libraries after two rounds of selection for selective binders of each of the five indicated caC-containing CpG dyads. Nucleobase colour code as in Figure 1a. (d, e) Phenotype enrichment (top) and diversity per degenerated residue (bottom) for mC/caC selection (d) and caC/caC selection (e) obtained by high throughput sequencing of libraries after two iterative selection rounds. Phenotype enrichment shows residues in order of sequence; i.e. wt MeCP2 corresponds to K/V/Y/S (compare to Figure 2a).

However, the engineering potential of MBDs is poorly understood,[13] and the presence of an unmodified C or oxidized mC repels MBDs, indicating that their canonical binding mode is intolerant for changes in size and/or hydrophobic character of the cytosine 5-substituent.[3] We have recently evolved a selective reader of the hmC/mC dyad based on the MeCP2 MBD.[13b] However, this dyad only minimally differs from canonical mC/mC in only one strand, and this MBD mutant retained a conserved mC pocket for canonical recognition.[14]

To evaluate the general design potential of the MBD scaffold for the selective recognition of completely unrelated CpG dyads via novel interaction modes, we here aimed to evolve readers of CpGs containing caC, the modification that differs most drastically from canonical mC: its carboxyl group is rigidly fixed in plane with the cytosine ring, it features the highest steric demand among the C modifications, and it replaces the hydrophobic surface of mC essential for canonical MBD recognition with a charged group.[15]

We employed a library of the MeCP2 MBD (aa 90–181) with four NNK-randomized residues in proximity to the C 5-substituents. In comparison to error-prone PCR diversification, this strategy allows to comprehensively sample a focused chemical space with a high likelihood to afford improvements in single-generation evolution experiments. The four residues constitute the major part of the surface of both mC pockets: S134, interacting with a phosphate in vicinity of one mC; Y123, interacting with the amino group of the opposite mC via a water molecule; and V122/K109 which are in overall vicinity of the dyad (Figure 2a).[11] In this library, the G-binding arginines R111 and R133 are conserved.

We conducted selections using a bacterial surface display protocol that allows to rapidly assign selectivities for multiple on- and off-target CpG dyads to single MBD mutants in a competitive fashion. This protocol is based on the AIDA-I protein that features a C-terminal autotransporter domain promoting the translocation of N-terminally attached passenger proteins across the cell envelopes of gram-negative bacteria. The system allows displaying MBDs on the surface of *E. coli* cells, and enables pooled, iterative sampling of MBD clones that bound to synthetic DNA probes using flow cytometry (FCM, see Figure 2b).[16] The probes contain a single CpG dyad in an oligo-dA/dT context, and are labeled with two different fluorophores for the target and off-target dyads, respectively (Figure 2d; see the Supporting Information for a description of the surface display and FCM procedures).

We carried out individual, iterative selections for all five possible, caC-containing CpG dyads, each involving all other CpG dyads as off-targets. Selections were conducted at a scale that fully covers the libraries' theoretical diversity of $>10^6$ mutants. After two selection rounds (each consisting of probe binding, FCM selection, and re-growth) with gating for high on-target and low off-target binding, we observed marked differences in the overall fluorescence distributions

of the library by FCM. Whereas the libraries shifted to off-target or promiscuous target-binding in the selections for C/caC, hmC/caC and fC/caC dyads, significant fractions of the libraries in the mC/caC and caC/caC dyad selections showed a marked shift to the on-target probes (Figure 2c). We therefore focused on the latter two targets, and determined the library dynamics on the genotype level by pool high throughput sequencing. Surprisingly, none of the randomized positions of the top 15 enriched clones in both selections showed a tendency to retain a wild type residue, indicating that these mutants completely abandoned the canonical recognition mode of MBDs (Figure 2d–e; see Table S1 for comprehensive enrichment data). Instead, many of the clones shared other specific residues at a subset of positions. In selections for mC/caC dyads, a striking feature was the occurrence of a positive charge by an S134R (but not S134K) mutation, together with a rather conservative steric alteration (L, C, A) of V122, a residue that we have previously identified to control the plasticity of the hydrophobic MBD core[14] (Figure 2d and S4). These preferences were also visible on the level of individual codons (Figure 2d bottom). In contrast, clones of the selection for caC/caC showed a clear preference for a positive charge at a different position (Y123R; again, Y123K did not occur) in strict combination with an S134N mutation (Figure 2e and S4). No R or K was found at the 134 position as in the mC/caC selection, and again there was no tendency to retain wild type residues.

We selected eight clones of the mC/caC selection for profiling experiments with all fifteen theoretical dyads via electromobility shift assays (EMSA, see Figure S4–14 for details). These included SLVR and ALMR, the two highest enriched clones of the selection, two related clones we identified in a final single clone sorting step (ACFR and KLTR), as well as four enriched clones with unrelated phenotypes (such as WPTG, FFYP or PVTP). Whereas the latter clones did not bind DNA in EMSA even at high MBD concentration (Figure S4), the four clones containing an R134 showed a very different selectivity profile compared to wt MeCP2 (Figure 3a), with a specific preference for caC-containing CpG dyads (Figure 3b and S4–14). All mutants showed highest binding to caC/caC and mC/caC dyads, followed by C/caC, hmC/caC and fC/caC. Other dyads were bound with lower affinity, with the lowest ones being fC/fC and ones containing a C or hmC. This latter observation reminds of the low affinities of wt MBDs for C and hmC-containing dyads.[3] We next picked the two highest enriched clones with Y123R/S134N phenotype (GLRN, KMRN) of the caC/caC selection together with an additional clone from a final sorting step (QLRN) for characterization.

Unlike the mutants of the mC/caC selections, these mutants showed a clear selectivity for caC/caC, and bound some of the other caC-containing dyads with only slightly higher affinity than other off-targets (Figure 3c). Although position 109 varied, the mutants showed similar overall profiles. Similar to the previous selection, hydrophobic residues (L and M) were dominant at the critical core residue V122.[14] We tested versions of the three mutants with position 122 replaced by either L or M (GMRN, KLRN, QMRN). Within this group of six mutants, QLRN and GMRN appeared to have the highest selectivity (Figure 3c).

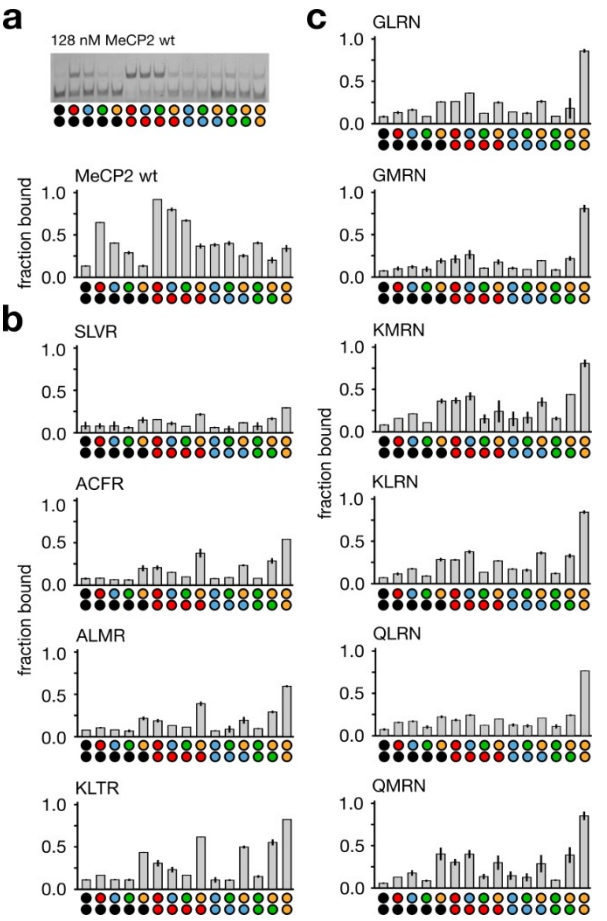

Figure 3. Selectivity profiling of MBD mutants identified in mC/caC and caC/caC selections for all possible CpG modification combinations. (a) EMSA-based selectivity profile of MeCP2 wt at 128 nM protein concentration. Color codes as in Figure 1a. Top shows exemplary EMSA gel (see Figure S25 for uncropped gel). (b) Same profiles for MBD mutants identified in mC/caC selection at 128 nM protein concentration. (c) Same profiles for MBD mutants identified in caC/caC selection at 128 nM protein concentration, including additional mutants of position 122. All data are from independent duplicate experiments, error bars in a, b and c mean ± SEM of two technical replicates (see Supporting Information for all raw data).

We next measured K_Ds of clones with the highest apparent affinities and selectivities (Figure S16) of both screens for binding of the canonical mC/mC dyad, as well as both on-target dyads mC/caC and caC/caC (Figure 4a and S17).

KLTR from the mC/caC selection had a drastically (95-fold) reduced affinity for mC/mC compared to wt MeCP2, but an increased, low nanomolar affinity for both caC dyads (Figure 4a; 48 and 85 nM, respectively). For the caC/caC selection, implying a greater alteration of the binding epitope compared to the canonical mC/mC target, mutants GMRN and QLRN showed an even greater loss of affinity for mC/mC than KLTR (117- and 145-fold), and bound caC/caC with even higher (~30 nM) affinity. Both mutants

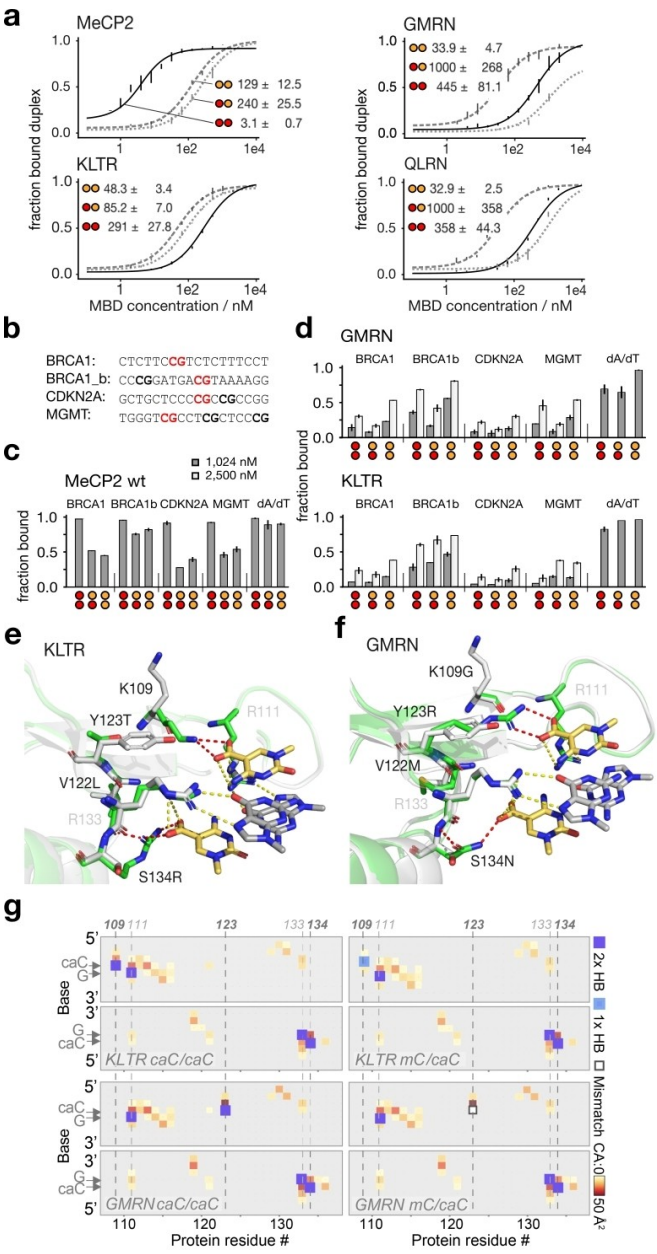

Figure 4. Characterization of caC-selective MeCP2 mutants. (a) K_Ds from EMSA for three most relevant CpG dyads (Error bars mean ± SEM of two technical replicates). (b) Sequences of dsDNA probes derived from cancer gene promoters used in Figure 4c–d (modified CpG in red and other CpGs in bold). (c) EMSA data for wt MeCP2 (duplicate measurements). (d) EMSA data for GMRN and KLTR (duplicate measurements). (e, f) Models of KLTR (e) and GMRN (f) interacting with caC/caC/dyad. Mutants in green, superimposed MeCP2 wt (pdb 3c2i)[11] in white. Retained, canonical interactions found in both wt and mutant in yellow, new interactions uniquely occurring in mutants in red. (g) Contact arrays for KLTR and GMRN with caC/caC and mC/caC. Note canonical, wt-like interactions of R111/133 with Gs (numbered in white), and new interactions of mutant-specific residues (numbered in black). Favorable interactions to C 5-substituents marked with green arrow–GMRN cannot interact with mC 5-methyl (red arrow), explaining its high selectivity for caC/caC. Interaction color code on the right (HB = hydrogen bond; CA: contact area).

showed a drastically increased selectivity over mC/caC, which was bound with a K_D of only >1 μM (Figure 4a). To study potential context dependencies of the mutant's affinities and selectivities, we conducted EMSAs with wt MeCP2, GMRN and KLTR in four additional sequence contexts. These covered parts of three cancer gene promoters (BRCA1, CDKN2A, MGMT), and differed in GC content and the presence of additional CpG dyads (Figure 4b). We observed slightly reduced overall affinities for GMRN and KLTR in these selected sequences, while the selectivity profiles where retained (Figure 4c–d).

To gain insights into the origin of the caC selectivity of the two new mutant classes, we conducted modelling studies with KLTR and GMRN using CHARMM.[17] KLTR retained the conserved wt recognition of the CpG guanines by R111 and R133 in this model (Figure 4e, g and S16). Interestingly, the 5-carboxyl groups of both caCs can be positioned under these arginines, leading to favorable interactions that are likely also present in wt MeCP2. Most striking, however, was the S134R mutation that likely accounts for both, the increased affinity to caC and the loss of mC binding. Specifically, it enables a salt bridge to one caC carboxyl group and removes S134, which forms part of the canonical hydrophobic pocket accommodating one mC 5-methyl group in wt MeCP2[11] (Figure 4e). Additionally, K109 (interacting with a phosphate in the wt, not shown in Figure 4b) can alter its orientation to form a salt bridge to the opposite caC carboxyl group. However, given the variability of this position in other enriched mutants (Figure 3b), this interaction seems not important, which may explain the rather promiscuous binding of mC/caC and caC/caC by KLTR.

In contrast, GMRN appears to gain strong interactions to the symmetric caC/caC dyad via its Y123R and S134N mutations, each recognizing one 5-carboxyl group (Figure 4f, g). In addition, S134–the canonical, mC-binding residue of wt MeCP2–is again removed, and unlike K109, R123 cannot swing out without leaving an unfilled hole, making mC enthalpically highly unfavorable in each of the two potential sites. This provides an explanation for the high selectivity of XXRN mutants for caC/caC over the asymmetric mC/caC dyad (Figure 4f, g). In both KLTR and GMRN mutants, the hydrophobic core residue 122 adopts a conserved conformation, but with fine-tuned sterics (Figure S19). A possible influence on the plasticity of the MBD core (as observed in previous studies for an hmC/mC-binding mutant)[14] cannot be excluded based on the model.

In conclusion, we report the first designed readers of symmetric and asymmetric caC-containing CpG dyads. Our directed evolution experiments reveal that–despite its high conservation for the distinct chemical properties of canonical mC/mC dyads and its small DNA contact area governing base specificity–the MBD scaffold is highly mutable exactly at this area. In the sampled four-position random region that covers the major part of the two canonical mC pockets, we did not find a single of the wt residues to be essential. Instead, the MBD is able to completely abandon its conserved mC recognition in favor of novel binding modes for two new, structurally distinct

epigenetic marks. Moreover, the overall recognition of CpG dyads encoded by the peculiar arrangement of the two spatially conserved arginine fingers R111 and R133 (Figure 2a) is surprisingly uncoupled from the recognition of new mC oxidation states, so that directed evolution towards noncanonical dyads occurs independently from the overall functionality of the fold. This orthogonality appears to be a unique, highly suitable fundament for the domains' compatibility with a rather unrestricted chemical space of epigenetic cytosine modifications, granting versatility for the emerging prospects of interrogating the combinatorial landscape of epigenetic CpG marks by designer readers, e.g. via affinity enrichment and sequencing alone or in combination with subsequent nucleobase conversion-based sequencing.

Supporting Information

The authors have cited additional references within the Supporting Information.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG)–325871075 and the Emmy Noether program – and Germany's Excellence Strategy–EXC 2033–390677874–RESOLV. Funded by the European Research Council (ERC CoG EPICODE, No. 723863 to D.S.). The work received funding from the CANTAR program "Netzwerke 2021", an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia. The sole responsibility for the content of this publication lies with the authors. Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Eldorado at not available yet, reference number 0.

Keywords: Epigenetic cytosine modifications · Methyl-CpG-binding domains · DNA recognition · Directed Evolution

[1] C. D. Allis, T. Jenuwein, *Nat. Rev. Genet.* **2016**, *17*, 487–500.
[2] a) X. Wu, Y. Zhang, *Nat. Rev. Genet.* **2017**, *18*, 517–534; b) T. Carell, M. Q. Kurz, M. Muller, M. Rossa, F. Spada, *Angew. Chem. Int. Ed. Engl.* **2018**, *57*, 4296–4312.
[3] a) H. Hashimoto, Y. Liu, A. K. Upadhyay, Y. Chang, S. B. Howerton, P. M. Vertino, X. Zhang, X. Cheng, *Nucleic Acids Res.* **2012**, *40*, 4841–4849; b) B. C. Buchmuller, B. Kosel, D. Summerer, *Sci. Rep.* **2020**, *10*, 4053.
[4] E. A. Raiber, G. Portella, S. Martinez Cuesta, R. Hardisty, P. Murat, Z. Li, M. Iurlaro, W. Dean, J. Spindel, D. Beraldi, Z. Liu, M. A. Dawson, W. Reik, S. Balasubramanian, *Nat. Chem.* **2018**, *10*, 1258–1266.
[5] G. Song, G. Wang, X. Luo, Y. Cheng, Q. Song, J. Wan, C. Moore, H. Song, P. Jin, J. Qian, H. Zhu, *Nat. Commun.* **2021**, *12*, 795.
[6] L. Wang, Y. Zhou, L. Xu, R. Xiao, X. Lu, L. Chen, J. Chong, H. Li, C. He, X. D. Fu, D. Wang, *Nature* **2015**, *523*, 621–625.
[7] a) C. G. Spruijt, F. Gnerlich, A. H. Smits, T. Pfaffeneder, P. W. Jansen, C. Bauer, M. Munzel, M. Wagner, M. Muller, F. Khan, H. C. Eberl, A. Mensinga, A. B. Brinkman, K. Lephikov, U. Muller, J. Walter, R. Boelens, H. van Ingen, H. Leonhardt, T. Carell, M. Vermeulen, *Cell* **2013**, *152*, 1146–1159; b) G. P. Pfeifer, P. E. Szabo, J. K. Song, *J. Mol. Biol.* **2020**, *432*, 1718–1730.
[8] M. J. Booth, E. A. Raiber, S. Balasubramanian, *Chem. Rev.* **2015**, *115*, 2240–2254.
[9] Q. Du, P. L. Luu, C. Stirzaker, S. J. Clark, *Epigenomics* **2015**, *7*, 1051–1073.
[10] a) A. B. Brinkman, F. Simmer, K. Ma, A. Kaan, J. Zhu, H. G. Stunnenberg, *Methods* **2010**, *52*, 232–236; b) D. Serre, B. H. Lee, A. H. Ting, *Nucleic Acids Res.* **2010**, *38*, 391–399.
[11] K. L. Ho, L. W. Mcnae, L. Schmiedeberg, R. J. Klose, A. P. Bird, M. D. Walkinshaw, *Mol. Cell* **2008**, *29*, 525–531.
[12] C. MayerJung, D. Moras, Y. Timsit, *J. Mol. Biol.* **1997**, *270*, 328–335.
[13] a) B. E. Tarn, K. J. Sung, H. D. Sikes, *Mol. Syst. Des. Eng.* **2016**, *1*, 273–277; b) B. C. Buchmuller, J. Droden, H. Singh, S. Palei, M. Drescher, R. Linser, D. Summerer, *J. Am. Chem. Soc.* **2022**, *144*, 2987–2993.
[14] H. Singh, C. K. Das, B. C. Buchmuller, L. V. Schafer, D. Summerer, R. Linser, *Nucleic Acids Res.* **2023**, *51*, 6495–6506.
[15] M. W. Szulik, P. S. Pallan, B. Nocek, M. Voehler, S. Banerjee, S. Brooks, A. Joachimiak, M. Egli, B. F. Eichman, M. P. Stone, *Biochemistry* **2015**, *54*, 1294–1305.
[16] J. Maurer, J. Jose, T. F. Meyer, *J. Bacteriol.* **1997**, *179*, 794–804.
[17] B. R. Brooks, C. L. Brooks, 3rd, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, *J. Comput. Chem.* **2009**, *30*, 1545–1614.

Manuscript received: December 7, 2023
Accepted manuscript online: January 29, 2024
Version of record online: March 19, 2024