All living beings must choose to enact those genetic instructions from their comprehensive set of genes which are relevant to their survival and repress unneeded information of other genes. For multicellular organisms this implies coordination of gene expression in a cell-specific manner to realize tissue and organ function despite each cell carrying the same genetic material. One way to achieve such stable differentiation is to modify the nucleobases of DNA, the carrier of the genetic information.

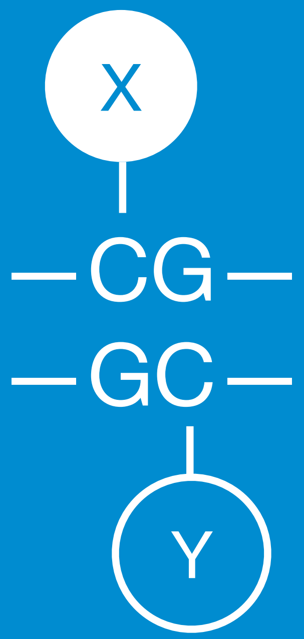
In mammals like human and mice, the methylation of the DNA nucleobase cytosine at carbon C5 of the pyrimidine ring exerts this function in a special way. The methylation takes place on both strands of the DNA double-helix in the short palindromic sequence CpG and often decisively changes the molecular interactions required to access the genetic information at these sites. Since the product 5-methylcytosine can be further oxidized enzymatically into 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxycytosine, different combinations of the cytosine derivatives can coexist in the DNA double-strand at such modified CpG dyads, each representing a unique chemical interaction interface. To date, their significance for gene regulation has escaped closer scrutiny because technological innovation to examine such combinations in native chromatin is missing. Here, I investigate the possibility to probe and decode strand-symmetric and strand-asymmetric combinations of modified cytosine derivatives in CpG dyads at the molecular level in the double-stranded DNA double-helix.

Starting from a set of homologous methyl-CpG-binding domain (MBD) proteins that interact with strand-symmetrically methylated CpG dyads, degenerated protein libraries were created based on structural contemplation and functional studies of MBD–DNA binding. A high-throughput screening assay specifically set up for this goal, recovered MBD variants with higher binding selectivity for one out of fifteen modified CpG dyads and allowed to draft first substitution profiles to accommodate some of these combinations. Specifically, several MBD variants were discovered that had novel binding selectivity not present in wildtype domains for 5-hydroxymethylcytosine- and 5-carboxycytosine-containing combinations. Further biochemical and structural analyses with respect to the basis of the binding specificity allowed for insights into the molecular recognition of strand-asymmetrically modified CpG dyads which will be key to unravel the epigenetic role of cytosine modifications in the human genome with such carefully tailored probes.

Buchmuller 2021

Deciphering strand-asymmetrically modified CpG dyads in the DNA double-helix

An evolutionary approach



Benjamin C. Buchmuller, M. Sc.
Faculty of Chemistry and Chemical Biology
TU Dortmund University
September 2021

Deciphering strand-asymmetrically modified CpG dyads in the DNA double-helix

An evolutionary approach

Inauguraldissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(*Dr. rer. nat.*)
der
Fakultät für Chemie und Chemische Biologie
der Technischen Universität Dortmund

vorgelegt von
Benjamin Christopher Buchmuller
aus Stuttgart

Dortmund 2021

Eidesstattliche Versicherung

Ich versichere hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel „Deciphering strand-asymmetrically modified CpG dyads in the DNA double-helix: An evolutionary approach“ selbstständig und ohne unzulässige fremde Hilfe angefertigt habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht.

Die Arbeit hat in gegenwärtiger oder in einer anderen Fassung weder an der Technischen Universität Dortmund noch einer anderen Hochschule im Zusammenhang mit einer staatlichen oder akademischen Prüfung vorgelegen.

Über die geltenden Rechtssätze zur vorsätzlichen Täuschung bei Prüfungsleistungen (nach § 63 Abs. 5 HG des Landes Nordrhein-Westfalen) und zur Abgabe einer falschen Versicherung an Eides statt (nach §§ 156, 161 StGB) wurde ich belehrt.

I hereby declare that I have completed the present dissertation independently and without illegitimate external support. I have not used any sources or tools other than those indicated throughout the text. The presence of any quotations, verbatim or according to their meaning, is clearly indicated.

This thesis has not been submitted, either wholly or substantially, to the TU Dortmund University or another university in connection with another state or academic examination.

Benjamin C. Buchmuller
Dortmund, am 9. September 2021

Von der Fakultät für Chemie und Chemische Biologie der Technischen Universität Dortmund als Dissertation angenommen.

Tag der Annahme:	18. Oktober 2021
Erstgutachter:	Prof. Dr. D. Summerer
Zweitgutachter:	Prof. Dr. R. Linser
Tag der mündlichen Prüfung:	26. Oktober 2021

"All appearances to the contrary, the only watchmaker in nature is the blind forces of physics, albeit deployed in a very special way."

— Richard Dawkins, The Blind Watchmaker (1986)

Abstract

All living beings must choose to enact those genetic instructions from their comprehensive set of genes which are relevant to their survival and repress unneeded information of other genes. For multicellular organisms this implies coordination of gene expression in a cell-specific manner to realize tissue and organ function despite each cell carrying the same genetic material. One way to achieve such stable differentiation is to modify the nucleobases of DNA, the carrier of the genetic information.

In mammals like human and mice, the methylation of the DNA nucleobase cytosine at carbon C5 of the pyrimidine ring exerts this function in a special way. The methylation takes place on both strands of the DNA double-helix in the short palindromic sequence CpG and often decisively changes the molecular interactions required to access the genetic information at these sites. Since the product 5-methylcytosine can be further oxidized enzymatically into 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxycytosine, different combinations of the cytosine derivatives can coexist in the DNA double-strand at such modified CpG dyads, each representing a unique chemical interaction interface. To date, their significance for gene regulation has escaped closer scrutiny because technological innovation to examine such combinations in native chromatin is missing. Here, I investigate the possibility to probe and decode strand-symmetric and strand-asymmetric combinations of modified cytosine derivatives in CpG dyads at the molecular level in the double-stranded DNA double-helix.

Starting from a set of homologous methyl-CpG-binding domain (MBD) proteins that interact with strand-symmetrically methylated CpG dyads, degenerated protein libraries were created based on structural contemplation and functional studies of MBD–DNA binding. A high-throughput screening assay specifically set up for this goal, recovered MBD variants with higher binding selectivity for one out of fifteen modified CpG dyads and allowed to draft first substitution profiles to accommodate some of these combinations. Specifically, several MBD variants were discovered that had novel binding selectivity not present in wildtype domains for 5-hydroxymethylcytosine- and 5-carboxycytosine-containing combinations. Further biochemical and structural analyses with respect to the basis of the binding specificity allowed for insights into the molecular recognition of strand-asymmetrically modified CpG dyads which will be key to unravel the epigenetic role of cytosine modifications in the human genome with such carefully tailored probes.

Zusammenfassung

Alle Lebewesen müssen dafür Sorge tragen, in ihrem umfangreichen Erbgut die gerade für sie überlebensnotwendigen Gene von denen zu unterscheiden, die nicht gebraucht werden. Insbesondere mehrzellige Organismen müssen diesen Prozess zellspezifisch koordinieren, trotz dessen, dass hier dieselbe Erbinformation in allen Zellen des Individuums vorliegt. Ein Mechanismus, welcher diesem Zwecke dient, ist die Modifikation von DNA-Nukleobasen, den Bausteinen des Trägers der Erbinformation.

In Säugetieren wie Mensch und Maus kommt hierbei der Methylierung der DNA-Nukleobase Cytosin am Kohlenstoffatom C5 des Pyrimidinrings eine besondere Rolle zu. Sie findet auf beiden Strängen der DNA-Doppelhelix innerhalb des kurzen Sequenzpalindroms CpG statt und trägt entscheidend dazu bei, dass hier ortsspezifisch andere molekulare Interaktionen für die Expression der Erbinformationen notwendig werden. Da das Produkt 5-Methylcytosin für weitere enzymatische Modifikationen wie der Oxidation zu 5-Hydroxymethylcytosin, 5-Formylcytosin oder 5-Carboxycytosin zur Verfügung steht, können unterschiedliche Kombinationen dieser Cytosinderivate mit gänzlich einzigartigen chemischen Eigenschaften an den komplementären CpG-Paaren im DNA-Doppelstrang vorliegen. Ein Aspekt, der unter dem Gesichtspunkt der epigenetischen Funktion dieser Derivate in Ermangelung technologischer Innovation sie in natürlichem Chromatin zu untersuchen, bislang kaum erschlossen werden konnte. Inwiefern es nun möglich ist, solche Strang-symmetrischen oder Strang-asymmetrischen Kombinationen von Cytosinderivaten in diesen CpG-Paaren auf molekularer Ebene in der DNA-Doppelhelix zu erkennen und somit gegebenenfalls zu entschlüsseln, ist Gegenstand der vorliegenden Arbeit.

Ausgehend von verschiedenen Homologen einer Proteindomäne, welche symmetrisch methylierte CpG-Paare erkennen, den Methyl-CpG-bindenden Domänen (MBD), wurden aufgrund struktureller Erwägungen und funktionaler Studien der MBD–DNA-Binding, degenerierte Proteinvariantenbibliotheken erstellt. Mithilfe eines hierfür eigens entwickelten Hochdurchsatzverfahrens gelang es, Varianten zu identifizieren, die nahezu selektiv eine aus fünfzehn Paarungen obiger Cytosinderivate im DNA-Doppelstrang erkennen. Neben allgemeinen Substitutionsprofilen für verschiedene Paarungen wurden im Speziellen mehrere MBD-Varianten entdeckt, die eine neue, natürlicherweise nicht vorhandene Selektivität für 5-Hydroxymethyl- und 5-Carboxymethylcytosin-haltige CpG-Paarungen aufwiesen. Aus der weiteren biochemischen und strukturellen Charakterisierung der Bindespezifität konnten einige Erkenntnisse über die molekulare Erkennung Strang-asymmetrisch modifizierter CpG-Paarungen gewonnen werden, welche in Zukunft als Schlüssel dienen können, die epigentische Funktion der Cytosinmodifizierung im humanen Genom mithilfe solcher speziell auf sie zugeschnittenen Sonden zu entschlüsseln.

Preface and Acknowledgements

This work was prepared from December 2017 to September 2021 in the group of Prof. Dr. Daniel Summerer (Chair in Chemical Biology of Nucleic Acids) at the TU Dortmund University. The idea to submit naturally modified cytosines in their biologically relevant context of the single DNA double-strand to closer scrutiny was conceived by Daniel in fall 2017. In the four years that followed, it was my great pleasure to pursue several strategies we came up with to probe such marks which could be of service to my colleagues in the future.

At the time I interviewed for a doctoral research position, I came to know that I enjoyed a reputation "to work like a watchmaker" at the bench. Though I still don't know whether or no this is accurate, I for sure could not have accomplished any of this without continuous support:

- I wish to express my sincere appreciation to *Prof. Dr. Daniel Summerer* for the exciting research question and the privilege to perform this study as a member of his team. I am obliged to him for his constant and constructive support, his trust and the creative freedom throughout the research process.
- I am grateful to *Prof. Dr. Dr. hc. Herbert Waldmann* and *Prof. Dr. Andrea Musacchio* for assuming the office of my thesis advisory committee, their guidance and support.
- I thank *Prof. Dr. Rasmus Linser* for kindly agreeing to give the second appraisal of this work.
- Special thanks go to *Brinja Kosel* (M. Sc.), *Christoph Hoppe* (M. Sc.), *David Horlacher* (intern), *Jessica Jankowski* (B. Sc.), *Katja Schamrin* (B. Sc.), and *Katrin Bigler* (B. Sc.) whom I had the pleasure to train and advise at different stages in their careers and who have contributed to this project's lasting success with their firm commitment.
- I am indebted to our excellent cooperation partners: *Jessica Dröden* and *Prof. Dr. Malte Drescher* (University of Konstanz) for contributing the DEER measurements; the impressive work of *Dr. Himanshu Singh* and *Prof. Dr. Rasmus Linser* on the NMR structure of the DNA–protein complex; as well as *Dr. Rosario Oliva* and *Prof. Dr. Roland Winter* for their help with the CD spectroscopy. Only some parts of this fruitful work can be presented here.
- I further thank the research groups of *Prof. Dr. Daniel Rauh*, *Prof. Dr. Guido Clever*, and *Dr. Malte Gersch* for offering their time and equipment as and when needed.
- A special thanks goes to the lab's research technicians *Simone Eppmann* and *Nadine Schmidt*, in particular for carrying out the large-scale recombinant protein expressions.

- I wish to thank all former and present coworkers for the welcoming, engaging and open-minded work atmosphere and the delightful time off the lab, namely *Álvaro Muñoz-López*, *Anna Witte*, *Anne Jung*, *Brinja Kosel*, *Damian Schiller*, *Dominic Kamps*, *Jan Wolffgramm*, *Lena Engelhardt*, *Mario D. Gieß*, *Sara Maurer*, *Dr. Shubhendu Palei*, *Suchet Nanda*, *Sudakshina Banerjee*, and *Tzu-Chen Lin*. It was so great to be around you!
- For handling all administrative affairs, I thank *Maria Sergani*, *Martina Reibner*, *Petra Alhorn*, *Ulrich Schoppe* and the coordination office of the International Max Planck Research School, *Christa Hornemann* and *Lucia Sironi*.

I extend my acknowledgements to everyone who has played a role in my academic and personal accomplishments: First of all to my parents and my family who supported me with love and understanding. The Heidelberger Life-Science Lab at the German Cancer Research Center which has shaped my skills and interests since my early high school years. My friends and colleagues who have grounded me, inspired me and cheered me up and who have not been mentioned yet: *M. Finke*, *K. Herbst*, *Y. Irtel von Brenndorff*, *C. D. Reinkemeier*, *F. Simonis*, *S.* and *S. Tischlik*. Thank you all for your unwavering support!

To carry out the research presented here, I gratefully acknowledge funding from:

- the TU Dortmund University,
- the European Union's Horizon 2020 research and innovation program under the grant agreement No 723863 entitled "Programmable Readers, Writers, and Erasers of the Epigenetic Cytosine Code" (ERC consolidator grant to Prof. Dr. Daniel Summerer),
- the International Max Planck Research School (IMPRS) in Molecular and Chemical Biology (2017–2020) and the IMPRS in Living Matter (2021) at the Max Planck Institute of Molecular Physiology Dortmund,
- the Leona M. and Harry B. Helmsley Charitable Trust through a travel grant for an advanced training at the Cold Spring Harbor Laboratory,

and

- the Joachim Herz Stiftung through a generous, two-year Add-on Fellowship for Interdisciplinary Life Science.

List of Publications

This work has furnished the following publications:

Buchmuller, B. C., Kosel, B., & Summerer, D. (2020). Complete profiling of methyl-CpG-binding domains for combinations of cytosine modifications at CpG dinucleotides reveals differential read-out in normal and Rett-associated states. *Sci. Rep.*, *10*(1), 4053–9.

Further publications:

Buchmuller, B. C., Jung, A., Muñoz-López, Á., & Summerer, D. (2021). Programmable tools for targeted analysis of epigenetic DNA modifications. *Curr. Opin. Chem. Biol.*, *63*, 1–10.

Wolffgramm, J., Buchmuller, B. C., Palei, S., Muñoz-López, Á., Kanne, J., Janning, P., ... Summerer, D. (2021). Light-activation of DNA-methyltransferases. *Angew. Chem., Int. Ed.*, *60*(24), 13507–13512.

Muñoz-López, Á., Buchmuller, B. C., Wolffgramm, J., Jung, A., Hussong, M., Kanne, J., ... Summerer, D. (2020). Designer receptors for nucleotide-resolution analysis of genomic 5-methylcytosine by cellular imaging. *Angew. Chem., Int. Ed.*, *59*(23), 8927–8931.

Palei, S., Buchmuller, B. C., Wolffgramm, J., Muñoz-López, Á., Jung, S., Czodrowski, P., & Summerer, D. (2020). Light-activatable TET-dioxygenases reveal dynamics of 5-methylcytosine oxidation and transcriptome reorganization. *J. Am. Chem. Soc.*, *142*(16), 7289–7294.

Gieß, M., Muñoz-López, Á., Buchmuller, B. C., Kubik, G., & Summerer, D. (2019). Programmable protein–DNA cross-linking for the direct capture and quantification of 5-formylcytosine. *J. Am. Chem. Soc.*, *141*(24), 9453–9457.

Maurer, S., Buchmuller, B. C., Ehrt, C., Jasper, J., Koch, O., & Summerer, D. (2018). Overcoming conservation in TALE–DNA interactions: a minimal repeat scaffold enables selective recognition of an oxidized 5-methylcytosine. *Chem. Sci.*, *9*, 7247–7252.

The MBD mutants disclosed in this work are subject to one or more pending or issued patent claims filed by the TU Dortmund University under PCT WO 19 220 082.2, 2020.

Table of contents

List of Publications x
List of Tables xiv
List of Figures xv
List of Abbreviations xviii

Introduction and Aim of This Work

1 Epigenetic cytosine modifications in genomic DNA 3
1.1 Abundance of modified cytosines 4
1.2 Origin and turnover of cytosine modifications 5
1.3 Distribution of modified cytosines in the mammalian genome 8
1.4 Biological consequences of DNA cytosine modification 11
1.5 Strand-symmetric and -asymmetric cytosine C5 modifications in CpG dyads 14
1.6 Synopsis 15

2 Technologies to detect modified cytosines in DNA 17
2.1 Detection of modified cytosines in the genomic sequence context 18
2.2 Detection of concurrently modified cytosines in CpG dyads 18
2.3 Synopsis 23

3 Recognition of modified cytosines in DNA 25
3.1 DNA base and shape readout by DNA-binding proteins 25
3.2 Protein domains that recognize C5-modified cytosines 29
3.3 The methyl-CpG-binding domain 34
3.4 Molecular determinants of MBD–DNA binding 35
3.5 Synopsis 37

4 Protein engineering and directed evolution platforms 39
4.1 *Ex vivo* surface display platforms 40
4.2 Fluorescence-activated high-throughput screening platforms 41
4.3 Synopsis 43
Aim 45

Original Work and Discussion

5 **Promise and prospect of the methyl-CpG-binding domain** 49

5.1 Sequence and structural conservation in the MBD 50

5.2 Specific differences in mammalian MBD binding 52

5.3 Candidates for protein engineering 57

5.4 Synopsis 58

6 **Decoding of modified CpG dyads by natural MBDs** 59

6.1 Interaction of wildtype MBDs with modified CpG dyads 59

6.2 Effect of MeCP2 Rett mutations on interactions with modified CpG dyads 62

6.3 Synopsis 65

7 **High-throughput screening for MBD variants with novel modified CpG dyad specificity** 67

7.1 Candidate MBD residues to target for altered DNA binding selectivity 67

7.2 Creation of site-saturated codon-degenerated MBD libraries 69

7.3 Bacterial AIDA-I-mediated MBD cell surface display 72

7.4 Fluorescence-activated MBD–DNA binding assay 76

7.5 Full library screening for MBDs with novel binding selectivity 84

7.6 Synopsis 89

8 **An artificial MBD with novel CpG binding selectivity** 91

8.1 MeCP2[K109T/V122A/S134N] is an artificial 5hmC/5mC-selective MBD 91

8.2 Structural characterization 96

8.3 Synopsis 100

9 **Fitting binding isotherms to electrophoretic mobility shift data** 103

9.1 Binding to a single site 104

9.2 Binding to multiple sites 105

9.3 Macroscopic apparent equilibrium constants 106

9.4 Microscopic equilibrium constants 109

9.5 Implementation in R and additional features 114
9.6 Synopsis 117

Conclusions and Outlook

10 Conclusions 121
10.1 Monitoring protein–DNA interactions on the bacterial cell-surface 121
10.2 Providing a framework to characterize MBD–DNA binding specificities 122
10.3 Engineering MBDs with novel DNA binding selectivity 122
11 Future development and outlook 125

Resources

12 Materials 129
12.1 Hosts and vectors 129
12.2 Oligonucleotides and probes 132
12.3 Instruments and consumables 139
12.4 Buffers, reagents and kits 142
12.5 Software and online tools 148
13 Methods 149
13.1 General procedures and routines 149
13.2 Recombinant MBD expression and protein purification 150
13.3 Electrophoretic mobility shift assays 156
13.4 Creation of NNK codon-degenerated MBD libraries 158
13.5 Cell surface display of MBDs 160
13.6 MBD–DNA binding assays and MBD library screening 161
A Supplementary Information 167
A.1 Supplementary Figures 167
A.2 Supplementary Tables 179
A.3 Source code 190
A.4 Credits and license information 198

B References 201

List of Tables

1.1 Distribution of cytosine in CpG and non-CpG contexts 8
1.2 Distribution of CpG dinucleotides 9
2.1 Reactivity and specificity of commercially available affinity agents 22
5.1 Root-mean-square distances between superimposed MBD structures 50
6.1 Dissociation constants for MeCP2 Rett mutants R133C and S134C 64
12.1 Bacterial strains 129
12.2 Plasmids 129
12.3 Oligonucleotides for DNA duplex probes 133
12.4 Oligonucleotides for DNA sequencing 134
12.5 Oligonucleotides for cloning 135
12.6 Gene fragments 137
12.7 Laboratory instruments 139
12.8 Consumables and disposable labware 141
12.9 Commercial kits and ready-to-use mixtures 142
12.10 Commercially available enzymes 143
12.11 Commercially available antibodies and affinity reagents 144
12.12 Chemicals 144
13.1 Consensus coding sequences and cloning of MBD wildtype domains 152
13.2 Missense mutations and cloning of MeCP2 Rett mutants 153
13.3 Cloning of MeCP2 cysteine variants 153
13.4 Physicochemical properties of MBD fusion proteins 155
13.5 Combinations of ODNs to create homopolymeric DNA probes for EMSA 157
13.6 Combinations of ODNs to create genomic DNA probes for EMSA 158
13.7 Degenerated MBD libraries created in this work 160
13.8 Combinations of ODNs to create DNA probes for FACS binding assays 163
A.1 Content of modified cytosines in various genomes 179
A.2 Sequences of DNA probes used in the literature 181
A.3 Reported binding affinities of MBD1 182

A.4 Reported binding affinities of MBD2 182
A.5 Reported binding affinities of MBD3 182
A.6 Reported binding affinities of MBD4 183
A.7 Reported binding affinities of MeCP2 183
A.8 Model estimates for one-site MeCP2 wildtype and Rett mutant binding 185
A.9 Model estimates for multi-site MBD–DNA binding 186
A.10 Model estimates for MeCP2 Ala-122 variants 188

List of Figures

1.1 Biochemical derivatives of cytosine 4
1.2 Content of modified cytosines in various genomes 5
1.3 Natural processes modifying cytosine in mammalian genomes 6
1.4 Biological effects of DNA methylation in the context of DNA-binding proteins . . . 13
1.5 CpG dinucleotide dyads in double-stranded DNA 15
2.1 Technologies to detect modified cytosines 17
2.2 Bisulfite conversion of C5-modified cytosines 19
2.3 Inferring strand-(a-)symmetry from conversion-based DNA sequencing methods 20
3.1 Base and shape readout of methylated DNA 26
3.2 Double-stranded DNA-binding domains and contact statistics 29
3.3 Three-dimensional structure of the MBD 35
3.4 CpG dyad-specific MBD–DNA interactions 36
4.1 Rational design and directed evolution 39
4.2 Surface-display with the adhesin involved in diffuse adherence (AIDA-I) autotransporter 41
4.3 Fluorescence-activated cell sorting 42
4.4 Kinetic and thermodynamic DNA binding assays 42
5.1 MBDs in MBD protein family members 50
5.2 Conservation of sequence and structure in the MBD 51
5.3 Protein–DNA contact surface of MBD proteins 53
5.4 Protein–DNA interactions in five human MBDs 54
6.1 Binding affinity of wildtype MBDs reported in the literature 60

6.2 Binding selectivity of wildtype MBDs for different modified CpG dyads 61
6.3 Protein architecture and positions of MeCP2 Rett mutants 62
6.4 Selectivity and affinity of MeCP2 Rett mutants 63
6.5 Affinity of MeCP2 Rett mutants R133C and S134C 64
7.1 Positions of the degenerated amino acid residues in five MBD libraries 68
7.2 Strategies to create site-saturated codon-degenerated MBD libraries 70
7.3 Degeneracy of the MBD libraries 71
7.4 Surface-display of various MBDs using pET- and pBAD-based expression system 73
7.5 Inducer concentration and survival in the pET-MBD-AIDA^{C} system 75
7.6 Fluorescence-based assays to assess DNA binding affinity and selectivity 77
7.7 Choice of fluorophore amount for MBD–DNA binding detection on FACS 79
7.8 Suppressing DNA probe relabeling in multi-color FACS stains 79
7.9 Sensitivity and specificity by fluorophore in the two-color assay 80
7.10 Discrimination of surface-displayed DNA binders by their binding selectivity in the two-color assay 81
7.11 One-color FACS binding assay with wildtype MBD 82
7.12 Two-color FACS binding assay with wildtype MBD2 83
7.13 Screening the degenerated MeCP2 library against 15 modified CpG dyads 85
7.14 Post-screening phenotype enrichment 86
7.15 Single-color assay of candidate 5caC/5mC CpG-selective MBDs 87
7.16 Single-color assay of candidate 5hmC/5mC CpG-selective MBDs 88
7.17 One-color FACS binding assay with MeCP2[K109T/V122A/S134N] 89
8.1 Electrophoretic mobility shift assay of MeCP2-derived 5hmC/5mC binders 91
8.2 Binding affinity of T/A/Y/N and T/C/Y/N towards modified CpG dyads 92
8.3 Full binding selectivity of MeCP2[K109T/V122A/S134N] 93
8.4 Binding affinity of T/A/Y/N towards modified CpG dyads in two gene promoters 94
8.5 Dependency of binding affinity of T/A/Y/N towards modified CpG dyads 96
8.6 Mutational analysis on Ala-122 97
8.7 Circular dichroism spectroscopy of wildtype MeCP2 and MeCP2[K109T/V122A/S134N] 99
8.8 DNA duplex orientation in the MBD of MeCP2 100
9.1 Typical DNA–protein EMSAs 103

9.2 Macroscopic and microscopic descriptors of multistep binding 105
9.3 Modeling and fitting binding isotherms based on macroscopic descriptors of multistep binding 108
9.4 Modeling and fitting binding isotherms based on microscopic descriptors of multistep binding 113
13.1 Overview of plasmid vectors 151
A.1 Free and DNA-bound form of MBD1 167
A.2 Sequence alignment of the MBDs 168
A.3 Compilation of primary EMSA data evaluated for MBD1, MBD2 and MBD3 169
A.4 Compilation of primary EMSA data evaluated for MBD4 and MeCP2 170
A.5 Compilation of primary EMSA data evaluated for MeCP2 Rett mutants 171
A.6 Positions of the degenerated amino acid residues in additional MeCP2 libraries . 172
A.7 Specificity of Anti-c-Myc epitope detection of AIDA-I surface display 172
A.8 Whole cell lysate and outer membrane analysis of surface-displayed MBD1 and MBD2 172
A.9 Binding of MeCP2 to methylated and unmethylated CpA dinucleotides 173
A.10 Surface-display of various MBDs using a pBAD-based expression system 173
A.11 Effect of maturation time on surface display levels of active MBD 174
A.12 Co-detection of surface display level and DNA binding 174
A.13 Preparation of staining mixes for fluorescence-activated DNA binding assays ... 175
A.14 Saturation of the FACS DNA probe labeling reactions 175
A.15 Amount of DNA probe required for surface-displayed MBD staining 176
A.16 One-color FACS binding assay controls and surface-displayed wildtype MBD3 .. 176
A.17 Screening the degenerated MeCP2 library against 15 modified CpG dyads 176
A.18 Phenotype enrichment with different screening strategies 177
A.19 Binding affinity of T/A/Y/N and T/C/Y/N towards modified CpA · TpG dyads 177
A.20 Binding selectivity of A117C and other mutants for MTSL labeling 178
A.21 Effect of TEV cleavage on SpA(Z)–MBD fusion proteins 179

List of Abbreviations

Units of measure and of physical and chemical quantities. Compliant with the recommendations of the *International Union of Pure and Applied Chemistry.* Metric prefixes of the International System of Units (SI, «Système International d'Unités»), SI base units, SI derived units and non-SI units accepted for use with SI are not listed here.

Å	ångström; 1 Å = 1×10^{-10} m
bp	base pair
eq	molar equivalents
g	standard gravity equivalents
nt	nucleotide
OD	optical density
ppm	parts per million
ppt	parts per thousand
rpm	revolutions per minute
U	enzyme unit; 1 U = 1 µmol/min
vol	volume equivalents

Designations of nucleic acids, polynucleotides, amino acids, polypeptides and their constituents. In accordance with the IUPAC-IUB nomenclature of 1974 and 1983 with exceptions:

- The modifier for 2'-deoxyribonucleosides is omitted. No reference to ribonucleosides is made. Therefore, CpG ≡ d(CpG) is a 3'-5'-linked dideoxynucleotide ('dinucleotide').
- The 'codon triplet rule' is applied implicitly to all oligodeoxynucleotide sequences, designating from left to right the 3'-5' linkages of 2'-deoxyribonucleosides even without the triplet grouping. Therefore, AAAAAA ≡ AAA␣AAA ≡ d(ApApApApApA).
- The pair of modified cytosines C^1 and C^2 at CpG dyads in the two associated DNA chains is designated by the shorthand $C^1/C^2 \equiv (C^1\,\mathrm{pG}) \cdot (C^2\,\mathrm{pG})$.

Great care has been taken to avoid ambiguous usage of single-letter nucleobase and amino acid residue codes.

Other abbreviations and acronyms.

AGE	agarose gel electrophoresis
AIC	Akaike information criterion
AIDA-I	adhesin involved in diffuse adherence
APC	allophycocyanin
ATP	adenosine triphosphate
BCA	bicinchoninic acid
BER	base excision repair
BIC	Bayesian information criterion
BSA	bovine serum albumin
Btn	biotin
5caC	5-carboxycytosine
CD	circular dichroism
CFU	colony-forming unit
CGI	CpG island
DEER	double electron-electron resonance
Δ	differential
DMR	differentially methylated region
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease

DNMT	DNA methyltransferase
EDTA	ethylenediaminetetraacetate
EMSA	electrophoretic mobility shift assay
EPR	electron paramagnetic resonance
ESC	embryonic stem cell
FACS	fluorescence-activated cell sorting
FAM	6-carboxyfluorescein phosphoramidite
5fC	5-formylcytosine
FITC	6-carboxyfluorescein isothiocyanate
HEPES	2-(4-(2-hydroxyethyl)piperazin-1-yl)-ethanesulfonic acid
5hmC	5-hydroxymethylcytosine
IPTG	β-D-1-thiogalactopyranoside
LC-MS	liquid chromatography–mass spectrometry
LTR	long terminal repeat
MBD	methyl-CpG-binding domain
MBP	maltose-binding protein
5mC	5-methylcytosine
MTSL	1-oxyl-2,2,5,5-tetramethylpyrroline-3-methyl methanethiosulfonate
MWCO	molecular weight cut-off
NGS	next-generation sequencing
NMR	nuclear magnetic resonance
NTA	Ni-nitriloacetic acid
ODN	oligodeoxynucleotide
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PE	phycoerythrin
PMSF	phneylmethanesufonyl fluoride
PMT	photomultiplier tube
SpA	staphylococcal protein A
RMSD	root-mean-square distance
SAH	*S*-adenosyl homocysteine
SAM	*S*-adenosyl methionine
SDS	sodium dodecyl sulfate
SELEX	systematic evolution of ligands by exponential enrichment
SEM	standard error of the mean
SRA	SET and RING finger associated
SRAP	SOS response-associated peptidase
TALE	transcription activator-like effector
TAMRA	5-carboxytetramethylrhodamine
TDG	thymine DNA glycosylase
TEG	tetraethylene glycol
TET	ten-eleven translocation methylcytosine dioxygenase
Tris	2-amino-2-(hydroxymethyl)propane-1,3-diol
UMI	unique molecular identifier
ZNF	zinc finger

Introduction

and

Aim of This Work

Chapter 1
Epigenetic cytosine modifications in genomic DNA

Every individual life form must carry a comprehensive set of instructions that warrants the organism to survive, respond and adapt to its environment and that can be passed down from one generation to the next. In abstract terms, these instructions are referred to as *genes* and the entirety of the *genetic information* is the organism's *genome* (Brown, 2007). In virtually all recent life forms, the genetic information is stored in form of deoxyribonucleic acid (DNA). DNA is a linear polymer whose monomeric building blocks contain the nucleobases adenine (A), cytosine (C), guanine (G) and thymine (T ; Kossel, 1884; Kossel & Neumann, 1893; 1894). The information is thereby conveyed in the arrangement of the nucleobases which, for example, translates in the sequence of other biological macromolecules during gene expression (Crick, et al., 1961) or determines the placement of factors that initiate, terminate or modulate this and other vital processes (Jacob & Monod, 1961; Maston, et al., 2006).

As a matter of fact, not every gene's information is needed at all times. Especially in multicellular organisms like ourselves, different cells carry out specialized functions that require different genes to be active. It is easy to imagine that this differentiation into more than 200 cell types and their layout into over 70 organs in the human body requires strict orchestration of gene expression and repression. Since all cells in an individual organism share (with a few exceptions) the same DNA sequence (Gurdon, et al., 1975; Wilmut, et al., 1997; Yizhak, et al., 2019), regulatory mechanisms are necessary that collectively determine which part of the genetic information is accessible (Jaenisch & Bird, 2003). Collectively, *epigenetic information* embraces all "structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states" (Bird, 2007). Interestingly, some of these activity states perpetuate—like genes—after cell division to ensure cellular identity and can even transgress the germ line from parent to offspring (Moran, et al., 2021; Morgan & Whitelaw, 2008).[a] The underlying molecular mechanisms are plentiful and involve different biological macromolecules (Allis, et al., 2015).

Among them, modified DNA nucleobases hold an inimitable position because of their physical coupling to the carriers of the genes. In particular the biochemical modification of cytosine is deployed in many species across all domains of life not only to protect the integrity of their genomic DNA, but also to exert sustained control over gene expression (Casadesús & Low, 2006; Martienssen & Colot, 2001). To date, four carbon C5-substituted cytosine derivatives have been discovered in the genomic DNA of humans and mice (**Figure 1.1**). These are 5-methylcytosine (5mC; Hotchkiss, 1948) and the oxidized analogs 5-hydroxymethylcytosine (5hmC; Kriauciönis & Heintz, 2009), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC; Ito, et al., 2011).

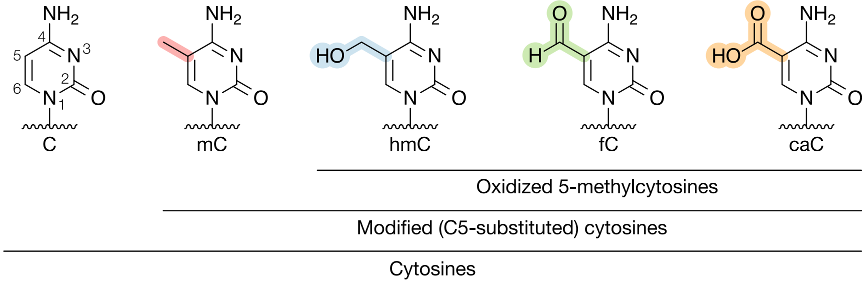

NH2
4
5
N3
2
6
N1
O
C
mC
HO
hmC
H
fC
caC
Oxidized 5-methylcytosines
Modified (C5-substituted) cytosines
Cytosines

Figure 1.1 Biochemical derivatives of cytosine. Unmodified cytosine (C) and the four modified cytosine derivatives with substitutions at carbon C5 in the pyrimidine ring, 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC). Naming and color conventions as applied throughout this work.

Of course, being a constituent (Section 1.1) is not per se equivalent to fulfilling a regulatory role. Yet a body of evidence has accumulated which supports an epigenetic function of cytosine modifications of which some even serve as transmissible marks that are passed along with the genomic DNA. This case is made namely by the presence of enzymes that 'write' and 'erase' these cytosine derivatives (Section 1.2) in a non-random, cell type-specific and long-lived manner (Section 1.3) at important sites in the genome where they promote distinct outcomes by various molecular mechanisms (Section 1.4) in still unexplored ways (Section 1.5).

In the ensuing chapters, their detection by technological means 'in the test tube' (Chapter 2) and by DNA-binding proteins in the nuclear proteome (Chapter 3) is summarized with a focus on their manifestation in the biologically relevant genomic DNA double-stand.

1.1 Abundance of modified cytosines

Like other non-canonical DNA nucleobases (Sood, et al., 2019), modified cytosines are a minor constituent of genomic DNA and account for less then 1% of all genomic cytosines in different species from all domains of life (**Figure 1.2a** and **Supplementary Table A.1**).

In *Mus musculus*. The most comprehensive data for modified cytosines in mammalian genomes has been collected from murine organs and cell lines (**Figure 1.2b**). The fraction of 5mC within all cytosines, modified or not, is 30 – 45 ppt (parts per thousand) which corresponds to 35 to 50 million 5mC sites across the haploid genome. The levels vary only little by cell type. 5hmC is one order of magnitude rarer than 5mC and most abundant in neuronal tissue such as the brain cortex with 6 – 7 ppt (6.5 to 7.5 million sites) and amounts to merely 0.6 – 0.8 ppt in other organs such as the liver. Embryonic stem cells (ESCs) have 5hmC levels around 1.2 – 1.6 ppt that tend to further decrease with replicative aging (Booth, et al., 2012). The level of the higher oxidized derivative 5fC is about 15 ppm (parts per million) in the brain (17,000 sites) and 6 ppm in the liver. This is two orders or one order of magnitude lower than 5hmC respectively. The total number of 5fC sites across different single ESCs at a time surmounts the number of common 5fC sites in the long-run which suggests a highly dynamic processes. 5caC is the rarest of the modifications with 3.5 ppm (3,800 sites) in ESCs and not rigorously observed in other somatic cell types (Carell, et al., 2018).

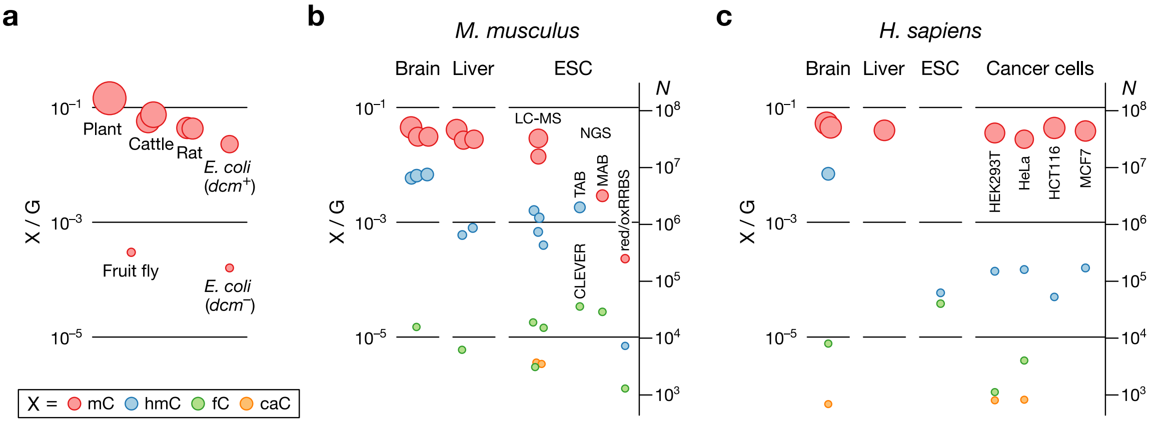


Figure 1.2 Content of modified cytosines in various genomes. (a) Abundance of 5-methylcytosine (5mC) in plant leaves *Arabidopsis thaliana*, calf thymus (*Bos taurus*), rat liver and brain *Rattus norvegicus*, the fruit fly *Drosophila melanogaster*, and a bacterium (*Escherichia coli*) harboring the DNA cytosine methyltransferase gene *dcm* or not. **(b)** Abundance of modified cytosines in different tissues and cell lines of the mouse (*Mus musculus*) as established by liquid chromatography–mass spectrometry (LC-MS) and by selected next-generation sequencing (NGS) protocols. **(c)** Abundance of modified cytosine in different human tissues, embryonic and cancer cell lines. Fraction of modified cytosines per guanine content; Absolute number *N* of sites in the haploid genome; Circles scaled with abundance; Raw data in Table A.1.

In *Homo sapiens*. The 5mC content in the human genome is relatively stable (40 – 45 ppt; 50 to 55 million sites; **Figure 1.2c**) even between unrelated individuals of different age (Eckhardt, et al., 2006). Similar to mice, the human brain contains slightly more 5mC than other tissues. Surprisingly little unbiased estimates for the oxidized 5-methylcytosines in human tissues and cell lines have been reported. One study (Liu, et al., 2013) finds 7.0 ppt 5hmC (8.5 million sites) in the brain, comparable to the levels observed in mice, while 5fC is about three order of magnitudes rarer (7.7 ppm, 9,300 sites) and 5caC nearly absent in this organ (0.8 ppm, 950 sites).

As compared to undifferentiated human ESCs, the content of oxidized cytosines varies between different cancer cell lines. The colorectal cancer cell line HCT116 has the lowest 5hmC content (48,000 sites) whereas the cervical cancer cell line HeLa has the highest 5fC content (3,800 sites).

1.2 Origin and turnover of cytosine modifications

Information emerges in absence of randomness (Hartley, 1928). Therefore it would be misleading to define the significance of modified DNA nucleobases only by their abundance (Breiling & Lyko, 2015). While the spontaneous deamination of cytosine into uracil, another rare, non-canonical DNA nucleobase, bears little information due to the randomness of this pervasive, naturally occurring, promutagenic DNA damage event, a locally confined enzymatic activity can signal specific states. A number of enzymes convert unmodified cytosine locally into 5mC, 5hmC, 5fC or 5caC and restore unmodified cytosine if need be. Thereby they act as 'writers' and 'erasers' of the potential epigenetic marks (**Figure 1.3a**).

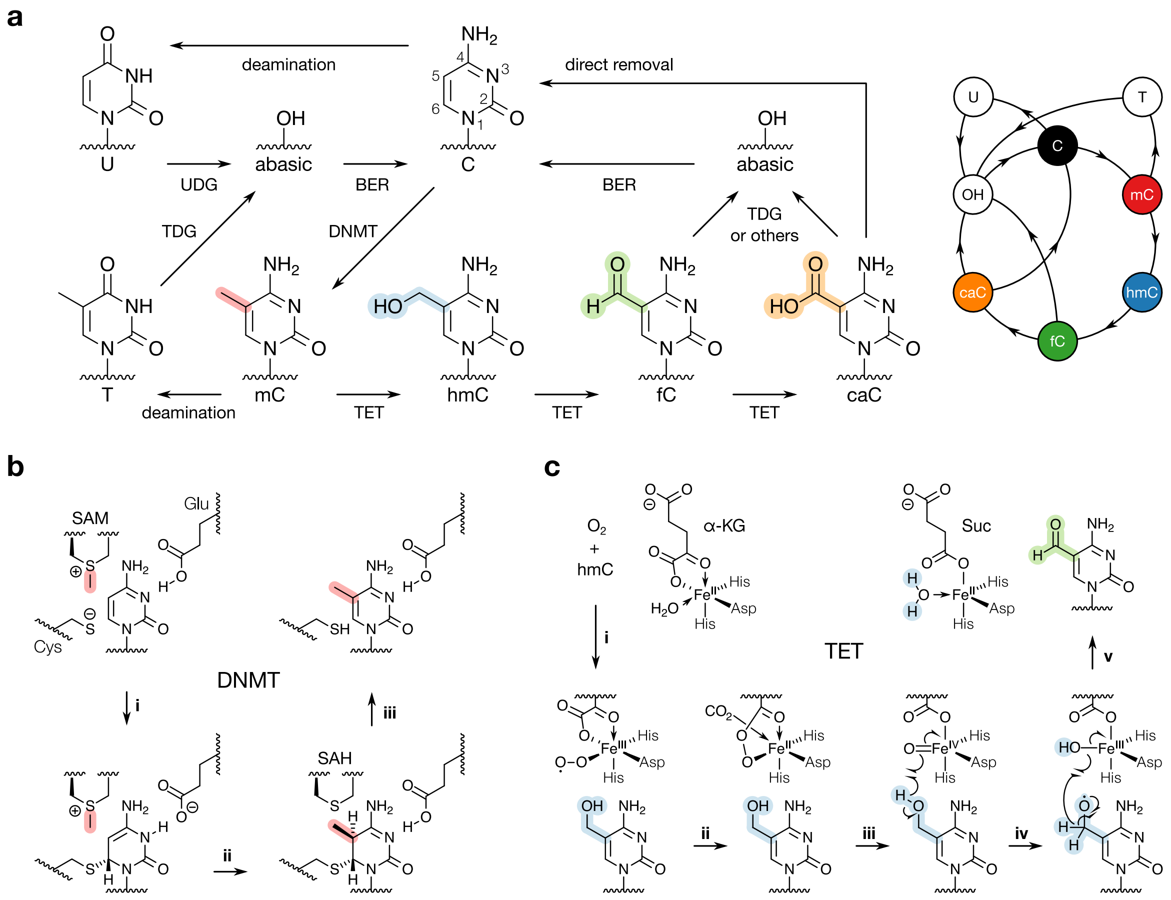


Figure 1.3 Natural processes modifying cytosine in mammalian genomes. (a) Cytosine (C) is converted by DNA methyltransferases (DNMTs) into 5-methylcytosine (5mC); Both, C and 5mC can spontaneously deaminate into uracil (U) or thymine (T), which calls for base excision and/or mismatch repair (not shown). The methyl group of 5mC is stepwise oxidized by ten-eleven translocation methylcytosine dioxygenases (TETs) into 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC), both of which are substrates for base excision repair (BER) mediated by thymine DNA glycosylase (TDG) or other processes (see main text) which finally restore C (Maiti & Drohat, 2011). **(b)** In the catalytic cycle of DNA methyltransferases (DNMTs), the methyl group of *S*-adenosyl methionine (SAM) is transferred onto cytosine in a three-step process. (i) Nucleophilic attack of a cysteine yields a protein–DNA conjugate as intermediate to which (ii) the methyl group of SAM is transferred, generating *S*-adenosyl homocysteine (SAH). (iii) Finally, aromaticity is restored by β-elimination and the next catalytic cycle can be initiated by deprotonation of cysteine (Du, et al., 2016; Song, et al., 2012). **(c)** The proposed catalytic cycle of ten-eleven translocation methylcytosine dioxygenase (TET) dioxygenases begins with (i) binding of the substrate (e.g., 5hmC), the cosubstrate α-ketoglutarate (α-KG) and molecular oxygen which oxidizes the Fe(II) center to Fe(III). (ii) Decarboxylation of the cosubstrate restores Fe(II) whilst generating a peroxy acid that (iii) undergoes heterolysis to yield succinate (Suc) and the active Fe(IV) center. (iv, v) The substrate is oxidized by sequential hydrogen abstraction and the next catalytic cycle is initiated after products and by-products are released (Hu, et al., 2015; Tarhonskaya, et al., 2019).

DNA methylation by DNMTs. 5mC is the product of three conserved enzymes, DNA methyltransferase 1 (DNMT1), DNMT3a and DNMT3b which use *S*-adenosyl methionine (SAM) as methyl donor to transfer a methyl group to position C5 of cytosine (**Figure 1.3b**; Du, et al., 2016; Song, et al., 2012). DNMTs act on cytosines in the short sequence palindrome CpG so that both cytosines in the double-stranded DNA duplex are modified in a strand-symmetric manner (Roy & Weissbach, 1975). These 'fully methylated' CpG dyads are the biochemical basis for the heredity of DNA methylation patterns after semiconservative DNA replication. The maintenance tandem DNMT1/UHRF1 is targeted to hemimethylated C/5mC dyads (reviewed in Bronner, et al., 2019) and is restored in a biphasic process, partly during S phase, partly post-replicative (Charlton, et al., 2018). The *de novo* DNA methyltransferases DNMT3a and DNMT3b can establish 5mC at non-methylated CpG dyads (Arand, et al., 2012; Ziller, et al., 2013) or the non-CpG sequences CpNpG (Clark, et al., 1995), CpA and CpT (Woodcock, et al., 1997), albeit to a much lesser degree.

Modification by TETs. 5hmC, 5fC, and 5caC are the product of stepwise oxidation by ten-eleven translocation methylcytosine dioxygenases (TETs; Ito, et al., 2011; Tahiliani, et al., 2009). The catalysis requires molecular oxygen and α-ketoglutarate (2-oxoglutarate) as cosubstrates and iron Fe(II) as cofactor (**Figure 1.3c**; Hu, et al., 2015). *In vitro*, TET enzymes oxidize modified and hemimodified CpG dyads in a non-processive manner (Hashimoto, et al., 2014; Tamanaha, et al., 2016; Xu, et al., 2014) independent of the modification on the complementary DNA strand (Crawford, et al., 2016; Hu, et al., 2013). However, TET1 and TET2 are about fourfold more active on 5mC-containing substrates than on 5hmC or 5fC (Hu, et al., 2015) because the pre-organization of the C5 substituents (see Chapter 3) elevates the energy barrier for hydrogen abstraction (Lu, et al., 2016). To date, there is no mechanism known to reconsolidate oxidized 5-methylcytosines after DNA replication in the newly synthesized strand. However, the modifications themselves are propagated with the DNA to the daughter cells (see Section 1.3).

Removal of modified cytosine bases. When maintenance methylation is missing or retarded, the number of fully modified CpG dyads will decrease from cell cycle to cell cycle, a process referred to as 'passive dilution'. Indeed, active modification can boost passive dilution as maintenance methylation by DNMT1 is slower at hemimodified C/5hmC, C/5fC or C/5caC dyads. This not often affects not only a single CpG dyad, but also of the ones in immediate neighborhood (Ji, et al., 2014; Valinluck & Sowers, 2007). The effect bears more on fast dividing cell.

Beyond this, 5fC and 5caC are subject to base excision repair (BER) by thymine DNA glycosylase (TDG; He, et al., 2011; Maiti & Drohat, 2011). The abasic sites are repaired according to the complementary DNA strand which reconstitutes cytosine, presenting an overall active demethylation pathway. The selectivity of TDG for 5fC and 5caC is due to the lower energy barrier for deglycosylation (Jeong, et al., 2020). Whether TDG is able to act on these substrates is determined by chromatin structure. At less accessible sites, their removal is impeded (Deckard III, et al., 2019). TDG is not the only 'eraser' of cytosine modifications. An additional pathway must be active particularly to erase cytosine modifications from pronuclear DNA after fertilization (Guo, et al., 2014; Song, et al., 2013). It may involve the SOS response-associated peptidase domain of Srap1 (Kweon, et al., 2017). Also DNMT1 could mediate decarboxylation of 5caC in absence of SAM by nucleophilic attack similar to the methylation mechanism (Liutkeviciute, et al., 2014). As intracellular levels of SAM are high, doubts have been raised whether this pathway is physiologically relevant *in vivo* (Carell, et al., 2018). Direct decarboxylation of 5caC (Schiesser, et al., 2012) and deformylation of 5fC (Iwan, et al., 2018) with other nucleophiles must also take place in cells as shown by stable isotope tracing (Carell, et al., 2018). The responsible enzymes are still to be identified (Kamińska, et al., 2021; Korytiaková, et al., 2021).

Taken together, writing and erasure of cytosine modifications is a dynamic process involving multiple biochemical reactions. This interplay may be crucial to establish not only the desired methylation pattern (Zhu, 2009) but also the desired pattern of other cytosine modifications.

1.3 Distribution of modified cytosines in the mammalian genome

Since DNMTs methylate cytosine almost exclusively in CpG dinucleotides and 5hmC, 5fC, and 5caC originate from 5mC, most modified cytosines are found in CpGs. 4 – 5% of all cytosines in the murine or human genome are part of CpGs, which amounts to 44 – 58 million sites (**Table 1.1**). In conjunction with the reported modification levels, 80 – 95% of all CpGs are modified on average, agreeing with earliest studies of dinucleotide digests (Sinsheimer, 1954; 1955).

Table 1.1 Distribution of cytosine in CpG and non-CpG contexts. Statistics on the double-stranded (dsDNA) haploid primary human and murine genome assemblies excluding unplaced contigs and mitochondrial DNA; H = A, C, T.

Organism	Base pairs	C	CHH	CHG	CpG	CpG dyads
H. sapiens[*]	3.09×10^9	1.20×10^9	8.88×10^8	2.54×10^8	5.80×10^7	2.94×10^7
	100. mol%	38.9 mol%	28.7 mol%	8.22 mol%	1.90 mol%	0.95 mol%
		100. mol%	74.0 mol%	21.1 mol%	4.89 mol%	4.89 mol%
M. musculus[†]	2.72×10^9	1.10×10^9	8.27×10^8	2.32×10^8	4.37×10^7	2.19×10^7
	100. mol%	40.5 mol%	30.4 mol%	8.51 mol%	0.80 mol%	0.40 mol%
		100. mol%	75.0 mol%	21.0 mol%	3.96 mol%	3.96 mol%

[*] Primary genome assembly for *Homo sapiens*, UCSC version hg38, with masked assembly gaps and intra-contig ambiguities; Bioconductor release 3.10, `BSgenome.Hsapiens.UCSC.hg38.masked`.

[†] As above, for *Mus musculus*; Bioconductor release 3.10, `BSgenome.Mmusculus.UCSC.mm10.masked`.

Overall however, the mammalian genome is depleted in CpG dinucleotides which is linked to deamination of 5mCpG which yields TpG dinucleotides that can persist unnoticed during DNA replication (Bird, 1980). The exception are CpG islands (CGIs), regions with significantly higher CpG levels than typical for the genome as whole (Gardiner-Garden & Frommer, 1987). Roughly 23,000 – 25,000 CGIs exist in the human and mouse genome that have a CpG every 10 – 15 bp (Illingworth, et al., 2010). Depending on the species, 10 – 20% of CpGs are found in CGIs or the adjacent 'shores' and 'shelves' (**Table 1.2**).

Table 1.2 Distribution of CpG dinucleotides. Number of CpG dinucleotides in CpG islands (CGIs) according to UCSC, the immediate flanking 2 kb (CGI shores), the next 2 kb (CGI shelves), and the remainder genomic segments (the 'open sea').

Organism	CpG islands*	CGI shores	CGI shelves	'Open sea'
H. sapiens	2.12×10^6	2.03×10^6	1.20×10^6	2.40×10^7
	7.22 mol%	6.91 mol%	4.08 mol%	81.6 mol%
M. musculus	1.05×10^6	9.41×10^5	5.77×10^5	1.93×10^7
	4.82 mol%	4.30 mol%	2.64 mol%	88.1 mol%

* Including potential CpG islands in repeat regions, i. e. the repeat unmasked version of the genome assemblies.

Various sequence elements of the genome—gene bodies, promoters, enhancers, transposons, and intergenic regions—can overlap with CpGs and therefore possibly contain a modified cytosine that may or may not modulate the physiological function of the element.

For example, the observation that more than 60% of all human gene promoters and more than 50% of all promoters in the mouse contain a CGI (see Data Source, page 190) which are active in most tissues (Saxonov, et al., 2006), has stimulated interest in the functional role CGIs since such coincidence is rather unlikely given that gene promoters hold only a small share of the total genomic sequences (Deaton & Bird, 2011). With the availability of whole-genome sequencing and methods that allow the mapping of cytosine modification at single-base resolution, differentially methylated regions (DMRs) have gained increasing importance. DMRs are genomic sites for which the DNA methylation state differs between cell types or changes dynamically during development and differentiation. 20% of DMRs in human fall into CGIs or CGI shores (Schultz, et al., 2015), leaving a vastly unexplored landscape of differentially methylated CpGs in the 'open sea'. In turn, non-CGI promoters constitute a large fraction of DMRs in various human tissues (Eckhardt, et al., 2006). The concept of DMRs can be extended to other modified cytosines as 'differentially modified regions.'

5-Methylcytosine. The evaluation of 5mC marks across the genome is complicated by the fact that bisulfite sequencing, the frequently employed 'gold standard' to evaluate DNA methylation on a genomic scale also detects 5hmC (Chapter 2.1). Given the lower abundance of 5hmC, the general trends might still hold true, though contradictory observations have been reported.

In general, all categories of sequence elements can be methylated (reviewed in Suzuki & Bird, 2008) and only CGIs remain largely unmethylated. Especially at transcription start sites CGIs bear less than 10% 5mC as well as low 5hmC. 5mC levels are inversely correlated with CpG density (Booth, et al., 2012). Since CGIs contain only 5 – 10% of all CpGs, mammalian genomes appear to be 'globally' methylated (Suzuki & Bird, 2008). In the mouse genome, varying levels of methylation (10 – 50%) are observed at distal regulatory regions coinciding with the repressive epigenetic marker histone H3K9me3 and protein p300 (Stadler, et al., 2011). Many of these regions are differentially methylated in different tissues (Hon, et al., 2013). DNA methylation can also coexist with the active H3K27ac marker in enhancers (Charlet, et al., 2016).

During development, about 6 – 8% of CGIs eventually become methylated in the course of differentiation, e. g., into blood, brain or muscle tissue (Illingworth & Bird, 2009; Illingworth, et al., 2008). In embryonic stem cells, a considerable 15 – 20% of 5mC (or 5hmC; my annotation) exists in non-CpG contexts, particularly at CpA dinucleotides (Chen, et al., 2011), suggesting that embryonic stem cells may use different methylation mechanisms to regulate gene expression (Lister, et al., 2009). Likewise, both modified cytosines accumulate at specific CpG and CpH sites (H = A, C, T) in the fetal cortex during brain development in neurons independent of the examined individual, but are absent in the supporting glia (Lister, et al., 2013). The most prominent sequence motifs of CpH methylation is TNCACN (N = A, C, G, T), specifically TACA**C** in neuronal cells and TACA**G** in embryonic stem cell and induced pluripotent stem cells (Schultz, et al., 2015).

5-Hydroxymethylcytosine. 5hmC is enriched at gene bodies, promoters, and transcription factor binding sites varying with cell type and developmental stage (reviewed in Shi, et al., 2017). Also during embryonic development, the distribution of 5hmC is highly dynamic and correlates with lineage-commitment and tissue-specific processes (reviewed in Zhu, et al., 2018). In mice, 5hmC accumulates during postnatal aging in gene bodies related to neurodegenerative disorders (Song, et al., 2011; Szulwach, et al., 2011).

5-Formylcytosine and 5-carboxycytosine. Not only the total amount of 5fC varies depending on the tissue (Bachman, et al., 2015), but also its genomic distribution is tissue-specific (Iurlaro, et al., 2016). Although global 5fC levels are lower than 5hmC, modification levels can rise comparably to 5hmC levels at specific genomic loci (Booth, et al., 2014). At the single-base level, 5fC and its precursor 5hmC overlap in only 20% of the cases, indicating that they may have different biological roles (Xia, et al., 2015).

In mouse embryonic stem cell (ESC), 5fC was relatively enriched around the transcriptional start sites of CGI promoters (along with 5hmC), but not at non-CGI promoters, especially when marked with H3K4me3 (Neri, et al., 2015). 5fC is found also at distinct long terminal repeats and satellite repeats (Raiber, et al., 2012). By single-cell sequencing along embryonal development, earlier stages (oocytes, sperm, pronuclei, and 2-cell) had more 5fC sites in common than later stages (4-cell, inner cell mass, trophectoderm). Further, intragenic 5fC sites seem more conserved than intergenic sites (Zhu, et al., 2017), confirming the higher intragenic 5fC levels observed in bulk measurements (Booth, et al., 2014).

5fC and 5caC are found in H3K4me1 marked regions associated with active or poised gene transcription (Shen, et al., 2013; Song, et al., 2013; Wu & Zhang, 2014), at active enhancers and in gene bodies (Lu, et al., 2015). Herein, 5fC is present in exons and promoters, whereas 5caC is found in introns. Also, 5fC has been found at gene promoters critical for development and metabolism prior to their expression during embryonic development (Zhu, et al., 2017).

5caC is enriched at promoters before their transcription in the course of lineage specification and differentiation (Lewis, et al., 2017).

Longevity at specific genomic sites. It may be argued that the observed genomic distributions could be but a snapshot of an active demethylation process rather than an active modification pathway. However at some loci, 5mC, 5hmC, and 5fC are stable (or semistable) modifications in ESCs *in vivo* as revealed by stable isotope labeling (Bachman, et al., 2015; Bachman, et al., 2014) and by hairpin bisulfite sequencing (Guo, et al., 2014). Also in human early preimplantation embryos, 5mC and 5hmC seem to be stable over several developmental stages; However, neither 5fC nor 5caC have been detected, potentially due their low abundance (Okamoto, et al., 2016).

1.4 Biological consequences of DNA cytosine modification

If modified cytosines represent meaningful biological information, their presence or absence will have specific consequences. Indeed, specific cytosine modifications are essential at certain loci and associated with imbalanced levels observed in disease.

(1.4.1) Correlative evidence

DNA methylation is required during major developmental transitions in mammals (reviewed in Smith & Meissner, 2013) for example in embryonic stem cells (ESCs) to suppress gene expression at germ line imprint control regions (Morgan, et al., 2005), transposable elements as well as pericentromeric repeats to allow proper chromosome alignment (Lehnertz, et al., 2003). Also in the adult, DNA methylation is tied with pluripotency and terminal differentiation, including *de novo* silencing of repetitive elements or dampening of transcriptional noise from intragenic regions (Bird, 1995; Huh, et al., 2013). Whereas DNA methylation in promoter regions leads to stable gene silencing (Korthauer & Irizarry, 2018), methylation of exons and introns can coexist with active transcription. So, 5mC is often but not exclusively a repressive epigenetic mark.

However, development cannot be thought without change. The oxidized 5-methylcytosines 5hmC and 5fC are specifically present during two waves of epigenetic reprogramming at early embryonic stages in mammals (reviewed by Zhu, et al., 2018). After fertilization, the maternal and especially the paternal genome of mice undergo active DNA demethylation by Tet3 and at later stages by Tet1 and Tet2 (Guo, et al., 2014; Wang, et al., 2014). Indeed, 5mC and 5hmC have been found to oscillate in circadian rhythms in mouse liver and lung at distinct genomic sites that may be lost during aging (Oh, et al., 2018). At certain genomic loci, fast methylation/ demethylation cycles are observed in response to external stimuli (Kallenberger, et al., 2019; Kangaspeska, et al., 2008; Zhang, et al., 2020).

5hmC is found near active (Ficz, et al., 2011; Stroud, et al., 2011) and repressed genes (Williams, et al., 2011; Wu & Zhang, 2011a; Zhang, et al., 2016). Its relationship with activation or repression of gene activity may therefore be contextual during development and dependent on the cell type, incorporating chromatin state, histone modification, and other epigenetic marks (Shi, et al., 2017). When present in gene bodies, 5hmC is typically correlated with gene activation (Lin, et al., 2017; Ponnaluri, et al., 2017; Wu & Zhang, 2011a). Also 5fC and 5caC are more frequent around transcription start sites of actively transcribed genes (Neri, et al., 2015).

Epigenetic aberrations in DNA methylation can overwrite physiological cell behavior in cancer (reviewed by Ortiz-Barahona, et al., 2020) and other diseases (reviewed by Wu, et al., 2020). For example, DNA methylation within a CpG island encompassing the bidirectional promoter of *BRCA1* and *NBR2* contributes to repression of *BRCA1*, a tumor suppressor gene relevant to non-hereditary, sporadic breast cancer (Rice, et al., 2000). Similarly, DNA hypermethylation of the *CDKN2A* promoter or its first exon has consequences for cell-cycle dysregulation in various cancers (Ran, et al., 2016) very similar to those mutational inactivation of this gene would have.

Reduced levels of 5hmC are associated with tumorigenesis in acute myeloid leukemia, multiple myeloma, and melanoma (Bonvin, et al., 2019; Chatonnet, et al., 2019; Han, et al., 2016; Jin, et al., 2011; Lian, et al., 2012), but not in glioma (Kraus, et al., 2015), and has therefore recently been proposed as tumor marker in liquid biopsies (Xu & Gao, 2020). In contrast, increased levels of 5hmC are found at genes involved of hippocampal cells upon exposure to early-life stress with a link to anxiety-related behavior (Papale, et al., 2017) and after exposure to acute stress situations (Li, et al., 2016). Further, aberrant 5hmC levels are linked to neurological and psychiatric disorders, including Rett syndrome (Szulwach, et al., 2011), Alzheimer's disease (Chouliaras, et al., 2013; Condliffe, et al., 2014; López, et al., 2017), Huntington's disease (Villar-

Menéndez, et al., 2013), Fragile X-associated tremor/ataxia syndrome (Yao, et al., 2014), and Ataxia-telangiectasia (Jiang, et al., 2015). All of these highlight the role of oxidized 5-methylcytosines in neuronal tissue, where highest abundances are observed.

(1.4.2) Mechanistic linkage

Modified DNA nucleobases can alter the interaction interface for DNA-biding proteins (reviewed in Zhu, et al., 2016), which will be examined in Chapter 3 for the case of modified cytosines. But which cellular events link those alterations to the observed biological effects?

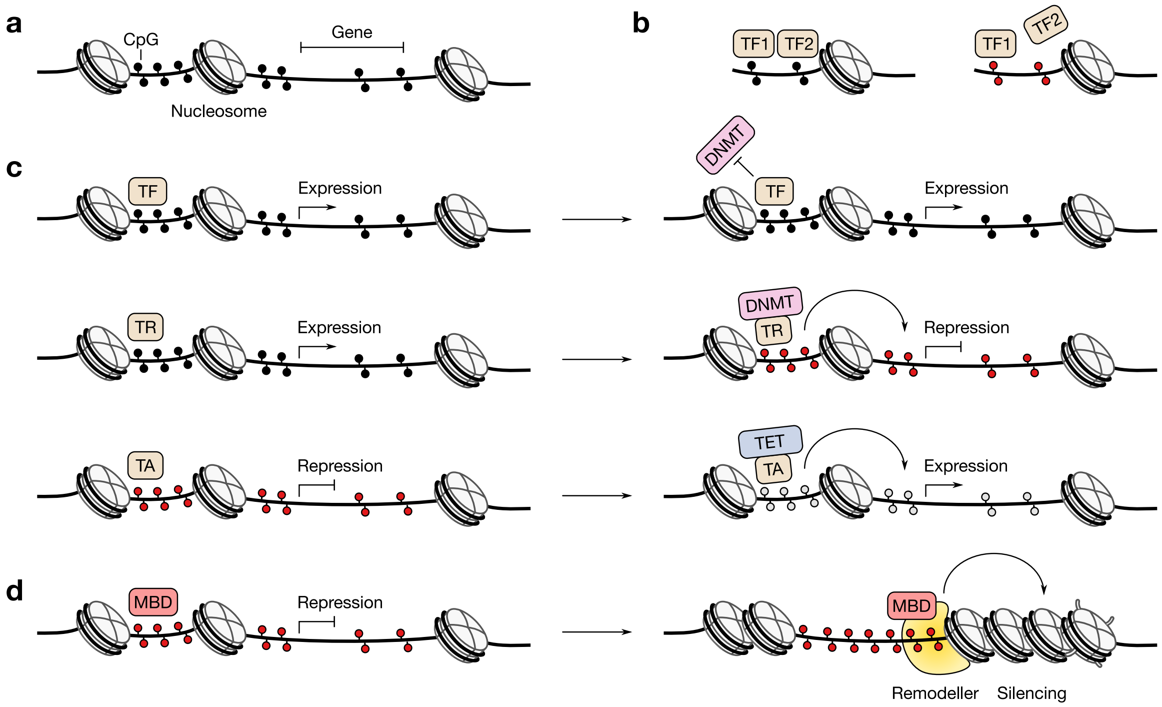


Figure 1.4 Biological effects of DNA methylation in the context of DNA-binding proteins. (a) Some genetic and epigenetic elements in genomic DNA. **(b)** Different effects of DNA methylation (red circles) on DNA recognition for an agnostic (TF1) and sensitive (TF2) transcription factor. **(c)** Local protection, reversal and reinforcement of epigenetic transcriptional states mediated via sequence-dependent factors and secondary effectors. **(d)** Sequence-independent recruitment of chromatin remodelers by methyl-CpG-binding domain (MBD) proteins, for example at highly methylated regions. Panel *c* after Blattler and Farnham (2013), CC BY 4.0 (full license in Appendix A.4).

Gene expression starts with local transcriptional activation. Out of 500 human transcription factors, less than 10% were found indifferent towards the presence of methylated CpGs in their consensus binding sequence while only 25% were repelled and 35% even preferred DNA methylation for binding (Yin, et al., 2017). A subset of these factors can elicit drastic chromatin changes, allowing for cell fate decisions during development or their reversal in cellular reprogramming (Iwafuchi-Doi & Zaret, 2014). Of these, Klf4 (Hu, et al., 2013; Wan, et al., 2017) and

Oct4 (Yin, et al., 2017) were found to have additional secondary or tertiary binding sites that appear only when these DNA sequences are methylated. Other factors important for cellular differentiation such as the GATA family are recruited indirectly, but through methylation-dependent transcription factors such as FOXA (Zhu, et al., 2016).

Indeed, there is extensive cross-talk between DNA methylation and factors recruited to protect, reverse or reinforce the DNA methylation status of neighboring sequences (Krebs, et al., 2014), often affecting the local chromatin state, i. e., how compact the genomic DNA is packed to restrict access to the genetic information. This involves DNA-binding proteins that have rare genomic recognition sequences (reviewed in Blattler & Farnham, 2013) as well as proteins that interact with modified cytosines in sequences that can be as short as a single CpG dinucleotide. Some members of the methyl-CpG-binding domain protein family, for example, recruit nucleosome remodeling complexes (Fuks, et al., 2003; Leighton & Williams Jr, 2019) and other modifiers of structural chromatin proteins to genomic regions that are heavily methylated in a sequence-independent manner (Baubec, et al., 2013; Menafra, et al., 2014).

Beyond 5mC, the oxidized cytosine derivatives 5fC and 5caC did perpetuate or even enhance transcriptional repression at a single gene promoter in a TDG-dependent manner (Kitsera, et al., 2017). In this regard, these modifications are not only passive intermediates of an active demethylation pathway but possess regulatory function by themselves.

1.5 Strand-symmetric and -asymmetric cytosine C5 modifications in CpG dyads

The intriguingly diverse biological roles of the five modified cytosine derivatives in the mammalian genome have been discussed so far almost exclusively as if each modified cytosine nucleobase would exist in isolation of other derivatives at a distinct genomic locus—not only for the sake of clarity in this introductory chapter, but also in the cited literature. However, cytosine modification takes place in CpG dinucleotides in the double-stranded DNA of the mammalian genomes where the short sequence palindrome CpG is present also on the reverse-complementary strand forming a *CpG dyad* (Wu & Zhang, 2017), i. e., a (CpG) · (CpG) pair [C/C for short] composed of two C · G base pairs (**Figure 1.5a**).[b] Since all 'writers' of modified cytosines in mammals can act on both cytosines present in a CpG dyad independent of the cytosine modification on the complementary strand (Section 1.2), several combinations of *strand-symmetrically* or *strand-asymmetrically* modified and/or unmodified cytosines in CpG dyads can exist, depending on whether the two C5 substituents in the dyad are identical or not (**Figure 1.5b**).

For the biological role of modified cytosines this means that in the context of the DNA double-helix 'signal hybrids' could exist or that several of these combinations exert a regulatory function by themselves—two vastly underexplored hypotheses.

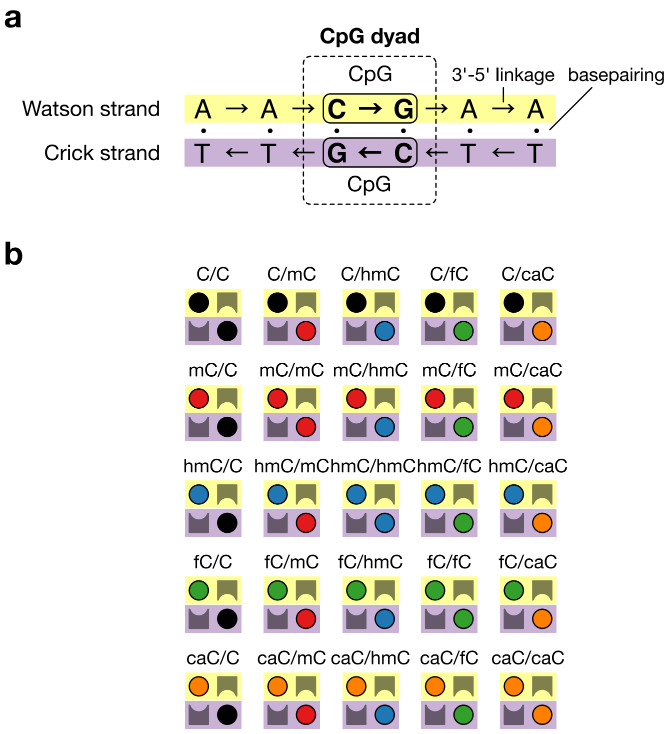


Figure 1.5 CpG dinucleotide dyads in double-stranded DNA. (a) In the two reverse-complementary strands of double-stranded DNA, the palindromic CpG dinucleotide is present on both, the Watson and the Crick strand, designated a 'CpG dyad'. **(b)** The enzymatic processes of Figure 1.3 in combination with DNA replication can potentially give rise to several distinct combinations of unmodified and modified cytosines in CpG dyads. 'Hemimodified' dyads bear one unmodified cytosine. Likewise, 'fully modified' dyads bear two C5 substituents.

The technological difficulties and recent advances in the simultaneous discrimination between two or more different cytosine C5 modifications will be addressed in Chapter 2 and their potential to elicit different biological outcomes through the proteins that interact with double-stranded genomic DNA in Chapter 3.

1.6 Synopsis

1. Some cytosine modifications are actively introduced into the double-stranded DNA of the mammalian genome, predominantly at CpG dinucleotides.
2. 5mC, 5hmC and 5fC can be stable or semistable modifications beyond cell cycle progression. In the case of 5mC, the modification is hereditable with a known biochemical basis.
3. Content and genomic distribution of the modified cytosines vary between tissue, cell type, degree of differentiation, and aging. This gives rise to 'differentially modified regions' and is characteristic for a carrier of biological information.
4. Modified cytosines are found at key regulatory regions for gene expression, namely pro-

moters, enhancers, introns, and exons. Many examples illustrate the regulatory capacity of these modifications by correlative and mechanistic analyses.

5. Modified cytosines integrate with other layers of epigenetic regulation.
6. Cytosine modification in CpG dyads gives rise to strand-symmetric and strand-asymmetric combinations of C5-modified cytosines. Their biological role is largely unexplored.

Endnotes

[a] The term ‘epigenetics’ has taken a change in meaning since its first introduction by Conrad Waddington in 1942 (reviewed in Deichmann, 2016). Some authors require the effect on the genotype–phenotype relation to be hereditable, but not by virtue of the canonical DNA sequence.

[b] This usage is distinct from the term ‘dyad symmetry’ which is used in molecular biology to designate the presence of a sequence and its reverse complement on the same DNA strand (Adhya, 2001) as well as the term ‘(dyad) asymmetry’ used in genetics to designate allelic differences in the sequence constitution or modification between sister chromatids, the dyads, i. e., paternally and maternally inherited genetic material (Vu, et al., 2000).

Chapter 2

Technologies to detect modified cytosines in DNA

Our ability to understand the function of cytosine modifications in their genomic context vastly depends on the performance of the technologies available for their detection. They can be grouped roughly into four categories (**Figure 2.1a–d**): (1) Methods that use physical separation or detection as technology to identify the different nucleobases such as two-dimensional thin-layer, gas or liquid chromatography (Breter, et al., 1976; Razin & Sedat, 1977; Wu & Zhang, 2011b), mass spectrometry (Carell, et al., 2018), redox- and electrochemical sensors or similar readouts (Korlach & Turner, 2012; Wescoe, et al., 2014); (2) Methods that promote C to T transitions by chemical or enzymatic means in order to reveal the modified nucleobases by DNA sequencing at nucleotide-resolution (reviewed in Raiber, et al., 2017; Zhao, et al., 2020); (3) Methods that exploit the unique chemical reactivity of the DNA modification (Gieß, et al., 2019; Okamoto, et al., 2006; Tanaka, et al., 2007) or selective enzymes (Robertson, et al., 2011; Song, et al., 2016) to conjugate for example affinity tags for the purification of modified DNA fragments; And (4) methods that use protein-based affinity reagents to directly detect the modified positions in the DNA strands (Nair, et al., 2011) with or without sequence context-specificity (further reading in Buchmuller, et al., 2021; Kubik & Summerer, 2015).

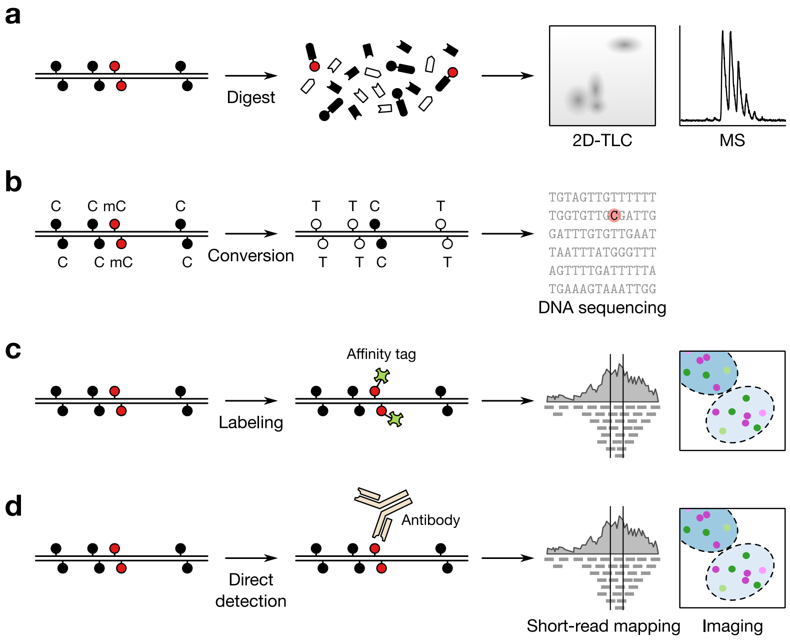


Figure 2.1 Technologies to detect modified cytosines. (a) Physical separation of DNA nucleobases, e. g., by enzymatic digest and resolution, e. g., by two-dimensional thin-layer chromatography (2D-TLC) and identification, e. g., by mass spectrometry (MS). **(b)** Chemical conversion of unmodified cytosine and identification by DNA sequencing. **(c)** Chemoenzymatic labeling of methylated cytosine to introduce, e. g., an affinity tag for enrichment of DNA fragments followed by short-read mapping or a fluorescence probe for imaging. **(d)** Direct detection of modified cytosines, e. g., using an antibody, which can be applied in similar ways as in *c*.

All these methods differ in sensitivity, specificity, resolution and scope of spacial, temporal or other contextual information that they can provide. They have been applied to reveal global, genome-wide modification landscapes (Iqbal, et al., 2011; Susan, et al., 1994) down to the resolution of epigenetic marks at defined genomic loci in individual cells (Muñoz-López, et al., 2020; Zhu, et al., 2017) and thereby allowed to study the dynamics of modified cytosines during development and in disease in various genomes (reviewed recently by Parry, et al., 2021).

Relevant to the problem addressed in this work are experimental strategies to determine the combination of modified cytosines concurrently present at individual CpG dyads. To aid the understanding of some catches in DNA sequencing-based approaches, the chemical conversion of modified cytosines is briefly explained in Section 2.1 before the relevant literature is presented in Section 2.2.

2.1 Detection of modified cytosines in the genomic sequence context

DNA sequencing technologies provide indispensable information about the arrangement in which the individual DNA nucleotides are linked together if a sufficient number of identical DNA strands can be analyzed (Liu, et al., 2012). The great fidelity and single-nucleotide resolution results from the unique base pairing between A · T and C · G during *in vitro* amplification of DNA by polymerase chain reaction (PCR; Saiki, et al., 1985) in presence of synthetic nucleobase analogs chosen according to the sequencing method (Fahnestock, et al., 1991; Ronaghi, et al., 1996; Sanger, et al., 1977). Since cytosine C5 substituents do not participate in base pairing, they are 'invisible' for this process. Therefore, preparatory chemical and/or enzymatic conversion is required to change the base pairing pattern according to the nature of the C5 substituent. For example, the treatment with bisulfite followed by a strong base can promote deamination of cytosine, but not 5-methylcytosine or 5-hydroxymethylcytosine. This yields uracil which pairs like thymine during PCR and DNA sequencing (**Figure 2.2a**; Frommer, et al., 1992). As milder alternatives, Friedländer synthesis (Zhu, et al., 2017), borane reduction (Liu, et al., 2019) and enzymes (Li, et al., 2018; Schutsky, et al., 2018) can be used for conversion.

To differentiate other sets of modified cytosines, additional chemical or enzymatic treatments must precede the bisulfite conversion step (**Figure 2.2b**; reviewed by Zhao, et al., 2020).

2.2 Detection of concurrently modified cytosines in CpG dyads

Various attempts have been made to elucidate the prevalence of concurrently modified cytosines in single CpG dyads (compare Figure 1.5 in Chapter 1.5) by different technologies:

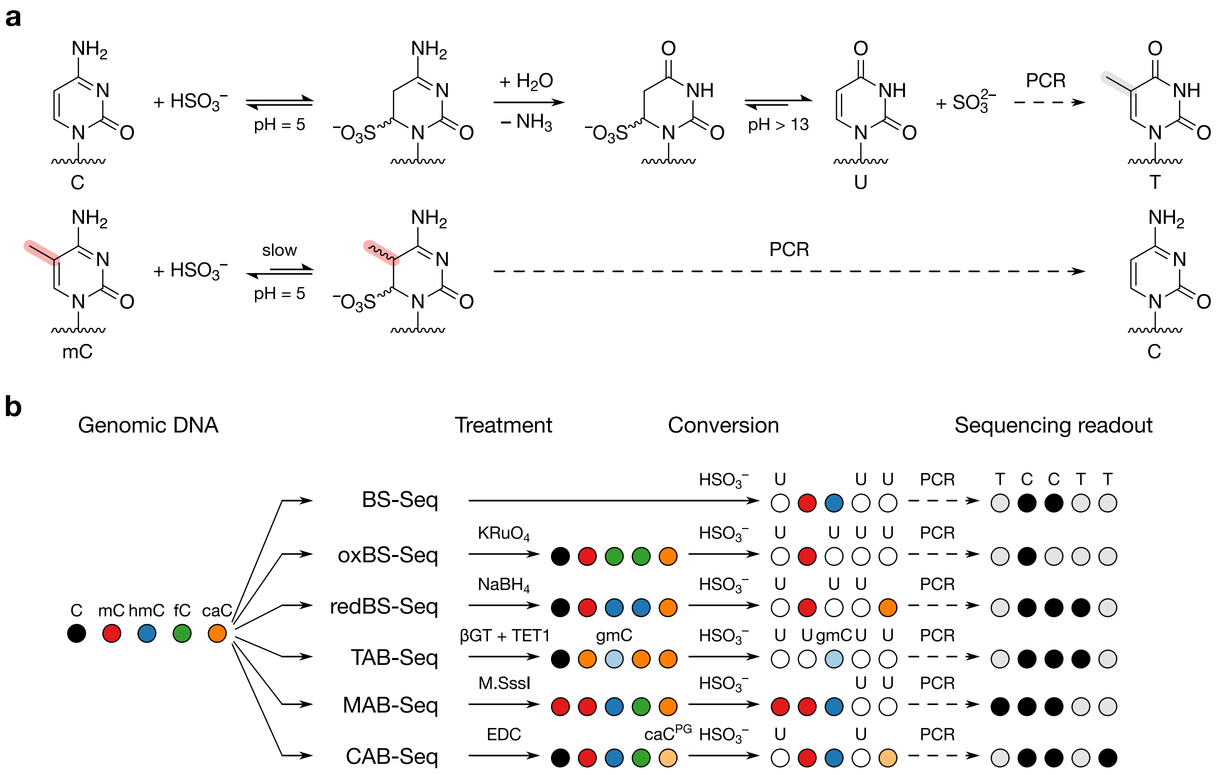


Figure 2.2 Bisulfite conversion of C5-modified cytosines. (a) Reaction of bisulfite with unmodified cytosine (C) and 5-methylcytosine; The addition at C6 is used to bring about deamination of the intermediate at C4, but the attack is slowed down by +I substituents such as a methyl or hydroxymethyl group at C5. **(b)** Examples of single-nucleotide resolution methods for mapping modified cytosines in genomic DNA; bisulfite sequencing (BS-Seq; Adey & Shendure, 2012); Chemical oxidation or reduction in oxidative BS-Seq (oxBS-Seq; Booth, et al., 2012) or reductive BS-Seq (redBS-Seq; Booth, et al., 2014); Enzymatic modification in TET-assisted BS-Seq (TAB-Seq; Yu, et al., 2012) or M.SssI-assisted BS-Seq (MAB-Seq; Wu, et al., 2014); Chemical protection using a carbodiimide (EDC; CAB-Seq; Lu, et al., 2013).

Using conversion-based DNA sequencing. The readout in conversion-based DNA sequencing approaches is binary. One or several modified cytosines are sequenced either as C or as T. In other terms, these approaches provide evidence about the presence or absence of an entire set of modifications, e. g., about the presence of 5mC or 5hmC when 'C' is detected in bisulfite sequencing, and C, 5hmC, 5fC or 5caC when 'T' appears in the sequencing trace. So, the 'true' modification at a genomic position cannot be resolved other than in combination with complementary information from an appropriate second conversion-based sequencing experiment that uses a different treatment (**Figure 2.3a**; Booth, et al., 2014; Liu, et al., 2021).

A complication to achieve DNA duplex-resolution in these experiments is the diploidy of somatic mammalian cells, i. e., the presence of a paternal and a maternal copy with almost the same DNA sequence. Each copy can be modified differently at the same genomic positions what is known as 'symmetric' or 'asymmetric' modification and has been observed for all four cytosine derivatives at imprinted genes (Vu, et al., 2000; Zeng, et al., 2019). So, an 'asymmetric' modification of the paternal and maternal genome is often indistinguishable from a strand-asymmetric modification present in both copies. To discriminate '(a-)symmetry' and 'strand-(a-)symmetry', hairpin adapters have been introduced to covalently link the complementary strands of each DNA duplex (Burden, et al., 2005; Giehr, et al., 2018; Laird, et al., 2004). In fact a probably indispensable improvement for the correct interpretation of such experiments.

No matter if or how the duplex information is preserved, a single DNA strand cannot be converted more than once neither in practical terms (Grunau, et al., 2001) nor in chemical terms due to the terminal defunctionalization of certain C5 substituents.[a] This implies that the frequency of each C5 modification at a single genomic position must be inferred from averages of many hundreds of genomes using probabilistic models (Xu & Corces, 2018; Äijö, et al., 2016). However, only if the modification levels within a CpG are fully penetrant, e. g., 100% for one of the cytosines and 0% (or 100%) for the cytosine in the complementary strand at the same CpG, then the 'true' combination can be unambiguously inferred (**Figure 2.3b**).

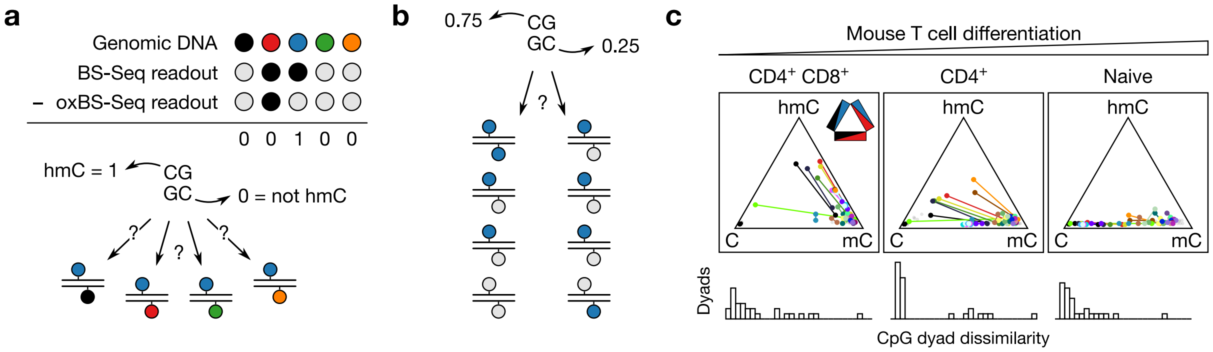


Figure 2.3 Inferring strand-(a-)symmetry from conversion-based DNA sequencing methods. (a) A combination of bisulfite sequencing (BS-Seq) and oxidative BS-Seq (oxBS-Seq) can identify 5hmC, but the identity for other cytosines remains elusive. **(b)** With intermediate cytosine modification levels in the bulk population for at a single CpG dyad, also the specific combination in the individual double-stranded DNA molecules is ambiguous. **(c)** Probabilistic models can integrate information from multiple sequencing experiment; Ternary plot of the likeliest cytosine modification present at a genomic position (dots) in 30 CpG dyads (connected with a line) at different stages during mouse T cell differentiation; Re-analysis of Äijö et al. (2016).

Nevertheless, also for intermediate modification levels said models allow some insights: For example, based on the likeliest C, 5mC and 5hmC frequencies across 30 CpG dyads during T cell differentiation, the number of 5hmC/5mC dyads tend to steadily regress whereas the number of C/5mC dyads increases before almost all dyads become either C/C or 5mC/5mC in the naive (i. e., more differentiated) T cells (Äijö, et al., 2016; **Figure 2.3c** [my own evaluation]).[b]

On a genome-wide scale, Neri et al. (2015) combined MAB-Seq with previously published TAB-Seq data for mouse embryonic stem cell (Habibi, et al., 2013) to find that—rather expectedly—5mC levels in CpG dyads on one strand correlated strongly (Pearson correlation $r = 0.82$) with 5mC levels on the complementary strand, arguing for predominant 'strand-symmetric' fully methylated 5mC/5mC dyads. Similarly, an analysis of human embryonic stem cell by TAB-Seq found 92% of all CpG dyads being strand-symmetrically methylated (Yu, et al., 2012). In contrast, 5hmC and 5fC or 5caC (the latter being indistinguishable in MAB-Seq) showed weak correlation ($r = 0.15$ and $r = 0.01$ respectively) hence strand-asymmetric CpG dyad modification (Neri, et al., 2015). Likewise, Yu et al. (2012) reported only 20% of all 5hmC-containing CpG dyads to be strand-symmetric 5hmC/5hmC dyads.

Similar conclusions were drawn from oxidative and reductive bisulfite sequencing (oxBS-Seq and redBS-Seq) of CpG islands in mouse embryonic stem cell. Booth et al. (2014) found that methylation levels within a CpG dyad differed by only 18% on average, whereas modification levels of 5hmC or 5fC differed by 43 – 46% on average. Yet even in these highly asymmetric dyads, 5mC levels differed but by only 20%, suggesting that 5mC oxidation by TET dioxygenases rather than reconstitution of 5mC by Dnmt1 would possibly drive this imbalance.

This idea has been revisited by Tamara L. Davis' lab focussing on CpG dyads at important differentially methylated regions (DMRs) during development in the mouse. By adopting a hairpin bisulfite protocol to disentangle the DNA duplexes from the different parental genomes, they found DNA hydroxymethylation in one DMR to correlate with hemimethylated C/5mC dyads in other DMRs in a parent-dependent manner (Guntrum, et al., 2017; Nechin, et al., 2019). Using a similar approach, Patiño-Parrado et al. (2017) found hemimethylated C/5mC and differentially modified 5mC/5fC dyads in DMRs of murine neuronal nuclei.

Considering that with conversion-based DNA sequencing approaches one must combine multiple data sets *in silico*, it is intriguing that similar patterns of strand-(a-)symmetrically modified CpG dyads can be detected at the same genomic positions across hundreds or thousands of genomes, despite some remaining ambiguity.

Using chemoselective labeling. More direct evidence comes from labeling modified cytosines *in situ*. Song et al. (2016) detected strand-asymmetrically modified CpG dyads in the double-strands of native genomic DNA using fluorescent probes that would specifically react either with 5mC or 5hmC, the two most prevalent modified cytosines in the mammalian genome (Ito, et al., 2011). The authors estimated that 60% of all 5hmC were present in strand-asymmetric 5hmC/5mC dyads in various murine tissues with different global 5hmC content. Fluorescent labeling of 5fC using a pyrene-hydrazine probe demonstrated that strand-symmetric 5fC/5fC dyads are formed more frequently during *in vitro* TET oxidation than hemimodified 5fC dyads (Xu, et al., 2014), yet their prevalence in genomic DNA still remains to be determined.[c]

Despite the irrefutable evidence in support of different combinations of cytosine C5 modifications at CpG dyads, these methods have not been applied to date to purify or these sites.

Using protein-based affinity reagents. In contrast to technologies that convert or modify DNA nucleobases, proteins allow to non-covalently probe the chromatin in its almost unperturbed, native state. Some of these proteins can even be transfected or expressed in cells to monitor or modulate the biological processes that contribute to or rely on DNA modification *in vivo*. Antibodies (Jin, et al., 2010; Mohn, et al., 2009), transcription activator-like effectors (TALEs; Rathi, et al., 2016; Zhang, et al., 2017), and methyl-CpG-binding domains (MBDs; Aberg, et al., 2017) are widely used to capture, purify, and analyze DNA fragments that contain modified cytosine nucleobases but have also been used to visualize their genomic distribution *in situ*. Alas, TALEs interact with the nucleobases on one of the DNA strands only and there is no antibody specific for a combinations of modified cytosines in CpG dyads to date (**Table 2.1**).

Table 2.1 Reactivity and specificity of commercially available affinity agents.

Clone	Epitope	Cat. No.	ssDNA*	dsDNA*	CpG dyad-specificity	Other reactivity
RM231	Anti-5mC	ab214727	yes	yes	no	none with C or 5hmC
33D3	Anti-5mC	ab10805	n/d	n/d	n/d	n/d
5MC-CD	Anti-5mC	ab73938	n/d	n/d	n/d	n/d
D3S2Z	Anti-5mC	cst28692	yes	yes	no	none with C, 5hmC, 5fC, 5caC
RM236	Anti-5hmC	ab214728	yes	yes	no	none with C or 5mC
AB3/63.3	Anti-5hmC	ab106918	n/d	n/d	n/d	none with C or 5mC
HMC31	Anti-5hmC	cst51660	yes	yes	no	none with C, 5mC, 5fC, 5caC
Polyclonal	Anti-5fC	ab231898	n/d	n/d	n/d	n/d
D5D4K	Anti-5fC	cst74178	yes	yes	no	none with C, 5mC; weak 5hmC, 5caC
Polyclonal	Anti-5caC	ab231801	n/d	n/d	n/d	none with C, 5hmC, 5fC, 5caC
D7S8U	Anti-5caC	cst36836	yes	yes	no	none with C, 5mC, 5hmC; weak 5fC

* Reactivity observed in single-stranded (ssDNA) and/or double-stranded DNA (dsDNA); n/d = not determined.

Although sequential affinity-enrichment with a combination of these reagents could in theory retrieve double-stranded DNA fragments that contain both modifications, the presence of multiple CpGs would often be detrimental to the accurate determination of the dyad modifications. There is no literature pertinent to this idea.

As an alternative to antibodies and TALEs, MBD protein family members have been used to enrich DNA fragments that contain fully methylated 5mC/5mC CpG dyads (Bock, et al., 2010; Brinkman, et al., 2010; Rauch & Pfeifer, 2005). The only engineered MBD2 variant for hemimethylated C/5mC CpG dyads (Heimer, et al., 2015) has not found application to date, maybe because of the small difference in binding affinity between C/5mC and 5mC/5mC CpG dyads and a weak, but noticeable affinity towards unmethylated C/C dyads.

2.3 Synopsis

1. Genome-wide and targeted conversion-based DNA sequencing indicates that unmodified and methylated cytosine occur most frequently in strand-symmetrical C/C and 5mC/5mC CpG dyads. Oxidized 5-methylcytosines are part of strand-asymmetrically modified dyads. The resolution of their exact constitution is limited by intrinsic analytical shortcomings.
2. Chemoselective labeling of genomic DNA finds 5hmC to face 5mC in 60% of all 5hmC-containing CpG dyads. The method has not been adapted to closer examine these sites.
3. No suitable biochemical reagent exists to date to examine specific combinations of cytosine C5 modifications in CpG dyads.

Endnotes

[a] One exception being that theoretically the presence of, e. g., 5hmC could be probed after bisulfite conversion using chemoenzymatic labeling. This has not been reported to the best of my knowledge.

[b] The original data set was reduced to CpG dyads with information available on the modification levels for both cytosines to connect these data points in Figure 2.3c.

[c] Given a detection limit of 10 nM for symmetric 5fC/5fC dyads in this method (Xu, et al., 2014) and a concentration of 2.5 pM for one copy of the genome in a mammalian nucleus (1,320 fL, Purkinje cell, BNID 103181), as little as 4,000 5fC/5fC dyads could theoretically be detected if the experiment were conducted at genomic DNA concentrations. This is just at the edge of total 5fC sites in the brain (9,300 , Table A.1).

Chapter 3

Recognition of modified cytosines in DNA

The recognition of nucleic acids by proteins is a fundamental molecular process in biology. Of particular importance are the interactions with double-stranded DNA, the carrier of the genetic information (Chapter 1). This information is conveyed in the arrangement of the four canonical deoxyribonucleotides and to our current knowledge interpreted in two ways:

In the first embodiment, the linear sequence instructs the making of other biological macromolecules including the replication of DNA itself. The information is transmitted through the unique pairing between a purine and a pyrimidine nucleobase; the same principles that hold together the complementary strands in the DNA double-helix (Watson & Crick, 1953). In consequence, transcribing or copying this layer of information requires separating the DNA strands. Proteins that participate in these processes interact with the nucleic acids in a mostly sequence-independent manner.

In the second embodiment, it is the subtle, sequence-dependent changes in the shape of the DNA double-helix and the spacial arrangement of the nucleobases with their unique physicochemical properties which expose this information along the DNA grooves. Readily accessible without dismantling the double-strand, this superficial interface can be used as a scaffold to coordinate the assembly of defined protein complexes or to regulate the access to specific parts of the genetic information prior to their expression.

How different C5 modifications of cytosine alter this latter interface in double-stranded DNA is summarized in the first part of this chapter (Section 3.1) and the second part highlights some examples how these alterations are recognized by naturally evolved proteins (Section 3.2). The methyl-CpG-binding domain protein family is examined in detail (Section 3.3).

3.1 DNA base and shape readout by DNA-binding proteins

The sequence-specificity of the protein–DNA interaction on the part of the DNA double-strand is determined by 'direct' (base-related) and 'indirect' (shape-related) factors. Proteins exploit both in order to create a tight, complementary protein–DNA interface (Gromiha, et al., 2004).

Base-specific factors include all physicochemical properties of the nucleobases that are exposed at the grooves of the DNA double-strand (Seeman, et al., 1976). The pattern of non-polar groups, hydrogen bond donors and acceptors are predominantly examined in the DNA major groove where they prove most different to each other (**Figure 3.1a–c**). At the minor groove, these patterns differ but in their electrostatic potential when seen in sequence context, thereby warranting sequence-specific minor groove recognition (Chiu, et al., 2017).

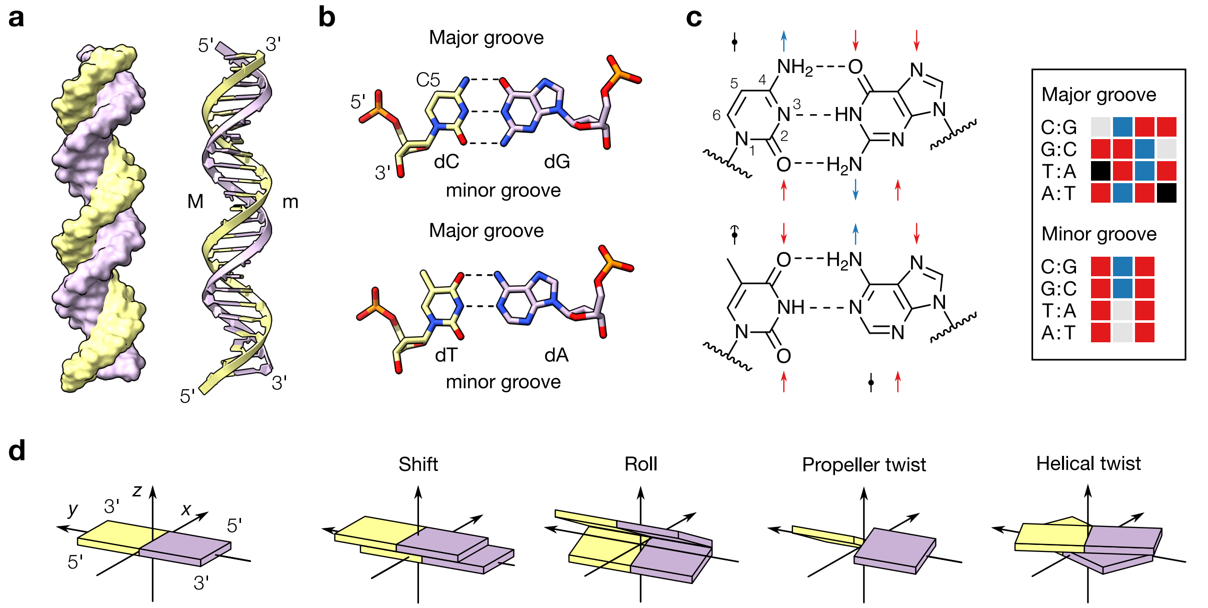


Figure 3.1 Base and shape readout of methylated DNA. (a) Van der Waals surface and helix skeleton in a fibre model of a 24 base pair double-stranded DNA in B-form conformation. The complementary single strands are colored and the position of a major (M) and minor (m) groove is exemplified. **(b)** Watson-Crick base pairing of deoxycytdidine (dC) with deoxyguanosine (dG) and deoxythymidine (dT) with deoxyadenosine (dA) via hydrogen bonds (dashed lines). The surface-exposed grooves in B-form DNA are indicated. Note the three-dimensional position of the 5'-phosphate an 3'-hydroxyl groups of the deoxyribose as well as the absence of functional groups at the C5 position in dC. **(c)** Skeletal formula of the base pairing in *b* with the pattern of hydrogen bond donors (blue) or acceptors (red) and non-polar groups (black) for each pyrimidine/purine nucleobase pair in either grooves. **(d)** Schematic diagrams of the rigid body transformations implied by the base pair parameter propeller twist (rotation along *y*) and the base step parameters slide (displacement along *y*), roll (rotation around *y*), and helix twist (rotation around *z*) that describe most base shape effects (Lu & Olson, 2003).

Such direct readout can involve tightly bound water molecules (Fuxreiter, et al., 2005; Joachimiak, et al., 1994) which originate from 'spine hydration' of either the major or the minor groove in B-form DNA (Dickerson, 1992).

Indirect readout of the DNA shape is based on the sequence-dependent structural non-uniformities of the double-helix. These are deviations from B-form DNA mostly in terms of groove width, roll, helix and propeller twist as shown in **Figure 3.1d** (Lawson & Berman, 2008). Also the flexibility of the DNA double-strand varies with sequence. Generally, purine-pyrimidine steps (RpY, i. e., GpC, GpT, ApC, ApT) are more rigid than the corresponding pyrimidine-purine steps (YpR, i. e., CpG, TpG, CpA, TpA) and facilitate bending of the helix (Gorin, et al., 1995; Olson, et al., 1998). As a consequence, a contiguous series of three or more ApA dinucleotides in the context of a YpR step can give rise to anomalously high curvatures towards the major groove (Beveridge, et al., 2004). Intriguingly, the relative affinities for different binding sequences of many DNA-binding proteins can be accurately described using a linear combination of the shape parameters associated with the dinucleotides, but not by the averaged mononucleotide properties (Rube, et al., 2018). It is therefore probably more appropriate to state DNA to be composed of ten dinucleotides rather than four monomers when it comes to DNA shape recognition.

(3.1.1) Effect of cytosine modifications on shape readout

In the literature various accounts on cytosine C5 modifications altering the conformation of the DNA duplex are documented, including a B–Z transition for 5-methylcytosine (5mC) (Béhé & Felsenfeld, 1981; Fujii, et al., 1982; Mooers, et al., 1995) and helical under-winding for 5-formylcytosine similar to A-form DNA (Raiber, et al., 2015). However, in none of these cases had the conformational changes been different from an (often omitted) unmodified analogue (Hardwick, et al., 2017; Hodges-Garcia & Hagerman, 1995) and they are hence assumed to be crystallization artifacts. Therefore, the notion has been probably overstressed that cytosine modifications operate *not* by altering DNA shape readout; yet, some effects are noticeable.

Effect of 5-methylcytosine. Owing to the close proximity of the methyl groups in fully methylated 5mC/5mC dinucleotides, the roll angle at this YpR step tends to increase (Tippin & Sundaralingam, 1997), leads to widening of the major groove.[a] Although these structural changes are minimal, they are sufficient to bias the rate of deoxyribonuclease I (DNase I) cleavage at methylated CpGs 10- to 20-fold (Dantas Machado, et al., 2015; Lazarovici, et al., 2013).

Effect of oxidized 5-methylcytosine. Whereas methylated DNA stretches are stiffer because the bulky methyl groups counteract DNA bending, 5hmC and 5fC significantly increase DNA flexibility (Ngo, et al., 2016). This is linked to the increased hydrophilicity of the oxidized variants and ensuing alteration of the DNA solvent shell (reviewed in Rausch, et al., 2019). The effect of 5mC oxidation has also been revisited by Fu et al. (2019), eliminating potential flaws from sequence context or crystallization conditions. The authors confirm that the overall conformation remains B-form DNA and any of the cytosine modifications (5mC, 5hmC, 5fC, and 5caC) have similar effects on the DNA double-helix with few exceptions. The roll angle widens along C, 5fC, 5hmC, 5mC and increases up to almost +5° for 5caC, leading to more bended structures. Also, major grooves containing 5mC or 5hmC are more open than the unmodified ones while those containing 5fC or 5caC are slightly more narrow. Both, the minor grooves at 5mC and even more at 5caC are significantly larger by about 0.5 Å and 1.0 Å respectively.

In summary, cytosine modifications affect DNA flexibility and DNA shape but moderately, potentially with the exception of 5caC which shows stronger distortion. These altered physical properties can be sensed by some DNA repair enzymes, including thymine DNA glycosylase (Fu, et al., 2019) and may affect other proteins that interact with double-stranded DNA.

(3.1.2) Effect of cytosine modifications on base readout

Modifications at position 4, 5 (or 6) in the pyrimidine ring of cytosine can alter the interaction interface at the DNA major groove in B-form DNA. Modifications at other positions disrupt the Watson-Crick base pairing and have only been described in the context of DNA damaging nucleobase adducts (Delaney & Essigmann, 2004; Saparbaev & Laval, 1998).

Effect of 5-methylcytosine. 5-Methylcytosine creates the opportunity for non-polar interactions to take place at the DNA major groove. It has therefore been argued that it would prevent the binding of water molecules and hence disrupt spine hydration (Frederick, et al., 1987; Tippin, et al., 1997). Later studies by Mayer-Jung et al. (1998) have shown that water can still be co-ordinated through non-conventional C–H···O hydrogen bonds, leading to a rearrangement of water molecules at the inner hydration shell. So, the methyl group can be either recognized directly by non-polar or directly by polar amino acid residues.

Effect of oxidized 5-methylcytosine. Clearly, the presence of other functional groups at carbon C5 installs specific physicochemical properties in the major groove: The hydroxymethyl group of 5hmC adds a hydrogen bond donor, a formyl group an additional hydrogen bond acceptor, and the carboxyl group of 5caC, which is deprotonated at physiological pH, creates a negative electrostatic surface potential. Steric effects and the presence of partial charges in the oxidized 5-methylcytosine derivatives dictate the stereochemical conformation of these groups. The C4–C5–C5α axis takes distinct dihedral conformations for different substituents. Whereas the methyl group of 5mC is configured anti-periplanar or synclinal, the hydroxyl group of 5hmC is most favorably placed synclinal, but has rotational freedom. As the carbonyl groups of 5fC and 5caC allow for hydrogen bonding with the N4 amino group, they prefer a syn-periplanar conformation. However, in context of the DNA double-strand, the repulsive forces produced by the negative charge of the carboxyl group of 5caC with the phosphodiester backbone will ‘squeeze’ the nucleobase away (see above; Fu, et al., 2019).

In addition, cytosine C5 modifications affect the strength of the base pairing within the DNA double-strand. Although the presence of an oxidized 5-methylcytosine modification is not followed by imino tautomerization of the exocyclic N4 amino group and therefore does not reverse the hydrogen bonding pattern at the major groove (Szulik, et al., 2015), 5fC and 5caC increase the acidity of N3 and thereby weaken the base pairing with guanine albeit not to the point of wobble base pairing (Dai, et al., 2016).

How these distinct physicochemical fingerprints at the DNA major groove are recognized on the molecular level, is best illustrated by their protein interaction partners.

3.2 Protein domains that recognize C5-modified cytosines

Proteins that replicate, erase or interpret the genetic or epigenetic information contained in DNA need necessarily to interact with the DNA double-strand. They do so using various structural motifs (**Figure 3.2a**) to engage with different parts of the double-helix (**Figure 3.2b**). Kind and number of these non-covalent contacts differ, but involve direct interactions between the two macromolecules such as electrostatic forces, van der Waals forces, hydrogen bonding, and cation–π interactions as well as indirect interactions mediated or driven by solvent or solutes. Herein, *recognition* will imply a tolerance or in best case a preference of the protein domain for interacting with a specific DNA nucleobase, although some biological effects, for example of modified cytosines, are also due to hinderance of protein engagement (see Chapter 1.4).

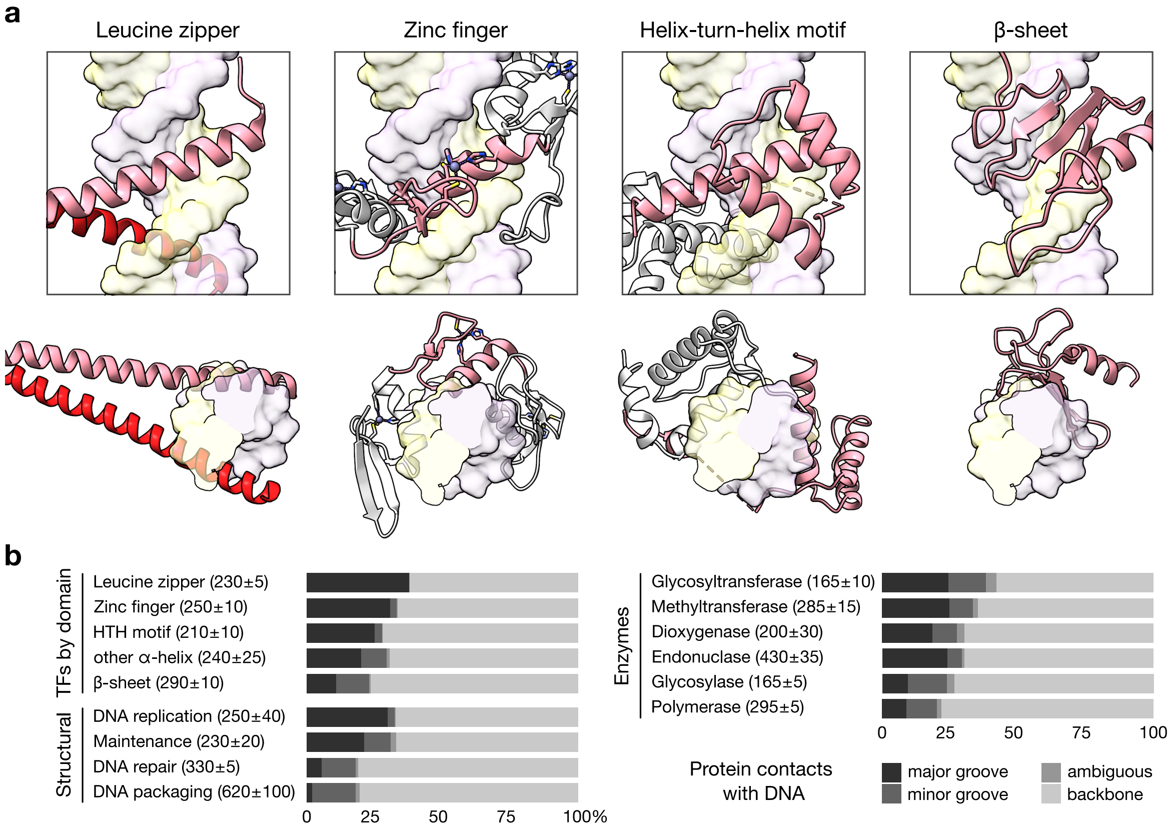


Figure 3.2 Double-stranded DNA-binding domains and contact statistics. (a) The basic leucine zipper of the Jun/Jun AP-1 complex (PDB 5t01), one of three zinc fingers of Kaiso (PDB 4f6n), the helix-turn-helix motif in PBX1 (PDB 1b72), and the β-sheet-dependent DNA-binding domain of Tn 916 integrase (PDB 1tn9). **(b)** Average protein–DNA contacts at the effective atomic interface for selected transcription factors, structural DNA-binding proteins and enzymes (mean ± SEM) as reported by Norambuena and Melo (2010).

A review of different protein domains that bind modified cytosines has been published by Ren et al. (2018) and is decently supplemented by Muñoz-López and Summerer (2018). My focus is on domains which simultaneously bind *both* strands of the DNA double-helix.

(3.2.1) Secondary structural elements to probe the DNA double-strand

An α-helix fits snugly into the major groove of B-form DNA (Zubay & Doty, 1959) and a variety of placements are observed (Suzuki, et al., 1995). This allows choosing the interacting residues more independently from another, which may facilitate structural diversification (Connolly, et al., 2000). Indeed, many eukaryotic transcription factors classify as either leucine zipper, helix-turn-helix, or zinc fingers protein (Lambert, et al., 2018) that recognize up to four nucleobases of their DNA target per α-helix (Suzuki & Gerstein, 1995). They form more contacts with the DNA major groove than those domains that harness β-strands (Norambuena & Melo, 2010).

Examples of proteins that use β-strands are indeed rare, maybe because multiple distant sites must be fitted to match the curvature of the DNA double-strand (Tateno, et al., 1997). With respect to the thermodynamics of the interaction, Connolly et al. (2000) pointed out that "the β-sheet as well as B-from DNA are rather rigid structures, which makes it harder to release ions and water molecules from the contact surface to drive the binding entropically." Base-specific interactions with the grooves alone are therefore often insufficient for high-affinity binding and the number of unspecific backbone contacts increases. However, this may allow for shorter DNA binding motifs. Many β-sheet domains bind to the narrow minor groove of B-DNA where a β-strand is just thin enough to fit (Church, et al., 1977; Tateno, et al., 1997). A prominent example of this sort is the TATA-box binding protein (TBP; PDB 1ytb, PDB 1tgh). However, in order to interact with C5 substituents of modified cytosine nucleobases directly, the β-sheet would have to face the DNA major groove.

Neither of these two secondary structural elements is employed by transcription activator-like effectors (TALEs) that probe the DNA major groove—although on one of the strands only—via two amino acid residues on a sharp pointed loop that is fixed between an array of α-helices.

(3.2.2) Leucine zippers

Some eukaryotic transcription factors of the basic leucine zipper (bZIP) and the helix-loop-helix (bHLH) superfamilies recognize modified cytosine bases via residues on at least one of the leucine zipper's α-helices that project into the DNA major groove on two diametrically opposed faces of the double-helix.

5-Methylcytosine. The bZIP Jun/Jun complex of AP-1 binds a 7 bp sequence starting either with thymine or 5-methylcytosine (5mC) while the ATF4/Jun (or ATF4/Fos) complex of AP-1 has even higher affinity when its target sequence starts with 5mC rather than thymine (reviewed in Ren, et al., 2018). The interaction takes place via a non-polar van der Waals contact of alanine (PDB 5t01). Methyl-SELEX (Yin, et al., 2017) did not reveal any of these proteins, but suggested to include the bZIP domains of MYF6, HLF, C/EBPε, C/EBPγ, and C/EBPβ (PDB 6mg2). The latter was also identified by CpG-methylated DNA microarrays (Mann, et al., 2013) and contacts four methyl groups, two in each strand, of which one resides in a hemimethylated 5mCpG dyad and is recognized via an methyl-arginine-guanine triad and the other one in a 5mCpA via a non-polar van der Waals contact of a valine side chain (Yang, et al., 2018).

5-Formylcytosine. The bZIP transcription factor NRF1 binds DNA only if its target sequence is non-methylated (Domcke, et al., 2015) and some data suggested that it would preferentially interact with 5fC (Spruijt, et al., 2013). A structure of this complex has not been reported; the molecular basis for this interaction is therefore still unknown.

5-Carboxycytosine. The bHLH MAX tolerates a 5caC modification of the CpG within its target sequence (Wang, et al., 2017). The interaction is mediated by hydrogen bonding with an arginine side chain (PDB 5eyo). Similarly, bHLH TCF4 adopts a conformation (PDB 6od5) in which Arg-576 contacts 5caC, which is relevant for the Pitt-Hopkins syndrome (Yang, et al., 2019).

(3.2.3) Zinc fingers

Members of the 'reader' (MBD1), 'writer' (DNMT1) and 'erasers' (TET1 and TET3) of 5mC in mammalian genomes contain a CXXC zinc finger domain in their full-length protein sequence. These domains, if functional, allow to recruit the protein to specific genomic loci.

5-Methylcytosine. At least 40 or 50 different zinc-finger containing transcription factors have been shown or are suspected by various experimental techniques to bind an 5mC-containing DNA (Hu, et al., 2013; Spruijt, et al., 2013; Yin, et al., 2017; Zhu, et al., 2016). Some recognize the target in a strand-asymmetric and sequence context-dependent manner.

The C2H2 class proteins Kaiso (ZBTB33), ZBTB38 and ZBTB4 contain three consecutive zinc fingers of which the first two target the DNA major groove and the third interacts with the minor groove. These zinc fingers bind fully methylated CpG dyads, but also hemimethylated C/5mC CpGs and TpGs (summarized in Sasai, et al., 2010). In the structures solved for Kaiso, 5mC (PDB 4f6n) or T (PDB 4f6m) are contacted via a methyl-arginine-guanine triad (Liu, et al., 2013). Another well-studied C2H2 zinc finger is the murine Zfp57 which is expressed during early embryogenesis. The second of its two consecutive zinc fingers contacts the fully methylated 5mC/5mC CpG dyad strand-asymmetrically (PDB 4gzn): One 5mC is contacted via a methyl-arginine-guanine triad, whereas the other one via a structured water network (Patel, 2016). A CpA upstream of this CpG, i. e., the opposite strand's TpG, is also recognized via a methyl-arginine-guanine triad (Ren, et al., 2018). YY1 (PDB 1ubd) contains multiple methyl-arginine-guanine triads: The initial CpG can be methylated, greatly reducing binding affinity. Further C2H2 examples are the pioneering transcription factor Klf4 (PDB 4m9e) and the transcriptional regulators WT1 (PDB 4r2e) and EGR1 (PDB 4r2a; murine Egr1, PDB 1a1g) which respond differentially to 5mC oxidation (see the following sections) and contain multiple methyl-arginine-guanine triads (Rooman, et al., 2002).

Like zinc fingers, p53 uses a zinc ion to coordinate to its DNA binding site and was found to bind stronger to a methylated target sequence containing T or 5mC by means of Arg-280 (Kribelbauer, et al., 2017). This Arg-280-TpG complex (PDB 1tsr) has been shown to form a methyl-arginine-guanine triad (Rooman, et al., 2002).

5-Hydroxymethylcytosine. The C2H2 class zinc finger of SALL4 prefers 5hmC over 5mC. The structural basis of this interaction has not yet been elucidated (Xiong, et al., 2016).

5-Formylcytosine. The C2H2 zinc finger EGR1 has been crystallized in complex with its target sequence containing a single 5fC (PDB 4r2d). The interaction is tolerated, but clearly destabilized as compared to the methylated target (Hashimoto, et al., 2014). Likewise for p53 which has been implicated in 5fC-recognition (Spruijt, et al., 2013).

5-Carboxycytosine. The CXXC class zinc finger of TET3 binds 5caC/5caC with threefold higher affinity than C/C in a CpG dyad (Jin, et al., 2016). A lysine side chain amino group and a backbone amide of an isoleucine contact the carboxyl group of one 5caC. The opposing 5caC is contacted by serine, threonine and glutamine side chains (PDB 5exh).

WT1 (PDB 4r2r) can recognize 5caC, while other C2H2 fingers cannot. This is because they have a negatively charged glutamate residue where WT1 interacts favorably with 5caC via Gln-369 (Hashimoto, et al., 2014). Indeed, the Zfp57[E182Q] mutant (PDB 4m9v) allows binding of 5caC (Liu, et al., 2013). CTCF, an eleven C2H2 zinc finger DNA-binding protein that plays a key role in the chromosomal architecture of mammalian genomes, was found to bind alternative 5caC-containing targets (Nanan, et al., 2019) which may be linked to its binding of 5mC-depleted regions and protection of unmethylated CpG islands from gaining methylation (Feldmann, et al., 2013).

(3.2.4) **Helix-turn-helix motifs**

5-Methylcytosine. The homeobox factors PBX1 and POU1F1 were predicted to bind a DNA target containing a methylated CpG (Yin, et al., 2017). In the available crystal structures with unmethylated DNA targets (PDB 1b72 and PDB 5wc9), the fold proofs to be able to recognize an (unrelated) TpG using a methyl-arginine-guanine triad (Rooman, et al., 2002). PBX1 and a couple of further homeobox and forkhead factors were also identified by quantitative mass-spectrometry (Spruijt, et al., 2013).

5-Hydroxymethylcytosine. The thymocyte nuclear protein 1 (Thy28) has been suggested to act as a specific 'reader' of 5hmC (Spruijt, et al., 2013). In the crystal structure of the human ortholog (PDB 5j3e, SGC), a 5mC-containing DNA double-strand is contacted via the minor groove with no base-specific contacts.

(3.2.5) **Base-flippers**

Base-flipping is operated during the enzymatic 'writing' of 5mC by DNMTs and the 'writing' of oxidized 5-methylcytosines by TETs as well as in a number of other eukaryotic and prokaryotic proteins that act on modified DNA nucleobases (Hong & Cheng, 2016). A structurally distinct, non-enzymatic base flipper is the SET and RING finger associated (SRA) domain (Arita, et al., 2008; Hashimoto, et al., 2008) which recognizes hemimodified CpG dyads. However, the plant protein SUVH5 (PDB 3q0b; Rajakumara, et al., 2011) and the structurally related bacterial McrBC (PDB 3ssc; Sukackaite, et al., 2012) can form homodimers to flip-out both 5mC in a fully methylated CpG dyad simultaneously (reviewed in Patel, 2016).

5-Methylcytosine. In the narrow binding pocket of the SRA domain of UHRF1 (PDB 2zo1), the modified cytosine is recognized via π stacking with the aromatic side chains of a phenylalanine and a tyrosine. The nitrogens N3 and N4 of the pyrimidine ring which were formerly engaged in base pairing, are now contacted by an aspartate side chain. Specificity for the methyl group of 5mC comes from a van der Waals contact with Cβ of a serine.

5-Hydroxymethylcytosine. The SRA of UHRF2 (PDB 4pw5) features a slightly larger binding pocket that accommodates 5hmC with somewhat higher preference than 5mC (Zhou, et al., 2014). The hydroxyl group of 5hmC is recognized by the backbone carbonyl of a threonine-glycine diresidue.

(3.2.6) **β-sheet-dependent DNA-binding domains**

Remarkably, all of these DNA-binding domains fold into a three- or four-stranded β-sheet that is supported by an α-helix despite none of them sharing any primary sequence homology (Connolly, et al., 2000). They all make base-specific contacts at the DNA major groove and contact the phosphodiester backbone of each of the DNA strands at two sites using a common mechanism: One of the unspecific anchors is formed around a glycine on a loop that connects two β-strands, while another basic residue at the opposing end of the β-sheet forms the second anchoring site (Connolly, et al., 2000). For example, Gly-53 and Arg-61 in the homing endonuclease I-*Ppo*I of *Physarum polycephalum* (PDB 1a74; Flick, et al., 1998), Gly-148 and Lys-156 in the ethylene-responsive transcription factor 4 of *A. thaliana* (PDB 1gcc; Allen, et al., 1998), Gly-16 and Arg-24 in the DNA-binding domain of Tn916 integrase (PDB 1tn9; Connolly, et al., 2000). Also, the more distantly related methyl-CpG-binding domain (MBD) makes similar contacts, such as Gly-25 and Lys-46 in MBD1 (see the following).

The MBD is the only known β-sheet-dependent DNA-binding domain to recognize modified cytosines, in particular strand-symmetrically methylated CpG dyads. The molecular basis of which is detailed in the following Section 3.3.

3.3 The methyl-CpG-binding domain

The MBD folds into an α-helix and a four-stranded, twisted, antiparallel β-sheet with protruding strands and has a compact, wedge-shaped three-dimensional appearance (**Figure 3.3a–b**). In contrast to other β-sheet-dependent binders, the β-sheet of the MBD is twisted rather than laying flat in the DNA major groove (Galvão & Thomas, 2005) and the α-helix is no longer oriented parallel at the back of the β-sheet, but slanting towards the DNA double-strand contributing mainly unspecific backbone interactions.

The strands β2 and β3 are longer than β1 and β4, and stretch over the major groove of the DNA where they form the 'thin end' of the MBD. β2 and β3 are connected by the highly mobile loop L1 which becomes well structured upon DNA binding (Ohki, et al., 1999) as per NMR studies of MBD1 (**Supplementary Figure A.1**). At the opposing 'thick end', the hydrophobic face of the β-sheet is tightly packed against the amphipathic α-helix, which is oriented roughly antiparallel to β1 and critical for the structural integrity of the domain. Non-conservative amino acid replacements in this region lead to partial or complete disruption of the MBD structure and loss of DNA binding (Ballestar, et al., 2000; Ohki, et al., 1999; Wakefield, et al., 1999).

A hydrophobic patch and a number of positively charged residues face the DNA (**Figure 3.3c**).

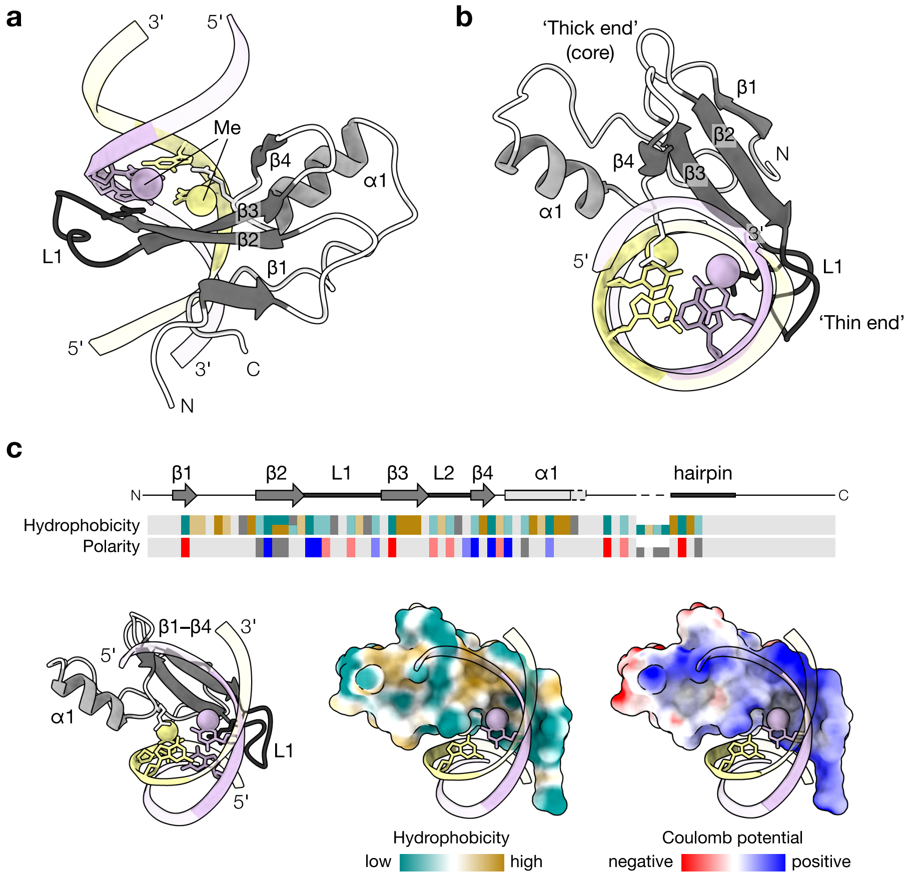


Figure 3.3 Three-dimensional structure of the MBD. Schematic ribbon drawings of methyl-CpG-binding domain (MBD) structures in complex with a 13 bp double-stranded DNA containing a fully methylated CpG. The arginine residues of the MBD interacting with the guanine nucleobases and the CpG nucleotides are shown as sticks; the sp^3 carbon of each 5-methylcytosine (5mC) is shown as van der Waals sphere. **(a)** The MBD of MeCP2 (residues 90–181; PDB 3c2i, Ho, et al., 2008) with the α-helix α1, the β-sheet β1–β4, and loop L1 in side view, and **(b)** in top view. **(c)** Molecular lipophilicity potential (Laguerre, et al., 1997) and Coulomb electrostatic potential (Luty, et al., 1995) projected onto the the linear sequence and the van der Waals surface of the MBD.

3.4 Molecular determinants of MBD–DNA binding

The MBD interacts as a monomer (Nan, et al., 1993; Wade, et al., 1999)[b] with the major groove of double-stranded DNA via residues on β2 and β3, the nearby part of the α-helix, and the loop L1. Although the MBD protects 12 – 14 nt in a DNase I footprinting assay (Klose, et al., 2005), its surface contact with the DNA measures only 550 – 640 Å^2. The aforementioned structural difference to other β-sheet-dependent DNA-binding domains allow for this small interaction surface which is almost exclusively limited to the CpG dyad itself. In contrast, AtERF4 contacts eight, the DNA-binding domain of Tn916 integrase seven, and I-*Ppo*I five base pairs.

Both, specific and unspecific DNA–protein interactions contribute to the MBD's high affinity towards methylated CpG dyads.

Unspecific DNA–MBD interactions. The sequence-independent interactions comprise the electrostatic contacts and the hydrogen bonding of conserved basic and polar residues with the negatively charged phosphate groups of the DNA backbone and several hydrophobic interactions of less conserved side chains with the deoxyribose. The electrostatic interactions with the backbone phosphates contribute with unusually high magnitude to the formation of the MBD–DNA complex and therefore specific binding is observed only at higher salt concentrations or in presence of an unspecific competitor (Khrapunov, et al., 2014).

Specific DNA–MBD interactions. The specificity for CpG dyads in double-stranded DNA has been suggested to involve several mechanisms (**Figure 3.4a**):

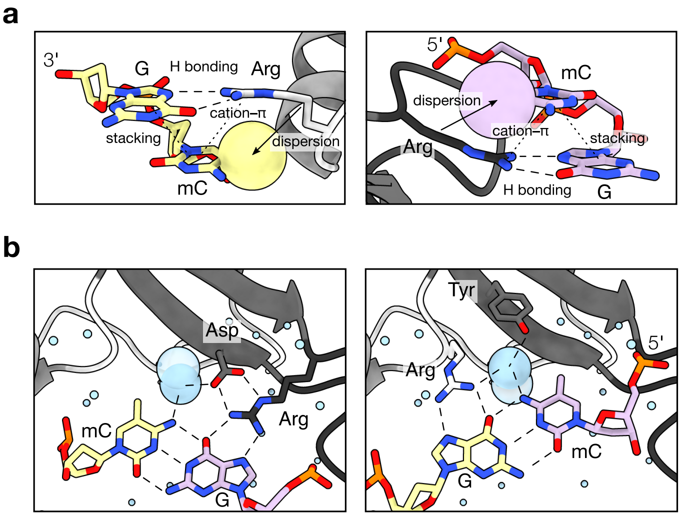


Figure 3.4 CpG dyad-specific MBD–DNA interactions. Binding principles in the prototype MBD of human MeCP2 (PDB 3c2i, Ho, et al., 2008) **(a)** The two cation–π/H-bond stair motifs in the MBD arise from hydrogen bonding (between guanine and arginine), nucleobase stacking (between the pyrimidine and guanine), and the cation–π interaction (between the pyrimidine and arginine). The non-polar interaction between the aliphatic side chain of the arginine and the methylated pyrimidine enhances the cation–π interaction (Zou, et al., 2012). **(b)** The hydrogen bonding network at each of the C · G base pairs in the CpG. Solvent molecules are shown as small spheres; structured water that forms tetrahedral hydrogen bonds in the DNA binding site is shown as van der Waals spheres; Hydrogen atoms were omitted for clarity.

1. Two highly conserved arginines form bidentate hydrogen bonds with each of the guanines in the CpG dyad(first shown by Ho, et al., 2008). However, the two arginines contribute unequally to DNA binding (Free, et al., 2001; Liu, et al., 2019). The more critical arginine is anchored and oriented by a nearby aspartate side chain and makes the initial and persistent contact with a guanine, while the other one is more flexible. The steric confinement of the conformational space in case of methylated DNA finally 'locks' this residue in the DNA binding site (Mezei & Csonka, 2016; Otani, et al., 2013; Sperlazza, et al., 2017).
2. Each arginine engages in a methyl-arginine-guanine triad or another cation–π interaction with a pyrimidine neighboring the bonded guanine. So, they form a sequence of cation–

π/H-bond stair motifs (Rooman, et al., 2002). This interaction is strengthened through an increased van der Waals interface (compare Figure 5.3), namely through cytosine methylation (Zou, et al., 2012).

3. Further, the aliphatic portion of the arginine side chains adds to the non-polar environment in direct vicinity to the methylated pyrimidine (Liu, et al., 2013). Though this dispersion is more important on the side of the more flexible arginine, its overall contribution to binding specificity remains small (Mezei & Csonka, 2016).[c]
4. The hydroxyl group of a conserved tyrosine can hydrogen bond with the structured water surrounding one of the methyl groups, allowing for a 10- to 20-fold increase in binding selectivity per 5mC in a CpG (Cramer, et al., 2014; Hashimoto, et al., 2012; Walavalkar, et al., 2014). These water molecules are coordinated through CH···O hydrogen bonding with the sp^3 methyl carbon in addition to the pyrimidine's conventional hydrogen bond at N4 (**Figure 3.4b**; Ho, et al., 2008; Mayer-Jung, et al., 1998).

How structural differences between the MBDs of the different MBD protein family members affect the recognition of modified CpG dyads is dissected in Chapter 5 and Chapter 6.

3.5 Synopsis

1. Cytosine C5 modifications can affect the shape of the DNA double-helix and constitute a unique physicochemical signal in the DNA major groove.
2. Protein α-helices, β-sheets and loops can probe the DNA major groove. β-sheets are rare in DNA nucleobase recognition and tend to employ more unspecific backbone interactions for high-affinity binding.
3. 5mC is recognized by a variety of proteins using methyl-arginine-guanine triads, non-polar van der Waals or polar interactions via structured water.
4. 5hmC can be recognized via hydrogen bonding with carbonyls, 5caC via arginine or lysine side chains. 5hmC and 5fC may be tolerated by 5mC- and 5caC-binders but usually destabilize the interaction.
5. Molecular recognition of fully methylated 5mC/5mC CpG dyads with distinct interactions at each 5mC exists in different structural classes of DNA-binding proteins, e. g., leucine zippers, zinc fingers and methyl-CpG-binding domains.
6. Some eukaryotic transcription factors require modified cytosine nucleobases for DNA binding, which hints of the regulatory importance of 5mC and its oxidized derivatives.

Endnotes

[a] It has been shown that the increased flexibility in the presence of a methyl group (5mC or T) facilitates DNA–DNA attraction (Yoo, et al., 2016). It is noteworthy that long-range contacts in nuclear organization, frequently involve AT-rich or methylated regions.

[b] Although homo- and heterodimer formation has been reported for full-length MBD2 and MBD3 (Tatematsu, et al., 2000), there is no evidence for the MBD itself undergoing dimerization. Also, the MBD MeCP2 has been shown to bind as a monomer (Nan, et al., 1993; Wade, et al., 1999). This is consistent with a study by Valinluck et al. (2004) as the binding curves show a single inflection point only. However, Khrapunov et al. (2014) report that the MBD itself may form a dimer when the protein is provided in excess. The dimerization is not linked to 5mC recognition.

[c] It has been noted in the early literature that the conserved tyrosine and a conserved phenylalanine form a superficial hydrophobic patch in the center of the DNA binding site of the MBD (Ohki, et al., 1999; Wakefield, et al., 1999). Thus, methylation would increase the buried hydrophobic surface and contribute to the stability of the protein–DNA complex (Zou, et al., 2012). However, the difference in the Gibbs energy of $-6.3\,\mathrm{kJ/mol}$ due to the increased hydrophobic surface is presumably too small to explain the preferred binding of methylated over non-methylated DNA (Mezei & Csonka, 2016).

Chapter 4
Protein engineering and directed evolution platforms

The natural process of evolution continuously yields biological macromolecules such as proteins that are able to adequately fulfill a function for the organism or system they are part of. In directed evolution, the fundamental principles of this process—variation, fitness and survival—are put to work in the laboratory to create proteins with new or improved properties that may or may not be required in nature. This approach has proven a fruitful addition to traditional protein engineering strategies (reviewed in Packer & Liu, 2015). In contrast to rational design, directed evolution can be adopted also when explicit understanding of the principles that govern a specific function is still missing or scarce (**Figure 4.1**).

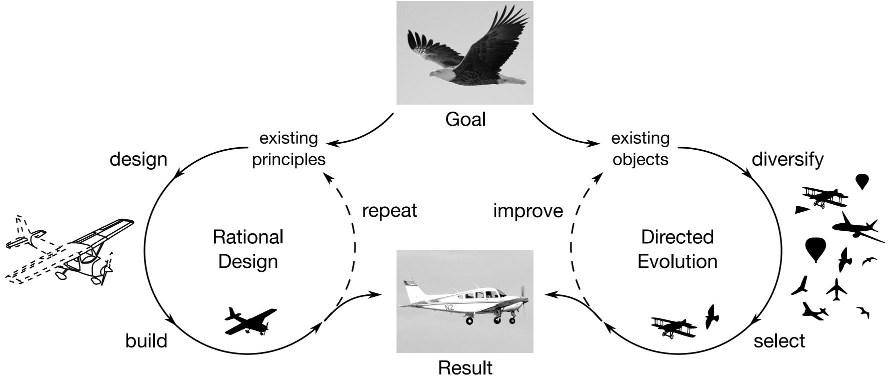


Figure 4.1 Rational design and directed evolution. Two strategies to 'making a flying object' as an analogy for protein engineering: Rational design based on established laws and principles and directed evolution of existing objects that might not (yet) be able to perform the desired task. Photographs in the Public Domain CC0 (full license in Appendix A.4).

Variation in the amino acid sequence and hence the phenotypic properties of the protein parent is created by manipulating the genetic information of the DNA sequence that encodes the protein. This can be achieved in a number of ways: At random positions, e. g., via error-prone PCR (Cadwell & Joyce, 1992; Wilson & Keefe, 2000), and at specific positions, e. g., via site-saturation mutagenesis (Miyazaki & Arnold, 1999; Wells, et al., 1985) or through the exchange, insertion, deletion or permutation of any number of subsequences (some examples in Gillam, 2014). After the protein variants are produced according to the diversified instructions, a specifically designed assay identifies members of this library that perform the desired function (*fitness*). The partition into desired and undesired phenotypes can be done serially (*screening*; manual or in high-throughput) or in a single step in parallel (*selection* by linking separation to survival). In particular during pooled screenings or selections it is crucial to stably propagate the genetic information along with the protein, i. e., to maintain the *genotype–phenotype linkage*, since this allows the successfully recovered variants (*survival*) to serve as the starting point for another cycle of directed evolution.

4.1 *Ex vivo* surface display platforms

Depending on the property to be evolved, the screening or selection assay can be performed *in vitro*, *in vivo* or *ex vivo* (examples in Tizei, et al., 2016). Protein–ligand interactions are typically evolved on *in vitro* (Hanes & Plückthun, 1997; Liu, et al., 2000) or on *ex vivo* display platforms (Boder & Wittrup, 1997; McCafferty, et al., 1990; Smith, 1985). Both platforms bypass the challenge of bringing the ligand into a compartment and allow to define and control the binding conditions.

In *ex vivo* display platforms such as phage display, bacterial, yeast or mammalian cell surface display, the protein of interest, the so-called *passenger*, is linked to a host protein that is naturally exposed to the environment. This allows to streamline the evolution process as the live (or infective) host can readily (be used to) amplify the desired genetic material for subsequent re-analysis, screening or selection. Further, the protein is expressed in the context of cellular factors that might be helpful or even required to assist its proper folding or post-translational modification. Also insoluble or otherwise hard to obtain proteins are often accessible (Kaeßler, et al., 2011). However, platform-specific restraints apply particularly in terms of size and folding of the passenger.

For example, the number of proteins that can be displayed on filamentous phage particle is limited to 5 copies via pIII in Ff bacteriophages (Hay & Lithgow, 2019) and the procedure requires an interim reinfection of a bacterial host. However, due to the lytic release of the phage particles, protein size is typically less problematic. In contrast, 10,000 to 70,000 proteins, sometimes even five- to tenfold more molecules, can be displayed on bacterial cells via autotransporters (Kaeßler, et al., 2011) or via intimin fusions (Salema, et al., 2013; Wentzel, et al., 2001). Similar levels can be displayed on yeast cells via Aga1p:Aga2p (Boder & Wittrup, 1997) despite the tenfold larger cell surface area of *Saccharomyces cerevisiae* as compared to *Escherichia coli*. Yet, these platforms are limited by the size of proteins that can be displayed.

The autotransporters of Gram-negative bacteria naturally expose protein domains with a size of 50 – 100 kDa on the bacterial cell surface (Henderson, et al., 2004). By replacing the passenger of the autotransporter used in this work (**Figure 4.2**), the adhesin involved in diffuse adherence (AIDA-I) of enteropathogenic *Escherichia coli* strain O126:H27 has been harnessed to display heterologous passengers of up to 50 – 70 kDa (Jose, et al., 2009; Schultheiss, et al., 2008) including a functional Klenow fragment of *E. coli* DNA polymerase I (Chung, et al., 2020). AIDA-I autotransport requires the passenger to remain unfolded during translocation to the outer membrane. In particular passage through the oxidizing environment of the periplasmic space is a major impediment to the successful display of various protein payloads prone to cysteine-cysteine disulfide bond formation (de Marco, 2009).

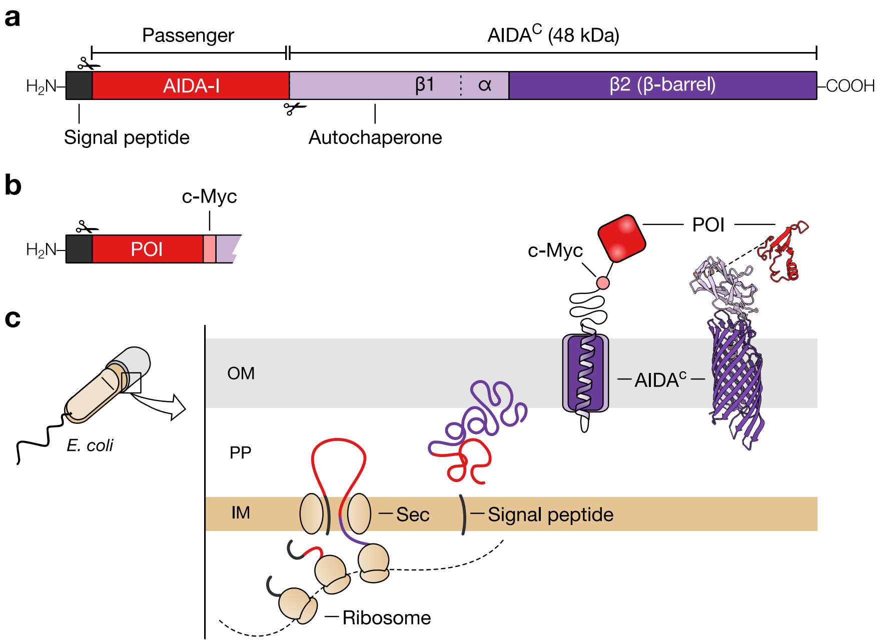


Figure 4.2 Surface-display with the adhesin involved in diffuse adherence (AIDA-I) autotransporter. (a) The wild-type autotransporter of *Escherichia coli* O126:H27 with N-terminal AIDA-I passenger (84 kDa) and C-terminal transporter unit AIDA^C. The autochaperone β1 has autoproteolytic activity via Asp-787 and cleaves AIDA^C off the mature AIDA-I which however remains tightly associated with the cell surface (Charbonneau, et al., 2009). **(b)** The AIDA-I can be replaced with different proteins of interest (POI), in this work the methyl-CpG-binding domain (MBD). **(c)** The autotransporter is cotranslationally secreted through the inner membrane (IM) into the periplasm (PP) where the signal peptide is removed and the autotransporter folds into the outer membrane (OM), bringing the passenger onto the cell surface if it remains unfolded; Structure of MBD of MeCP2 from PDB 3c2i and β2 from PDB 4mee; β1-α was modeled with RaptorX (Xu, et al., 2021).

4.2 Fluorescence-activated high-throughput screening platforms

The use of cells (rather than phages) as display platform permits high-throughput fluorescence-activated cell sorting (FACS) to screen (rather than to select) the individual library members (**Figure 4.3**). Here, the stringency of the screen can be controlled not only, e. g., by different washing conditions, but also by varying the fluorescence threshold that triggers the cell isolation (VanAntwerp & Wittrup, 2000). If multiple fluorescence channels and multiple collection lines are available, different parameters can be assessed for each single clone at one time and instruct the separation into different subpopulations in a single round.

For affinity maturation of protein–ligand interactions, in particular peptide and antibody fragment libraries (Boder & Wittrup, 1997; Daugherty, et al., 1998), fluorescently labeled ligands are used and high-affinity variants retrieved either in a kinetic or thermodynamic equilibrium binding assay (reviewed in Cherf & Cochran, 2015). In the kinetic setup (**Figure 4.4a**), samples are incubated with a single fluorescently labeled probe which is removed from less affine binding sites by extensive washing and/or addition of an excess of dark (unlabeled and cheap)

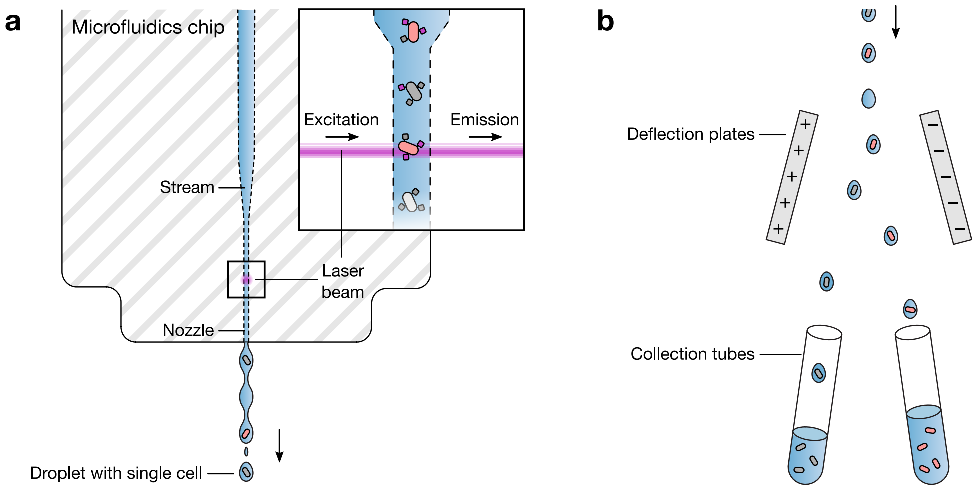


Figure 4.3 Fluorescence-activated cell sorting. (a) In modern fluorescence-activated cell sorting (FACS) instruments, a microfluidics device ('chip', e. g., 70 – 100 µm channel diameter) lines up cells in a sample stream to analyze thousands of individual cells per second using (multi-color) fluorescence intensity optics. At the nozzle, the sample stream is broken down into droplets embedding one cell per single droplet. **(b)** The charged droplets can be deflected such that the sample is separated physically into individual subpopulations based on the recorded fluorescence measurements.

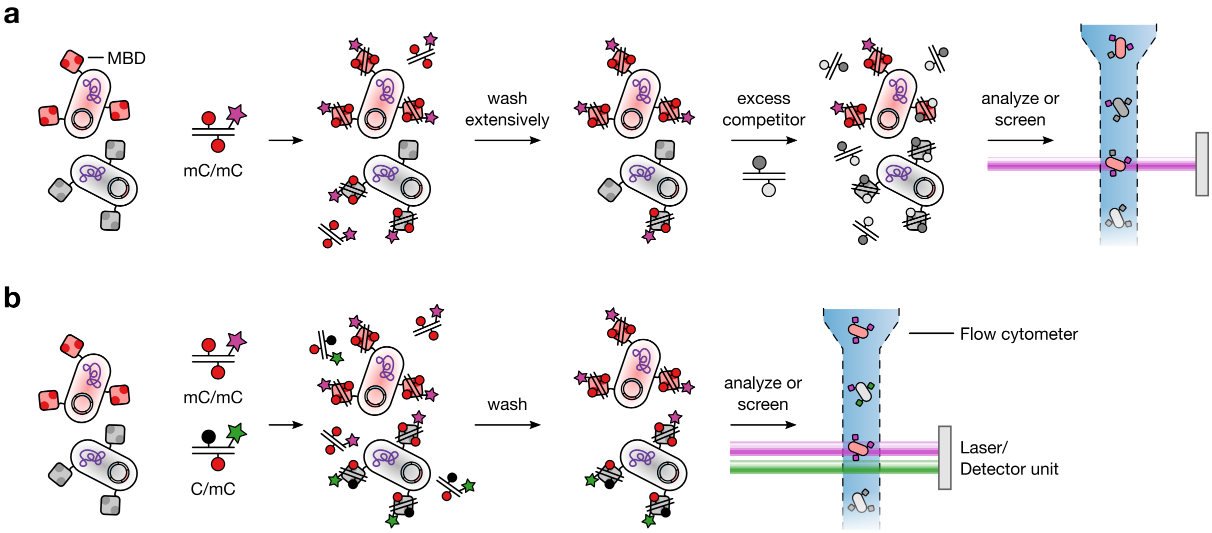


Figure 4.4 Kinetic and thermodynamic DNA binding assays. (a) Kinetic screening of surface-displayed proteins, e. g., an MBD, with a fluorescently-labeled DNA, here showing a single fully methylated CpG dyad; A dark, unlabeled competitor is used to remove unspecific binding. **(b)** Thermodynamic screening with two different labeled probes. Unbound probes are removed shortly before the samples are assayed on a multi-color flow cytometer.

competitor. Differences in the dissociation kinetics can then be monitored by monochromatic flow cytometry. In the thermodynamic setup (**Figure 4.4b**), samples are simultaneously exposed to different DNA probes. Both, association and dissociation kinetics determine the relative amounts of bound probes on the surface-displayed proteins when equilibrium is reached. However, in contrast to the kinetic setup, a five- to tenfold excess of ligand must be provided to saturate all ligand binding sites and to avoid ligand depletion during the binding reaction.

After the desired cell population has been sorted on the FACS instrument, the cells can either be amplified in culture and screened again to purify the highest affine phenotypes or their genotypes can be subjected to another round of diversification by random or targeted mutagenesis which starts another cycle of directed evolution to potentially improve the protein candidates.

4.3 Synopsis

1. Directed evolution is a powerful strategy to create or enhance specific properties of proteins. Out of a large number of protein variants, only the 'fittest' variants are taken forward.
2. In combination, surface display systems and fluorescence-activated cell sorting allow to assess millions of protein variants in high-throughput for probing protein–ligand interactions.

Aim

Problem. Deciphering distinct combinations of modified cytosine nucleobases such as 5mC, 5hmC, 5fC, and 5caC simultaneously in both strands of the DNA double-helix at a single CpG dyad without irreversibly destroying these sites is an unmet challenge to date. Yet, such combinations occur naturally in the genomic DNA of human, mice and other mammalian species as a result of an active DNA modification pathway. They constitute a unique physicochemical interaction interface at the DNA major groove which could be interpreted as distinct regulatory, potentially even epigenetic marks by cellular factors.

Although the recognition of strand-symmetrically modified 5mC/5mC CpG dyads by different nuclear proteins is well described in the literature, the molecular basis for the recognition of strand-asymmetrically modified CpG dyads is unexplored.

Objective. With this work I seek to create proteins that can recognize or ideally discriminate specific combinations of C5-modified cytosines in a single CpG dyad in the DNA double-helix.

Impact. The proteins proposed in this work would be first-of-their-kind DNA-binding proteins demonstrating the possibility of strand-asymmetric molecular recognition of modified cytosine combinations in the DNA double-helix and provide valuable insights into the requirements and limitations that such recognition might face. Beyond this, the proteins can serve as molecular probes to examine combinatorial marks of cytosine modifications at CpG dyads in genomic DNA at the level of the biologically relevant, single DNA duplex.

Original Work

and

Discussion

Chapter 5

Promise and prospect of the methyl-CpG-binding domain

Protein sequence, protein structure and protein function are intimately related. Therefore, proteins that exert a function similar to the function one desires to obtain can serve as a starting point or *parent* for protein engineering (Arnold, 1996).

To obtain a protein-based reagent that would specifically recognize distinct combinations of modified cytosines at CpG dyads in the DNA double-strand, several structural classes of DNA-binding proteins were considered (Chapter 3). Those scaffolds however that required or would likely require to arrange two independent monomers in order to simultaneously recognize both C5 substituents, such as leucine zippers and base-flippers, were not followed up in this work. Of the remaining choices, zinc finger (ZNF) had undoubtedly demonstrated that they could be engineered for the recognition of different, albeit unmodified DNA nucleobase combinations (Rebar & Pabo, 1994). Yet, ZNF binding is sequence context-dependent and their cooperative binding mode requires sophisticated engineering approaches due to the low affinity of a single ZNF (Chou, et al., 2017; Dutta, et al., 2016). So, the MBD which is far less explored in terms of its evolvability, but naturally had the required geometry to interact with two modified cytosines on the DNA double-helix, namely fully methylated 5mC/5mC CpG dyads, was chosen as parent for protein engineering.

Of course, *the* protein fold exists only as an abstract model which generalizes some common properties of structurally and often functionally related three-dimensional folds found within a set of proteins, be they evolutionarily related or not. However, if available, a structural and functional understanding of its particular members can rationally guide the choice both of the members to start with as parents and of candidate positions in the parent sequences to consider for alteration. This can be valuable since mutations that benefit the desired function are generally much more rare than deleterious or neutral ones in highly specialized proteins and the search for fitter variants in large sequence spaces is more resource-intensive.

As regards content, this first chapter contrasts the structural particularities of five MBDs (Section 5.1) and their consequences for the binding of DNA duplexes with modified CpG dyads (Section 5.2). Beyond a literature survey, I systematically add quantitative measures that guide further evaluation. Finally, this comparison will suggest candidate members and candidate sites within each domain that either are of such critical structural importance as they probably must be preserved during the engineering or that function as key for a specific ligand selectivity such that they could be critical for the protein engineer to degenerate in order to obtain the desired new function (Section 5.3).

5.1 Sequence and structural conservation in the MBD

Five mammalian proteins contain a 70 to 80 amino acid MBD which was identified in the full-length sequences based on the hidden Markov model of the MBD family (PF01429, Methods). In the human proteome these are MBD1[2–81], MBD2[146–225], MBD3[2–81], MBD4[76–167] and MeCP2[90–181] (**Figure 5.1**). For the sake of clarity, I refer to the respective domains in the full-length proteins simply by the name of the full-length proteins and shall write of the 'full-length' proteins otherwise.

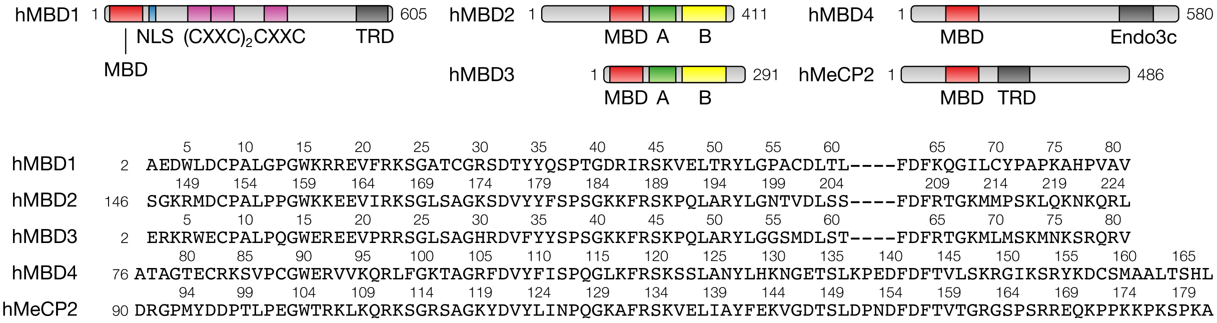

hMBD1 1 605
NLS (CXXC)$_2$CXXC TRD
MBD
hMBD2 1 411
MBD A B
hMBD3 1 291
MBD A B
hMBD4 1 580
MBD Endo3c
hMeCP2 1 486
MBD TRD
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80
hMBD1 2 AEDWLDCPALGPGWKRREVFRKSGATCGRSDTYYQSPTGDRIRSKVELTRYLGPACDLTL----FDFKQGILCYPAPKAHPVAV
149 154 159 164 169 174 179 184 189 194 199 204 209 214 219 224
hMBD2 146 SGKRMDCPALPPGWKKEEVIRKSGLSAGKSDVYYFSPSGKKFRSKPQLARYLGNTVDLSS----FDFRTGKMMPSKLQKNKQRL
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80
hMBD3 2 ERKRWECPALPQGWEREEVPRRSGLSAGHRDVFYYSPSGKKFRSKPQLARYLGGSMDLST----FDFRTGKMLMSKMNKSRQRV
80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165
hMBD4 76 ATAGTECRKSVPCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRYKDCSMAALTSHL
94 99 104 109 114 119 124 129 134 139 144 149 154 159 164 169 174 179
hMeCP2 90 DRGPMYDDPTLPEGWTRKLKQRKSGRSAGKYDVYLINPQGKAFRSKVELIAYFEKVGDTSLDPNDFDFTVTGRGSPSRREQKPPKKPKSPKA

Figure 5.1 MBDs in MBD protein family members. Position within the full-length proteins and enumeration of human and mouse methyl-CpG-binding domains (MBDs).

The superimposition of the available crystal structures choosing one of the methylated 5mC · G base pairs of the CpG dyad as arbitrary reference point (**Figure 5.2a**) showed noticeable structural variation among the domains. Whereas the positional deviations at the 'thin end' of the MBD (with loop L1 near the reference point) was expectedly small (**Table 5.1**), the positioning of helix α1 at the 'thick end' varied more drastically than the placement of the second 5mC · G base pair in the dyad did. This was most pronounced for the MBD of MBD4 and MeCP2 and could be understood either as different solutions in a continuous structural space towards the fulfillment of an (almost) identical function, or as a consequence of specific evolutionary adaptations towards more specialized roles for each of the MBDs.

Table 5.1 Root-mean-square distances between superimposed MBD structures. As shown in Figure 5.2a; RMSDs with respect to the MBD1 protein–DNA complex and backbone atoms if not otherwise declared.

MBD	Registry	Res.*	'Upper' C · G	Arg-44	'Lower' C · G	Arg-22	Protein	Loop L1	Helix α1
MBD1	PDB 6d1t	2.25 Å	0.00 Å	0.00 Å	0.00 Å	0.00 Å	0.00 Å	0.00 Å	0.00 Å
MBD2	PDB 6cnq	2.15 Å	0.89 Å	1.67 Å	1.42 Å	1.14 Å	3.32 Å	2.32 Å	3.76 Å
MBD3	PDB 6ccg	1.90 Å	0.90 Å	1.46 Å	1.39 Å	1.64 Å	2.66 Å	2.27 Å	3.40 Å
MBD4	PDB 4lg7	2.50 Å	0.47 Å	1.88 Å	1.02 Å	0.40 Å	3.96 Å	0.61 Å	4.45 Å
MeCP2	PDB 3c2i	2.50 Å	0.88 Å	2.02 Å	2.03 Å	1.90 Å	4.99 Å	1.88 Å	5.20 Å

* For differences in resolution larger than 0.15 Å, the RMSD differences were corrected according to Carugo (2003).

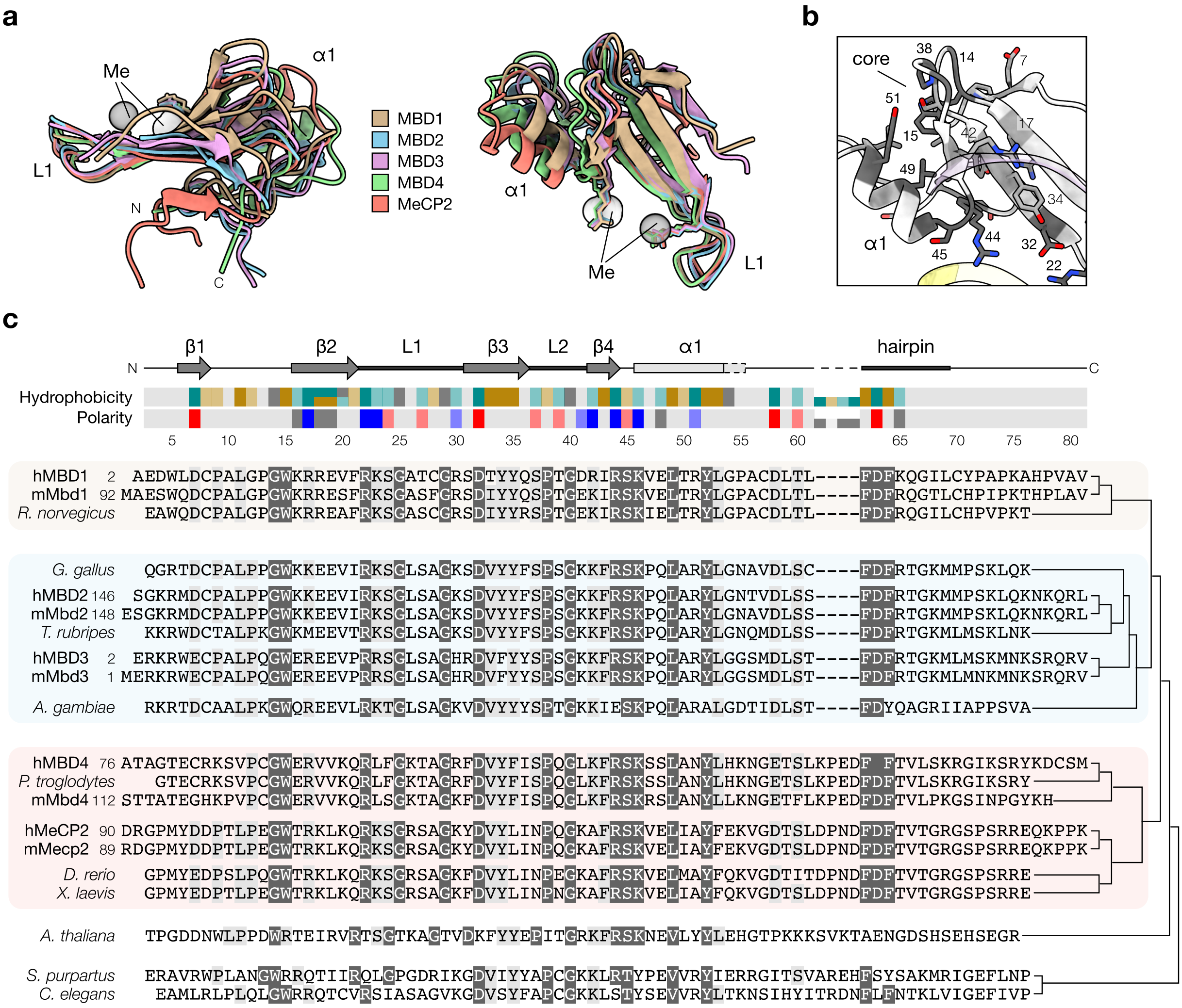

a
Me
α1
L1
N
C
MBD1
MBD2
MBD3
MBD4
MeCP2
α1
Me
L1
b
core
38
14
7
51
17
15
42
49
34
α1
44
32
45
22
c
β1
β2
L1
β3
L2
β4
α1
hairpin
N
C
Hydrophobicity
Polarity
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
hMBD1 2 AEDWLDCPALGPGWKRREVFRKSGATCGRSDTYYQSPTGDRIRSKVELTRYLGPACDLTL----FDFKQGILCYPAPKAHPVAV
mMbd1 92 MAESWQDCPALGPGWKRRESFRKSGASFGRSDIYYQSPTGEKIRSKVELTRYLGPACDLTL----FDFRQGTLCHPIPKTHPLAV
R. norvegicus EAWQDCPALGPGWKRREAFRKSGASCGRSDIYYRSPTGEKIRSKIELTRYLGPACDLTL----FDFRQGILCHPVPKT
G. gallus QGRTDCPALPPGWKKEEVIRKSGLSAGKSDVYYFSPSGKKFRSKPQLARYLGNAVDLSC----FDFRTGKMMPSKLQK
hMBD2 146 SGKRMDCPALPPGWKKEEVIRKSGLSAGKSDVYYFSPSGKKFRSKPQLARYLGNTVDLSS----FDFRTGKMMPSKLQKNKQRL
mMbd2 148 ESGKRMDCPALPPGWKKEEVIRKSGLSAGKSDVYYFSPSGKKFRSKPQLARYLGNAVDLSS----FDFRTGKMMPSKLQKNKQRL
T. rubripes KKRWDCTALPKGWKMEEVTRKSGLSAGKSDVYYFSPSGKKFRSKPQLARYLGNQMDLSS----FDFRTGKMLMSKLNK
hMBD3 2 ERKRWECPALPQGWEREEVPRRSGLSAGHRDVFYYSPSGKKFRSKPQLARYLGGSMDLST----FDFRTGKMLMSKMNKSRQRV
mMbd3 1 MERKRWECPALPQGWEREEVPRRSGLSAGHRDVFYYSPSGKKFRSKPQLARYLGGSMDLST----FDFRTGKMLMNKMNKSRQRV
A. gambiae RKRTDCAALPKGWQREEVLRKTGLSAGKVDVYYYSPTGKKIESKPQLARALGDTIDLST----FDYQAGRIIAPPSVA
hMBD4 76 ATAGTECRKSVPCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDF FTVLSKRGIKSRYKDCSM
P. troglodytes GTECRKSVPCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRY
mMbd4 112 STTATEGHKPVPCGWERVVKQRLSGKTAGKFDVYFISPQGLKFRSKRSLANYLLKNGETFLKPEDFDFTVLPKGSINPGYKH
hMeCP2 90 DRGPMYDDPTLPEGWTRKLKQRKSGRSAGKYDVYLINPQGKAFRSKVELIAYFEKVGDTSLDPNDFDFTVTGRGSPSRREQKPPK
mMecp2 89 RDGPMYDDPTLPEGWTRKLKQRKSGRSAGKYDVYLINPQGKAFRSKVELIAYFEKVGDTSLDPNDFDFTVTGRGSPSRREQKPPK
D. rerio GPMYEDPSLPQGWTRKLKQRKSGRSAGKFDVYLINPEGKAFRSKVELMAYFQKVGDTITDPNDFDFTVTGRGSPSRRE
X. laevis GPMYEDPTLPEGWTRKLKQRKSGRSAGKFDVYLINPNGKAFRSKVELIAYFQKVGDTSLDPNDFDFTVTGRGSPSRRE
A. thaliana TPGDDNWLPPDWRTEIRVRTSGTKAGTVDKFYYEPITGRKFRSKNEVLYYLEHGTPKKKSVKTAENGDSHSEHSEGR
S. purpartus ERAVRWPLANGWRRQTIIRQLGPGDRIKGDVIYYAPCGKKLRTYPEVVRYIERRGITSVAREHFSYSAKMRIGEFLNP
C. elegans EAMLRLPLQLGWRRQTCVRSIASAGVKGDVSYFAPCGKKLSTYSEVVRYLTKNSIHYITRDNFLFNTKLVIGEFIVP

Figure 5.2 Conservation of sequence and structure in the MBD. (a) Best-fit superimposition of the five human MBDs of MBD1 (residues 2–81; PDB 6d1t, SGC[a]), MBD2 (residues 146–225; PDB 6cnq, Liu, et al., 2018), MBD3 (residues 2–81; PDB 6ccg, Liu, et al., 2019), MBD4 (residues 76–167; PDB 4lg7, SGC[a]), and MeCP2 (residues 90–181; PDB 3c2i, Ho, et al., 2008). The nucleobases of one C · G base pair were superimposed; only the sp^3 carbon of each 5mC (Me) is shown for clarity. **(b)** Spacial organization of the conserved residues in the MBD; the enumeration follows MBD1 but MeCP2 is shown. **(c)** Amino acid sequence alignment and cladogram (Clustal Omega) of the five human MBDs, their murine homologs and phylogenetically more distantly related examples (UniProt identifiers Q66HB8, H2QNC3, Q2T2T7, H2SUB2, Q9YGC6, Q5EFL0, W4YH51, F5HM38, Q23590, and AtMBD5) of the MBD protein family; full chart in Supplementary Figure A.2. Identical residues in human and mouse shown in dark gray, residues with similar physicochemical properties in light gray.

On the primary sequence level, mouse and human MBDs share 40 – 55% sequence identity with 16 identical positions and mainly highly conservative substitutions between β2 to α1. Most of the conserved residues are part of the hydrophobic core at the 'thick end' (**Figure 5.2b**). Protein domains with highly similar sequence also exist in other vertebrates, invertebrates and in plants (**Figure 5.2c**, **Supplementary Figure A.2**). An evolutionary relationship has been proposed for invertebrate and vertebrate MBDs (Hendrich & Tweedie, 2003), suggesting that this MBD fold is at least 570 million years old. Plant MBDs in contrast must have followed a separate evolutionary trajectory, either emerging independently or succumbing a distinct selective pressure that accelerated its diversification as compared to the concurrently evolving DNA methyltransferases (Springer & Kaeppler, 2005). Whether all of these domains bind methylated CpGs or serve other purposes is not known.

Like in the structural superimposition, MBD1, MBD2 and MBD3 were also more similar to each other on the primary sequence level than MBD4 and MeCP2. The latter group has an insertion of four amino acids between α1 and a hairpin loop at the C-terminus as well as a cluster of more hydrophobic residues on the 'thin end' of β2, where MBD1, MBD2 and MBD3 have a glutamate residue instead (Wakefield, et al., 1999). Both alterations are thought to increase the hydrophobicity of the core and appear to stabilize the isolated MBD (Ho, et al., 2008; Otani, et al., 2013; Walavalkar, et al., 2014). Indeed, most of the highly conserved residues contribute to the core (**Figure 5.2c**) whereas only four (Arg-22, Asp-32, Ser-45 and Arg-44) participate in DNA binding. Judging from their importance for the structural integrity, substitutions at one or more positions in the hydrophobic core could be problematic, whereas replacing even conserved residues in the DNA binding site might be required to reengineer the domain.

5.2 Specific differences in mammalian MBD binding

In general, the specificity of the MBD for the recognition of CpG dyads critically depends on the two highly conserved arginine residues (Chapter 3.3). However, the observed structural variations link different MBDs to a particular ligand spectrum. To get a closer view onto the individual DNA binding sites, **Figure 5.3** presents the surface-contacts of the MBD with each of the DNA strands and **Figure 5.4** the organization of the individual DNA binding sites.

MBD1. The 'thick' end of this MBD is rotated slightly towards the binding site with β1 and β2 approaching the DNA more closely. This allows for Arg-18 and the amide of Ala-26 in L1 to bond with the phosphodiester backbone of one DNA strand and a tighter packing of the Val-47 in α1 against the deoxyribose of the other strand (Scarsdale, et al., 2011). In consequence, Val-20 is more prone to contact one CpG dinucleotide than in other MBDs (**Figure 5.3**). Further, the hydroxyl group of Tyr-34 is able to directly form a hydrogen bond with the N4 of the methylated cytosine (**Figure 5.4b–c**, Ohki, et al., 2001) whereas molecular dynamics simulations favor the bonding with N4 of an adjacent pyrimidine (Rauch, et al., 2005).[b] A state in which Asp-32 directly bridges the two methylated cytosines via their N4 amino groups has been observed in molecular dynamics simulations (**Figure 5.4d**). In this state, the hydroxyl group of Tyr-34 also engages in a hydrogen bond between the carboxyl group and Asp-32. The contribution of Tyr-34 to a local hydrophobic environment as suggested by Ohki et al. (2001) is questioned in

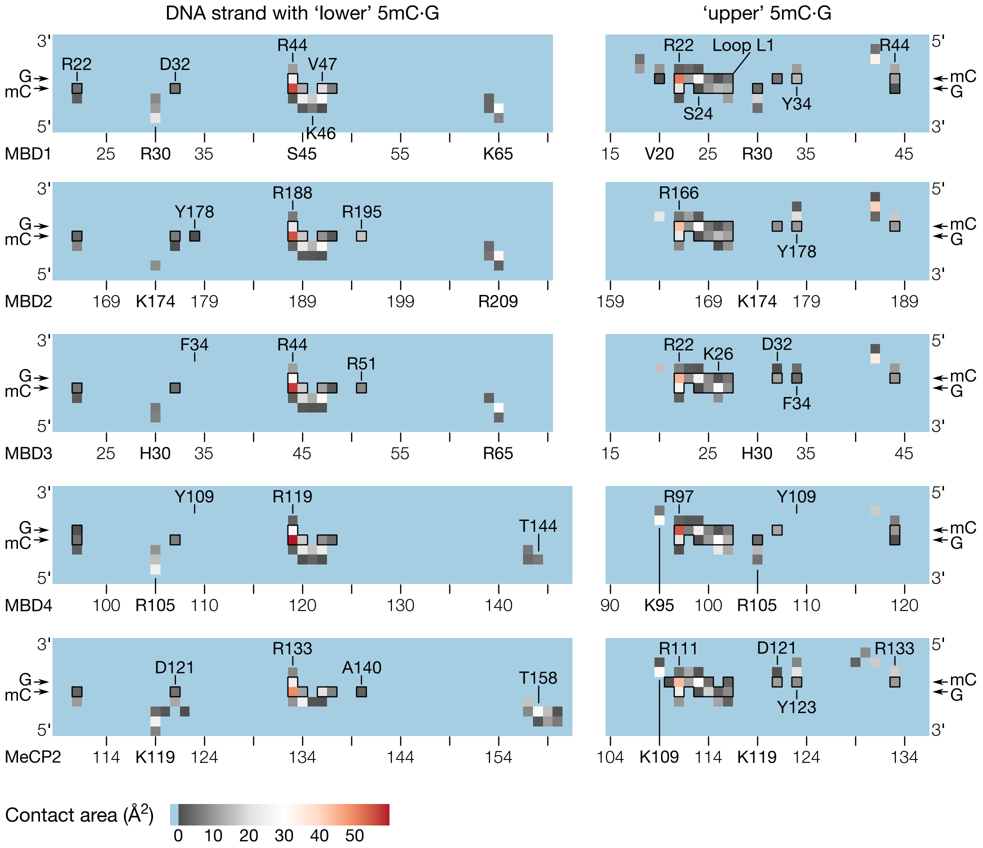


Figure 5.3 Protein–DNA contact surface of MBD proteins. Interatomic contact surface area by amino acid residue resolved for the 'lower' 5mC · G base pair (*left*, yellow in other figures) and the strand with the 'upper' 5mC · G base pair (*right*, purple in other figures). The contacts involving the (modified) nucleotides at the CpG are framed; residues with a particularly high (or differential) contact surface are indicated. Calculated according to Ribeiro et al. (2019) based on the refined crystal structures (Joosten, et al., 2011) of MBD1 (residues 2–81; PDB 6d1t, SGC[a]), MBD2 (residues 146–225; PDB 6cnq, Liu, et al., 2018) MBD3 (residues 2–81; PDB 6ccg, Liu, et al., 2019) MBD4 (residues 76–167; PDB 4lg7, SGC[a]), and MeCP2 (residues 90–181; PDB 3c2i, Ho, et al., 2008). Source code available.

favor of the aliphatic groups in Asp-32 (Rauch, et al., 2005), though Tyr-34 makes the highest surface contact of all human MBDs. There is no evidence for a stabilizing salt bridge between Glu-48 and Arg-44 in MBD1 as observed in MeCP2, neither in the solution structure of the free protein (PDB 1d9n, Ohki, et al., 1999) nor in the solution structure (PDB 1ig4, Ohki, et al., 2001) or in the crystal structure of the protein–DNA complex (PDB 6d1t, SGC[a]).

MBD2. Although the majority of structural differences in this MBD involve residues outside DNA the binding site, the most notable is the interaction of Arg-209 (Lys-65 in MBD1) with the phosphate backbone of the DNA (Scarsdale, et al., 2011) that takes place via the (secondary) ε-nitrogen of the guanidinium group while the η-nitrogens are engaged in hydrogen bonding with Ser-175 (**Figure 5.4f**) similar to Lys-65 and Ser-31 in MBD1. The analogous residue in MeCP2, Thr-158, plays a role in stabilizing two consecutive turns of the protein backbone.

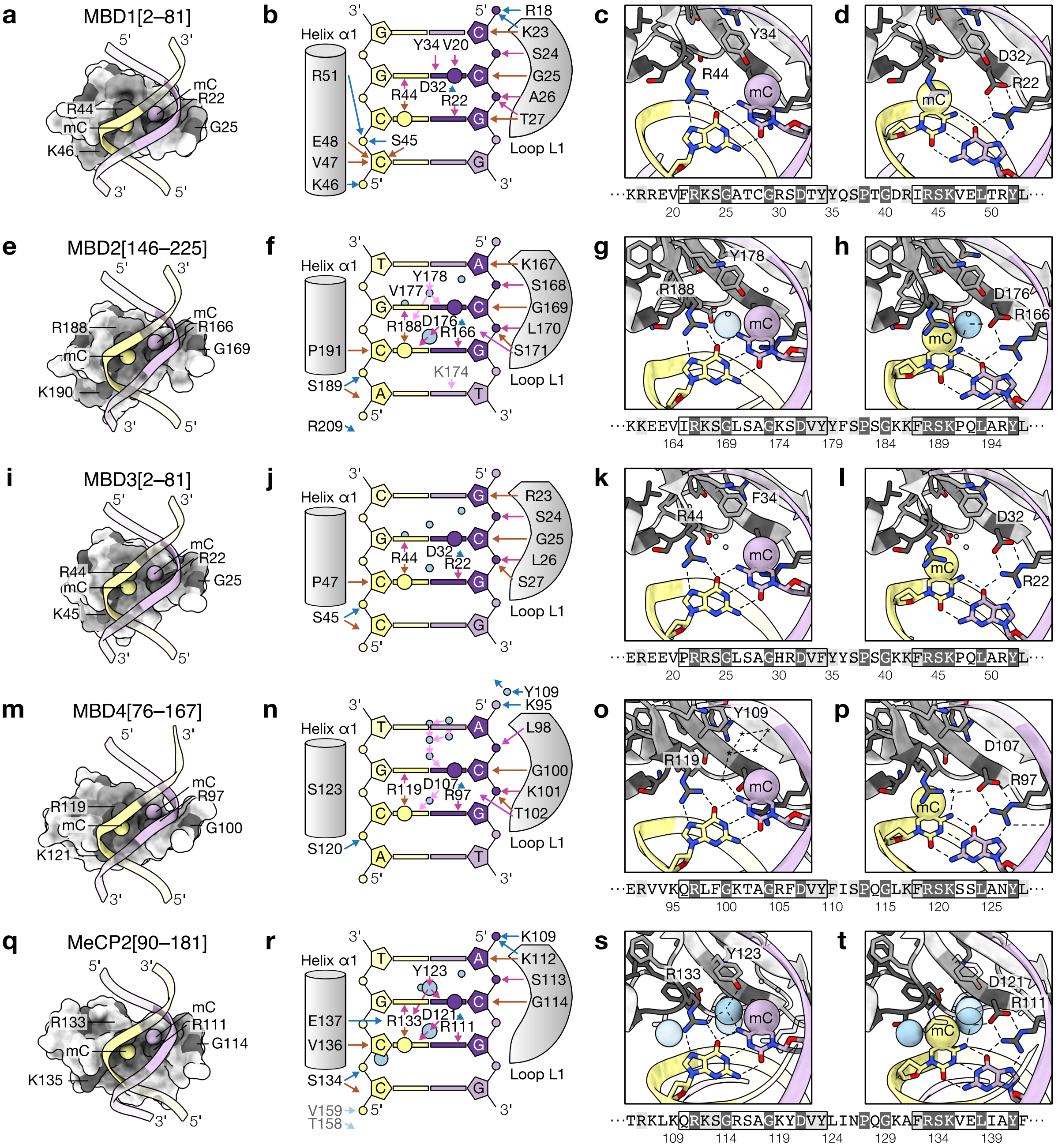


Figure 5.4 Protein–DNA interactions in five human MBDs. Details in the main text. The methyl-CpG-binding domains of *a–d* MBD1[2–81] (PDB 6d1t, SGC[a])[b], *e–h* MBD2[146–225] (PDB 6cnq, Liu, et al., 2018), *i–l* MBD3[2–81] (PDB 6ccg, Liu, et al., 2019), *m–p* MBD4[76–167] (PDB 4lg7, SGC[a]), and *o–t* MeCP2[90–181] (PDB 3c2i, Ho, et al., 2008) were superimposed on the 'upper' 5mC · G base pair (5-methylcytosine, 5mC, DNA strand in purple) and are shown from the same angle. **(a, e, i, m, q)** Van der Waals surface colored by residue conservation as in Figure 5.2. The sp^3 carbon of the 5-methylcytosine (5mC) is shown as van der Waals sphere. **(b, f, j, n, r)** London dispersion (brown), hydrogen bonding (pink), and electrostatic interactions (blue) between the MBD and the DNA double-strand as reported and extended in the literature. **(c, g, k, o, s)** 'Upper' 5mC · G base pair, the flexible arginine, interacting residues and structured water if annotated in the crystal structure in blue, spheres if closer than 4.4 Å to the sp^3 carbon of 5mC. **(d, h, l, p, t)** 'Lower' 5mC · G base pair and anchored arginine.

Low-affinity binding of MBD2 to methylated and unmethylated 5mCpApC/GpTpG trinucleotides is due to a more flexible Arg-188 that interacts with the GpT preceding the guanine bonded by Arg-166. Other methylated 5mCpApD/GpTpH trinucleotides are weakly bound similar to 5mCpG/5mCpG dyads but with thymine taking the role of 5-methylcytosine (Liu, et al., 2018).

MBD3. The human and murine isoforms of this MBD bind to methylated DNA very weakly and with less specificity, reflecting a rapid exchange between 5mC-specific and unspecific CpG binding modes (Cramer, et al., 2014). The main substitutions in the DNA binding site are histidine-30 (instead of lysine) and phenylalanine-34 (instead of tyrosine). However, the interactions of Arg-22 and Arg-44 with the CpG during methylation-specific binding are still observed (**Figure 5.4j–l**) and residues of L1 (Gly-25 and Ala-28) show chemical-shift differences consistent with DNA binding in NMR structures, albeit less well structured since they are averaged over an ensemble of conformations (Cramer, et al., 2014).

Phenylalanine-34 can not form hydrogen bonds and therefore prevents the MBD from engaging in additional solvent-mediated DNA contacts. The reconstitution MBD3[F34Y] is sufficient to recover methylation selectivity (Fraga, et al., 2003; Liu, et al., 2019; Saito & Ishikawa, 2002) but with lower affinity as compared to the MBD3[H30K,F34Y] mutant (Cramer, et al., 2014; Fraga, et al., 2003; Saito & Ishikawa, 2002). In contrast, the MBD3[H30K] mutant engages in unspecific DNA binding (Saito & Ishikawa, 2002). This demonstrates that selective recognition of 5mC/5mC depends necessarily, but not sufficiently on the conserved tyrosine.

MBD4. Full-length MBD4 participates in DNA repair of mismatched thymine bases in CpG dyads (Du, et al., 2015). Its MBD has a flipped Tyr-109 side chain that points away from the DNA major groove (**Figure 5.4o**) and contributes to the hydrophobic core. The hydroxyl group engages in solvent-mediated hydrogen bonding with the DNA backbone which allows for an extensive water network to finely tune the binding of CpG dyads with 5mC/5mC or 5mC/5hmC modification as well as with 5mCpA/TpG and mismatched 5mCpG/TpG dinucleotides (Otani, et al., 2013). With exception of 5mC/5hmC, each of these sequences has two symmetrically positioned methyl groups that can engage in binding.

The interactions around Arg-97 which is anchored by Asp-107 are similar to the ones observed in MBD1, MBD2 and MeCP2 (**Figure 5.4p**). Interestingly, the unique flexibility of Arg-119 created by the absence of a negatively charged side chain with Ser-123 (the analogous positions being Glu-137 in MeCP2) as well as the additional space left by the flipped Tyr-96 facilitates unspecific DNA interaction in 'sliding mode' prior to target recognition (Otani, et al., 2013; Walavalkar, et al., 2014). The contact surface at both 5mC-arginine contacts is almost equal in this MBD and highest of all MBDs (**Figure 5.3**).

In the MBD4 complex with a 5mC/5hmC dyad (PDB 3vyb), the hydroxyl group is found in its preferred synclinal orientation and participates in water-mediated hydrogen bonding with Asp-94 and Arg-84. Whereas 5mC/5fC is tolerated, yet with lower affinity than 5mC/5mC, 5mC/5caC dyads are not effectively bound, probably because of the vicinity of the negatively charged 5caC to Asp-94 (Otani, et al., 2013).

MeCP2. Structurally, the α-helix of this MBD has an additional turn and in consequence a more oblique angle onto the DNA, presumably contributing to the stability of the domain (Scarsdale, et al., 2011). Both arginines make almost equal contact with the respective 5mC. However, especially the contact at Arg-133, the more flexible arginine, is lowest of all MBDs, whereas the contribution of Val-136 is highest (**Figure 5.3**). In the DNA-bound complex, Glu-137 is involved in a unique salt bridge with the more flexible Arg-133 (**Figure 5.4r**).

The binding of 5mCpApC/GpTpG and 5hmCpApC/GpTpG trinucleotides has been reported by Lagger et al. (2017) albeit with lower affinity. Also unmethylated CpApC/GpTpG are recognized very weakly (Lei, et al., 2019) but the interaction irrelevant to the localization of the domain or the full-length protein *in vivo* (Connelly, et al., 2020). While the hydrogen bonding around the arginine anchored to Asp-121 (Arg-111, **Figure 5.4s**) remains largely unaffected in these scenarios, the more flexible Arg-133 forms hydrogen bonds with the GpT preceding the guanine that is bonded by Arg-111.

MBD5 and MBD6. The domains of these human MBD protein family members essentially lack L1 and do not bind methylated DNA (Laget, et al., 2010).

(5.2.1) Context-dependence of MBD binding

The CpG dyad is sequence-symmetric (palindromic) in double-stranded DNA which allows the MBD to bind theoretically in two possible orientations. However, the orientation and affinity can be affected by the sequence context in which the CpG dinucleotide appears.

MBD1. A TXGCA or TGXGCA context (X = 5mC) favors high-affinity binding as determined by SELEX, but other contexts are only four- to fivefold lower (Clouaire, et al., 2010).

MBD2. It has been noted that Lys-174 (Lys-32 in MBD2, PDB 2ky8) makes a base-specific contact to a guanine following a CpG contributing to the preferred orientation on a 5mCpGpG trinucleotide (Scarsdale, et al., 2011). This is in accordance with targets enriched by SELEX sharing a XGG consensus (Klose, et al., 2005).

MBD3. No binding preference has been observed (Liu, et al., 2019).

MBD4. It has been noted by Walavalkar et al. (2014) that Arg-105 in MBD4 makes a base-specific contact to a guanine preceding the 5mCpG on the opposite strand and one base apart, i. e., CpNp5mCpG, via the DNA major groove. However, MBD4 has not been reported to show any context-dependence of DNA binding (Buck-Koehntop & Defossez, 2014).

MeCP2. MeCP2 selects for targets with an A/T stretch adjacent to the 5mC site. This is independent of an AT hook C-terminal of the MBD in the full-length protein (Klose, et al., 2005). Indeed, the C-terminal region of its MBD itself comprises an unusual tandem 'Asx-ST motif' which consists of an Asx turn (^{156}DFT^{158}) followed by an ST motif (^{158}TVTG^{161}) that is stabilized through hydrogen bonding at Thr-158. If the DNA contains a $(dA,dT)_{4\text{–}6}$ run which causes narrowing of the minor groove and bending of the B-DNA, the Asx-ST motif may stabilize the DNA–MBD contact. Two mutants implicated in Rett syndrome, which break the hydrogen bonding in the Asx-ST motif, MeCP2[T158M] and MeCP2[T158A], reduce affinity to methylated DNA twofold while retaining selectivity (Ballestar, et al., 2000; Free, et al., 2001). When the bonding is preserved by a conservative substitution MeCP2[T158S], the affinity for methylated DNA is retained (Ho, et al., 2008).

5.3 Candidates for protein engineering

Comparison of the primary sequence and the three-dimensional structure of the five human MBDs suggested that despite all similarity, these domains interact with modified cytosines in CpG dyads in slightly different ways as supported by biochemical characterizations in the literature. All five scaffolds were therefore taken forward, including the non-binder MBD3, which might offer the possibility to explore a distinct evolutionary trajectory to engineer the domain.

Of minor concern at this initial stage was the possible sequence context-dependence for the engagement with a CpG dyad since all five MBDs show some, often weak preferences. Among them, MBD4 would probably be the preferred choice, irrespective of its comparably promiscuous interaction with other (un-)modified dinucleotide combinations similar to 5mC/5mC dyads.

With respect to the choice of candidate residues that contribute to the recognition of the C5 methyl groups, but not to the selectivity for the CpG dyad or the overall structural integrity of the domain, it seemed reasonable not to degenerate conserved residues of the hydrophobic core or the two arginines that bidentate the guanine nucleobases in the CpG. At least not in the first trials. The tyrosine which contacts directly or indirectly one of the C5 methyl substituents,

i. e., Tyr-52 in MBD1 (Tyr-123 in MeCP2), was an apparent candidate to degenerate. Beyond this, one finds among the less conserved residues in immediate vicinity to the DNA binding site some residues C-terminal at β2, e. g., Val-20 (Lys-109), some residues in loop L1, e. g., Thr-27 and Arg-30 (Lys-119, Lys-119). The highly conserved Asp-32 (Asp-121) which is critical for the pre-organization of one of the arginines, and Lys-46 (Ser-134) at the N-terminal end of α1.

Beyond these residues in close proximity of the CpG dyad, remote, in particular 'second-contact shell' mutations (Wilding, et al., 2019) could be taken into consideration.

5.4 Synopsis

1. Although many amino acid positions are highly conserved between five mammalian wild-type MBDs, their three-dimensional arrangement around the CpG dyad varies noticeably with specific consequences for the molecular recognition of methylated cytosines and structurally similar nucleobases.
2. Based on rationalizations of sequence conservation and binding mode, candidate residues could be proposed as starting points for protein engineering.

Endnotes

[a] Unpublished structure released by The Structural Genomics Consortium (Jones, et al., 2014).

[b] In the original solution structure published by Ohki et al. (2001), water molecules were not considered. The crystallographic structure deposited under PDB 6d1t has no explicitly annotated water molecules either, although it contains 'unknown' atoms that may be structural water similar to the ones observed MBD2 or MeCP2. Unclassified atoms are removed by default during model optimization with PDB_REDO (Joosten, et al., 2011) and therefore not shown in Figure 5.4.

Chapter 6
Decoding of modified CpG dyads by natural MBDs

In this chapter the recognition of symmetrically and asymmetrically modified cytosines in CpG dyads by the five human wildtype methyl-CpG-binding domains (MBDs) is investigated on a purely biochemical basis using electrophoretic mobility shift assays (Section 6.1) and further extended to some recurrent mutations in the MBD of MeCP2 which have been associated with the Rett syndrome, a neurodevelopmental disorder in children (Section 6.2).

The results of this chapter were published in

> Buchmuller, B. C., Kosel, B., & Summerer, D. (2020). Complete profiling of methyl-CpG-binding domains for combinations of cytosine modifications at CpG dinucleotides reveals differential read-out in normal and Rett-associated states. *Sci. Rep.*, *10*(1), 4053–9.

I thank Brinja Kosel for helping me carrying out the EMSAs for this study.

6.1 Interaction of wildtype MBDs with modified CpG dyads

MBDs are the central readers of methylated CpG dyads in the mammalian genome. Their interaction with modified cytosines in CpG dyads has therefore been characterized partially by different researchers using different biochemical assays, but also different MBD proteins, unspecific binding traps and, critically, different DNA duplexes (**Supplementary Table A.2**). This has produced considerable data (**Figure 6.1**, **Supplementary Tables A.3–A.7**), but hampered comparisons between different members of the family.

So, in order to understand to which degree differences in the sequence and structure of each MBD affect their interaction with modified CpG dyads, a systematic, well defined and ideally utmost reductionistic approach was needed. To this end, the 'core' MBD of the five human MBD protein family members MBD1, MBD2, MBD3, MBD4 and MeCP2 was identified by homology in the sequence alignments (see Figure 5.1) and the domains expressed as fusion proteins with a solubilizing protein tag that could be removed *in situ* prior to the experiment. The DNA probes to characterize the binding interactions were 24-mer oligodeoxynucleotide duplexes that contained a single central CpG in an oligo(A) · oligo(T) context with the cognate duplex that lacked the CpG serving as well-defined unspecific binding traps. The fraction of the bound DNA probe was determined for the ten strand-asymmetric and five strand-symmetric combinations of cytosine modifications at CpG dyads.

The binding selectivity of each MBD for the differentially modified CpG dyads was evaluated on electrophoretic mobility shift assays (EMSAs) at two protein concentrations (**Figure 6.2a**).

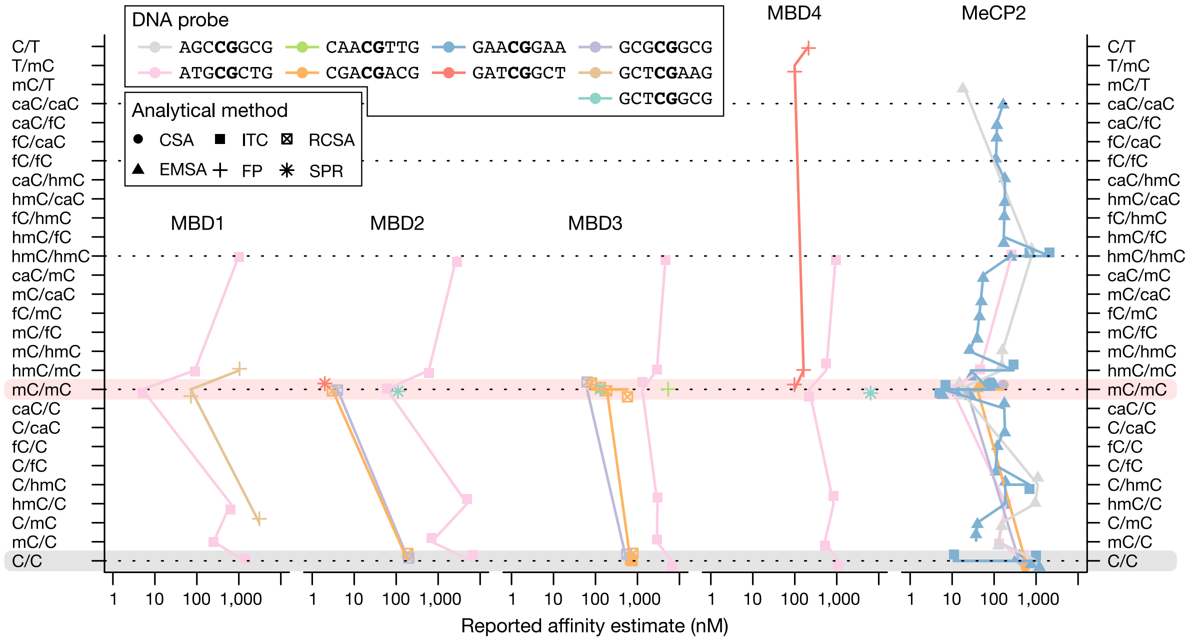


Figure 6.1 Binding affinity of wildtype MBDs reported in the literature. The DNA–MBD binding among the members of the MBD protein family has only been characterized in part using different DNA probes and analytical methods; (Reverse-phase) capillary shift assay (RCSA; CSA), electrophoretic mobility shift assay (EMSA), isothermal calorimetry (ITC); fluorescence polarization (FP); surface plasmon resonance (SPR). References in Supplementary Tables A.3– A.7.

MBD1. In agreement with previous studies (Hashimoto, et al., 2012; Otani, et al., 2013), the MBD of MBD1 (**Figure 6.2b**) strongly bound 5mC/5mC and 5mC/5hmC even at low protein concentrations. C/5mC was bound with markedly reduced affinity and also 5mC/5fC, but not 5mC/5caC, were recognized. Strikingly, none of the other combinations was bound, indicating a strict dependence of MBD1 on the presence of at least one 5mC. Any combination of an oxidized 5mC at both positions or in combination with C or 5caC abolished CpG binding.

MBD2. In contrast to MBD1, the MBD of MBD2 interacted with multiple combinations containing one or even two modified cytosines (including 5caC; **Figure 6.2c**). Again, if one of these modifications was 5mC, the binding was stronger (with the exception of 5mC/5caC) and strongest for 5mC/5mC. These observations agree with a study on murine Mbd2 in complex with the transcriptional repressor p66α which reported that affinity decreased along 5mC/5mC > C/5mC ≈ C/5hmC > 5hmC/5hmC (Hashimoto, et al., 2012).

MBD3. The MBD of MBD3 shares 70% amino acid sequence similarity with MBD2, but contains the mutations K30H and Y34F that reduce the binding to methylated CpGs. The murine ortholog has been shown to interact with 5mC/5mC and other combinations involving C, 5mC and 5hmC only very weakly (Hashimoto, et al., 2012). For the human MBD, we neither observed binding to these combinations, including other previously not evaluated combinations.

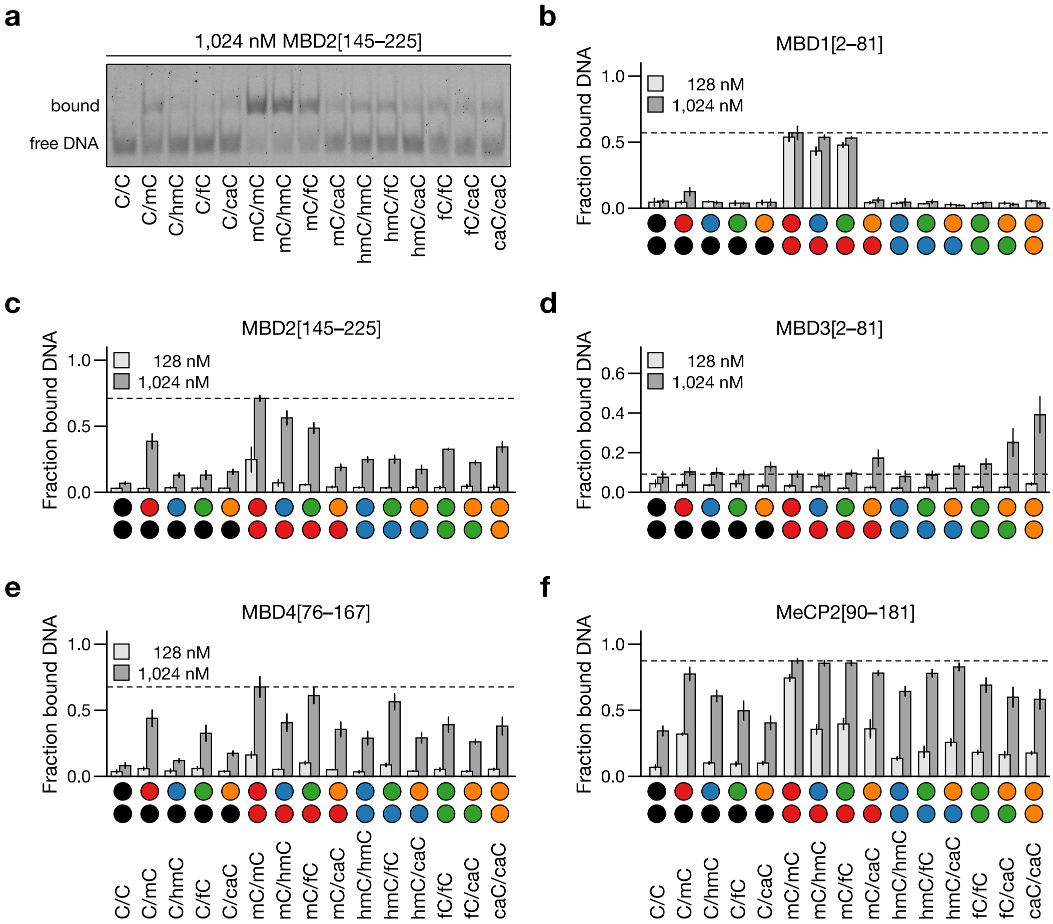


Figure 6.2 Binding selectivity of wildtype MBDs for different modified CpG dyads. (a) Representative electrophoretic mobility shift assay (EMSA) with fluorescently-labeled DNA duplexes and MBD2[145–225]. **(b)** Bar diagram of the fraction of MBD-bound DNA duplex for a sequence with a single modified CpG (as indicated) for MBD1[2-81], **(c)** for MBD2[146–225], **(d)** for MBD3[2–81], **(e)** for MBD4[76–167] and **(f)** for MeCP2[90–181]. Mean ± SEM of three independent experiments. Full data in Supplementary Figures A.3 and A.4. Modified from Buchmuller et al. (2020), CC BY 4.0 (full license in Appendix A.4).

However in our assay, binding was slightly less impaired in presence of a 5caC nucleobases in the CpG, preferentially when paired with a second 5caC or an 5fC (**Figure 6.2d**).

MBD4. The MBD of MBD4 which is part of a protein that exerts DNA glycosylase activity during base excision repair (Du, et al., 2015), is known to preferentially bind 5mC/5mC but with comparably low selectivity. The combinations 5mC/5hmC and 5mC/5fC were reported to be bound with similar affinity and higher affinity than 5hmC/5hmC or 5mC/5caC (Otani, et al., 2013). Our binding data were in agreement with these findings (**Figure 6.2e**), except that we observed higher binding to 5mC/5fC than to 5mC/5hmC. Moreover, our extended interaction profiles revealed 5hmC/5fC as a new preferential combination. Also C/5fC, 5fC/5fC and 5caC/5caC were recognized albeit with lower affinity.

The decreased binding to 5caC could be explained by vicinity of the negatively charged carboxyl group to Asp-107 (Asp-94 in the murine homolog, PDB 3vxx) where nonpolar nucleobases such as 5fC might be better accommodated (Muñoz-López & Summerer, 2018).

MeCP2. MeCP2 showed the highest affinity of the tested MBDs and a pronounced 5mC/5mC selectivity at lower protein concentrations (**Figure 6.2f**). The second highest affinities were observed for 5mC-containing combinations, including 5mC/5caC in contrast to MBD1 or MBD2. The presence of an unmodified cytosine diminished binding affinity in general, stronger than other modified cytosines did, confirming an earlier study by Yang et al. (2016).

Also, the asymmetric 5mC/5hmC was bound more tightly than 5hmC/5hmC, and our measurements at low protein concentrations further confirmed the earlier observation that symmetric 5hmC/5hmC and hemimodified C/5hmC are bound equally weakly, arguing for a potential modulating epigenetic role of these combinations (Khrapunov, et al., 2014).

6.2 Effect of MeCP2 Rett mutations on interactions with modified CpG dyads

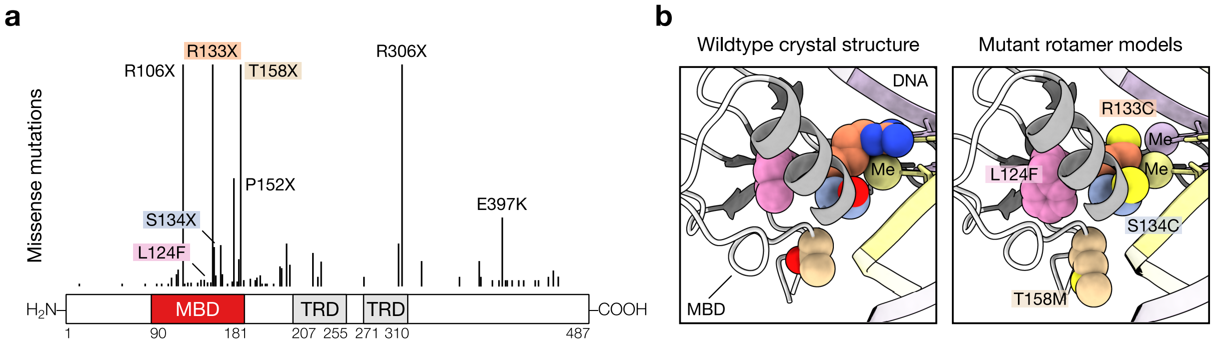


Figure 6.3 Protein architecture and positions of MeCP2 Rett mutants. (a) Protein architecture of full-length MeCP2 and frequency of Rett-associated missense mutations (RettBASE, Krishnaraj, et al., 2017). **(b)** Position of amino acid residues mutated in selected Rett genotypes highlighted in the wildtype MBD crystal structure in complex with a DNA duplex containing a 5mC/5mC CpG (PDB 3c2i, left) and artistic illustration of mutated residues based on the Dunbrack rotamer library (right); The methyl groups both 5mC are shown as van der Waals spheres; Models created and visualized with ChimeraX.

MeCP2 plays an important role in neurons, where it is nearly as abundant as histone octamers (Skene, et al., 2010) and therefore their selectivity at high concentrations might be physiologically more relevant.[a] Various mutations in the MBD of MeCP2 have been linked to the severe neurodevelopmental disorder Rett syndrome (reviewed by Ip, et al., 2018, **Figure 6.3a**). For some of these mutations, their effect on the recognition of fully methylated CpG dyads has been characterized (Ballestar, et al., 2000; Franklin, 2019; Yang, et al., 2016) and more recently a similar loss in binding affinity hypothesized for the interaction with methylated or hydroxymethylated CpA dinucleotides (Ibrahim, et al., 2021). Specifically, mutations in the DNA-contacting residues Arg-111, Tyr-120, Arg-133, Ser-134, Lys-135, Glu-137, Ala-140, and Thr-158 are some of the most frequently observed ones in Rett syndrome (Agarwal, et al., 2011). Their effect on the interaction with oxidized 5-methylcytosines in CpGs is however unexplored despite their high abundance in the brain (Wen & Tang, 2014) and a putative connection between Rett syndrome, MeCP2 and 5hmC binding (Kinde, et al., 2015).

To shed some light into this direction, we evaluated the four Rett mutants MeCP2[L124F], MeCP2[T158M], MeCP2[R133C] and MeCP2[S134C] (**Figure 6.3b**; **Figure 6.4a–d**).

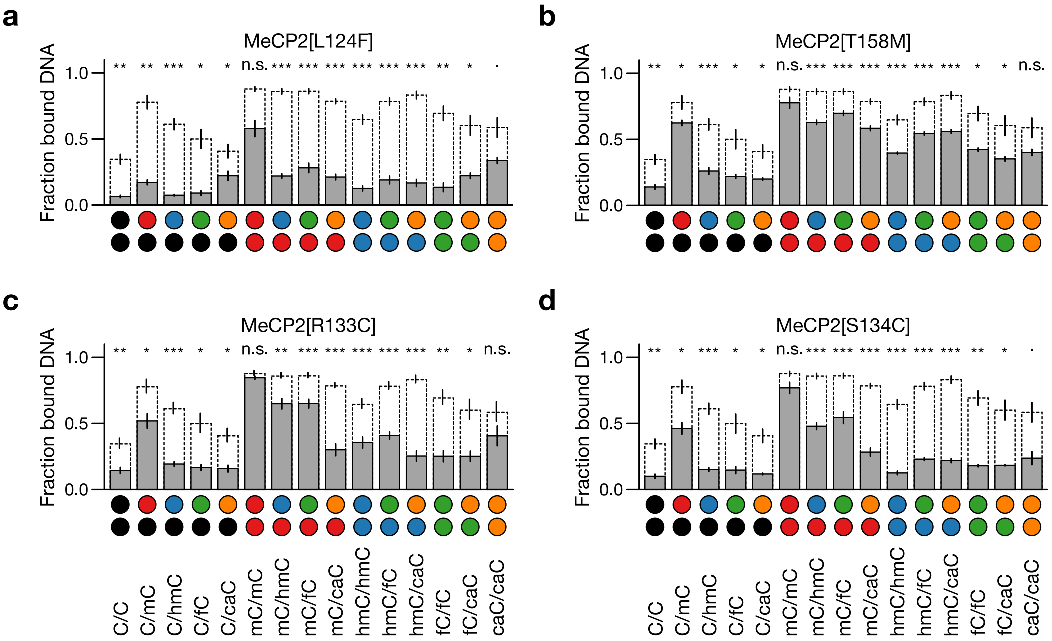


Figure 6.4 Selectivity and affinity of MeCP2 Rett mutants. Bar diagrams of the fraction of MBD-bound DNA duplex in an EMSA at 1,024 nM MBD concentration for different Rett mutants (gray) underlaid with wildtype MeCP2. **(a)** For the MeCP2[90–181] mutant L124F, **(b)** T158M, **(c)** R133C, and **(d)** S134C. Means ± SEM from three independent experiments; Student's *t*-test of mean differences between wildtype and mutant, n. s. indicates statistically not significant different means. Full data in Supplementary Figure A.5. Modified from Buchmuller et al. (2020), CC BY 4.0 (full license in Appendix A.4).

Overall, the mutants exhibited lower binding affinity than wildtype MeCP2 except for 5mC/5mC although a similar trend was appreciable. Especially L124F showed only very weak binding of 5mC-containing combinations other than 5mC/5mC. The mutants T158M, R133C and S134C, bound to CpGs that contained an 5mC still stronger than to other combinations but showed differences in their individual binding selectivities. Particularly, 5mC/5caC was bound with much lower affinity by the R133C and S134C mutants as compared to wildtype MeCP2 or T158M.

Moreover, whereas T158M, R133C and S134C exhibited selectivity profiles comparable to the wildtype for the five C-containing combinations (with C/5mC as the preferred combination), L124F slightly preferred C/5caC over C/5mC. Similarly, L124F and R133C preferentially retained the interaction with 5caC/5caC as compared to other higher oxidized combinations.

A particularly noteworthy difference was the seemingly higher loss in affinity for a CpG with a 5hmC/5hmC dyad in the case of the S134C mutant as compared to its loss in affinity for 5mC/5mC or 5mC/5hmC. Since lower oxidized combinations are more frequent in neurons (Wen & Tang, 2014), we determined the apparent dissociation constants K_d for four the interaction with four representative combinations of modified cytosines of TET oxidation, 5mC/5hmC, 5hmC/5hmC, 5mC/5fC and 5hmC/5fC, and C/5hmC as an expected product of the 'active modification-passive dilution' demethylation pathway (Wu & Zhang, 2017).

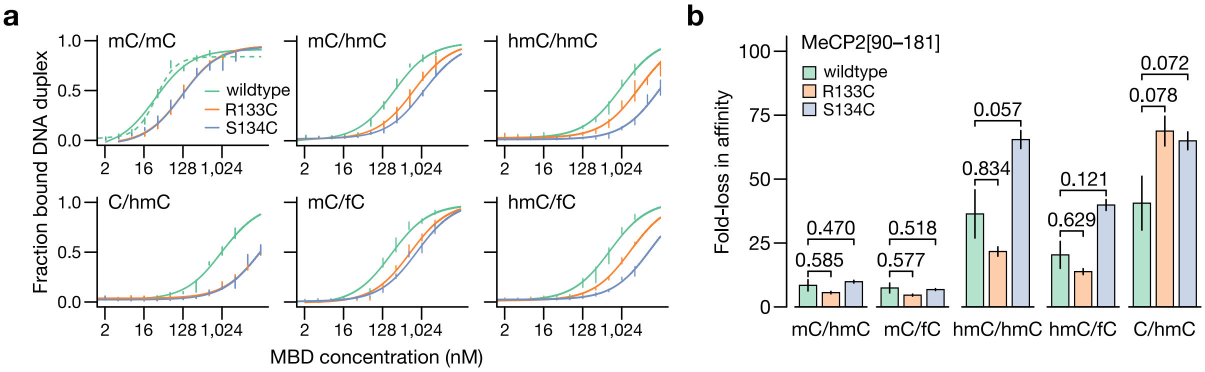


Figure 6.5 Affinity of MeCP2 Rett mutants R133C and S134C. (a) K_d-log diagrams of wildtype and Rett mutants; The wildtype binding with 5mC/5mC could also be modeled by a two-site binding isotherm (dashed line) as suggested by Khrapunov et al. (2014). **(b)** Fold-loss in affinity on basis of K_d for wildtype MeCP2 and Rett mutants for different oxidized 5-methylcytosine with respect to affinity for the 5mC/5mC CpG; Mean ± SEM, *p*-value from Student's *t*-test with Bonferroni correction. Modified from Buchmuller et al. (2020), CC BY 4.0 (full license in Appendix A.4).

Table 6.1 Dissociation constants for MeCP2 Rett mutants R133C and S134C. Estimates extracted from the models fitted in Figure 6.5. Details in Supplementary Table A.8.

Combination	Wildtype	R133C	S134C
5mC/5mC	27 ± 4 nM	118 ± 12 nM	117 ± 9 nM
5mC/5hmC	224 ± 25 nM	660 ± 70 nM	1,150 ± 70 nM
5mC/5fC	197 ± 24 nM	540 ± 50 nM	790 ± 50 nM
5hmC/5hmC	970 ± 90 nM	2,600 ± 300 nM	7,600 ± 600 nM
5hmC/5fC	540 ± 60 nM	1,630 ± 90 nM	4,600 ± 400 nM
C/5hmC	1,100 ± 100 nM	8,100 ± 500 nM	7,600 ± 700 nM

The R133C and the S134C mutant of MeCP2 lost about fourfold in affinity for the 5mC/5mC combination and about threefold for 5mC/5hmC and 5mC/5fC (**Figure 6.5a**, **Table 6.1**). Expectedly, the fold-loss difference in affinity against 5mC/5mC, a measure which is essentially unaffected by any potential differences in protein concentration estimates between the two specimen (**Figure 6.5b**), confirmed that there was no significant difference between the loss in binding of 5mC/5hmC or 5mC/5fC. The difference for the loss in binding affinity towards the higher oxidized combinations 5hmC/5hmC and 5hmC/5fC however was striking: While the S134C mutant lost affinity to a much greater extent than the wildtype, the R133C mutant seemed to even reverse this effect. For the non-5mC combination C/5hmC, the deficit between the mutants was again higher, but indifferent for the mutants.

Indeed, R133C and S134C belong to two different clusters in terms of folding stability observed across a panel of 13 Rett mutants with respect to the binding of unmethylated and methylated DNA (Yang, et al., 2016). R133C, the second most frequent Rett syndrome mutation, like T158M belong to a cluster that have folding stabilities similar to the wildtype, but disrupt certain interactions between the domain with the CpG (Arg-133 is part of the cation–π/H-bond stair motif) or within the domain (Thr-158 is part of the Asx-ST motif). S134C on the contrary belongs to a cluster with the less stable mutants and has a particularly decreased stability in loop L2 (upstream of position 134, Figure ??) due to a loss of two intra-protein hydrogen bonds. Apparently, this had important consequences for the ability of the domain to interact with some combinations of higher oxidized CpGs such as 5hmC/5hmC and 5hmC/5fC which probably needs a less flexible loop L2 which was lost in the S134C mutant. Combinations with methylated or unmethylated cytosines seemed not affected.

In summary, the presence of modified cytosines in CpG dyads affects their interaction with specific members of the MBD protein family in an MBD-specific manner. Therefore these modified CpG sites could potentially act in the genome as attenuated recruitment signals and the MBD family proteins as decoders that not only differ in their loading with additional effector domains or proteins but also in their propensity to recruit to the same modified CpG. This recruitment seems further dependent on their tissue expression level with higher concentrations potentially increasing the recruitment to CpGs with oxidized 5-methylcytosines. In addition, it must be taken into consideration that CpG sequence context preferences have been described for MeCP2 and several other MBDs (Chapter ??) which may further influence their binding behavior (Chapter ??) and genomic localization.

6.3 Synopsis

1. The four human MBDs of MBD1, MBD2, MBD4 and MeCP2 preferentially interact with fully methylated 5mC/5mC CpG dyads. The MBD of MBD3 interacts weakly with 5caC-containing CpG dyads, but not with 5mC/5mC CpG dyads.
2. Each MBD has specific binding preferences for different combinations of oxidized 5-methylcytosines at CpG dyads including previously uncharacterized interactions. These become relevant at high MBD concentrations as encountered for MeCP2 in neuronal tissue.
3. Mutations in the MBD that associated with Rett syndrome can specifically affect the interaction with distinct groups of these modified CpG dyads.

Endnote

[a] The number of MeCP2 was estimated to be 16×10^6 molecules (26.5 amol). Given an average nuclear volume of about 1,320 fL for a Purkinje neuron (BNID 103181), this corresponds to a nominal concentration of 20 μM. The nominal molar concentration of 40×10^6 methylated CpGs in a diploid murine genome in such nuclei is 50 μM.

Chapter 7

High-throughput screening for MBD variants with novel modified CpG dyad specificity

One strategy to tailor a protein for a specific, often novel purpose is to alter some part of its sequence either systematically or at random. In both cases, hundreds, thousands or even millions of protein variants must be tested for the desired property and conforming mutants retrieved. In this chapter, I present a platform that allowed to characterize the DNA-binding properties of wildtype and mutant methyl-CpG-binding domains (MBD) in a fast assay based on multicolor flow cytometry. Specifically, this assay was optimized to screen libraries with millions of MBD mutants for sequence alterations that lead to DNA binding selectivities not observed in wildtype MBDs, namely towards modified CpG dyads other than 5mC/5mC CpG dyads.

A set of candidate sites within the MBD sequence are identified in Section 7.1 based on the structural and functional considerations of Chapter 5 and Chapter 6. Section 7.2 describes how the sequence space at these sites was thoroughly randomized technically and how the respective combinatorial libraries were created as abstract genetic instructions. In Section 7.3, I put a method into praxis to translate these instructions into pooled collections of ‘tangible' proteins by using bacteria that express the MBD and physically attach the protein on the outside of their cell surface. The testing procedure to screen the individual MBD variants was established based on well-defined mixtures of such surface-displayed wildtype MBDs in form of a fluorescence-activated DNA binding assay (Section 7.4). Some of the library screenings and post-screening analyses are presented in Section 7.5, culminating in the discovery of MBD mutants with altered binding selectivity. I conclude with a summary of additional screening and diversification efforts.

7.1 Candidate MBD residues to target for altered DNA binding selectivity

The inspection of the DNA binding sites of the five human MBDs revealed several residues in immediate proximity of the 5mC nucleobases that could be reasonable targets to address in a protein engineering effort (Chapter 5.3). As discussed there, Arg-22, Asp-32 and Arg-44 (Arg-111, Asp-121 and Arg-133 in MeCP2) are essential residues for the recognition of the CpG dyad, independent of cytosine modification and evolutionary highly conserved across all domains of life. Therefore, other residues with more specialized role in the recognition of the methyl group were sought out for the libraries discussed in this work (**Figure 7.1**).

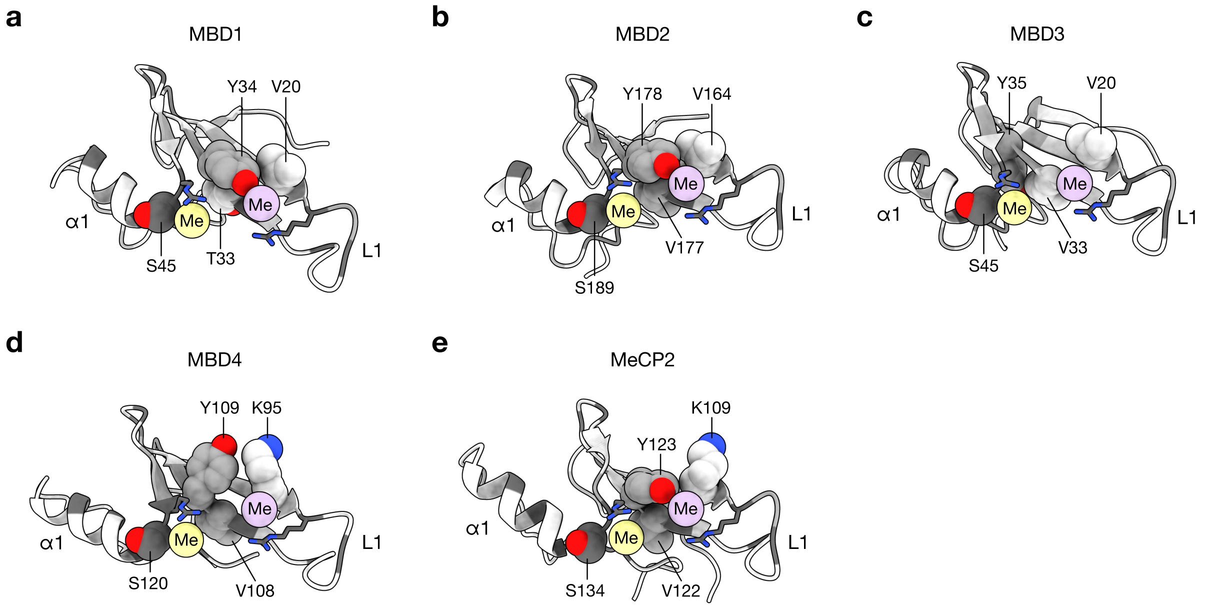


Figure 7.1 Positions of the degenerated amino acid residues in five MBD libraries. (a) In MBD1[2–81] (PDB 6d1t, SGC), **(b)** MBD2[146–225] (PDB 6cnq, Liu, et al., 2018), **(c)** MBD3[2–81] (PDB 6ccg, Liu, et al., 2019), **(d)** MBD4[76–167] (PDB 4lg7, SGC), and **(e)** MeCP2[90–181] (PDB 3c2i, Ho, et al., 2008). α-helix α1 and loop L1 of the MBD; Only the sp^3 carbon of each 5mC (Me) is indicated for clarity.

– Tyrosine-34 (Tyr-123 in MeCP2) participates in the recognition of the methyl group through interaction with structured water in the binding site of many MBDs and its unique dislocation in MBD4 allows other modified and unmodified DNA nucleobases to engage with the MBD. Although a non-conservative replacement with a purely aromatic phenylalanine in MBD3 abolishes binding to methylated CpGs, weak binding towards 5caC/5caC CpGs was observed. It was therefore assumed that substitutions at this position could be critical for novel DNA binding selectivity.

In MBD3, Tyr-35 was targeted instead of Phe-34 to probe a slightly different sequence space.

– Valine-20 contributes to the hydrophobic patch in the DNA binding pocket nearby the 'upper' 5mC · G base pair in MBD1, MBD2 and MBD3. Likewise, the analogous lysine-109 in MeCP2 (and MBD4) interacts with the negatively charged phosphodiester backbone at this site, maybe contributing to the readout of changes in DNA shape. Thus, if the methyl group of 5mC was replaced with another functional group, breaking these interactions or even creating other more favorable interactions at this site could be required.

– Threonine-/Valine-33 (Val-122 in MeCP2) is located on the far side of the β-sheet so that its side chain is not engaged in DNA binding. However, it was speculated that the increasing steric demands of 5hmC, 5fC, and 5caC could be addressed with a possibly better suited residue at this position.

- Serine-45 (Ser-134 in MeCP2) participates in nonspecific DNA backbone interactions around the 'lower' 5mC · G base pair in all five MBDs. From our structure-function analysis of the Rett syndrome-associated MeCP2[S134C] mutant, which reduced affinity for higher oxidized CpGs, in particular 5hmC/5hmC, it's role for nucleobase selectivity seemed apparent.

Within this set, two residues, Tyr-34 and Val-20 (Lys-109 and Tyr-123) were located in vicinity of the 'upper' 5mC · G base pair in the DNA binding site, whereas Thr-/Val-33 and Ser-45 (Ser-134 and Val-122) were in vicinity of the 'lower' 5mC · G base pair. If the DNA interactions at these sites were mutually independent of each other, this could provide a means to also engineer selective interactions for strand-asymmetrically modified CpG dyads.

Libraries with alternative degeneration schemes are available (**Supplementary Figure A.6**).

7.2 Creation of site-saturated codon-degenerated MBD libraries

In the language of the standard genetic code, all 20 canonical amino acids are encoded by the degenerated triplet codon NNK (N = A, C, T, G; and K = G, T). In the physical world however, there is no such thing as degeneracy of a single DNA molecule: Degeneracy is a property of mixtures. Such mixtures are accessible to the synthetic chemist using appropriately mixed monomer building blocks during DNA synthesis and nowadays available commercially. The concern of the protein engineer hence is to maintain the mixtures degenerated throughout all preparatory steps until the library is screened for the first time. With four degenerated NNK sites in the MBD library design (Section 7.1), a complexity of one million different genotypes that encoded more than 194,000 phenotypes[a] had to be maintained. This required more than 3 million transformants for a statistical coverage at 95% completeness (Reetz, et al., 2008).

To get hold of such high numbers of transformants in the least time-consuming and labor-intensive way, two cloning strategies to propagate the complexity of the degenerated DNA oligonucleotide mixtures into vectors that encode the wildtype MBD (**Figure 7.2a**) were tested: (1) Amplification of the entry vector by polymerase chain reaction (PCR) with oligonucleotide primers that each contain a number of NNK degenerated sites and a type IIS restriction enzyme recognition sequence for traceless religation of the amplicon (Beck & Burtscher, 1994). (2) Isothermal enzymatic assembly (Gibson, et al., 2009) of the PCR-linearized entry vector with a short, degenerated double-stranded oligonucleotide.

Both strategies yielded degenerated mixtures of entry vectors (data not shown). Yet, only the modified Gibson assembly strategy (Methods)[b] afforded sufficiently high numbers of transformants to create the five MBD libraries in 10 to 15 transformations each (**Figure 7.2c**). In contrast, the traceless restriction-ligation strategy yielded 10-times less colony-forming units

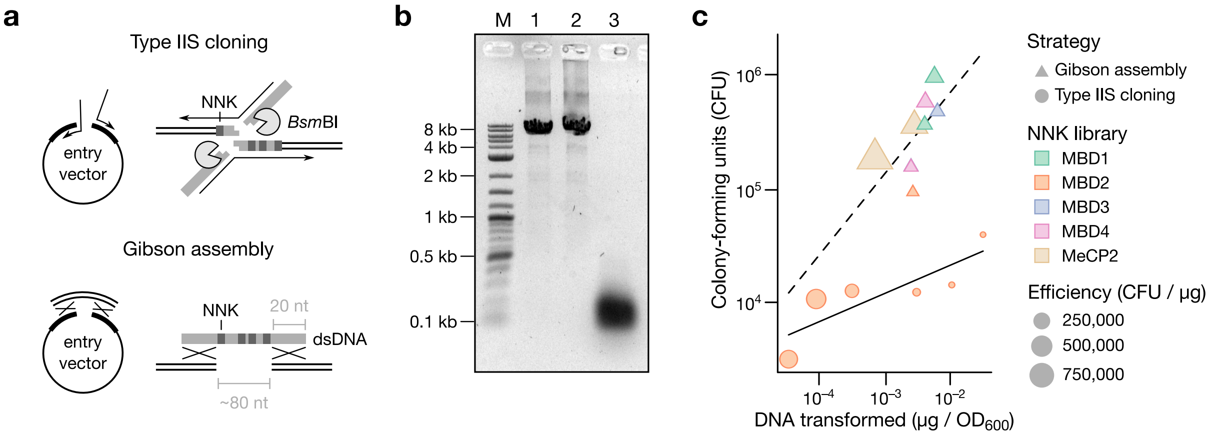


Figure 7.2 Strategies to create site-saturated codon-degenerated MBD libraries. (a) Schematic representation of the introduction of four NNK degenerated codons in the coding sequence of the MBD using either the type IIS cutter *Bsm*BI (top) or Gibson assembly (lower panel). **(b)** PCR-linearized entry vectors (lane 1, type IIS cloning; lane 2, Gibson assembly) and the degenerated primer after Klenow fragment primer extension (lane 3, Gibson assembly); 1% agarose gel electrophoresis, 1 kb Plus DNA Ladder (New England Biolabs; lane M). **(c)** Yield in colony-forming units (CFU) per microgram of DNA transformed into high-density electrocompetent *E. coli* DH10B.

per reaction, impeding effective scaling of library construction. Further, the requirement to dilute the *Bsm*BI-digested amplicons for self-circularization and to concentrate the material again, added another day to complete the procedure. The isothermal assembly protocol however was ready for transformation within four hours. So, it was still economic to generate 8 to 10 million clones per library using the Gibson strategy which raised the oversampling factor to 8- to 9-fold and the nominal completeness above 99% (Reetz, et al., 2008).

Empirically, the degeneracy was assessed by DNA sequencing. All MBD libraries showed the expected nucleobase degeneration in the Sanger sequencing chromatograms (**Figure 7.3a**). For the MeCP2 library, the underlying genotype composition was gauged from two non-exhaustive next-generation sequencing (NGS) runs at 0.5% sequencing depth (**Figure 7.3b–d**): 89% of the identified genotypes (35,000) accounted for 65% of the sequenced clones (48,000); the 20 most abundant genotypes together for 6.8% including 3.6% wildtype sequences. Considering that 11% of the genotypes were frequent enough to be detected in both samples, a 4- to 5-fold oversampling should suffice to recover the genotypes almost completely.

The presence of wildtype was likely the result of an undesired yet inevitable carryover of some side-products of the PCR linearization, or, due to incomplete removal of the plasmid template by *Dpn*I treatment. In reverse conclusion, the low frequency of 3.6% suggested that both processes must have completed at an average 98.5% success rate. For the throughput of the screening, encountering one wildtype in every 30 clones analyzed imposed little burden if a high-throughput platform could be used. Quite the contrary, this might ensure the wildtype is not missed if it met the screening criteria to an equal degree.

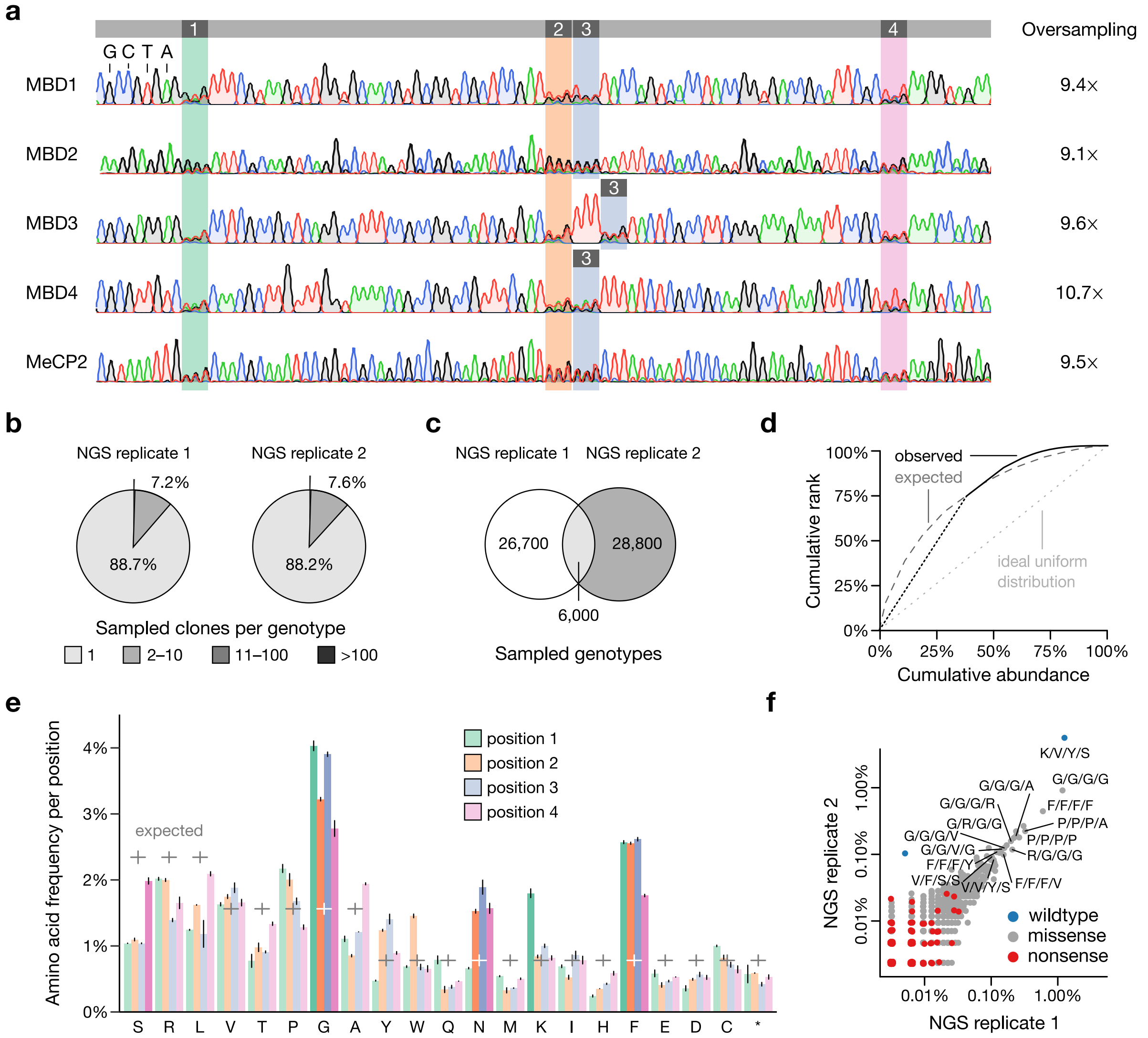


Figure 7.3 Degeneracy of the MBD libraries. (a) Sanger sequencing chromatograms of the codon-degenerated MBD libraries created via the modified Gibson assembly strategy. **(b)** Fraction of genotypes in MeCP2 library by clonal abundance based on short-read next-generation sequencing and unique molecular identifier (UMI) counting in 48,000 reads per replicate. **(c)** Number of shared and disjoint genotypes sampled in *b*. **(d)** Cumulative distribution of phenotypes (combined samples of *b*) ranked by clonal abundance; Observed and expected trends for an ideal NNK degenerated library with four sites and a trimer-based combinatorial library (uniform abundance of all amino acid combinations); The dotted section in the observed trend are phenotypes with 1 UMI. **(e)** Observed frequency of encoded amino acids by degenerated position, expected abundance for an NNK degenerated library (crosses), major deviations in flushed colors. **(f)** Phenotype representation based on 6,000 recurrent genotypes in *c*; 88.5% missense mutants (expected: 88.1%), 7.8% nonsense mutants (expected: 11.9%), 3.7% wildtype MeCP2 (expected: <0.01%), disallowed stop codons (TAA, TGA) <0.09%; Pearson correlation $\rho = 0.7$ (p-value < 0.001).

In terms of codon frequency per degenerated position (excluding exact matches to the wildtype genotype; **Figure 7.3e–f**), the glycine (GGG, GGT), asparagine (AAT) and phenylalanine (TTT) codons occurred almost twice as often as expected for an ideal NNK library. In about 12% of all genotypes more than two of such codons were present simultaneously. The origin of this overrepresentation remained however unclear. The relative overrepresentation of the lysine wildtype AAG at the first position and the wildtype serine AGT codon at the last position were a consequence of the slower exonucleic resection rate in presence of the aforementioned wildtype sequences for isothermal assembly. Nonsense codons were as frequent as expected.

In summary, the NGS analysis suggested that more than 80% of the clones created with the modified Gibson assembly strategy contained a unique genotype. Relying on a simple primer extension with no subsequent cycling to amplify the degenerated DNA, the initial degeneracy of the oligonucleotide mixture was supposedly maintained with only minor deviations. When screened with appropriate oversampling, these libraries were deemed representative for a mixture expected after NNK codon degeneration.[c]

7.3 Bacterial AIDA-I-mediated MBD cell surface display

To assess the phenotypic consequences within the created sequence space for the MBD–DNA interaction, an *ex vivo* screening platform was established that could be used to partition MBD mutants with desired DNA binding properties. Hadley D. Sikes has used a yeast-based cell-surface platform to display the MBDs of MBD1, MBD2, MBD4, and MeCP2 and to select variants generated by error-prone PCR with improved binding affinity for fully methylated 5mC/5mC CpGs (Heimer, et al., 2015) and hemimethylated C/5mC CpGs (Tam, et al., 2016). Here, we opted for a bacterial platform which in general has similar properties to the yeast surface display platform but unlike yeast does not glycosylate the surface-displayed payload (Boder & Wittrup, 1997). Therefore, the state of the screened protein would be more similar to a recombinantly expressed MBD used in future diagnostic procedures.[d] In the adhesin involved in diffuse adherence AIDA-I autotransporter cassette designed for this work (compare Figure 4.2) the wildtype signal peptide was replaced with the more effective one of cholera toxin B (*Vibrio cholereae* O1) bearing a I2V mutation (Maurer, et al., 1997) and featured an in-frame c-Myc epitope downstream of the N-terminal passenger to verify the transport of the payload to the outer membrane via antibody staining (**Supplementary Figure A.7**).

The gene to express such MBD-AIDA^{C} autotransporters was brought under control of one of two inducible promoters, namely the β-ᴅ-1-thiogalactopyranoside (IPTG)-inducible *T7lac* promoter (pET vectors) or the ʟ-arabinose-inducible *araBAD* promoter (pBAD vectors). Both of the realized expression systems resulted in functionally displayed MBD passengers (**Figure 7.4**). Surface-display was highest for MBD2, MBD3 and MBD4, whereas MBD1 and MeCP2 showed little or no exposed or accessible c-Myc epitope. However, low levels of DNA binding for these two constructs was still detectable, arguing for low display levels rather than dysfunctional protein states. MBD3 in contrast showed, as expected, negligible levels of DNA binding in comparison to its high surface display level.

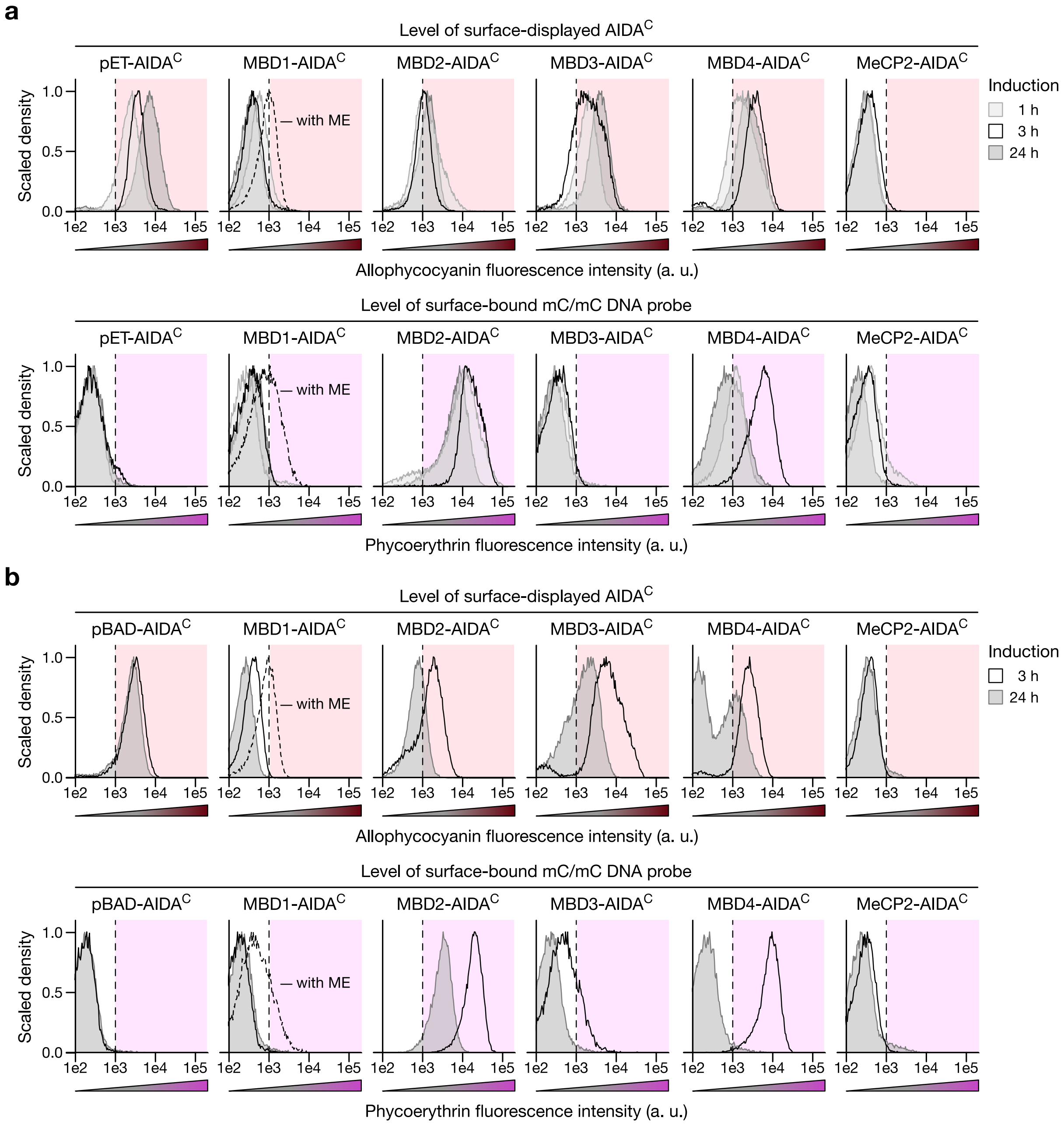


Figure 7.4 Surface-display of various MBDs using pET- and pBAD-based expression system. Display levels and MBD–DNA binding with B strain Tuner™(DE3) cells. After induction, the cells were probed for the presence of a surface-exposed c-Myc epitope using an allophycocyanin-coupled antibody and for binding of a DNA probe containing a fully methylated 5mC/5mC CpG that was labeled with phycoerythrin. MBD3 is a non-binding MBD. **(a)** Induction of *T7lac* promoter-based pET vectors with 50 µM IPTG for 1 hour, 3 hours or overnight. **(b)** Induction of *araBAD* promoter-based pBAD vectors with 0.1% L-arabinose for 3 hours or overnight. For MBD1-AIDA^C, 10 mM 2-mercaptoethanol (ME) was added to the growth medium where indicated.

For MBD1-AIDA^C which carried the only MBD with three cysteines, display and DNA binding levels increased after 2-mercaptoethanol was added to the expression medium, potentially relieving a prematurely folded protein state from the periplasmic space. Further, its outer-membrane localization and full-length expression were confirmed (**Supplementary Figure A.8**).

For the cysteine-free MeCP2-AIDA^{C}, surface display levels could vary drastically between experiments (as will become apparent in this chapter). Despite several attempts, no growth or induction conditions could be identified that would promote reproducibly high levels of a surface-exposed MBD. Since MeCP2 could also bind to CpA dinucleotides, it was possible that the MBD became trapped inside the cells when exposed to the bacterial nucleoid. In support of this hypothesis, a non-binding double-arginine mutant (R111A, R133A) of MeCP2 displayed as high as MBD2 with no affinity towards CpA dinucleotides (**Supplementary Figure A.9**).

Besides high and ideally constant surface display levels, a successful screening campaign of a randomized protein library also must prevent the library's diversity from perishing of cytotoxic stress so that the mutants of interest alive can be recovered alive. Such optimization could be carried out only with the vector-host combination pET and B strain Tuner™(DE3)[e] but neither with pBAD and K-12 strain DH10B (no functional MBD display; **Supplementary Figure A.10**) nor with pBAD and B strain cells (no control by inducer titration).[f] Therefore, I focused on working with pET vectors in a B strain host.

(7.3.1) Maximizing MBD-AIDA^{C} surface display levels and survival

Contrary to induction of the empty pET-AIDA^{C} vector, surface display with an MBD passenger stagnated after one hour of induction (see again **Figure 7.4a**), suggesting the system had reached its capacity of producing or displaying the payload. To possibly maximize the yield of active surface-expressed MBDs, cultivation and induction conditions were varied.

Optimization of inducer concentration. The genetic constitution of the chosen expression host for the pET-AIDA^{C} system allowed to control the production of payload via the inducer concentration in the growth medium, thereby providing a means to tune display level and survival. Indeed, surface display levels of MBD2, MBD4 and MeCP2 increased steadily over a range of IPTG concentrations up to 100 µM (**Figure 7.5**). Higher concentrations reduced the levels of surface-exposed MBD-AIDA^{C}. Importantly, no increased toxicity was observed even at higher IPTG regimes up to 500 µM, suggesting that another pathway must have brought down the number of displayed MBDs such as intracellular aggregation.

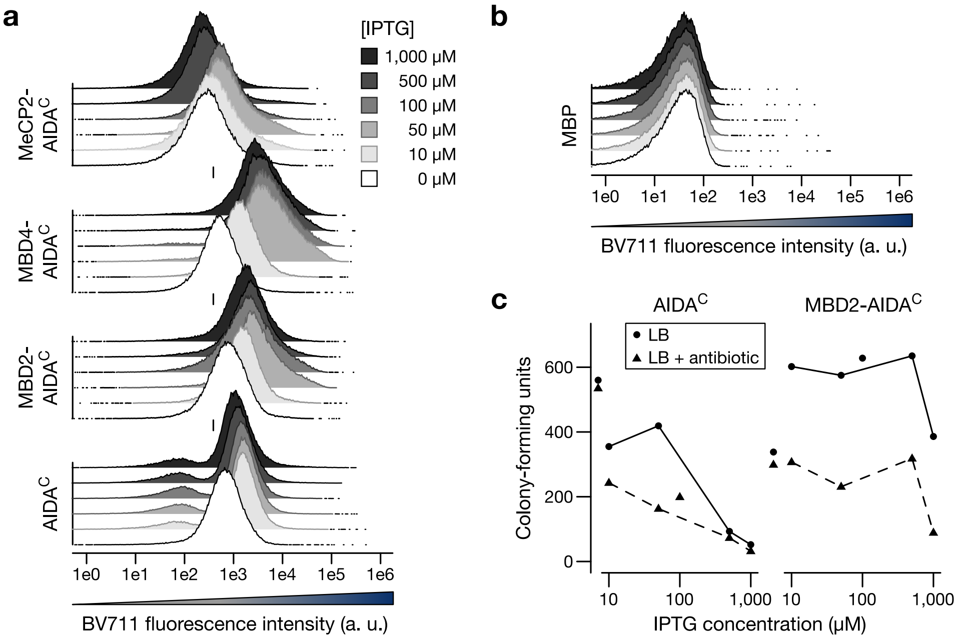


Figure 7.5 Inducer concentration and survival in the pET-MBD-AIDA^C system. Expression of recombinant proteins was induced with IPTG in BL21 Tuner™(DE3) for 1 hour. **(a)** With pET-AIDA^C vectors; Presence of a surface-exposed c-Myc epitope was probed with a Brilliant Violet™ 711 (BV711)-coupled antibody as proxy for surface display of a passenger. **(b)** As in *a* with pET-maltose-binding protein (MBP), an intracellular protein. **(c)** Survival and plasmid loss after sorting from the upper 50% quantile of induced cells using a flow cytometer and spread-plating onto LB agar plates. Survival in the population transformed with pET-MBP in a similar procedure resulted in 70% survival upon treatment with 1,000 µM IPTG.

In contrast, the display of the the empty AIDA^C autotransporter did not increase much with IPTG levels and higher concentrations drastically diminished survival, possibly because of the disintegration of the bacterial envelope (Chung, et al., 2020). In turn, a population that had escaped cell surface display became apparent, probably the only recovered viable fraction at the highest IPTG concentrations tested.

Whether or not an MBD was present, expression of the AIDA-I autotransporter from the pET vectors was leaky raising basal levels of surface-exposed c-Myc epitope significantly above background. Since this made little difference for the survival in MBD-AIDA^C strains, no further optimization was undertaken with the exception that pre-cultures were supplemented with 20 mM glucose to further subdue *lacO* derepression. However, the almost 70% reduction of surviving clones with an empty AIDA-I cassette at as little as 10 µM IPTG for one hour accentuated that for protein payloads other than MBDs tighter expression systems might be required to prevent demise of library diversity prior to screening or selection. For the pET-MBD-AIDA^C system, an acceptable balance of display level and survival was 50 µM IPTG for one hour at 30 °C.

Optimization of maturation time. Maturation of the cells in absence of inducer slightly increased the level of functional MBDs displayed on the cell surface (**Supplementary Figure A.11**). Therefore, maturation was carried out for one hour on ice or longer if the schedule permitted.

(7.3.2) Handling cultures during and after the screening assay

As important as the preservation of library diversity prior to a screening is the proper separation of a target population from the heterogenous starting material and its preservation for rescreening or post-screening assays.

Cells displaying the autotransporter had a propensity to flocculate ('diffuse adherence') by which false positive 'parasites' might be carried along the screen. Besides thorough resuspension of the cells, treating the samples with 0.1% ultrapure bovine serum albumin (BSA) was found to mitigate this phenomenon. The simultaneous detection of the c-Myc epitope as a proxy for surface display and the bound DNA probes had no benefit to rejecting false positive events (**Supplementary Figure A.12**).

Various recovery conditions including outgrowth in LB medium or on solid LB agar plates had no apparent effect on library composition. Neither had inoculation from glycerol stocks a significant effect on the fraction of surface-displayed cells.

7.4 Fluorescence-activated MBD–DNA binding assay

Fluorescence-activated cell sorting (FACS) is a convenient means to separate mixtures of small particles or cells if the biochemical properties of interest can be probed with fluorescent substrates or fluorescently-labeled ligands. Here, the opportunity of a multi-color instrument was taken advantage of to assay the binding selectivity of surface-displayed MBDs, i. e., to probe the presence of specific ratios of different DNA probes on the cell surface. Two assays were set up (**Figure 7.6**): One in which the binding of each target was measured in a separate reaction using the same fluorophore ('one-color assay') and one in which this assessment was carried out simultaneously using mixtures of fluorescently-labeled DNA probes ('two-color assay'). The two-color assay would be used for the screening of binders with altered binding selectivity, the one-color assay could be used to preliminarily characterize candidates without the need to recombinantly express and purify the protein.

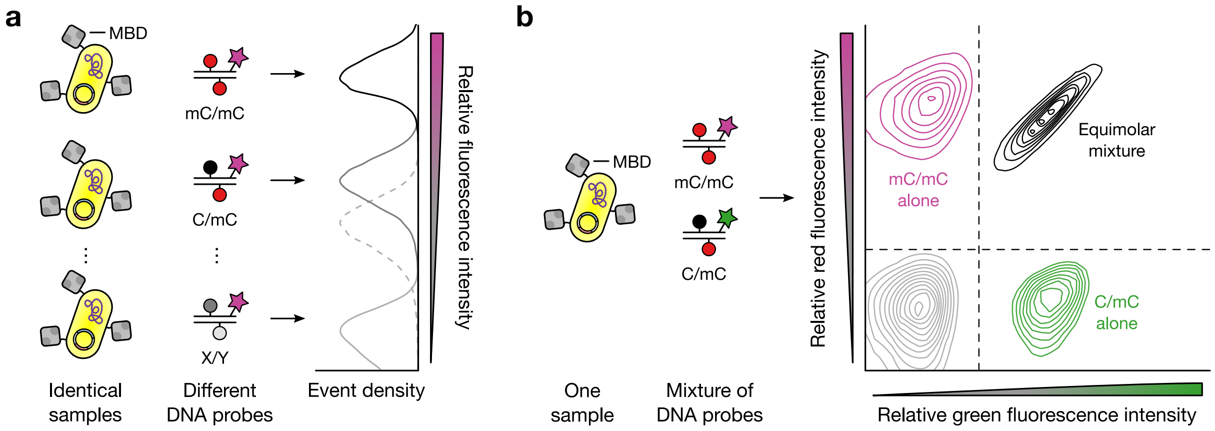


Figure 7.6 Fluorescence-based assays to assess DNA binding affinity and selectivity. The relative binding affinities of surface-displayed methyl-CpG-binding domains (MBDs) can be determined in **(a)** an 'one-color assay' after splitting the population in separate samples that are treated each with different DNA probes but labeled with the same fluorophore, or **(b)** a 'two-color assay' using one sample treated with a mixture of DNA probes in which different probes are labeled with a different fluorophores; Binary fluorophore combination shown.

In contrast to the kinetic screening of Heimer et al. (2015), an equilibrium screening for MBD–DNA binding was established to effectuate ligand competition (see Figure 4.4). Although it is possible to carry out such assays labeling only one of the competing ligands, all ligands in the mixture were fluorescently labeled. This allowed to distinguish between surface-displayed proteins with generally low display level and those with low binding selectivity.

(7.4.1) Establishing a competitive single-cell equilibrium binding assay

Within the limits of MBD surface display tolerated by the host cells, conditions had to be established under which an MBD–DNA binding assay could be conducted on the surface of a single bacterial cell so that the signal would still be strong enough to differentiate binding strengths.

Signal amplification: Choice of fluorophores and staining procedure. Covalently labeled DNA probes with a single fluorophore (fluorescein, tetramethylrhodamine, Pacific Blue) did not emit sufficient light to detect MBD–DNA binding on the FACS instrument (data not shown). Therefore, signal amplification with a secondary fluorochrome-conjugated reagent was considered. Since multiple fluorochromes can be attached to such reagents, e. g., a single streptavidin tetramer can carry 20 to 21 fluorescein molecules on average, they offer a higher brightness and hence lower the limit of detection. Further, a biotinylated probe could be used with different fluorophores, lowering material costs and increasing versatility.

Although it was possible to choose a staining procedure in which the biotinylated DNA probes were first bound to the surface-displayed MBDs and later labeled with a streptavidin-fluorochrome conjugate (data not shown), such proceeding was not compatible with the differentially labeling of multiple ligands to probe binding selectivity. Therefore, a staining protocol was conceived (**Supplementary Figure A.13**) in which the different ligands were labeled separately, excess reagents quenched and the labeled ligands mixed at the desired molar ratios before being applied at once to the surface-displayed proteins to compete for the binding sites.

Two biotin tags per DNA probe were found to suffice the detection on the FACS instrument. Although a probe with a single biotin which was labeled with a streptavidin conjugated to a fluorochrome of high quantum yield such as phycoerythrin (Johnson, 2010) was sufficient to detect the binding of MBD2 to a fully methylated 5mC/5mC CpG (**Figure 7.7a**), it merely outshined the background for a hemimethylated 5mC/C target, even less so with MeCP2 as an MBD that generally achieved lower surface display levels. With fluorochromes of lower quantum yield (**Figure 7.7b**) such interactions were detectable only when two or three biotin tags per DNA probe were installed. Such probes adequately reflected the differential binding affinity of the surface-displayed MBDs and on the account of costs and benefits, the doubly biotinylated probes were taken forward, i. e., those with one biotin tag per modified DNA strand. An appropriate excess of the streptavidin reagent ensured a homogeneously labeled DNA probe using these twofold multivalent reagents, the tetravalent streptavidin on the one hand and the bivalent DNA on the other hand (**Supplementary Figure A.14**; Methods).

Equilibrium considerations and reagent economy. To meet the requirements for an equilibrium binding assay, a concentration 5- to 10-fold above the expected equilibrium dissociation constant K_d of the highest binding interaction is required (Boder & Wittrup, 1998). Further, all DNA probes must be supplied in at least tenfold excess over the number of displayed proteins to avoid ligand depletion (Cherf & Cochran, 2015). Between 10,000 to 300,000 AIDA-I autotransporters can be present on the outer membrane of a single cell (Kaeßler, et al., 2011). With a balanced estimate of 50,000 displayed molecules per cell and 20 million cells to be screened per library (to assert tenfold oversampling), 1.7 pmol of surface-displayed MBD proteins are present per staining reaction if all of the MBD mutants were functional.

In order to experimentally verify this number, the minimum amount of DNA probe required to homogeneously stain a sample of 2 million cells displaying MBD2 under optimal conditions was determined (**Supplementary Figure A.15**). This minimum was about 0.2 pmol, thus in good agreement with the estimate. If only a small fraction of the MBD mutants would be functional DNA binders, which seemed to be a reasonable assumption, this amount would also suffice for the excess required in equilibrium binding. The final concentration of the DNA probe hence was set to 64 nM, which was about five to tenfold above the K_d reported for most MBDs towards fully methylated 5mC/5mC CpGs.

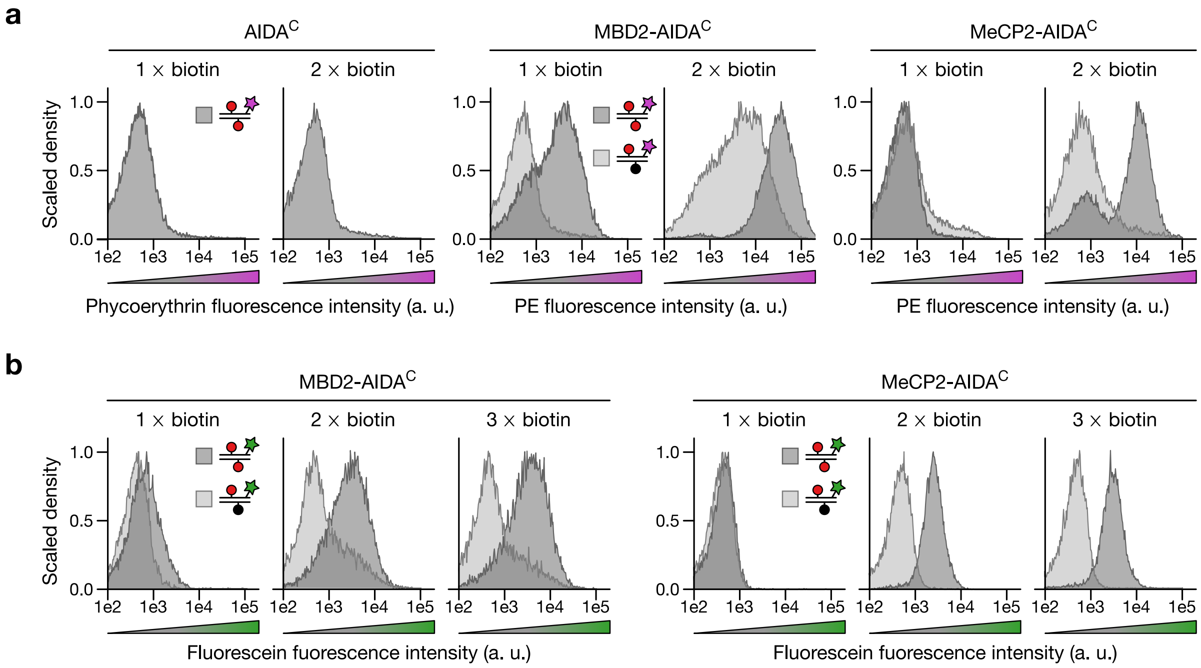


Figure 7.7 Choice of fluorophore amount for MBD–DNA binding detection on FACS. Surface-displayed MBDs (20 million BL21 Tuner™(DE3) cells, pET-AIDA^C, 50 µM IPTG) were stained with 7.5 pmol labeled DNA probes containing a single 5mC/5mC or 5mC/C CpG and 1, 2 or 3 biotin tags. The DNA probe was labeled with a fluorochrome-streptavidin conjugate before staining, stained cells extensively washed and analyzed on a multi-color flow cytometer (10,000 events shown). **(a)** Labeling with 3-fold excess phycoerythrin (PE)-streptavidin (1.2:1 conjugation ratio). **(b)** Labeling with 3-fold excess 6-carboxyfluorescein isothiocyanate (FITC)-streptavidin (20.6:1).

Assay performance. Since two or more DNA probes labeled with different fluorophores were present simultaneously in the binding reaction of the two-color assay, any inadvertently blending of active streptavidin agents was effectively subdued by quenching with a high molar excess of biotin (**Figure 7.8**; fully optimized conditions).

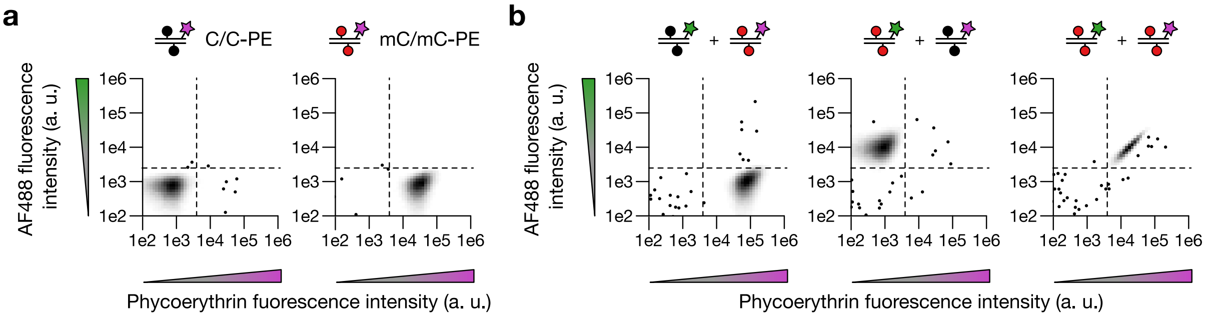


Figure 7.8 Suppressing DNA probe relabeling in multi-color FACS stains. Surface-displayed MBD2 (2 million BL21 Tuner™(DE3) cells, pET-AIDA^C, 50 µM IPTG) stained with 2 pmol labeled DNA probes containing a single C/C or 5mC/5mC CpG and two biotin tags. The DNA probes were labeled with 6 pmol fluorochrome-streptavidin conjugate and quenched with 120 pm biotin. The stained cells were extensively washed and analyzed on a multi-color flow cytometer (10,000 events shown). **(a)** Staining with 1 of dsDNA probes labeld with a single fluorophore. **(b)** Staining with a mixture of 0.5 of each dsDNA probe labeld with streptavidin conjugated to either phycoerythrin (PE) or the fluorescein derivative Alexa Fluor 488 (AF488). The sparsest 0.1% of the events are shown as dots.

From the evaluation of this probably most extreme case in terms of MBD binding selectivity, discriminating an unmodified C/C from a fully methylated 5mC/5mC CpG, the sensitivity and specificity of the two-color assay was established (**Figure 7.9**). When the fluorescence intensity thresholds ('gates') were set such that the fraction of double-positive events was as high as 1% and the phycoerythrin-labeled probe was the desired 5mC/5mC 'on-target', the specificity (rejection of true negative events) for the costaining was 99.9% and the sensitivity (detection of true positive events) was 99.9%. For the Alexa Fluor 488-based labeling, the specificity and sensitivity were 99.1% and 97.9% respectively. The false discovery rate was 0.12% with the red fluorophore and 0.91% with the green fluorophore. This suggested a high performance of the optimized two-color assay to reliably discriminate binding events from non-binding events.

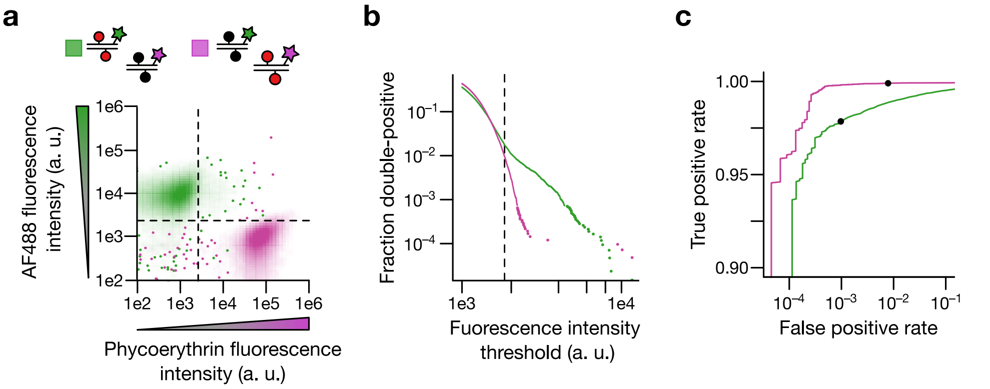


Figure 7.9 Sensitivity and specificity by fluorophore in the two-color assay. Same data as in Figrue 7.8 b. **(a)** Overlay of two separate staining reactions with reciprocal DNA probe–fluorophore combination for surface-displayed MBD2 with an equimolar mixture of C/C or 5mC/5mC DNA probes (green: 5mC/5mC-AF488-streptavidin and C/C-phycoerythrin (PE)-streptavidin; magenta: C/C-AF488-streptavidin and 5mC/5mC-PE-streptavidin). **(b)** Quantitation of the fraction of double-positive cells in dependence of the fluorescence intensity thresholds in *a*. **(c)** True positive and false positive rate in dependence of the fluorescence intensity thresholds in *a*. 50,000 events analyzed, the sparsest 0.2% are shown as dots. Source code available.

A more difficult scenario was met with less extreme cases. To assess the assay's power to discriminate binders of different selectivity from each other, e. g., a binder with 20-fold selectivity from a binder with 10-fold selectivity or no selectivity, cells were stained with mixtures reflecting the expected fluorophore fractions of 5%, 10% and 50% on the cell surface. To ensure that in this experiment the differences in probe affinity could be neglected when interpreting the result, the same 5mC/5mC DNA probe was used in for both labeling reactions, i. e., once with a red fluorophore and once with a green fluorophore. Further, the assessment was carried out for pure and dilute mixtures of a binder population in a non-binder population (**Figure 7.10**).

Independent of the binder dilution, the distinct populations of different selectivity ratios were discernible. Whereas the highly selective populations resided almost exclusively in the lower-right quadrant when phycoerythrin was used to label the 'on-target' probe, the 10- and 20-fold selective populations labeled with a Alexa Fluor 488 fluorophore occupied rather the top-right

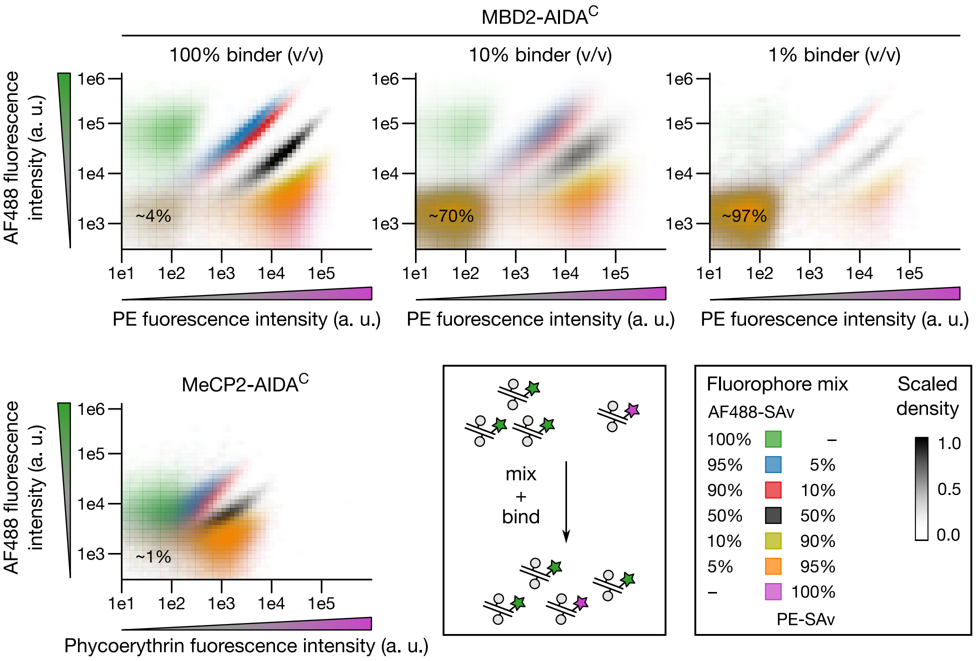


Figure 7.10 Discrimination of surface-displayed DNA binders by their binding selectivity in the two-color assay. Surface-displayed MBD2 or MeCP2 (50 µM IPTG) were stained with mixtures made from the same DNA probe, yet different amounts of the labeling reagents AF488-streptavidin (SAv) or phycoerythrin (PE)-SAv to simulate a defined binding selectivity. The MBD2 population was further diluted in a population displaying the empty AIDA-I; Percentages indicate the fraction of double-negative events, gates not shown. 50,000 events per condition are displayed.

quadrant, suggesting that this area would have to be screened in such a screening stratagem.

Another apparent consequence of the lower brightness of the fluorescein-like fluorophore was that when used as an ‘off-target’ label in an ‘on-target’ screen with phycoerythrin, fully, 10- and 20-fold selective binders could not be discriminated from each other very well. In contrast, the reverse labeling scheme provided excellent resolution in the upper-left and upper-right quadrants. So, an initial screen of a library should use phycoerythrin as label for the ‘on-target’ probe to enrich binders with acceptable selectivity. In a later stage, Alexa Fluor 488 could be useful as ‘on-target’ probe to better stratify the population according to binder selectivity.

If however the MBD lacked high binding activity on the cell surface either because display levels were not sufficiently high or the displayed payload was inactive, discriminating the more selective binders from the less selective ones became increasingly difficult as illustrated by the example of surface-displayed MeCP2.

Overall, the collected data suggested that the established conditions for the fluorescence-activated DNA binding assay were suitable to identify selective binders with high specificity and sensitivity with potentially nano-molar affinity even if present in low number in a mixed population of binding and non-binding entities.

(7.4.2) Validating the FACS binding assay with surface-displayed wildtype MBD

To better inform the screening of the degenerated MBD libraries, the behavior of surface-displayed wildtype MBD was examined in the one-color and the two-color assay.

One-color assay. The binding of surface-displayed MBDs to DNA probes with different modified CpG dyads was investigated using the optimized FACS DNA binding assay (**Figure 7.11**).

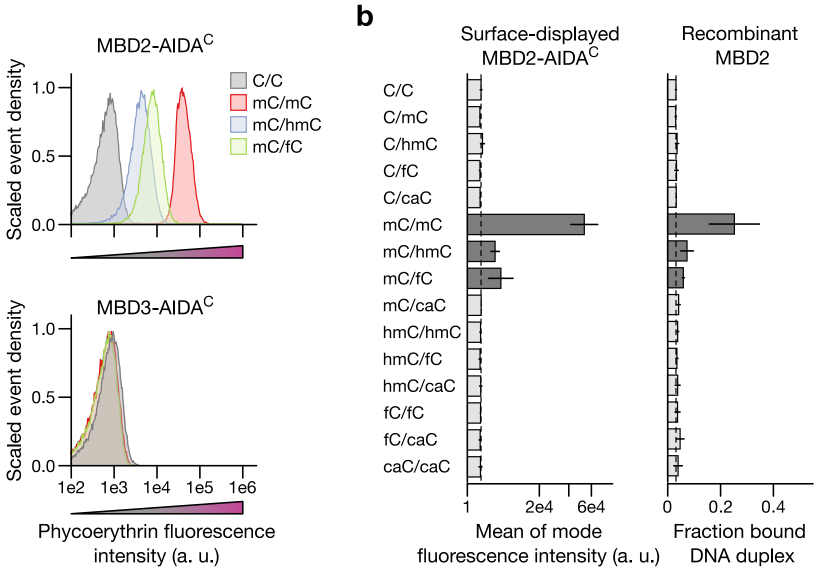


Figure 7.11 One-color FACS binding assay with wildtype MBD. Surface display of MBDs was induced with 50 µM IPTG and display levels confirmed by Anti-c-Myc epitope staining (Figure A.16). **(a)** Fluorescence intensity retained on surface-displayed MBD2 and MBD3 after staining with the selected modified DNA probes. **(b)** Full profile of two biological duplicates for surface-displayed MBD2 (data for MBD3 in Figure A.16) against 15 combinations of modified CpG dyads and comparison with the selectivity profile observed on an electrophoretic mobility shift assay for recombinantly expressed MBD2 at 128 nM.

The binding profile of surface-displayed MBD2 recapitulated the selectivity observed for recombinantly expressed MBD2 at a concentration of 128 nM. Surface-displayed MBD3 expectedly did not bind any of the modified DNA probes (**Figure A.16**). This suggested that the one-color assay could indeed be used like an 'EMSA on FACS' for MBDs that would show similar display levels and binding activity on the cell-surface as MBD2. This condition seemed to be met for binding interactions that were strong enough to prevail at nanomolar concentrations, hence for high affinity binders. If no such interaction would exist in the sequence space investigated, the screening conditions would have likely to be adjusted, e. g., by increasing the concentration of the DNA probe to reveal also less affine interactions.

Two-color assay. A binary competitive staining with two differentially labeled DNA probes for wildtype MBD2 indicated similar trends (**Figure 7.12**). The fully methylated 5mC/5mC probe effectively competed the strongest against all of the tested modified CpG dyads independent

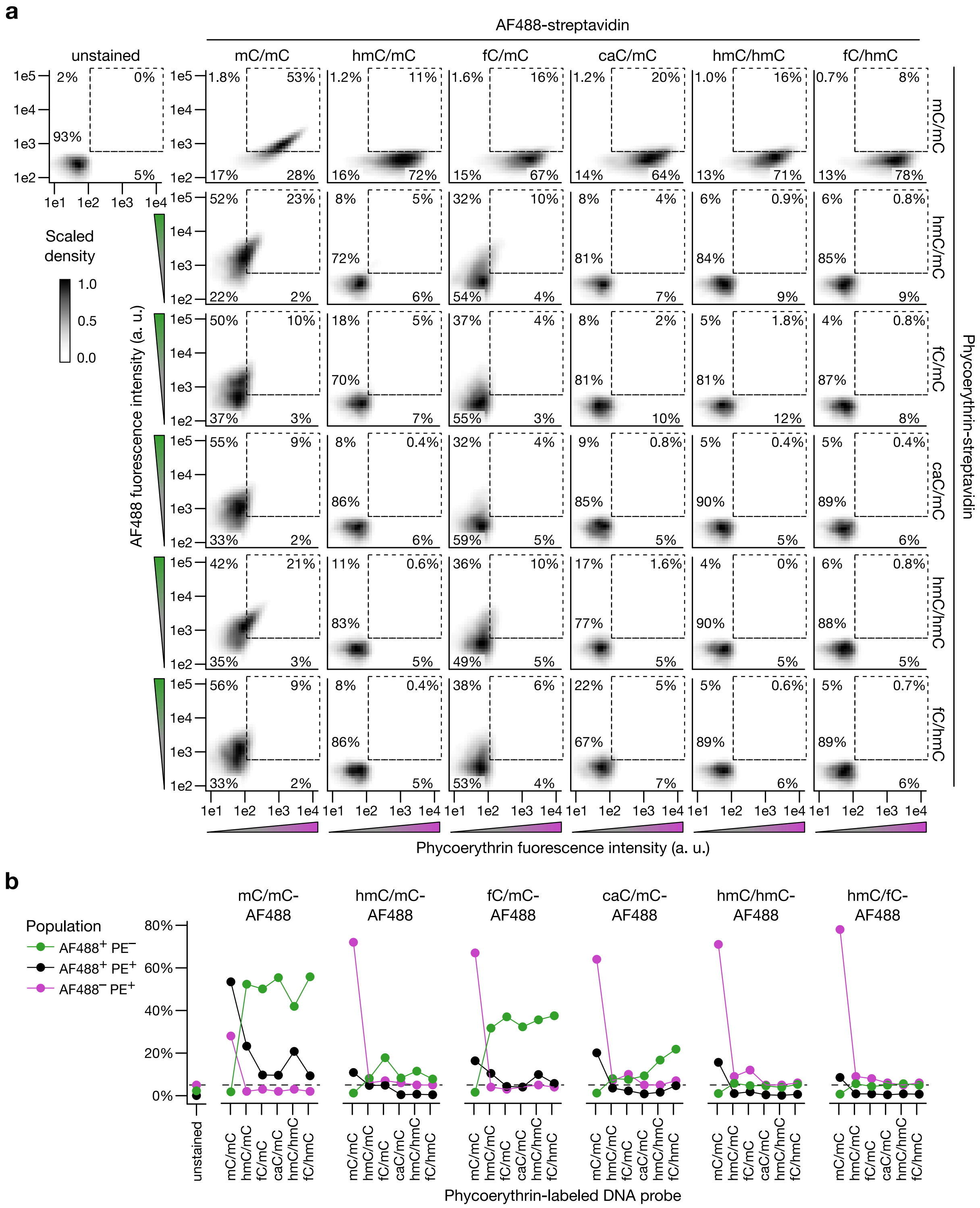


Figure 7.12 Two-color FACS binding assay with wildtype MBD2. (a) Cell surface display was induced with 50 µM IPTG and the displayed MBD2 stained in a binary combination of two DNA probes that had a single modified CpG and were labeled differentially for the analysis on the flow cytometer; Fraction of events per quadrant gate and scaled event density of 50,000 gated events per condition indicated. **(b)** Fraction of events per quadrant gate and DNA probe combination.

of whether Alexa Fluor488 or phycoerythrin were used.

The second most effective competitor was 5fC/5mC, which was visible quite clearly with the brighter phycoerythrin labeling, but hardly detected using Alexa Fluor 488. Although the one-color assay had indicated this trend too, the magnitude in the competitive setting was probably unexpected. Certainly, one must not over-interpret such an observation, but it emphasized that binding selectivity was read out differently in each of the assays. Whereas in the one-color setting, isolated binding reaction converged toward their equilibrium, the competition between the ligands in the two-color assay could be dominated over shorter or longer periods by the kinetic differences in the respective association and dissociation rates of each ligand. Notwithstanding the necessity that such hypotheses would need to be evaluated in a more controlled setup than a protein cell surface display assay, it seemed evident that screening in a competitive setup would be ultimately relevant to any future application of the MBD binder that involves the retrieval of modified DNA from complex mixtures.

7.5 Full library screening for MBDs with novel binding selectivity

A two-color screening of all five degenerated MBD libraries against 5hmC/C, 5fC/C, 5caC/C and 5hmC/5mC, 5fC/5mC, 5caC/5mC in which the given 'on-target' DNA duplex with the desired modified CpG dyad was labeled with the bright red phycoerythrin fluorophore and all other DNA probes with the fluorescein-like green Alexa Fluor 488 (not shown) indicated that the degenerated MeCP2 library had a particularly high potential to yield MBD variants with altered binding selectivity. Indeed, an exhaustive screening with this library against 15 different modified CpG dyads (**Figure 7.13**, controls in **Supplementary Figure A.17**) showed a particularly high enrichment for the 5mC/5mC, 5hmC/5mC, 5caC/5mC and 5caC/5caC dyads. For other modified CpG dyads, no binders with sufficient surface display level or sufficient 'on-target' affinity could be detected or only some binders retrieved that appeared promiscuous to a sizable degree.

The post-screening analyses presented in the following therefore will focus on MBD subpopulations that appeared to be selective binders of 5hmC- or 5caC-containing CpG dyads.

(7.5.1) Assessing global phenotype enrichments in the MBD of MeCP2

To assess the complexity in the enriched subpopulations from the screening of the degenerated MeCP2 library, the underlying genotypes were determined by next-generation sequencing and the fold-enrichment with respect to the initial library was calculated for each phenotype (**Figure 7.14**). Since the initial library had been sequenced shallowly (Section 7.2), it was assumed

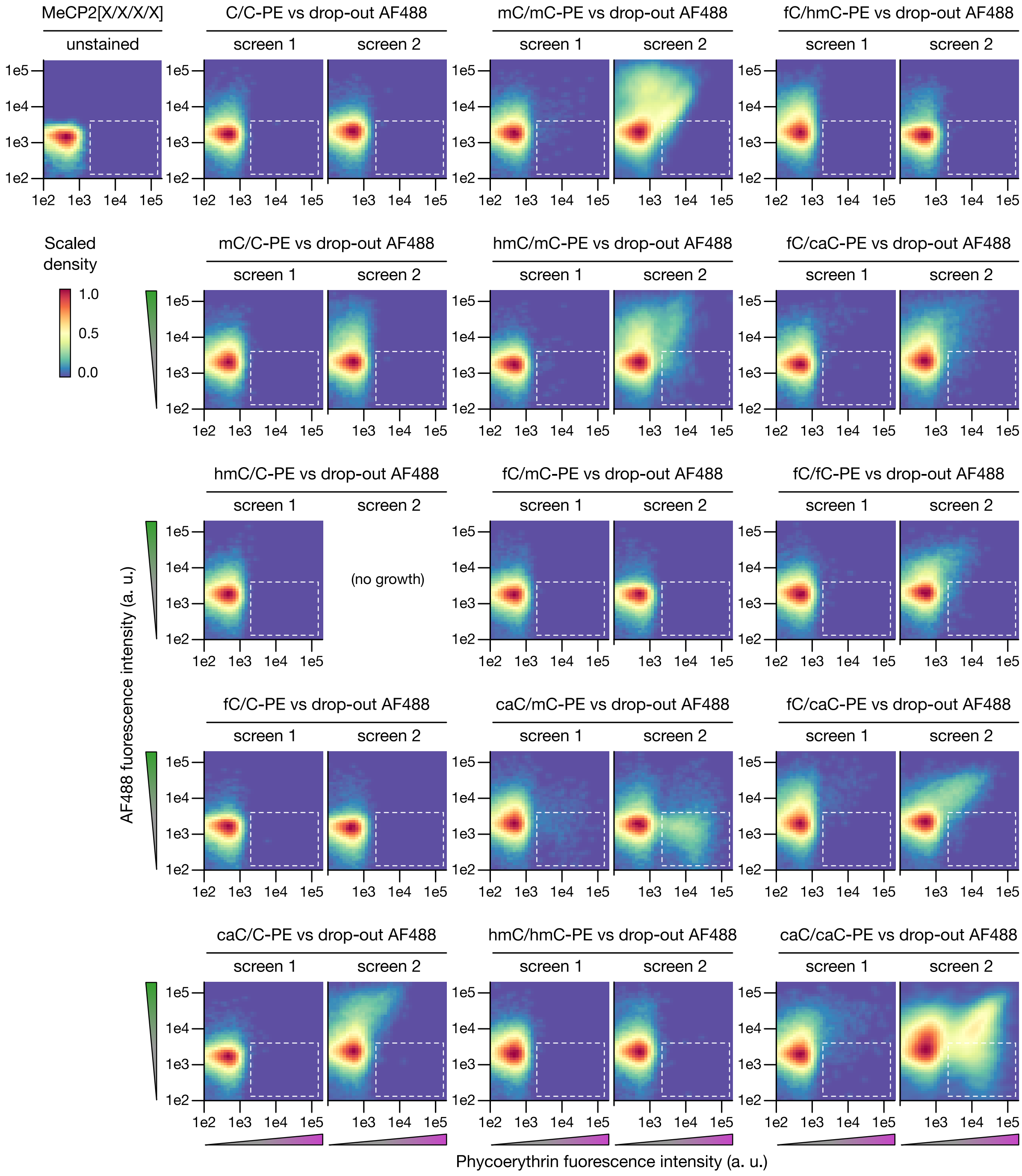


Figure 7.13 Screening the degenerated MeCP2 library against 15 modified CpG dyads. Two-color screen after the first and after the second screening round with the ‘on-target’ DNA probe labeled with phycoerythrin and the other DNA probes labeled with Alexa Fluor (AF) 488. Each probe, independent of its label, had the same molar concentration.

that genotypes that were enriched but not detected in the sequencing of the library were 'just missed by one read' which was a worst-case assumption.

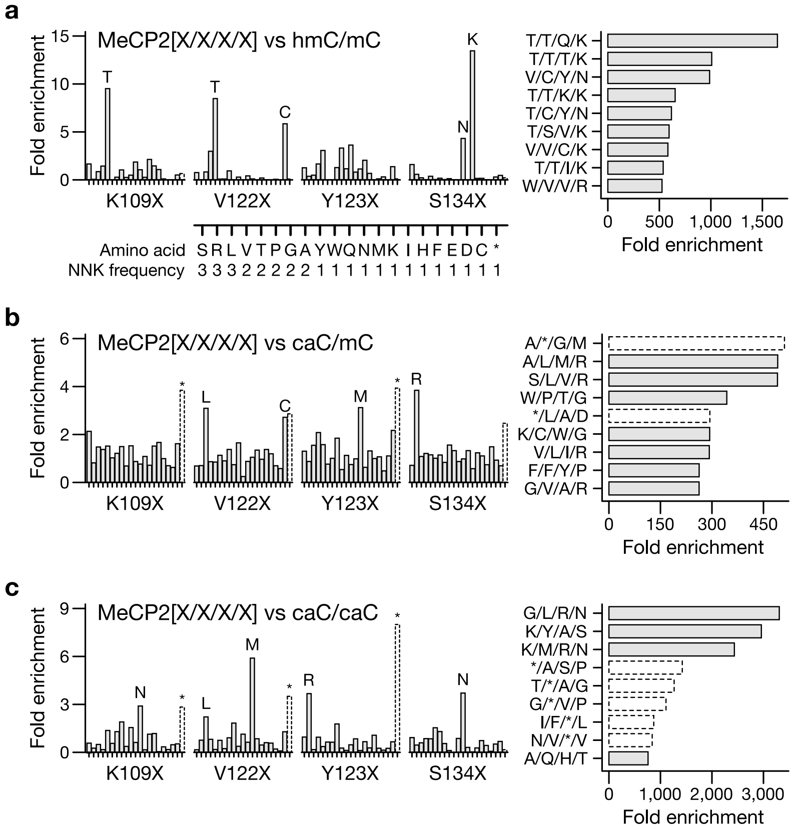


Figure 7.14 Post-screening phenotype enrichment. Subpopulations retrieved after two consecutive rounds of screening and enrichment of the MeCP2 NNK codon-degenerated library were analyzed by next-generation sequencing and the relative prevalence of amino acid substitutions per degenerated position and full-length phenotypes compared to the original library. **(a)** Screening for 5hmC/5mC-selective, **(b)** for 5caC/5mC-selective and **(c)** for 5caC/5caC-selective MBD mutants. Source code available.

After two consecutive screening and purification rounds for 5hmC/5mC, the first two degenerated sites in the MeCP2 sequence, lysine-109 and valine-122, were substituted predominantly to threonine. Substitutions of tyrosine-123 seemed to be at random. Serine-134 was typically replaced with lysine or asparagine. K109T/V122T/Y123Q/S134K (T/T/Q/K for short) and T/T/T/K were the strongest enriched phenotypes in this screen. T/T/T/K and T/T/I/K followed the same pattern. A second substitution pattern seemed to be K109X/V122C/S134N with examples such as V/C/Y/N and T/C/Y/N. Intriguingly, the latter substitution pattern was also found in an alternative screening strategy (**Supplementary Figure A.18**).

For 5caC/5mC the fold-enrichment was not as strong as in the previously analyzed subpopulation, maybe because of two dominant nonsense mutants. They could have been carried along the pre-purification procedure as parasites that easily overgrew one of the selected subpopulations. Since they harbor a plasmid conferring resistance to the selective growth medium, but carry no burden of expressing a protein, they can be hard to remove from the bulk. Nonetheless, a common substitution pattern for accommodating a 5caC interaction could be recognized as K109X/V122L/Y123X/S134R (confirmed in absence of nonsense mutants). The most prevalent phenotypes were A/L/M/R, V/L/I/R and S/L/V/R.

The 5caC/5caC screen recovered three MBD variants that dominated the enriched subpopulation: G/L/R/N, K/Y/A/S and K/M/R/N with K109X/V122X/Y123R/S134N as the presumed substitution profile to accommodate two 5caC interactions in MeCP2.

From the ensemble of enriched phenotypes, this analysis seemed to reveal meaningful instructions for at least some of the substitutions required in a MBD wildtype sequence to bring about new interactions with certain oxidized CpGs. These predictions were assessed on the level of the individual phenotypes.

(7.5.2) Characterizing mutant MBD candidates on the surface display platform

The binding selectivity of isolated single clones and predicted candidates from the NGS analysis was assessed on the 'one-color' DNA binding assay for a panel of relevant CpG dyads.

Three randomly picked clones after the 5caC/5mC screen of the MeCP2 library shared the substitution of serine-134 to arginine. Although none of the picked clones was among the top ten of the enriched phenotypes in the NGS analysis, the common serine-to-arginine substitution was also the predominant feature in the prediction. Reassuringly, the three clones bound 5caC/5mC CpGs the strongest in the surface display DNA binding assay, followed by 5caC/5caC CpGs and the canonical 5mC/5mC CpG modification.

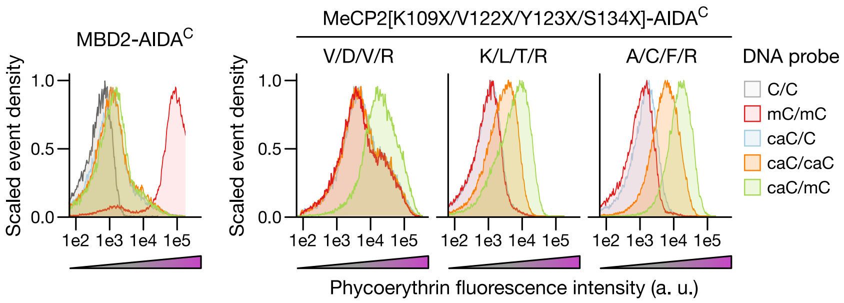


Figure 7.15 Single-color assay of candidate 5caC/5mC CpG-selective MBDs. Surface-displayed MBD2 as staining control with candidates from the NNK codon-degenerated MeCP2 library. Surface-display was confirmed by staining with an allophycocyanin-coupled Anti-c-Myc epitope antibody; All cells induced at 50 µM IPTG, 10,000 events shown.

Two candidates form the 5hmC/5mC screen that followed the K109T/V122T/Y123X/S134K substitution pattern showed a higher affinity towards the 'on-target' probe 5hmC/5mC than an MBD wildtype, but bound 5mC/5mC equally well in this assay. Although one of the picked clones, S/I/T/N, almost conformed the second substitution pattern K109X/V122X/S134N, no DNA binding was observed. The two other MeCP2 mutants, T/C/Y/N and T/A/Y/N showed the expected increase in affinity towards 5hmC/5mC CpGs at a decreased affinity for 5mC/5mC, hence reversing the MBD's natural binding selectivity.

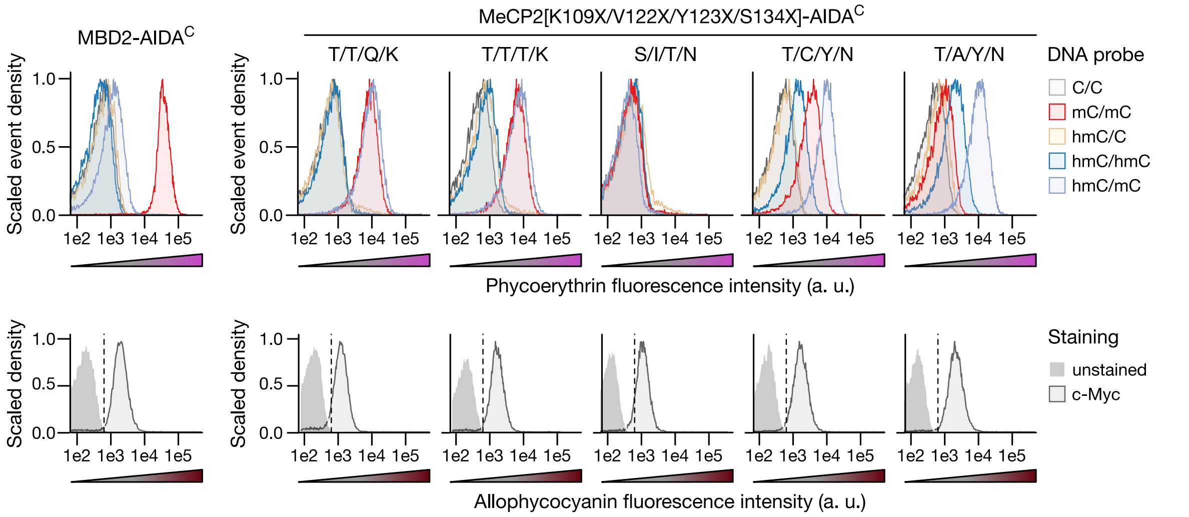


Figure 7.16 Single-color assay of candidate 5hmC/5mC CpG-selective MBDs. Surface-displayed MBD2 as staining control with candidates from the NNK codon-degenerated MeCP2 library. Surface-display was confirmed by staining with an allophycocyanin-coupled Anti-c-Myc epitope antibody; All cells induced at 50 µM IPTG, 10,000 events shown.

For the MeCP2[K109T/V122A/S134N] MBD, 5hmC/5mC was confirmed as the modified CpG dyad that was bound the strongest out of a panel of 15 different CpG dyads with different modified cytosines against which the initial library had been screened (**Figure 7.17**).

Overall, these validations indicated that the established surface display of MBDs and the conditions of the two-color DNA binding assay could identify a number of candidates from the MeCP2 library with the desired binding properties.

(7.5.3) Tweaking MeCP2[K109T/V122A/S134N] performance

To test if some of the binding properties could still be improved, for example selectivity of the MeCP2[K109T/V122A/S134N] mutant (T/A/Y/N hereafter), a small randomized library MeCP2[K109X/V122X/S134N] was generated. In contrast to the original screen with equimolar mixtures of the DNA probes, a fourfold excess of 5mC/5mC over 5hmC/5mC was chosen using the a fluorescein-like fluorophore as 'on-target' probe according because of the probably higher sensitivity of this combination to discriminate binding selectivity (page 81). However, no enrichment was observed (data not shown), suggesting that the substitution of these positions would not allow for a higher selectivity.

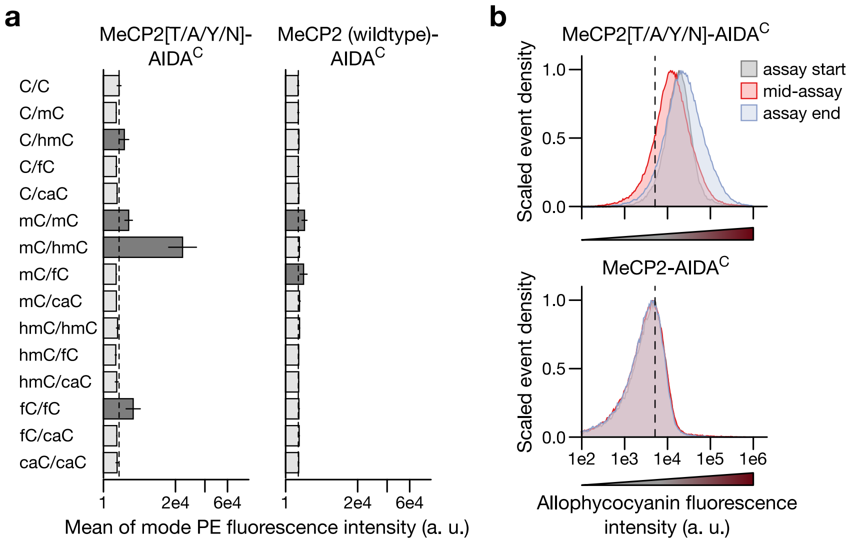


Figure 7.17 One-color FACS binding assay with MeCP2[K109T/V122A/S134N]. Surface display of wildtype and mutant MeCP2 was induced with 50 µM IPTG. **(a)** Fluorescence intensity retained on surface-displayed MBDs after staining with the selected modified DNA probes; averages of modes from three independent experiments. **(b)** Surface-display level over the course of the one-color fluorescence-activated DNA binding assay.

In the work of Katrin Bigler (Bigler, 2020), T/A/Y/N and similar mutants were subjected to error-prone PCR in several rounds to identify additional sites in the sequence that could afford a more selective MBD. However, besides some modest improvement by the inactivation of aspartic acid-90 to glycine, none of the candidates could further increase the selectivity so far.

7.6 Synopsis

Technologies to imitate the processes of variation and selection akin to natural evolution enable the creation of novel function in biological macromolecules. To probe the functional consequence of the alterations in the sequence space of the methyl-CpG-binding domain (MBD), the work presented in this chapter has resulted in the following platforms:

1. An efficient isothermal library assembly protocol to simultaneously degenerate multiple key residues in the DNA–protein interface of the MBD.
2. The inducible cell surface display of human MBDs in a functional form on a bacterial host.
3. A DNA binding assay conducted on a the surface of a single bacterium with multiple fluorescently-labeled DNA probes under thermodynamically controlled conditions to measure and screen MBD binding selectivity for modified CpG dyads on a flow cytometer.

This has enabled to find:

4. The primary sequence of MeCP2 can be substituted such that the domain selectively or promiscuously binds CpG dyads with different combinations of cytosine modifications.
5. Based on bulk sequencing and the characterization of individual mutants, the most critical substitutions involved Ser-134 and Val-122.
6. MeCP2[K109T/V122A/S134N]-AIDA^{C} is a surface-displayed MBD which specifically binds 5hmC/5mC CpG dyads.

Endnotes

[a] In addition to the 20 canonical amino acids, NNK also encodes the amber stop codon TAG.

[b] The assembly of oligonucleotides shorter than 200 bp in Gibson-like methods is inherently difficult (Birla & Chou, 2015). In the modification of the manufacturer's protocol presented here, the reaction temperature is lowered to 45 °C which probably reduces the activity of the 5'→3' exonuclease and slows down the erosion of the oligonucleotides.

[c] It may be doubted that the alternative type IIS strategy could have yielded a more uniform library: As the PCR linearization for type IIS cloning strategy uses the degenerated oligonucleotides as primers, the cycling introduces and amplifies bias in the clonal representation of the individual genotypes similar to classical saturation mutagenesis where 'complementary' degenerated primer pairs are used (Acevedo-Rocha, et al., 2015). In contrast, preparing the backbone for Gibson assembly via PCR yields an essentially homogenous product, depleting the wildtype genotype.

[d] Of course, the higher bacterial growth rate is another advantage over yeast-based surface display systems in terms of screening throughput. In addition, many laboratories are well experienced and equipped to culture the bacterium *Escherichia coli* whereas the culture of the yeast *Saccharomyces cerevisiae* requires some additional provisions.

[e] The *E. coli* B strain Tuner™(DE3) lacks a functional lactose symporter *lacY*, breaking the positive feedback loop that would lead to intracellular accumulation of lactose or its analogs. Thus, the amount of protein produced under *lacO* control increases proportional with the concentration of the lactose analog IPTG. (Turner, et al., 2005)

[f] Titration of expression levels with L-arabinose requires an *araBADC*⁻ *araEFGH*$^{+}$ genotype which lets the cells take up L-arabinose but not metabolize it such as K-12 DH10B; B strains are *araBADC*$^{+}$.

Chapter 8

An artificial MBD with novel CpG binding selectivity

The screening of the degenerated MeCP2 MBD library has produced several candidates with putative 5hmC/5mC binding selectivity. The biochemical (Section 8.1) and structural characterization (Section 8.2) of the recombinantly expressed proteins is presented in this chapter.

Some important insights about the putative molecular recognition could not have been made without the know-how and support of our cooperation partners; their contribution to the content of this chapter is clearly indicated.

8.1 MeCP2[K109T/V122A/S134N] is an artificial 5hmC/5mC-selective MBD

A survey of MBD selectivity at low nanomolar concentration of the recombinantly expressed proteins (**Figure 8.1**) confirmed in good agreement with the observations in the one-color fluorescence-activated DNA binding assay on FACS (Chapter 7.5.2) that within the group of 5hmC/5mC candidates those binders that followed the K109T/V122T/Y123X/S134K substitution scheme did not preferentially bind 5hmC/5mC CpGs whereas K109T/V122X/S134N preferred those sites over 5mC/5mC CpGs.

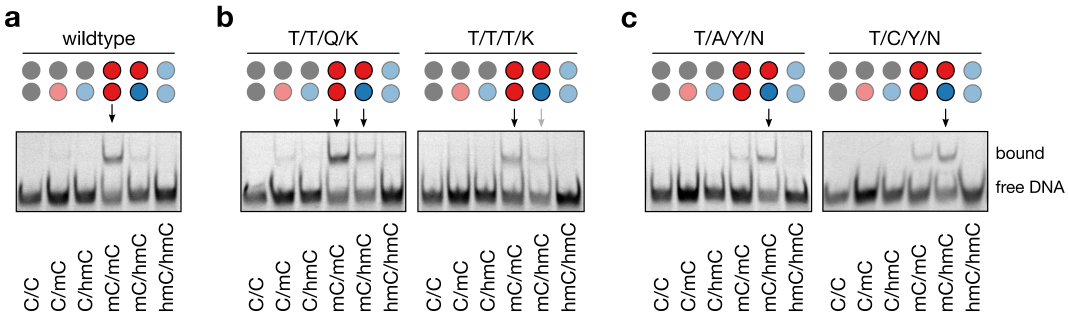


Figure 8.1 Electrophoretic mobility shift assay of MeCP2-derived 5hmC/5mC binders. (a) Wildtype MeCP2, **(b)** Mutants of the K109T/V122T/Y123X/S134K group and **(c)** of the K109T/V122X/S134N group. Assays at 10 nM protein concentration.

The biochemical characterization of the interactions in terms of their apparent dissociation constant K_d using the electrophoretic mobility shift assay (EMSA) framework (Chapter 6) revealed a tenfold higher affinity of MeCP2[K109T/V122A/S134N] (T/A/Y/N hereafter) for 5hmC/5mC (8 ± 2 nM) than for 5mC/5mC (80 ± 15 nM; **Figure 8.2a**). This was, quite unexpectedly, the reversed binding selectivity of wildtype MeCP2 (47 ± 12 nM for 5hmC/5mC, 4 ± 1 nM for 5mC/5mC; not shown[a]).

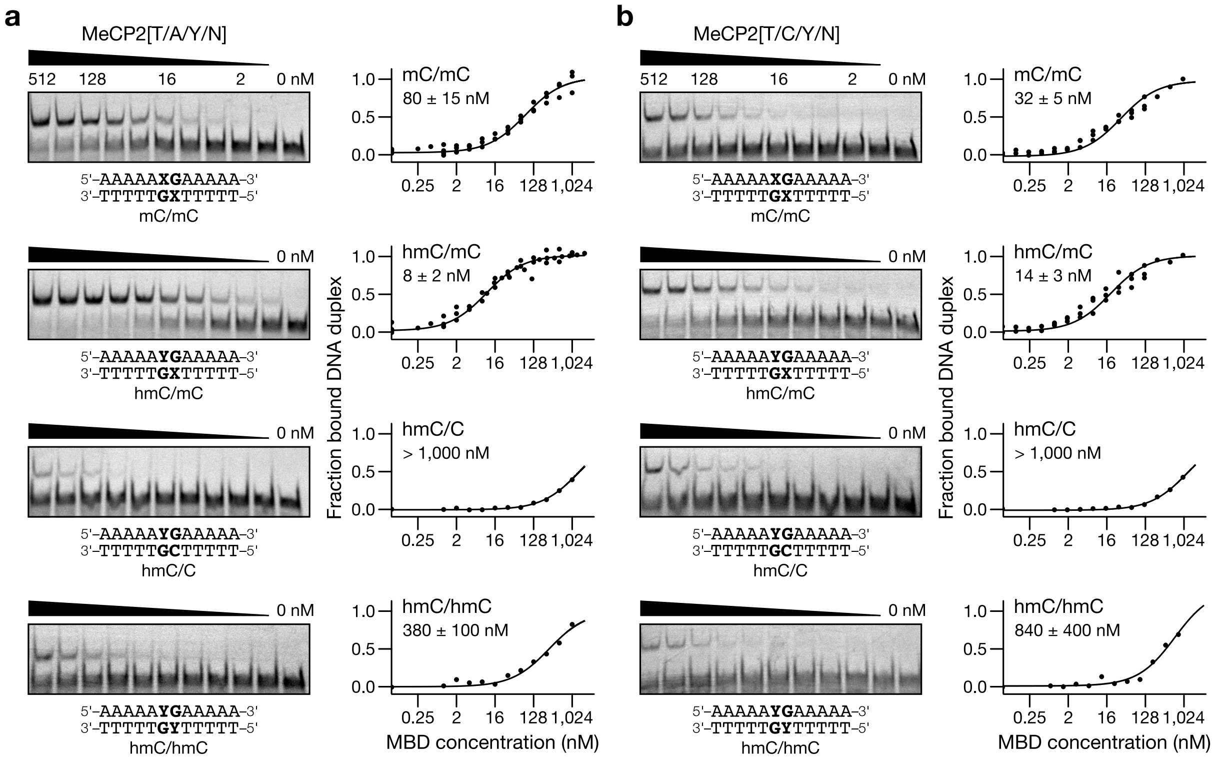


Figure 8.2 Binding affinity of T/A/Y/N and T/C/Y/N towards modified CpG dyads. The apparent dissociation constants K_d were determined using electrophoretic mobility shift assays with artificial 24-mer DNA probes that contained a central CpG with the indicated modified cytosine combinations for **(a)** MeCP2[K109T/V122A/S134N] (T/A/Y/N) and **(b)** T/C/Y/N.

The T/C/Y/N mutant had a twofold binding selectivity of 5hmC/5mC over 5mC/5mC (**Figure 8.2b**) with nanomolar binding affinity again successfully cross-validating the conditions of the screening campaign and the results obtained from the one-color FACS binding assay. In contrast to wildtype MeCP2 neither T/A/Y/N nor T/C/Y/N showed a propensity to interact with unmodified CpA dyads and the interaction with methylated 5mCpA dyads was markedly decreased (**Supplementary Figure A.19**).

A comprehensive characterization of the fraction of bound DNA duplex selectivity of T/A/Y/N with all 25 combinations of modified and unmodified cytosines in the CpG dyad, confirmed that 5mC/5mC and 5hmC/5mC were the two combinations that were bound the strongest (**Figure 8.3**). Also 5caC/5caC was bound albeit weakly. This was in line with T/A/Y/N partially fulfilling the substitution profile found of the 5caC/5caC screen (compare Figure 7.14).

Rather unexpected, whereas 5hmC/5mC was bound the strongest by T/A/Y/N, the identical combination of modified cytosines appearing in the reversed sequence context 5mC/5hmC was not bound as strongly. Indeed, 5hmC/5mC was the context of the probe in which the mutant was identified during the screen. A similar preference was observable for wildtype MeCP2 while MBD2 did not discriminate between 5hmC/5mC and 5mC/5hmC. One could speculate that an additional, maybe sequence-specific interaction, could lock MeCP2 in a preferred ori-

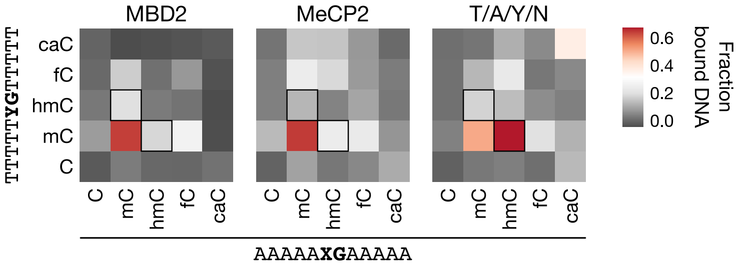


Figure 8.3 Full binding selectivity of MeCP2[K109T/V122A/S134N]. Fraction of bound DNA duplexes containing a single CpG dyad in the context of the indicated strands for wildtype MBD2, MeCP2 and MeCP2[K109T/V122A/S134N] at 53 nM; 5mC/5hmC and 5hmC/5mC dyads framed.

entation on the DNA double-strand or a that the artificial oligo(A) context in which the 5hmC-modified CpG appeared caused a (compare Section 3.1).

Indeed, MeCP2 has an additional interface with one of the DNA strands around the Asx-ST motif with Thr-158 (compare Figure 5.3). However, a T158M mutation did not alleviate directional binding. The deletion of the entire $\Delta(^{158}\text{TVTG}^{161})$ or of the preceding $\Delta(^{150}\text{LDPND}^{154})$ which is only present in MBD4 and MeCP2, but not in the other MBDs (compare Figure 5.2), abrogated DNA binding entirely (data not shown). Whether alleviation of the observed binding preference is desired at all or not, likely depends on the intended application of the protein and whether or not comparisons to the wildtype are desired.

(8.1.1) Binding selectivity in genomic sequence contexts

The binding affinity of MeCP2[K109T/V122A/S134N] for 5mC/5mC- and 5hmC/5mC-modified CpG dyads was assessed in more complex sequence contexts derived from genomic regions in presence of a competitor with low sequence complexity to reveal additional interactions of the MBD with the DNA duplex. This experimental setup which can generate multiple shifts on the EMSA requires specific computational analysis to discriminate the contributions of each binding site (implemented in Chapter 9). Although the shifts observed in the following often argued for third or fourth order interactions, i. e., the MBD interacted with two or three additional sites on the duplex in addition to the modified CpG dyad, only a second order polynomial was evaluated since not enough data points could be collected at higher protein concentrations. It was further assumed that any additional binding interactions on a duplex would occur with the same affinity independent of the modification state allowing to evaluate the model simultaneously for both combinations of cytosine modifications.

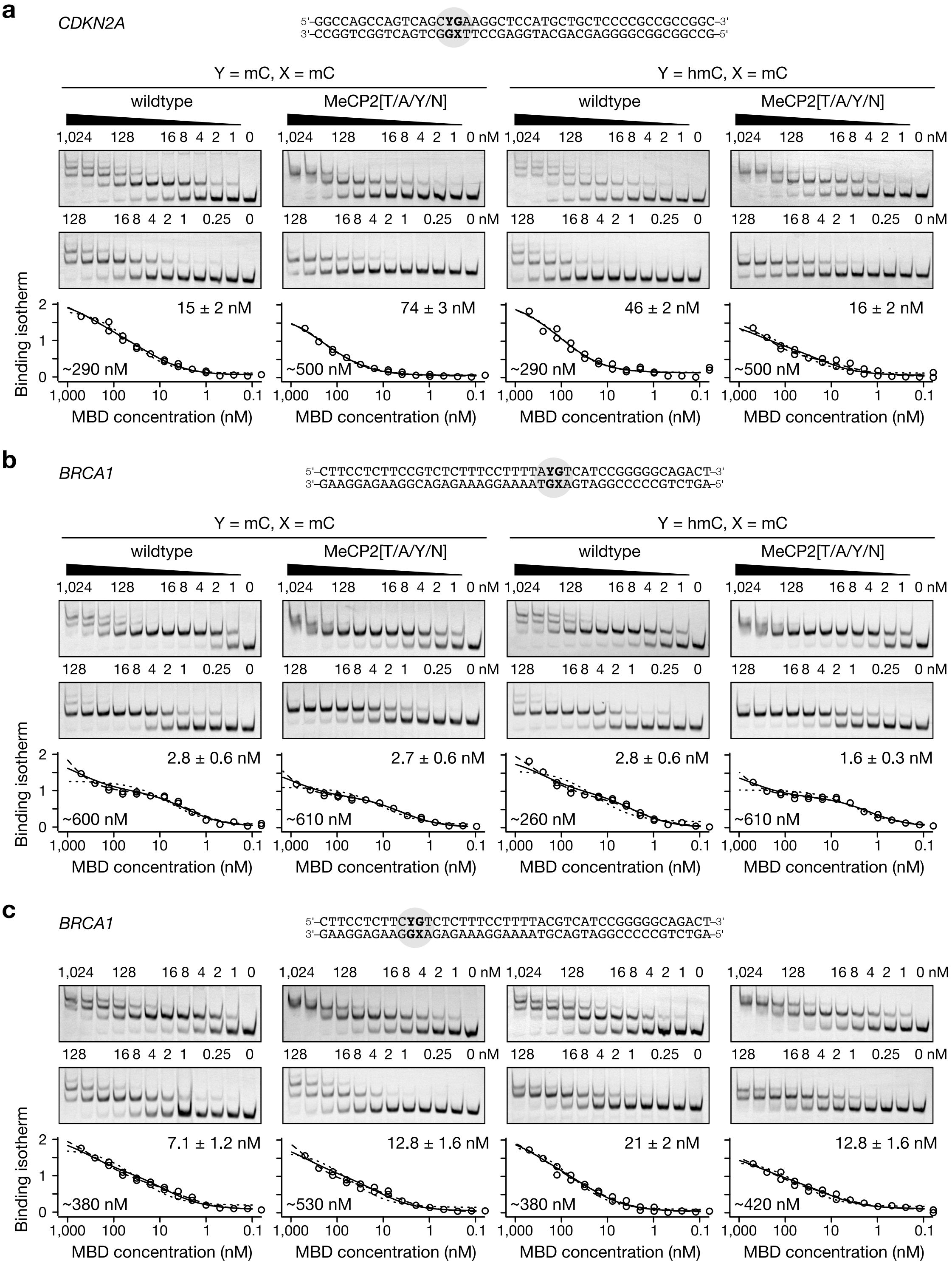


Figure 8.4 Binding affinity of T/A/Y/N towards modified CpG dyads in two gene promoters. Electrophoretic mobility shift assays with fluorescently-labeled probes in presence of a homopolymeric oligo(A) · oligo(T) competitor and the multiple shifts assessed as described in Chapter 9. First (dotted), second (solid; estimates reported) and third (dashed) order polynomial fits, macroscopic binding constants, one shared estimate (light-faced). **(a)** With a single modified CpG in the sequence context of the first exon of *CDKN2A* (chr9:21,974,777–822; hg38). **(b, c)** With one out of three CpGs present in the sequence of the bidirectional promoter of *BRCA1* and *NBR2* (chr17:43,125,546–590; hg38). Source code available.

For a single modified CpG dyad in the sequence context of *CDKN2A*, the wildtype MeCP2 showed an about threefold selectivity of binding 5mC/5mC over 5hmC/5mC and the T/A/Y/N mutant a reversed fivefold selectivity (**Figure 8.4a**; **Supplementary Table A.9**). The two preferred interactions had an apparent dissociation constant of about 15 nM. The predicted additional interaction was about twofold weaker for the T/A/Y/N mutant than for the wildtype as evaluated in the binding model.

Next, the differences in binding selectivity was assessed if one of several CpG dyads in a DNA double-strand was modified while the others remained unmodified. A sequence from the human *BRCA1/NBR2* locus that contained three CpG dyads served as model (**Figure 8.4b–c**). If the CpG dyad in *b* was modified, all interactions had low nanomolar dissociation constants. While the mutant had an about twofold preference for the 5hmC/5mC dyad, there was no measurable preference for wildtype since both binding isotherms were almost indistinguishable. As before, the wildtype but not the T/A/Y/N mutant showed a number of additional, defined interactions with the DNA duplex. If the CpG dyad in *c* was modified, the wildtype showed a threefold binding preference for the fully methylated dyad over the 5hmC/5mC dyad similar to *CDKN2A*. However, no difference in the binding isotherms could be detected for the mutant. A potential reason could be that in contrast to all previous examples, the second binding site on the DNA was occupied almost immediately after the first one, hence violating some important assumptions to disentangle the individual contributions of each binding site to the macroscopic isotherm in the underlying binding model (see Chapter 9). In other words, as soon as equally appropriate solutions would fit the observed data, the correct one cannot be determined. An additional biochemical reason could be that the second determinant in the MeCP2 scaffold did not align with the determinant of the engineered binding site, i. e., the orientation of the 5hmC/5mC dyad in the duplex did not align with a secondary property of the duplex, e. g., a sequence motif, to selectively engage which could be why the unspecific interaction dominated. In any case, both examples illustrate that sequences exist in which differential modification of a CpG dyad can be difficult to discriminate, at least in a non-competitive setting.

The impact of additional sequence context is further illustrated by a single modified CpG dyad presented in an intronic sequence of the *Hey2* ortholog in the zebrafish *Danio rerio* on the one hand and in a limited double-stranded sequence context at the same position on the other hand (**Figure 8.5a–b**). Whereas the individual microscopic binding affinities of wildtype and T/A/Y/N mutant are obscured in the fully double-stranded duplex, i. e., they cannot be determined on the basis of the macroscopic binding isotherm (see above), a reduction to the 12-mer context reveals a twofold preference of the wildtype and an almost tenfold, but reversed, preference of 5hmC/5mC over 5mC/5mC for the T/A/Y/N mutant.

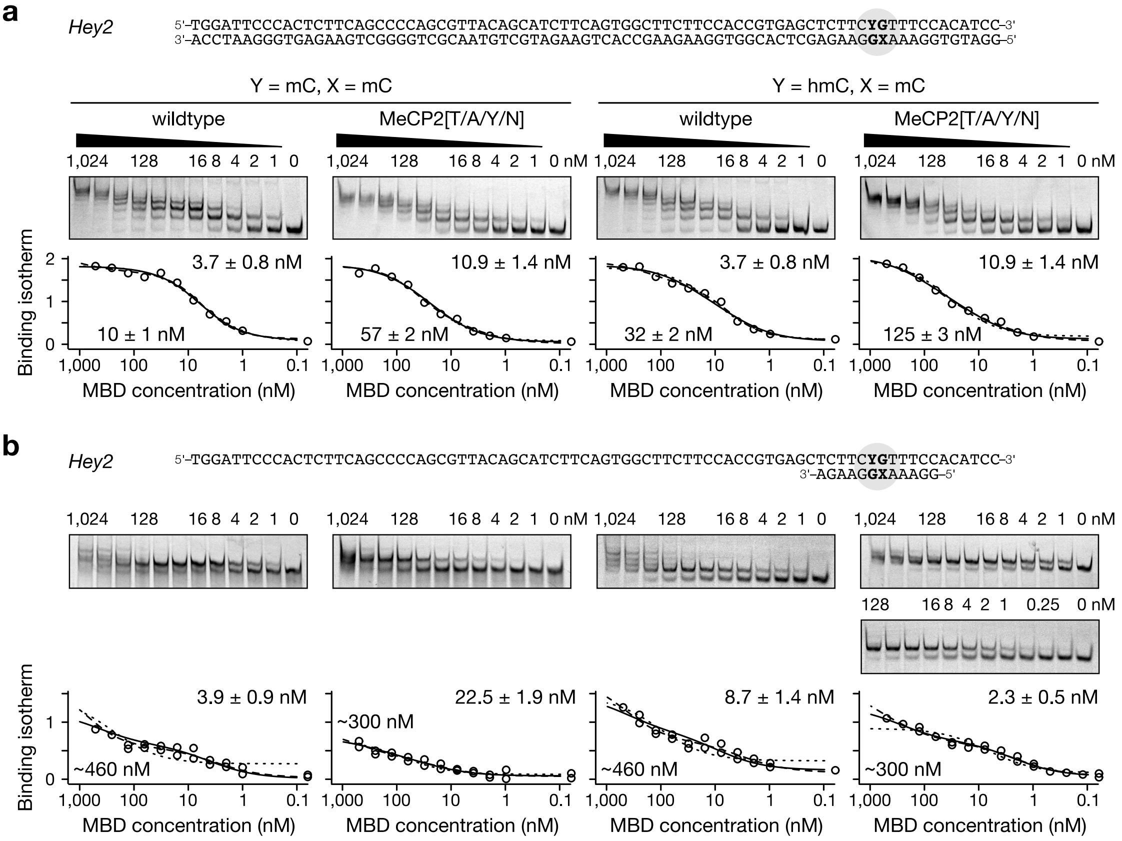


Figure 8.5 Dependency of binding affinity of T/A/Y/N towards modified CpG dyads. Same experimental conditions as in Figure 8.4. With a single modified CpG in the fully double-stranded sequence of a *Hey2* ortholog in the zebrafish *Danio rerio* (chr:20:39,589,641-719; danRer11) using **(a)** the full complement or **(b)** a short complement. Source code available.

Of course, the stringency under which the EMSAs were conducted was, within limits, an arbitrary choice. However, whether or not the conditions can be optimized for a particular sequence, it seems unlikely to satisfy all possible sequence contexts or duplex lengths at once.

In summary, the evaluation of wildtype MeCP2 and the MeCP2[K109T/V122A/S134N] mutant suggested that complex mixtures of DNA sequences can pose a challenge for the detection of their binding selectivity for a single modified CpG dyad. The absence of evidence however is not evidence for absence. While in some contexts binding selectivity may be absent, in other cases it may be more challenging to reveal, and in other cases readily demonstrated.

8.2 Structural characterization

To shed light onto the structural principles that underlie the recognition of 5hmC/5mC CpG dyads by the MeCP2[K109T/V122A/S134N] mutant, mutational studies were carried out and several spectroscopic measurements were initiated to resolve the engagement at different levels.

(8.2.1) Mutational analyses

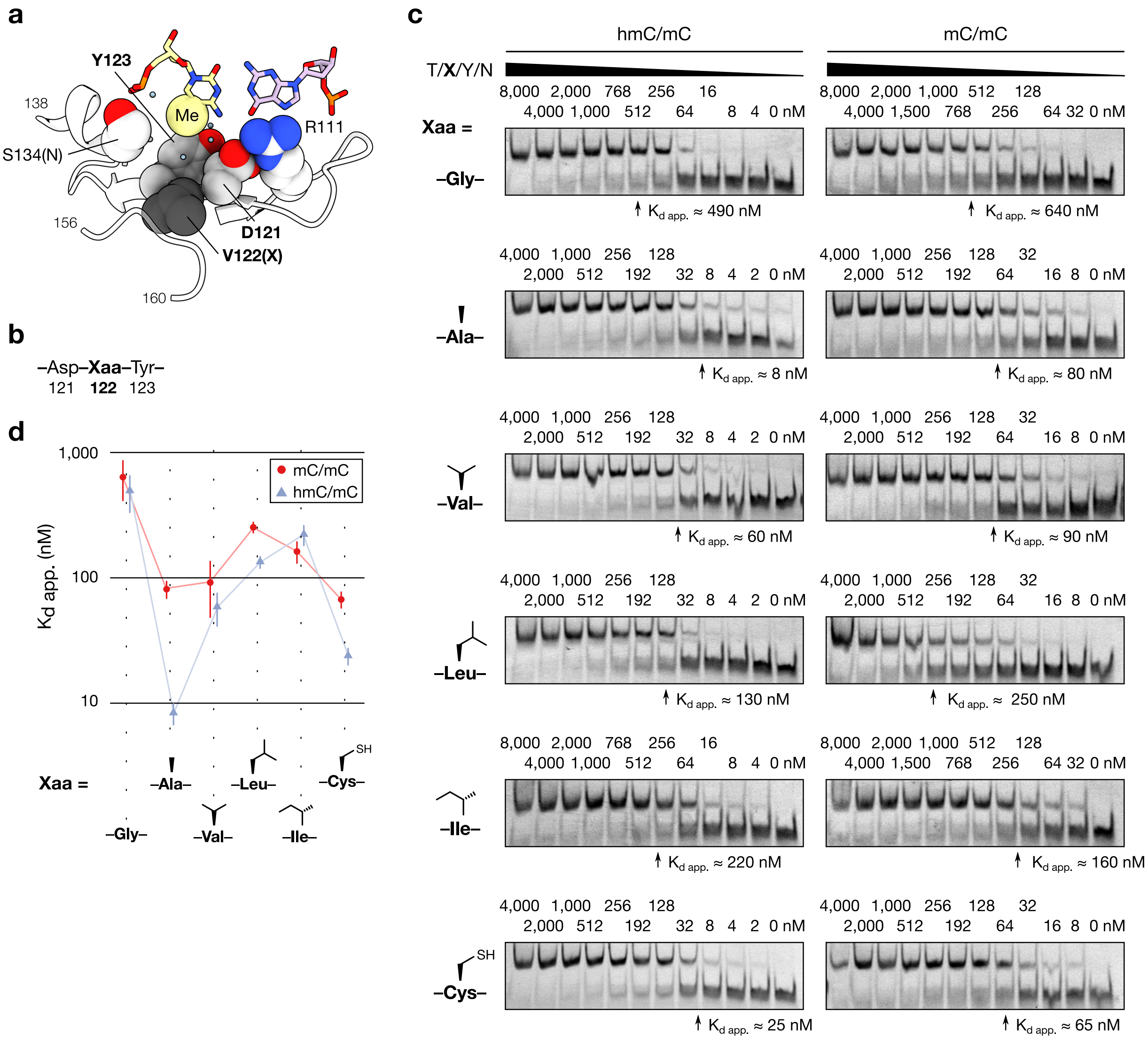


Figure 8.6 Mutational analysis on Ala-122. (a) Position of the cognate wildtype Val-122 in a crystal structure of MeCP2 (PDB 3c2i; Ho, et al., 2008) with important residues shown as van der Waals spheres, including the methyl group of one of the modified cytosines in the CpG dyad. **(b)** Sequence context around position 122. **(c)** Representative electrophoretic mobility shift assays using DNA duplexes with 5hmC/5mC or 5mC/5mC CpG dyads. **(d)** Apparent dissociation constants determined from the EMSAs of *c*.

While the substituted residues Thr-109 and Asn-134 in the MeCP2[K109T/V122A/S134N] mutant faced the DNA-binding interface of the MBD, the role of the substituted Ala-122 was less clear as this residue lies on the other face of β3 and engages with the ST motif ^{158}TVTG^{161} of the Asx-ST motif in MeCP2. At the same time, it is placed between two critical residues for the recognition of fully methylated CpG dyads in the wildtype protein, Asp-121 and Tyr-123 (**a-b8.6**). The K109T/V122C/S134N mutant retrieved from the screen had a cysteine residue at this position and showed reduced binding selectivity. To understand whether indeed an alanine was required at this position or, e. g., the wildtype valine was sufficient, a mutational screening with other aliphatic amino acids was carried out (**Figure 8.6c**).

Only the Ala-122 variant had tenfold binding selectivity followed by twofold selectivity of the Cys-122 variant (**Figure 8.6d**; **Supplementary Table A.10**). Neither the wildtype Val-122 nor smaller (Gly-122) or bulkier substituents (leucine, isoleucine) showed such significant specificity for 5hmC/5mC over 5mC/5mC CpG dyads. Possibly Ala-22 played a role for positioning other parts of the domain in an appropriate geometry to selectively engage with the 5hmC modification.

The cognate MBD2[V164T/V177A/S189N] was unable to bind DNA (data not shown), suggesting that additional sequence features of the MeCP2 domain were also necessary for a productive binding interaction.

(8.2.2) Spectroscopic analyses

Structural alterations upon DNA binding. Specific alterations in the secondary structure of MeCP2 are concomitant with binding of fully methylated 5mC/5mC dyads, but not for the MBD–DNA complex with unmethylated DNA; in particular an increase in the α-helical fraction as revealed by circular dichroism (CD) spectroscopy (Ghosh, et al., 2008). Indeed, wildtype and T/A/Y/N mutant MeCP2 showed the expected changes in the spectra around 195 nm (**Figure 8.7a**) in presence of a DNA duplex that contained the 12-mer sequence context of a *BRCA1* promoter CpG (the one of Figure 8.4b). However, while the wildtype showed such changes but moderately and also seemed to form a complex with an unmodified DNA duplex, the T/A/Y/N mutant spectra only peaked with a duplex containing a single 5mC/5mC or 5hmC/5mC CpG dyad. The fact that both modified duplexes seemed to evoke similar alterations in both proteins whether or no 5mC/5mC or 5hmC/5mC dyads were present could be a consequence of the high protein and DNA concentrations required for the CD measurement.

In fact, both proteins assumed a more defined, globular structure when a modified CpG dyad was present in this duplex based on comparison with other structured and unstructured proteins (**Figure 8.7b**) with an increase in the α-helical fraction (**Figure 8.7c**) which seemed to be more drastic in the T/A/Y/N mutant than in the wildtype.

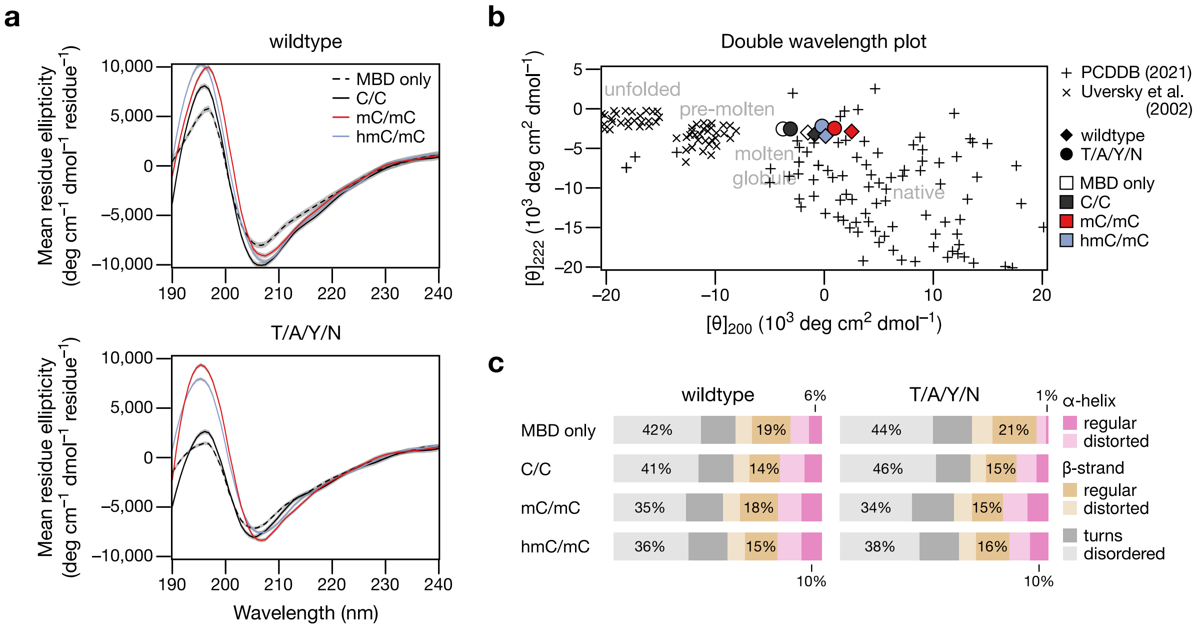


Figure 8.7 Circular dichroism spectroscopy of wildtype MeCP2 and MeCP2[K109T/V122A/S134N]. (a) Far UV circular dichroism (CD) protein spectra at 22 °C of the wildtype and T/A/Y/N mutant of MeCP2 with and without a *BRCA1* DNA duplex that contains a single CpG bearing the indicated cytosines; Average of three measurements; Contributions of buffer and DNA corrected. **(b)** Double wavelength plot of the mean molar ellipticity [θ] at 222 nm and 200 nm obtained from each spectrum in *a* in comparison with reference protein structures from Uversky (2003) and Whitmore et al. (2017) as reported by CAPITO (Methods). **(c)** Deconvolution of secondary structural elements from the spectra in *a* using DichroWeb (Miles, et al., 2021).

Overall, the CD measurements suggested that the sequence alterations in the T/A/Y/N mutant rendered the domain much less structured in absence of DNA. However, in presence of a DNA duplex with specifically modified CpG dyads such as 5hmC/5mC, the domain assumed a more structured fold of similar composition to the wildtype.

Binding orientation in the MBD–DNA complex. The MBD is a single, asymmetric protein domain with an α-helix flanking one side of the DNA binding site and a loop flanking the other site. Therefore, the binding of a strand-asymmetrically modified CpG dyad must take place in a specific orientation which can be revealed by measuring distances between two sites in the two macromolecules that would be differ depending on the binding orientation (**Figure 8.8a–b**) and do not affect the binding selectivity of the proteins (**Supplementary Figure A.20**). These were Ala-117 in the loop L1 of MeCP2 (to be substituted with a cysteine) and the phosphodiester five linkage upstream the CpG dyad (to be substituted with a phosphorothioate).

The spin labeling and double electron-electron resonance (DEER) measurements were performed by Jesscia Dröden (Prof. Dr. Malte Drescher, University of Konstanz) and will be presented and discussed in a place other than this work. In brief, the T/A/Y/N mutant in combination with a 5hmC/5mC duplex strongly preferred orientation of Figure 8.8b which places the 5hmC substituent in vicinity to the helix α1, hence the S134N substitution. This orienta-

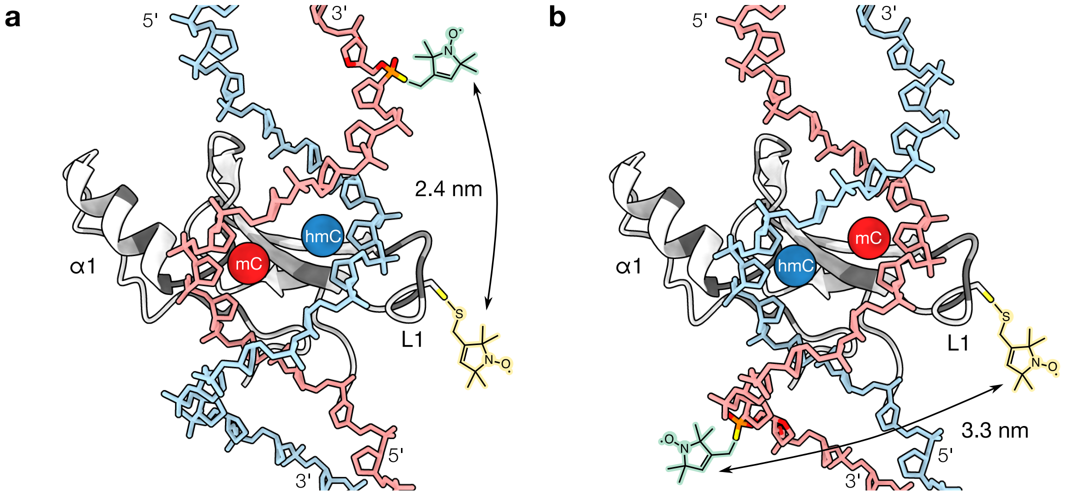


Figure 8.8 DNA duplex orientation in the MBD of MeCP2. Model based on PDB 3c2i (Ho, et al., 2008) indicating the placement of two MTSL electron spin labels, one in one strand of the DNA duplex and one in the protein domain that result in two distinct distance measures depending on the orientation of the DNA duplex in the DNA binding site. **(a)** Orientation with short distance which places the 5hmC modification on the unlabeled strand in the vicinity of loop L1. **(b)** Orientation with long distance which places the 5hmC modification in vicinity to the helix α1.

tion could suggest that the Asn-134 engages in hydrogen bonding with the hydroxyl group of 5hmC, while 5mC is recognized as in the wildtype MeCP2 via Tyr-123 which was rarely found substituted in the screen.

Another insightful observation that was made is that also wildtype MeCP2 with a fully methylated 5mC/5mC duplex (which theoretically could be bound in either orientation) showed a preference that placed the unlabeled strand in the same orientation as for the T/A/Y/N mutant. This corroborates the hypothesis of a second binding determinant must present at least in such homopolymeric sequence contexts as the oligo(A) one.

Overall the EPRspectroscopy contributed a highly valuable hypothesis about the possible recognition of strand-asymmetrically modified CpG dyads. To gain further insights at the atomic level and to elucidate the role of the required V122A mutation however much higher resolution is required which can be obtained, e. g., by nuclear magnetic resonance spectroscopy.

8.3 Synopsis

1. MeCP2[K109T/V122A/S134N] is an artificial MBD that preferentially binds to 5hmC/5mC CpG dyads with low nanomolar affinity in different sequence contexts.
2. In other sequence contexts, different modified CpG dyads cannot be well discriminated using wildtype or mutant MeCP2 for inherent structural or methodological reasons.

3. In contrast to wildtype MeCP2, spurious interactions with other dinucleotides, e. g., CpA dinucleotides, are weakened in the engineered domain.
4. In contrast to wildtype MeCP2, the engineered domain has a smaller α-helical contribution to the secondary structure but acquires wildtype-like levels upon DNA binding.
5. Ala-122 substitutions is essential for the binding selectivity of the engineered domain.
6. Asn-134 is located nearby the 5hmC nucleobase in the engineered DNA–MBD complex.

Endnote

[a] Compare also Table 6.1 for the N-terminal SpA(Z) fusion proteins. In the experiments referred to in this chapter, the protein solubility tag was removed *in situ* by proteolytic digest.

Chapter 9

Fitting binding isotherms to electrophoretic mobility shift data

A DNA molecule can present multiple binding sites to a DNA-binding protein in form of any recognizable biophysical quality of the biopolymer, such as its conformation, curvature or strandedness, as well as in form of different or repeated DNA sequence motifs. If the binding sites are well defined and the protein–DNA complexes sufficiently stable, the number of bound proteins can be experimentally revealed as discrete bands upon separation of the free and protein-bound states in an electrophoretic mobility shift assay.

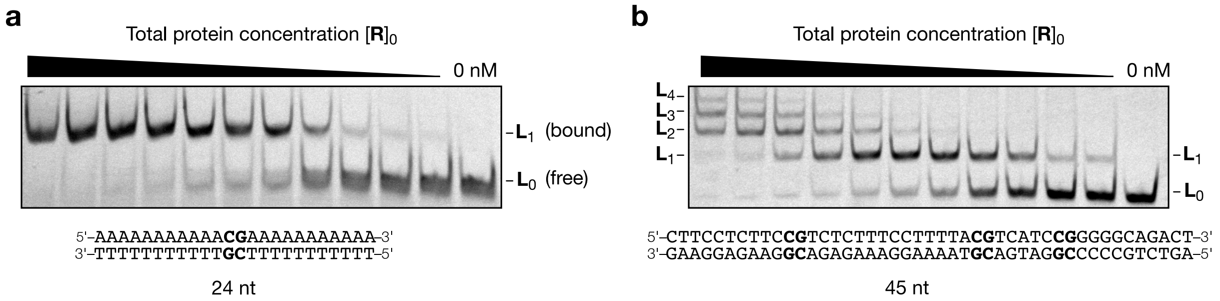

a
Total protein concentration [R]$_0$
0 nM
L$_1$ (bound)
L$_0$ (free)
5'–AAAAAAAAAAACGAAAAAAAAAAA–3'
3'–TTTTTTTTTTTGCTTTTTTTTTTT–5'
24 nt
b
Total protein concentration [R]$_0$
0 nM
L$_4$
L$_3$
L$_2$
L$_1$
L$_1$
L$_0$
5'–CTTCCTCTTCCGTCTCTTTCCTTTTACGTCATCCGGGGGCAGACT–3'
3'–GAAGGAGAAGGCAGAGAAAGGAAAATGCAGTAGGCCCCCGTCTGA–5'
45 nt

Figure 9.1 Typical DNA–protein EMSAs. Titration of an MBD **R** against 2 nM of a labeled DNA **L**. **(a)** Fractional binding observed for a DNA with a single modified CpG binding site. **(b)** Fractional binding observed for a DNA with three CpGs (one modified CpG) provoking multiple band shifts. Both assays in presence of poly(dA)·poly(dT) competitor.

Although such observations are particularly rewarding to the experimentalist—they are direct prove of additional binding interactions—more sophisticated experimentation is needed to determine the intrinsic binding affinities of each site (Brenowitz, et al., 1986). However, under certain assumptions a guess may be made based on the observed binding isotherm. This chapter addresses this challenge by introducing a computational framework that models the stepwise equilibrium constants as well as the underlying microscopic binding constants on a linear DNA molecule. I will start with the analysis of a single binding interaction (Section 9.1). Then, I recapitulate a model that describes the binding at multiple different sites conceived by Adair et al. (1925) and generalized by Wyman (1948; 1964). For a review in terms of statistical mechanics, I refer to Ben-Naim (2001a). The outline here is limited to the general concepts. which I will apply to the band shifts observed in an electrophoretic mobility shift assay in Section 9.2.

Since a convenient means to fit such models to the experimental data is missing to my knowledge, I will also outline the relevant code for the computational implementation in *R*. A package has been publicly released as 'summerr*band*' on GitHub (doi:10.5281/zenodo.5348399).[1]

[1] The package also provides other amenities such as importing from ImageQuant TL and adding column data.

9.1 Binding to a single site

The bimolecular *association* of a protein **R** and a ligand **L** to form the complex **RL** is a reversible reaction with the equilibrium constant K according to

$$\mathbf{R} + \mathbf{L} \rightleftharpoons \mathbf{RL}; \quad K = \frac{[\mathbf{RL}]}{[\mathbf{R}][\mathbf{L}]} = \frac{[\mathbf{RL}]}{([\mathbf{R}]_0 - [\mathbf{RL}])([\mathbf{L}]_0 - [\mathbf{RL}])} \tag{9.1}$$

where $[\mathbf{R}]$ and $[\mathbf{L}]$ refer to the concentration of the free, unbound protein and ligand in equilibrium.[a] They are inferred from the known initial total concentrations as $[\mathbf{L}]_0$ and $[\mathbf{R}]_0$. The experimentally accessible measure of protein-bound fraction over total ligand thus becomes

$$\frac{[\mathbf{RL}]}{[\mathbf{L}]_0} = \frac{[\mathbf{R}]_0 + [\mathbf{L}]_0 + 1/K - \sqrt{([\mathbf{R}]_0 + [\mathbf{L}]_0 + 1/K)^2 - 4[\mathbf{R}]_0[\mathbf{L}]_0}}{2[\mathbf{L}]_0} \tag{9.2}$$

Note that the assignment of **L** and **R** is interchangeable. In practice, one relates to the labeled entity that is held constant, here **L**, while the other entity, **R**, is titrated.

As a quadratic term, $1/K$ in Equation 9.2 must be fitted using non-linear models. In *R*, the following function is fed with the tabular data object `x` that has column variables that represent the recorded $[\mathbf{RL}]/[\mathbf{L}]_0$ for the various $[\mathbf{R}]_0$. The variables are specified, e. g. similar to `formula = bound_fraction ~ titrated_conc`. One must also provide the total constant concentration of $[\mathbf{L}]_0$, `L0`.[b] Further, the starting parameters and limits for the estimates `K`, `upper` and `lower` as well as the function `FUN` that implements the fitting algorithm are provided.[c]

```r
fit_K <- function(x, formula, L0 = NaN, ..., FUN = nls) {

  RL <- formula.tools::lhs(formula)
  L0 <- formula.tools::rhs(formula)

  if (is.finite(L0)) {  # formula for quadratic model

    FML <- substitute(RL ~ I(lower + (upper - lower) * ((R0 + L0 + 1 / K) - sqrt(
      (R0 + L0 + 1 / K)^2 - 4 * R0 * L0)) / (2 * L0)), list(RL = RL, R0 = R0, L0 = L0))

  }

  eval(rlang::call2(FUN, x, formula = stats::as.formula(FML), ...))

}
```

Two additional estimates, `lower` and `upper - lower`, account for any background and saturation offsets present in the quantitated signals (Altschuler, et al., 2012).

If a computer is not at hand, Equation 9.2 cannot be fitted easily to the experimental data. However, if (1) the total concentration of **L** was held constant at concentrations below $1/K$ at a still reasonable signal-to-noise ratio for accurate quantitation,[d] and (2) $\mathbf{R}$ is titrated over a concentration range well above and below $1/K$, then $[\mathbf{R}] \approx [\mathbf{R}]_0$ holds and one may simplify

$$\frac{[\mathbf{RL}]}{[\mathbf{L}]_0} = \frac{[\mathbf{R}]_0}{[\mathbf{R}]_0 + 1/K} \quad \left(= \frac{K[\mathbf{R}]_0}{K[\mathbf{R}]_0 + 1} \right) \tag{9.3}$$

This relationship becomes linear upon double-logarithmic transformation so that $1/K$ is the intersection with the abscissa and a linear regression could extract $-\log K$. It must be warned however that the variance in the measured quantities becomes heteroskedastic during the transform and although the estimate of $-\log K$ is unaffected, the estimate of its variance is biased. As a consequence, I implement Equation 9.3 using again the non-linear least square approach:

```
if (!is.finite(L0)) {  # formula for infinite receptor pool

  FML <- substitute(RL ~ I(lower + (upper - lower) * (R0 / (R0 + 1 / K))),
            list(RL = RL, R0 = R0))

}
```

9.2 Binding to multiple sites

On the molecular level, the association of two molecular entities is fully described by the above models. Any deviating macroscopic observation must thus be interpreted in terms of additional interactions between more than two entities even if they seem uniform at first glance: A body of theorems can describe such *binding cooperativity* of identical binding sites (Adair, et al., 1925; Hill, 1910). Here, I follow a similarly simple, combinatorial approach to examine the non-identical binding sites on a DNA molecule presented to a homogenous protein binding partner.

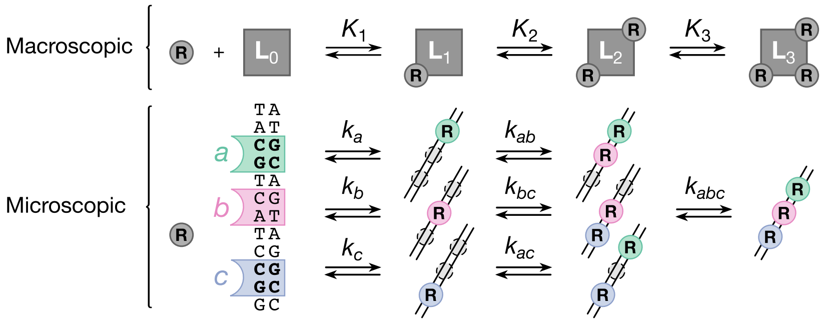


Figure 9.2 Macroscopic and microscopic descriptors of multistep binding. A DNA molecule with three binding sites for a DNA-binding protein can be present in one of four macroscopic, or eight microscopic, complexes. The respective equilibrium constants *K* and *k* take the name from the complex they form. Note that, e. g., complex *ab* can form via two processes.

This association can be dissected step by step, e. g., for three distinct binding sites *a*, *b*, *c* (Figure 9.2): The first binding equilibria are described by the intrinsic microscopic equilibrium constants k_a, k_b, k_c, the second equilibria by k_{ab}, k_{bc}, k_{ac} and the last one by k_{abc}. The values of these constants may be identical or dissimilar within or between the steps, but they are connected to each other: For example, k_{ab} is connected through an interaction term g_{ab} to k_a and k_b etc. so that $g_i = 1$ when independent. Experimentally however, only the sequential macroscopic binding constants K_1, K_2, K_3 etc. are accessible. They describe the association of one, two, three etc. proteins given that no, one, two etc. proteins are already bound. The configuration of the binding sites remains unspecified.

9.3 Macroscopic apparent equilibrium constants

Let's designate the adsorbent DNA molecule with no proteins bound $\mathbf{L}_0$, the complex **RL** (containing one **R**) be $\mathbf{L}_1$, the complex with two **R** be $\mathbf{L}_2$ etc. Then, the association of another **R** to the existing complex $\mathbf{L}_{i-1}$ is, akin to Equation 9.1,

$$\mathbf{L}_{i-1} + \mathbf{R} \rightleftharpoons \mathbf{L}_i \,; \quad K_i = \frac{[\mathbf{L}_i]}{[\mathbf{R}][\mathbf{L}_{i-1}]} \tag{9.4}$$

From the law of mass conservation, one obtains the total concentration of DNA molecules **L** in terms of $x = [\mathbf{R}]$ (Equation 9.5) and the average concentration of engaged DNA–protein complexes (Equation 9.6) as proposed by Wyman (1964). The term $\xi_n'(x)$ is called the *binding polynomial* and reflects the grand partition function in **R** relative to the reference state $\mathbf{L}_0$.

$$\begin{aligned}\sum_{i=0}^{n}[\mathbf{L}_i] &= [\mathbf{L}_0](1 + K_1x + K_1K_2x^2 + K_1K_2K_3x^3 + \cdots + K_1\cdots K_nx^n)\\ &= [\mathbf{L}_0]\textstyle\sum_{i=0}^{n}\left(\prod_{j=0}^{i}(K_j)\cdot x^i\right) = [\mathbf{L}_0]\xi_n'(x) \end{aligned} \tag{9.5}$$

$$\sum_{i=0}^{n} i[\mathbf{L}_i] = [\mathbf{L}_0](K_1x + 2K_1K_2x^2 + 3K_1K_2K_3x^3 + \cdots + nK_1\cdots K_nx^n) \tag{9.6}$$

Hence, the average number of bound proteins per DNA molecule becomes

$$\vartheta_n(x) = \frac{\sum_i^n i[\mathbf{L}_i]}{\sum_i^n [\mathbf{L}_i]} = \frac{K_1x + 2K_1K_2x^2 + 3K_1K_2K_3x^3 + \cdots + nK_1\cdots K_nx^n}{1 + K_1x + K_1K_2x^2 + K_1K_2K_3x^3 + \cdots + K_1\cdots K_nx^n} \tag{9.7}$$

$$= \frac{\partial\xi_n'(x)}{\partial x}\cdot\frac{x}{\xi_n'(x)} \tag{9.8}$$

and is a function solely of the concentration of the unbound protein x for which $x = [\mathbf{R}] \approx [\mathbf{R}]_0$ (page 105). This measure is called the *binding isotherm* $\vartheta_n(x)$ with $0 \leq \vartheta_n(x) \leq n$. For $n = 1$, one obtains the bracketed term in Equation 9.3.

Thus the macroscopic Adair-Klotz constants K_j can be determined by fitting the observed average occupancy per DNA molecule to Equation 9.7. For an EMSA with n discrete bands, this is the sum over all bands in a lane multiplied each with the number of associated proteins (starting at $i = 0$) and weighted by their relative intensities s_i:

$$\vartheta_n(x) = \sum_{i=0}^{n} i \cdot s_i \tag{9.9}$$

First, I implement in *R* a function to generate the verbose expression for the grand partition function $\xi'_n(x)$ using the stepwise equilibrium constants K_j as coefficients.

```
gpf_macro <- function(degree = 2, params = as.character(seq(degree)), xname = "x",
                      kname = "K_") {
  parse(text = paste0("1 + ", paste0(sapply(seq_along(params), function(m) paste0(
        paste0(kname, params[1:m], collapse = " * "), " * ", xname, "^", m)),
        collapse = " + ")))
}
```

Then, the following function will evaluate a suitable $\xi'_n(x)$ for all concentrations `x` given a (named) vector `binding_constants` with the values of all K_j. For the macroscopic case, the degree n of the polynomial is determined from the length of this vector.

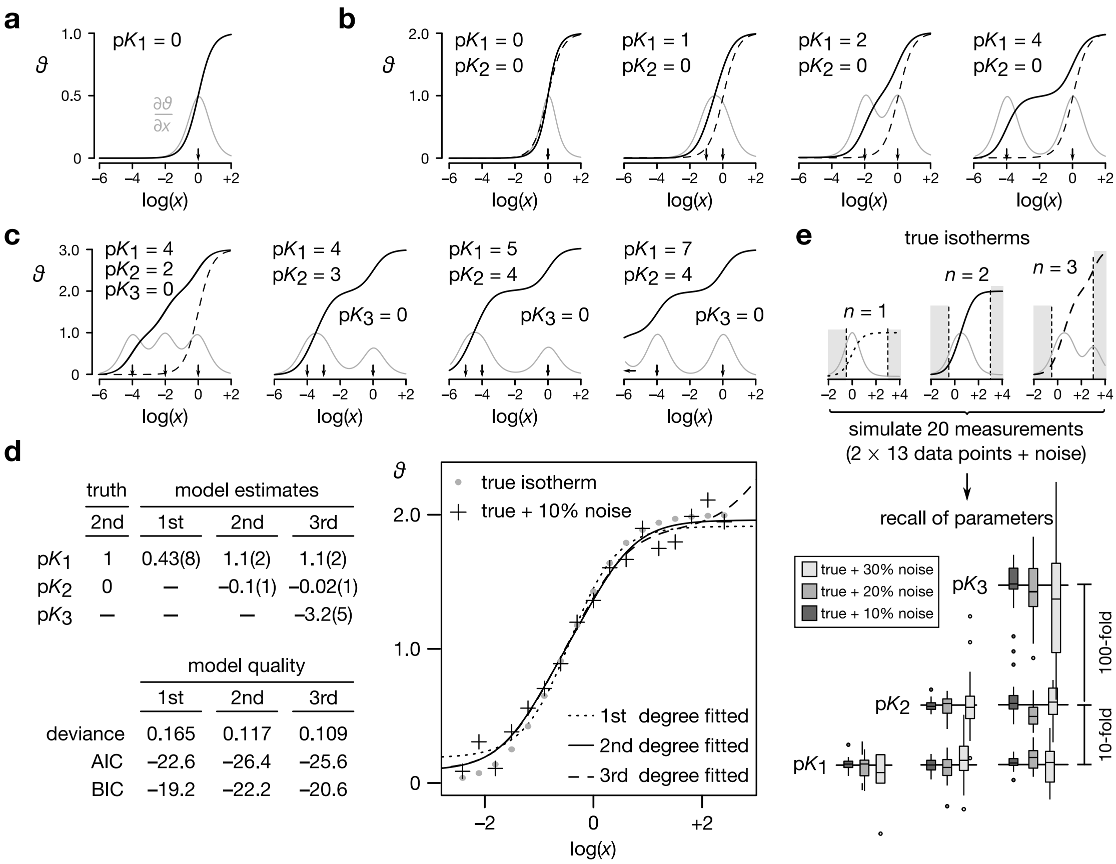
```
gpf_fraction_bound <- function(x, binding_constants, type = "macro") {
  if (type == "macro") {
    binding_constants <- sort(binding_constants[which(binding_constants > 0)],
                              decreasing = TRUE, na.last = NA)
    bd <- length(binding_constants)  # degree of the polynomial
    if (is.null(names(binding_constants))) {
      kname <- "K_"; bn <- paste0(kname, as.character(seq_along(binding_constants)))
    } else {
      names(binding_constants)[which(names(binding_constants) == "x")] <- "..x"
      kname <- " "; bn <-  names(binding_constants)
    }
  }  # omitted code is on page 110
  names(binding_constants) <- bn
  gpf <- do.call(paste0("gpf_", type), list(xname = "x", kname = kname, params = sub(
    pattern = kname, replacement = "", names(binding_constants)[1:bd], fixed = TRUE)))
  dpf <- stats::D(gpf, "x")

  env <- c(list(x = x), as.list(binding_constants))

  eval(dpf, envir = env) * x / eval(gpf, envir = env)

}
```

Now, the properties of the macroscopic binding isotherm can be explored in *R* (Figure 9.3).



	truth	model estimates		
	2nd	1st	2nd	3rd
pK_1	1	0.43(8)	1.1(2)	1.1(2)
pK_2	0	–	−0.1(1)	−0.02(1)
pK_3	–	–	–	−3.2(5)

	model quality		
	1st	2nd	3rd
deviance	0.165	0.117	0.109
AIC	−22.6	−26.4	−25.6
BIC	−19.2	−22.2	−20.6

Figure 9.3 Modeling and fitting binding isotherms based on macroscopic descriptors of multistep binding. (a) Modeling the binding isotherm ϑ for a single-step process, i. e., one binding site; **(b)** Modeling of a two-step process with identical or different macroscopic equilibrium constants *K* according to Equation 9.7; dashed lines are the (scaled) single-step binding isotherms of *a*; gray lines indicate the first derivative of the modeled binding isotherms. **(c)** Same as *b* for the three-step process. **(d)** Fitting models of different degree to simulated noisy data for a two-step process with close-by macroscopic binding constants. The fitted estimates (last digit standard errors), deviance, as well as each model's Akaike and Bayesian information criterion (AIC, BIC) are tabulated. All *K* and *x* normalized to the units of concentration of the titrated species and given as their negative decadic logarithm. **(e)** Recall as in *d* for noisy data under difficult parameter combinations. Source code available.

As compared to the single-step binding reaction, the isotherm of the multistep binding process has a somewhat increased steepness if all secondary associations take place with the same apparent 'affinity' as the first ones. The more dissimilar they become, the more ligand is required to saturate all sites simultaneously as seen from the plateaus where $K_{i-1} \gtrsim 100\,K_i$.

To test whether the implemented non-linear least square fitting was able to pick up the correct underlying model, isotherms were constructed that had $K_1 \approx 10\,K_2$ (Figure 9.3 d).[e] Indeed, a simple model with $n = 1$ (first degree) underestimated K_1 about fourfold. Models of second and third degree estimated K_1 and K_2 accurately within the limits of error. The third degree model had the lowest deviance, proposing an additional association with a K_3 around 1,000-fold weaker than K_2. Although this is theoretically possible, such a claim remains unsupported by the data. Based on the three models' relative information content (Akaike and Bayesian information criterion), one would have however correctly selected the second degree model.

From a larger cohort of simulated binding isotherms in which but part of the plateaux were 'measured' (Figure 9.3 e), the estimates of the macroscopic binding constants were still fairly accurate even if the level of noise amounted to 20% of the fractional binding. Further, in about 60 – 80% of the simulated cases, the underlying binding model among three alternatives was correctly identified based on AIC and 70 – 90% by BIC. However, the more the shape of the isotherms is blurred by random noise, the higher the propensity to favor the simpler models (data not shown). Experimentally however, the degree of the underlying model can often be clearly identified if the titration spans a suitable range of ligand concentrations.

So, by implementing a framework in *R*, the macroscopic binding isotherms of molecular associations at multiple binding sites can be adequately evaluated, offering a valuable alternative to fitting simple single-site binding models. Despite 'similar shape of the isotherms' the value of these parameters can often be accurately estimated even from noisy data given the degree of the binding polynomial is known by experiment.

9.4 Microscopic equilibrium constants

The macroscopic binding constants K_i are averages over all ways to add a single protein **R** to a DNA $\mathbf{L}_i$ that has $i - 1$ proteins bound. So, the binding polynomial $\xi'_3(x)$ for three individual sites a, b, c is expressed as summation over the microscopic intrinsic binding constants k_a, k_b, k_c, k_{ab}, k_{bc}, k_{ac}, k_{abc} (Figure 9.2). For n sites, one generalizes $\xi'_n(x)$:

$$\xi'_3(x) = 1 + (k_a + k_b + k_c)x + (k_{ab} + k_{bc} + k_{ac})x^2 + k_{abc}x^3 \tag{9.10}$$

$$\xi'_n(x) = \textstyle\sum_{i=0}^{n} \big(\sum_{\sum j=i} (k_j) \cdot x^i \big) \tag{9.11}$$

The binding isotherm for one specific of three binding sites, e. g. a, is

$$\vartheta_{a,3}(x) = \frac{k_a x + (k_{ab} + k_{ac})x^2 + k_{abc}x^3}{\xi_3'(x)} \tag{9.12}$$

and the isotherm for the entire macromolecule becomes

$$\begin{aligned}\vartheta_3(x) &= \vartheta_{a,3}(x) + \vartheta_{b,3}(x) + \vartheta_{c,3}(x) \\ &= \frac{(k_a + k_b + k_c)x + 2(k_{ab} + k_{bc} + k_{ac})x^2 + 3k_{abc}x^3}{1 + (k_a + k_b + k_c)x + (k_{ab} + k_{bc} + k_{ac})x^2 + k_{abc}x^3}\end{aligned} \tag{9.13}$$

Again, I implement Equation 9.11 for any degree n in *R*

```
gpf_micro <- function(degree = 2, params = letters[seq(degree)], xname = "x",
                      kname = "k_") {

  parse(text = paste0("1 + ", paste0(sapply(seq_along(params), function(m) paste0("(",
        paste0(kname, utils::combn(params, m, FUN = paste0, collapse = ""),
        collapse = " + "), ") * ", xname, "^", m)), collapse = " + ")))

}
```

and expand `gpf_fraction_bound(...)` to calculate the binding isotherm.

```
  # code to be inserted on page 107

  if (type == "micro") {

    bd <- which(sapply(1:10, function(i) sum(choose(i, 1:i))) == length(
                       binding_constants))  # user provides less than 1000 parameters

    if (length(bd) == 0) stop("Some binding constants are missing (or superfluous).")

    kname <- "k_"
    bn <- unlist(sapply(seq(bd), function(m) paste0(kname, utils::combn(letters[1:bd],
                                                    m, FUN = paste0, collapse = ""))))

  }
```

(9.4.1) Modeling microscopic binding constants from the binding isotherm

Ideally, one would like to extract the intrinsic binding constants k_a, k_b, k_c etc. for the individual binding sites from the observed binding isotherm. From Equation 9.7 and Equation 9.13, one identifies for a macromolecule with three binding sites:

$$\begin{aligned} K_1 &= k_a + k_b + k_c \\ K_1 K_2 &= k_{ab} + k_{ac} + k_{bc} = g_{ab}k_a k_b + g_{ac}k_a k_c + g_{bc}k_b k_c \\ K_1 K_2 K_3 &= k_{abc} \qquad\qquad\quad = g_{abc}k_a k_b k_c \end{aligned} \tag{9.14}$$

So, the macroscopic binding constants K are determined by the intrinsic constants k. It is commonly assumed that the microscopic constants for the second and further associations, i. e. k_{ab} etc., are derived from k_a, k_b etc. through correlation factors g_{ab} etc.

Clearly though, it is not possible to derive an analytical expression for the seven microscopic unknowns from the three macroscopic ones.[f] To determine these values, one must obtain the individual binding isotherms $\vartheta_{a,3}(x)$, $\vartheta_{b,3}(x)$, $\vartheta_{c,3}(x)$ according to Equation 9.12 under the same experimental conditions as $\vartheta_3(x)$. This means to follow the fractional binding at each site while the other sites are unrestrained to becoming occupied or not (Ben-Naim, 2001b). This is experimentally difficult in particular when some associations are weak. In such cases, the identification of the occupied binding sites themselves, e. g., by DNase footprinting titration (Brenowitz, et al., 1986), significantly interferes with the ligand binding.

In absence of access to such empirical data, one must come to reasonable assumptions about the studied system in order to simplify the Equation System 9.14 and derive a model with less unknowns to determine the intrinsic binding constants k_a, k_b, k_c.

In terms of statistical mechanics g_{ab} describes the free energy change $W(a, b)$ for any process that is equivalent to the following transformation: Starting with two systems of which one has all sites a and the other one all sites b occupied; ending with two systems, one has both sites fully occupied, the other has empty binding sites (Ben-Naim, 2001b). For this process

$$W(a,b) = W_{\text{conf}\,\mathbf{L}}(a,b) + W_{\text{conf}\,\mathbf{R}}(a,b) + W_{\text{tr}}(a,b) + W_{\text{rot}}(a,b) = -\mathrm{k_B}T \ln g_{ab} \qquad (9.15)$$

One must hence consider the following contributions to this change in free energy in the context of MBD–DNA binding:

Assumptions about conformational changes in the adsorbent DNA. $W_{\text{conf}\,\mathbf{L}(a,b)}$ reflects the impact of occupying a on the binding site b through conformational changes in the DNA itself. For DNA-binding proteins, one may consider short- and long-range effects, depending on the protein. As extremum, one includes the possibility of physical obstruction when a and b are neighboring or overlapping motifs.

For the MBD, the small conformational changes in the DNA duplex are essentially limited to the 12 – 14 bp covered by the domain (Ho, et al., 2008) and probably negligible. Considering also the winding and rise of the B-form DNA double-helix, one can assume that MBD binding at a site will not obstruct another site that is more than 4 – 5 bp apart.

Assumptions about the configurations of the adsorbed protein. Generally, it is assumed that the ligand binds in a single orientation to the macromolecule. In the case of DNA-binding proteins, especially the MBDs as non-self symmetric single-domain binders, this may not always be the case. Whereas the association with a symmetrically modified CpG, 5mC/5mC, 5hmC/5hmC, etc. can take place in one of two possible configurations, the binding of asymmetrically modified CpG dyads is strictly directional. It may thus be that a specific arrangement of binding sites demands a specific orientation of the MBD or not so that the number of configurations of such a system may be larger or smaller than the number of binding sites.

However, one would have $W_{\text{conf}\,\mathbf{R}}(a,b) \neq 0$ only for neighboring binding sites through pair cooperativity (Ben-Naim, 2001b).

Assumptions about mass ratios and inertia effects. For large macromolecules, the total complex $\mathbf{L}_i$ is significantly larger than $\mathbf{L}_{i-1}$ and thereby contributes not only to nonadditivity of the correlation functions g, but also to negative cooperativity between the binding sites (Ben-Naim, 2001b). The effect is of the magnitude

$$\begin{aligned} W_{\text{tr}}(a,b) = W_{\text{tr}}(a,c) = W_{\text{tr}}(b,c) &= -\frac{3}{2}\text{k}_\text{B}T\ln\left[\frac{1+2\,m_\mathbf{R}/m_\mathbf{L}}{(1+m_\mathbf{R}/m_\mathbf{L})^2}\right] \\ W_{\text{tr}}(a,b,c) &= -\frac{3}{2}\text{k}_\text{B}T\ln\left[\frac{1+3\,m_\mathbf{R}/m_\mathbf{L}}{(1+m_\mathbf{R}/m_\mathbf{L})^3}\right] \end{aligned} \tag{9.16}$$

which for an MBD ($m_\mathbf{R} \approx 10\,\text{kDa}$) and a 45 nt DNA duplex ($m_\mathbf{L} \approx 30\,\text{kDa}$) gives a hypothetical, purely translational cooperativity factor of $g_{\text{tr}} \approx 0.91$ for the $\mathbf{L}_1 \rightarrow \mathbf{L}_2$ reactions and $g_{\text{tr}} \approx 0.78$ to the $\mathbf{L}_2 \rightarrow \mathbf{L}_3$ process.

Besides the translational partition function, also the rotational partition function is affected by the non-negligible mass differences. For a linear macromolecule with three approximately equally distributed binding sites, of which b lies between a and c, this is

$$\begin{aligned} W_{\text{rot}}(a,b) &\approx W_{\text{rot}}(b,c) = 0 \\ W_{\text{rot}}(a,c) = W_{\text{rot}}(a,b,c) &= -\text{k}_\text{B}T\ln\left[\frac{4(1+3\,m_\mathbf{R}/m_\mathbf{L}}{(2+3\,m_\mathbf{R}/m_\mathbf{L})^3}\right] \end{aligned} \tag{9.17}$$

One calculates a hypothetical, purely rotational contribution to cooperativity for the processes with non-zero free energy change of $g_{\text{rot}} \approx 0.89$.

Combination. Overall, if the three binding sites do not overlap or situate next to each other, one can estimate for an MBD–DNA system with three binding sites based on the assumptions above

$$\begin{aligned} g_{ab} = g_{bc} &\approx 0.91 \\ g_{ac} &\approx 0.81 \\ g_{abc} &\approx 0.68 \end{aligned} \tag{9.18}$$

which are small effects as compared to the binding of the λ operator where some $g_i \gg 10$ reflect the strong positive cooperation between the adjacent binding sites (Ben-Naim, 2001b).

What does such a simplified model with most $g_i \approx 1$ means for k_a etc. and the derived macroscopic constants K_1 etc. mean for MBD interaction? Again, different parameters of the system can be explored using the *R* implementation (Figure 9.4).

The isotherms of the microscopic models with $g_{ab} = 1$ (independence) follow the macroscopic models of Figure 9.3 very closely when $K_1 = k_a$ and $K_2 = k_b$, which is also expected from Equation 9.14. Also, for small negative or positive cooperativity, $0.5 \leq g_{ab} \leq 2.0$, one expectedly finds only small deviations at higher x, i. e., for the weaker secondary associations.

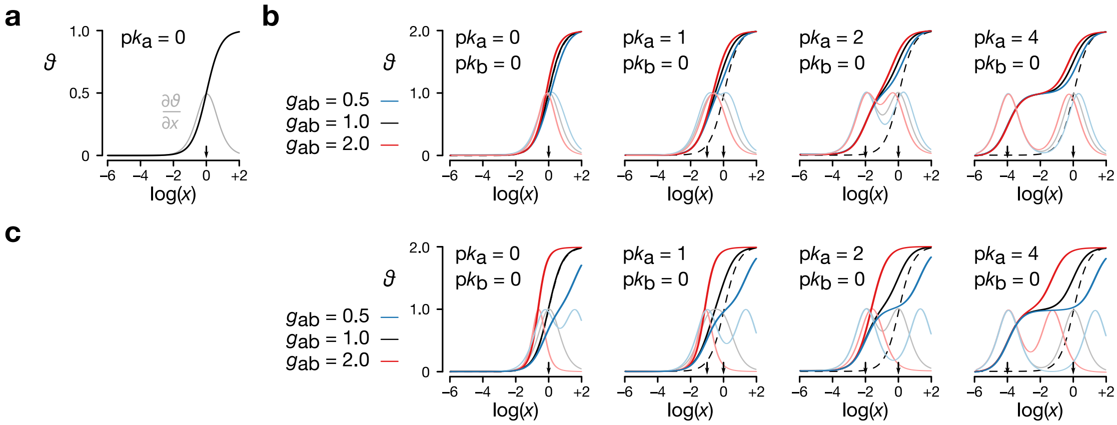


Figure 9.4 Modeling and fitting binding isotherms based on microscopic descriptors of multistep binding. (a) Modeling the binding isotherm ϑ for binding a single binding site; **(b)** Modeling binding of identical or different binding sites with intrinsic binding constants *k* according to Equation 9.13 given the correlations *g*; dashed lines are the (scaled) single-step binding isotherms of Figure 9.3 a; light lines indicate the first derivative of the modeled binding isotherms. **(c)** Same as *b* with more extreme values for *g*.

Hence, given a binding isotherm, the larger the differences between k_a and k_b, the better

$$K_1 \approx k_a \tag{9.19}$$

for small correlation factors. Even if k_a and k_b differ only five- to tenfold, this seems reasonable.

If the correlation factors become larger, here $g_{ab} \ll 0.5$ for negative interactions and $g_{ab} \gg 2$ for positive interactions, approximation 9.19 holds only if k_a and k_b differ at least 100-fold or more.

Note that in all cases, one may certainly not state $K_2 \approx k_b$.

9.5 Implementation in R and additional features

Similar to `fit_K(...)` (on page 104), one feeds the function `fit_binding_isotherm(...)` in *R* with a tabular data object x. This table has side-by-side the measured (fractional) band intensities for each $\mathbf{R}_0$ along with other variables that identify the measurement. One provides a formula to calculate the observed binding isotherm (Equation 9.9) in dependence of the titrated $\mathbf{R}_0$ such as `formula = 1 * band_1 + 2 * band_2 ~ titrated_conc`. From this formula, the degree of the binding polynomial (Equation 9.5 or 9.11) is established if unstated.

For the fitting, there are several options:

- Fitting of the macroscopic stepwise binding constants K or the microscopic intrinsic binding constants k. For the latter, the correlation coefficients g must be provided as named vector `correlation = c(...)`.
- Fitting the constants for two groups specified in a column with the name declared with `INDEX =` under the premise that k_a (or K_1 if reasonable) is group-specific, but binding to b (and c where applicable) should be the same for both groups.
  This is particularly useful when one fits data from experiments, in which only the DNA is locally modified at site a, e. g. by DNA methylation, but the sequence context and hence the correlation functions are unlikely to change.

To minimize the chance of getting trapped in non-global minima across all combinations of variables to estimate, one uses a grid-start approach for a non-linear least square fitting implemented by the `nls.multstart` package (Padfield, et al., 2021). To evenly sample this space, the binding constants are log-transformed.

NOTE: The binding constants are log-transformed dissociation constants $K_d = 1/K$ or $K_d = 1/k$.

```
fit_binding_isotherm <- function(x, formula, degree = NULL, type = "macro", correlation,
                                 INDEX = NULL, ..., start_K_d = c(-1, 4)) {

  # required helper functions are on page 116

  RL_isotherm <- function(...)
  RL_isotherm_shared <- function(...)

  # generate observed binding isotherm according to formula virtually

  x <- dplyr::mutate(.data = x, .RL = !!formula.tools::lhs(formula))
```

```
# determine degree of binding polynomial to fit
if (is.null(degree)) degree <- length(formula.tools::lhs.vars(formula))
degree <- as.integer(degree); stopifnot(all(is.finite(degree), degree <= 3))
# further parameter parsing
L0 <- formula.tools::rhs(formula); INDEX <- rlang::enquo(INDEX)
# construct formula according to shared/grouped or ungrouped evaluation
if (rlang::quo_is_null(INDEX)) {
  FML <- substitute(.RL ~ RL_isotherm(conc_L = L0, pK_d1, pK_d2, pK_d3, upper, lower,
                                      type = T0), list(L0 = L0, T0 = type))
  # setup of start ranges, grid ranges and parameters not shown
  starts <- list(...); iters <- list(...); params <- list(...)
} else {
  INDEX <- rlang::as_name(INDEX); stopifnot(INDEX %in% colnames(x))
  FML <- substitute(.RL ~ RL_isotherm_shared(conc_L = L0, INDEX = I0,
           pK_d1.x, pK_d1.y, pK_d2, pK_d3, upper.x, lower.x, upper.y, lower.y,
           type = T0), list(L0 = L0, I0 = x[[INDEX]], T0 = type))
  #  setup of start ranges, grid ranges and parameters not shown
  starts <- list(...); iters <- list(...); params <- list(...)
}
# further parameter removal/expansion
if (degree < 3) {
  iters$pK_d3 <- starts$pK_d3 <- params$pK_d3 <- NULL
  FML <- do.call("substitute", list(FML, list(pK_d3 = Inf)))
}
if (degree < 2) {
  iters$pK_d2 <- starts$pK_d2 <- params$pK_d2 <- NULL
  FML <- do.call("substitute", list(FML, list(pK_d2 = Inf)))
}
ll <- sapply(params, min, na.rm = TRUE, USE.NAMES = TRUE)
ul <- sapply(params, max, na.rm = TRUE, USE.NAMES = TRUE)
li <- sapply(starts, min, na.rm = TRUE, USE.NAMES = TRUE)
```

```
    ui <- sapply(starts, max, na.rm = TRUE, USE.NAMES = TRUE)

40  eval(rlang::call2(.fn = "nls_multstart", .ns = "nls.multstart", data = x,
                      formula = stats::as.formula(FML), iter = unlist(iters),
                      lower = ll, upper = ul, start_lower = li, start_upper = ui))

  }
```

In terms of the programming technique to implement the broad set of model specifications, `nls.multstart` requires a formula argument in which the variables to estimate must appear as unassigned arguments of a function call (lines 40–42). In order to avoid creating three independent functions to fit $\vartheta_1(x)$, $\vartheta_2(x)$, $\vartheta_3(x)$ etc., one can reduce the effort to a single function for the highest degree, `RL_isotherm(...)`, and create the required formula (lines 15,16 or 21–23) from which one later removes any unwanted parts (lines 28–35).

`RL_isotherm(...)` allows to calculate binding isotherms up to $n = 3$. As for the single-site binding models, I include two additional variables, `lower` and `upper - lower`, to account for background and saturation offsets present in the quantitated signals. The realization of the function is straightforward.

```
RL_isotherm <- function(conc_L, pK_d1, pK_d2, pK_d3, upper, lower, type = "macro") {

  if (type == "micro") {

    correlation <- c(correlation, c(ab = 0, bc = 0, ac = 0, abc = 0)[setdiff(
                                  c("ab", "bc", "ac", "abc"), names(correlation))])

    a <- 10^(-pK_d1); b <- 10^(-pK_d2); c <- 10^(-pK_d3)

    ab <- unname(correlation["ab"] * a * b)
    ac <- unname(correlation["ac"] * a * c)
    bc <- unname(correlation["bc"] * b * c)

    abc <- unname(correlation["abc"] * a * b * c)

    params <- c(a = a, b = b, c = c, ab = ab, ac = ac, bc = bc, abc = abc)

  } else {

    params <- c(K1 = 10^(-pK_d1), K2 = 10^(-pK_d2), K3 = 10^(-pK_d3))

  }

  lower + (upper - lower) * gpf_fraction_bound(x = conc_L, binding_constants = params,
                                               type = type)

}
```

Moreover, one can resort to this function, when one requires to share some of the estimates between two groups.

```
RL_isotherm_shared <- function(conc_L, INDEX = NULL,
                               pK_d1.x, pK_d1.y = pK_d1.x, pK_d2, pK_d3,
                               upper.x, lower.x, upper.y = upper.x,
                               lower.y = lower.y, type = "macro") {

  INDEX <- as.factor(INDEX)

  if (length(levels(INDEX)) == 2) {

    conc_L <- split(conc_L, INDEX, drop = FALSE)

    rlx <- RL_isotherm(conc_L[[1]], pK_d1.x, pK_d2, pK_d3, upper.x, lower.x, type)
    rly <- RL_isotherm(conc_L[[2]], pK_d1.y, pK_d2, pK_d3, upper.y, lower.y, type)

    res <- list(rlx, rly); names(res) <- names(conc_L)
    res <- unsplit(res, INDEX, drop = TRUE)

  } else {

    if(length(levels(INDEX)) > 2) warning("Data splits into more than 2 groups;
                                           not grouping at all now.")

    res <- RL_isotherm(conc_L, pK_d1.x, pK_d2, pK_d3, upper.x, lower.x, type = type)

  }

  res

}
```

9.6 Synopsis

Binding of ligands to multiple binding sites on a macromolecule such as proteins binding to DNA makes the analysis of binding isotherms more complex.

1. Macroscopic stepwise equilibrium constants can be derived from Wyman's a generalized binding polynomial for which I present here an implementation in *R*.
2. The exact modeling of the complete set of underlying microscopic equilibrium constants requires experimentation in addition to band shift or filter binding assays.
3. However, if the differences in the microscopic association constants are large and the binding reactions almost independent of each other, K_1 of the macroscopic binding model is a good predictor for the most affine intrinsic association constant k_a in the microscopic model.
4. For the binding of a MBD to DNA with a single modified CpG and other non-modified CpG dyads or further DNA binding motifs, this assumption may hold true.

Endnotes

[a] Strictly speaking, one assumes that the activity of both entities is proportional to their concentration; which appropriate for sufficiently dilute solutions.

[b] The parameter and function names used in 'summerr*band*' differ from the shown ones: The dissociation constant $K_d = 1/K$ with `fit_Kd(...)` is fitted and in practice the concentration of the ligand under the name `R0 = ...` provided. The argument names were chosen such that it is generally more intuitive to titrate a ligand than a receptor. Only a part of the function implementation is shown.

[c] This setup gives the user full control the fitting algorithm. The default choice is the Levenberg-Marquat algorithm implemented in `minpack.lm::nlsLM` rather than the shown Gauss-Newton algorithm of `base::nls`.

[d] Altschuler et al. (2012) notes that "using a nucleic acid concentration that is too high is one of the most commonly made mistakes in EMSA and filter binding experiments".

[e] Note that during model fitting, one introduces a saturation factor such that the isotherm $\vartheta_n(x)$ may be scaled and its absolute maximum value, i. e. the number of bound ligands n, be freely chosen.

[f] One solution of the system of equations are the harmonic means for $K_2 = (k_{ab} + k_{ac} + k_{bc})/K_1$ with $k_{ab} = 2k_a k_b$, $k_{ac} = 2k_a k_c$, $k_{bc} = 2k_b k_c$ and for $K_3 = k_{abc}/(K_1 K_2)$ with $k_{abc} = 6k_a k_b k_c$; suggesting for no reason high positive cooperativity.

Conclusions

and

Future Development and Outlook

Chapter 10

Conclusions

With the work presented in this thesis, I have contributed (1) a platform to screen millions of protein–DNA interactions for the specific recognition of modified cytosines on the bacterial cell surface, (2) a systematic biochemical characterization and compilation of available reports on the binding specificity of different methyl-CpG-binding domains towards specific combinations of modified cytosines in CpG dyads, and (3) several engineered proteins based on this domain which, for the first time, allow to probe the presence of specific combinations of modified cytosines in CpG dyads as well as to gain further knowledge about the rules that can govern such interactions. These contributions will aid our further understanding of the role and relevance of strand-symmetrically and strand-asymmetrically modified CpG dyads in the genome where they may serve as distinct, probably epigenetic signals for the organism.

10.1 Monitoring protein–DNA interactions on the bacterial cell-surface

An enabling technology for the discovery of MBDs with novel DNA binding selectivity was an assay by which the DNA–protein interactions could be probed for surface-displayed protein passengers. Particularly successful was the screening of a degenerated MeCP2 library.

AIDA-I-mediated MBD cell surface display. With the bacterial cell surface display platform established for this work, I was able to display all five human MBDs in a functional state and inducible manner on the bacterial cell surface with minor adjustments to the specific passenger. In the best case encountered with MBD2, an average cell displayed about 50,000 functional MBD molecules. Yet the usage of *bacterial* cell surface display platforms to probe protein–DNA interactions is rare in the literature. In the present case, this choice had very likely an unexpected advantage over yeast and other eukaryotic cell surface display platforms for the screening of interactions with non-canonical DNA nucleobases. Surprisingly, all candidate MeCP2 variants tested to date showed significantly reduced binding to CpA dinucleotides whereas wildtype MeCP2, capable of engaging with this combination, remained but poorly displayed. If this was not out of intrinsic biophysical necessity or pure coincidence, it could be hypothesized that the temporary exposure to the bacterial nucleoid could have served as a counterselection to sequester passengers with binding affinity towards undesired combinations of canonical DNA nucleobases such as to CpA dinucleotides in the cell. Mechanistically, the passenger must remain unfolded during the autotransport. However, this applies only to the passage of the *outer* membrane; The translocation of the *inner* membrane is mediated by the Sec translocon for which co- and post-translational pathways have been described during which a folded protein can be unfolded again (Denks, et al., 2014). By whichever means, counterselections that precede the display of the protein variant are extremely valuable for directed evolution as they increase ligand (or substrate) specificity as well as activity of the retrieved candidates. Therefore they are often purposefully engineered into the displaying host (Yi, et al., 2013). An important corollary is though that general or sequence-specific binders of DNA could be difficult to display on the bacterial platform.

Fluorescence-activated MBD–DNA binding assay. In contrast to the kinetic screening of protein–ligand interactions commonly used in conjunction with surface display platforms, a thermodynamically controlled, competitive screening assay was established in this work. This assay could be sufficiently optimized to conform with the display levels and requirements of MBD proteins. Importantly, DNA concentrations were chosen such that relevant low nanomolar binding selectivities could be determined both in a 'one-color' and a 'two-color' setup with high specificity and high sensitivity. Indeed, this screening setup has afforded a MeCP2 variant with nanomolar binding affinity comparable to the wildtype protein but with reversed binding selectivity for 5mC/5mC and 5hmC/5mC dyads. In the future, these protocols could serve as a blueprint for characterizing other protein–DNA interactions.

10.2 Providing a framework to characterize MBD–DNA binding specificities

Although the first member of the MBD protein family was discovered more than 30 years ago (Meehan, et al., 1989) and many studies have characterized the various members of the family using different techniques and probes, a systematic compilation of this data and a unified approach that comprehensively characterized the different domains has been missing.

With this work, I provide a summary on the current knowledge about the common and distinct properties of individual MBDs as reported in the literature and along with this a compendium of biochemical parameters such as their binding affinity towards different modified and unmodified DNA duplexes. This motivated to close some gaps for missing or ambiguous pieces of data. In particular, I have contributed a comprehensive study on five human MBDs for their binding specificity towards different combinations of modified and unmodified cytosines in CpG dyads (Buchmuller, et al., 2020). This study has revealed remarkable differences between the different members of this overall conserved domain. Also, a previously overlooked interaction of MBD3 with 5caC/5caC CpG dyads could be identified which might be of some biological relevance.

10.3 Engineering MBDs with novel DNA binding selectivity

This work presents the first man-made proteins that simultaneously recognize strand-asymmetrically modified cytosines in single CpG dyads in double-stranded DNA. Importantly, they are not only the first proteins that engage with *two different* modified cytosines in a DNA duplex, but also some of the rare examples for an engineered protein that employs a ‘positive mode’ of recognition towards modified DNA nucleobases (Liu, et al., 2020; Maurer, et al., 2018; Zhang, et al., 2017) while ‘negative’ (Kubik, et al., 2015; Maier, et al., 2017) or ‘neutral modes’ (Gieß, et al., 2018; Maurer, et al., 2016; Tam, et al., 2016) are more commonly found, probably because it is easier to obliterate an interaction than to create one.

All engineered proteins in this work could be derived from the MBD of MeCP2 and were retrieved from a screen of degenerated MeCP2 variants, a natural protein that engages with strand-symmetrically modified 5mC/5mC CpG dyads. Of course, there is no guarantee that an (artificial) evolutionary trajectory from the wildtype to a protein with the desired function exists. In light of the sequence conservation and variation observed within the five human MBDs and related homologs, this might have seemed as improbable as possible. Many positions in the primary sequence of the MBD are highly conserved and of structural importance for the domain or its specific engagement with CpG dyads. Our analysis of disease-associated (MeCP2-)MBD mutants confirmed that single amino acid substitutions at these sites had detrimental effects especially for their engagement with oxidized 5-methylcytosines (Buchmuller, et al., 2020). For example, MeCP2[S134C] replaces a single oxygen atom with a sulfur atom and lost more than 50-fold in relative affinity for 5hmC/5hmC and CpG dyads with higher oxidized cytosines. On the other hand, the present substitutions at less conserved positions in the primary sequence did not affect the binding preference for fully methylated 5mC/5mC CpG dyads over any other combination of oxidized 5-methylcytosine or unmodified cytosines (with a single non-binding exception). However, the ensuing subtle three-dimensional structural differences in the five domains lead to remarkably different affinities towards non-5mC/5mC CpG dyads, suggesting that trajectories exist to ‘tune’ the binding selectivity of the domain.

On the basis of the data presented from the high-throughput screening of the degenerated MeCP2 domain at four positions in close vicinity to the CpG dyad in the DNA binding site, several (short) trajectories towards the accommodation of 5hmC/5mC, 5caC/5mC and 5caC/5caC existed for which common substitution patterns could be determined. These often involved serine-134, a residue which was also replaced in the more closely characterized 5hmC/5mC-selective mutant MeCP2[K109T/V122A/S134N] and found in vicinity of the 5hmC nucleobase in the CpG dyad (DEER measurements in collaboration with J. Dröden and Prof. Dr. M. Drescher, University of Konstanz). The same position was substituted with arginine in 5caC/5mC-selective MeCP2 mutants, offering in both cases the possibility for direct molecular interactions via hydrogen bonding or electrostatic interactions.

In the MeCP2[K109T/V122A/S134N] mutant, two known mechanisms to engage with 5hmC and 5mC nucleobases are likely deployed based on the currently available data. The methyl group of 5mC could take place like in the wildtype via CH···O hydrogen bonding involving potentially a structured water, while the hydroxyl group of 5hmC might be engaged via a carbonyl hydrogen bonding as reported for the base-flipper UHRF2. Further, the Ala-122 substitution was critical for the selective interaction with 5hmC/5mC over 5mC/5mC CpG dyads. Although the structural basis remains to be elucidated, one can already indulge in speculation. Since the residue resides at some distance to the dyad, it could be involved in shape recognition of the altered DNA double-helix or it could involve the correct positioning of the strand β3 relative to the Asx-ST motif and the helix α1. Probably, interactions around this Asx-ST motif with the backbone DNA double-strand contributed in addition to the determinants at the DNA binding site to a preferred orientation of wildtype and mutant MeCP2 as revealed by DEER. Only if both requirements were sufficiently satisfied in a given DNA sequence context, modification-specific low nanomolar binding affinities could be observed. In some sequence contexts, modified cytosines in a single CpG dyad could therefore contribute more or less strongly to the overall affinity with which a single DNA duplex is occupied.

This finding has probably important implications for the detection and discrimination of individual modified CpG dyads in a DNA duplex, both for the biological role of wildtype MeCP2 in an organism and for the technological application of MeCP2 mutants to reveal genomic sites. In the latter case, it might be necessary to carry out such assays under stringent binding conditions and/or at a limiting concentration of the MBD to preferentially engage with those CpG dyads that contain the desired modified cytosines. Although a differential enrichment using wildtype and MeCP2[K109T/V122A/S134N] in separate reactions would need particular analytical focus on correcting (McCarthy, et al., 2016) the wildtype's binding of CpA dinucleotides which is almost absent in the reported mutants, a venue to control spurious binding in non-CpG contexts could be to simultaneously apply the wildtype and mutant MBD in a mixture in which only the mutant carries a second affinity tag to purify the DNA–MBD complexes that contain the desired CpG dyads. If in the future a higher binding affinity was needed for one of the variants, then a concatemeric MBD could be used (Jørgensen, et al., 2006). Given that the presented MeCP2[K109T/V122A/S134N] mutant has similar but reversed binding affinity as the successfully commercialized wildtype MeCP2 in kits for the genomic enrichment of fully methylated CpGs (Kangaspeska, et al., 2008), a biotechnological application seems possible.

Irrespective of additional improvements that potentially benefit the performance of the binders, the creation of MBD variants with novel selectivity towards individual combinations of modified cytosines in particular to strand-asymmetrically modified CpG dyads provided already exciting insights into the requirements for implementing such interactions on the confined interaction surface on the DNA major groove. Similarly exciting discoveries for other combinations lay ahead in the future.

Chapter 11
Future development and outlook

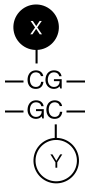
The examination of the MeCP2[K109T/V122A/S134N]–DNA complex at the biochemical and molecular level fostered our understanding of principles that guide such recognitions by means of different molecular interactions in a single CpG dyad on the DNA double-strand. Additional MBD variants with other binding specificities would therefore add valuable information to complete our understanding of strand-symmetric and strand-asymmetric recognition. Whether or not an MBD variant exists with binding specificity for any arbitrary combination of modified cytosines, e. g., one that involves 5fC, can only be determined by experimentation. Potentially, a more gradual directed evolutionary engineering approach with less substitutions as have been tested here could be helpful. In such an approach promiscuous variants could be considered as intermediates of an evolutionary trajectory towards more selective variants. The site-saturated screening of additional sites in the secondary shell of the DNA binding site could also be essential towards this goal (such as Val-122 was an essential substitution for 5hmC/5mC specificity). To identify these sites in the first place however, probably a random search in the sequence space using error-prone PCR could be necessary. Motivated by this initial work, such screenings are a logical next step.

By similar means to different ends, more selective variants or variants that are less dependent on additional sequence contexts could be screened for. Also an MeCP2 domain that retained 5mC/5mC CpG-selectivity but had significantly reduced affinity for CpA, 5mCpA and 5hmCpA dinucleotides would be an invaluable asset to investigate the molecular underpinnings of Rett syndrome (Tillotson, et al., 2021) and its contribution to physiological function in neurons (Ibrahim, et al., 2021). Even the use of the cell surface display system for proteins other than the MBD seems possible and could reveal maybe even natural binders of modified CpG dyads.

In general, binders that selectively recognize specific combinations of modified cytosines in CpG dyads would enable us for the first time to directly examine the distribution of these sites in the genomic DNA and thereby contribute largely to the elucidation of their biological role. Very likely, they could be involved in transcription and chromatin regulation (Iurlaro, et al., 2013; Spruijt, et al., 2013), serving either as distinct regulatory signals or creating 'poised' sites which can recruit cellular factors that engage with one of both modifications. However, such undertaking must also be guided by the fact that modified cytosines are rare constituents of genomic DNA even in neural tissue and certain combinations hence even rarer. It would therefore not come as a surprise to see such engineered binders first being used in a targeted manner (Lungu, et al., 2017) or in combination with ultra-sensitive chromatin enrichment strategies (Kaya-Okur, et al., 2019). As an alternative to these physiological systems, one could consider transient TET overexpression *in vivo* or the examination of TET oxidation *in vitro* where distinct combinations could be more frequent (Kizaki & Sugiyama, 2014). In these experiments, the engineered MBDs with different binding specificity would block a CpG dyad that has acquired a certain combination of modified cytosines from further modification and thereby reveal the processivity of the enzyme for an individual target. In contrast to chemical labeling strategies this protein-based detection could be carried out simultaneously to the oxidation reaction under physiologically relevant conditions.

Since modified cytosine nucleobases have been linked to a number of diseases including neurodevelopmental disorders and cancer, also a biomedical, diagnostic application of engineered MBD variants is conceivable. Particularly, of course, in cases in which distinct combinations of modified cytosines at a genomic locus would be relevant. An area that is still underexplored. In this context it is noteworthy that MBDs have been part of procedures with purified or pre-enriched genomic DNA in real-time single-molecule detection systems for epigenomics (Cipriany, et al., 2012; Yu, et al., 2010).

In summary, this work presents the first MBDs with a strand-asymmetric DNA binding mode and offers valuable insights into the molecular recognition of distinct combinations of modified cytosines in CpG dyads. It also serves as a methodological blueprint for the screening of further MBD variants or other DNA-binding proteins with respect to the recognition of modified DNA nucleobases. Beyond this, the engineered MBDs presented in this work can be used as molecular probes to decipher previously hardly accessible information about combinations of DNA modification in the single DNA double-helix in native chromatin and thus open an exciting way to study their epigenetic role in the future.



Resources

Chapter 12

Materials

12.1 Hosts and vectors

All bacterial strains used in this work are non-enteropathogenic *Escherichia coli* (Castellani and Chalmers, 1919) B or K-12 isolates (**Table 12.1**). The genetic materials transformed into the recipient hosts conformed with biosafety level 1 regulations.

Table 12.1 Bacterial strains.

Strain	Genotype, Origin and Notes
BL21-Gold(DE3)	F^- *hsdS*$_B$ (r_B^- m_B^-) *gal dcm*$^+$ *endA1 ompT* λ(DE3) Tet^r Hte Agilent Technologies (Waldbronn, Germany), cat. no. 230132 *NOTES:* B isolate engineered for high transformation efficiency and high protein yields with T7 RNA polymerase; Lacks the Lon and the OmpT protease; Resistant to tetracycline. The Dcm methylase, naturally missing in *E. coli* B, is inserted.
Tuner™(DE3)	F^- *hsdS*$_B$ (r_B^- m_B^-) *gal dcm ompT lacY1*(DE3) Novagen™ Merck (Darmstadt, Germany), cat. no. 70623. *NOTES:* BL21 derivate for titratable exogenous expression of proteins due to *lacZY* deletion breaking the positive feedback regulation within the *lac* operon.
DH5α	F^- *hsdR17* (r_K^- m_K^+) φ80*lacZ*ΔM15 Δ(*lacZYA-argF*)U169 *recA1 endA1 relA1 phoA supE44 thi-1 gyrA96* λ^- Invitrogen™ Thermo Fisher Scientific (Schwerte, Germany), cat. no. 18265017. *NOTES:* K-12 isolate that prevents unwanted recombination of transformed DNA due to *recA1*; Cells are sensitive to UV irradiation; *gyrA96* confers resistance to CcdB colicin.
DH10B (TOP10™)	F^- *mcrA* Δ(*mrr-hsdRMS-mcrBC*) φ80*lacZ*ΔM15 Δ(*lac*)X74 *recA1 endA1 araD139* Δ(*araA-leu*)7697 *galU galK rpsL*(Str^r) *nupG* Invitrogen™ Thermo Fisher Scientific (Schwerte, Germany), cat. no. C404003. *NOTES:* K-12 isolate for uptake of large plasmids since deoxyribose is constantly synthesized; Used with recombinant mammalian and plant DNA and during library construction of such. Titratable *araBCD*$^-$ phenotype since *araBA* deleted and *araC* and *araD* have inactivating point mutation.
GH371	F^- *mcrA* Δ(*mrr-hsdRMS-mcrBC*) φ80*lacZ*ΔM15 Δ(*lac*)X74 *recA1 endA1 araD139* Δ(*araA-leu*)7697 *galU galK rpsL*(Str^r) *nupG fhuA::IS2 upp*$^-$ Obtained from J. W. Chin. *NOTES:* Derived from DH10B; Resistant to 5-fluorouracil, allows negative genetic selection.

Table 12.2a Plasmids.

Identifier	Purpose	Gene*	Backbone	Marker
p1680	Template		pBAD33.1	Cm^r
p1733	Template		pBluescript SK(+)	Amp^r
p2606	Other	His$_6$–MBP–TEV	pOPIN	Amp^r
p1379	Entry (expression)	MBP–His$_6$	pET21d(+)	Amp^r
p1380	Entry (expression)	SpA(Z)–His$_6$	pET21d(+)	Amp^r
p1705	Entry (expression)	SP–AIDA^C	pBAD33.1	Cm^r

Table 12.2b Plasmids.

Identifier	Purpose	Gene*	Backbone	Marker
p1780	Entry (expression)	SpA(Z)–His$_6$	pET21d(+)	Cm^r
p1785	Entry (expression)	MBP–His$_6$	pET21d(+)	Cm^r
p2720	Entry (expression)	GST–His$_6$	pGEX-6P-1	Amp^r
p1383	Entry (surface display)	SP–AIDA^C	pET21d(+)	Amp^r
p1384	Expression	MBP–MBD1–His$_6$	pET21d(+)	Amp^r
p1385	Expression	MBP–MBD2–His$_6$	pET21d(+)	Amp^r
p1386	Expression	MBP–MBD3–His$_6$	pET21d(+)	Amp^r
p1387	Expression	MBP–MBD4–His$_6$	pET21d(+)	Amp^r
p1388	Expression	MBP–MeCP2–His$_6$	pET21d(+)	Amp^r
p1389	Expression	SpA(Z)–MBD1–His$_6$	pET21d(+)	Amp^r
p1390	Expression	SpA(Z)–MBD2–His$_6$	pET21d(+)	Amp^r
p1391	Expression	SpA(Z)–MBD3–His$_6$	pET21d(+)	Amp^r
p1392	Expression	SpA(Z)–MBD4–His$_6$	pET21d(+)	Amp^r
p1393	Expression	SpA(Z)–MeCP2–His$_6$	pET21d(+)	Amp^r
p1781	Expression hit	SpA(Z)–MeCP2[K109S/V122I/Y123T/S134N]–His$_6$	pET21d(+)	Amp^r
p1782	Expression hit	SpA(Z)–MeCP2[K109T/V122T/Y123T/S134K]–His$_6$	pET21d(+)	Amp^r
p1783	Expression hit	SpA(Z)–MeCP2[K109T/V122A/S134N]–His$_6$	pET21d(+)	Amp^r
p1856	Expression hit	MBP–MeCP2[K109T/V122T/Y123T/S134K]–His$_6$	pET21d(+)	Amp^r
p1857	Expression hit	SpA(Z)–MeCP2[K109T/V122A/S134N]–His$_6$	pET21d(+)	Cm^r
p1859	Expression hit	MBP–MeCP2[K109T/V122A/S134N]–His$_6$	pET21d(+)	Cm^r
p2080	Expression hit	MBP–MeCP2[V122C/Y123S/S134Q]–His$_6$	pET21d(+)	Amp^r
p2081	Expression hit	MBP–MeCP2[K109A/V122C/Y123F/S134R]–His$_6$	pET21d(+)	Amp^r
p2082	Expression hit	MBP–MeCP2[V122L/Y123T/S134R]–His$_6$	pET21d(+)	Amp^r
p2083	Expression hit	MBP–MeCP2[K109V/Y123D/S134R]–His$_6$	pET21d(+)	Amp^r
p2084	Expression hit	MBP–MBD2[V164R/S183V]–His$_6$	pET21d(+)	Amp^r
p2526	Expression hit	MBP–MeCP2[K109T/V122C/S134N]–His$_6$	pET21d(+)	Amp^r
p2527	Expression hit	MBP–MeCP2[K109T/V122T/Y123Q/S134K]–His$_6$	pET21d(+)	Amp^r
p2528	Expression hit	MBP–MeCP2[K109A/V122L/Y123M/S134R]–His$_6$	pET21d(+)	Amp^r
p1642	Expression mutant	SpA(Z)–MeCP2[S134C]–His$_7$	pET21d(+)	Amp^r
p1643	Expression mutant	SpA(Z)–MeCP2[L124F]–His$_7$	pET21d(+)	Amp^r
p1644	Expression mutant	SpA(Z)–MeCP2[R133C]–His$_7$	pET21d(+)	Amp^r
p1645	Expression mutant	SpA(Z)–MeCP2[T158M]–His$_7$	pET21d(+)	Amp^r
p2090	Expression mutant	MBP–MeCP2[R133A]–His$_6$	pET21d(+)	Amp^r
p2091	Expression mutant	MBP–MeCP2[R111A/R133A]–His$_6$	pET21d(+)	Amp^r
p2469	Expression mutant	MBP–MBD2[R22A/R44A]–His$_6$	pET21d(+)	Amp^r
p2470	Expression mutant	MBP–MBD2[R44A]–His$_6$	pET21d(+)	Amp^r
p2525	Expression mutant	MBP–MeCP2[K109T/V122C/S134T]–His$_6$	pET21d(+)	Amp^r
p2529	Expression mutant	MBP–MeCP2[Y123M]–His$_6$	pET21d(+)	Amp^r

Table 12.2c Plasmids.

Identifier	Purpose	Gene*	Backbone	Marker
p2603	Expression mutant	MBP–MBD2[V164T/V177A/S189N]–His$_6$	pET21d(+)	Amp^r
p2623	Expression mutant	MBP–MBD2[V164T]–His$_6$	pET21d(+)	Amp^r
p2635	Expression mutant	MBP–MBD2[V164T/V177A/S189N]–His$_6$	pET21d(+)	Amp^r
p2636	Expression mutant	MBP–MBD2[S189N]–His$_6$	pET21d(+)	Amp^r
p2653	Expression mutant	MBP–MeCP2[K109T/V122A/S134N/T158M]–His$_6$	pET21d(+)	Cm^r
p2654	Expression mutant	MBP–MeCP2[T158M]–His$_6$	pET21d(+)	Amp^r
p2657	Expression mutant	MBP–MeCP2[K109T/V122A/S134N/ΔTVTG]–His$_6$	pET21d(+)	Cm^r
p2658	Expression mutant	MBP–MeCP2[ΔTVTG]–His$_6$	pET21d(+)	Amp^r
p2668	Expression mutant	MBP–MeCP2[K109T/V122A/S134N/ΔLDPND]–His$_6$	pET21d(+)	Cm^r
p2669	Expression mutant	MBP–MeCP2[ΔLDPND]–His$_6$	pET21d(+)	Amp^r
p2687	Expression mutant	MBP–MeCP2[K109T/V122I/S134N]–His$_6$	pET21d(+)	Cm^r
p2688	Expression mutant	MBP–MeCP2[K109T/V122L/S134N]–His$_6$	pET21d(+)	Amp^r
p2689	Expression mutant	MBP–MeCP2[K109T/V122G/S134N]–His$_6$	pET21d(+)	Cm^r
p2706	Expression mutant	MBP–MeCP2[K109T/S134N]–His$_6$	pET21d(+)	Cm^r
p2573	Expression variant	MBP–MeCP2[A117C]–His$_6$	pET21d(+)	Amp^r
p2574	Expression variant	MBP–MeCP2[A140C]–His$_6$	pET21d(+)	Amp^r
p2575	Expression variant	MBP–MeCP2[G161C]–His$_6$	pET21d(+)	Amp^r
p2576	Expression variant	MBP–MeCP2[G146C]–His$_6$	pET21d(+)	Amp^r
p2577	Expression variant	MBP–MeCP2[K109T/V122A/S134N/A117C]–His$_6$	pET21d(+)	Amp^r
p2578	Expression variant	MBP–MeCP2[K109T/V122A/S134N/A140C]–His$_6$	pET21d(+)	Amp^r
p2579	Expression variant	MBP–MeCP2[K109T/V122A/S134N/G161C]–His$_6$	pET21d(+)	Amp^r
p2580	Expression variant	MBP–MeCP2[K109T/V122A/S134N/G146C]–His$_6$	pET21d(+)	Amp^r
p1727	Library	SP–MeCP2[K109X/V122X/Y123X/S134X]–AIDA^C	pET21d(+)	Amp^r
p1728	Library	SP–MBD2[V164X/V177X/Y178X/S189X]–AIDA^C	pET21d(+)	Amp^r
p1938	Library	SP–MBD1[V20X/T33X/Y34X/S45X]–AIDA^C	pET21d(+)	Amp^r
p1941	Library	SP–MBD3[V20X/V33X/Y35X/S45X]–AIDA^C	pET21d(+)	Amp^r
p1944	Library	SP–MBD4[K82X/K108X/K109X/K120X]–AIDA^C	pET21d(+)	Amp^r
p2594	Library	SP–MeCP2[K109X/Q110X/S113X/S116X]–AIDA^C	pET21d(+)	Amp^r
p2596	Library	SP–MeCP2[Y120X/V122X/Y123Φ/F132X/S134X]–AIDA^C	pET21d(+)	Amp^r
p2598	Library	SP–MeCP2[S116X/G118X/K119X/Y120X]–AIDA^C	pET21d(+)	Amp^r
p1566	Surface display	SP–MBD1–AIDA^C	pET21d(+)	Amp^r
p1567	Surface display	SP–MBD2–AIDA^C	pET21d(+)	Amp^r
p1568	Surface display	SP–MBD3–AIDA^C	pET21d(+)	Amp^r
p1569	Surface display	SP–MBD4–AIDA^C	pET21d(+)	Amp^r
p1570	Surface display	SP–MeCP2–AIDA^C	pET21d(+)	Amp^r
p1706	Surface display	SP–MBD1–AIDA^C	pBAD33.1	Cm^r
p1707	Surface display	SP–MBD2–AIDA^C	pBAD33.1	Cm^r
p1708	Surface display	SP–MBD3–AIDA^C	pBAD33.1	Cm^r

Table 12.2d Plasmids.

Identifier	Purpose	Gene*	Backbone	Marker
p1709	Surface display	SP–MBD4–AIDA^{C}	pBAD33.1	Cm^{r}
p1710	Surface display	SP–MeCP2–AIDA^{C}	pBAD33.1	Cm^{r}
p2370	Surface display	SP–MeCP2[R111A/R133A]–AIDA^{C}	pET21d(+)	Amp^{r}
p1777	Surface display hit	SP–MeCP2[K109S/V122I/Y123T/S134N]–AIDA^{C}	pET21d(+)	Amp^{r}
p1778	Surface display hit	SP–MeCP2[K109T/V122T/Y123T/S134K]–AIDA^{C}	pET21d(+)	Amp^{r}
p1779	Surface display hit	SP–MeCP2[K109T/V122A/S134N]–AIDA^{C}	pET21d(+)	Amp^{r}
p2060	Surface display hit	SP–MeCP2[K109D/V122C/S134N]–AIDA^{C}	pET21d(+)	Amp^{r}
p2061	Surface display hit	SP–MeCP2[K109R/V122Y/Y123V/S134H]–AIDA^{C}	pET21d(+)	Amp^{r}
p2062	Surface display hit	SP–MBD2[V164R/V177F/S183V]–AIDA^{C}	pET21d(+)	Amp^{r}
p2064	Surface display hit	SP–MeCP2[V122C/Y123S/S134Q]–AIDA^{C}	pET21d(+)	Amp^{r}
p2065	Surface display hit	SP–MeCP2[K109A/V122C/Y123F/S134R]–AIDA^{C}	pET21d(+)	Amp^{r}
p2066	Surface display hit	SP–MeCP2[V122L/Y123T/S134R]–AIDA^{C}	pET21d(+)	Amp^{r}
p2073	Surface display hit	SP–MeCP2[K109V/Y123D/S134R]–AIDA^{C}	pET21d(+)	Amp^{r}
p2644	Surface display hit	SP–MeCP2[K109T/V122T/Y123Q/S134K]–AIDA^{C}	pET21d(+)	Amp^{r}
p2621	Surface display mutant	SP–MBD2[V164T]–AIDA^{C}	pET21d(+)	Amp^{r}
p2633	Surface display mutant	SP–MBD2[V164T/V177A/S189N]–AIDA^{C}	pET21d(+)	Amp^{r}
p2634	Surface display mutant	SP–MBD2[S189N]–AIDA^{C}	pET21d(+)	Amp^{r}
p2637	Surface display mutant	SP–MeCP2[K109T/V122I/S134N]–AIDA^{C}	pET21d(+)	Amp^{r}
p2638	Surface display mutant	SP–MeCP2[K109T/V122L/S134N]–AIDA^{C}	pET21d(+)	Amp^{r}
p2639	Surface display mutant	SP–MeCP2[K109T/V122G/S134N]–AIDA^{C}	pET21d(+)	Amp^{r}
p2462	Application	SpA(Z)–MBD2-MNase–His$_{6}$	pET21d(+)	Amp^{r}
p2483	Application	MBP–MBD2-MNase–His$_{6}$	pET21d(+)	Amp^{r}
p2729	Application	GST–MeCP2–His$_{6}$	pGEX-6P-1	Amp^{r}
p2730	Application	GST–MeCP2[K109T/V122A/S134N]–His$_{6}$	pGEX-6P-1	Amp^{r}
p2731	Application	GST–MeCP2[R111A/R133A]–His$_{6}$	pGEX-6P-1	Amp^{r}
p2774	Application	MBP–MeCP2–GST-His$_{6}$	pET21d(+)	Amp^{r}
p2775	Application	MBP–MeCP2[K109T/V122A/S134N]–GST-His$_{6}$	pET21d(+)	Cm^{r}
p2776	Application	MBP–MeCP2[R111A/R133A]–GST-His$_{6}$	pET21d(+)	Amp^{r}

* SP = signal peptide

12.2 Oligonucleotides and probes

Oligonucleotides were synthesized by Merck KGaA (Darmstadt, Germany) or by metabion GmbH (Planegg/Steinkirchen, Germany) if they contained oxidized 5-methylcytosines. The desalted or for spectroscopic measurements HPLC-purified oligonucleotides were stocked at 100 μM in ultrapure water. Gene fragments were purchased from Integrated DNA Technologies (Leuven, Belgium) or GENEWIZ Germany GmbH (Leipzig, Germany).

Table 12.3a Oligonucleotides for DNA duplex probes.

Identifier	Sequence	Modifications
o2968	AAAAAAAAAAAAAAAAAAAAAAAA	
o2969	TTTTTTTTTTTTTTTTTTTTTTTT	
o3416	AAAAAAAAAAADHAAAAAAAAAAA	
o3417	TTTTTTTTTTTDHTTTTTTTTTTT	
o2906	AAAAAAAAAAACGAAAAAAAAAAA	5'-FAM
o2967	AAAAAAAAAAAXGAAAAAAAAAAA	5'-FAM; X = 5mC
o3115	AAAAAAAAAAAXGAAAAAAAAAAA	5'-FAM; X = 5hmC
o3116	AAAAAAAAAAAXGAAAAAAAAAAA	5'-FAM; X = 5fC
o3117	AAAAAAAAAAAXGAAAAAAAAAAA	5'-FAM; X = 5caC
o2907	TTTTTTTTTTTCGTTTTTTTTTTT	5'-FAM
o2904	TTTTTTTTTTTCGTTTTTTTTTTT	
o2909	TTTTTTTTTTTXGTTTTTTTTTTT	X = 5mC
o3112	TTTTTTTTTTTXGTTTTTTTTTTT	X = 5hmC
o3113	TTTTTTTTTTTXGTTTTTTTTTTT	X = 5fC
o3114	TTTTTTTTTTTXGTTTTTTTTTTT	X = 5caC
o2905	AAAAAAAAAAACGAAAAAAAAAAA	5'-Pacific Blue
o2908	AAAAAAAAAAACGAAAAAAAAAAA	5'-TAMRA
o3082	TTTTTTTTTTTXGTTTTTTTTTZT	5'-Btn-TEG; X = 5mC, Z = Btn-dT
o3083	TZTTTTTTTTTXGTTTTTTTTTZT	5'-Btn-TEG; X = 5mC, Z = Btn-dT
o3244	AAAAAAAAAAACGAAAAAAAAAAA	5'-Btn
o3081	AAAAAAAAAAAXGAAAAAAAAAAA	5'-Btn-TEG; X = 5mC
o3211	AAAAAAAAAAAXGAAAAAAAAAAA	5'-Btn; X = 5hmC
o3212	AAAAAAAAAAAXGAAAAAAAAAAA	5'-Btn; X = 5fC
o3213	AAAAAAAAAAAXGAAAAAAAAAAA	5'-Btn; X = 5caC
o3245	TTTTTTTTTTTCGTTTTTTTTTTT	5'-Btn-TEG
o3214	TTTTTTTTTTTXGTTTTTTTTTTT	5'-Btn; X = 5mC
o3215	TTTTTTTTTTTXGTTTTTTTTTTT	5'-Btn; X = 5hmC
o3216	TTTTTTTTTTTXGTTTTTTTTTTT	5'-Btn; X = 5fC
o3217	TTTTTTTTTTTXGTTTTTTTTTTT	5'-Btn; X = 5caC
o2903	AAAAAAAAAAACGAAAAAAAAAAA	
o4345	AAAAAAAAAAAXGAAAAAAAAAAA	X = 5hmC
o4277	AAAAAAAAAAATGAAAAAAAAAAA	5'-FAM
o4278	TTTTTTTTTTTXATTTTTTTTTTT	X = 5mC
o4279	TTTTTTTTTTTCATTTTTTTTTTT	
o4839	TTTTTTTTTTTXATTTTTTTTTTT	X = 5hmC
o4850	AAAAAAAAAAAXAAAAAAAAAAAA	X = 5hmC
o4328	TTTTTTTTTTTCGTTTT*TTTTTTT	5'-FAM; * = 3'-5' phosphorothioate
o4329	AAAAAAAAAAACGAAAA*AAAAAAA	5'-FAM; * = 3'-5' phosphorothioate

Table 12.3b Oligonucleotides for DNA duplex probes.

Identifier	Sequence	Modifications
o1516	CTTCCTCTTCCGTCTCTTTCCTTTTACGTCATCCGGGGGCAGACT	
o1529	AGTCTGCCCCCGGATGACGTAAAAGGAAAGAGACGGAAGAGGAAG	
o4497	AGTCTGCCCCCGGATGACGTAAAAGGAAAGAGAXGGAAGAGGAAG	5'-FAM; X = 5mC
o1517	CTTCCTCTTCXGTCTCTTTCCTTTTACGTCATCCGGGGGCAGACT	X = 5mC
o1520	CTTCCTCTTCXGTCTCTTTCCTTTTACGTCATCCGGGGGCAGACT	X = 5hmC
o4498	AGTCTGCCCCCGGATGAXGTAAAAGGAAAGAGACGGAAGAGGAAG	5'-FAM; X = 5mC
o1518	CTTCCTCTTCCGTCTCTTTCCTTTTAXGTCATCCGGGGGCAGACT	X = 5mC
o1521	CTTCCTCTTCCGTCTCTTTCCTTTTAXGTCATCCGGGGGCAGACT	X = 5hmC
o1591	GGCCAGCCAGTCAGCCGAAGGCTCCATGCTGCTCCCCGCCGCCGGC	
o1617	GCCGGCGGCGGGGAGCAGCATGGAGCCTTCGGCTGACTGGCTGGCC	
o4499	GCCGGCGGCGGGGAGCAGCATGGAGCCTTXGGCTGACTGGCTGGCC	5'-FAM; X = 5mC
o1592	GGCCAGCCAGTCAGCXGAAGGCTCCATGCTGCTCCCCGCCGCCGGC	X = 5mC
o1593	GGCCAGCCAGTCAGCXGAAGGCTCCATGCTGCTCCCCGCCGCCGGC	X = 5hmC
o476	TGGATTCCCACTCTTCAGCCCCAGCGTTACAGCATCTTCAGTGGCTTCTTCCACCTGAGCTCTTCCGTTTCCACATCC	
o1152	GGATGTGGAAACGGAAGAGCTCACGGTGGAAGAAGCCACTGAAGATGCTGTAACGTGGGGCTGAAGAGTGGGAATCCA	
o4379	GGAAAXGGAAGA	5'-FAM; X = 5mC
o517	TGGATTCCCACTCTTCAGCCCCAGCGTTACAGCATCTTCAGTGGCTTCTTCCACCTGAGCTXTTCXGTTTCCACATCC	X = 5mC
o520	TGGATTCCCACTCTTCAGCCCCAGCGTTACAGCATCTTCAGTGGCTTCTTCCACCTGAGCTCTTCXGTTTCCACATCC	X = 5hmC
o4823	GGATGTGGAAAXGGAAGAGCTCACGGTGGAAGAAGCCACTGAAGATGCTGTAACGTGGGGCTGAAGAGTGGGAATCCA	5'-FAM; X = 5mC
o4524	GATGAXGTAAAGTTTTCTTTAZGTCATC	X = 5mC, Z = 5hmC
o4687	GATGAXGTAAAGTTTTCTTTAZGTCATC	X = 5mC, Z = 5mC
o4552	GATGACGTAAAGTTTTCTTTACGTCATC	

Table 12.4a Oligonucleotides for DNA sequencing.

Identifier	Sequence	Purpose
o315	CGTAGAGGATCGAGATC	Sanger sequencing: pET T7 promoter
o1590	CTAGTTATTGCTCAGCGG	Sanger sequencing: pET T7 terminator
o3177	CCAAGTCCTCTTCAGAAATGAGC	Sanger sequencing: c-Myc epitope
o1412	TCCATAAGATTAGCGGATC	Sanger sequencing: pBAD araBAD promoter
o40	TAATCTGTATCAGGCTG	Sanger sequencing: pBAD T1 terminator
o3861	CTTCCTGGCACGAGNNNNNNAGGGCTGGACCCGTA	NGS UMI: adapter
o3862	GAAACAGCTATGACNNNNNNTATGCGATCAACTCCACC	NGS UMI: adapter
o2363	ATCACGCTTCCTGGCACGAG	NGS barcoding: forward barcode J01
o2364	CGATGTCTTCCTGGCACGAG	NGS barcoding: forward barcode J02
o2365	TTAGGCCTTCCTGGCACGAG	NGS barcoding: forward barcode J03

Table 12.4b Oligonucleotides for DNA sequencing.

Identifier	Sequence	Purpose
o2366	TGACCACTTCCTGGCACGAG	NGS barcoding: forward barcode J04
o2367	ACAGTGCTTCCTGGCACGAG	NGS barcoding: forward barcode J05
o2368	GCCAATCTTCCTGGCACGAG	NGS barcoding: forward barcode J06
o2369	CAGATCCTTCCTGGCACGAG	NGS barcoding: forward barcode J07
o2370	ACTTGACTTCCTGGCACGAG	NGS barcoding: forward barcode J08
o3033	ATCACGGAAACAGCTATGAC	NGS barcoding: reverse barcode J01
o3034	CGATGTGAAACAGCTATGAC	NGS barcoding: reverse barcode J02
o3035	TTAGGCGAAACAGCTATGAC	NGS barcoding: reverse barcode J03
o3036	TGACCAGAAACAGCTATGAC	NGS barcoding: reverse barcode J04
o3037	ACAGTGGAAACAGCTATGAC	NGS barcoding: reverse barcode J05
o3038	GCCAATGAAACAGCTATGAC	NGS barcoding: reverse barcode J06
o3039	CAGATCGAAACAGCTATGAC	NGS barcoding: reverse barcode J07
o3040	ACTTGAGAAACAGCTATGAC	NGS barcoding: reverse barcode J08

Table 12.5a Oligonucleotides for cloning.

Identifier	Sequence	Purpose
o2872	CTTTAAGAAGGAGATATACATATGGACAACAAATTCAACAAAGAACACAAAACGC	Gibson: Z domain of SpA into pET
o2873	AGTGGTGGTGGTGGTGGTGCTCGAGAGACTGAAAATAAAGATTTTCAGCCTTCCCTCGATAGAACCACTGCCAGATCCCGCGTC	Gibson: Z domain of SpA into pET
o2879	AAATCTTTATTTTCAGTCTCTCGAGGCAGAAGACTGGTTGGACTG	Gibson: MBD1 into pET
o2880	GTGGTGGTGGTGGTGGTGACTAGTCACTGCTACGGGATGCG	Gibson: MBD1 into pET
o2882	AAATCTTTATTTTCAGTCTCTCGAGTCAGGCAAACGTATGGATTG	Gibson: MBD2 into pET
o2883	GTGGTGGTGGTGGTGGTGACTAGTCAAGCGTTGCTTATTCTTC	Gibson: MBD2 into pET
o2885	AAATCTTTATTTTCAGTCTCTCGAGGAACGCAAACGCTGGGAATG	Gibson: MBD3 into pET
o2886	GTGGTGGTGGTGGTGGTGACTAGTCACGCGTTGACGAGATTTATTC	Gibson: MBD3 into pET
o2888	AAATCTTTATTTTCAGTCTCTCGAGGCGACCGCAGGTACAGAG	Gibson: MBD4 into pET
o2889	GTGGTGGTGGTGGTGGTGACTAGTGAGATGGGATGTTAATGCTGCC	Gibson: MBD4 into pET
o2891	AAATCTTTATTTTCAGTCTCTCGAGGATCGTGGTCCTATGTATG	Gibson: MeCP2 into pET
o2892	GTGGTGGTGGTGGTGGTGACTAGTAGCTTTGGGTGATTTTGG	Gibson: MeCP2 into pET
o2893	CTCACGGCGTTTCCAACCTG	PCR: MBD1 linearization
o2894	AAGGTAGAATTGACCCGCTATC	PCR: MBD1 linearization
o2896	CTCCTCCTTTTTCCATCCGG	PCR: MBD2 linearization
o2897	GGCAAGAAGTTTCGCTCAAAAC	PCR: MBD2 linearization
o4440	AAACCACAGTTAGCACGTTATTTGG	PCR: MBD2 linearization
o2899	CAATTTACGGGTCCAGCCCTC	PCR: MeCP2 linearization
o2900	AAGGTGGAGTTGATCGCATAC	PCR: MeCP2 linearization
o3004	GGATGCATATGGTTAAATTAAAATTTGGTGTTTTTTTTAC	Gibson: AIDA-I into pET
o3005	CGATCGTCGACAAGCTTCAGAAGCTGTATTTTATCC	Gibson: AIDA-I into pET

Table 12.5b Oligonucleotides for cloning.

Identifier	Sequence	Purpose
o3110	AGGGCTGGACCCGTA	PCR: MeCP2 amplfiication
o3111	TATGCGATCAACTCCACC	PCR: MeCP2 amplfiication
o3118	AAGGGAAGGCATTTCGCTGTAAGGTGGAGTTGATC	QuikChange: Rett variant MeCP2[S134C]
o3119	GATCAACTCCACCTTACAGCGAAATGCCTTCCCTT	QuikChange: Rett variant MeCP2[S134C]
o3120	CGGCAAGTATGATGTGTATTTTATCAATCCCCAAGGGAA	QuikChange: Rett variant MeCP2[L124F]
o3121	TTCCCTTGGGGATTGATAAAATACACATCATACTTGCCG	QuikChange: Rett variant MeCP2[L124F]
o3122	CCCCAAGGGAAGGCATTTTGCAGTAAGGTGG	QuikChange: Rett variant MeCP2[R133C]
o3123	CCACCTTACTGCAAAATGCCTTCCCTTGGGG	QuikChange: Rett variant MeCP2[R133C]
o3124	CCTAATGACTTCGATTTTACCGTAATGGGGCGTGGAAGCC	QuikChange: Rett variant MeCP2[T158M]
o3125	GGCTTCCACGCCCCATTACGGTAAAATCGAAGTCATTAGG	QuikChange: Rett variant MeCP2[T158M]
o3173	CTCCTCGCGTTCCCATCCTTG	PCR: MBD3 linearization
o3174	AAACCACAGTTAGCGCGTTACC	PCR: MBD3 linearization
o3175	GACAACGCGCTCCCACCCAC	PCR: MBD4 linearization
o3176	AAGTCATCACTTGCTAACTAT	PCR: MBD4 linearization
o3178	TTTGAGCGAAACTTCTTGCC	PCR: MBD2 amplification
o3179	GGTAACGCGCTAACTGTG	PCR: MBD3 amplification
o3180	AAATAGTTAGCAAGTGATG	PCR: MBD4 amplification
o3006	GCCCAAGATAGCGGGTCAATTCTAC	PCR: MBD1 amplfication
o3894	GGTGAGCGTCGACGCAACATCCACGAAAAAACTTCATAAAG	PCR: MNase with downstream *Sal*I
o3895	CGTGAGCCTCGAGTTGCCCACTATCCGCGTTG	PCR: MNase with upstream *Xho*I
o4254	CAAATCCGGTCGCAGCTGCGGCAAGTATGATGTG	QuikChange: MeCP2[A117C]
o4255	CACATCATACTTGCCGCAGCTGCGACCGGATTTG	QuikChange: MeCP2[A117C]
o4256	CGCAGTAAGGTGGAGTTGATCTGCTACTTCGAGAAAGTGGGTGAT	QuikChange: MeCP2[A140C]
o4257	ATCACCCACTTTCTCGAAGTAGCAGATCAACTCCACCTTACTGCG	QuikChange: MeCP2[A140C]
o4258	TAATGACTTCGATTTTACCGTAACTTGCCGTGGAAGCCCTTC	QuikChange: MeCP2/mutants[G146C]
o4259	AAGGGCTTCCACGGCAAGTTACGGTAAAATCGAAGTCATTAG	QuikChange: MeCP2/mutants[G146C]
o4260	GCATACTTCGAGAAAGTGTGTGATACATCTCTGGACC	QuikChange: MeCP2/mutants[G161C]
o4261	GGTCCAGAGATGTATCACACACTTTCTCGAAGTATGC	QuikChange: MeCP2/mutants[G161C]
o4280	CAAATCCGGTCGCAGCTGCGGCAAGTATGATGCG	QuikChange: MeCP2(V122A)[A117C]
o4281	CGCATCATACTTGCCGCAGCTGCGACCGGATTTG	QuikChange: MeCP2(V122A)[A117C]
o4282	CGCAATAAGGTGGAGTTGATCTGCTACTTCGAGAAAGTGGGTGAT	QuikChange: MeCP2(S134N)[A140C]
o4283	ATCACCCACTTTCTCGAAGTAGCAGATCAACTCCACCTTATTGCG	QuikChange: MeCP2(S134N)[A140C]
o4402	GGAAAAAGGAGGAGGACCTTCGCAAGTCAGGCCTTAGTG	QuikChange: MBD2[V164T]
o4403	CACTAAGGCCTGACTTGCGAAGGTCCTCCTCCTTTTTCC	QuikChange: MBD2[V164T]
o4443	GTGGCAAGAAGTTTCGCAACAAACCACAGTTAGCACGT	QuikChange: MBD2[S189N]
o4444	ACGTGCTAACTGTGGTTTGTTGCGAAACTTCTTGCCAC	QuikChange: MBD2[S189N]
o4442	CCGGATGGAAAAAGGAGGAGACCATTCGCAAGTCAGGCCTTAGTGCGGCAAATCCGATGCGTATTATTTCTCACCTAGTGGCAAGAAGTTTCCAACAAACCACAGTTAGCACGTTATT	Gibson: MBD2[V164T/V177A/S189N]

Table 12.5c Oligonucleotides for cloning.

Identifier	Sequence	Purpose
o4619	GGGGATTGATAAGATACCCATCATACTTGCCGGCG	QuikChange: MeCP2[A122G]
o4620	CGCCGGCAAGTATGATGGGTATCTTATCAATCCCC	QuikChange: MeCP2[A122G]
o4621	GGGGATTGATAAGATACACATCATACTTGCCGGCG	QuikChange: MeCP2[A122V]
o4622	CGCCGGCAAGTATGATGTGTATCTTATCAATCCCC	QuikChange: MeCP2[A122V]
o4623	CCCTTGGGGATTGATAAGATATATATCATACTTGCCGGCGCTGCG	QuikChange: MeCP2[A122I]
o4624	CGCAGCGCCGGCAAGTATGATATATATCTTATCAATCCCCAAGGG	QuikChange: MeCP2[A122I]
o4625	CCTTGGGGATTGATAAGATATAGATCATACTTGCCGGCGCTGC	QuikChange: MeCP2[A122L]
o4626	GCAGCGCCGGCAAGTATGATCTATATCTTATCAATCCCCAAGG	QuikChange: MeCP2[A122L]
o4649	CCCTGGGATCCCCGGAATTCGAAGCTTCTGCCTCCCCCAAACAGCG CGCTCCATCATCCGTGATCGTGGTCCTATGTATGATGATC	PCR: MeCP2 in pGEX-6P-1
o4670	GTCAGTCACGATGCGGCCGCTGGTGGTGGTGGTGGTGGTGACTAGT GCTTTGGGTGATTTTGGC	PCR: MeCP2 in pGEX-6P-1
o4763	GCCAAAATCACCCAAAGCTAGCTCCCCTATACTAGGTTATTG	PCR: C-terminal GST
o4764	GGTGGTGGTGGTGGTGGTGAGCCGATTTTGGAGGATGGTC	PCR: C-terminal GST

Table 12.6a Gene fragments.

Identifier	Sequence	Gene
o2878	GCAGAAGACTGGTTGGACTGTCCAGCTTTAGGTCCAGGTTGGAAACGCCGTGAGG GTTTCGTAAGTCTGGTGCAACGTGCGGTCGCTCCGATACCTACTACCAGTCACCT CCGGTGACCGCATTCGCTCTAAGGTAGAATTGACCCGCTATCTTGGGCCGGCCTG GATCTGACCTTATTCGATTTCAAACAGGGTATCTTGTGTTACCCCGCGCCCAAAG GCATCCCGTAGCAGTG	MBD1[2–81], CCDS59318.1; codon-optimized for *E. coli*
o2881	TCAGGCAAACGTATGGATTGCCCCGCACTGCCTCCCGGATGGAAAAAGGAGGAGG AATTCGCAAGTCAGGCCTTAGTGCCGGCAAATCCGATGTATATTATTTCTCACCT GTGGCAAGAAGTTTCGCTCAAAACCACAGTTAGCACGTTATTTGGGTAATACGGT GACTTGTCCTCGTTCGATTTCCGTACCGGTAAAATGATGCCATCGAAATTACAGA GAATAAGCAACGCTTG	MBD2[145–225], CCDS11953.1; codon-optimized for *E. coli*
o2884	GAACGCAAACGCTGGGAATGCCCAGCTTTACCTCAAGGATGGGAACGCGAGGAGG TCCTCGCCGCTCAGGCCTTTCCGCCGGCCATCGTGACGTATTTTACTACTCGCCT CGGGAAAGAAATTCCGCTCAAAACCACAGTTAGCGCGTTACCTGGGAGGATCTAT GATTTGTCTACCTTCGACTTTCGCACCGGTAAAATGCTGATGAGCAAGATGAATA ATCTCGTCAACGCGTG	MBD3[2–83], CCDS12072.1; codon-optimized for *E. coli*
o2887	GCGACCGCAGGTACAGAGTGCCGTAAGTCAGTTCCCTGTGGGTGGGAGCGCGTTG CAAGCAACGCTTGTTTGGTAAAACAGCCGGCCGTTTCGATGTTTACTTTATCTCG CGCAAGGCCTAAAGTTCCGTTCCAAGTCATCACTTGCTAACTATTTACACAAAAA GGTGAAACCTCCCTTAAACCGGAAGACTTTGACTTCACTGTGTTAAGCAAGCGTG TATCAAAAGCCGCTACAAGGACTGCTCAATGGCAGCATTAACATCCCATCTC	MBD4[76–167], CCDS3058.1; codon-optimized for *E. coli*
o2890	GATCGTGGTCCTATGTATGATGATCCAACACTTCCTGAGGGCTGGACCCGTAAAT GAAGCAACGCAAATCCGGTCGCAGCGCCGGCAAGTATGATGTGTATCTTATCAAT CCCAAGGGAAGGCATTTCGCAGTAAGGTGGAGTTGATCGCATACTTCGAGAAAGT GGTGATACATCTCTGGACCCTAATGACTTCGATTTTACCGTAACTGGGCGTGGAA CCCTTCGCGTCGCGAGCAGAAACCCCCTAAAAAGCCAAAATCACCCAAAGCT	MeCP2[92–169], CCDS14741.1; codon-optimized for *E. coli*

Table 12.6b Gene fragments.

Identifier	Sequence	Gene
o2912	ATGGTTAAATTAAAATTTGGTGTTTTTTTTACAGTTTTACTATCTTCAGCCTATGACATGGAACACTCGAGGTAGCAGTGACTAGTGCGGAGGAGCAAAAGCTCATTTCTAAGAGGACTTGGGTACCCTTAATCCTACAAAAGAAAGTGCAGGTAATACTCTTACGTGTCAAATTATACTGGGACACCGGGAAGTGTTATTTCTCTTGGTGGTGTGCTTGAGGAGATAATTCACTTACGGACCGTCTGGTGGTGAAAGGTAATACCTCTGGTCAAGTGACATCGTTTACGTCAATGAAGATGGCAGTGGTGGTCAGACGAGAGATGGTATAACATTATTTCTGTAGAGGGAAATTCTGATGCAGAATTTTCTCTGAAGAACCGCGAGTTGCCGGAGCTTATGATTACACACTGCAGAAAGGAAACGAGAGTGGGACAGATATAAGGGATGGTATTTAACCAGTCATCTTCCCACATCTGATACCCGGCAATACAGCCGGAGAACGGAAGTTATGCTACCAATATGACACTGGCTAACTCACTGTTCCTCAGGATTTGAATGAGCGTAAGCAATTCAGGGCAATGAGTGATAATACACAGCCTGAACTGCATCCGTGTGGATGAGGATTACTGGAGGAAGAAGCTCTGGTAAACTTAATGAGGGCAAAATAAAACAACAACCAATCAGTTTATCAATCAGCTCGGGGGGGATATTTCAAATTCCATGCTGAACAACTGGGTGATTTTACCTTAGGGATTATGGGAGGATACCGAATGCAAAAGGTAAAACGATAAATTACACGAGCAACAAAGCTGCCAGAAACACCTGGATGGTTATTCTGTCGGGGTATATGGTACGTGGTATCAGAATGGGGAAAATGAACAGGGCTCTTTGCTGAAACTTGGATGCAATATAACTGGTTTAATGCCTCGGTGAAGGTGACGGACTGGAAGAAGAAAAATATAATCTGAATGGTTTAACCGCTTCTGCGGTGGGGGATATAACCTGAATGTGCACACATGGACATCACCTGAAGGAATAACAGTGAATTTTGGTTGCAGCCTCATTTGCAGGCTGTCTGGATGGGGGTTACACCGGATCACACCAGGAGGATAACGGAACGGTGGTGCAGGGAGCAGGGAAAAATAACATTCAACAAAAGCAGGTATTCGTGCATCCTGGAAGGTGAAAAGCACCCTGGATAAGGATACGGGCGGGAGTTCAGTCCGTATATAGAGGCAAACTGGATTCATAACACGCATGAATTGGTGTTAAAATGAGTGATGACAGCCAGTTGTTGTCAGGTAGCCGAAATCAGGGGAGATAAAGACAGGTATTGAAGGGGTGATTACTCAAAACTTGTCAGTGAATGGCGAGTCGCATATCAGGCAGGAGGTCACGGGAGCAATGCCATCTCGGGAGCACTGGGGTAAAATACAGCTTCTGA	AIDA-I surface display cassette with c-Myc epitope
o3893	GCAACATCCACGAAAAAACTTCATAAAGAACCTGCCACTCTTATTAAAGCTATTGTGGAGACACCGTTAAGTTAATGTACAAGGGCCAGCCCATGACGTTCCGCTTACTTTGGTGGACACCCCGGAAACAAAGCACCCAAAAAAAGGCGTAGAGAAGTACGGACCGAGGCGAGCGCGTTTACAAAAAAAATGGTGGAGAATGCAAAAAAAATCGAGGTTGGTTCGACAAAGGCCAACGCACGGACAAATATGGCCGTGGACTGGCCTACATCTATCGGATGGTAAGATGGTTAATGAGGCTCTTGTACGTCAGGGGTTAGCGAAGGTGGCTATGTTTACAAACCCAACAACACGCACGAACAACATTTACGCAAGTCTGAAGCCCAGCTAAGAAAGAGAAATTGAACATTTGGAGCGAAGACAACGCGGATAGTGGGCAA	MNase of *Staphylococcus aureus*; codon-optimized for *E. coli*
o2895	CAGGTTGGAAACGCCGTGAGNNKTTTCGTAAGTCTGGTGCAACGTGCGGTCGCTCGATNNKNNKTACCAGTCACCTACCGGTGACCGCATTCGCNNKAAGGTAGAATTGACCGCTA	MBD1[V20X/T33X/Y34X/S45X]
o2898	CCGGATGGAAAAAGGAGGAGNNKATTCGCAAGTCAGGCCTTAGTGCCGGCAAATCGATNNKNNKTATTTCTCACCTNNKGGCAAGAAGTTTCGCTCAAA	MBD2[V164X/V177X/Y178X/S183X]
o4441	CCGGATGGAAAAAGGAGGAGNNKATTCGCAAGTCAGGCCTTAGTGCCGGCAAATCGATNNKNNKTATTTCTCACCTAGTGGCAAGAAGTTTCGCNNKAAACCACAGTTAGACGTTATT	MBD2[V164X/V177X/Y178X/S189X]
o3181	AAGGATGGGAACGCGAGGAGNNKCCTCGCCGCTCAGGCCTTTCCGCCGGCCATCGGACNNKTTTNNKTACTCGCCTTCGGGAAAGAAATTCCGCNNKAAACCACAGTTAGGCGTTA	MBD3[V20X]

Table 12.6c Gene fragments.

Identifier	Sequence	Gene
o3182	GTGGGTGGGAGCGCGTTGTCNNKCAACGCTTGTTTGGTAAAACAGCCGGCCGTTTGATNNKNNKTTTATCTCGCCGCAAGGCCTAAAGTTCCGTNNKAAGTCATCACTTGTAACTA	MBD4[K95X/V108X/Y109X/S120X]
o2901	AGGGCTGGACCCGTAAATTGNNKCAACGCAAATCCGGTCGCAGCGCCGGCAAGTAGATNNKNNKCTTATCAATCCCCAAGGGAAGGCATTTCGCNNKAAGGTGGAGTTGACGCATA	MeCP2[K109X/V122X/Y123X/S134X]
o3966	AGGGCTGGACCCGTAAATTGNNKNNKCGCAAANNKGGTCGCNNKGCCGGCAAGTAGATGTGTATCTTATCAATCCCCAAGGGAAGGCATTTCGCAGTAAGGTGGAGTTGACGCATA	MeCP2[K109X/Q110X/S113X/S116X]
o3967	AGGGCTGGACCCGTAAATTGAAGCAACGCAAATCCGGTCGCAGCGCCGGCAAGNNGATNNKTDKCTTATCAATCCCCAAGGGAAGGCANNKCGCNNKAAGGTGGAGTTGACGCATA	MeCP2[Y120X/V122X/Y123Φ/F132X/S134X]
o3968	AGGGCTGGACCCGTAAATTGAAGCAACGCAAATCCGGTCGCNNKGCCNNKNNKNNGATGTGTATCTTATCAATCCCCAAGGGAAGGCATTTCGCAGTAAGGTGGAGTTGACGCATA	MeCP2[S116X/G118X/K119X/Y120X]
o4425	CGTAGTCGTCTCTCCGGCGCTGCGACCGGATTTGCGTTGMNNCAATTTACGGGTCAGCCCT	MeCP2[K109X]
o4427	CGTAGTCGTCTCACCGGCAAGTATGATNNKTATCTTATCAATCCCCAAGGGAAGGATTTCGCAATAAGGTGGAGTTGATCGCATA	MeCP2[V122X]

12.3 Instruments and consumables

Table 12.7a Laboratory instruments.

Part	Model	Company
Balance	PM400	Mettler-Toledo (Gießen, Germany)
Balance, analytical	M-Pact AX224	Sartorius (Göttingen, Germany)
Bunsen burner	1040/1	Carl Friedrich Usbeck KG (Radevormwald, Germany)
Camera	PowerShot G10	Canon (Krefeld, Germany)
Centrifuge	Mini centrigure ROTILABO®	Carl Roth (Karlsruhe, Germany)
Centrifuge, benchtop with cooling	5810 R	Eppendorf (Hamburg, Germany)
Centrifuge, benchtop with cooling	5424 R	Eppendorf (Hamburg, Germany)
Chromatography system	ÄKTA FPLC™ Fast Protein Liquid Chromatograph	GE Healthcare (Solingen, Germany)
Concentrator	Concentrator plus	Eppendorf (Hamburg, Germany)
Electrophoresis system, horizontal	EC-330 Primo™ Midicell™	Thermo Fisher (Schwerte, Germany)
Electrophoresis system, horizontal	kuroGEL Mini Plus 10	VWR (Darmstadt, Germany)
Electrophoresis system, vertical	Mini-PROTEAN® Tetra Cell	Bio-Rad (Munich, Germany)
Electroporator	Eporator®	Eppendorf (Hamburg, Germany)
Fluorescence-activated cell sorter	SH800 SGP	Sony Biotechnology (Weybridge, U. K.)
Freezer, –20 °C	Premium GGU 1500	Liebherr (Biberach, Germany)

Table 12.7b Laboratory instruments.

Part	Model	Company
Freezer, –20 °C	ProfiLine GG 4010	Liebherr (Biberach, Germany)
Freezer, –86 °C	New Brunswick™ HEF® U410	Eppendorf (Hamburg, Germany)
Heating block	ThermoStat™ plus	Eppendorf (Hamburg, Germany)
Ice flake maker	Scotsman AF20	Fisher Scientific (Loughborough, U. K.)
Incubator	INE 600	Memmert GmbH (Schwabach, Germany)
Incubator shaker	New Brundwick™ I26	Eppendorf (Hamburg, Germany)
Laser scanner, variable mode	Typhoon™ FLA 9500	GE Healthcare (Solingen, Germany)
Magnetic stand	MagRack 6	GE Healthcare (Solingen, Germany)
Magnetic stirrer	RCT classic	IKA-Werke (Staufen, Germany)
Magnetic stirrer	MR Hei-Standard, -Mix	Heidolph (Schwabach, Germany)
Micropipette, 0.1–2.5 µL	Research plus	Eppendorf (Hamburg, Germany)
Micropipette, 0.5–10 µL	Research plus	Eppendorf (Hamburg, Germany)
Micropipette, 10–100 µL	Research plus	Eppendorf (Hamburg, Germany)
Micropipette, 100–1,000 µL	Research plus	Eppendorf (Hamburg, Germany)
Micropipette, 12-channel, 0.5–10 µL	Xplorer	Eppendorf (Hamburg, Germany)
Microplate reader	Infinite® M1000	Tecan (Männedorf, Switzerland)
Microwave oven	Tecnolux ED 8525 exquisit	Verbeken & Fils (Drogenbos, Belgium)
Multistep pipette	Multipette® plus	Eppendorf (Hamburg, Germany)
PCR workstation enclosure	PCR Workstation Pro	PEQLAB (Erlangen, Germany)
Real-time PCR system	CFX384 Touch™	Bio-Rad (Munich, Germany)
PH electrode	LE410	Mettler-Toledo (Gießen, Germany)
PH meter	FiveEasy™ F20	Mettler-Toledo (Gießen, Germany)
Photometer	BioPhotometer® plus	Eppendorf (Hamburg, Germany)
Pipetting aid	accu-jet® pro	Brand GmbH (Wertheim, Germany)
Power supply	PowerPac™ Basic	Bio-Rad (Munich, Germany)
Refrigerator, 2–8 °C	ProfiLine FKU 1800	Liebherr (Biberach, Germany)
Scanner	CanonScan 9000F	Canon (Krefeld, Germany)
Shaker, orbital	Unimax 1010	Heidolph (Schwabach, Germany)
Shaker, overhead	Loopster Digital	IKA-Werke (Staufen, Germany)
Shaker, overhead	Tube Revolver	Thermo Fisher (Schwerte, Germany)
Size-exclusion column	HiPrep™ 26/60 Sephacryl® S-200HR	Merck (Darmstadt, Germany)
Sonicator	Bioruptor® Plus	Diagenode (Seraing, Belgium)
Spectrophotometer, UV-Vis	NanoDrop™ 2000	Thermo Fisher (Schwerte, Germany)
Thermocycler	T-Personal	Biometra (Göttingen, Germany)
Thermocycler	SimpliAmp™	Thermo Fisher (Schwerte, Germany)
Thermomixer	ThermoMixer® F	Eppendorf (Hamburg, Germany)
Thermomixer	ThermoMixer® C	Eppendorf (Hamburg, Germany)
Ultracentrifuge	Sorvall™ LYNX™ 6000	Thermo Fisher (Schwerte, Germany)

Table 12.7c Laboratory instruments.

Part	Model	Company
Transilluminator, UV	UVstar Plus	Biometra (Göttingen, Germany)
Vacuum pump	VNC 2	Vacuubrand (Wertheim, Germany)
Vortex mixer	Vortex-Genie 2	Scientific Industries (Bohemia, NY, U. S.)
Water bath, unstirred	JB Aqua 12 Plus	Grant Instruments (Shepreth, U. K.)

Table 12.8a Consumables and disposable labware.

Product (Brand)	Cat. No.	Company
384-well lightcycler plate, PP	72.1985.202	Sarstedt (Nümbrecht, Germany)
96-well plate for PCR, skirted	732-2387	VWR (Darmstadt, Germany)
96-well plate, clear, flat, for BCA	15045	Thermo Fisher (Schwerte, Germany)
Bottle-top filter, 500 mL (Filtropur)	83.1823.101	Sarstedt (Nümbrecht, Germany)
Centrifugal filter device, MWCO 3.5 kDa (Amicon®)	UFC9003	Merck (Darmstadt, Germany)
Column with frit, PP, 1.0 mL	34924	Qiagen (Hilden, Germany)
Column with frit, PP, 5.0 mL	34964	Qiagen (Hilden, Germany)
Cuvettes, standard	67.742	Sarstedt (Nümbrecht, Germany)
Cuvettes, UV-transparent	67.758	Sarstedt (Nümbrecht, Germany)
Dialysis unit, MWCO 10 kDa (Slide-A-Lyzer™ Cassettes)	66381	Thermo Fisher (Schwerte, Germany)
Dialysis unit, MWCO 10 kDa (Slide-A-Lyzer™ MINI)	88041	Thermo Fisher (Schwerte, Germany)
Electroporation cuvettes, 1 mm	PP38.1	Carl Roth (Karlsruhe, Germany)
Filter paper, qualitative, folded	no. 301	VWR (Darmstadt, Germany)
Filter paper, qualitative, folded	no. 305	VWR (Darmstadt, Germany)
Filter tips, 100–1,000 µL, low retention (Sorenson Mµlti)	732-3254	VWR (Darmstadt, Germany)
Filter tips, low retention, 0.1–10 µL (Sorenson Mµlti)	732-3249	VWR (Darmstadt, Germany)
Filter tips, low retention, 1–200 µL (Sorenson Mµlti)	732-3253	VWR (Darmstadt, Germany)
Glass beads, 5 mm	MARI4901005	VWR (Darmstadt, Germany)
Gloves, nitrile	816781635	Semperit (Vienna, Austria)
Injection needle, 18 G (Sterican®)	4665120	B. Braun (Melsungen, Germany)
Injection needle, 20 G (Sterican®)	4667093	B. Braun (Melsungen, Germany)
Microcentrifuge tubes, 1.5 mL	72.695.500	Sarstedt (Nümbrecht, Germany)
Microcentrifuge tubes, 1.5 mL, low retention (Protein LoBind™)	30108116	Eppendorf (Hamburg, Germany)
Microcentrifuge tubes, 1.5 mL, low retention nucleic acids	72.695.700	Sarstedt (Nümbrecht, Germany)
Microcentrifuge tubes, 2.0 mL	76.706	Sarstedt (Nümbrecht, Germany)
Microfluidic sorting chip, 100 µm	LE-C3610	Sony Biotechnology (Weybridge, U. K.)
Microfluidic sorting chip, 70 µm	LE-C3207	Sony Biotechnology (Weybridge, U. K.)
Microtubes for Bioruptor Pico, 1.5 mL	C30010016	Diagenode (Seraing, Belgium)
One-well plate, clear, PS, 127.8 x 85.5 mm	670190	Greiner Bio-one
Paper boxes (9 x 9)	95.64.981	Sarstedt (Nümbrecht, Germany)

Table 12.8b Consumables and disposable labware.

Product (Brand)	Cat. No.	Company
Paper towels, lint-free, 213 x 114 mm (Kimtech)	AA64.2	Carl Roth (Karlsruhe, Germany)
Parafilm® M	PM-996	Bemis (Oshkosh, WI, U.S.)
PCR tube, thick-walled, 0.2 mL (Multiply Pro)	72.737.002	Sarstedt (Nümbrecht, Germany)
PCR tube, thin-walled, 0.2 mL	72.737.005	Sarstedt (Nümbrecht, Germany)
Petri dishes for agar plates, 92 x 16 mm	82.1473.001	Sarstedt (Nümbrecht, Germany)
Petri dishes for agar plates, 150 x 20 mm	82.1184	Sarstedt (Nümbrecht, Germany)
Pipette tips, 0.1–10 µL	70.1130	Sarstedt (Nümbrecht, Germany)
Pipette tips, 10–200 µL	70.760	Sarstedt (Nümbrecht, Germany)
Pipette tips, 100–1,000 µL	70.762	Sarstedt (Nümbrecht, Germany)
Plate seals, adhesive, clear	600238	Biozym Scientific (Hessisch Oldendorf)
Scalpel	5518075	B. Braun (Melsungen, Germany)
Serological pipette, 10 mL	86.1254.001	Sarstedt (Nümbrecht, Germany)
Serological pipette, 25 mL	86.1685.001	Sarstedt (Nümbrecht, Germany)
Syringe filter, 0.2 µm, PES	16532-K	Sartorius (Göttingen, Germany)
Syringe filters, CA, 30 mm, sterile (Schleicher & Schuell Puradisc™ FP 30)	10462240	GE Healthcare (Solingen, Germany)
Syringes, single-use (Omnifix 10 mL, Luer Lock)	4617100V	B. Braun (Melsungen, Germany)
Syringes, single-use (Omnifix 50 mL, Luer Lock)	4617509F	B. Braun (Melsungen, Germany)
Syringes, single-use (Omnifix-F 1 mL, Luer Lock)	9166017V	B. Braun (Melsungen, Germany)
Transfer pipets, PP, 3.5 mL	86.1171	Sarstedt (Nümbrecht, Germany)
Tubes, conical, 15 mL, PP	62.547.254	Sarstedt (Nümbrecht, Germany)
Tubes, conical, 50 mL, PP	62.554.502	Sarstedt (Nümbrecht, Germany)
Tubes, round-bottom, 5 mL	10100151	Thermo Fisher (Schwerte, Germany)
Tubes, round-bottom, with sieve, 5mL	10585801	Thermo Fisher (Schwerte, Germany)
Waste bags, 200 x 300 mm	E706.1	Carl Roth (Karlsruhe, Germany)
Waste bags, 700 x 1120 mm	86.1204	Sarstedt (Nümbrecht, Germany)
Weigh boats, 41 x 41 mm	1-1124	Neolab (Heidelberg, Germany)
Weigh boats, 89 x 89 mm	1-1125	Neolab (Heidelberg, Germany)

12.4 Buffers, reagents and kits

Table 12.9a Commercial kits and ready-to-use mixtures.

Product (Brand)	Cat. No.	Company
Acrylamide/bisacrylamide 37.5:1, 40% (Rotiphorese® Gel 40)	T802	Carl Roth (Karlsruhe, Germany)
Automatic set-up beads for FACS	LE-B3001	Sony Biotechnology (Weybridge, U.K.)
DNA marker, 2-log DNA Ladder	N3200S	New England Biolabs (Frankfurt am Main)
DNTP Mix, 10 mM each	N0447L	New England Biolabs (Frankfurt am Main)

Table 12.9b Commercial kits and ready-to-use mixtures.

Product (Brand)	Cat. No.	Company
Illustra Ready-To-Go GenomiPhi V3	25-6601-24	GE Healthcare (Solingen, Germany)
LB agar (Lennox)	X965	Carl Roth (Karlsruhe, Germany)
LB broth (Lennox)	X964	Carl Roth (Karlsruhe, Germany)
NucleoSpin® Gel and PCR Clean-Up	740609.250	Macherey-Nagel (Düren, Germany)
NucleoSpin® Plasmid EasyPure	740727.250	Macherey-Nagel (Düren, Germany)
Pierce™ BCA Protein Assay Kit	23227	Thermo Fisher (Schwerte, Germany)
Protein ladder, pre-stained (PageRuler™)	26616	Thermo Fisher (Schwerte, Germany)
ThermoPol Reaction Buffer	B9004S	New England Biolabs (Frankfurt am Main)

Table 12.10 Commercially available enzymes. Enzymes provided in storage buffer; Pure enzymes in Table 12.12.

Product (Brand)	Cat. No.	Company
Alkaline phosphatase, calf intestine (CIP)	M0525S	New England Biolabs (Frankfurt am Main)
Alkaline phosphatase, shrimp (rSAP)	M0371S	New England Biolabs (Frankfurt am Main)
DNase I	M0303L	New England Biolabs (Frankfurt am Main)
*Dpn*I	R0176L	New England Biolabs (Frankfurt am Main)
*Eco*RI-HF	R3101S	New England Biolabs (Frankfurt am Main)
Klenow Fragment, exo–	EP0421	Thermo Fisher (Schwerte, Germany)
KOD Hot Start DNA Polymerase	71086-3	Merck (Darmstadt, Germany)
Mung Bean Nuclease	M0250S	New England Biolabs (Frankfurt am Main)
*Nde*I	R0111S	New England Biolabs (Frankfurt am Main)
NEBuilder HiFi DNA Assembly Master Mix	E2621S	New England Biolabs (Frankfurt am Main)
*Not*I-HF	R3189S	New England Biolabs (Frankfurt am Main)
Pfu DNA Polymerase	M7741	Promega (Walldorf, Germany)
Phusion DNA Polymerase	M0530S	New England Biolabs (Frankfurt am Main)
Plasmid-Safe™ ATP-dependent DNase	E3101K	Biozym Scientific (Hessisch Oldendorf, Germany)
Q5 High-Fidelity DNA Polymerase	M0491S	New England Biolabs (Frankfurt am Main)
*Sal*I	R0138S	New England Biolabs (Frankfurt am Main)
*Spe*I-HF	R3133S	New England Biolabs (Frankfurt am Main)
T4 DNA Ligase	M0202L	New England Biolabs (Frankfurt am Main)
T4 Polynucleotide Kinase	EK0031	Thermo Fisher (Schwerte, Germany)
T5 Exonuclease	M0363S	New England Biolabs (Frankfurt am Main)
Taq DNA Ligase	M0208S	New England Biolabs (Frankfurt am Main)
*Xba*I	R0145S	New England Biolabs (Frankfurt am Main)
*Xho*I	R0146S	New England Biolabs (Frankfurt am Main)

Table 12.11 Commercially available antibodies and affinity reagents.

Product (Brand)	Cat. No.	Company
Anti-c-Myc epitope (9E10) APC SureLight, mouse monoclonal antibody	ab72580	Abcam (Cambridge, U. K.)
Anti-c-Myc epitope (9E10), biotinylated, mouse monoclonal antibody	MA5-12077	Thermo Fisher (Schwerte, Germany)
Alexa Fluor® 488-streptavidin	BLD-405235	Biozol (Eching, Germany)
Brilliant Violet 711™-streptavidin	BLD-405241	Biozol (Eching, Germany)
FITC-strepatavidin	BLD-405201	Biozol (Eching, Germany)
R-phycoerythrin (PE)-strepatvidin	BLD-405204	Biozol (Eching, Germany)

Table 12.12a Chemicals.

CAS No.	Compound	Cat. No.	Company
64-19-7	acetic acid	6755	Carl Roth (Karlsruhe, Germany)
9012-36-6	agarose LE, molecular biology grade	840006	Biozym Scientific (Hessisch Oldendorf)
7727-54-0	ammonium persulfate (APS)	9592	Carl Roth (Karlsruhe, Germany)
5328-37-0	L(+)-arabinose	5118	Carl Roth (Karlsruhe, Germany)
58-85-5	D(+)-biotin	3822	Carl Roth (Karlsruhe, Germany)
34725-61-6	bromophenol blue, sodium salt	B8026	Merck (Darmstadt, Germany)
9048-46-8	bovine serum albumine (BSA)	9998	Cell Signaling Technology (Danvers, MA, U. S.)
10043-52-4	calcium chloride	10043-52-4	Fisher Scientific (Nidderau, Germany)
4800-94-6	carbenicillin, disodium salt	6344	Carl Roth (Karlsruhe, Germany)
56-75-7	chloramphenicol	3886	Carl Roth (Karlsruhe, Germany)
67-66-3	chloroform	C2432	Merck (Darmstadt, Germany)
6104-58-1	Coomassie Brilliant Blue G 250	9598	Carl Roth (Karlsruhe, Germany)
3483-12-3	1,4-dithiothreitol (DTT)	6908	Carl Roth (Karlsruhe, Germany)
64-17-5	ethanol, absolute, p. a.	32221	Merck (Darmstadt, Germany)
	ethanol, denatured with 1% methylethyl ketone	15835054	Fisher Scientific (Nidderau, Germany)
1239-45-8	ethidium bromide, 1% solution	2218	Carl Roth (Karlsruhe, Germany)
6381-92-6	ethylenediaminetetraacetate (EDTA)	E5134	Alfa Aeser (Landau, Germany)
50-99-7	D(+)-glucose, molecular biology grade	6887	Carl Roth (Karlsruhe, Germany)
56-81-5	glycerol, anhydrous, p.a. (Rotipuran®)	6962	Carl Roth (Karlsruhe, Germany)
56-40-6	glycine	HN07	Carl Roth (Karlsruhe, Germany)
7365-45-9	2-(4-(2-hydroxyethyl-) piperazin-1-yl)ethanesulfonic acid (HEPES)	9105	Carl Roth (Karlsruhe, Germany)
7647-01-0	hydrochloric acid, 37%	30721	Merck (Darmstadt, Germany)
288-32-4	imidazole	109053	abcr (Karlsruhe, Germany)
67-63-0	isopropanol	10315720	Fisher Scientific (Nidderau, Germany)
367-93-1	isopropyl-β-D-thiogalactopyranosid	2316	Carl Roth (Karlsruhe, Germany)
12650-88-3	lysozyme, from chicken egg white	11384029	Fisher Scientific (Nidderau, Germany)
7791-18-6	magnesium chloride, hexahydrate	197530010	Fisher Scientific (Nidderau, Germany)

Table 12.12b Chemicals.

CAS No.	Compound	Cat. No.	Company
67-56-1	methanol	34860	Merck (Darmstadt, Germany)
60-24-2	2-mercaptoethanol	M7154	Merck (Darmstadt, Germany)
108-95-2	phenol	328111	Merck (Darmstadt, Germany)
329-98-6	phneylmethanesufonyl fluoride (PMSF)	6367	Carl Roth (Karlsruhe, Germany)
7447-40-7	potassium chloride	A137	Carl Roth (Karlsruhe, Germany)
7778-77-0	potassium dihydrogen phosphate	3904	Carl Roth (Karlsruhe, Germany)
7647-14-5	sodium chloride	31434	Merck (Darmstadt, Germany)
151-21-3	sodium dodecyl sulfate, for biochemistry	CN30	Carl Roth (Karlsruhe, Germany)
1310-73-2	sodium hydroxide	30620	Merck (Darmstadt, Germany)
51805-45-9	Tris(2-carboxyethyl)-phosphine (TCEP), hydrochloride	HN95	Carl Roth (Karlsruhe, Germany)
110-18-9	N,N,N',N'-tetramethylethane-1,2-diamine (TEMED)	2367	Carl Roth (Karlsruhe, Germany)
77-86-1	2-amino-2-(hydroxymethyl)propane-1,3-diol (Tris, Trizma base), buffer grade	AE15	Carl Roth (Karlsruhe, Germany)
9002-93-1	Triton® X-100	A1388	Applichem (Darmstadt, Germany)
91079-40-2	tryptone (peptone ex casein)	8952	Carl Roth (Karlsruhe, Germany)
9005-64-5	Tween® 20	P9416	Merck (Darmstadt, Germany)
57-13-6	urea, p. a.	3941	Carl Roth (Karlsruhe, Germany)
7732-18-5	water, nuclease-free	P1193	Promega (Walldorf, Germany)
8013-01-2	yeast extract	2363	Carl Roth (Karlsruhe, Germany)

(12.4.1) Stock solutions

Carbenicillin
50 mg/mL in ethanol

Chloramphenicol
34 mg/mL in ethanol

DTT (dithiothreitol), 1 M
155 mg/mL in water, filter sterilized, stored frozen

EDTA, 0.5 M
186.1 mg/mL

Glucose, 1 M
180 mg/mL in water, filter sterilized

Glycerol, 50% (v/v)
0.5 mL/mL in water

PMSF (phneylmethanesufonyl fluoride), 100 mM
17.4 mg/mL in ethanol or isopropanol, stored frozen

SDS (sodium dodecyl sulfate), 20% (w/v)
20 mg/mL in water

Sodium acetate (NaOAc), 3 M
408 mg/mL sodium acetate trihydrate, adjusted to pH = 5.0 with acetic acid

Triton X-100, 10% (w/v)
100 mg/mL

(12.4.2) Buffer and media compositions

CutSmart® buffer, New England Biolabs, 1 ×
20 mM Tris-acetate, 50 mM KOAc, 10 mM $Mg(OAc)_2$, 0.1 mg/mL BSA, pH = 7.9

Dialysis buffer for recombinant MBD expression, 1 ×
20 mM HEPES, 100 mM NaCl, 10% glycerol, 0.1% Triton X-100, pH = 7.3

EMSA buffer, 10 ×
200 mM HEPES, 300 mM KCl, 10 mM EDTA, 10 mM $(NH_4)_2SO_4$, pH = 7.3

Extraction buffer for recombinant MBD expression, 1 ×

20 mM Tris-HCl, 250 mM NaCl, 10% glycerol, 10 mM 2-mercaptoethanol, 5 mM imidazole, 1 mM PMSF, 0.1% Triton X-100, pH = 8.0

LB broth (Miller)

10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl, pH = 7.0

NEBuffer3.1™ buffer, New England Biolabs, 1 ×

50 mM Tris-HCl, 100 mM NaCl, 10 mM magnesium acetate, 0.1 mg/mL BSA, pH = 7.9

Outer membrane extraction buffer, 1 ×

200 mM Tris-HCl, 20 mM glucose, 0.2 mM EDTA, pH = 8.0

Phosphate-buffered saline (PBS), 1 ×

10 mM Na_2HPO_4, 1.8 mM K_2HPO_4, 137 mM NaCl, 2.7 mM KCl, adjust to pH = 7.2

SDS-glycine PAGE running buffer, 1 ×

3 g/L Tris base, 14 g/L glycine, 1% (w/v) SDS, pH = 8.3 when stock is diluted

SDS PAGE sample buffer, 1 ×

62.5 mM Tris-HCl, 2% (w/v) SDS, 10% (v/v) glycerole, 0.02% bromophenol blue, 100 mM DTT, pH = 6.8

S. O. C. medium

20 g/L tryptone, 5 g/L yeast extract, 0.5 g/L NaCl, 20 mM glucose

Spectroscopy buffer for CD spectroscopy, 1 ×

50 mM NaH_2PO_4/Na_2HPO_4, 50 mM NaF, pH = 7.2

Spectroscopy buffer for NMR spectroscopy, 1 ×

50 mM NaH_2PO_4/Na_2HPO_4, 50 mM NaCl, 10 mM KCl, pH = 7.2

T4 DNA ligase buffer, New England Biolabs, 1 ×

50 mM Tris-HCl, 10 mM $MgCl_2$, 10 mM dithioerythritol, 1 mM adenosine triphosphate, pH = 7.5

Tris-acetate EDTA (TAE), 1 ×

40 mM Tris base, 20 mM acetic acid, 1 mM EDTA, pH = 7.8

Tris-borate EDTA (TBE), 1 ×

89 mM Tris base, 89 mM boric acid, 2 mM EDTA, pH = 8.3

Tris-EDTA (TE), 1 ×

10 mM Tris-HCl, 1 mM EDTA, pH = 8.0

12.5 Software and online tools

Non-exhaustive list. Additional resources are mentioned as applied.

Molecular cloning

SnapGene® v4.3 (GSL Biotech, LLC, Chicago, IL, U. S.); IDT Codon Optimization Tool as of 12/2017 (Integrated DNA Technologies Inc., Coralville, IA, U. S.); OligoCalc v3.19 (Kibbe, 2007); Tm Calculator v1.13 (New England Biolabs, Inc., Ipswich, MA, U. S.); NEBuilder® Assembly Tool v1.12 and v2.5 (New England Biolabs)

Instrument control

Cell Sorter Software v2.1 (Sony Biotechnology, Inc., San Jose, CA, U. S.); i-control v1.10 (Tecan Schweiz AG, Männedorf, Switzerland); NanoDrop 2000 v1.6 (Thermo Fisher Scientific, Waltham, MA, U. S.)

Image acquisition

BioDoc Analyze v2.1 (Biometra, Göttingen, Germany); Typhoon FLA 9500 Control Software v1.21 (Cytiva Europe GmbH, Freiburg, Germany)

Image processing and image analysis

ImageQuant TL v8.1 (Cytiva Europe); ImageJ v2.1/1.53c (Fiji distribution; Schindelin, et al., 2012);

Data management and data analysis

Github (GitHub, Inc., San Fransisco, CA, U. S.); Microsoft® Excel for Mac v16.52 (Microsoft Corporation, Redmond, WA, U. S.); *R* (R Core Team, 2021); RStudio v1.4 (Rstudio, PBC, Boston, MA, U. S.)

Visualization and graphics

Inkscape v1.1 (Inkscape Developers); UCSF ChimeraX v1.2 (Pettersen, et al., 2021)

Manuscript preparation

Papers v4.25 (Digital Science & Research Solutions, Inc., London, U. K.); LuaMetaTEX v2.09 (PRAGMA Advanced Document Engineering, Hasselt, The Netherlands)

Chapter 13

Methods

13.1 General procedures and routines

(13.1.1) Cell culture and transgenesis

Bacterial cultivation. If not otherwise indicated, bacterial cultures were cultivated in LB broth or on LB agar (Lennox) at 37 °C. Liquid cultures were grown aerobically at 220 rpm in a temperature-controlled shaker so that the culture volume was one fifth of the vessel.

Transformed bacteria were maintained on selective growth media with 50 µg/mL carbenicillin (an ampicillin substitute) or 30 µg/mL chloramphenicol depending on the resistance marker.

Competent bacterial cells for chemically transformation. Bacterial strains (see Table 12.1) were streaked to single clones on LB agar and expanded in 1 L LB broth from a fresh overnight culture until the OD_{600} had reached 0.5 cm^{-1}. Then, the culture was cooled on ice for 10 min and the cells harvested at 4 °C, 3,000 × *g* for 10 min. The pellet was washed once with 100 mL ice-cold sterilized 100 mM $MgCl_2$ and once with 50 mM $CaCl_2$. After resuspension in 5 mL sterile-filtered 50 mM $CaCl_2$ 15% (v/v) glycerol, the suspension was aliquoted into fractions of 50 µL, snap-frozen in liquid nitrogen and stored at −80 °C until further use.

Routinely, 1 – 100 ng DNA were transferred in a maximum of 10 µL to one aliquot of 50 µL (thawed on ice) and gently mixed by flicking the tube (do not pipet up and down). The mixture was incubated on ice for 20 min and placed at 42 °C for exactly 30 s without shaking before putting back on ice for 2 min. The cells were rescued with 1 mL pre-warmed S. O. C. medium for 30 – 60 min at 37 °C. 20 µL and 200 µL of the transformation was plated onto LB agar containing antibiotics, the plates allowed to air-dry and were incubated inverted at 37 °C overnight.

Competent bacterial cells for electroporation. For the preparation of competent cells, strains were expanded and harvested as above, but resuspended once in 400 mL ice-cold sterile water then again once in 200 mL before the pellet was finally resuspended in 100 m 10% (v/v) glycerol, the supernatant discarded and the pellet resuspended in 5 mL 10% (v/v) glycerol. Aliquots of 25 µL were snap-frozen and stored at −80 °C.

Routinely, 0.1 – 10 ng plasmid (or 10 – 100 ng linearized, purified PCR products) in maximal 5 µL were added to one 25 µL aliquot of competent cells and transferred to a pre-chilled electroporation cuvette (Carl Roth, Karlsruhe, Germany), electroporated at 1.8 kV, and immediately rescued with 1 mL pre-warmed S. O. C. medium for 30 – 60 min depending on the resistance marker at 37 °C. Per transformation, spread-plating 10 µL yield enough colonies for routines.

Electroporation cuvettes could be reused up to five times after thoroughly rinsing with 70% ethanol and distilled water.

(13.1.2) Analytical biochemistry

Isolation of plasmids from bacterial strains. Plasmids were isolated using silica column-purification with a commercially available kit (NucleoSpin® Plasmid EasyPure; Macherey-Nagel, Düren, Germany) according to the manufacturer's instruction.

Sanger sequencing of plasmids. The identity of all plasmids was routinely checked by restriction digest where possible and ultimately confirmed by Sanger sequencing by Microsynth Seqlab GmbH (Göttingen, Germany) or Eurofins Genomics Germany GmbH (Munich, Germany; formerly GATC Biotech AG, Konstanz, Germany).

Purification of PCR products. PCR products and other double-stranded DNA were purified using NucleoSpin® Gel and PCR Clean-Up (Machery-Nagel).

Agarose gel electrophoresis and gel extraction. Preparative and analytical agarose gels of 0.8 – 2.0% (w/v) agarose in 0.5 × TBE buffer were used to resolve DNA samples at 6 – 12 V/cm. The gels were stained in 0.5 µg/mL ethidium bromide, destained with water and the DNA was visualized by UV fluorescence, documented with a camera. For extracting DNA from the gel, the NucleoSpin® Gel and PCR Clean-Up kit was used.

Glycine-SDS polyacrylamide gel electrophoresis. Analytical glycine-SDS PAGE gels of 12 – 15% (v/v) acrylamide were used to resolve MBD or MBD-fusion proteins. Samples were separated at 36 mA (240 V) for 40 min and stained with Coomassie brilliant blue for 20 min, destained with 45% (v/v) methanol in 10% (v/v) acetic aced until protein bands were clearly visible. Gels were documented using an image scanner.

13.2 Recombinant MBD expression and protein purification

The vectors for expression of recombinant methyl-CpG-binding domains (MBDs) are based on pET-21d(+) and allow shuffling the MBD insert between the expression vectors and/or the cell surface display constructs (Section) using *Xho*I and *Spe*I (**Figure 13.1**).

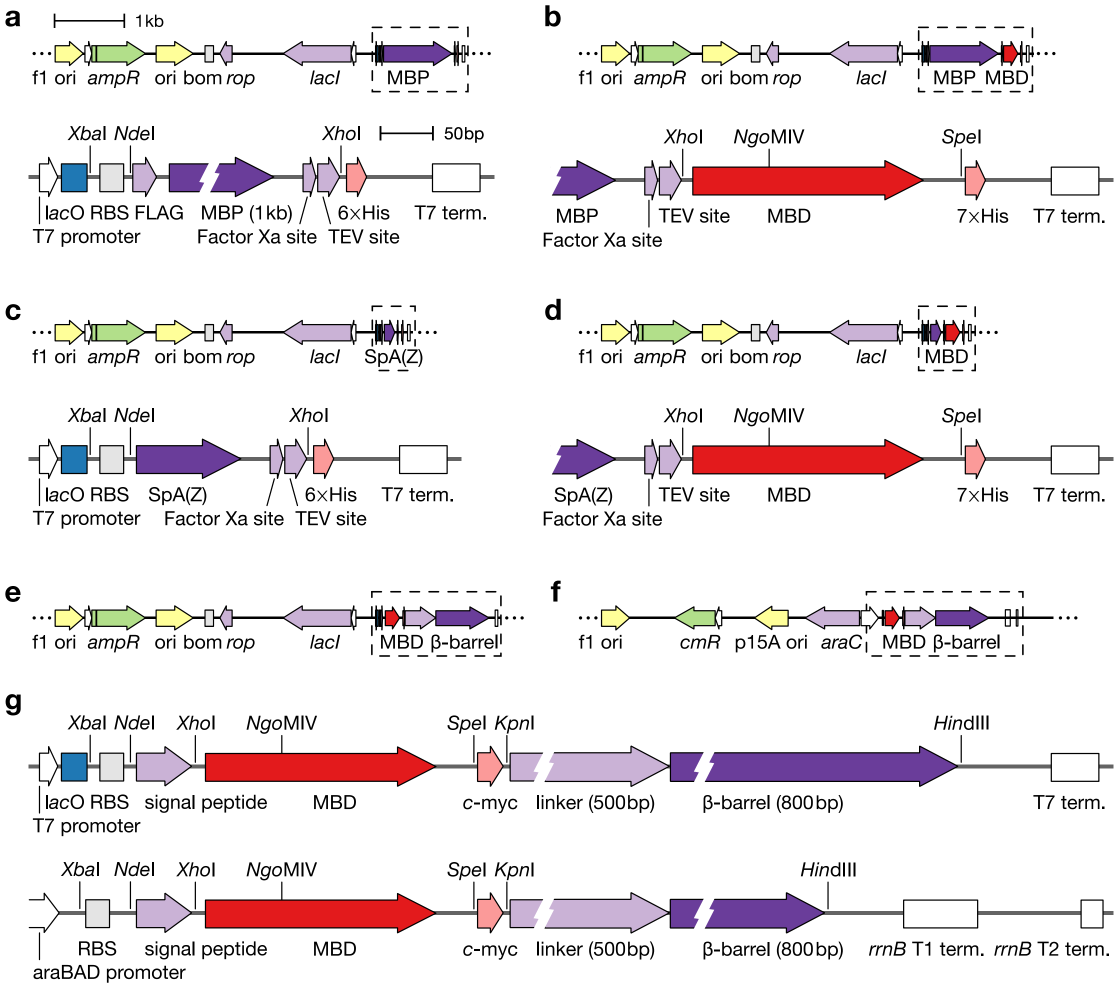

a
1 kb
f1 ori ampR ori bom rop lacI MBP
XbaI NdeI XhoI 50bp
lacO RBS FLAG MBP (1kb) 6×His T7 term.
T7 promoter Factor Xa site TEV site
b
f1 ori ampR ori bom rop lacI MBP MBD
XhoI NgoMIV SpeI
MBP TEV site MBD 7×His T7 term.
Factor Xa site
c
f1 ori ampR ori bom rop lacI SpA(Z)
XbaI NdeI XhoI
lacO RBS SpA(Z) 6×His T7 term.
T7 promoter Factor Xa site TEV site
d
f1 ori ampR ori bom rop lacI MBD
XhoI NgoMIV SpeI
SpA(Z) TEV site MBD 7×His T7 term.
Factor Xa site
e
f1 ori ampR ori bom rop lacI MBD β-barrel
f
f1 ori cmR p15A ori araC MBD β-barrel
g
XbaI NdeI XhoI NgoMIV SpeI KpnI HindIII
lacO RBS signal peptide MBD c-myc linker (500bp) β-barrel (800bp) T7 term.
T7 promoter
XbaI NdeI XhoI NgoMIV SpeI KpnI HindIII
RBS signal peptide MBD c-myc linker (500bp) β-barrel (800bp) rrnB T1 term. rrnB T2 term.
araBAD promoter

Figure 13.1 Overview of plasmid vectors. (a) Entry vector for recombinant expression with N-terminal MBP tag p1379 with *amp*^{r} (p1780 with *cm*^{r}). **(b)** Recombinant vector for MBP-MBD expression, e.g., p1388. **(c)** Entry vector for recombinant expression with N-terminal SpA(Z) tag p1380 (p1785). **(d)** Recombinant vector for SpA(Z)-MBD expression, e.g., p1393. **(e)** pET21d(+)-based, β-D-1-thiogalactopyranoside-inducible recombinant vector for surface display of MBD. **(f)** pBAD33.1-based, arabinose-inducible recombinant vector for surface display of MBD. **(g)** Detail of *e* (top) and *f* (bottom).

(13.2.1) Cloning of expression vectors

Entry vectors. pET-21d(+) (Merck KGaA, Darmstadt, Germany) was digested with *Xho*I and *Nco*I (New England Biolabs GmbH, Frankfurt am Main, Germany; 'NEB' hereafter) to replace the T7 tag by Gibson assembly (Gibson, et al., 2009) with a maltose-binding protein (MBP) tag or the synthetic Z domain of staphylococcal protein A (SpA) (Nilsson, et al., 1987). SpA(Z) was amplified from an accessory plasmid (gift from P. Bieling, MPI Dortmund) using the primers o2872/o2873. The resulting vectors were p1379 and p1380 respectively, in which the N-terminal tag is followed by a factor Xa and a TEV recognition and cleavage site and the target protein by a non-cleavable C-terminal His$_6$ tag.

Entry vectors with alternative selection markers. To enable stringent shuffling, a second pair of vectors was created by exchanging the ampicillin resistance marker (*amp*^{r}) of p1379 and p1380 for the chloramphenicol resistance marker (*cm*^{r}) of pBAD33.1 (Chung & Raetz, 2010). The plasmids were linearized by PCR (KOD Hot Start DNA Polymerase, Merck) using o1260/o3539 on pBAD33.1 or o1260/o3540 on the pET derivates, digested with *Dpn*I and *Bam*HI (NEB) and ligated with T4 DNA ligase (NEB). Only the antisense orientation of the marker cassette was retrieved. The resulting plasmids were p1780 and p1785.

Subcloning of MBDs. The consensus coding sequences (CCDS) of the human MBD proteins were obtained form the CCDS project (Farrell, et al., 2013). The MBD domain within the coding sequences was identified by alignment to Pfam PF01429 and flanked with about 5 to 15 additional amino acids at the N- and C-terminus of the domain respectively. This sequence was codon-optimized for bacterial expression using the IDT Codon Optimization Tool. Unwanted restriction sites were removed. All MBD coding sequences contained a *Ngo*MIV restriction site, and the coding sequence of MBD3 contained an *Bgl*I site in addition.

p1379 and p1380 were linearized with *Xho*I, and the codon-optimized sequences of the human MBDs obtained as gBlocks (Integrated DNA Technologies, IDT, Leuven, Belgium) were amplified by PCR (Phusion® High-fidelity DNA Polymerase, NEB) and introduced by Gibson assembly (NEBuilder®, NEB) following the manufacturer's protocol (**Table 13.1**). Due to the repetitive sequence encoding the His$_6$ tag, assembling resulted in His$_7$ tagged fusion proteins.

Table 13.1 Consensus coding sequences and cloning of MBD wildtype domains.

MBD	CCDS*	gBlock	PCR primer pair	PCR ϑ_a	MBP–MBD	SpA(Z)–MBD
MBD1[2–81]	59318.1	o2878	o2879/o2880	65 °C	p1384	p1389
MBD2[146–225]	11953.1	o2881	o2882/o2883	60 °C	p1385	p1390
MBD3[2–81]	12072.1	o2884	o2885/o2886	65 °C	p1386	p1391
MBD4[76–167]	03058.1	o2887	o2888/o2889	65 °C	p1387	p1392
MeCP2[90–181]	14741.1	o2890	o2891/o2892	60 °C	p1388	p1393

* Consensus CDS Database Release 22, https://www.ncbi.nlm.nih.gov/projects/CCDS/, retrieved on December 18, 2018.

Rett mutants. Rett-associated MeCP2 mutations were introduced using a modified QuikChange site-directed mutagenesis (Agilent) protocol on p1393. Using 10 ng template, a KOD PCR (KOD Hot Start DNA Polymerase, Merck) was carried out following the manufacturer's protocol in presence of 300 nM of each primer (**Table 13.2**) over 30 cycles allowing 4 min for elongation. 5 µL were transformed into chemically competent DH5α without purification.

Table 13.2 Missense mutations and cloning of MeCP2 Rett mutants.

MBD Rett mutant	RettBASE* Frequency	RettBASE Percentage	PCR primer pair	PCR ϑ_a	MBP–MBD	SpA(Z)–MBD
MeCP2[T158M]	420	8.74%	o3124/o3125	71 °C	n/a	p1645
MeCP2[R133C]	217	4.52%	o3122/o3123	71 °C	n/a	p1644
MeCP2[S134C]	21	0.44%	o3118/o3119	71 °C	n/a	p1642
MeCP2[L124F]	3	0.06%	o3120/o3121	71 °C	n/a	p1643

* RettBASE: RettSyndrome.org Variation Database, http://mecp2.chw.edu.au/, retrieved on July 9, 2018.

Variants compatible with MTSL labeling. MBP–MeCP2 are a cysteine-free fusion proteins. To introduce a cysteine residue for labeling the protein, e. g., with a MTSL spin label, the 50 ng expression vectors was subjected to the following QuikChange site-directed mutagenesis protocol using 2 U *Pfu* DNA polymerase (Promega, Walldorf, Germany) in a total reaction volume of 50 µL (20 mM Tris-HCl, 50 mM KCl, pH = 8.4, 0.3 mM dNTP mix, 500 nM each primer, **Table 13.3**): Following an initial denaturation at 95 °C for 2 min, 20 cycles of 95 °C for 0.5 min, 58 °C for 1 min, 68 °C for 15 min. 5 µL of each reaction were transformed into chemically competent DH5α without purification.

Table 13.3 Cloning of MeCP2 cysteine variants.

MeCP2 Cys variant	Template	PCR primer pair	PCR ϑ_a	MBP–MBD	SpA(Z)–MBD
MeCP2[A117C]	p1388	o4254/o4255	58 °C	p2573	n/a
MeCP2[A140C]	p1388	o4256/o4257	58 °C	p2574	n/a
MeCP2[G161C]	p1388	o4260/o4261	58 °C	p2575	n/a
MeCP2[G146C]	p1388	o4258/o4259	58 °C	p2576	n/a
MeCP2[K109T/V122A/S134N/A117C]	p1859	o4280/o4281	58 °C	p2577	n/a
MeCP2[K109T/V122A/S134N/A140C]	p1859	o4282/o4283	58 °C	p2578	n/a
MeCP2[K109T/V122A/S134N/G161C]	p1859	o4260/o4261	58 °C	p2579	n/a
MeCP2[K109T/V122A/S134N/G146C]	p1859	o4258/o4259	58 °C	p2580	n/a

(13.2.2) Expression and purification

A protocol of Free et al. (2001) and Valinluck et al. (2004) was used with modifications for recombinant expression and purification of MBD fusion proteins.

Expression and harvest. *E. coli* BL21-Gold(DE3) (Agilent) were transformed with plasmids for protein expression and fresh overnight cultures of single clones diluted the next morning to an optical density (OD_{600}) of 0.05 in 30 mL LB broth (Miller) supplemented with 50 µg/mL carbenicillin, 1 mM $MgCl_2$ and 1 mM $ZnSO_4$ (Hashimoto, et al., 2012). Cultures were grown at

37 °C (220 rpm) to an OD_{600} of 0.5 – 0.6 , briefly chilled on ice, and then induced by supplying 1 µM β-D-1-thiogalactopyranoside (IPTG). Cultures were incubated at 25 °C (150 rpm) for at least 6 h or overnight, and cells were harvested and washed once by resuspension in 0.25 vol ice-cold 20 mM Tris-HCl (pH = 8.0).

Extraction. Pellets were resuspended in 2 mL extraction buffer (20 mM Tris-HCl, 250 mM NaCl, 10% glycerol, adjusted to pH = 8.0, then supplemented with 10 mM 2-mercaptoethanol, 5 mM imidazole, 1 mM PMSF and 0.1% Triton X-100) and the proteins extracted by pulse-sonication. Suspensions were treated with 0.1 mg/mL lysozyme (Merck) and 10 U/mL DNase I (NEB) overnight. Insoluble debris was collected by centrifugation at 14,000 × *g* for 20 min at 4 °C.

Small-scale purification. The cleared supernatants were retained, diluted with 1 vol (2 mL) extraction buffer, mixed with 450 µL 50% Ni-nitriloacetic acid (NTA) HisPur™ Superflow Agarose (Thermo Fisher) and incubated at 4 °C for 2 h. The resins were washed twice with 1 mL extraction buffer containing 90 mM imidazole (20 min at 4 °C) and the fusion proteins were eluted by gravity flow in 2 × 0.2 mL and 1 × 0.4 mL extraction buffer with 500 mM imidazole after 10 min incubation at 4 °C.

Large-scale purification. When starting with a 2 L expression culture, the pellet was resuspended in 30 mL extraction buffer and treated with lysozyme in presence of PMSF for 60 min on a wheel-shaker at 4 °C. The suspension was extracted by pulse-sonication or on a high-shear microfluidizer homogenizer. Insoluble debris was removed at 30,000 × *g* for 45 min at 4 °C.

The cleared supernatant was sterile-filtered (0.2 µm syringe filter) and loaded on a self-packed 10 mL column HisPur™ Ni-NTA Resin (Thermo Fisher) connected to an ÄKTA FPLC™ Fast Protein Liquid Chromatograph (GE Healthcare, Solingen, Germany). The mixture was separated at 1 mL/min flow rate of binding buffer containing 5 – 90 mM imidazole (0 – 100%) in 80 min. Fractions containing the MBP–MBD fusion protein were combined for dialysis.

Dialysis. Fractions judged to be more than 90% pure by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS–PAGE) were combined and dialyzed against 3 × 15 mL 20 mM HEPES, 100 mM NaCl, 10% glycerol, adjusted to pH = 7.3, and 0.1% Triton X-100 (volumes for small-scale expression) in Slide-A-Lyzer MINI devices (3.5 kDa MWCO, Thermo Fisher Scientific, Schwerte, Germany). An additional 1:2 – 1:5 dilution was found beneficial when scaling up this procedure to avoid precipitation during dialysis.

Protein quantification and storage. The protein concentration was determined with a bicinchoninic acid (BCA) assay (Thermo Fisher) and the proteins stocked at 15 µM after snap-freezing in liquid nitrogen, at (stable for several months). Typically, 1 pmol MBP–MBD fusion protein or 3 – 4 pmol SpA(Z)–MBD fusion protein were obtained per milliliter culture.

The SpA(Z) tag can be efficiently removed with 0.25 µM TEV protease at 4 °C overnight *in situ*. Uncleaved or cleaved SpA(Z) tag and/or the TEV protease do not interfere with DNA binding; Further purification is not necessary. However, it has been noted that prolonged storage of the tag-free MBDs can result in spontaneous precipitation.

Table 13.4 Physicochemical properties of MBD fusion proteins. Calculated values.

		MBD domain					
N-/C-terminal tag	**Property**	**none**	**MBD1**	**MBD2**	**MBD3**	**MBD4**	**MeCP2**
MBP/His$_{6-7}$	expressed from	p1380	p1384	p1385	p1386	p1387	p1388
	molecular weight	45.2 kDa	53.9 kDa	54.0 kDa	54.4 kDa	55.5 kDa	56.4 kDa
	isoelectric point	4.83	5.23	6.63	6.41	6.29	5.96
SpA(Z)/His$_7$	expressed from	p1381	p1389	p1390	p1391	p1392	p1393
	molecular weight	10.1 kDa	18.8 kDa	18.9 kDa	19.3 kDa	20.4 kDa	20.3 kDa
	isoelectric point	6.22	6.66	9.89	9.73	8.56	9.40
tag-free	expressed from	n/a	n/a	n/a	n/a	n/a	n/a
	molecular weight	n/a	8.9 kDa	9.0 kDa	9.4 kDa	10.3 kDa	10.4 kDa
	isoelectric point	n/a	7.24	10.74	10.93	10.35	10.33
	$\varepsilon_{280\,nm}$ (mM^{-1} cm^{-1})	n/a	16.96	9.97	9.73	9.97	11.56

Removal of solubility tags for spectroscopic analyses. For spectroscopic CD and NMR analyses (large-scale expressions) the solubility tag was cleaved by adding His-free TEV protease to the combined eluates of the column purification after a first dialysis (3.5 kDa MWCO) against binding buffer without imidazole (20 mL eluate against 2 L buffer). The dialysis buffer was exchanged once. Then, the His-free TEV and the solubility tag were removed over the same Ni-NTA column as before using a gradient of 0 – 30% over 150 min. The MBDs were eluted with 100% 90 mM imidazole in binding buffer and again the desired fractions combined.

The combined fractions were concentrated over an Amicon® centrifugal filter device (3.5 kDa MWCO; Merck) to 1 – 2 mL for size-exclusion chromatography and loaded onto a HiPrep™ 26/60 Sephacryl® S-200HR dextran (Merck) column. The MBD domain was polished at a flow rate of 1 mL/min of the final buffer; the desired fractions combined and concentrated as before.

Determination of protein concentration. Purity of was assessed by SDS-PAGE and protein concentrations were determined with the Pierce™ BCA Protein Assay Kit (Thermo Fisher). For spectroscopic analyses, the protein concentration could be reliably determined by absorbance measurements at 280 nm.

13.3 Electrophoretic mobility shift assays

Annealing of probes. 24-mer oligodeoxynucleotide (ODN) pairs (**Table 13.5** and **Table 13.6**) were combined at 1.5 µM of the labeled strand and 1.8 µM of the unlabeled strand to ensure complete assimilation of the labeled strand in the duplex. The concentration of the ODNs and the annealed DNA probes must be determined spectrophotometrically and adjusted where needed. For the 24-mers, an extinction coefficient of 307 mM^{-1} cm^{-1} was used for the FAM-labeled oligo(A) context and 197 mM^{-1} cm^{-1} for the unlabeled oligo(T) context.

The ODNs were annealed in 100 µL 2 × EMSA buffer (10 × stock is 200 mM HEPES, 300 mM KCl, 10 mM EDTA, 10 mM $(NH_4)_2SO_4$, pH = 7.3) in boiling water and slowly brought to room temperature. The duplex was diluted to 300 nM (A260 = 0.09 cm^{-1} for the A/T 24-mers) and the concentration confirmed again. A final dilution to 30 nM was aliquoted and stocked at −20 °C.

The unspecific binding traps were prepared by annealing a 24-mer oligo(A) with a 24-mer oligo(T) or unmodified complementary strands at equimolar ratios at 50 µM.

Assay. To determine the binding affinity of MBDs, a well-established protocol (Free, et al., 2001; Khrapunov, et al., 2014; Valinluck, et al., 2004; Yang, et al., 2016) was followed with minor modifications. Typically, the purified MBDs were diluted 1,024 nM 512 nM, 256 nM, 128 nM, 64 nM, 32 nM, 16 nM, 8 nM, 4 nM, 2 nM and 0 nM in at least 20 µL MBD dialysis and storage buffer with 0.1 mg/mL BSA to minimize loss due to surface adsorption. The dilution series was incubated with 2 nM labeled duplex and 50 ng/µL of the oligo(A) · oligo(T) binding trap in a final volume of 15 µL EMSA buffer (20 mM HEPES, 30 mM KCl, 1 mM EDTA, 1 mM $(NH_4)_2SO_4$, pH = 7.3) containing 1 mM dithioerythritol or tris(2-carboxyethyl)phosphin as reducing agent and 0.2% Tween 20. If several different duplexes were compared, the procedure was carried out at a single concentration.

The binding reaction was allowed to equilibrate for 20 min at room temperature before 3 µL of a 6× loading dye (1.5× TBE buffer, pH = 7.5, 40% glycerol, 70 pg/mL bromophenol blue) were added on ice. These samples (10 µL) were loaded on 0.25×, TBE 12% polyacrylamide gels (gels must be pre-run at 70 V overnight or 200 for 120 min, 4 °C) and the electrophoresis carried out at 240 V for 35 min (24-mers) or 40 min (longer DNA probes) at 4 °C in Mini-PROTEAN vertical

Table 13.5 Combinations of ODNs to create homopolymeric DNA probes for EMSA. Hybridization of the Watson and the Crick strand gives fluorescently-labeled 24-mer DNA duplexes.

Combination*	Context†	Sequence source	Watson strand	Crick strand	Other modifications
A/T	AAAAAAAAAAAA	artificial	o2968	o2969	dark competitor
D/D	AAAAADHAAAAA	artificial	o3416	o3417	dark competitor
C/C	AAAAACGAAAAA	artificial	o2906	o2904	1 × FAM
C/5mC	AAAAACGAAAAA	artificial	o2906	o2909	1 × FAM
5mC/C	AAAAACGAAAAA	artificial	o2967	o2904	1 × FAM
C/5hmC	AAAAACGAAAAA	artificial	o2906	o3112	1 × FAM
5hmC/C	AAAAACGAAAAA	artificial	o3115	o2904	1 × FAM
C/5fC	AAAAACGAAAAA	artificial	o2906	o3113	1 × FAM
5fC/C	AAAAACGAAAAA	artificial	o3116	o2904	1 × FAM
C/5caC	AAAAACGAAAAA	artificial	o2906	o3114	1 × FAM
5caC/C	AAAAACGAAAAA	artificial	o3117	o2904	1 × FAM
5mC/5mC	AAAAACGAAAAA	artificial	o2967	o2909	1 × FAM
5mC/5hmC	AAAAACGAAAAA	artificial	o2967	o3112	1 × FAM
5hmC/5mC	AAAAACGAAAAA	artificial	o3115	o2909	1 × FAM
5mC/5fC	AAAAACGAAAAA	artificial	o2967	o3113	1 × FAM
5fC/5mC	AAAAACGAAAAA	artificial	o3116	o2909	1 × FAM
5mC/5caC	AAAAACGAAAAA	artificial	o2967	o3114	1 × FAM
5fC/5mC	AAAAACGAAAAA	artificial	o3117	o2909	1 × FAM
5hmC/5hmC	AAAAACGAAAAA	artificial	o3115	o3112	1 × FAM
5hmC/5fC	AAAAACGAAAAA	artificial	o3115	o3113	1 × FAM
5fC/5hmC	AAAAACGAAAAA	artificial	o3116	o3112	1 × FAM
5hmC/5caC	AAAAACGAAAAA	artificial	o3115	o3114	1 × FAM
5caC/5hmC	AAAAACGAAAAA	artificial	o3117	o3112	1 × FAM
5fC/5fC	AAAAACGAAAAA	artificial	o3116	o3113	1 × FAM
5fC/5caC	AAAAACGAAAAA	artificial	o3116	o3114	1 × FAM
5caC/5fC	AAAAACGAAAAA	artificial	o3117	o3113	1 × FAM
5caC/5caC	AAAAACGAAAAA	artificial	o3117	o3114	1 × FAM
C/T	AAAAACAAAAAA	artificial	o4277	o4279	1 × FAM
5mC/T	AAAAACAAAAAA	artificial	o4278	o4279	1 × FAM
5hmC/T	TTTTTCATTTTT	artificial	o4839	o4277	2 × FAM
5hmC/T	AAAAACAAAAAA	artificial	o4840	o4279	1 × FAM

* Strand-diametral cytosines X/Y in a Xp$\overline{\text{Y}}$ · Yp$\overline{\text{X}}$ dyad, i. e., C/T = CpA · TpG and C/C = CpG · CpG.

† Refers to the Watson strand.

electrophoresis cells (Bio-Rad, Munich, Germany). Gels were recorded on a Typhoon FLA-9500 laser scanner (GE Healthcare, Solingen, Germany) equipped with a 473 nm laser and a 510 LP filter at 700 V photomultiplier tube (PMT) gain.

Table 13.6 Combinations of ODNs to create genomic DNA probes for EMSA. The 79-mer *Hey2* sequence is located at chr20:39,589,641-39,589,719 (danRer11) in an intronic region. The 45-mer *BRCA1* at chr17:43,125,546-43,125,590 (hg38) in the first exon of *NBR2* which is part of the bidirectional promoter of *BRCA1*. The 45-mer *CDKN2A* sequence is located at chr9:21,974,777-21,974,822 (hg38) in the gene promoter (or first exon of an alternative start site).

Combination*	Context†	Sequence source	Watson strand	Crick strand	Other modifications
C/C	*(multiple)*	*Hey2* (*Danio rerio*)	o476	o1152	dark competitor
5mC/5mC	TCTTCCGTTTCC	*Hey2* (*Danio rerio*)	o4379	o517	1 × FAM
5hmC/5mC	TCTTCCGTTTCC	*Hey2* (*Danio rerio*)	o4379	o520	1 × FAM
C/C	*(multiple)*	*BRCA1* (*H. sapiens*)	o1516	o1529	dark competitor
5mC/5mC	TCTTCCGTCTCT	*BRCA1* (*H. sapiens*)	o4497	o1517	1 × FAM
5hmC/5mC	TCTTCCGTCTCT	*BRCA1* (*H. sapiens*)	o4497	o1520	1 × FAM
5mC/5mC	TTTTACGTCATC	*BRCA1* (*H. sapiens*)	o4498	o1518	1 × FAM
5mC/5mC	TTTTACGTCATC	*BRCA1* (*H. sapiens*)	o4497	o1521	1 × FAM
C/C	*(multiple)*	*CDKN2A* (*H. sapiens*)	o1591	o1617	dark competitor
5mC/5mC	TCAGCCGAAGGC	*CDKN2A* (*H. sapiens*)	o4499	o1592	1 × FAM
5hmC/5mC	TCAGCCGAAGGC	*CDKN2A* (*H. sapiens*)	o4499	o1593	1 × FAM

* Strand-diametral cytosines X/Y in a $\mathrm{Xp\overline{Y}} \cdot \mathrm{Yp\overline{X}}$ dyad, i. e., C/T = CpA · TpG and C/C = CpG · CpG.

† Refers to the Watson strand.

Quantification. The fraction of bound duplex was determined using ImageQuant TL v8.1 1D Gel Analysis (GE Healthcare) applying rubber band background subtraction and manual peak detection with approximately equal peak areas across all lanes. The fraction of bound duplex was established with *R* using the 'summerr*band*' package (written for this work) and the dissociation constant $K_d = 1/K$ determined by non-linear curve fitting using an exact binding model and the Levenberg-Marquardt algorithm if not otherwise specified as detailed in Section 9. The package is available under doi:10.5281/zenodo.5348399.

13.4 Creation of NNK codon-degenerated MBD libraries

(13.4.1) Library creation via type IIS cloning

This is a working protocol to create NNK codon-degenerated libraries. However, the yield of transformants is unsatisfactory to create highly diverse libraries (Chapter 7.2). The procedure is exemplified for NNK codon degeneration of MBD2.

Backbone linearization and degeneration. The backbone p1567 was linearized in 30 PCR cycles using 0.04 ng/µL template and 0.3 µM of each primer (o3280 and o3281; annealing temperature: 64 °C) in a KOD Hot Start DNA Polymerase PCR (Toyobo, Osaka, Japan; distributed by Merck Millipore, Darmstadt, Germany) according to the manufacturer's instructions. The template was digested *in situ* with *Dpn*I at 37 °C for 15 min. A single reaction purified by silica column-chromatography (Machery-Nagel, Düren, Germany) yielded about 1.2 µg. The reaction was digested with B s mB at 0.2 U/µL in NEBuffer™ 3.1 (50 mM Tris-HCl, 100 mM NaCl, 10 mM $MgCl_2$, 0.1 mg/m BSA, pH = 7.9; NEB) at 55 °C for 60 min and purified as before.

Self-circularization and plasmid pool creation. The digested product was diluted to 2.5 ng/µL in T4 DNA ligase buffer (50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM ATP, 10 mM dithiothreitol, pH = 7.5; NEB) and T4 DNA ligase (NEB) was added to a final concentration of 0.4 U/µL. The ligation was allowed to proceed at 16 °C for 6 h before the circularized DNA was recovered by ice-cold ethanol precipitation. Transformation of 0.10 vol of the reaction in DH10B (TOP10™, Thermo Fisher Scientific) at a target voltage of 1.8 kV over 4 ms in 100 µL (25 OD_{600}) competent cells yielded about 21,000 colonies.

(13.4.2) Library creation via Gibson assembly

This is an adaptation of the 'NEBuilder HiFi DNA Assembly Reaction (E2621) V.1' protocol (doi:10.17504/protocols.io.cwaxad) to meet the demands for NNK library creation with short oligodeoxynucleotides (ODN). A plasmid library is obtained via Gibson assembly of a NNK codon-degenerated 118-mer ODN and the linearized plasmid backbone, e. g., an AIDA-I surface display entry vector. The procedure is exemplified for NNK codon degeneration of MeCP2.

Primer extension and backbone linearization. About 25 µg of the ODN was readily obtained in a 100 µL reaction by primer extension of the freshly annealed complex of o2901 (2.5 µM) and o3111 (2.0 µM) in T4 DNA ligase buffer (50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM ATP, 10 mM dithiothreitol, pH = 7.5; NEB) using 100 µM dNTP Mix (NEB) and 5 U Klenow Fragment (3'→5' exo$^-$; NEB) over 60 min at 37 °C. The reaction was quenched with 10 mM EDTA, concentrated by ice-cold ethanol precipitation and the unincorporated nucleotides removed by silica column-chromatography (Machery-Nagel, Düren, Germany).

The backbone p1570 was linearized in 30 PCR cycles using 0.04 ng/µL template and 0.3 µM of each primer (o2899 and o2900; annealing temperature: 66 °C) in a KOD Hot Start DNA Polymerase PCR (Toyobo, Osaka, Japan; distributed by Merck Millipore, Darmstadt, Germany) according to the manufacturer's instructions. For some backbones, the addition of 2% (v/v)

DMSO was beneficial. The reaction was purified by silica column-chromatography and eluted in 5 mM Tris-HCl (pH = 8.5) pre-heated to 56 °C. Two 50 µL reactions typically yielded 8 µg of the linearized backbone.

Fragment assembly, plasmid pool and library creation. 440 ng of the linearized backbone were assembled with 72 ng of the insert in a 20 µL NEBuilder HiFi DNA Assembly (NEB) reaction at 45 °C for 40 min. Any remainder PCR template was afterwards digested at 37 °C for 60 min with 1 U/µL *Dpn*I. The degenerated library was recovered by ethanol precipitation and transformed in electrocompetent DH10B (TOP10™, Thermo Fisher Scientific) at a target voltage of 1.8 kV over 4 ms. Transformation of 0.15 vol of the reaction in 100 µL (25 OD_{600}) competent cells yielded about 400,000 – 800,000 colonies. A total of 9,000,000 clones was collected in total.

The plasmid pool (p1727) was purified from the spread-plated colonies using a commercial plasmid DNA purification protocol (Machery-Nagel). Degeneracy of the targeted sites was confirmed by Sanger sequencing before the library was transformed into electrocompetent Tuner™(DE3) (Merck Millipore) at sufficiently low multiplicity.

Table 13.7 Degenerated MBD libraries created in this work.

MBD	Degenerated sites*	Primer extension	Backbone	Linearization	Plasmid pool
MBD1[2–81]	V20X, T33X, Y34X, S45X	o2895/o3006	p1566	o2893/o2894	p1938–40
MBD2[146–225]	V164X, V177X, Y178X, S189X	o4441/o3178	p1567	o2896/o4440	p1728-32
MBD3[2–81]	V20X, V33X, Y35X, S45X	o3181/o3179	p1568	o3173/o3174	p1941–3
MBD4[76–167]	K95X, V108X, Y109X, S120X	o3182/o3180	p1569	o3175/o3176	p1944–6
MeCP2[90–181]	K109X, V122X, Y123X, S134X	o2901/o3111	p1570	o2899/o2900	p1727
MeCP2[90–181]	K109X, Q110X, S113X, S116X	o3966/o3111	p1570	o2899/o2900	p2594–5
MeCP2[90–181]	Y120X, V122X, Y123Φ, F132X, S134X	o3967/o3111	p1570	o2899/o2900	p2596–7
MeCP2[90–181]	S116X, G118X, K119X, Y120X	o3968/o3111	p1570	o2899/o2900	p2598

* X = any canonical amino acid or amber stop (TAG); Φ = Cys, Phe, Leu, Trp, Tyr or amber stop.

13.5 Cell surface display of MBDs

Cloning of entry vectors and subcloning for MBD cell surface display. The entry vector for β-D-1-thiogalactopyranoside (IPTG)-inducible cell surface display p1383 was obtained from p1380, a pET-21d(+) (Merck, Darmstadt, Germany) derivate, by replacing the N-terminal SpA(Z)-tag with the AIDA-I surface display cassette (o2912, FragmentGene, GENEWIZ, Leipzig, Germany; amplified with o3004 and o3005) using T4 DNA ligation after *Nde*I/*Xho*I restriction of the plasmid backbone and *Nde*I/*Sal*I restriction of the amplified cassette (all enzymes from New England Biolabs, Frankfurt am Main, Germany). The entry vector for arabinose-inducible cell surface display p1705 was based on pBAD33.1 (Chung & Raetz, 2010) using *Nde*I/*Hin*dIII restriction on both, plasmid backbone and the previously amplified cassette. All vectors had a N-terminal signal peptide from *Vibrio cholerae* enterotoxin binding subunit CtxB followed by an in-frame *Spe*I/*Xho*I cloning site, a human c-Myc epitope encoding sequence and the AIDA-I autotransporter (linker and β-barrel domain) from *Escherichia coli*.

Wildtype MBDs were subcloned from the respective expression vectors using *Spe*I/*Xho*I restriction-ligation cloning. The degenerated libraries were obtained by plasmid linearization and a modified Gibson assembly protocol (see elsewhere).

Outer-membrane extraction. To verify the translocation of the AIDA-I payload to the outer membrane, a protocol of Park et al. (2015) was followed. In brief, an OD_{600} of 0.8 cm^{-1} of the induced bacteria was treated in 100 µL 200 mM Tris-HCl, 20 mM glucose, 0.2 mM EDTA (pH = 8.0) with 2 µg lysozyme for 10 min at room temperature before PMSF was added to a final concentration of 1 mM. The outer membranes were disrupted by adding 100 µL 50 mM Tris-HCl, 10 mM $MgCl_2$, 2% Triton X-100 and 1 U DNase I (NEB), 30 min on ice and recovered as supernatant after centrifugation at 1500 × *g* for 5 min s. This fraction was washed twice with phosphate-buffered saline (PBS), recovering the membrane-associated protein fractions at 20,000 × *g*, 15 min.

13.6 MBD–DNA binding assays and MBD library screening

Bacterial cell surface display of MBDs with pET-AIDA-I. From a single clone of a freshly transformed strain of *E. coli* Tuner™(DE3) (Novagen, Merck Millipore) an overnight culture was prepared in LB medium supplemented with antibiotics and 20 mM glucose (0.36% w/v). The next morning, 1.5 mL LB medium with antibiotic were inoculated at an OD_{600} of 0.05 cm^{-1} for the monoclonal strains. Glycerol stocks of the pooled libraries were inoculated directly from aliquots of suitable density. The cultures were grown to exponential phase at 220 rpm, 37 °C for 2 h until the OD_{600} reached 0.4 – 0.6 cm^{-1}. Overgrown cultures and clones kept at 4 °C for more than 2 weeks may give unsatisfactory results. After placing the cultures briefly on ice, the expression of the AIDA-I cassette was induced with 50 µM β-D-1-thiogalactopyranoside. Higher inducer concentrations lead to higher display levels, but decreased survival of the cells. Expression was allowed to proceed with the cultures shaking at 150 rpm, 30 °C for 1 – 2 h.

The OD_{600} to collect from these cultures was chosen according to the type of downstream assay, the number of staining reactions and the events required: Per 10 reactions that yield about 600,000 to 800,000 cells each on the FACS instrument, an OD_{600} of 0.4 cm^{-1} was harvested, pelletized at 8,000 × *g* for 2 min and washed twice with the same volume phosphate-buffered saline

(PBS). The pellet was resuspended in 200 µL PBS with 0.1% bovine serum albumin (ultra pure grade, else omit; Cell Signaling Technology, Danvers, MA, U. S.) and incubated for 30 – 60 min on ice. Required volumes:

- Per binding or selectivity measurement, 1.0 vol corresponded to 20 µL, but as little as 5 – 10 µL could be handled with experience.
- For the initial screening of a pooled library, the volume was scaled to oversample the nominal degeneracy at least tenfold, typically, 400 µL of the suspension were used.
- In the subsequent screenings, 20 µL were sufficient.

General staining and DNA labeling procedure. Per staining, 1.0 vol of the cell suspension (described above) and 0.5 vol of the staining reagent (described below) were combined and incubated for 20 – 60 min at 21 °C, 700 rpm to keep the cells in suspension. Then, the bacteria were pelletized at 8,000 × *g* for 2 min, washed with 2.0 vol, but at least 80 µL, ice-cold PBS and resuspended in 6.5 vol PBS (minimum 60 µL) and kept on ice for flow cytometry or FACS. In a typical staining reaction, the labeled DNA probe had a concentration of 64 nM and was carried out in presence of 500 nM competitor DNA.

The staining reagents were prepared as follows: Two complementary oligodeoxynucleotides (ODN; **Table 13.8**), typically 5'-biotinylated including a triethyleneglycol spacer, were hybridized to give the unlabeled DNA probe. It was crucial for the accurate assessment of binding selectivity to confirm the concentration of all stocks spectrophotometrically and to adjust the stocks at the indicated stages if needed. For the 24-mers, an extinction coefficient of 286 $\mathrm{mM^{-1}\,cm^{-1}}$ was used for the oligo(A) context and 197 $\mathrm{mM^{-1}\,cm^{-1}}$ for the oligo(T) context.

First, the ODN were brought to a concentration of 50 µM in water and 1 nmol (20 µL) of the complementary single-strands annealed in 100 µL 2 × EMSA buffer by immersion in boiling water slowly brought to room temperature. These stocks were adjusted to 5 µM (A260 = 1.5 $\mathrm{cm^{-1}}$ for the 24-mers) with water (ideally, 100 µL), aliquoted and stored at −20 °C.

For the labeling of the DNA probe, two reaction components (A and B) were prepared. Per 10 labeling reactions (with 20 µL for 1 vol), component A was 30 µL 10 × EMSA buffer and 6 µL 50 m A/T competitor (o3416 and o3417) supplemented with 3.84 µL of the 5 µM stocks brought to 50 µL with water. Component B contains the fluorochrome-streptavidin conjugate (all from Biozol, Eching, Germany) at a final concentration of the streptavidin component of 1.15 µM in 0.4 g/L BSA, 6 mM tris(2-carboxyethyl)phosphin. This component must not be frozen and is stable for up to two months when stored at 4 °C shielded from light. Equal volumes of A and B are combined on ice to label the DNA probe by adding A slowly into B to effect that the diluted DNA probe with multiple biotin sites is saturated in the concentrated solution of excess labeling

Table 13.8 Combinations of ODNs to create DNA probes for FACS binding assays.

Combination*	Context†	Sequence source	Watson strand	Crick strand	Other modifications
C/C	AAAAACGAAAAA	artificial	o3244	o3245	2 × biotin
C/5mC	AAAAACGAAAAA	artificial	o2905	o2909	1 × Pacific Blue
C/5mC	AAAAACGAAAAA	artificial	o2906	o2909	1 × FAM
C/5mC	AAAAACGAAAAA	artificial	o2908	o2909	1 × TAMRA
C/5mC	AAAAACGAAAAA	artificial	o2906	o3081	1 × biotin
C/5mC	AAAAACGAAAAA	artificial	o2906	o3082	2 × biotin, 1 × FAM
C/5mC	AAAAACGAAAAA	artificial	o2906	o3083	3 × biotin, 1 × FAM
5mC/C	AAAAACGAAAAA	artificial	o3081	o3245	2 × biotin
5hmC/C	AAAAACGAAAAA	artificial	o3211	o3245	2 × biotin
5fC/C	AAAAACGAAAAA	artificial	o3212	o3245	2 × biotin
5caC/C	AAAAACGAAAAA	artificial	o3213	o3245	2 × biotin
5mC/5mC	AAAAACGAAAAA	artificial	o2967	o3081	1 × biotin
5mC/5mC	AAAAACGAAAAA	artificial	o2967	o3082	2 × biotin, 1 × FAM
5mC/5mC	AAAAACGAAAAA	artificial	o3081	o3214	2 × biotin
5mC/5mC	AAAAACGAAAAA	artificial	o2967	o3083	3 × biotin, 1 × FAM
5hmC/5mC	AAAAACGAAAAA	artificial	o3211	o3214	2 × biotin
5fC/5mC	AAAAACGAAAAA	artificial	o3212	o3214	2 × biotin
5fC/5mC	AAAAACGAAAAA	artificial	o3213	o3214	2 × biotin
5hmC/5hmC	AAAAACGAAAAA	artificial	o3211	o3215	2 × biotin
5fC/5hmC	AAAAACGAAAAA	artificial	o3212	o3215	2 × biotin
5caC/5hmC	AAAAACGAAAAA	artificial	o3213	o3215	2 × biotin
5fC/5fC	AAAAACGAAAAA	artificial	o3212	o3216	2 × biotin
5caC/5fC	AAAAACGAAAAA	artificial	o3213	o3216	2 × biotin
5caC/5caC	AAAAACGAAAAA	artificial	o3213	o3217	2 × biotin

* Strand-diametral cytosines X/Y in a $\mathrm{Xp}\overline{\mathrm{Y}} \cdot \mathrm{Yp}\overline{\mathrm{X}}$ dyad, i. e., C/T = CpA · TpG and C/C = CpG · CpG.

† Refers to the Watson strand.

agent in order to prevent the formation of alternating DNA-streptavidin chains. Labeling was allowed to proceed for 60 min. The labeling reaction was 192 nM DNA probe, 576 nM streptavidin conjugate, 1.5 μM competitor DNA, 0.2 g/L BSA, 3 mM tris(2-carboxyethyl)phosphin in 1 × EMSA buffer.

One-color DNA binding assay. Labeling and staining was carried out as described above with 20 μL bacterial cell suspension and 10 μL of the staining reagent. No quenching was required. The cells were washed and analyzed on the flow cytometer to assess relative binding affinities.

Two-color DNA binding assay. According to the screening goal, staining mixes were prepared as described above using different streptavidin conjugates. It was convenient to prepare the drop-out mix with equal volumes of the DNA stocks diluted in component A before labeling with component B. The labeled mixtures were quenched with a total of 0.1 vol 120 µM biotin before combining at the desired molar ratios. Typically, if a single 'on-target' labeled with streptavidin-phycoerythrin was screened against 14 'off-target's labeled with streptavidin-FITC, 1/15 parts of the first was combined with 14/15 parts of the latter. The combined mixes were immediately added to the cell suspension.

Determination of surface display levels by antibody staining. To determine the surface display level of full-length payloads, the entry vector p1383 contained an in-frame c-Myc epitope sequence C-terminal of the *Xho*I/*Spe*I cloning site. Detection of the Myc epitope was carried out either in a single step with an allophycocyanin (APC)-coupled Anti-c-Myc epitope antibody (Abcam, Cambridge, U. K.; 1:15) or sequentially by staining with a biotinylated Anti-c-Myc epitope antibody (Thermo Fisher Scientific, Schwerte, Germany) first and then labeling with a fluorochrome-streptavidin conjugate, e. g., Brilliant Violet™ 711-streptavidin (Biozol, Eching, Germany). After each staining, the cells were washed twice with 6.5 vol PBS.

Flow cytometry and fluorescence-activated cell sorting. The bacteria were processed on a SH800SFP Cell Sorter (Sony Biotechnology, Weybridge, U. K.) using a 70 µm (for analysis) or 100 µm (for sorting) microfluidic sorting chip (Sony Biotechnology). The flow cytometer was equipped with a 405 nm, 488 nm, 561 nm and 638 nm laser. The "Filter Pattern 2" was used to split the collinear fluorescence beam. Fluorescence of fluorescein or AF488 was acquired in detector "FL2" at 65% photomultiplier tube (PMT) voltage, and the optimized setups for phycoerythrin ("FL3", 45% PMT voltage), allophycocyanin ("FL4", 45% PMT voltage), Brilliant Violet™ 711 ("FL5", 40% PMT voltage). Regular cells from a control population were gated at a forward scatter intensity of 16 a. u. and a backward scatter 40% PMT voltage. Before loading, samples were resuspended by pipetting up and down and analyzed at about 10,000 events per second or sorted at about 100 events per second.

Cells were sorted in single-cell mode either directly onto one-well plates with selective LB agar or using the sorting mode indicated in the text in 1.5 mL reaction tubes pre-filled with S. O. C. medium and later preferably spread-plated onto LB agar plates, scraped and stocked. Outgrowth in liquid LB medium at 37 °C and 150 rpm overnight with immediate proceeding the following morning did not have adverse effects if the population size is small. Glycerol stocks were prepared after the first outgrowth by combining 300 µL 50% sterile-filtered glycerol with 700 µL liquid culture or the scraped material.

Optimization of the number of biotin sites. A modified, unoptimized single-color DNA binding assay was performed using probes with a hemimethylated 5mC/C or fully methylated 5mC/5mC CpG based on o3081, o3082 and o3083 (**Table 13.8**).

Determination of assay specificity and sensitivity. To determine the false-positive and true-negative rate two-color DNA binding assay was carried out as described above using bacteria displaying MBD2 (p1567) of an empty pET-

AIDA (p1383) and an equimolar mixture of a DNA probe containing an unmodified C/C CpG (o3244, o3245) or a fully methylated 5mC/5mC CpG (o3801, o3214) labeled as indicated in Figure 7.9. A total of 15,000 regular cells was sampled from the recorded data and the fraction of double-positive events was determined in dependence of the threshold levels in *R*. Typically, a random sample of 10,000 gated events was analyzed per condition.

General analysis of flow cytometric data. Gating was carried out using the instrument's Cell Sorter Software v2.1.5 (Sony Biotechnology). The data was exported for processing and analysis using *R*. The flowCore v1.11.20 package (Hahne, et al., 2009) was used for extracting and reprocessing Flow Cytometry Standard (FCS) files. A particular gating strategy was rebuilt if needed computationally using openCyto v2.0.0 (Finak, Greg, et al., 2014). Typically, 60 – 70% of all events acquired under the above acquisition settings were 'cells', so that also for rebuilding an ellipsoid gate based on forward and sideward scatter data, the 70% quantile was used. The ggplot2 v3.3.5 (Wickham, 2016) extensions ggCyto v1.16.0 (Van, et al., 2018) and ggridges v0.5.3 produced the data visualizations presented in this work. Where monochromatic color scales were used, the scaled event densities in 2-dimensional FACS dot plots were based on the square root of the event count per bin to improve the visibility of low-density regions.

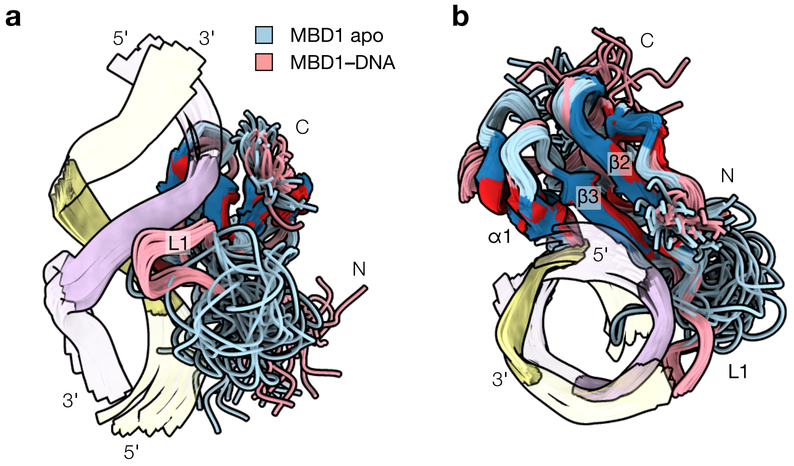
Next-generation sequencing data after FACS. For plasmid extraction, glycerol stocks of the selected populations were inoculated in LB medium with antibiotic and grown at 30 °C, 150 rpm to an OD_{600} of 4.0 cm^{-1}. The degenerated region of all plasmids (50 ng) was flanked in a 25 µL Phusion® PCR (NEB) with o3861 and o3862 which introduced random hexamers (NNNNNN) as unique molecular identifiers (UMIs) in 2 cycles with an annealing temperature of 60 °C and 20 s elongation. 2 µL of this reaction were transferred to a fresh 25 µL Phusion PCR to barcode the samples in 25 cycles using o2363/o3033, o2364/o3034, etc. The reactions were purified and pooled according to their clone number.

Illumina NGS was carried out by an external service provider (GENEWIZ Germany, Leipzig, Germany). The paired-end reads were merged using PANDAseq (v2.11 Masella, et al., 2012) and aligned with bbmap (v36.86; Bushnell, 2019) in semiperfect mode to the reference sequence which contained "NNN" at the positions of the degenerated codons, the position of the UMIs and the position of the barcodes of the second PCR amplification. The bam files were filtered, imported, and the sequences at the degenerated sites extracted using packages from the *R* Bioconductor suite (Lawrence, et al., 2013; Pagès, et al., 2016). Only reads that had a mapping quality of 13 or higher and which showed the expected barcode pairs on both ends of their ends (which applied to roughly 70% of all reads) were kept. Distinct sequences were established based on UMIs and the codons present at the degenerated sites.

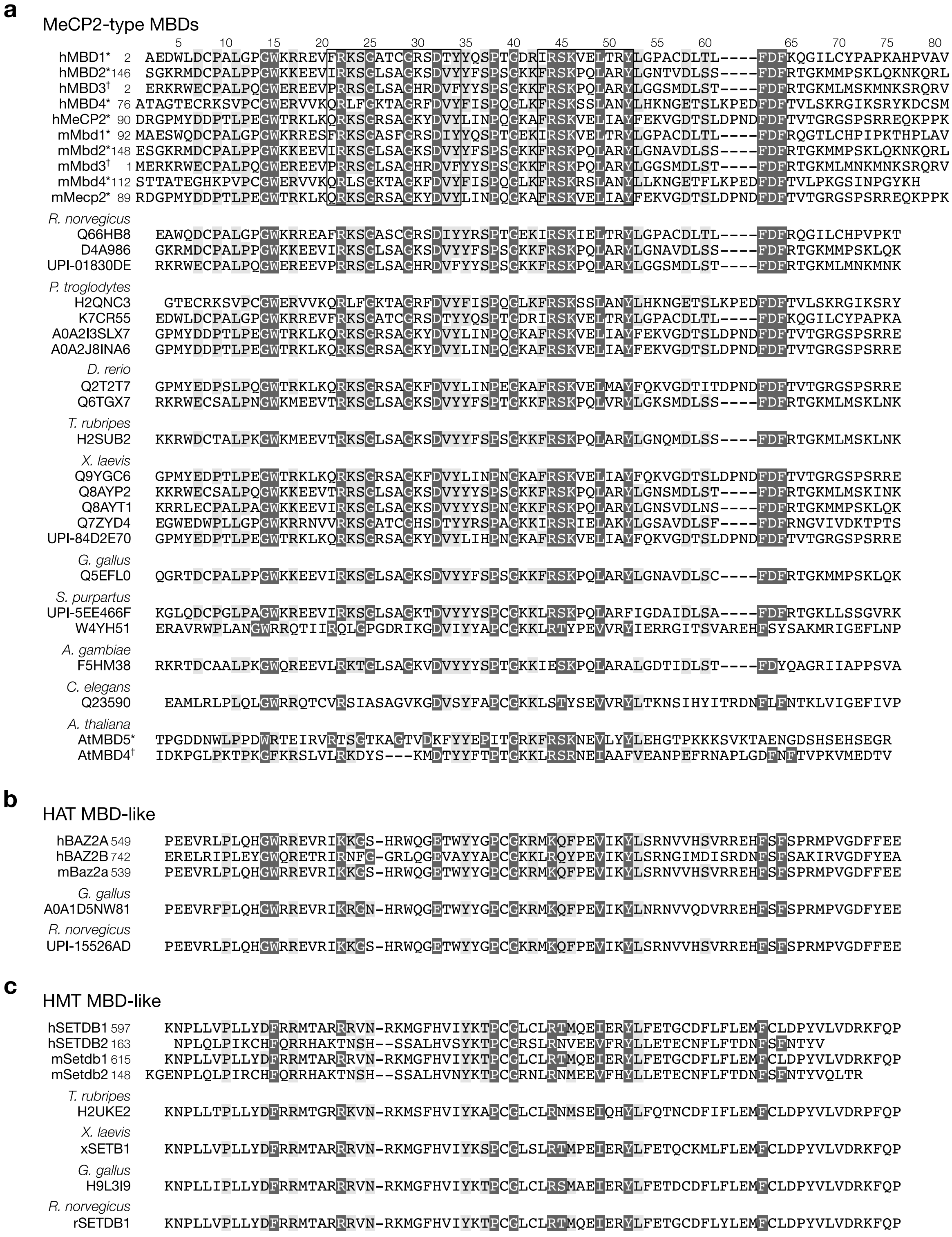
To calculate the abundance of a genotype or the respective phenotype shown in Figure 7.14, the fraction of UMI counts within the number of distinct sequences was used. For genotypes or phenotypes not present in the sequencing of the initial library (due to under sampling of more than 1 million clones with 2 × 50,000 reads), a missed-by-one assumption was made. The amino acid enrichment per position was determined from the total of distinct codon–UMI combinations. Source code at page 193.

A
Supplementary Information

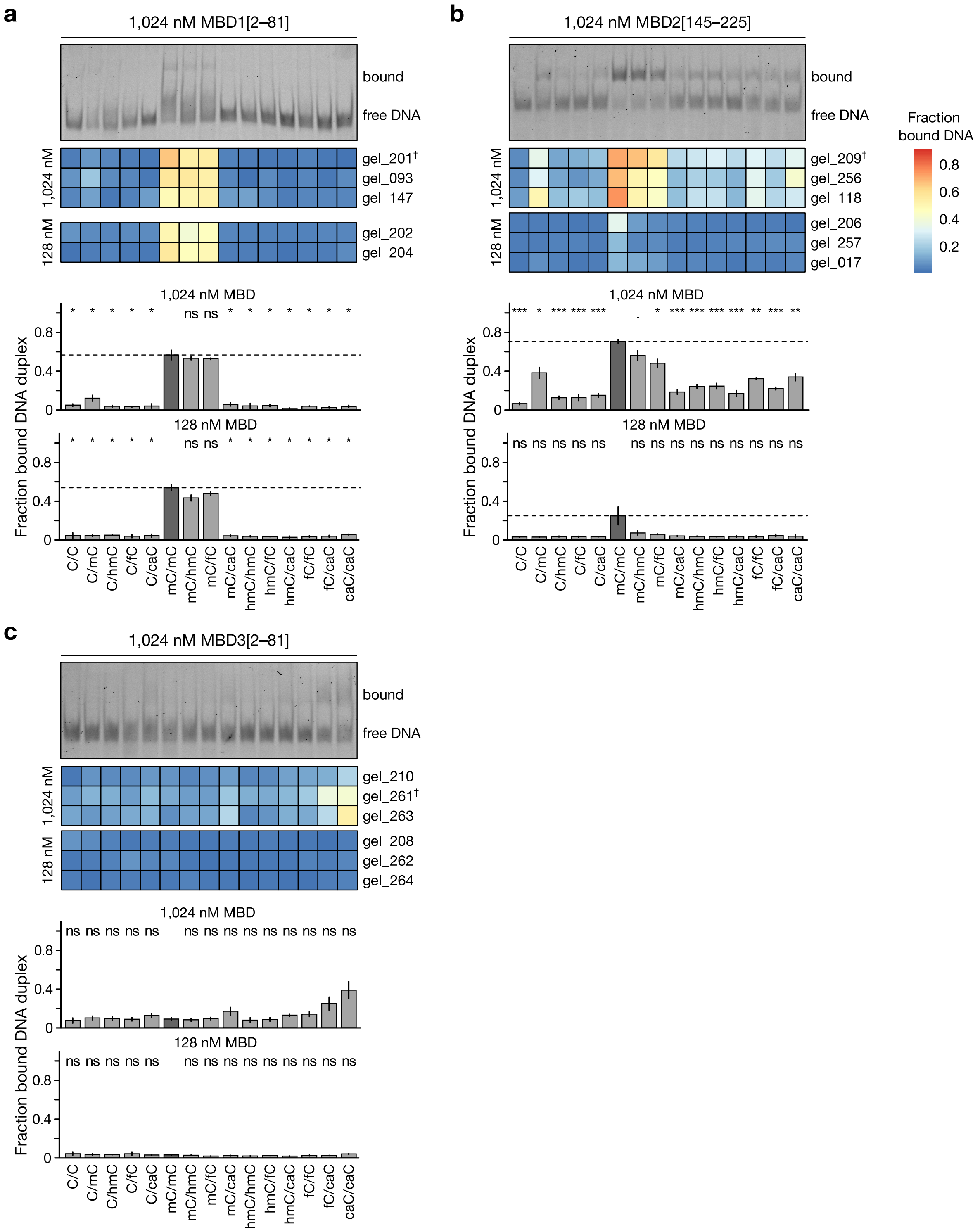
A.1 Supplementary Figures



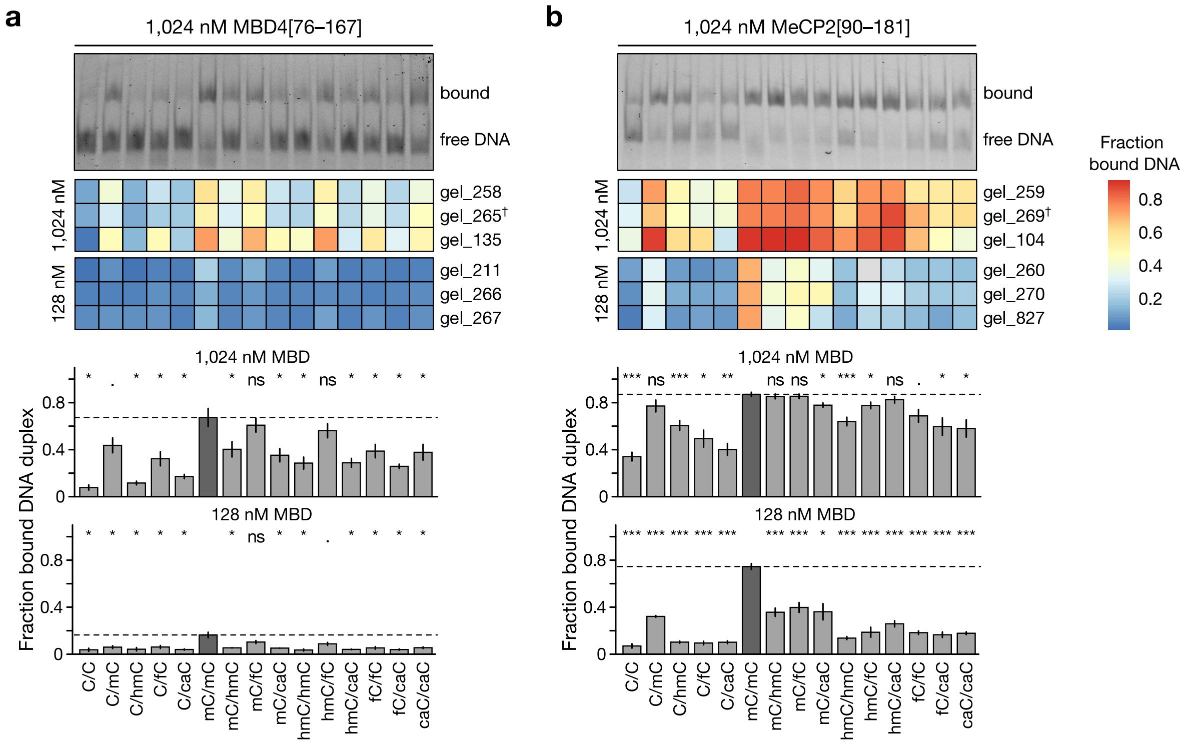
Supplementary Figure A.1 Free and DNA-bound form of MBD1. Superimpositions of NMR structures of MBD1 determined in presence (PDB 1ig4, Ohki, et al., 1999) or absence of a methylated double-stranded DNA duplex (PDB 1d9n, *ibid*).


a
MeCP2-type MBDs
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80
hMBD1* 2 AEDWLDCPALGPGWKRREVFRKSGATCGRSDTYYQSPTGDRIRSKVELTRYLGPACDLTL----FDFKQGILCYPAPKAHPVAV
hMBD2*146 SGKRMDCPALPPGWKKEEVIRKSGLSAGKSDVYYFSPSGKKFRSKPQLARYLGNTVDLSS----FDFRTGKMMPSKLQKNKQRL
hMBD3† 2 ERKRWECPALPQGWEREEVPRRSGLSAGHRDVFYYSPSGKKFRSKPQLARYLGGSMDLST----FDFRTGKMLMSKMNKSRQRV
hMBD4* 76 ATAGTECRKSVPCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRYKDCSM
hMeCP2* 90 DRGPMYDDPTLPEGWTRKLKQRKSGRSAGKYDVYLINPQGKAFRSKVELIAYFEKVGDTSLDPNDFDFTVTGRGSPSRREQKPPK
mMbd1* 92 MAESWQDCPALGPGWKRRESFRKSGASFGRSDIYYQSPTGEKIRSKVELTRYLGPACDLTL----FDFRQGTLCHPIPKTHPLAV
mMbd2*148 ESGKRMDCPALPPGWKKEEVIRKSGLSAGKSDVYYFSPSGKKFRSKPQLARYLGNAVDLSS----FDFRTGKMMPSKLQKNKQRL
mMbd3† 1 MERKRWECPALPQGWEREEVPRRSGLSAGHRDVFYYSPSGKKFRSKPQLARYLGGSMDLST----FDFRTGKMLMNKMNKSRQRV
mMbd4*112 STTATEGHKPVPCGWERVVKQRLSGKTAGKFDVYFISPQGLKFRSKRSLANYLLKNGETFLKPEDFDFTVLPKGSINPGYKH
mMecp2* 89 RDGPMYDDPTLPEGWTRKLKQRKSGRSAGKYDVYLINPQGKAFRSKVELIAYFEKVGDTSLDPNDFDFTVTGRGSPSRREQKPPK
R. norvegicus
Q66HB8 EAWQDCPALGPGWKRREAFRKSGASCGRSDIYYRSPTGEKIRSKIELTRYLGPACDLTL----FDFRQGILCHPVPKT
D4A986 GKRMDCPALPPGWKKEEVIRKSGLSAGKSDVYYFSPSGKKFRSKPQLARYLGNAVDLSS----FDFRTGKMMPSKLQK
UPI-01830DE RKRWECPALPQGWEREEVPRRSGLSAGHRDVFYYSPSGKKFRSKPQLARYLGGSMDLST----FDFRTGKMLMNKMNK
P. troglodytes
H2QNC3 GTECRKSVPCGWERVVKQRLFGKTAGRFDVYFISPQGLKFRSKSSLANYLHKNGETSLKPEDFDFTVLSKRGIKSRY
K7CR55 EDWLDCPALGPGWKRREVFRKSGATCGRSDTYYQSPTGDRIRSKVELTRYLGPACDLTL----FDFKQGILCYPAPKA
A0A2I3SLX7 GPMYDDPTLPEGWTRKLKQRKSGRSAGKYDVYLINPQGKAFRSKVELIAYFEKVGDTSLDPNDFDFTVTGRGSPSRRE
A0A2J8INA6 GPMYDDPTLPEGWTRKLKQRKSGRSAGKYDVYLINPQGKAFRSKVELIAYFEKVGDTSLDPNDFDFTVTGRGSPSRRE
D. rerio
Q2T2T7 GPMYEDPSLPQGWTRKLKQRKSGRSAGKFDVYLINPEGKAFRSKVELMAYFQKVGDTITDPNDFDFTVTGRGSPSRRE
Q6TGX7 RKRWECSALPNGWKMEEVTRKSGLSAGKSDVYYFSPTGKKFRSKPQLVRYLGKSMDLSS----FDFRTGKMLMSKLNK
T. rubripes
H2SUB2 KKRWDCTALPKGWKMEEVTRKSGLSAGKSDVYYFSPSGKKFRSKPQLARYLGNQMDLSS----FDFRTGKMLMSKLNK
X. laevis
Q9YGC6 GPMYEDPTLPEGWTRKLKQRKSGRSAGKFDVYLINPNGKAFRSKVELIAYFQKVGDTSLDPNDFDFTVTGRGSPSRRE
Q8AYP2 KKRWECSALPQGWKKEEVTRRSGLSAGKSDVYYYSPSGKKFRSKPQLARYLGNSMDLST----FDFRTGKMLMSKINK
Q8AYT1 KRRLECPALPAGWKKEEVIRKSGLSAGKSDVYYYSPNGKKFRSKPQLARYLGNSVDLNS----FDFRTGKMMPSKLQK
Q7ZYD4 EGWEDWPLLGPGWKRRNVVRKSGATCGHSDTYYRSPAGKKIRSRIELAKYLGSAVDLSF----FDFRNGVIVDKTPTS
UPI-84D2E70 GPMYEDPTLPEGWTRKLKQRKSGRSAGKYDVYLIHPNGKAFRSKVELIAYFQKVGDTSLDPNDFDFTVTGRGSPSRRE
G. gallus
Q5EFL0 QGRTDCPALPPGWKKEEVIRKSGLSAGKSDVYYFSPSGKKFRSKPQLARYLGNAVDLSC----FDFRTGKMMPSKLQK
S. purpartus
UPI-5EE466F KGLQDCPGLPAGWKREEVIRKSGLSAGKTDVYYYSPCGKKLRSKPQLARFIGDAIDLSA----FDFRTGKLLSSGVRK
W4YH51 ERAVRWPLANGWRRQTIIRQLGPGDRIKGDVIYYAPCGKKLRTYPEVVRYIERRGITSVAREHFSYSAKMRIGEFLNP
A. gambiae
F5HM38 RKRTDCAALPKGWQREEVLRKTGLSAGKVDVYYYSPTGKKIESKPQLARALGDTIDLST----FDYQAGRIIAPPSVA
C. elegans
Q23590 EAMLRLPLQLGWRRQTCVRSIASAGVKGDVSYFAPCGKKLSTYSEVVRYLTKNSIHYITRDNFLFNTKLVIGEFIVP
A. thaliana
AtMBD5* TPGDDNWLPPDWRTEIRVRTSGTKAGTVDKFYYEPITGRKFRSKNEVLYYLEHGTPKKKSVKTAENGDSHSEHSEGR
AtMBD4† IDKPGLPKTPKGFKRSLVLRKDYS---KMDTYYFTPTGKKLRSRNEIAAFVEANPEFRNAPLGDFNFTVPKVMEDTV
b
HAT MBD-like
hBAZ2A 549 PEEVRLPLQHGWRREVRIKKGS-HRWQGETWYYGPCGKRMKQFPEVIKYLSRNVVHSVRREHFSFSPRMPVGDFFEE
hBAZ2B 742 ERELRIPLEYGWQRETRIRNFG-GRLQGEVAYYAPCGKKLRQYPEVIKYLSRNGIMDISRDNFSFSAKIRVGDFYEA
mBaz2a 539 PEEVRLPLQHGWRREVRIKKGS-HRWQGETWYYGPCGKRMKQFPEVIKYLSRNVVHSVRREHFSFSPRMPVGDFFEE
G. gallus
A0A1D5NW81 PEEVRFPLQHGWRREVRIKRGN-HRWQGETWYYGPCGKRMKQFPEVIKYLNRNVVQDVRREHFSFSPRMPVGDFYEE
R. norvegicus
UPI-15526AD PEEVRLPLQHGWRREVRIKKGS-HRWQGETWYYGPCGKRMKQFPEVIKYLSRNVVHSVRREHFSFSPRMPVGDFFEE
c
HMT MBD-like
hSETDB1 597 KNPLLVPLLYDFRRMTARRRVN-RKMGFHVIYKTPCGLCLRTMQEIERYLFETGCDFLFLEMFCLDPYVLVDRKFQP
hSETDB2 163 NPLQLPIKCHFQRRHAKTNSH--SSALHVSYKTPCGRSLRNVEEVFRYLLETECNFLFTDNFSFNTYV
mSetdb1 615 KNPLLVPLLYDFRRMTARRRVN-RKMGFHVIYKTPCGLCLRTMQEIERYLFETGCDFLFLEMFCLDPYVLVDRKFQP
mSetdb2 148 KGENPLQLPIRCHFQRRHAKTNSH--SSALHVNYKTPCGRNLRNMEEVFHYLLETECNFLFTDNFSFNTYVQLTR
T. rubripes
H2UKE2 KNPLLTPLLYDFRRMTGRRKVN-RKMSFHVIYKAPCGLCLRNMSEIQHYLFQTNCDFIFLEMFCLDPYVLVDRPFQP
X. laevis
xSETB1 KNPLLVPLLYDFRRMTARRRVN-RKMGFHVIYKSPCGLSLRTMPEIERYLFETQCKMLFLEMFCLDPYVLVDRKFQP
G. gallus
H9L3I9 KNPLLIPLLYDFRRMTARRRVN-RKMGFHVIYKTPCGLCLRSMAEIERYLFETDCDFLFLEMFCLDPYVLVDRKFQP
R. norvegicus
rSETDB1 KNPLLVPLLYDFRRMTARRRVN-RKMGFHVIYKTPCGLCLRTMQEIERYLFETGCDFLYLEMFCLDPYVLVDRKFQP

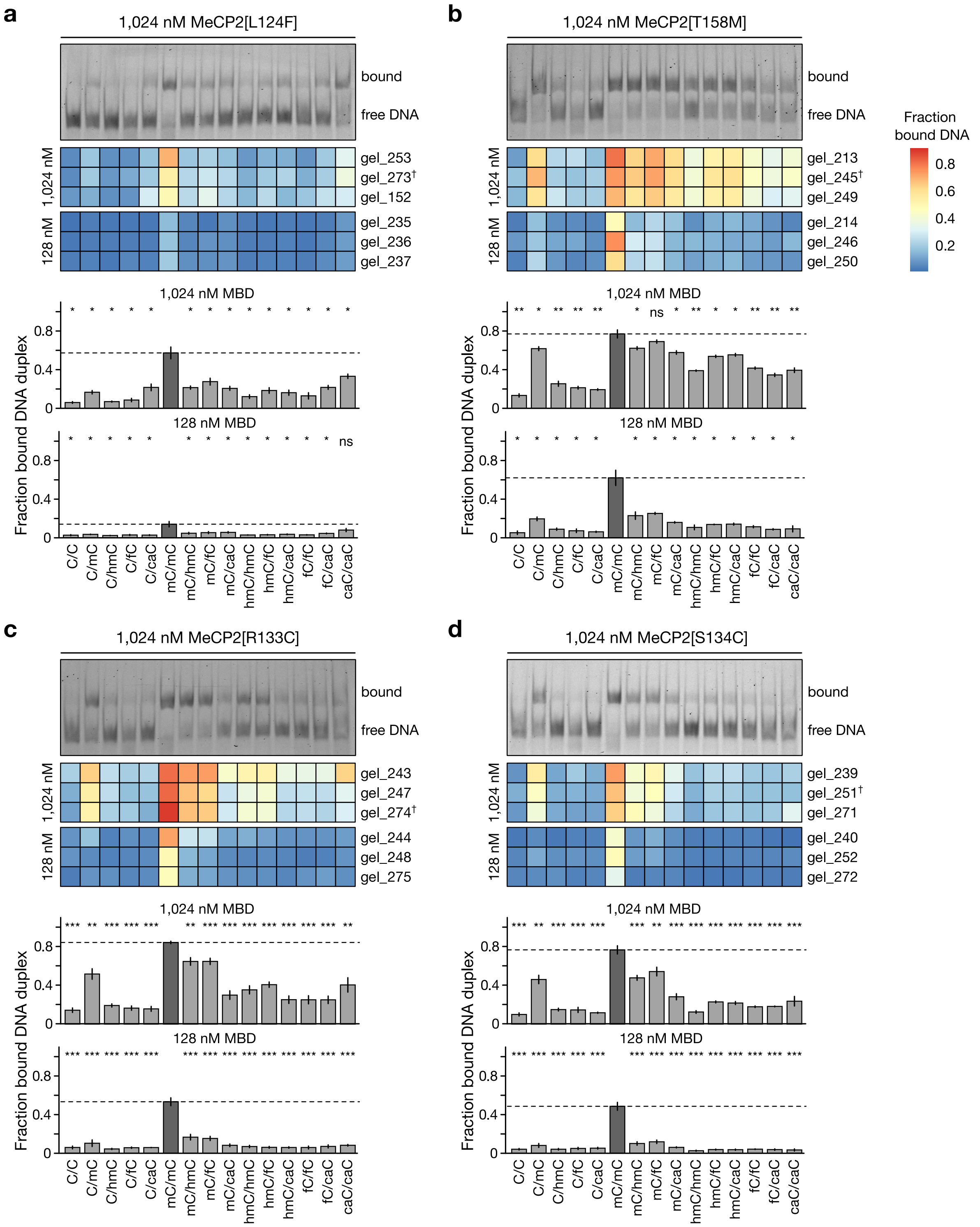
Supplementary Figure A.2 Sequence alignment of the MBDs. Sequence alignment of **(a)** MeCP2-like MBDs, **(b)** HAT MBD-type MBDs and **(c)** HMT MBD-type MBDs. Note the presence of vertebrate human (h), mouse (m), invertebrate and plant proteins in the MeCP2-like MBD group. Not all sequences retrieved under Pfam PF01429 included, highly similar ones omitted in favor of displaying more dissimilar ones. Sequences with proven (*) or disproven (†) ability to bind methylated CpGs are indicated.



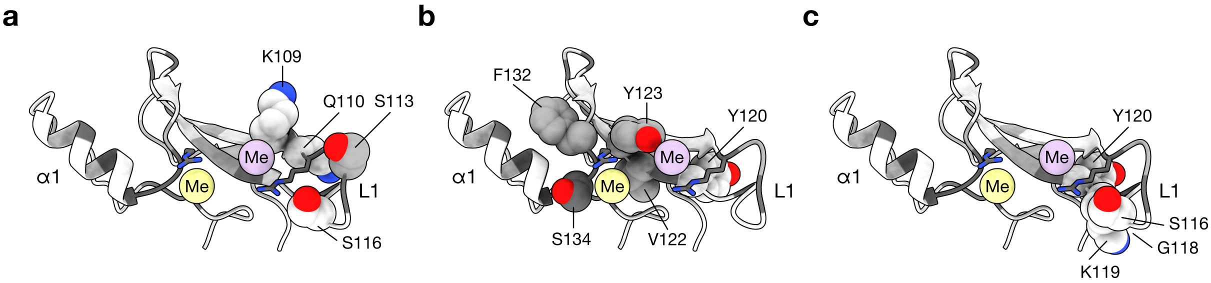
Supplementary Figure A.3 Compilation of primary EMSA data evaluated for MBD1, MBD2 and MBD3. Representative gel images of an electrophoretic mobility shift assay at 1,024 nM MBD concentration is shown in the top panel, all other measurements as heat maps in the middle (dagger indicates the gel image shown); The fractions of bound DNA duplex were averaged and summarized as bar graphs in the lower panels. Error bars indicate standard error of the mean (SEM) with a two-sides Student's *t*-test against the fraction of bound 5mC/5mC duplex; False-discovery rate was controlled using the Benjamini-Hochberg procedure to correct the *p*-values for multiple comparisons; ns: *p*-value in (0.1, 1], . (0.05, 1], * (0.01, 0.05], ** (0.001, 0.01], *** (0, 0.001]. Reprinted from Buchmuller et al. (2020), CC BY 4.0 (full license in Appendix A.4).



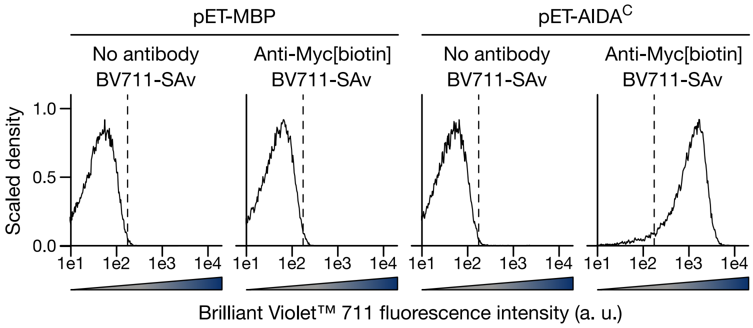
Supplementary Figure A.4 Compilation of primary EMSA data evaluated for MBD4 and MeCP2. Details see Supplementary Figure A.3. Reprinted from Buchmuller et al. (2020), CC BY 4.0 (full license in Appendix A.4).



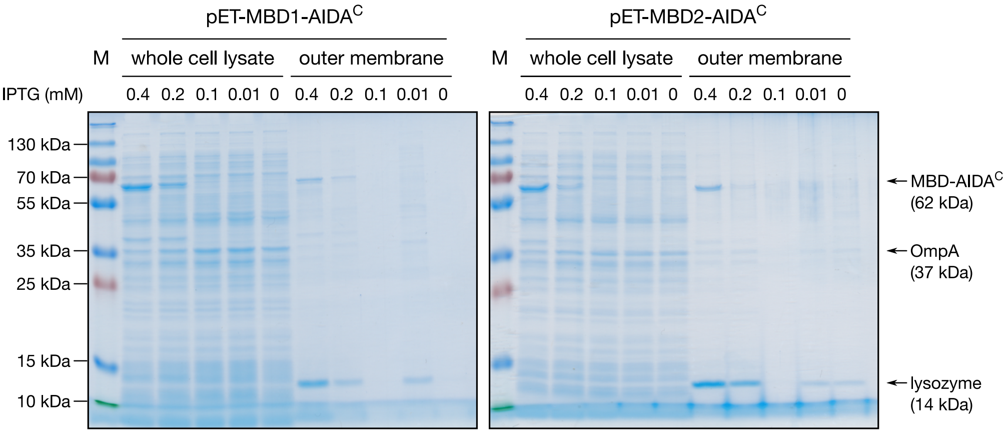
Supplementary Figure A.5 Compilation of primary EMSA data evaluated for MeCP2 Rett mutants. Representative gel images of an electrophoretic mobility shift assay at 1,024 nM MBD concentration is shown in the top panel, all other measurements as heat maps in the middle (dagger indicates the gel image shown); The fractions of bound DNA duplex were averaged and summarized as bar graphs in the lower panels. Error bars indicate standard error of the mean (SEM) with a two-sides Student's *t*-test against the fraction of bound 5mC/5mC duplex; False-discovery rate was controlled using the Benjamini-Hochberg procedure to correct the *p*-values for multiple comparisons; ns: *p*-value in (0.1, 1], . (0.05, 1], * (0.01, 0.05], ** (0.001, 0.01], *** (0, 0.001]. Reprinted from Buchmuller et al. (2020), CC BY 4.0 (full license in Appendix A.4).



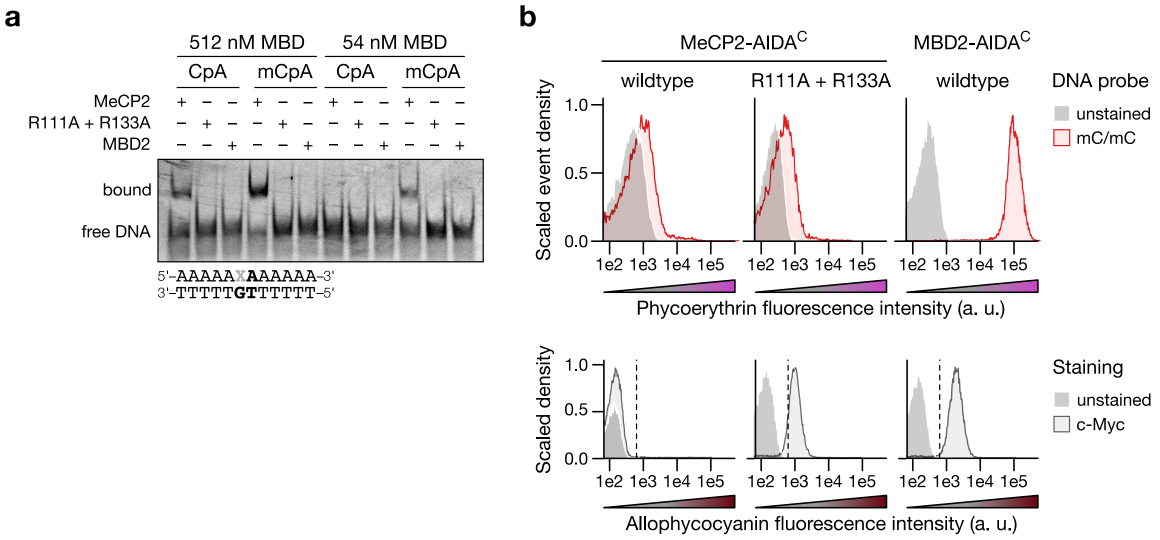
Supplementary Figure A.6 Positions of the degenerated amino acid residues in additional MeCP2 libraries. Three libraries based on MeCP2[90–181] (PDB 3c2i, Ho, et al., 2008) in **(a–c)** with α-helix α1 and loop L1 of the MBD; Only the *sp*^{3} carbon of each 5mC (Me) is indicated for clarity.



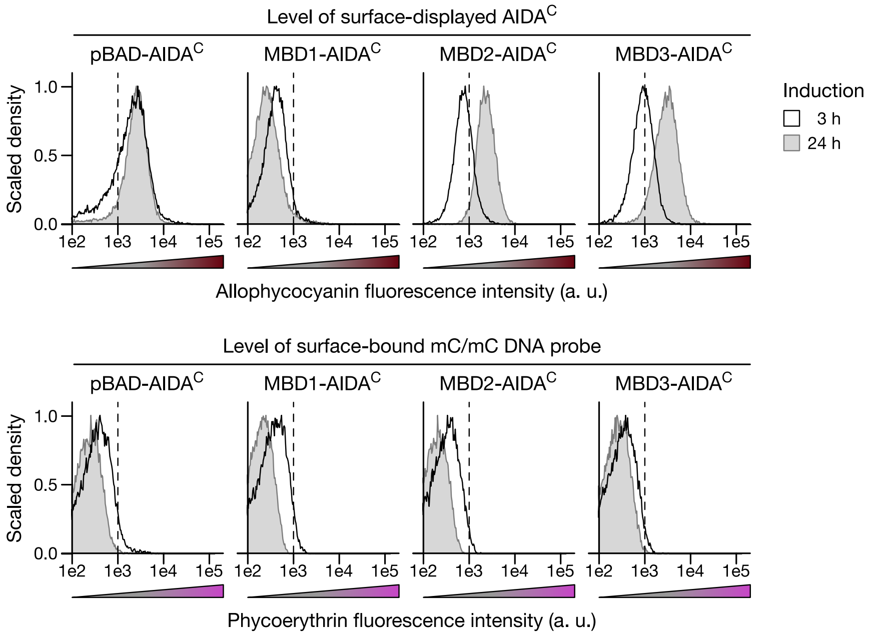
Supplementary Figure A.7 Specificity of Anti-c-Myc epitope detection of AIDA-I surface display. *E. coli* B strain Tuner™(DE3) cells expressing either maltose-binding protein (MBP) or the empty c-Myc epitope-containing AIDA^C autotransporter were stained with Brilliant Violet™ 711-labeled streptavidin (SAv) after treatment with a biotinylated Anti-c-Myc epitope antibody or not.



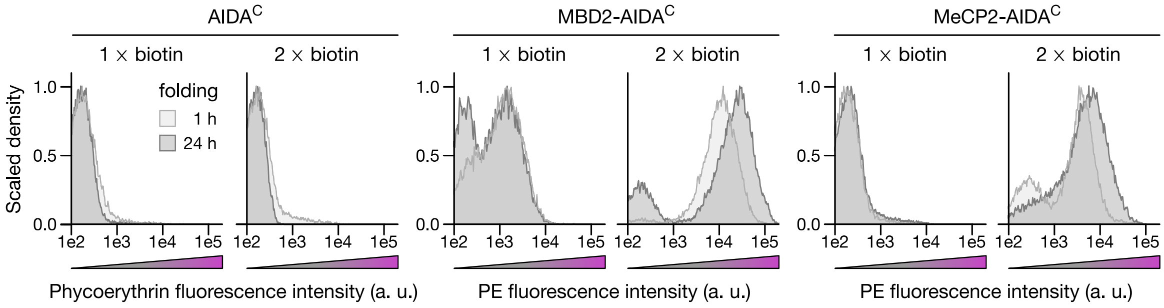
Supplementary Figure A.8 Whole cell lysate and outer membrane analysis of surface-displayed MBD1 and MBD2. Surface-display of MBDs was induced for 60 min at 30 °C and cells harvested for outer-membrane extraction which uses lysozyme to digest the bacterial cell wall; Not all extractions were successful. Separation on 12% glycine-SDS PAGE; PageRuler™ Plus Prestained Protein Ladder (Lane M).



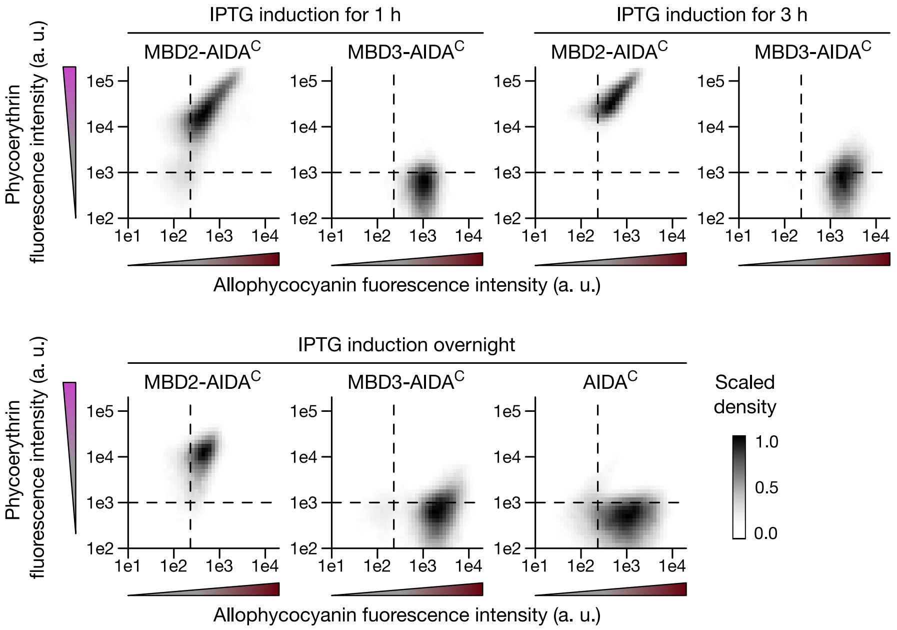
Supplementary Figure A.9 Binding of MeCP2 to methylated and unmethylated CpA dinucleotides. (a) Electrophoretic mobility shift assay with recombinantly expressed MBDs and fluorescently-labeled DNA probes containing a single modified or unmodified CpA. **(b)** Cell surface display levels of MBD-AIDA^C proteins and binding of a probe with the same DNA sequence as in *a* labeled with phycoerythrin using a one-color fluorescence-activated DNA binding assay. The MeCP2-AIDA^C double-arginine mutant was cloned by Jankowski (2020).



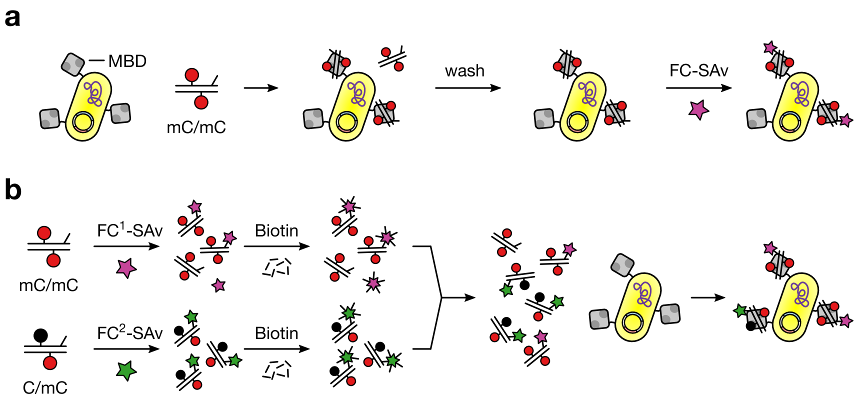
Supplementary Figure A.10 Surface-display of various MBDs using a pBAD-based expression system. Display levels and MBD–DNA binding in K-12 derivate DH10B *E. coli* and *araBAD* promoter-based vectors induced with 0.1% L-arabinose for 3 hours or overnight. The cells were probed for the presence of a surface-exposed c-Myc epitope using an allophycocyanin-coupled antibody and for binding of a DNA probe containing a fully methylated 5mC/5mC CpG that was labeled with phycoerythrin. Although DH10B is a suitable host to titrate protein expression levels with increasing L-arabinose concentration in the medium, it is not *ompT*$^-$ which meant that surface-displayed payloads could be cleaved by the outer membrane-associated omptin OmpT at dibasic residues as present in the N-terminus of the MBD.



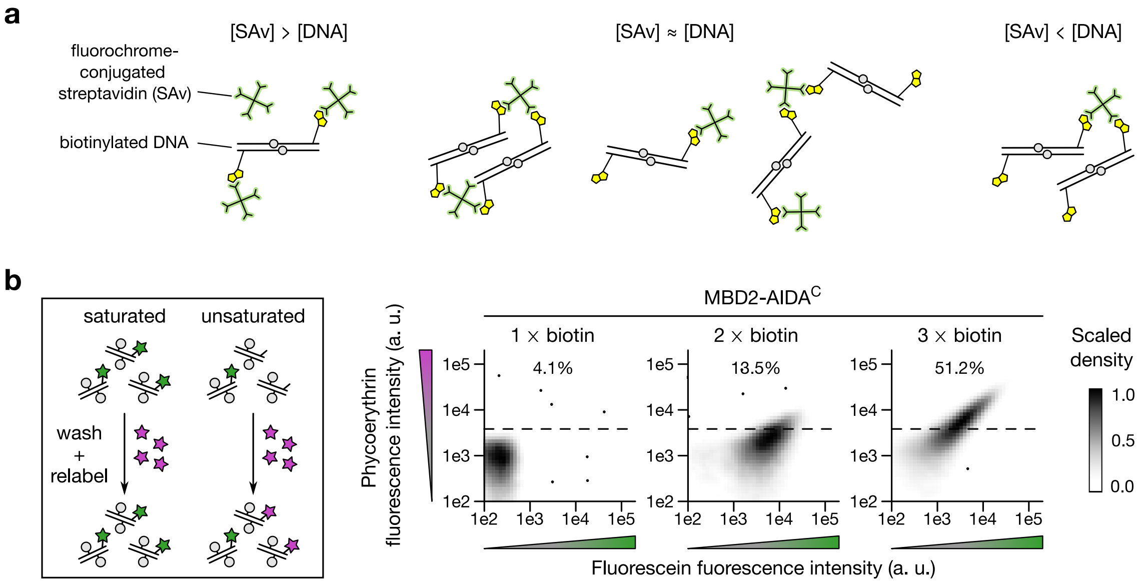
Supplementary Figure A.11 Effect of maturation time on surface display levels of active MBD. Cells were induced with 50 µM IPTG for one hour, the inducer removed and the cell incubated for 1 h or 24 h at 4 °C. Surface-display of functional MBDs was determined by staining with fluorescently-labeled DNA probes varying the amount of fluorophore as detailed in Section 7.4.



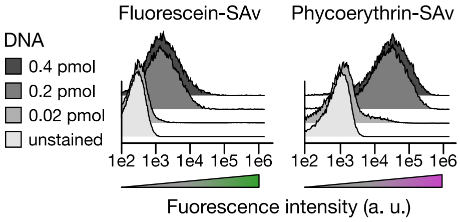
Supplementary Figure A.12 Co-detection of surface display level and DNA binding. Cells were induced with 50 µM IPTG for the indicated period of time and stained with a allophycocyanin-labeled Anti-c-Myc epitope antibody and phycoerythrin-labeled DNA probes that contained a single fully methylated 5mC/5mC CpG. MBD3 is a non-binding MBD.



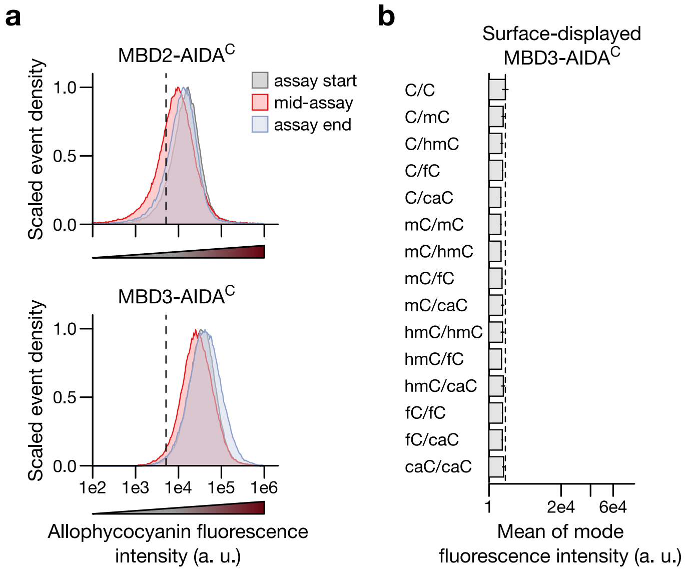
Supplementary Figure A.13 Preparation of staining mixes for fluorescence-activated DNA binding assays. (a) Sequential staining protocol (not used) in which the biotinylated DNA probe is first bound to the surface-displayed methyl-CpG-binding domain (MBD), excess probe removed by washing and then labeled with a fluorochrome (FC)-conjugated streptavidin (SAv) reagent. **(b)** Differentially labeled DNA probes are prepared separately using two (or more) fluorochrome-conjugated streptavidin reagents; excess streptavidin is quenched with biotin, then the mixture is applied to the cells and excess stain removed by washing.



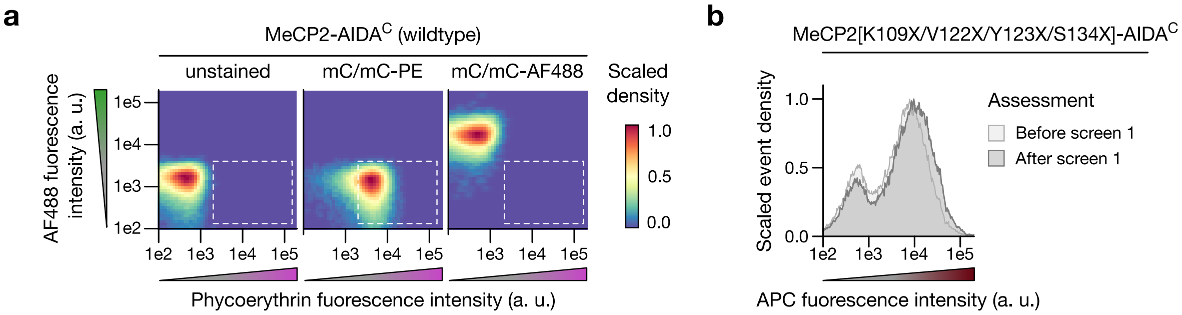
Supplementary Figure A.14 Saturation of the FACS DNA probe labeling reactions. (a) Possible labeling outcomes at different molar ratios with multivalent reagents; Entities drawn to scale; Bracketed terms denote the concentration of an entity. **(b)** Labeling and staining as in Figure 7.7 b followed by washing and relabeling with 0.4 pmol PE-streptavidin to probe saturation of the first labeling reaction. DNA probes with a single biotin were completely reluctant to relabeling and almost 85% of the biotin in the bivalent DNA probes were saturated even at a low molar excess of the fluorophore over all biotin tags. At the same concentration, the trivalent DNA probes showed the expected relabeling for 50% of the unsaturated tags, confirming that a high excess of the fluorochrome-conjugated streptavidin was required.



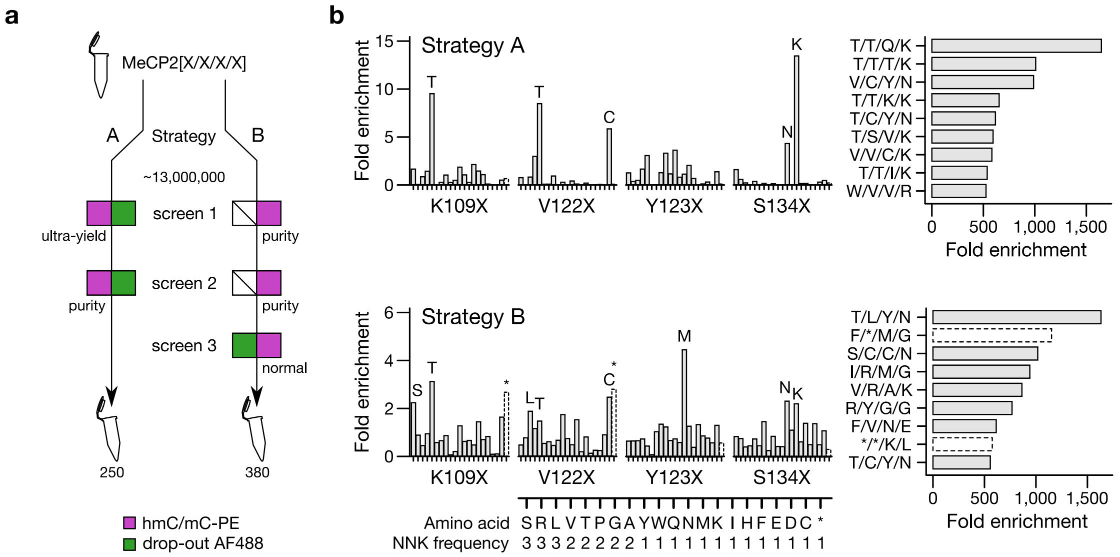
Supplementary Figure A.15 Amount of DNA probe required for surface-displayed MBD staining. Surface-displayed MBD2 (2 million BL21 Tuner™(DE3) cells, pET-AIDA^C, 50 µM IPTG) stained with the indicated amounts of labeled DNA probes containing a single 5mC/5mC CpG and two biotin tags. The DNA probes were labeled with 6 pmol fluorochrome-streptavidin (SAv) conjugate and quenched with 120 pmol biotin, washed and analyzed on a multi-color flow cytometer (10,000 events shown).



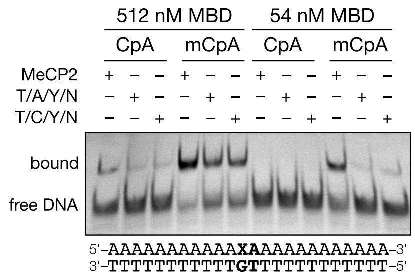
Supplementary Figure A.16 One-color FACS binding assay controls and surface-displayed wildtype MBD3. Same conditions as in Figure 7.11. **(a)** Confirmation of surface display with an allophycocyanin-labeled Anti-c-Myc epitope antibody for MBD2 and MBD3. **(b)** Full selectivity profile of surface-displayed MBD3.



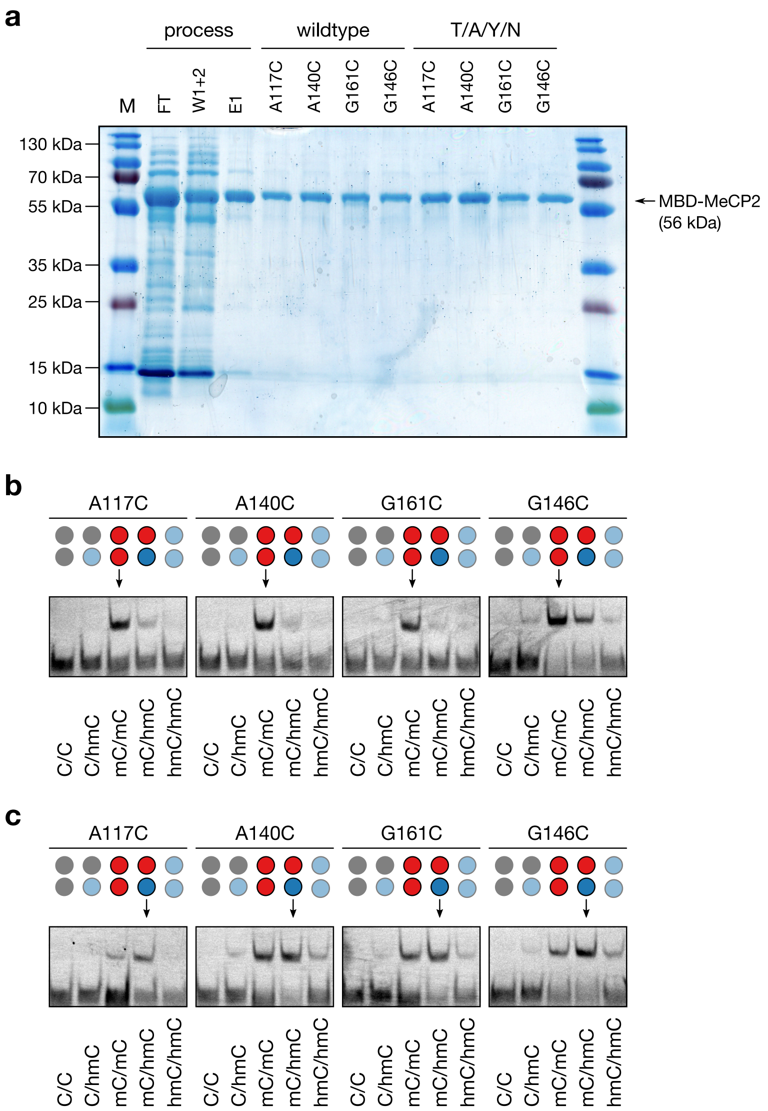
Supplementary Figure A.17 Screening the degenerated MeCP2 library against 15 modified CpG dyads. (a) Wildtype controls on the day of the first screen. **(b)** Display levels of the library. The same instrument settings were used in Figure 7.13.



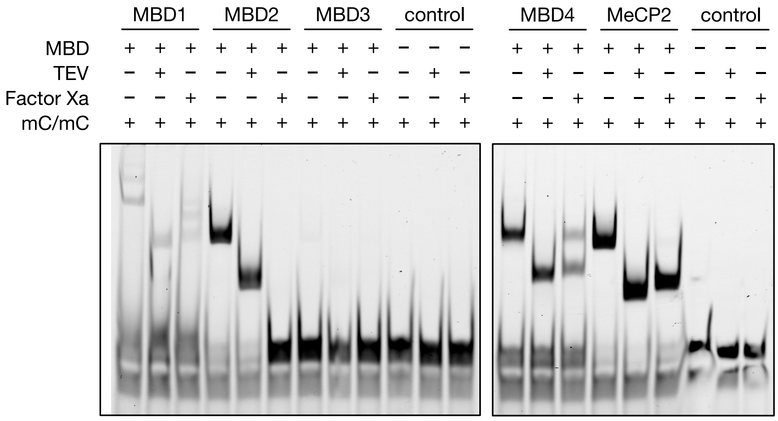
Supplementary Figure A.18 Phenotype enrichment with different screening strategies. **(a)** The same MeCP2-based NNK codon-degenerated library was screened for 5mC/5hmC-selective MBDs with the 'on-target' DNA probe labeled with phycoerythrin and the drop-out mix of 14 other probes labeled with Alexa Fluor (AF) 488. **(b)** Enriched amino-acid substitutions after the final screening round indicated in *a* per degenerated position (left) and in combination for distinct full-length phenotypes (right). The top panel is shown in Figure 7.14.


512 nM MBD
54 nM MBD
CpA
mCpA
CpA
mCpA
MeCP2 + − − + − − + − − + − −
T/A/Y/N − + − − + − − + − − + −
T/C/Y/N − − + − − + − − + − − +
bound
free DNA
5'–AAAAAAAAAAAXAAAAAAAAAAAAA–3'
3'–TTTTTTTTTTTGTTTTTTTTTTTT–5'

Supplementary Figure A.19 Binding affinity of T/A/Y/N and T/C/Y/N towards modified CpA · TpG dyads. Electrophoretic mobility shift assays with artificial 24-mer DNA probes that contained a central CpA with the indicated modified cytosine combinations at two different MBD concentrations.



Supplementary Figure A.20 Binding selectivity of A117C and other mutants for MTSL labeling. (a) Recombinant expression of four cysteine mutants of MeCP2 wildtype or MeCP2[K109T/V122A/S134N]; Flow-through (FT), wash (W1+2) and eluate 1 (E1) as reference, only eluate 2 were used; Proteins expressed as maltose-binding protein-fusion. **(b)** Binding selectivity at 256 nM for MeCP2 wildtype cysteine mutants. **(c)** Binding selectivity at 256 nM for MeCP2[K109T/V122A/S134N] cysteine mutants.


MBD1
MBD2
MBD3
control
MBD4
MeCP2
control
MBD + + + + + + + + + − − − + + + + + + − − −
TEV − + − − + − − + − − + − − + − − + − − + −
Factor Xa − − + − − + − − + − − + − − + − − + − − +
mC/mC +

Supplementary Figure A.21 Effect of TEV cleavage on SpA(Z)–MBD fusion proteins. Electrophoretic mobility shift assay (EMSA) at 4 nM labeled DNA duplex (with excess labeled strand, lowest immobile band) containing an mC/mC modified CpG with the indicated SpA(Z)-MBD fusion proteins at 1 µM MBD (except for MBD1, which was re-expressed for all other assays in this study). Reprinted from Buchmuller et al. (2020), CC BY 4.0 (full license in Appendix A.4).

A.2 Supplementary Tables

Supplementary Table A.1a Content of modified cytosines in various genomes.

Organism	Tissue/Organ	X	X / G*	*N* sites (ds)†	Method	Reference
B. taurus	Calf thymus	5mC	58 ± 5 ppt	7.0×10^{7}	Gravimetry	[Wya51]
B. taurus	Calf thymus	5mC	75 ± 4 ppt	9.1×10^{7}	UV-Vis	[Sin54]
A. thaliana	Leaves	5mC	145 ± 15 ppt	7.0×10^{6}	LC-SRM‡	[CMK+14]
D. melanogaster	Whole fly	5mC	0.30 ± 0.13 ppt	1.8×10^{4}	LC-SRM‡	[CMK+14]
E. coli (*dcm*$^{-}$)	Whole cell	5mC	0.16 ± 0.02 ppt	3.8×10^{2}	LC-SRM‡	[CMK+14]
E. coli (*dcm*$^{+}$)	Whole cell	5mC	23 ± 1 ppt	5.4×10^{4}	LC-SRM‡	[CMK+14]
M. musculus	Liver	5mC	76 ± 8 ppt	8.4×10^{7}	LC-SRM‡	[CMK+14]
M. musculus	Brain cortex	5mC	45 ± 2 ppt	4.9×10^{7}	LC-MS	[MGB+10]
M. musculus	Liver	5mC	41 ± 7 ppt	4.5×10^{7}	LC-MS	[GMM+10]
M. musculus	Brain	5mC	31 ± 2 ppt	3.4×10^{7}	*Hha*I + *Hpa*II	[SRL+79]
M. musculus	Liver	5mC	27 ± 4 ppt	3.0×10^{7}	*Hha*I + *Hpa*II	[SRL+79]
M. musculus	Brain	5mC	31 ppt	3.4×10^{7}	LC-MS	[ISD+11]
M. musculus	Liver	5mC	28 ppt	3.1×10^{7}	LC-MS	[ISD+11]
M. musculus	ESC	5mC	14 ppt	1.5×10^{7}	LC-MS	[BUY+14]
M. musculus	ESC E14	5mC	29 ppt	3.2×10^{7}	LC-MS	[ISD+11]
M. musculus	ESC E14	5mC	n/d	3.2×10^{6}	MAB-Seq	[NIK+15]
M. musculus	ESC BL/6	5mC	n/d	2.5×10^{5}	redBS/oxBS-Seq	[BMB+14]
R. norvegicus	Brain	5mC	44 ± 8 ppt	5.5×10^{7}	UV-Vis	[VMV+73]
R. norvegicus	Liver	5mC	43 ± 2 ppt	5.4×10^{7}	UV-Vis	[VMV+73]
H. sapiens	Brain	5mC	45 ± 1 ppt	5.4×10^{7}	RP-HPLC	[EGH+82]
H. sapiens	Liver	5mC	40 ± 1 ppt	4.8×10^{7}	RP-HPLC	[EGH+82]
H. sapiens	Brain	5mC	53 ± 1 ppt	6.4×10^{7}	LC-MS	[LWS+13]

Supplementary Table A.1b Content of modified cytosines in various genomes.

Organism	Tissue/Organ	X	X / G*	*N* sites (ds)†	Method	Reference
H. sapiens	HEK293T**	5mC	36 ± 3 ppt	4.3×10^{7}	LC-MS	[LWS+13]
H. sapiens	HeLa**	5mC	28 ± 2 ppt	3.4×10^{7}	LC-MS	[LWS+13]
H. sapiens	HCT116**	5mC	44 ppt	5.3×10^{7}	LC-MS	[BUY+14]
H. sapiens	MCF-7**	5mC	39 ppt	4.7×10^{7}	LC-MS	[BUY+14]
Bacteriophage	T4, (T2, T6)	5hmC	fully modified	4.6×10^{4}	UV-Vis	[WC52]
M. musculus	Granule cells	5hmC	2.3 ± 0.1 ppt	2.5×10^{6}	NNA††	[KH09]
M. musculus	Purkinje cells	5hmC	5.9 ± 0.5 ppt	6.5×10^{6}	NNA††	[KH09]
M. musculus	Brain cortex	5hmC	6.5 ± 0.4 ppt	7.2×10^{6}	LC-MS	[MGB+10]
M. musculus	Liver	5hmC	0.6 ± 0.1 ppt	6.6×10^{5}	LC-MS	[GMM+10]
M. musculus	Brain	5hmC	6.8 ppt	7.5×10^{6}	LC-MS	[ISD+11]
M. musculus	Liver	5hmC	0.8 ppt	8.8×10^{5}	LC-MS	[ISD+11]
M. musculus	ESC	5hmC	0.68 ± 0.05 ppt	7.5×10^{5}	LC-MS	[LWS+13]
M. musculus	ESC	5hmC	0.4 ppt	4.4×10^{5}	LC-MS	[BUY+14]
M. musculus	ESC J1	5hmC	1.6 ppt	1.8×10^{6}	LC-MS	[BBF+12]
M. musculus	ESC E14	5hmC	1.2 ppt	1.4×10^{6}	LC-MS	[ISD+11]
M. musculus	ESC E14Tg2a	5hmC	n/d	2.0×10^{6}	TAB-Seq	[YHS+12]
M. musculus	ESC BL/6	5hmC	n/d	7.3×10^{3}	redBS/oxBS-Seq	[BMB+14]
H. sapiens	Brain	5hmC	7.0 ± 0.9 ppt	8.5×10^{6}	LC-MS	[LWS+13]
H. sapiens	HEK293T**	5hmC	0.14 ± 0.01 ppt	1.7×10^{5}	LC-MS	[LWS+13]
H. sapiens	HeLa**	5hmC	0.15 ± 0.02 ppt	1.8×10^{5}	LC-MS	[LWS+13]
H. sapiens	HCT116**	5hmC	0.05 ppt	4.8×10^{4}	LC-MS	[BUY+14]
H. sapiens	MCF-7**	5hmC	0.16 ppt	1.9×10^{5}	LC-MS	[BUY+14]
H. sapiens	ESC H1	5hmC	n/d	6.9×10^{4}	TAB-Seq	[YHS+12]
M. musculus	Brain	5fC	15 ppm	1.7×10^{4}	LC-MS	[ISD+11]
M. musculus	Liver	5fC	6 ppm	6.5×10^{3}	LC-MS	[ISD+11]
M. musculus	ESC E14	5fC	18 ± 1 ppm	2.0×10^{4}	LC-MS	[ISD+11]
M. musculus	ESC	5fC	3 ± 1 ppm	3.2×10^{3}	LC-MS	[PHT+11]
M. musculus	ESC	5fC	14.5 ± 2.9 ppm	1.6×10^{4}	LC-MS	[LWS+13]
M. musculus	ESC E14	5fC	n/d	2.9×10^{4}	MAB-Seq	[NIK+15]
M. musculus	ESC BL/6	5fC	n/d	1.3×10^{3}	redBS/oxBS-Seq	[BMB+14]
M. musculus	ESC (inhib.)	5fC	32 ppm	3.6×10^{4}	CLEVER-Seq‡‡	[ZGG+17]
M. musculus	ESC (serum)	5fC	n/d	1.2×10^{4}	CLEVER-Seq‡‡	[ZGG+17]
H. sapiens	Brain	5fC	7.7 ± 1.5 ppm	9.3×10^{3}	LC-MS	[LWS+13]
H. sapiens	HEK293T**	5fC	1.1 ± 0.1 ppm	1.2×10^{3}	LC-MS	[LWS+13]
H. sapiens	HeLa**	5fC	3.9 ± 0.3 ppm	3.8×10^{3}	LC-MS	[LWS+13]
H. sapiens	ESC	5fC	n/d	4.1×10^{4}	CLEVER-Seq‡‡	[ZGG+17]
M. musculus	ESC E14	5caC	3.6 ± 0.3 ppm	3.9×10^{3}	LC-MS	[ISD+11]
M. musculus	ESC	5caC	3.4 ± 0.3 ppm	3.7×10^{3}	LC-MS	[LWS+13]

Supplementary Table A.1c Content of modified cytosines in various genomes.

Organism	Tissue/Organ	X	X / G[*]	*N* sites (ds)[†]	Method	Reference
H. sapiens	Brain	5caC	0.68 ± 0.05 ppm	820	LC-MS	[LWS+13]
H. sapiens	HEK293T[**]	5caC	0.79 ± 0.14 ppm	950	LC-MS	[LWS+13]
H. sapiens	HeLa[**]	5caC	0.81 ± 0.17 ppm	970	LC-MS	[LWS+13]

[*] Relative frequency *f* = X / G or *f* =X / (X + C), the more conservative estimate is used if possible; For sequencing-based approaches within the covered genomic regions; n/d = not determined. A study on 5hmC in rodent brain by Penn et al. (1972) was excluded since 5mC was not detectable and the result not reproducible by others (Kothari & Shankar, 1976).

[†] Calculated as *N*(bp) × %(G + C) × *f* or *N*(C, ds) × *f* based on the reference genome assemblies; For sequence-based approaches *N*(sites) or *N*(sites) × median modification level if provided by the authors.

[‡] Liquid chromatography selective reaction monitoring.

[**] Cancer cell line.

[††] Nearest-neighbor analysis (Ramsahoye, 2002).

[‡‡] Total number of uniquely covered sites from eight to twelve single cells.

Supplementary Table A.2 Sequences of DNA probes used in the literature.

Source	CpG (mod.)	CpA	Sequence	References
‘GAM1’	12 (1)	0	GATCCGACGACGACGA**CG**ACGACGACGACGACGATC	[NMB93] and others
‘GAM3’	12 (3)	0	GATCCGACGACGA**CG**A**CG**A**CG**ACGACGACGACGACGATC	[NMB93]
‘GAM4’	12 (4)	0	GATCCGACGACGA**CG**A**CG**A**CG**ACGACGACGACGACGATC	[NMB93]
‘GAM5’	12 (5)	0	GATCCGACGA**CG**A**CG**A**CG**A**CG**ACGACGACGACGACGATC	[NMB93]
‘GAM6’	12 (6)	0	GATCCGACGA**CG**A**CG**A**CG**A**CG**A**CG**ACGACGACGACGATC	[NMB93]
‘GAM12’	12 (12)	0	GATC**CG**A**CG**A**CG**A**CG**A**CG**A**CG**A**CG**A**CG**A**CG**A**CG**A**CG**A**CG**ATC	[NMB93] and others
artificial	4 (1)	2	AGCTTATCGCAGC**CG**GCGCGAATCTGA	[VTR+04]
artificial	3 (1)	1	GAGGCGCT**CG**GCGGCAG	[CSW+14; WCB+14]
artificial	1 (1)	2	GCCAA**CG**TTGGC	[LLW+19]
artificial	1 (1)	0	AAAAAAAAAA**CG**AAAAAAAAAA	[BKS20] (this work)
BDNF1	1 (1)	2	GCCCTGGAA**CG**GAACTCTTCTGGCC	[YKL+16]
BRCA1	6 (1)	2	AAAACTGCGACTGCG**CG**GCGTGAGCTCGCTGAGACTTCCTGGACGGGGA	[FBM+03]
BRCA1	6 (2)	2	AAAACTGCGACTGCG**CG**G**CG**TGAGCTCGCTGAGACTTCCTGGACGGGGA	[FBM+03]
BRCA1	6 (2)	2	AAAACTGCGACTGCG**CG**GCGTGAGCTCGCTGAGACTTCCTGGA**CG**GGGA	[FBM+03]
GSTP1	6 (6)	3	CCCTCCAGAAGAG**CG**GC**CG**G**CG**C**CG**TGACTCAGCACTGGGG**CG**GAG**CG**GG	[FBM+03]
MLH1	3 (3)	4	GAA**CG**TGAGCA**CG**AGGCACTGAGGTGATTGGCTGAAGGCACTTC**CG**TTGA	[FBM+03]
CDKN2A	7 (7)	2	G**CG**CT**CG**G**CG**GCTG**CG**GAGAGGGGGAGAGCAGGCAG**CG**GG**CG**G**CG**GGGAG	[FBM+03]
ρ-globin	1 (1)	0	GGAT**CG**GCTC	[SWG+11]

Supplementary Table A.3 Reported binding affinities of MBD1. Only probes with a single modified CpG dyad reported.

Combination	Context	Protein sequence	Affinity	Ion strength	Method	Reference
C/C	CCATGCGCTGAC	1–105 (*H. sapiens*)	1,400 ± 300 nM	150 mM NaCl	FP	[HLU+12]
C/C	GGGCTCGAAGTG	1–75 (*H. sapiens*)	33,000 ± 6,900 nM	100 mM NaCl	ITC	[OAK+13]
5mC/C	CCATGCGCTGAC	1–105 (*H. sapiens*)	250 ± 90 nM	150 mM NaCl	FP	[HLU+12]
C/5mC	GGGCTCGAAGTG	1–75 (*H. sapiens*)	3,080 ± 830 nM	100 mM NaCl	ITC	[OAK+13]
5hmC/C	CCATGCGCTGAC	1–105 (*H. sapiens*)	640 ± 100 nM	150 mM NaCl	FP	[HLU+12]
5mC/5mC	CCATGCGCTGAC	1–105 (*H. sapiens*)	5 ± 1 nM	150 mM NaCl	FP	[HLU+12]
5mC/5mC	GGGCTCGAAGTG	1–75 (*H. sapiens*)	72 ± 11 nM	100 mM NaCl	ITC	[OAK+13]
5hmC/5mC	CCATGCGCTGAC	1–105 (*H. sapiens*)	90 ± 10 nM	150 mM NaCl	FP	[HLU+12]
5hmC/5mC	GGGCTCGAAGTG	1–75 (*H. sapiens*)	1,040 ± 422 nM	100 mM NaCl	ITC	[OAK+13]
5hmC/5hmC	CCATGCGCTGAC	1–105 (*H. sapiens*)	1,000 ± 100 nM	150 mM NaCl	FP	[HLU+12]

Supplementary Table A.4 Reported binding affinities of MBD2. Only probes with a single modified CpG dyad reported.

Combination	Context	Protein sequence	Affinity	Ion strength	Method	Reference
C/C	CCATGCGCTGAC	153–414/p66α (*M. m.*)	6,500 ± 3,000 nM	150 mM NaCl	FP	[HLU+12]
C/C	GACGACGACGAC	full length (*M. musculus*)	189 ± 47 nM	50 mM NaCl	RCSA	[FBM+03]
C/C	CTGCGCGGCGTG	full length (*M. musculus*)	200 ± 54 nM	50 mM NaCl	RCSA	[FBM+03]
5mC/C	CCATGCGCTGAC	153–414/p66α (*M. m.*)	700 ± 100 nM	150 mM NaCl	FP	[HLU+12]
5hmC/C	CCATGCGCTGAC	153–414/p66α (*M. m.*)	4,900 ± 1,300 nM	150 mM NaCl	FP	[HLU+12]
5mC/5mC	CCATGCGCTGAC	153–414/p66α (*M. m.*)	60 ± 20 nM	150 mM NaCl	FP	[HLU+12]
5mC/5mC	GACGACGACGAC	full length (*M. musculus*)	3 ± 1 nM	50 mM NaCl	RCSA	[FBM+03]
5mC/5mC	CTGCGCGGCGTG	full length (*M. musculus*)	4 ± 2 nM	50 mM NaCl	RCSA	[FBM+03]
5mC/5mC	-GGATCGGCTC-	2–72 (*G. gallus*)	2 ± 0 nM	50 mM NaCl	SPR	[SWG+11]
5mC/5mC	GCGCTCGGCGGC	1–70 (*G. gallus*)	110 ± 1 nM	50 mM NaCl	SPR	[CSW+14]
5hmC/5mC	CCATGCGCTGAC	153–414/p66α (*M. m.*)	600 ± 100 nM	150 mM NaCl	FP	[HLU+12]
5hmC/5hmC	CCATGCGCTGAC	153–414/p66α (*M. m.*)	2,800 ± 700 nM	150 mM NaCl	FP	[HLU+12]
5hmC/5hmC	GCGCTCGGCGGC	1–70 (*G. gallus*)	54,000 ± 8,000 nM	50 mM NaCl	SPR	[CSW+14]

Supplementary Table A.5a Reported binding affinities of MBD3. Only probes with a single modified CpG dyad reported.

Combination	Context	Protein sequence	Affinity	Ion strength	Method	Reference
C/C	CCATGCGCTGAC	1–265ΔE/p66β (*M. m.*)	6,600 ± 1,600 nM	150 mM NaCl	FP	[HLU+12]
C/C	GACGACGACGAC	full length (*X. laevis*)	779 ± 257 nM	50 mM NaCl	RCSA	[FBM+03]
C/C	CTGCGCGGCGTG	full length (*X. laevis*)	555 ± 128 nM	50 mM NaCl	RCSA	[FBM+03]
C/C	GACGACGACGAC	full length (*M. musculus*)	684 ± 124 nM	50 mM NaCl	RCSA	[FBM+03]
C/C	GACGACGACGAC	full length, F34Y (*M. m.*)	751 ± 146 nM	50 mM NaCl	RCSA	[FBM+03]
C/C	GACGACGACGAC	full length, H30K (*M. m.*)	682 ± 126 nM	50 mM NaCl	RCSA	[FBM+03]
5mC/C	CCATGCGCTGAC	1–265ΔE/p66β (*M. m.*)	2,900 ± 900 nM	150 mM NaCl	FP	[HLU+12]
5hmC/C	CCATGCGCTGAC	1–265ΔE/p66β (*M. m.*)	3,000 ± 500 nM	150 mM NaCl	FP	[HLU+12]
5mC/5mC	CCATGCGCTGAC	1–265ΔE/p66β (*M. m.*)	1,300 ± 100 nM	150 mM NaCl	FP	[HLU+12]
5mC/5mC	GACGACGACGAC	full length (*X. laevis*)	186 ± 42 nM	50 mM NaCl	RCSA	[FBM+03]

Supplementary Table A.5b Reported binding affinities of MBD3. Only probes with a single modified CpG dyad reported.

Combination	Context	Protein sequence	Affinity	Ion strength	Method	Reference
5mC/5mC	CTGCGCGGCGTG	full length (*X. laevis*)	63 ± 16 nM	50 mM NaCl	RCSA	[FBM+03]
5mC/5mC	GACGACGACGAC	full length (*M. musculus*)	580 ± 105 nM	50 mM NaCl	RCSA	[FBM+03]
5mC/5mC	GACGACGACGAC	full length, F34Y (*M. m.*)	81 ± 12 nM	50 mM NaCl	RCSA	[FBM+03]
5mC/5mC	GACGACGACGAC	full length, H30K (*M. m.*)	132 ± 20 nM	50 mM NaCl	RCSA	[FBM+03]
5mC/5mC	GCGCTCGGCGGC	1–70 (*H. sapiens*)	54,000 ± 7,000 nM	50 mM NaCl	SPR	[CSW+14]
5mC/5mC	GCGCTCGGCGGC	1–70, F34Y,H30K (*H. s.*)	130 ± 10 nM	50 mM NaCl	SPR	[CSW+14]
5mC/5mC	GCCAACGTTGGC	1–71 (*H. sapiens*)	5,400 ± 1,200 nM	150 mM NaCl	ITC	[LLW+19]
5hmC/5mC	CCATGCGCTGAC	1–265ΔE/p66β (*M. m.*)	2,900 ± 800 nM	150 mM NaCl	FP	[HLU+12]
5hmC/5hmC	CCATGCGCTGAC	1–265ΔE/p66β (*M. m.*)	4,700 ± 1,000 nM	150 mM NaCl	FP	[HLU+12]
5hmC/5hmC	GCGCTCGGCGGC	1–70 (*H. sapiens*)	not detected	50 mM NaCl	SPR	[CSW+14]
5mC/5mC	GCCAACGTTGGC	1–71 (*H. sapiens*)	not detected	150 mM NaCl	ITC	[LLW+19]

Supplementary Table A.6 Reported binding affinities of MBD4. Only probes with a single modified CpG dyad reported.

Combination	Context	Protein sequence	Affinity	Ion strength	Method	Reference
C/C	CCATGCGCTGAC	49–187 (*M. musculus*)	1,070 ± 110 nM	150 mM NaCl	FP	[HLU+12]
C/C	-GGATCGGCTC-	69–136 (*H. sapiens*)	not detected	100 mM NaCl	ITC	[OAK+13]
C/C	GCGCTCGGCGGC	80–148 (*H. sapiens*)	17,200 ± 2,000 nM	50 mM NaCl	SPR	[WCB+14]
5mC/C	CCATGCGCTGAC	49–187 (*M. musculus*)	520 ± 60 nM	150 mM NaCl	FP	[HLU+12]
C/5mC	-GGATCGGCTC-	69–136 (*H. sapiens*)	not detected	100 mM NaCl	ITC	[OAK+13]
5hmC/C	CCATGCGCTGAC	49–187 (*M. musculus*)	840 ± 70 nM	150 mM NaCl	FP	[HLU+12]
5mC/5mC	CCATGCGCTGAC	49–187 (*M. musculus*)	220 ± 10 nM	150 mM NaCl	FP	[HLU+12]
5mC/5mC	-GGATCGGCTC-	69–136 (*H. sapiens*)	98 ± 76 nM	100 mM NaCl	ITC	[OAK+13]
5mC/5mC	GCGCTCGGCGGC	80–148 (*H. sapiens*)	6,400 ± 1,500 nM	50 mM NaCl	SPR	[WCB+14]
5hmC/5mC	CCATGCGCTGAC	49–187 (*M. musculus*)	560 ± 100 nM	150 mM NaCl	FP	[HLU+12]
5hmC/5mC	-GGATCGGCTC-	69–136 (*H. sapiens*)	162 ± 58 nM	100 mM NaCl	ITC	[OAK+13]
5hmC/5hmC	CCATGCGCTGAC	49–187 (*M. musculus*)	950 ± 40 nM	150 mM NaCl	FP	[HLU+12]
5hmC/5hmC	GCGCTCGGCGGC	80–148 (*H. sapiens*)	14,200 ± 1,900 nM	50 mM NaCl	SPR	[WCB+14]
T/5mC	-GGATTGGCTC-	69–136 (*H. sapiens*)	99 ± 42 nM	100 mM NaCl	ITC	[OAK+13]
C/T	-GGATCAGCTC-	69–136 (*H. sapiens*)	213 ± 58 nM	100 mM NaCl	ITC	[OAK+13]

Supplementary Table A.7a Reported binding affinities of MeCP2. Only probes with a single modified CpG dyad reported.

Combination	Context	Protein sequence	Affinity	Ion strength	Method	Reference
C/C	CCATGCGCTGAC	77–205 (*H. sapiens*)	500 ± 100 nM	150 mM NaCl	FP	[HLU+12]
C/C	GACGACGACGAC	full length (*M. musculus*)	458 ± 88 nM	50 mM NaCl	CSA	[FBM+03]
C/C	CTGCGCGGCGTG	full length (*M. musculus*)	441 ± 138 nM	50 mM NaCl	CSA	[FBM+03]
C/C	GACGACGACGAC	full length (*X. laevis*)	556 ± 77 nM	50 mM NaCl	CSA	[FBM+03]
C/C	CTGCGCGGCGTG	full length (*X. laevis*)	395 ± 100 nM	50 mM NaCl	CSA	[FBM+03]
C/C	GCAGCCGGCGCG	77–165 (*M. musculus*)	1,030 ± 20 nM	30 mM KCl	EMSA	[VTR+04]
C/C	TGGAACGGAACT	77–167 (*H. sapiens*)	398 ± 49 nM	30 mM KCl	EMSA	[YKL+16]

Supplementary Table A.7b Reported binding affinities of MeCP2. Only probes with a single modified CpG dyad reported.

Combination	Context	Protein sequence	Affinity	Ion strength	Method	Reference
C/C	TGGAACGGAACT	77–167, R133C (*H. s.*)	1,203 ± 392 nM	30 mM KCl	EMSA	[YKL+16]
C/C	TGGAACGGAACT	77–167, S134C (*H. s.*)	759 ± 184 nM	30 mM KCl	EMSA	[YKL+16]
C/C	TGGAACGGAACT	77–167, T158M (*H. s.*)	351 ± 54 nM	30 mM KCl	EMSA	[YKL+16]
C/C	TGGAACGGAACT	76–167 (*H. sapiens*)	11 ± 4 nM	25 mM KCl	FP	[KWC+14]
C/C	TGGAACGGAACT	76–167 (*H. sapiens*)	1,000 nM	100 mM KCl	FP	[KWC+14]
C/C	TGGAACGGAACT	76–167 (*H. sapiens*)	10,300 nM	150 mM KCl	FP	[KWC+14]
5mC/C	CCATGCGCTGAC	77–205 (*H. sapiens*)	130 ± 20 nM	150 mM NaCl	FP	[HLU+12]
5mC/C	GCAGCCGGCGCG	77–165 (*M. musculus*)	127 ± 3 nM	30 mM KCl	EMSA	[VTR+04]
C/5mC	GCAGCCGGCGCG	77–165 (*M. musculus*)	152 ± 4 nM	30 mM KCl	EMSA	[VTR+04]
C/5mC	TGGAACGGAACT	77–167 (*H. sapiens*)	40 ± 1 nM	30 mM KCl	EMSA	[YKL+16]
5mC/C	TGGAACGGAACT	77–167 (*H. sapiens*)	37 ± 1 nM	30 mM KCl	EMSA	[YKL+16]
5hmC/C	CCATGCGCTGAC	77–205 (*H. sapiens*)	190 ± 20 nM	150 mM NaCl	FP	[HLU+12]
5hmC/C	GCAGCCGGCGCG	77–165 (*M. musculus*)	950 ± 10 nM	30 mM KCl	EMSA	[VTR+04]
C/5hmC	GCAGCCGGCGCG	77–165 (*M. musculus*)	1,100 ± 10 nM	30 mM KCl	EMSA	[VTR+04]
C/5hmC	TGGAACGGAACT	77–167 (*H. sapiens*)	185 ± 25 nM	30 mM KCl	EMSA	[YKL+16]
5hmC/C	TGGAACGGAACT	77–167 (*H. sapiens*)	185 ± 20 nM	30 mM KCl	EMSA	[YKL+16]
C/5hmC	TGGAACGGAACT	76–167 (*H. sapiens*)	700 nM	100 mM KCl	FP	[KWC+14]
C/5fC	TGGAACGGAACT	77–167 (*H. sapiens*)	110 ± 5 nM	30 mM KCl	EMSA	[YKL+16]
5fC/C	TGGAACGGAACT	77–167 (*H. sapiens*)	120 ± 5 nM	30 mM KCl	EMSA	[YKL+16]
C/5caC	TGGAACGGAACT	77–167 (*H. sapiens*)	180 ± 50 nM	30 mM KCl	EMSA	[YKL+16]
5caC/C	TGGAACGGAACT	77–167 (*H. sapiens*)	175 ± 35 nM	30 mM KCl	EMSA	[YKL+16]
5mC/5mC	CCATGCGCTGAC	77–205 (*H. sapiens*)	10 ± 1 nM	150 mM NaCl	FP	[HLU+12]
5mC/5mC	GACGACGACGAC	full length (*X. laevis*)	48 ± 9 nM	50 mM NaCl	CSA	[FBM+03]
5mC/5mC	CTGCGCGGCGTG	full length (*X. laevis*)	22 ± 8 nM	50 mM NaCl	CSA	[FBM+03]
5mC/5mC	GACGACGACGAC	full length (*M. musculus*)	172 ± 23 nM	50 mM NaCl	CSA	[FBM+03]
5mC/5mC	CTGCGCGGCGTG	full length (*M. musculus*)	161 ± 40 nM	50 mM NaCl	CSA	[FBM+03]
5mC/5mC	GCAGCCGGCGCG	77–165 (*M. musculus*)	15 ± 1 nM	30 mM KCl	EMSA	[VTR+04]
5mC/5mC	TGGAACGGAACT	77–167 (*H. sapiens*)	6 ± 1 nM	30 mM KCl	EMSA	[YKL+16]
5mC/5mC	TGGAACGGAACT	77–167, R133C (*H. s.*)	100 ± 12 nM	30 mM KCl	EMSA	[YKL+16]
5mC/5mC	TGGAACGGAACT	77–167, S134C (*H. s.*)	28 ± 1 nM	30 mM KCl	EMSA	[YKL+16]
5mC/5mC	TGGAACGGAACT	77–167, T158M (*H. s.*)	14 ± 2 nM	30 mM KCl	EMSA	[YKL+16]
5mC/5mC	TGGAACGGAACT	76–167 (*H. sapiens*)	7 ± 3 nM	25 mM KCl	FP	[KWC+14]
5mC/5mC	TGGAACGGAACT	76–167 (*H. sapiens*)	5 nM	100 mM KCl	FP	[KWC+14]
5mC/5mC	TGGAACGGAACT	76–167 (*H. sapiens*)	75 ± 1 nM	150 mM KCl	FP	[KWC+14]
5mC/5mC	GCAGCCGGCGCG	77–165 (*M. musculus*)	15 nM	30 mM KCl	EMSA	[FWS+01]
5mC/5mC	GACGACGACGAC	1–467 (*X. laevis*)	40 nM	50 mM NaCl	EMSA	[BYW00]
5hmC/5mC	CCATGCGCTGAC	77–205 (*H. sapiens*)	46 ± 9 nM	150 mM NaCl	FP	[HLU+12]
5hmC/5mC	GCAGCCGGCGCG	77–165 (*M. musculus*)	151 ± 7 nM	30 mM KCl	EMSA	[VTR+04]

Supplementary Table A.7c Reported binding affinities of MeCP2. Only probes with a single modified CpG dyad reported.

Combination	Context	Protein sequence	Affinity	Ion strength	Method	Reference
5mC/5hmC	GCAGCCGGCGCG	77–165 (*M. musculus*)	157 ± 8 nM	30 mM KCl	EMSA	[VTR+04]
5mC/5hmC	TGGAACGGAACT	77–167 (*H. sapiens*)	26 ± 1 nM	30 mM KCl	EMSA	[YKL+16]
5hmC/5mC	TGGAACGGAACT	77–167 (*H. sapiens*)	28 ± 1 nM	30 mM KCl	EMSA	[YKL+16]
5hmC/5mC	TGGAACGGAACT	76–167 (*H. sapiens*)	290 nM	100 mM KCl	FP	[KWC+14]
5mC/5fC	TGGAACGGAACT	77–167 (*H. sapiens*)	40 ± 1 nM	30 mM KCl	EMSA	[YKL+16]
5fC/5mC	TGGAACGGAACT	77–167 (*H. sapiens*)	45 ± 1 nM	30 mM KCl	EMSA	[YKL+16]
5mC/5caC	TGGAACGGAACT	77–167 (*H. sapiens*)	50 ± 5 nM	30 mM KCl	EMSA	[YKL+16]
5caC/5mC	TGGAACGGAACT	77–167 (*H. sapiens*)	55 ± 5 nM	30 mM KCl	EMSA	[YKL+16]
5hmC/5hmC	CCATGCGCTGAC	77–205 (*H. sapiens*)	260 ± 20 nM	150 mM NaCl	FP	[HLU+12]
5hmC/5hmC	GCAGCCGGCGCG	77–165 (*M. musculus*)	777 ± 15 nM	30 mM KCl	EMSA	[VTR+04]
5hmC/5hmC	TGGAACGGAACT	77–167 (*H. sapiens*)	250 ± 44 nM	30 mM KCl	EMSA	[YKL+16]
5hmC/5hmC	TGGAACGGAACT	76–167 (*H. sapiens*)	700 nM	100 mM KCl	FP	[KWC+14]
5hmC/5hmC	TGGAACGGAACT	76–167 (*H. sapiens*)	2,100 ± 1,500 nM	150 mM KCl	FP	[KWC+14]
5hmC/5fC	TGGAACGGAACT	77–167 (*H. sapiens*)	170 ± 5 nM	30 mM KCl	EMSA	[YKL+16]
5fC/5hmC	TGGAACGGAACT	77–167 (*H. sapiens*)	175 ± 5 nM	30 mM KCl	EMSA	[YKL+16]
5hmC/5caC	TGGAACGGAACT	77–167 (*H. sapiens*)	180 ± 25 nM	30 mM KCl	EMSA	[YKL+16]
5caC/5hmC	TGGAACGGAACT	77–167 (*H. sapiens*)	180 ± 25 nM	30 mM KCl	EMSA	[YKL+16]
5fC/5fC	TGGAACGGAACT	77–167 (*H. sapiens*)	110 ± 30 nM	30 mM KCl	EMSA	[YKL+16]
5fC/5caC	TGGAACGGAACT	77–167 (*H. sapiens*)	115 ± 20 nM	30 mM KCl	EMSA	[YKL+16]
5caC/5fC	TGGAACGGAACT	77–167 (*H. sapiens*)	117 ± 15 nM	30 mM KCl	EMSA	[YKL+16]
5caC/5caC	TGGAACGGAACT	77–167 (*H. sapiens*)	165 ± 30 nM	30 mM KCl	EMSA	[YKL+16]
5mC/T	GCAGCCAGCGCG	77–165 (*M. musculus*)	18 ± 2 nM	30 mM KCl	EMSA	[VTR+04]

Supplementary Table A.8a Model estimates for one-site MeCP2 wildtype and Rett mutant binding. Fitted estimates for macroscopic descriptors of the binding isotherms in Figure 6.4a.

Degree	Protein	Probe	Combination	Term	Estimate	Std. Error	*p*-value
1	wildtype	artificial	C/5hmC	K_{d1}	1,078 nM	100 nM	2.00×10^{-13}
1	wildtype	artificial	5mC/5mC	K_{d1}	26 nM	4 nM	2.23×10^{-08}
2	wildtype	artificial	5mC/5mC	K_{d1}	29 nM	2 nM	5.20×10^{-16}
1	wildtype	artificial	5mC/5hmC	K_{d1}	224 nM	24 nM	2.10×10^{-11}
1	wildtype	artificial	5mC/5fC	K_{d1}	197 nM	24 nM	3.57×10^{-10}
1	wildtype	artificial	5hmC/5hmC	K_{d1}	967 nM	89 nM	1.51×10^{-13}
1	wildtype	artificial	5hmC/5fC	K_{d1}	541 nM	62 nM	9.21×10^{-11}
1	R133C	artificial	C/5hmC	K_{d1}	8,113 nM	526 nM	1.18×10^{-18}
1	R133C	artificial	5mC/5mC	K_{d1}	117 nM	11 nM	1.43×10^{-12}
1	R133C	artificial	5mC/5hmC	K_{d1}	660 nM	72 nM	2.47×10^{-11}
1	R133C	artificial	5mC/5fC	K_{d1}	543 nM	47 nM	3.40×10^{-14}
1	R133C	artificial	5hmC/5hmC	K_{d1}	2,555 nM	261 nM	2.76×10^{-12}

Supplementary Table A.8b Model estimates for one-site MeCP2 wildtype and Rett mutant binding. Fitted estimates for macroscopic descriptors of the binding isotherms in Figure 6.4a.

Degree	Protein	Probe	Combination	Term	Estimate	Std. Error	*p*-value
1	R133C	artificial	5hmC/5fC	K_{d1}	1,627 nM	93 nM	1.48×10^{-20}
1	S134C	artificial	C/5hmC	K_{d1}	7,569 nM	660 nM	2.33×10^{-14}
1	S134C	artificial	5mC/5mC	K_{d1}	116 nM	18 nM	1.70×10^{-07}
1	S134C	artificial	5mC/5hmC	K_{d1}	1,149 nM	71 nM	2.81×10^{-19}
1	S134C	artificial	5mC/5fC	K_{d1}	790 nM	74 nM	3.05×10^{-13}
1	S134C	artificial	5hmC/5hmC	K_{d1}	7,627 nM	568 nM	1.37×10^{-16}
1	S134C	artificial	5hmC/5fC	K_{d1}	4,644 nM	354 nM	3.11×10^{-16}

Supplementary Table A.9a Model estimates for multi-site MBD–DNA binding. Fitted estimates for macroscopic descriptors of the binding isotherms with shared estimates as shown in Figure 8.4 and Figure 8.5.

Degree	Protein	Probe	Combination	Term	Estimate	Std. Error	*p*-value
1	wildtype	*CDKN2A*	5hmC/5mC	–p(K_1 / nM)	1.642	0.056	1.48×10^{-18}
1	wildtype	*CDKN2A*	5mC/5mC	–p(K_1 / nM)	2.012	0.073	5.13×10^{-18}
1	wildtype	*BRCA1 (b)*	5hmC/5mC	–p(K_1 / nM)	0.783	0.169	1.76×10^{-04}
1	wildtype	*BRCA1 (b)*	5mC/5mC	–p(K_1 / nM)	1.275	0.138	7.26×10^{-09}
1	wildtype	*BRCA1 (c)*	5hmC/5mC	–p(K_1 / nM)	1.503	0.097	1.21×10^{-12}
1	wildtype	*BRCA1 (c)*	5mC/5mC	–p(K_1 / nM)	1.794	0.068	1.57×10^{-17}
1	wildtype	*Hey2 (short)*	5hmC/5mC	–p(K_1 / nM)	2.387	0.243	1.15×10^{-08}
1	wildtype	*Hey2 (short)*	5mC/5mC	–p(K_1 / nM)	2.036	0.155	5.44×10^{-11}
1	wildtype	*Hey2 (long)*	5hmC/5mC	–p(K_1 / nM)	0.784	0.066	8.04×10^{-07}
1	wildtype	*Hey2 (long)*	5mC/5mC	–p(K_1 / nM)	1.006	0.081	5.98×10^{-07}
1	T/A/Y/N	*CDKN2A*	5hmC/5mC	–p(K_1 / nM)	2.223	0.042	6.82×10^{-24}
1	T/A/Y/N	*CDKN2A*	5mC/5mC	–p(K_1 / nM)	1.804	0.108	1.42×10^{-13}
1	T/A/Y/N	*BRCA1 (b)*	5hmC/5mC	–p(K_1 / nM)	0.875	0.153	1.08×10^{-05}
1	T/A/Y/N	*BRCA1 (b)*	5mC/5mC	–p(K_1 / nM)	0.493	0.138	1.83×10^{-03}
1	T/A/Y/N	*BRCA1 (c)*	5hmC/5mC	–p(K_1 / nM)	1.754	0.109	2.95×10^{-13}
1	T/A/Y/N	*BRCA1 (c)*	5mC/5mC	–p(K_1 / nM)	1.690	0.093	2.28×10^{-14}
1	T/A/Y/N	*Hey2 (short)*	5hmC/5mC	–p(K_1 / nM)	1.920	0.132	3.95×10^{-12}
1	T/A/Y/N	*Hey2 (short)*	5mC/5mC	–p(K_1 / nM)	0.840	0.118	6.83×10^{-07}
1	T/A/Y/N	*Hey2 (long)*	5hmC/5mC	–p(K_1 / nM)	1.417	0.058	1.65×10^{-09}
1	T/A/Y/N	*Hey2 (long)*	5mC/5mC	–p(K_1 / nM)	1.547	0.078	9.62×10^{-09}
2	wildtype	*CDKN2A*	5hmC/5mC	–p(K_1 / nM)	1.662	0.081	2.98×10^{-23}
2	wildtype	*CDKN2A*	5mC/5mC	–p(K_1 / nM)	1.166	0.082	1.80×10^{-17}
2	wildtype	*CDKN2A*	n/a	–p(K_2 / nM)	2.458	0.175	2.73×10^{-17}
2	wildtype	*BRCA1 (b)*	5hmC/5mC	–p(K_1 / nM)	2.424	0.258	1.50×10^{-11}
2	wildtype	*BRCA1 (b)*	5mC/5mC	–p(K_1 / nM)	2.775	0.238	2.66×10^{-14}
2	wildtype	*BRCA1 (b)*	n/a	–p(K_2 / nM)	0.452	0.100	5.44×10^{-05}

Supplementary Table A.9b Model estimates for multi-site MBD–DNA binding. Fitted estimates for macroscopic descriptors of the binding isotherms with shared estimates as shown in Figure 8.4 and Figure 8.5.

Degree	Protein	Probe	Combination	Term	Estimate	Std. Error	*p*-value
2	wildtype	*BRCA1 (c)*	5hmC/5mC	–p(K_1 / nM)	1.314	0.070	2.09×10^{-21}
2	wildtype	*BRCA1 (c)*	5mC/5mC	–p(K_1 / nM)	0.853	0.091	1.15×10^{-11}
2	wildtype	*BRCA1 (c)*	n/a	–p(K_2 / nM)	2.583	0.205	1.54×10^{-15}
2	wildtype	*Hey2 (short)*	5hmC/5mC	–p(K_1 / nM)	0.939	0.250	6.05×10^{-04}
2	wildtype	*Hey2 (short)*	5mC/5mC	–p(K_1 / nM)	0.595	0.274	3.65×10^{-02}
2	wildtype	*Hey2 (short)*	n/a	–p(K_2 / nM)	2.665	0.516	9.04×10^{-06}
2	wildtype	*Hey2 (long)*	5hmC/5mC	–p(K_1 / nM)	1.500	0.141	6.26×10^{-09}
2	wildtype	*Hey2 (long)*	5mC/5mC	–p(K_1 / nM)	1.009	0.133	7.48×10^{-07}
2	wildtype	*Hey2 (long)*	n/a	–p(K_2 / nM)	0.572	0.112	9.05×10^{-05}
2	T/A/Y/N	*CDKN2A*	5hmC/5mC	–p(K_1 / nM)	1.219	0.118	6.33×10^{-13}
2	T/A/Y/N	*CDKN2A*	5mC/5mC	–p(K_1 / nM)	1.868	0.079	1.64×10^{-25}
2	T/A/Y/N	*CDKN2A*	n/a	–p(K_2 / nM)	2.699	0.312	8.35×10^{-11}
2	T/A/Y/N	*BRCA1 (b)*	5hmC/5mC	–p(K_1 / nM)	0.189	0.098	6.04×10^{-02}
2	T/A/Y/N	*BRCA1 (b)*	5mC/5mC	–p(K_1 / nM)	0.435	0.095	4.62×10^{-05}
2	T/A/Y/N	*BRCA1 (b)*	n/a	–p(K_2 / nM)	2.787	0.142	2.12×10^{-22}
2	T/A/Y/N	*BRCA1 (c)*	5hmC/5mC	–p(K_1 / nM)	2.627	0.451	7.83×10^{-07}
2	T/A/Y/N	*BRCA1 (c)*	5mC/5mC	–p(K_1 / nM)	2.728	0.412	5.64×10^{-08}
2	T/A/Y/N	*BRCA1 (c)*	n/a	–p(K_2 / nM)	1.106	0.119	1.13×10^{-11}
2	T/A/Y/N	*Hey2 (short)*	5hmC/5mC	–p(K_1 / nM)	0.362	0.119	4.26×10^{-03}
2	T/A/Y/N	*Hey2 (short)*	5mC/5mC	–p(K_1 / nM)	1.352	0.151	5.06×10^{-11}
2	T/A/Y/N	*Hey2 (short)*	n/a	–p(K_2 / nM)	2.485	0.231	3.13×10^{-13}
2	T/A/Y/N	*Hey2 (long)*	5hmC/5mC	–p(K_1 / nM)	2.096	0.155	1.53×10^{-10}
2	T/A/Y/N	*Hey2 (long)*	5mC/5mC	–p(K_1 / nM)	1.757	0.128	1.27×10^{-10}
2	T/A/Y/N	*Hey2 (long)*	n/a	–p(K_2 / nM)	1.037	0.082	4.28×10^{-10}
3	wildtype	*CDKN2A*	5hmC/5mC	–p(K_1 / nM)	1.629	0.103	9.06×10^{-19}
3	wildtype	*CDKN2A*	5mC/5mC	–p(K_1 / nM)	1.073	0.192	1.82×10^{-06}
3	wildtype	*CDKN2A*	n/a	–p(K_2 / nM)	2.217	0.496	6.22×10^{-05}
3	wildtype	*CDKN2A*	n/a	–p(K_3 / nM)	3.621	1.293	7.82×10^{-03}
3	wildtype	*BRCA1 (b)*	5hmC/5mC	–p(K_1 / nM)	2.269	0.142	2.07×10^{-18}
3	wildtype	*BRCA1 (b)*	5mC/5mC	–p(K_1 / nM)	3.062	0.157	2.15×10^{-21}
3	wildtype	*BRCA1 (b)*	n/a	–p(K_2 / nM)	0.358	0.069	6.73×10^{-06}
3	wildtype	*BRCA1 (b)*	n/a	–p(K_3 / nM)	3.072	0.143	6.58×10^{-23}
3	wildtype	*BRCA1 (c)*	5hmC/5mC	–p(K_1 / nM)	1.035	0.093	1.08×10^{-13}
3	wildtype	*BRCA1 (c)*	5mC/5mC	–p(K_1 / nM)	0.506	0.111	5.27×10^{-05}
3	wildtype	*BRCA1 (c)*	n/a	–p(K_2 / nM)	2.082	0.106	7.33×10^{-22}
3	wildtype	*BRCA1 (c)*	n/a	–p(K_3 / nM)	2.850	0.234	7.55×10^{-15}
3	wildtype	*Hey2 (short)*	5hmC/5mC	–p(K_1 / nM)	2.497	0.214	1.25×10^{-13}
3	wildtype	*Hey2 (short)*	5mC/5mC	–p(K_1 / nM)	2.833	0.335	5.76×10^{-10}

Supplementary Table A.9c Model estimates for multi-site MBD–DNA binding. Fitted estimates for macroscopic descriptors of the binding isotherms with shared estimates as shown in Figure 8.4 and Figure 8.5.

Degree	Protein	Probe	Combination	Term	Estimate	Std. Error	*p*-value
3	wildtype	*Hey2 (short)*	n/a	$-\mathrm{p}(K_2 / \mathrm{nM})$	0.500	0.129	4.34×10^{-04}
3	wildtype	*Hey2 (short)*	n/a	$-\mathrm{p}(K_3 / \mathrm{nM})$	2.833	0.269	2.24×10^{-12}
3	wildtype	*Hey2 (long)*	5hmC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	2.179	0.292	1.38×10^{-06}
3	wildtype	*Hey2 (long)*	5mC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	3.481	0.277	1.04×10^{-09}
3	wildtype	*Hey2 (long)*	n/a	$-\mathrm{p}(K_2 / \mathrm{nM})$	n/d	n/d	$1.00 \times 10^{+00}$
3	wildtype	*Hey2 (long)*	n/a	$-\mathrm{p}(K_3 / \mathrm{nM})$	n/d	n/d	$1.00 \times 10^{+00}$
3	T/A/Y/N	*CDKN2A*	5hmC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	0.684	0.135	9.18×10^{-06}
3	T/A/Y/N	*CDKN2A*	5mC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	1.473	0.120	4.14×10^{-15}
3	T/A/Y/N	*CDKN2A*	n/a	$-\mathrm{p}(K_2 / \mathrm{nM})$	n/d	n/d	6.74×10^{-01}
3	T/A/Y/N	*CDKN2A*	n/a	$-\mathrm{p}(K_3 / \mathrm{nM})$	n/d	n/d	6.72×10^{-01}
3	T/A/Y/N	*BRCA1 (b)*	5hmC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	0.217	0.064	1.55×10^{-03}
3	T/A/Y/N	*BRCA1 (b)*	5mC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	0.449	0.062	7.50×10^{-09}
3	T/A/Y/N	*BRCA1 (b)*	n/a	$-\mathrm{p}(K_2 / \mathrm{nM})$	n/d	n/d	9.96×10^{-01}
3	T/A/Y/N	*BRCA1 (b)*	n/a	$-\mathrm{p}(K_3 / \mathrm{nM})$	n/d	n/d	9.96×10^{-01}
3	T/A/Y/N	*BRCA1 (c)*	5hmC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	2.166	0.144	4.21×10^{-18}
3	T/A/Y/N	*BRCA1 (c)*	5mC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	2.582	0.145	1.42×10^{-20}
3	T/A/Y/N	*BRCA1 (c)*	n/a	$-\mathrm{p}(K_2 / \mathrm{nM})$	0.765	0.081	1.16×10^{-11}
3	T/A/Y/N	*BRCA1 (c)*	n/a	$-\mathrm{p}(K_3 / \mathrm{nM})$	2.900	0.147	2.97×10^{-22}
3	T/A/Y/N	*Hey2 (short)*	5hmC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	0.320	0.115	8.17×10^{-03}
3	T/A/Y/N	*Hey2 (short)*	5mC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	1.195	0.174	3.63×10^{-08}
3	T/A/Y/N	*Hey2 (short)*	n/a	$-\mathrm{p}(K_2 / \mathrm{nM})$	2.482	0.196	3.41×10^{-15}
3	T/A/Y/N	*Hey2 (short)*	n/a	$-\mathrm{p}(K_3 / \mathrm{nM})$	3.344	0.257	1.50×10^{-15}
3	T/A/Y/N	*Hey2 (long)*	5hmC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	2.049	0.192	1.12×10^{-08}
3	T/A/Y/N	*Hey2 (long)*	5mC/5mC	$-\mathrm{p}(K_1 / \mathrm{nM})$	1.644	0.117	2.03×10^{-10}
3	T/A/Y/N	*Hey2 (long)*	n/a	$-\mathrm{p}(K_2 / \mathrm{nM})$	1.701	0.107	3.33×10^{-11}
3	T/A/Y/N	*Hey2 (long)*	n/a	$-\mathrm{p}(K_3 / \mathrm{nM})$	0.612	0.112	5.46×10^{-05}

Supplementary Table A.10a Model estimates for MeCP2 Ala-122 variants. Fitted estimates as shown in Figure 8.6.

Degree	Protein	Probe	Combination	Term	Estimate	Std. Error	*p*-value
1	T/G/Y/N	artificial	5mC/5mC	$K_{\mathrm{d}1}$	786 nM	260 nM	6.80×10^{-03}
1	T/G/Y/N	artificial	5hmC/5mC	$K_{\mathrm{d}1}$	630 nM	188 nM	2.62×10^{-03}
1	T/A/Y/N	artificial	5mC/5mC	$K_{\mathrm{d}1}$	92 nM	13 nM	2.40×10^{-08}
1	T/A/Y/N	artificial	5hmC/5mC	$K_{\mathrm{d}1}$	9 nM	2 nM	5.94×10^{-06}
1	T/V/Y/N	artificial	5mC/5mC	$K_{\mathrm{d}1}$	109 nM	48 nM	3.82×10^{-02}
1	T/V/Y/N	artificial	5hmC/5mC	$K_{\mathrm{d}1}$	72 nM	20 nM	3.44×10^{-03}
1	T/L/Y/N	artificial	5mC/5mC	$K_{\mathrm{d}1}$	253 nM	24 nM	2.91×10^{-08}

Supplementary Table A.10b Model estimates for MeCP2 Ala-122 variants. Fitted estimates as shown in Figure 8.6.

Degree	Protein	Probe	Combination	Term	Estimate	Std. Error	*p*-value
1	T/L/Y/N	artificial	5hmC/5mC	K_{d1}	133 nM	14 nM	1.08×10^{-07}
1	T/I/Y/N	artificial	5mC/5mC	K_{d1}	166 nM	38 nM	5.29×10^{-04}
1	T/I/Y/N	artificial	5hmC/5mC	K_{d1}	221 nM	39 nM	8.37×10^{-06}
1	T/C/Y/N	artificial	5mC/5mC	K_{d1}	77 nM	11 nM	1.20×10^{-08}
1	T/C/Y/N	artificial	5hmC/5mC	K_{d1}	26 nM	4 nM	1.82×10^{-08}

A.3 Source code

Data source to page 9. The definition of a CpG island may vary. I relate to the terms implicit in the latest release of the UCSC genome annotation.

```r
library(annotatr)

ref_genomes <- c("hg38", "mm10")

genes_with_promoter_CGI <- function(g) {

  promoters <- annotatr::build_annotations(paste0(g, "_genes_promoters"), genome = g)
  annot_cgi <- annotatr::build_annotations(paste0(g, "_cpg_islands"), genome = g)

  ol <- findOverlaps(query = promoters, subject = annot_cgi)

  all_genes <- length(unique(mcols(promoters)$symbol))
  cgi_proms <- length(unique(mcols(promoters[queryHits(ol)])$symbol))

  cgi_proms / all_genes

}

sapply(ref_genomes, genes_with_promoter_CGI)
```

The result obtained for the human reference genome "hg38" of 60% is close to the 56% estimated from experimental data (Antequera & Bird, 1993).

Source code to create Figure 5.3.

The surface-contact areas between the protein and nucleic acid objects in a .pdb structure file were determined using a local version of dr_dasa. In addition to the specifications of the van der Waals radii for each atom in the canonical amino acids and DNA nucleobase, the following non-standard residues were included.

```
RESIDUE       ATOM   MSE    8
ATOM   N      1.65   1
ATOM   CA     1.87   0
ATOM   CB     1.87   0
ATOM   CG     1.76   0
ATOM   SE     1.85   0
ATOM   CE     1.87   0
ATOM   C      1.76   0
ATOM   O      1.40   1
RESIDUE       NUCL   5CM    20
ATOM   P      1.90   0
ATOM   OP1    1.40   1
ATOM   OP2    1.40   1
ATOM   O5'    1.40   1
ATOM   C5'    1.80   0
ATOM   C4'    1.80   0
ATOM   O4'    1.40   1
ATOM   C3'    1.80   0
ATOM   O3'    1.40   1
ATOM   C2'    1.80   0
ATOM   C1'    1.80   0
ATOM   N1     1.60   1
ATOM   C2     1.80   0
ATOM   O2     1.40   1
ATOM   N3     1.60   1
ATOM   C4     1.80   0
ATOM   N4     1.60   1
ATOM   C5     1.80   0
ATOM   C6     1.80   0
ATOM   C5A    1.80   0
```

In the `.pdb` files, the protein was chain A and the DNA strands were chain B and chain C.

```bash
#!/bin/bash
mkdir -p results
cd results
find ../*.pdb | xargs -L1 -I {} dr_sasa -m 1 -i {} -v vdw.radii.pdb -chain A -chain BC
&> error_contact.log > analysis_contact.log
```

The total buried surface on chain A by B and C is designated `A<---BC` in the log file. The visualization of the calculated data was done with *R*.

```r
#!/usr/bin/env R
library(tidyverse)

read_csa <- function(file) {

  readr::read_tsv(file) %>%
    tidyr::pivot_longer(-1) %>%
    magrittr::set_colnames(c("target", "buried_by", "value")) %>%
    dplyr::mutate(
      tar = stringr::str_extract(target, "^.+(?=/.+/)"),
      tar_chain = stringr::str_extract(target, "(?<=/)[A-Z]+(?=/)"),
      tar_pos  = as.integer(stringr::str_extract(target, "[0-9]+$")),
      bur = stringr::str_extract(buried_by, "^.+(?=/.+/)"),
      bur_chain = stringr::str_extract(buried_by, "(?<=/)[A-Z]+(?=/)"),
      bur_pos  = as.integer(stringr::str_extract(buried_by, "[0-9]+$"))
    )

}

csa_by_res <- sapply(list.files(pattern = "A_vs_BC.by_res.tsv"), read_csa,
  simplify = FALSE, USE.NAMES = TRUE) %>% bind_rows(.id = "file") %>%
  mutate(protein = substr(file, 1, 4))
```

To keep track of the amino acid positions (`aa_pos`) of interest for each structure, the following object was created.

```
included_aa <- list(
  `6d1t` = 2:81,
  `6cnq` = 146:225,
  `6ccg` = 2:81,
  `4lg7` = 76:167,
  `3c2i` = 90:181) %>% enframe() %>% unnest(cols = "value") %>%
  rename(protein = "name", aa_pos = "value") %>%
  mutate(chain = "A") %>% mutate(include_aa = TRUE)
```

Likewise, an object `included_nt` for each `nt_pos` and `chain` (not shown).

The output format was dependent on whether the protein or the nucleic acid was the buried object, but in either case, the protein should always be plotted along the horizontal axis.

```
# assuming these residues are part of the sequence; else NULL
if ("GLY" %in% csa_by_res$bur) aa_match <- c("bur_pos" = "aa_pos", "bur_chain" =
    "chain", "protein" = "protein")
if ("GLY" %in% csa_by_res$tar) aa_match <- c("tar_pos" = "aa_pos", "tar_chain" =
    "chain", "protein" = "protein")
if ("DT"  %in% csa_by_res$bur) nt_match <- c("bur_pos" = "nt_pos", "bur_chain" =
    "chain", "protein" = "protein")
if ("DT"  %in% csa_by_res$tar) nt_match <- c("tar_pos" = "nt_pos", "tar_chain" =
    "chain", "protein" = "protein")

csa_by_res <- csa_by_res %>%
  left_join(included_aa, by = aa_match) %>%
  left_join(included_nt, by = nt_match) %>%
  filter(include_aa == TRUE & include_nt == TRUE) %>%
  # renumber to increase legibility
  rename(aa_pos = names(aa_match)[[1]], aa_chain = names(aa_match)[[2]],
         nt_pos = names(nt_match)[[1]], nt_chain = names(nt_match)[[2]]) %>%
  # consider relative offset for the nucleotides
  group_by(protein, nt_chain) %>%
  mutate(rel_nt_pos = nt_pos - min(nt_pos))

ggplot(cas_by_res, aes(x = aa_pos, y = rel_nt_pos, fill = value)) +
  geom_tile() +
  scale_fill_distiller(type = "div", palette = "RdGy") +
  coord_fixed() +
  facet_grid(vars(protein), vars(nt_chain))
```

Source code to create Figure 7.9. For calculating the values of panel *c*, use:

```
library(tidyverse)

df  # grouped data frame with fluorescence intensities for each event as rows and
    # each fluorescence channel as column variables ("channel_G" and "channel_R")

events_by_threshold <- function(threshold = 0, x = df, cols = c("FSC-A"), ...) {

  x %>%
    drop_na(any_of(cols)) %>%
    summarize(across(any_of(cols), list(pos = ~ sum(. >= threshold),
                                        neg = ~ sum(. <  threshold))),
              .groups = "keep")

}

df_roc <- tibble(threshold = unique(c(seq(3.0, 3.5, 1e-2), seq(3.5, 4.5, 1e-3)))) %>%
  mutate(tmp = pmap(., events_by_threshold, cols = c("channel_G", "channel_R"))) %>%
  unnest(tmp)

df_roc <- df_roc %>%
  mutate(TP = case_when(type == "MM_G_CC_R" ~ channel_G_pos,
                        type == "MM_R_CC_G" ~ channel_R_pos),
         TN = case_when(type == "MM_G_CC_R" ~ channel_R_neg,
                        type == "MM_R_CC_G" ~ channel_G_neg),
         FP = case_when(type == "MM_G_CC_R" ~ channel_R_pos,
                        type == "MM_R_CC_G" ~ channel_G_pos),
         FN = case_when(type == "MM_G_CC_R" ~ channel_G_neg,
                        type == "MM_R_CC_G" ~ channel_R_neg),
         )

df_roc %>%
  filter(threshold == 3.5) %>%
  transmute(sensitivity = TP / (TP + FN) * 100,
            specificity = TN / (TN + FP) * 100,
            FDR = FP / (FP + TP) * 100)
```

Source code to create Figure 7.14.

```
library(tidyverse)

df  # data frame with sequences of reads per position (pos1, pos2, pos3, pos4)
    # per UMI (umi1, umi2) and the demultiplexed barcode (bc1 -> bc_assigned)

rearrange_gen_fen <- function(x) {

  # fast helper to convert X_ABC/Y_DEF/Z_GHI (provided as X_ABC_Y_DEF_Z_GHI)
  # into X/Y/Z_ABC/DEF/GHI; although column names have to be specified, i.e.
  # only works here with 4 degenerate positions

  data.table::data.table(x)[, data.table::tstrsplit(x, split = "_")][, .(paste(
      paste(V1, V3, V5, V7, sep = "/"),
      paste(V2, V4, V6, V8, sep = "/"), sep = "_"))]$V1

}

compare_after_before <- function(x, barcode_after, barcode_before, by = "site") {

  res <- x %>% select(!any_of("pos")) %>%
    # subset and associate A -> "after screen"; B -> "before screen"
    filter(bc_assigned %in% c(barcode_after, barcode_before)) %>%
    left_join(tibble(bc_assigned = c(barcode_after, barcode_before),
                     bc_function = c("after", "before")), by = "bc_assigned") %>%
    # translate and keep record of genotype
    mutate(across(starts_with("pos") ~ str_c(Biostrings::GENETIC_CODE[.], "_", .)))
  if (by == "site") {

    res <- res %>%
      pivot_longer(starts_with("pos"), names_to = "pos", values_to = "value") %>%
      group_by(bc_function, pos, value)

  } else {

    res <-  res %>%
      unite(value, starts_with("pos"), sep = "_") %>%
      mutate(across(value, ~ rearrange_gen_fen)) %>%
      group_by(bc_function, value)

  }

  res <- res %>%
    summarize(n = n_distinct(umi1, umi2, na.rm = TRUE), .groups = "keep") %>%
    # number of genotypes that support the phenotype
    separate(value, into = c("value", "codon"), sep = "_", remove = FALSE) %>%
    summarize(n = sum(n), n_gen = n_distinct(codon), .groups = "keep") %>%
    # fractional composition
    ungroup(value) %>% mutate(pc = n / sum(n)) %>%
    # arrange nicely in wide format
    pivot_wider(id_cols = c(group_vars(.), any_of(c("value", "codon"))),
                values_from = c("n", "n_gen", "pc"), names_from = "bc_function",
                values_fn = list, names_sep = ".") %>%
    unnest(where(is.list)) %>%
    # ignore disappearing phenotypes
    filter(!is.na(n.after)) %>%
    # make sure observations below the limit of detection in initial pool are not
    # lost (NA) when calculating the enrichment factor
    mutate(pc.before = replace_na(pc.before, min(pc.before, na.rm = TRUE))) %>%
    # enrichment between fractions
    mutate(fold_enrichment = pc.after / pc.before) %>%
    arrange(desc(fold_enrichment))

  attr(res, "barcode_before") <- barcode_before
  attr(res, "barcode_after")  <- barcode_after

  res

}
```

Source code to create Figure 8.4.

```
library(summerrband)
library(magrittr)

# data preparation

aff_gels_context <- list(
  # gels with two or more bands quantified in ImageQuant TL
  list(file = "gel_XXX.txt",
       protein = "p1388",
       probe = "Ah", # shorthand
       conc  = c(2^(10:0), 0),
       exclude = c("lane_1")),
  # (etc. for all quantitated gels)
  NULL
)

aff_gels_context <- summerrband::iqtl_import_all(aff_gels_context, path = ".")

aff_gels_for_fit <- aff_gels_context %>%
  group_by(protein, probe) %>% filter(!is.na(vol_frac)) %>%
  pivot_wider(names_from = band_id, values_from = vol_frac,
              id_cols = c(conc, group_vars(.)), values_fn = list) %>%
  unnest(cols = where(is.list)) %>%
  mutate(probe_major = substr(probe, 1, 1), probe_minor = substr(probe, 2, 2)) %>%
  group_by(protein, probe_major) %>%
  replace_na(list(band_0 = 0, band_1 = 0, band_2 = 0, band_3 = 0))

aff_gels_for_fit %>%
  ggplot(aes(x = conc, y = band_1 + 2 * band_2 + 3 * band_3, color = protein)) +
  geom_line(stat = "summary", fun.data = mean_se) +
  scale_x_log10() + facet_wrap(vars(probe))

# test different combinations of parameters for fitting

aff_gels_models <- tibble(expand.grid(shared = c(TRUE, FALSE), degree = 2:3,
                                      type = c("micro", "macro"))) %>%
  bind_rows(tibble(shared = FALSE, degree = 1, type = "macro")) %>%
  group_by(shared, degree, type)

aff_gels_models$x <- list(aff_gels_for_fit)

# helper function to pass the parameters accordingly

fit_x <- function(x, shared = FALSE, type = "macro", degree = 3, ...) {

  ARGS <- list(FUN = fit_binding_isotherm,
               formula = band_1 + 2 * band_2 + 3 * band_3 ~ conc,
               limits_K_d = c(1e0, 1e6), start_K_d = c(1e-1, 1e6),
               correlation = c(ab = 0.91, ac = 0.81, abc = 0.68),
               newdata = data.frame(conc = 10^seq(-3, 3, length.out = 100)))

  if (shared == TRUE) {

    EXPR <- substitute(do.call(model_cleanly_groupwise, args = c(list(
      x, type = T0, degree = D0, INDEX = I0), ARGS)),
      list(T0 = type, D0 = degree, I0 = "probe_minor"))

  } else {

    x <- group_by(x, probe_minor, .add = TRUE)

    EXPR <- substitute(do.call(model_cleanly_groupwise, args = c(list(
      x, type = T0, degree = D0, INDEX = NULL), ARGS)), list(T0 = type, D0 = degree))

  }

  res <- eval(EXPR)

  if (shared == TRUE) {

    res$augment_new[which(lengths(res$model) > 0)] <- map(
      res$model[which(lengths(res$model) > 0)], augment_shared_isotherms,
      newdata = data.frame(ARGS$newdata, probe_minor = rep(c("m", "h"), each = nrow(
        ARGS$newdata))), INDEX = probe_minor)

  }

  res

}

aff_gels_models$model <- pmap(aff_gels_models, fit_x); aff_gels_models  # fitted
```

Source code to create Figure 9.3.

```
library(tidyverse)
library(summerrband)

# ---- panel a ----

summerrband:::gpf_fraction_plot(binding_constants = c(K1 = 1e0), type = "macro")

# ---- panel b ----

op <- par(no.readonly = TRUE); par(mfrow = c(2, 4))

lapply(c(1e0, 1e1, 1e2, 1e4), function(K2) summerrband:::gpf_fraction_plot(
  binding_constants = c(K1 = 1e0, K2 = K2), type = "macro"))
lapply(c(1e2, 1e3, 1e5, 1e7), function(K3) summerrband:::gpf_fraction_plot(
  binding_constants = c(K1 = 1e0, K2 = 1e4, K3 = K3), type = "macro"))

par(op)

# ---- panel c ----

set.seed(113917)

df <- data.frame(x = 2^(seq(-8, 8)))
df$y_real <- summerrband::gpf_fraction_bound(df$x, c(K1 = 1, K2 = 10), type = "macro")
df$y_noise <- df$y_real + rnorm(length(df$y_real), sd = 0.1)

plot(df$x, df$y_real, log = "x", col = "gray"); points(df$x, df$y_noise, pch = 3)

mf <- lapply(1:3, function(d) summerrband::fit_binding_isotherm(df,
             formula = y_noise ~ x, degree = d))  # fitting to different polynom degrees
```

```r
lapply(mf, broom::glance)  # goodness of fit measueres

mapply(function(x, col) points(broom::augment(x, newdata = data.frame(
  x = 10^seq(-6, 3, length.out = 100))), col = col, type = "l"), mf, c("red", "blue",
                                                                       "green"))

# ---- panel d -----

m_truth = c(pK_d1 = 0, pK_d2 = 1, pK_d3 = 3)

noisy_data <- function(x = df$x, degree, noise, ..., truth = unname(m_truth^-1)) {

  data.frame(x = x, y = summerrband::gpf_fraction_bound(x, truth[1:degree],
     type = "macro") + rnorm(x, sd = noise))

}

df2 <- tibble(expand.grid(noise = c(0.1, 0.2, 0.3), degree = 1:3, run = 1:100)) %>%
  mutate(df = pmap(., .f = noisy_data),
         mf_1 = map(df, summerrband::fit_binding_isotherm, formula = y ~ x, degree = 1),
         mf_2 = map(df, summerrband::fit_binding_isotherm, formula = y ~ x, degree = 2),
         mf_3 = map(df, summerrband::fit_binding_isotherm, formula = y ~ x, degree = 3))

pf2 <- pivot_longer(df2, starts_with("mf_"), names_prefix = "mf_", names_to = "m_degree",
                    values_to = "m_fit") %>% group_by(noise, degree, run) %>%
  mutate(glance = map(m_fit, broom::glance), tidy = map(m_fit, broom::tidy)) %>%
  unnest(glance) %>% mutate(m_selected = min(AIC) == AIC)

# number of "correct" model choices based on AIC/BIC

ggplot(pf2, aes(x = paste(degree, m_degree), fill = m_degree, alpha = m_selected)) +
  geom_bar(position = "stack") + facet_grid(rows = vars(noise))

# recall accuracy

assign_closest_permutation <- function(x, truth) {

  length(x) <- length(truth)

  per <- map(1:length(truth), ~ names(truth)) %>%
    cross() %>% keep(~ length(unique(.x)) == length(truth)) %>% map(unlist)

  res <- per[which.min(sapply(per, function(y) sum((x - truth[y])^2, na.rm = TRUE)))]

  res[[1]][1:length(na.omit(x))]

}

pf2 %>% filter(m_degree == degree) %>%  # alternatively, check: m_selected == TRUE
  unnest(tidy) %>% filter(startsWith(term, "pK_")) %>%
  mutate(closest_term = assign_closest_permutation(estimate, truth = m_truth[1:unique(
                                                   m_degree)])) %>%
  ggplot(aes(x = paste(m_degree, noise), y = estimate, color = closest_term)) +
  geom_hline(yintercept = m_truth) + geom_boxplot()
```

Source code to create Figure 9.4.

```r
library(tidyverse)
library(summerrband)

# ---- panel a ----

summerrband:::gpf_fraction_plot(binding_constants = c(a = 1e0), type = "micro")

# ---- panel b, c ----

op <- par(no.readonly = TRUE); par(mfrow = c(2, 4))

for (g_ab in c(0.5, 2.0, 0.05, 20, 1.0)) {

    lapply(c(1e0, 1e1, 1e2, 1e4), function(b) summerrband:::gpf_fraction_plot(
      binding_constants = c(a = 1e0, b = b, ab = 1e0 * b * g_ab), type = "micro"))

}

par(op)
```

A.4 Credits and license information

The author is grateful to the following for permission to reproduce copyright material:

W. W. Norton & Company, Inc. for the epigraph at page v from *The Blind Watchmaker* by Richard Dawkins. Reproduced with permission of W. W. Norton & Company on behalf of the author.

License of Figure 1.4, page 13. Figure 1.4 is redrawn based on Figure 1 in Blattler and Farnham (2013). The adapted image is licensed like the original under the Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

License of Figure 4.1, page 39. Figure 4.1 is my own work assembled from material released in the Public Domain (CC0): The eagle was photographed by Jean Beaufort and the small flying plane in the sky was photographed by Linda Thomas.

License of Figure 6.2, page 61. Figure 6.2 is reprinted in part from Figure 2 in Buchmuller et al. (2020): Panels *a* and *b* were removed, panels *c*–*h* renumbered and the axis labels and titles reformatted. The adapted image is licensed like the original under the Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

License of Figure 6.4, page 63. Figure 6.4 is reprinted from Figure 3 in Buchmuller et al. (2020): Panels *a*, *b*, *g*–*i* were removed and panels *c*–*f* renumbered, the axis labels and titles reformatted and a legend added. The adapted image is licensed like the original under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

License of Figure 6.5, page 64. Figure 6.5 was adapted from Figure 3 in Buchmuller et al. (2020): Panels *a–f* and *h* were removed and panel *g* and *i* rearranged. The test statistic has been added to panel *i*. The adapted image is licensed like the original under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

License of Figure A.21, page 179. Figure A.21 appeared as Supplementary Figure 14 in Buchmuller et al. (2020). The image is licensed under the Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

License of Supplementary Figure A.3, page 169. Supplementary Figure A.3 appeared as Supplementary Figures 1–3 in Buchmuller et al. (2020): Redundant axis labels were removed and the label texts adapted. The images are licensed like the originals under the Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

License of Supplementary Figure A.4, page 170. Supplementary Figure A.4 appeared as Supplementary Figures 4–5 in Buchmuller et al. (2020): Redundant axis labels were removed and the label texts adapted. The images are licensed like the originals under the Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

License of Supplementary Figure A.5, page 171. Supplementary Figure A.5 appeared as Supplementary Figures 6–9 in Buchmuller et al. (2020): Redundant axis labels were removed and the label texts adapted. The images are licensed like the originals under the Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

B

References

Aberg, K. A., Chan, R. F., Shabalin, A. A., Zhao, M., Turecki, G., Staunstrup, N. H., ... van den Oord, E. J. (2017). A MBD-seq protocol for large-scale methylome-wide studies with (very) low amounts of DNA. *Epigenetics*, *12*(9), 743–750.

Acevedo-Rocha, C. G., Reetz, M. T., & Nov, Y. (2015). Economical analysis of saturation mutagenesis experiments. *Sci. Rep.*, *5*(1), 10654.

Adair, G. S., Bock, A. V., & Field, H. (1925). The hemoglobin system VI. The oxygen dissociation curve of hemoglobin. *J. Biol. Chem.*, *63*(2), 529–545.

Adey, A. & Shendure, J. (2012). Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res.*, *22*(6), 1139–1143.

Adhya, S. (2001). Regulatory Genes. In S. Brenner & J. H. Miller (Eds.), *Encyclopedia of Genetics*. (pp. 1652–1657). Cambridge (MA): Academic Press.

Agarwal, N., Becker, A., Jost, K. L., Haase, S., Thakur, B. K., Brero, A., ... Cardoso, M. C. (2011). MeCP2 Rett mutations affect large scale chromatin organization. *Hum. Mol. Genet.*, *20*(21), 4187–4195.

Äijö, T., Huang, Y., Mannerström, H., Chavez, L., Tsagaratou, A., Rao, A., & Lähdesmäki, H. (2016). A probabilistic generative model for quantification of DNA modifications enables analysis of demethylation pathways. *Genome Biol.*, *17*(1), 49.

Äijö, T., Yue, X., Rao, A., & Lähdesmäki, H. (2016). LuxGLM: a probabilistic covariate model for quantification of DNA methylation modifications with complex experimental designs. *Bioinformatics*, *32*(17), i511–i519.

Allen, M. D., Yamasaki, K., Ohme-Takagi, M., Tateno, M., & Suzuki, M. (1998). A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. *EMBO J.*, *17*(18), 5484–5496.

Allis, C. D., Caparros, M. L., Jenuwein, T., & Reinberg, D. (2015). *Epigenetics*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Altschuler, S. E., Lewis, K. A., & Wuttke, D. S. (2012). Practical strategies for the evaluation of high-affinity protein/nucleic acid interactions. *J. Nucleic Acids Invest.*, *4*(1), 3.

Antequera, F. & Bird, A. P. (1993). Number of CpG Islands and Genes in Human and Mouse. *Proc. Natl. Acad. Sci. U.S.A.*, *90*(24), 11995–11999.

Arand, J., Spieler, D., Karius, T., Branco, M. R., Meilinger, D., Meissner, A., ... Lustik, P. (2012). In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet.*, *8*(6), e1002750.

Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y., & Shirakawa, M. (2008). Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature*, *455*(7214), 818–821.

Arnold, F. H. (1996). Directed evolution: creating biocatalysts for the future. *Chem. Eng. Sci.*, *51*(23), 5091–5102.

Bachman, M., Uribe-Lewis, S., Yang, X., Burgess, H. E., Iurlaro, M., Reik, W., ... Balasubramanian, S. (2015). 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.*, *11*(8), 555–557.

Bachman, M., Uribe-Lewis, S., Yang, X., Williams, M., Murrell, A., & Balasubramanian, S. (2014). 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.*, *6*(12), 1049–1055.

Ballestar, E., Yusufzai, T. M., & Wolffe, A. P. (2000). Effects of Rett syndrome mutations of the methyl-CpG-binding domain of the transcriptional repressor MeCP2 on selectivity for association with methylated DNA. *Biochemistry*, *39*(24), 7100–7106.

Baubec, T., Ivánek, R., Lienert, F., & Schübeler, D. (2013). Methylation-dependent and -independent genomic targeting principles of the MBD protein family. *Cell*, *153*(2), 480–492.

Beck, R. & Burtscher, H. (1994). Introduction of arbitrary sequences into genes by use of class IIs restriction enzymes. *Nucleic Acids Res.*, *22*(5), 886–887.

Béhé, M. & Felsenfeld, G. (1981). Effects of methylation on a synthetic polynucleotide: the B–Z transition in poly(dG-m5dC)·poly(dG-m5dC). *Proc. Natl. Acad. Sci. U.S.A.*, *78*(3), 1619–1623.

Ben-Naim, A. (2001a). The binding isotherm. In *Cooperativity and Regulation in Biochemical Processes*. (pp. 25–49). Boston, MA: Springer.

— (2001b). Three-site systems: nonadditivity and long-range correlations. In *Cooperativity and Regulation in Biochemical Processes*. (pp. 143–191). Boston, MA: Springer.

Beveridge, D. L., Dixit, S. B., Barreiro, G., & Thayer, K. M. (2004). Molecular dynamics simulations of DNA curvature and flexibility: helix phasing and premelting. *Biopolymers*, *73*(3), 380–403.

Bigler, K. (2020, 9). *Affinity maturation by directed evolution of engineered MBD domains for the detection of mod-*

ified CpG dyads. (Bachelor thesis). Technische Universität Dortmund, Dortmund.

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.*, *8*(7), 1499–1504.

— (1995). Gene number, noise reduction and biological complexity. *Trends Genet.*, *11*(3), 94–100.

— (2007). Perceptions of epigenetics. *Nature*, *447*(7143), 396–398.

Birla, B. S. & Chou, H. H. (2015). Rational design of high-number dsDNA fragments based on thermodynamics for the construction of full-length genes in a single reaction. *PLoS One*, *10*(12), e0145682.

Blattler, A. & Farnham, P. J. (2013). Cross-talk between site-specific transcription factors and DNA methylation states. *J. Biol. Chem.*, *288*(48), 34287–34294.

Bock, C., Tomazou, E. M., Brinkman, A. B., Müller, F., Simmer, F., Gu, H., ... Meissner, A. (2010). Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.*, *28*(10), 1106–1114.

Boder, E. T. & Wittrup, K. D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.*, *15*(6), 553–557.

— (1998). Optimal screening of surface-displayed polypeptide libraries. *Biotechnol. Prog.*, *14*(1), 55–62.

Bonvin, E., Radaelli, E., Bizet, M., Luciani, F., Calonne, E., Putmans, P., ... Fuks, F. (2019). TET2-dependent hydroxymethylome plasticity reduces melanoma initiation and progression. *Cancer Res.*, *79*(3), 482–494.

Booth, M. J., Branco, M. R., Ficz, G., Oxley, D., Krueger, F., Reik, W., & Balasubramanian, S. (2012). Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, *336*(6083), 934–937.

Booth, M. J., Marsico, G., Bachman, M., Beraldi, D., & Balasubramanian, S. (2014). Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat. Chem.*, *6*(5), 435–440.

Breiling, A. & Lyko, F. (2015). Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenet. Chromatin*, *8*(1), 24.

Brenowitz, M., Senear, D. F., Shea, M. A., & Ackers, G. K. (1986). Quantitative DNase footprint titration: a method for studying protein–DNA interactions. In *Enzyme Structure Part K*. (Vol. 130, pp. 132–181). Academic Press.

Breter, H. J., Seibert, G., & Zahn, R. K. (1976). The use of high-pressure liquid cation-exchange chromatography for determination of the 5-methylcytosine content of DNA. *J. Chromatogr. A*, *118*(2), 242–249.

Brinkman, A. B., Simmer, F., Ma, K., Kaan, A., Zhu, J., & Stunnenberg, H. G. (2010). Whole-genome DNA methylation profiling using MethylCap-seq. *Methods*, *52*(3), 232–236.

Bronner, C., Alhosin, M., Hamiche, A., & Mousli, M. (2019). Coordinated dialogue between UHRF1 and DNMT1 to ensure faithful inheritance of methylated DNA patterns. *Genes*, *10*(1).

Brown, T. A. (2007). *Genomes 3*. New York: Garland Science.

Buchmuller, B. C., Jung, A., Muñoz-López, Á., & Summerer, D. (2021). Programmable tools for targeted analysis of epigenetic DNA modifications. *Curr. Opin. Chem. Biol.*, *63*, 1–10.

Buchmuller, B. C., Kosel, B., & Summerer, D. (2020). Complete profiling of methyl-CpG-binding domains for combinations of cytosine modifications at CpG dinucleotides reveals differential read-out in normal and Rett-associated states. *Sci. Rep.*, *10*(1), 4053–9.

Buck-Koehntop, B. A. & Defossez, P. A. (2014). On how mammalian transcription factors recognize methylated DNA. *Epigenetics*, *8*(2), 131–137.

Burden, A. F., Manley, N. C., Clark, A. D., Gartler, S. M., Laird, C. D., & Hansen, R. S. (2005). Hemimethylation and non-CpG methylation levels in a promoter region of human LINE-1 (L1) repeated elements. *J. Biol. Chem.*, *280*(15), 14413–14419.

Bushnell, B. (2019). *BBMap short read aligner*. The Joint Genome Institute. Retrieved from https://sourceforge.net/projects/bbmap/

Cadwell, R. C. & Joyce, G. F. (1992). Randomization of genes by PCR mutagenesis.. *Genome Res.*, *2*(1), 28–33.

Capuano, F., Mülleder, M., Kok, R., Blom, H. J., & Ralser, M. (2014). Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species. *Anal. Chem.*, *86*(8), 3697–3702.

Carell, T., Kurz, M. Q., Müller, M., Rossa, M., & Spada, F. (2018). Non-canonical bases in the genome: the regulatory information layer in DNA. *Angew. Chem., Int. Ed.*, *57*(16), 4296–4312.

Carugo, O. (2003). How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. *J. Appl. Crystallogr.*, *36*(1), 125–128.

Casadesús, J. & Low, D. (2006). Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.*, *70*(3), 830–856.

Charbonneau, M. È., Janvore, J., & Mourez, M. (2009). Autoprocessing of the Escherichia coli AIDA-I Autotransporter: A new mechanism involving acidic residues in the junction region. *J. Biol. Chem.*, *284*(25), 17340–17351.

Charlet, J., Duymich, C. E., Lay, F. D., Mundbjerg, K., Dalsgaard Sørensen, K., Liang, G., & Jones, P. A. (2016). Bivalent regions of cytosine methylation and H3K27 acetylation suggest an active role for DNA methylation at enhancers. *Mol. Cell*, *62*(3), 422–431.

Charlton, J., Downing, T. L., Smith, Z. D., Gu, H., Clement, K., Pop, R., ... Meissner, A. (2018). Global delay in nascent strand DNA methylation. *Nat. Struct. Mol. Biol.*, *25*(4), 327–332.

Chatonnet, F., Pignarre, A., Sérandour, A. A., Caron, G., Avner, S., Robert, N., ... Salbert, G. (2019). The hydroxymethylome of multiple myeloma identifies FAM72D as a 1q21 marker linked to proliferation. *Haematologica*, haematol.2019.222133.

Chen, P. Y., Feng, S., Joo, J. W. J., Jacobsen, S. E., & Pellegrini, M. (2011). A comparative analysis of DNA methylation across human embryonic stem cell lines. *Genome Biol.*, *12*(7), R62.

Cherf, G. M. & Cochran, J. R. (2015). Applications of yeast surface display for protein engineering. In B. Liu (Ed.), *Yeast Surface Display: methods, protocols, and applications*. (Vol. 1319, pp. 155–175). Humana Press, New York, NY.

Chiu, T. P., Rao, S., Mann, R. S., Honig, B., & Rohs, R. (2017). Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein–DNA binding. *Nucleic Acids Res.*, *45*(21), 12565–12576.

Chou, C. C., Wei, S. Y., Lou, Y. C., & Chen, C. (2017). In-depth study of DNA binding of Cys2His2 finger domains in testis zinc-finger protein. *PLoS One*, *12*(4), e0175051.

Chouliaras, L., Mastroeni, D., Delvaux, E., Grover, A., Kenis, G., Hof, P. R., ... van den Hove, D. L. A. (2013). Consistent decrease in global DNA methylation and hydroxymethylation in the hippocampus of Alzheimer's disease patients. *Neurobiol. Aging*, *34*(9), 2091–2099.

Chung, H. S. & Raetz, C. R. H. (2010). Interchangeable domains in the Kdo transferases of Escherichia coli and Haemophilus influenzae. *Biochemistry*, *49*(19), 4126–4137.

Chung, M., Goroncy, K., Kolesnikova, A., Schönauer, D., & Schwaneberg, U. (2020). Display of functional nucleic acid polymerase on Escherichia coli surface and its application in directed polymerase evolution. *Biotechnol. Bioeng.*, *117*(12), 3699–3711.

Church, G. M., Sussman, J. L., & Kim, S. H. (1977). Secondary structural complementarity between DNA and proteins. *Proc. Natl. Acad. Sci. U.S.A.*, *74*(4), 1458–1462.

Cipriany, B. R., Murphy, P. J., Hagarman, J. A., Cerf, A., Latulippe, D., Levy, S. L., ... Craighead, H. G. (2012). Real-time analysis and selection of methylated DNA by fluorescence-activated single molecule sorting in a nanofluidic channel. *Proc. Natl. Acad. Sci. U.S.A.*, *109*(22), 8477–8482.

Clark, S. J., Harrison, J., & Frommer, M. (1995). CpNpG methylation in mammalian cells. *Nat. Genet.*, *10*(1), 20–27.

Clouaire, T., de las Heras, J. I., Merusi, C., & Stancheva, I. (2010). Recruitment of MBD1 to target genes requires sequence-specific interaction of the MBD domain with methylated DNA. *Nucleic Acids Res.*, *38*(14), 4620–4634.

Condliffe, D., Wong, A., Troakes, C., Proitsi, P., Patel, Y., Chouliaras, L., ... Lunnon, K. (2014). Cross-region reduction in 5-hydroxymethylcytosine in Alzheimer's disease brain. *Neurobiol. Aging*, *35*(8), 1850–1854.

Connelly, J. C., Cholewa-Waclaw, J., Webb, S., Steccanella, V., Waclaw, B., & Bird, A. P. (2020). Absence of MeCP2 binding to non-methylated GT-rich sequences in vivo. *Nucleic Acids Res.*, *48*(7), 3542–3552.

Connolly, K. M., Ilangovan, U., Wojciak, J. M., Iwahara, M., & Clubb, R. T. (2000). Major groove recognition by three-stranded beta-sheets: affinity determinants and conserved structural features. *J. Mol. Biol.*, *300*(4), 841–856.

Cramer, J. M., Scarsdale, J. N., Walavalkar, N. M., Buchwald, W. A., Ginder, G. D., & Williams, D. C. (2014). Probing the dynamic distribution of bound states for methylcytosine-binding domains on DNA. *J. Biol. Chem.*, *289*(3), 1294–1302.

Crawford, D. J., Liu, M. Y., Nabel, C. S., Cao, X. J., Garcia, B. A., & Kohli, R. M. (2016). Tet2 catalyzes stepwise 5-methylcytosine oxidation by an iterative and de novo mechanism. *J. Am. Chem. Soc.*, *138*(3), 730–733.

Crick, F. H. C., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, *192*(4809), 1227–1232.

Dai, Q., Sanstead, P. J., Peng, C. S., Han, D., He, C., & Tokmakoff, A. (2016). Weakened N3 hydrogen bonding by 5-formylcytosine and 5-carboxylcytosine reduces their base-pairing stability. *ACS Chem. Biol.*, *11*(2), 470–477.

Dantas Machado, A. C., Zhou, T., Rao, S., Goel, P., Rastogi, C., Lazarovici, A., ... Rohs, R. (2015). Evolving insights on how cytosine methylation affects protein–DNA binding. *Brief. Funct. Genomics*, *14*(1), 61–73.

Daugherty, P. S., Chen, G., Olsen, M. J., Iverson, B. L., & Georgiou, G. (1998). Antibody affinity maturation using bacterial surface display. *Protein Eng.*, *11*(9), 825–832.

de Marco, A. (2009). Strategies for successful recombinant expression of disulfide bond-dependent proteins in Escherichia coli. *Microb. Cell Fact.*, *8*(1), 26.

Deaton, A. M. & Bird, A. P. (2011). CpG islands and the regulation of transcription. *Genes Dev.*, *25*(10), 1010–1022.

Deckard III, C. E., Banerjee, D. R., & Sczepanski, J. T. (2019). Chromatin structure and the pioneering transcription factor FOXA1 regulate TDG-mediated removal of 5-formylcytosine from DNA. *J. Am. Chem. Soc.*, *141*(36), 14110–14114.

Deichmann, U. (2016). Epigenetics: the origins and evolution of a fashionable topic. *Dev. Biol.*, *416*(1), 249–254.

Delaney, J. C. & Essigmann, J. M. (2004). Mutagenesis, genotoxicity, and repair of 1-methyladenine, 3-alkylcytosines, 1-methylguanine, and 3-methylthymine in alkB Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, *101*(39), 14051–14056.

Denks, K., Vogt, A., Sachelaru, I., Petriman, N. A., Kudva, R., & Koch, H. G. (2014). The Sec translocon mediated protein transport in prokaryotes and eukaryotes. *Mol. Membr. Biol.*, *31*(2-3), 58–84.

Dickerson, R. E. (1992). DNA structure from A to Z. *Meth. Enzymol.*, *211*, 67–111.

Domcke, S., Bardet, A. F., Adrian Ginno, P., Hartl, D., Burger, L., & Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, *528*(7583), 575–579.

Du, Q., Luu, P. L., Stirzaker, C., & Clark, S. J. (2015). Methyl-CpG-binding domain proteins: readers of the epigenome. *Epigenomics*, *7*(6), 1051–1073.

Du, Q., Wang, Z., & Schramm, V. L. (2016). Human DNMT1 transition state structure. *Proc. Natl. Acad. Sci. U.S.A.*, *113*(11), 2916–2921.

Dutta, S., Madan, S., & Sundar, D. (2016). Exploiting the recognition code for elucidating the mechanism of zinc finger protein–DNA interactions. *BMC Genomics*, *17*(Suppl 13), 1037.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., ... Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, *38*(12), 1378–1385.

Ehrlich, M., Gama-Sosa, M. A., Huang, L. H., Midgett, R. M., Kuo, K. C., McCune, R. A., & Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res.*, *10*(8), 2709–2721.

Fahnestock, M., Johnston, A. J., Ross, P., & Tsien, R. Y. (1991). *DNA sequencing*. *World Patent No. WO9106678A1.* WIPO (PCT).

Farrell, C. M., O'Leary, N. A., Harte, R. A., Loveland, J. E., Wilming, L. G., Wallin, C., ... Pruitt, K. D. (2013). Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.*, *42*(D1), D865–D872.

Feldmann, A., Ivánek, R., Murr, R., Gaidatzis, D., Burger, L., & Schübeler, D. (2013). Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet.*, *9*(12), e1003994.

Ficz, G., Branco, M. R., Seisenberger, S., Santos, F., Krueger, F., Hore, T. A., ... Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, *473*(7347), 398–402.

Finak, Greg, Frelinger, Jacob, Jiang, Wenxin, Newell, Evan W., Ramey, John, Davis, Mark M., ... Gottardo, Raphael (2014). OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Comput. Biol.*, *10*(8), e1003806.

Flick, K. E., Jurica, M. S., Monnat, R. J., & Stoddard, B. L. (1998). DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI. *Nature*, *394*(6688), 96–101.

Fraga, M. F., Ballestar, E., Montoya, G., Taysavang, P., Wade, P. A., & Esteller, M. (2003). The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. *Nucleic Acids Res.*, *31*(6), 1765–1774.

Franklin, D. (2019). P152R Mutation Within MeCP2 Can Cause Loss of DNA-Binding Selectivity. *Interdiscip. Sci.*, *9*, 2365–11.

Frederick, C. A., Saal, D., van der Marel, G. A., van Boom, J. H., Wang, A. H., & Rich, A. (1987). The crystal structure of d(GGm5CCGGCC): the effect of methylation on A-DNA structure and stability. *Biopolymers*, *26 Suppl*(S0), S145–60.

Free, A., Wakefield, R. I., Smith, B. O., Dryden, D. T., Barlow, P. N., & Bird, A. P. (2001). DNA recognition by the methyl-CpG-binding domain of MeCP2. *J. Biol. Chem.*, *276*(5), 3353–3360.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., ... Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U.S.A.*, *89*(5), 1827–1831.

Fu, T., Liu, L., Yang, Q. L., Wang, Y., Xu, P., Zhang, L., ... Zhang, L. (2019). Thymine DNA glycosylase recognizes the geometry alteration of minor grooves induced by 5-formylcytosine and 5-carboxylcytosine. *Chem. Sci.*, *10*(31), 7407–7417.

Fujii, S., Wang, A. H., van der Marel, G., van Boom, J. H., & Rich, A. (1982). Molecular structure of (m5·dC-dG)3: the role of the methyl group on 5-methyl cytosine in stabilizing Z-DNA. *Nucleic Acids Res.*, *10*(23), 7879–7892.

Fuks, F., Hurd, P. J., Wolf, D., Nan, X., Bird, A. P., & Kouzarides, T. (2003). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *J. Biol. Chem.*, *278*(6), 4035–4040.

Fuxreiter, M., Mezei, M., Simon, I., & Osman, R. (2005). Interfacial water as a "hydration fingerprint" in the noncognate complex of BamHI. *Biophys. J.*, *89*(2), 903–911.

Galvão, T. C. & Thomas, J. O. (2005). Structure-specific binding of MeCP2 to four-way junction DNA through its methyl CpG-binding domain. *Nucleic Acids Res.*, *33*(20), 6603–6609.

Gardiner-Garden, M. & Frommer, M. (1987). CpG Islands in vertebrate genomes. *J. Mol. Biol.*, *196*(2), 261–282.

Ghosh, R. P., Horowitz-Scherer, R. A., Nikitina, T., Gierasch, L. M., & Woodcock, C. L. (2008). Rett syndrome-causing mutations in human MeCP2 result in diverse structural changes that impact folding and DNA interactions. *J. Biol. Chem.*, *283*(29), 20523–20534.

Gibson, D. G., Young, L., Chuang, R. Y., Venter, J. C., Hutchison, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, *6*(5), 343–345.

Giehr, P., Kyriakopoulos, C., Lepikhov, K., Wallner, S., Wolf, V., & Lustik, P. (2018). Two are better than one: hPoxBS - hairpin oxidative bisulfite sequencing. *Nucleic Acids Res.*, *46*(15), e88–e88.

Gieß, M., Muñoz-López, Á., Buchmuller, B. C., Kubik, G., & Summerer, D. (2019). Programmable protein–DNA cross-linking for the direct capture and quantification of 5-formylcytosine. *J. Am. Chem. Soc.*, *141*(24), 9453–9457.

Gieß, M., Witte, A., Jasper, J., Koch, O., & Summerer, D. (2018). Complete, programmable decoding of oxidized 5-methylcytosine nucleobases in DNA by chemoselective blockage of universal transcription-activator-like effector repeats. *J. Am. Chem. Soc.*.

Gillam, E. M. J. (2014). Error-prone PCR and effective generation of gene variant libraries for directed evolution. In E. M. J. Gillam, J. N. Copp, & D. F. Ackerley (Eds.), *Directed Evolution Library Creation: methods and protocols*. (No. 1179, 2 ed.). New York: Humana Press, New York, NY.

Globisch, D., Münzel, M., Müller, M., Michalakis, S., Wagner, M., Koch, S., ... Carell, T. (2010). Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One*, *5*(12), e15367.

Gorin, A. A., Zhurkin, V. B., & Olson, W. K. (1995). B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.*, *247*(1), 34–48.

Gromiha, M. M., Siebers, J. G., Selvaraj, S., Kono, H., & Sarai, A. (2004). Intermolecular and intramolecular readout mechanisms in protein–DNA recognition. *J. Mol. Biol.*, *337*(2), 285–294.

Grunau, C., Clark, S. J., & Rosenthal, A. (2001). Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res.*, *29*(13), e65–e65.

Guntrum, M., Vlasova, E., & Davis, T. L. (2017). Asymmetric DNA methylation of CpG dyads is a feature of secondary DMRs associated with the Dlk1/Gtl2 imprinting cluster in mouse. *Epigenet. Chromatin*, *10*(1), 457.

Guo, F., Li, X., Liang, D., Li, T., Zhu, P., Guo, H., ... Xu, G. l. (2014). Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. *Cell Stem Cell*, *15*(4), 447–459.

Gurdon, J. B., Laskey, R. A., & Reeves, O. R. (1975). The developmental capacity of nuclei transplanted from keratinized skin cells of adult frogs.. *J. Embryol. Exp. Morphol.*, *34*(1), 93–112.

Habibi, E., Brinkman, Arie B., Arand, J., Kroeze, Leonie I., Kerstens, Hindrik H. D., Matarese, F., ... Stunnenberg, Hendrik G. (2013). Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell*, *13*(3), 360–369.

Hahne, F., LeMeur, N., Brinkman, R. R., Ellis, B., Haaland, P., Sarkar, D., ... Gentleman, R. (2009). FlowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinf.*, *10*(1), 106.

Han, D., Lu, X., Shih, A. H., Nie, J., You, Q., Xu, M. M., ... He, C. (2016). A Highly Sensitive and Robust Method for Genome-wide 5hmC Profiling of Rare Cell Populations. *Mol. Cell*, *63*(4), 711–719.

Hanes, J. & Plückthun, A. (1997). In vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl. Acad. Sci. U.S.A.*, *94*(10), 4937–4942.

Hardwick, J. S., Ptchelkine, D., El-Sagheer, A. H., Tear, I., Singleton, D., Phillips, S. E. V., ... Brown, T. (2017). 5-Formylcytosine does not change the global structure of DNA. *Nat. Struct. Mol. Biol.*, *24*(6), 544–552.

Hartley, R. V. L. (1928). Transmission of Information1. *Bell Syst. Tech. J.*, *7*(3), 535–563.

Hashimoto, H., Horton, J. R., Zhang, X., Bostick, M., Jacobsen, S. E., & Cheng, X. (2008). The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature*, *455*(7214), 826–829.

Hashimoto, H., Liu, Y., Upadhyay, A. K., Chang, Y., Howerton, S. B., Vertino, P. M., ... Cheng, X. (2012). Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res.*, *40*(11), 4841–4849.

Hashimoto, H., Olanrewaju, Y. O., Zheng, Y., Wilson, G. G., Zhang, X., & Cheng, X. (2014). Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence. *Genes Dev.*, *28*(20), 2304–2313.

Hashimoto, H., Pais, J. E., Zhang, X., Saleh, L., Fu, Z. Q., Dai, N., ... Cheng, X. (2014). Structure of a Naegleria Tet-like dioxygenase in complex with 5-methylcytosine DNA. *Nature*, *506*(7488), 391–395.

Hay, I. D. & Lithgow, T. (2019). Filamentous phages: masters of a microbial sharing economy. *EMBO Rep.*, *20*(6).

He, Y. F., Li, B. Z., Li, Z., Liu, P., Wang, Y., Tang, Q., ... Xu, G. l. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, *333*(6047), 1303–1307.

Heimer, B. W., Tam, B. E., & Sikes, H. D. (2015). Characterization and directed evolution of a methyl-binding domain protein for high-sensitivity DNA methylation analysis. *Protein Eng., Des. Sel.*, *28*(12), 543–551.

Henderson, I. R., Navarro-Garcia, F., Desvaux, M., Fernandez, R. C., & Ala'Aldeen, D. (2004). Type V protein secretion pathway: the autotransporter story. *Microbiol. Mol. Biol. Rev.*, *68*(4), 692–744.

Hendrich, B. & Tweedie, S. (2003). The methyl-CpG-binding domain and the evolving role of DNA methylation in animals. *Trends Genet.*, *19*(5), 269–277.

Hill, A. (1910). The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J. Physiol. (Oxford, U. K.)*, *40*, iv–vii.

Ho, K. L., McNae, I. W., Schmiedeberg, L., Klose, R. J., Bird, A. P., & Walkinshaw, M. D. (2008). MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Mol. Cell*, *29*(4), 525–531.

Hodges-Garcia, Y. & Hagerman, P. J. (1995). Investigation of the influence of cytosine methylation on DNA flexibility. *J. Biol. Chem.*, *270*(1), 197–201.

Hon, G. C., Rajagopal, N., Shen, Y., McCleary, D. F., Yue, F., Dang, M. D., & Ren, B. (2013). Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat. Genet.*, *45*(10), 1198–1206.

Hong, S. & Cheng, X. (2016). DNA Base Flipping: a General Mechanism for Writing, Reading, and Erasing DNA Modifications. *Adv. Exp. Med. Biol.*, *945*, 321–341.

Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J. Biol. Chem.*, *175*(1), 315–332.

Hu, L., Li, Z., Cheng, J., Rao, Q., Gong, W., Liu, M., ... Xu, Y. (2013). Crystal structure of TET2–DNA complex: insight into TET-mediated 5mC oxidation. *Cell*, *155*(7), 1545–1555.

Hu, L., Lu, J., Cheng, J., Rao, Q., Li, Z., Hou, H., ... Xu, Y. (2015). Structural insight into substrate preference for TET-mediated oxidation. *Nature*, *527*(7576), 118–122.

Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H. N., ... Zhu, H. (2013). DNA methylation presents distinct binding sites for human transcription factors. *eLife*, *2*, e00726.

Huh, I., Zeng, J., Park, T., & Yi, S. V. (2013). DNA methylation and transcriptional noise. *Epigenet. Chromatin*, *6*(1), 9–10.

Ibrahim, A., Papin, C., Mohideen-Abdul, K., Gras, S. L., Stoll, I., Bronner, C., ... Hamiche, A. (2021). MeCP2 is a microsatellite binding protein that protects CA repeats from nucleosome invasion. *Science*, *372*(6549), eabd5581.

Illingworth, R., Kerr, A., Desousa, D., Jørgensen, H., Ellis, P., Stalker, J., ... Bird, A. P. (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.*, *6*(1), e22.

Illingworth, R. S. & Bird, A. P. (2009). CpG islands–'a rough guide'. *FEBS Lett.*, *583*(11), 1713–1720.

Illingworth, R. S., Gruenewald-Schneider, U., Webb, S., Kerr, A. R. W., James, K. D., Turner, D. J., ... Bird, A. P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.*, *6*(9), e1001134.

Ip, J. P. K., Mellios, N., & Sur, M. (2018). Rett syndrome: insights into genetic, molecular and circuit mechanisms. *Nat. Rev. Neurosci.*, *19*(6), 368–382.

Iqbal, K., Jin, S. G., Pfeifer, G. P., & Szabó, P. E. (2011). Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proc. Natl. Acad. Sci. U.S.A.*, *108*(9), 3642–3647.

Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., ... Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, *333*(6047), 1300–1303.

Iurlaro, M., Ficz, G., Oxley, D., Raiber, E. A., Bachman, M., Booth, M. J., ... Reik, W. (2013). A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.*, *14*(10), R119.

Iurlaro, M., McInroy, G. R., Burgess, H. E., Dean, W., Raiber, E. A., Bachman, M., ... Reik, W. (2016). In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. *Genome Biol.*, *17*(1), 141–9.

Iwafuchi-Doi, M. & Zaret, K. S. (2014). Pioneer transcription factors in cell Reprogramming. *Genes Dev.*, *28*(24), 2679–2692.

Iwan, K., Rahimoff, R., Kirchner, A., Spada, F., Schröder, A. S., Kosmatchev, O., ... Carell, T. (2018). 5-Formylcytosine to cytosine conversion by C-C bond cleavage in vivo. *Nat. Chem. Biol.*, *14*(1), 72–78.

Jacob, F. & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, *3*(3), 318–356.

Jaenisch, R. & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, *33*(Suppl 3), 245–254.

Jankowski, J. (2020, 2). *Gerichtete Evolution der Methyl-CpG-Bindungsdomäne für die Analyse von oxidierten 5-Methylcytosin-Guanin-Dinukleotiden*. (Bachelor thesis). Technische Universität Dortmund, Dortmund.

Jeong, Y. E. R., Lenz, S. A. P., & Wetmore, S. D. (2020). DFT study on the deglycosylation of methylated, oxidized, and canonical pyrimidine nucleosides in water: Implications for epigenetic regulation and DNA repair. *J. Phys. Chem. B*, *124*(12), 2392–2400.

Ji, D., Lin, K., Song, J., & Wang, Y. (2014). Effects of Tet-induced oxidation products of 5-methylcytosine on Dnmt1- and DNMT3a-mediated cytosine methylation. *Mol. BioSyst.*, *10*(7), 1749–1752.

Jiang, D., Zhang, Y., Hart, R. P., Chen, J., Herrup, K., & Li, J. (2015). Alteration in 5-hydroxymethylcytosine-mediated epigenetic regulation leads to Purkinje cell vulnerability in ATM deficiency. *Brain*, *138*(Pt 12), 3520–3536.

Jin, S. G., Jiang, Y., Qiu, R., Rauch, T. A., Wang, Y., Schackert, G., ... Pfeifer, G. P. (2011). 5-Hydroxymethylcytosine is strongly depleted in human cancers but its levels do not correlate with IDH1 mutations. *Cancer Res.*, *71*(24), 7360–7365.

Jin, S. G., Kadam, S., & Pfeifer, G. P. (2010). Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res.*, *38*(11), e125–e125.

Jin, S. G., Zhang, Z. M., Dunwell, T. L., Harter, M. R., Wu, X., Johnson, J., ... Pfeifer, G. P. (2016). Tet3 reads 5-carboxylcytosine through its CXXC domain and is a potential guardian against neurodegeneration. *Cell Rep.*, *14*(3), 493–505.

Joachimiak, A., Haran, T. E., & Sigler, P. B. (1994). Mutagenesis supports water mediated recognition in the trp repressor-operator system. *EMBO J.*, *13*(2), 367–372.

Johnson, I. D. (2010). *Molecular Probes Handbook: A Guide to Fluorescent Probes and Labeling Technologies*. (11 ed.). Life Technologies Corporation.

Jones, M. M., Castle-Clarke, S., Brooker, D., Nason, E., Huzair, F., & Chataway, J. (2014). The Structural Genomics Consortium: a knowledge platform for drug discovery: a summary. *Rand Health Q.*, *4*(3), 19.

Joosten, R. P., Joosten, K., Cohen, S. X., Vriend, G., & Perrakis, A. (2011). Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. *Bioinformatics*, *27*(24), 3392–3398.

Jørgensen, H. F., Adie, K., Chaubert, P., & Bird, A. P. (2006). Engineering a high-affinity methyl-CpG-binding protein. *Nucleic Acids Res.*, *34*(13), e96–e96.

Jose, J., Chung, J. W., Jeon, B. J., Maas, R. M., Nam, C. H., & Pyun, J. C. (2009). Escherichia coli with autodisplayed Z-domain of protein A for signal amplification of SPR biosensor. *Biosens. Bioelectron.*, *24*(5), 1324–1329.

Kaeßler, A., Olgen, S., & Jose, J. (2011). Autodisplay of catalytically active human hyaluronidase hPH-20 and testing of enzyme inhibitors. *Eur. J. Pharm. Sci.*, *42*(1-2), 138–147.

Kallenberger, L., Erb, R., Kralickova, L., Patrignani, A., Stöckli, E., & Jiricny, J. (2019). Ectopic methylation of a single persistently unmethylated CpG in the promoter of the vitellogenin gene abolishes its inducibility by estrogen through attenuation of upstream stimulating factor binding. *Mol. Cell. Biol.*, *39*(23).

Kamińska, E., Korytiaková, E., Reichl, A., Müller, M., & Carell, T. (2021). Intragenomic decarboxylation of 5-carboxy-2′-deoxycytidine. *Angew. Chem., Int. Ed.*. Advanced online publication. doi:10.1002/anie.202109995

Kangaspeska, S., Stride, B., Métivier, R., Polycarpou-Schwarz, M., Ibberson, D., Carmouche, R. P., ... Reid, G. (2008). Transient cyclical methylation of promoter DNA. *Nature*, *452*(7183), 112–115.

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., ... Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.*, *10*(1), 1930.

Khrapunov, S., Warren, C., Cheng, H., Berko, E. R., Greally, J. M., & Brenowitz, M. (2014). Unusual characteristics of the DNA binding domain of epigenetic regulatory protein MeCP2 determine its binding specificity. *Biochemistry*, *53*(21), 3379–3391.

Kibbe, W. A. (2007). OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Res.*, *35*(suppl_2), W43–W46.

Kinde, B., Gabel, H. W., Gilbert, C. S., Griffith, E. C., & 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*(22), 6800–6806.

Kitsera, N., Allgayer, J., Parsa, E., Geier, N., Rossa, M., Carell, T., & Khobta, A. (2017). Functional impacts of 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxycytosine at a single hemi-modified CpG dinucleotide in a gene promoter. *Nucleic Acids Res.*, *45*(19), 11033–11042.

Kizaki, S. & Sugiyama, H. (2014). CGmCGCG is a versatile substrate with which to evaluate Tet protein activity. *Org. Biomol. Chem.*, *12*(1), 104–107.

Klose, R. J., Sarraf, S. A., Schmiedeberg, L., McDermott, S. M., Stancheva, I., & Bird, A. P. (2005). DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. *Mol. Cell*, *19*(5), 667–678.

Korlach, J. & Turner, S. W. (2012). Going beyond five bases in DNA sequencing. *Curr. Opin. Struct. Biol.*, *22*(3), 251–261.

Korthauer, K. & Irizarry, R. A. (2018). Genome-wide repressive capacity of promoter DNA methylation is revealed through epigenomic manipulation. *bioRxiv*, 381145.

Korytiaková, E., Kamińska, E., Müller, M., & Carell, T. (2021). Deformylation of 5-Formylcytidine in Different Cell Types. *Angew. Chem., Int. Ed.*, *60*(31), 16869–16873.

Kossel, A. (1884). Ueber Guanin. *Hoppe-Seyler's Z. Physiol. Chem.*, *8*, 404–410.

Kossel, A. & Neumann, A. (1893). Ueber das Thymin, ein Spaltungsproduct der Nucleïnsäure. *Ber. Dtsch. Chem. Ges.*, *26*(3), 2753–2756.

— (1894). Darstellung und Spaltungsprodukte der Nucleïnsäure (Adenylsäure). *Ber. Dtsch. Chem. Ges.*, *27*(2), 2215–2222.

Kothari, R. M. & Shankar, V. (1976). 5-Methylcytosine content in the vertebrate deoxyribonucleic acids: species specificity. *J. Mol. Evol.*, *7*(4), 325–329.

Kraus, T. F. J., Greiner, A., Steinmaurer, M., Dietinger, V., Guibourt, V., & Kretzschmar, H. A. (2015). Genetic characterization of ten-eleven-translocation methylcytosine dioxygenase alterations in human glioma. *J. Cancer*, *6*(9), 832–842.

Krebs, A. R., Dessus-Babus, S., Burger, L., & Schübeler, D. (2014). High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *eLife*, *3*, 1074.

Kriaucionis, S. & Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, *324*(5929), 929–930.

Kribelbauer, J. F., Laptenko, O., Chen, S., Martini, G. D., Freed-Pastor, W. A., Prives, C., ... Bussemaker, H. J. (2017). Quantitative analysis of the DNA methylation sensitivity of transcription factor complexes. *Cell Rep.*, *19*(11), 2383–2395.

Krishnaraj, R., Ho, G., & Christodoulou, J. (2017). RettBASE: rett syndrome database update. *Hum. Mutat.*, *38*(8), 922–931.

Kubik, G., Batke, S., & Summerer, D. (2015). Programmable sensors of 5-hydroxymethylcytosine. *J. Am. Chem. Soc.*, *137*(1), 2–5.

Kubik, G. & Summerer, D. (2015). Deciphering epigenetic cytosine modifications by direct molecular recognition. *ACS Chem. Biol.*, *10*(7), 1580–1589.

Kweon, S. M., Zhu, B., Chen, Y., Aravind, L., Xu, S. Y., & Feldman, D. E. (2017). Erasure of Tet-oxidized 5-methylcytosine by a SRAP nuclease. *Cell Rep.*, *21*(2), 482–494.

Laget, S., Joulie, M., Le Masson, F., Sasai, N., Christians, E., Pradhan, S., ... Defossez, P. A. (2010). The human proteins MBD5 and MBD6 associate with heterochromatin but they do not bind methylated DNA. *PLoS One*, *5*(8), e11982.

Lagger, S., Connelly, J. C., Schweikert, G., Webb, S., Selfridge, J., Ramsahoye, B. H., ... Bird, A. P. (2017). MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. *PLoS Genet.*, *13*(5), e1006793.

Laguerre, M., Saux, M., Dubost, J. P., & Carpy, A. (1997). MLPP: a program for the calculation of molecular lipophilicity potential in proteins. *Pharm. Pharmacol. Commun.*, *3*(5-6), 217–222.

Laird, C. D., Pleasant, N. D., Clark, A. D., Sneeden, J. L., Hassan, K. M. A., Manley, N. C., ... Stöger, R. (2004). Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *Proc. Natl. Acad. Sci. U.S.A.*, *101*(1), 204–209.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., ... Weirauch, M. T. (2018). The human transcription factors. *Cell*, *172*(4), 650–665.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., ... Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, *9*(8), e1003118.

Lawson, C. L. & Berman, H. M. (2008). Indirect readout of DNA sequence by proteins. In P. A. Rice & C. C. Correll (Eds.), *Protein–Nucleic Acid Interactions: structural biology*. (pp. 66–90). Cambridge: Royal Society of Chemistry.

Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A. C., Riley, T. R., Sandstrom, R., ... Bussemaker, H. J. (2013). Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl. Acad. Sci. U.S.A.*, *110*(16), 6376–6381.

Lehnertz, B., Ueda, Y., Derijck, A. A. H. A., Braunschweig, U., Perez-Burgos, L., Kubicek, S., ... Peters, A. H. F. M. (2003). Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr. Biol.*, *13*(14), 1192–1200.

Lei, M., Tempel, W., Chen, S., Liu, K., & Min, J. (2019). Plasticity at the DNA recognition site of the MeCP2 mCG-binding domain. *Biochim. Biophys. Acta, Gene Regul. Mech.*, 194409.

Leighton, G. & Williams Jr, D. C. (2019). The methyl-CpG-binding domain 2 and 3 proteins and formation of the nucleosome remodeling and deacetylase complex. *J. Mol. Biol.*.

Lewis, L. C., Lo, P. C. K., Foster, J. M., Dai, N., Corrêa, I. R., Durczak, P. M., ... Ruzov, A. (2017). Dynamics of 5-carboxylcytosine during hepatic differentiation: potential general role for active demethylation by DNA repair in lineage specification. *Epigenetics*, *12*(4), 277–286.

Li, Q. Y., Xie, N. B., Xiong, J., Yuan, B. F., & Feng, Y. Q. (2018). Single-nucleotide resolution analysis of 5-hydroxymethylcytosine in DNA by enzyme-mediated deamination in combination with sequencing. *Anal. Chem.*, *90*(24), 14622–14628.

Li, S., Papale, L. A., Zhang, Q., Madrid, A., Chen, L., Chopra, P., ... Alisch, R. S. (2016). Genome-wide alterations in hippocampal 5-hydroxymethylcytosine links plasticity genes to acute stress. *Neurobiol. Dis.*, *86*, 99–108.

Lian, C. G., Xu, Y., Ceol, C., Wu, F., Larson, A., Dresser, K., ... Shi, Y. G. (2012). Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. *Cell*, *150*(6), 1135–1146.

Lin, I. H., Chen, Y. F., & Hsu, M. T. (2017). Correlated 5-hydroxymethylcytosine (5hmc) and gene expression profiles underpin gene and organ-specific epigenetic regulation in adult mouse brain and liver. *PLoS One*, *12*(1), e0170779.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., ... Ecker, J. R. (2013). Global epigenomic reconfiguration during mammalian brain development. *Science*, *341*(6146), 1237905–1237905.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., ... Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–322.

Liu, K., Lei, M., Wu, Z., Gan, B., Cheng, H., Li, Y., & Min, J. (2019). Structural analyses reveal that MBD3 is a methylated-CG binder. *FEBS J.*, *34*, 654.

Liu, K., Xu, C., Lei, M., Yang, A., Loppnau, P., Hughes, T. R., & Min, J. (2018). Structural basis for the ability of MBD domains to bind methyl-CG and TG sites in DNA. *J. Biol. Chem.*, *293*(19), 7344–7354.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., ... Law, M. (2012). Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.*, *2012*, 251364.

Liu, L., Zhang, Y., Liu, M., Wei, W., Yi, C., & Peng, J. (2020). Structural insights into the specific recognition of 5-methylcytosine and 5-hydroxymethylcytosine by TAL effectors. *J. Mol. Biol.*, *432*(4), 1035–1047.

Liu, R., Barrick, J. E., Szostak, J. W., & Roberts, R. W. (2000). Optimized synthesis of RNA-protein fusions for in vitro protein selection. In *RNA-Ligand Interac-*

tions Part B. (Vol. 318, pp. 268–293). Academic Press.

Liu, S., Wang, J., Su, Y., Guerrero, C., Zeng, Y., Mitra, D., ... Wang, Y. (2013). Quantitative assessment of Tet-induced oxidation products of 5-methylcytosine in cellular and tissue DNA. *Nucleic Acids Res.*, *41*(13), 6421–6429.

Liu, Y., Hu, Z., Cheng, J., Siejka-Zielińska, P., Chen, J., Inoue, M., ... Song, C. X. (2021). Subtraction-free and bisulfite-free specific sequencing of 5-methylcytosine and its oxidized derivatives at base resolution. *Nat. Commun.*, *12*(1), 618.

Liu, Y., Siejka-Zielińska, P., Velikova, G., Bi, Y., Yuan, F., Tomkova, M., ... Song, C. X. (2019). Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.*, *37*(4), 424–429.

Liu, Y., Olanrewaju, Y. O., Zhang, X., & Cheng, X. (2013). DNA recognition of 5-carboxylcytosine by a Zfp57 mutant at an atomic resolution of 0.97 Å. *Biochemistry*, *52*(51), 9310–9317.

Liu, Y., Zhang, X., Blumenthal, R. M., & Cheng, X. (2013). A common mode of recognition for methylated CpG. *Trends Biochem. Sci.*, *38*(4), 177–183.

Liutkeviciute, Z., Kriukienė, E., Ličytė, J., Rudytė, M., Urbanavičiūtė, G., & Klimašauskas, S. (2014). Direct decarboxylation of 5-carboxylcytosine by DNA C5-methyltransferases. *J. Am. Chem. Soc.*, *136*(16), 5884–5887.

López, V., Fernández, A. F., & Fraga, M. F. (2017). The role of 5-hydroxymethylcytosine in development, aging and age-related diseases. *Ageing Res. Rev.*, *37*, 28–38.

Lu, J., Hu, L., Cheng, J., Fang, D., Wang, C., Yu, K., ... Luo, C. (2016). A computational investigation on the substrate preference of ten-eleven-translocation 2 (TET2). *Phys. Chem. Chem. Phys.*, *18*(6), 4728–4738.

Lu, X., Han, D., Zhao, B. S., Song, C. X., Zhang, L. S., Doré, L. C., & He, C. (2015). Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Res.*, *25*(3), 386–389.

Lu, X., Song, C. X., Szulwach, K., Wang, Z., Weidenbacher, P., Jin, P., & He, C. (2013). Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. *J. Am. Chem. Soc.*, *135*(25), 9315–9317.

Lu, X. J. & Olson, W. K. (2003). 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, *31*(17), 5108–5121.

Lungu, C., Pinter, S., Broche, J., Rathert, P., & Jeltsch, A. (2017). Modular fluorescence complementation sensors for live cell detection of epigenetic signals at endogenous genomic sites. *Nat. Commun.*, *8*(1), 649.

Luty, B. A., Wasserman, Z. R., Stouten, P. F. W., Hodge, C. N., Zacharias, M., & McCammon, J. A. (1995). A molecular mechanics/grid method for evaluation of ligand-receptor interactions. *J. Comput. Chem.*, *16*(4), 454–464.

Maier, J. A. H., Möhrle, R., & Jeltsch, A. (2017). Design of synthetic epigenetic circuits featuring memory effects and reversible switching based on DNA methylation. *Nat. Commun.*, *8*(1), 15336.

Maiti, A. & Drohat, A. C. (2011). Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J. Biol. Chem.*, *286*(41), 35334–35338.

Mann, I. K., Chatterjee, R., Zhao, J., He, X., Weirauch, M. T., Hughes, T. R., & Vinson, C. (2013). CG methylated microarrays identify a novel methylated sequence bound by the CEBPB/ATF4 heterodimer that is active in vivo. *Genome Res.*, *23*(6), 988–997.

Martienssen, R. A. & Colot, V. (2001). DNA methylation and epigenetic inheritance in plants and filamentous fungi. *Science*, *293*(5532), 1070–1074.

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., & Neufeld, J. D. (2012). PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinf.*, *13*(1), 31.

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, *7*(1), 29–59.

Maurer, J., Jose, J., & Meyer, T. F. (1997). Autodisplay: one-component system for efficient surface display and release of soluble recombinant proteins from Escherichia coli. *J. Bacteriol.*, *179*(3), 794–804.

Maurer, S., Buchmuller, B. C., Ehrt, C., Jasper, J., Koch, O., & Summerer, D. (2018). Overcoming conservation in TALE–DNA interactions: a minimal repeat scaffold enables selective recognition of an oxidized 5-methylcytosine. *Chem. Sci.*, *9*, 7247–7252.

Maurer, S., Gieß, M., Koch, O., & Summerer, D. (2016). Interrogating key positions of size-reduced tale repeats reveals a programmable sensor of 5-carboxylcytosine. *ACS Chem. Biol.*, *11*(12), 3294–3299.

Mayer-Jung, C., Moras, D., & Timsit, Y. (1998). Hydration and recognition of methylated CpG steps in DNA. *EMBO J.*, *17*(9), 2709–2718.

McCafferty, J., Griffiths, A. D., Winter, G., & Chiswell,

D. J. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. *Nature*, *348*(6301), 552–554.

McCarthy, D., Pulverer, W., Weinhaeusel, A., Diago, O. R., Hogan, D. J., Ostertag, D., & Hanna, M. M. (2016). MethylMeter: bisulfite-free quantitative and sensitive DNA methylation profiling and mutation detection in FFPE samples. *Epigenomics*, *8*(6), 747–765.

Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L., & Bird, A. P. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell*, *58*(3), 499–507.

Menafra, R., Brinkman, A. B., Matarese, F., Franci, G., Bartels, S. J. J., Nguyen, L., ... Stunnenberg, H. G. (2014). Genome-wide binding of MBD2 reveals strong preference for highly methylated loci. *PLoS One*, *9*(6), e99603.

Mezei, P. D. & Csonka, G. I. (2016). Features of the interactions between the methyl-CpG motif and the arginine residues on the surface of MBD proteins. *Struct. Chem.*, *27*(4), 1317–1326.

Miles, A. J., Ramalli, S. G., & Wallace, B. A. (2021). DichroWeb, a website for calculating protein secondary structure from circular dichroism spectroscopic data. *Protein Sci.*.

Miyazaki, K. & Arnold, F. H. (1999). Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function. *J. Mol. Evol.*, *49*(6), 716–720.

Mohn, F., Weber, M., Schübeler, D., & Roloff, T. C. (2009). Methylated DNA Immunoprecipitation (MeDIP). In J. Tost (Ed.), *DNA Methylation: methods and protocols*. (Vol. 507, pp. 55–64). Humana Press, New York, NY.

Mooers, B. H., Schroth, G. P., Baxter, W. W., & Ho, P. S. (1995). Alternating and non-alternating dG-dC hexanucleotides crystallize as canonical A-DNA. *J. Mol. Biol.*, *249*(4), 772–784.

Moran, K. L., Shlyakhtina, Y., & Portal, M. M. (2021). The role of non-genetic information in evolutionary frameworks. *Crit. Rev. Biochem. Mol. Biol.*, *56*(3), 1–29.

Morgan, D. K. & Whitelaw, E. (2008). The case for transgenerational epigenetic inheritance in humans. *Mamm. Genome*, *19*(6), 394–397.

Morgan, H. D., Santos, F., Green, K., Dean, W., & Reik, W. (2005). Epigenetic reprogramming in mammals. *Hum. Mol. Genet.*, *14 Spec No 1*(suppl_1), R47–58.

Muñoz-López, Á. & Summerer, D. (2018). Recognition of oxidized 5-methylcytosine derivatives in DNA by natural and engineered protein scaffolds. *Chem. Rec.*, *18*(1), 105–116.

Muñoz-López, Á., Buchmuller, B. C., Wolffgramm, J., Jung, A., Hussong, M., Kanne, J., ... Summerer, D. (2020). Designer receptors for nucleotide-resolution analysis of genomic 5-methylcytosine by cellular imaging. *Angew. Chem., Int. Ed.*, *59*(23), 8927–8931.

Münzel, M., Globisch, D., Brückl, T., Wagner, M., Welzmiller, V., Michalakis, S., ... Carell, T. (2010). Quantification of the sixth DNA base hydroxymethylcytosine in the brain. *Angew. Chem., Int. Ed.*, *49*(31), 5375–5377.

Nair, S. S., Coolen, M. W., Stirzaker, C., Song, J. Z., Statham, A. L., Strbenac, D., ... Clark, S. J. (2011). Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG-binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics*, *6*(1), 34–44.

Nan, X., Meehan, R. R., & Bird, A. P. (1993). Dissection of the methyl-CpG-binding domain from the chromosomal protein MeCP2. *Nucleic Acids Res.*, *21*(21), 4886–4892.

Nanan, K. K., Sturgill, D. M., Prigge, M. F., Thenoz, M., Dillman, A. A., Mandler, M. D., & Oberdoerffer, S. (2019). TET-catalyzed 5-carboxylcytosine promotes CTCF binding to suboptimal sequences genome-wide. *iScience*, *19*, 326–339.

Nechin, J., Tunstall, E., Raymond, N., Hamagami, N., Pathmanabhan, C., Forestier, S., & Davis, T. L. (2019). Hemimethylation of CpG dyads is characteristic of secondary DMRs associated with imprinted loci and correlates with 5-hydroxymethylcytosine at paternally methylated sequences. *Epigenet. Chromatin*, *12*(1), 457.

Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Anselmi, F., Parlato, C., ... Oliviero, S. (2015). Single-base resolution analysis of 5-formyl and 5-carboxyl cytosine reveals promoter DNA methylation dynamics. *Cell Rep.*, *10*(5), 674–683.

Ngo, T. T. M., Yoo, J., Dai, Q., Zhang, Q., He, C., Aksimentiev, A., & Ha, T. (2016). Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.*, *7*(1), 10813–9.

Nilsson, B., Moks, T., Jansson, B., Abrahmsén, L., Elmblad, A., Holmgren, E., ... Uhlén, M. (1987). A synthetic IgG-binding domain based on staphylococcal protein A. *Protein Eng., Des. Sel.*, *1*(2), 107–113.

Norambuena, T. & Melo, F. (2010). The protein–DNA interface database. *BMC Bioinf.*, *11*(1), 1–12.

Oh, G., Ebrahimi, S., Carlucci, M., Zhang, A., Nair, A., Groot, D. E., ... Petronis, A. (2018). Cytosine modifications exhibit circadian oscillations that are involved in epigenetic diversity and aging. *Nat. Commun.*, *9*(1), 644.

Ohki, I., Shimotake, N., Fujita, N., Nakao, M., & Shirakawa, M. (1999). Solution structure of the methyl-CpG-binding domain of the methylation-dependent transcriptional repressor MBD1. *EMBO J.*, *18*(23), 6653–6661.

Ohki, I., Shimotake, N., Fujita, N., Jee, J. G., Ikegami, T., Nakao, M., & Shirakawa, M. (2001). Solution structure of the methyl-CpG-binding domain of human MBD1 in complex with methylated DNA. *Cell*, *105*(4), 487–497.

Okamoto, A., Tainaka, K., & Kamei, T. (2006). Sequence-selective osmium oxidation of DNA: efficient distinction between 5-methylcytosine and cytosine. *Org. Biomol. Chem.*, *4*(9), 1638–1640.

Okamoto, Y., Yoshida, N., Suzuki, T., Shimozawa, N., Asami, M., Matsuda, T., ... Takada, T. (2016). DNA methylation dynamics in mouse preimplantation embryos revealed by mass spectrometry. *Sci. Rep.*, *6*(1), 19134–9.

Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M., & Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, *95*(19), 11163–11168.

Ortiz-Barahona, V., Joshi, R. S., & Esteller, M. (2020). Use of DNA methylation profiling in translational oncology. *Semin. Cancer Biol.*.

Otani, J., Arita, K., Kato, T., Kinoshita, M., Kimura, H., Suetake, I., ... Shirakawa, M. (2013). Structural basis of the versatile DNA recognition ability of the methyl-CpG-binding domain of methyl-CpG-binding domain protein 4. *J. Biol. Chem.*, *288*(9), 6351–6362.

Packer, M. S. & Liu, D. R. (2015). Methods for the directed evolution of proteins. *Nat. Rev. Genet.*, *16*(7), 379–394.

Padfield, D., O'Sullivan, H., & Pawar, S. (2021). *RTPC and nls.multstart: a new pipeline to fit thermal performance curves in r*. (Vol. 12, No. 6, pp. 1138-1143). Author.

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2016). *Biostrings: String objects representing biological sequences, and matching algorithms*. Bioconductor. Retrieved from https://bioconductor.org/packages/Biostrings

Palei, S., Buchmuller, B. C., Wolffgramm, J., Muñoz-López, Á., Jung, S., Czodrowski, P., & Summerer, D. (2020). Light-activatable TET-dioxygenases reveal dynamics of 5-methylcytosine oxidation and transcriptome reorganization. *J. Am. Chem. Soc.*, *142*(16), 7289–7294.

Papale, L. A., Madrid, A., Li, S., & Alisch, R. S. (2017). Early-life stress links 5-hydroxymethylcytosine to anxiety-related behaviors. *Epigenetics*, *12*(4), 264–276.

Park, M., Yoo, G., Bong, J. H., Jose, J., Kang, M. J., & Pyun, J. C. (2015). Isolation and characterization of the outer membrane of Escherichia coli with autodisplayed Z-domains. *Biochim. Biophys. Acta, Biomembr.*, *1848*(3), 842–847.

Parry, A., Rulands, S., & Reik, W. (2021). Active turnover of DNA methylation during cell fate decisions. *Nat. Rev. Genet.*, *22*(1), 59–66.

Patel, D. J. (2016). A structural perspective on readout of epigenetic histone and DNA methylation marks. *Cold Spring Harbor Perspect. Biol.*, *8*(3), a018754.

Patiño-Parrado, I., Gómez-Jiménez, Á., López-Sánchez, N., & Frade, J. M. (2017). Strand-specific CpG hemimethylation, a novel epigenetic modification functional for genomic imprinting. *Nucleic Acids Res.*, *45*(15), gkx518.

Penn, N. W., Suwalski, R., O'Riley, C., Bojanowski, K., & Yura, R. (1972). The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. *Biochem. J.*, *126*(4), 781–790.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., ... Ferrin, T. E. (2021). UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.*, *30*(1), 70–82.

Pfaffeneder, T., Hackner, B., Truss, M., Münzel, M., Müller, M., Deiml, C. A., ... Carell, T. (2011). The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem., Int. Ed.*, *50*(31), 7008–7012.

Ponnaluri, V. K. C., Ehrlich, K. C., Zhang, G., Lacey, M., Johnston, D., Pradhan, S., & Ehrlich, M. (2017). Association of 5-hydroxymethylation and 5-methylation of DNA cytosine with tissue-specific gene expression. *Epigenetics*, *12*(2), 123–138.

R Core Team (2021). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Raiber, E. A., Hardisty, R., van Delft, P., & Balasubramanian, S. (2017). Mapping and elucidating the function of modified bases in DNA. *Nat. Rev. Chem.*, *1*, 0069.

Raiber, E. A., Beraldi, D., Ficz, G., Burgess, H. E., Branco, M. R., Murat, P., ... Balasubramanian, S. (2012). Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol.*, *13*(8), R69.

Raiber, E. A., Murat, P., Chirgadze, D. Y., Beraldi, D., Luisi, B. F., & Balasubramanian, S. (2015). 5-Formylcytosine alters the structure of the DNA double helix. *Nat. Struct. Mol. Biol.*, *22*(1), 44–49.

Rajakumara, E., Law, J. A., Simanshu, D. K., Voigt, P., Johnson, L. M., Reinberg, D., ... Jacobsen, S. E. (2011). A dual flip-out mechanism for 5mC recognition by the Arabidopsis SUVH5 SRA domain and its impact on DNA methylation and H3K9 dimethylation in vivo. *Genes Dev.*, *25*(2), 137–152.

Ramsahoye, B. H. (2002). Nearest-neighbor analysis. In K. I. Mills & B. H. Ramsahoye (Eds.), *DNA Methylation Protocols*. (Vol. 200, pp. 9–15). Humana Press, New York, NY.

Ran, Z., Young Choi, B., Mee-Hyun, L., M. Bode, A., & Zigang, D. (2016). Implications of genetic and epigenetic alterations of CDKN2A (p16INK4a) in cancer. *EBioMedicine*, *8*, 30–39.

Rathi, P., Maurer, S., Kubik, G., & Summerer, D. (2016). Isolation of human genomic DNA sequences with expanded nucleobase selectivity. *J. Am. Chem. Soc.*, *138*(31), 9910–9918.

Rauch, C., Trieb, M., Wibowo, F. R., Wellenzohn, B., Mayer, E., & Liedl, K. R. (2005). Towards an understanding of DNA recognition by the methyl-CpG-binding domain 1. *J. Biomol. Struct. Dyn.*, *22*(6), 695–706.

Rauch, T. & Pfeifer, G. P. (2005). Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. *Lab. Invest.*, *85*(9), 1172–1180.

Rausch, C., Hastert, F. D., & Cardoso, M. C. (2019). DNA modification readers and writers and their interplay. *J. Mol. Biol.*, *432*(6), 1731–1746.

Razin, A. & Sedat, J. (1977). Analysis of 5-methylcytosine in DNA II. Gas chromatography. *Anal. Biochem.*, *77*(2), 370–377.

Rebar, E. & Pabo, C. (1994). Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science*, *263*(5147), 671–673.

Reetz, M. T., Kahakeaw, D., & Lohmer, R. (2008). Addressing the numbers problem in directed evolution. *ChemBioChem*, *9*(11), 1797–1804.

Ren, R., Horton, J. R., Zhang, X., Blumenthal, R. M., & Cheng, X. (2018). Detecting and interpreting DNA methylation marks. *Curr. Opin. Struct. Biol.*, *53*, 88–99.

Ribeiro, J., Ríos-Vera, C., Melo, F., & Schüller, A. (2019). Calculation of accurate interatomic contact surface areas for the quantitative analysis of non-bonded molecular interactions. *Bioinformatics*, *35*(18), 3499–3501.

Rice, J. C., Ozcelik, H., Maxeiner, P., Andrulis, I., & Futscher, B. W. (2000). Methylation of the BRCA1 promoter is associated with decreased BRCA1 mRNA levels in clinical breast cancer specimens. *Carcinogenesis*, *21*(9), 1761–1765.

Robertson, A. B., Dahl, J. A., Vågbø, C. B., Tripathi, P., Krokan, H. E., & Klungland, A. (2011). A novel method for the efficient and selective identification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic Acids Res.*, *39*(8), e55–e55.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., & Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.*, *242*(1), 84–89.

Rooman, M., Liévin, J., Buisine, E., & Wintjens, R. (2002). Cation-pi/H-bond stair motifs at protein–DNA interfaces. *J. Mol. Biol.*, *319*(1), 67–76.

Roy, P. H. & Weissbach, A. (1975). DNA methylase from HeLa cell nuclei. *Nucleic Acids Res.*, *2*(10), 1669–1684.

Rube, H. T., Rastogi, C., Kribelbauer, J. F., & Bussemaker, H. J. (2018). A unified approach for quantifying and interpreting DNA shape readout by transcription factors. *Mol. Syst. Biol.*, *14*(2), e7902.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, *230*(4732), 1350–1354.

Saito, M. & Ishikawa, F. (2002). The mCpG-binding Domain of Human MBD3 Does Not Bind to mCpG but Interacts with NuRD/Mi2 Components HDAC1 and MTA2. *J. Biol. Chem.*, *277*(38), 35434–35439.

Salema, V., Marín, E., Martínez-Arteaga, R., Ruano-Gallego, D., Fraile, S., Margolles, Y., ... Fernández, L. Á. (2013). Selection of Single Domain Antibodies from Immune Libraries Displayed on the Surface of E. coli Cells with Two β-Domains of Opposite Topologies. *PLoS ONE*, *8*(9), e75126.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, *74*(12), 5463–5467.

Saparbaev, M. & Laval, J. (1998). 3,N4-ethenocytosine, a highly mutagenic adduct, is a primary substrate for Escherichia coli double-stranded uracil-DNA glycosylase and human mismatch-specific thymine-DNA glycosylase. *Proc. Natl. Acad. Sci. U.S.A.*, *95*(15), 8508–8513.

Sasai, N., Nakao, M., & Defossez, P. A. (2010). Sequence-specific recognition of methylated DNA by human zinc-finger proteins. *Nucleic Acids Res.*, *38*(15), 5015–5022.

Saxonov, S., Berg, P., & Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U.S.A.*, *103*(5), 1412–1417.

Scarsdale, J. N., Webb, H. D., Ginder, G. D., & Williams Jr, D. C. (2011). Solution structure and dynamic analysis of chicken MBD2 methyl binding domain bound to a target-methylated DNA sequence. *Nucleic Acids Res.*, *39*(15), 6741–6752.

Schiesser, S., Hackner, B., Pfaffeneder, T., Müller, M., Hagemeier, C., Truss, M., & Carell, T. (2012). Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing. *Angew. Chem., Int. Ed.*, *51*(26), 6516–6520.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., ... Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, *9*(7), 676–682.

Schultheiss, E., Weiss, S., Winterer, E., Maas, R., Heinzle, E., & Jose, J. (2008). Esterase autodisplay: enzyme engineering and whole-cell activity determination in microplates with ph sensors. *Appl. Environ. Microbiol.*, *74*(15), 4782–4791.

Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., ... Ecker, J. R. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, *523*(7559), 212–216.

Schutsky, E. K., DeNizio, J. E., Hu, P., Liu, M. Y., Nabel, C. S., Fabyanic, E. B., ... Kohli, R. M. (2018). Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat. Biotechnol.*, *36*, 1083–1090.

Seeman, N. C., Rosenberg, J. M., & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. U.S.A.*, *73*(3), 804–808.

Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, A. C., Fung, H. L., ... Zhang, Y. (2013). Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell*, *153*(3), 692–706.

Shi, D. Q., Ali, I., Tang, J., & Yang, W. C. (2017). New insights into 5hmc DNA modification: generation, distribution and function. *Front. Genet.*, *8*, 100.

Singer, J., Roberts-Ems, J., Luthardt, F. W., & Riggs, A. D. (1979). Methylation of DNA in mouse early embryos, teratocarcinoma cells and adult tissues of mouse and rabbit. *Nucleic Acids Res.*, *7*(8), 2369–2385.

Sinsheimer, R. L. (1954). The action of pancreatic desoxyribonuclease. I. Isolation of mono- and dinucleotides. *J. Biol. Chem.*, *208*(1), 445–459.

— (1955). The action of pancreatic deoxyribonuclease. II. Isomeric dinucleotides. *J. Biol. Chem.*, *215*(2), 579–583.

Skene, P. J., Illingworth, R. S., Webb, S., Kerr, A. R. W., James, K. D., Turner, D. J., ... Bird, A. P. (2010). Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Mol. Cell*, *37*(4), 457–468.

Smith, G. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, *228*(4705), 1315–1317.

Smith, Z. D. & Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, *14*(3), 204–220.

Song, C. X., Diao, J., Brunger, A. T., & Quake, S. R. (2016). Simultaneous single-molecule epigenetic imaging of DNA methylation and hydroxymethylation. *Proc. Natl. Acad. Sci. U.S.A.*, *113*(16), 4338–4343.

Song, C. X., Szulwach, K. E., Dai, Q., Fu, Y., Mao, S. Q., Lin, L., ... He, C. (2013). Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*, *153*(3), 678–691.

Song, C. X., Szulwach, K. E., Fu, Y., Dai, Q., Yi, C., Li, X., ... He, C. (2011). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.*, *29*(1), 68–72.

Song, J., Teplova, M., Ishibe-Murakami, S., & Patel, D. J. (2012). Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. *Science*, *335*(6069), 709–712.

Sood, A. J., Viner, C., & Hoffman, M. M. (2019). DNAmod: the DNA modification database. *J. Cheminform.*, *11*(1), 30.

Sperlazza, M. J., Bilinovich, S. M., Sinanan, L. M., Javier, F. R., & Williams Jr, D. C. (2017). Structural basis of MeCP2 distribution on non-CpG methylated and hydroxymethylated DNA. *J. Mol. Biol.*, *429*(10), 1581–1594.

Springer, N. M. & Kaeppler, S. M. (2005). Evolutionary divergence of monocot and dicot methyl-CpG-binding domain proteins. *Plant Physiol.*, *138*(1), 92–104.

Spruijt, C. G., Gnerlich, F., Smits, A. H., Pfaffeneder, T., Jansen, P. W. T. C., Bauer, C., ... Vermeulen, M. (2013). Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell*, *152*(5), 1146–1159.

Stadler, M. B., Murr, R., Burger, L., Ivánek, R., Lienert, F., Schöler, A., ... Schübeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, *480*(7378), 490–495.

Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S., & Jacobsen, S. E. (2011). 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.*, *12*(6), R54–8.

Sukackaite, R., Grazulis, S., Tamulaitis, G., & Siksnys, V. (2012). The recognition domain of the methyl-specific endonuclease McrBC flips out 5-methylcytosine. *Nucleic Acids Res.*, *40*(15), 7552–7562.

Susan, J. C., Harrison, J., Paul, C. L., & Frommer, M. (1994). High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.*, *22*(15), 2990–2997.

Suzuki, M., Brenner, S. E., Gerstein, M., & Yagi, N. (1995). DNA recognition code of transcription factors. *Protein Eng.*, *8*(4), 319–328.

Suzuki, M. & Gerstein, M. (1995). Binding geometry of alpha-helices that recognize DNA. *Proteins*, *23*(4), 525–535.

Suzuki, M. M. & Bird, A. P. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, *9*(6), 465–476.

Szulik, M. W., Pallan, P. S., Nocek, B., Voehler, M., Banerjee, S., Brooks, S., ... Stone, M. P. (2015). Differential stabilities and sequence-dependent base pair opening dynamics of Watson-Crick base pairs with 5-hydroxymethylcytosine, 5-formylcytosine, or 5-carboxylcytosine. *Biochemistry*, *54*(5), 1294–1305.

Szulwach, K. E., Li, X., Li, Y., Song, C. X., Wu, H., Dai, Q., ... Jin, P. (2011). 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat. Neurosci.*, *14*(12), 1607–1616.

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., ... Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, *324*(5929), 930–935.

Tam, B. E., Sung, K., & Sikes, H. D. (2016). Engineering affinity agents for the detection of hemi-methylated CpG sites in DNA. *Mol. Syst. Des. Eng.*, *1*(3), 273–277.

Tamanaha, E., Guan, S., Marks, K., & Saleh, L. (2016). Distributive processing by the iron(II)/α-ketoglutarate-dependent catalytic domains of the TET enzymes is consistent with epigenetic roles for oxidized 5-methylcytosine bases. *J. Am. Chem. Soc.*, *138*(30), 9345–9348.

Tanaka, K., Tainaka, K., Kamei, T., & Okamoto, A. (2007). Direct Labeling of 5-Methylcytosine and Its Applications. *J. Am. Chem. Soc.*, *129*(17), 5612–5620.

Tarhonskaya, H., Rydzik, A. M., Leung, I. K. H., Loik, N. D., Chan, M. C., Kawamura, A., ... Schofield, C. J. (2019). Non-enzymatic chemistry enables 2-hydroxyglutarate-mediated activation of 2-oxoglutarate oxygenases. *Nat. Commun.*, 1–10.

Tatematsu, K. I., Yamazaki, T., & Ishikawa, F. (2000). MBD2-MBD3 complex binds to hemi-methylated DNA and forms a complex containing DNMT1 at the replication foci in late S phase. *Genes Cells*, *5*(8), 677–688.

Tateno, M., Yamasaki, K., Amano, N., Kakinuma, J., Koike, H., Allen, M. D., & Suzuki, M. (1997). DNA recognition by β-sheets. *Biopolymers*, *44*(4), 335–359.

Tillotson, R., Cholewa-Waclaw, J., Chhatbar, K., Connelly, J. C., Kirschner, S. A., Webb, S., ... Bird, A. (2021). Neuronal non-CG methylation is an essential target for MeCP2 function. *Mol. Cell*, *81*(6), 1260–1275.e12.

Tippin, D. B., Ramakrishnan, B., & Sundaralingam, M. (1997). Methylation of the Z-DNA decamer d(GC)5 potentiates the formation of A-DNA: crystal structure of d(Gm5CGm5CGCGCGC). *J. Mol. Biol.*, *270*(2), 247–258.

Tippin, D. B. & Sundaralingam, M. (1997). Nine polymorphic crystal structures of d(CCGGGCCCGG), d(CCGGGCCm5CGG), d(Cm5CGGGCCm5CGG) and d(CCGGGCC(Br)5CGG) in three different conformations: effects of spermine binding and methylation on the bending and condensation of A-DNA. *J. Mol. Biol.*, *267*(5), 1171–1185.

Tizei, P. A. G., Csibra, E., Torres, L., & Pinheiro, V. B. (2016). Selection platforms for directed evolution in synthetic biology. *Biochem. Soc. Trans.*, *44*(4), 1165–1175.

Turner, P., Holst, O., & Karlsson, E. N. (2005). Optimized expression of soluble cyclomaltodextrinase of thermophilic origin in Escherichia coli by using a soluble fusion-tag and by tuning of inducer concentration. *Protein Expression Purif.*, *39*(1), 54–60.

Uversky, V. N. (2003). Protein folding revisited. A polypeptide chain at the folding – misfolding – nonfolding cross-roads: which way to go?. *Cell. Mol. Life Sci.*, *60*(9), 1852–1871.

Valinluck, V. & Sowers, L. C. (2007). Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1. *Cancer Res.*, *67*(3), 946–950.

Valinluck, V., Tsai, H. H., Rogstad, D. K., Burdzy, A., Bird, A. P., & Sowers, L. C. (2004). Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG-binding domain (MBD) of methyl-CpG-binding protein 2 (MeCP2). *Nucleic Acids Res.*, *32*(14), 4100–4108.

Van, P., Jiang, W., Gottardo, R., & Finak, G. (2018). GgCyto: next generation open-source visualization software for cytometry. *Bioinformatics*, *34*(22), 3951–3953.

VanAntwerp, J. J. & Wittrup, K. D. (2000). Fine affinity discrimination by yeast surface display and flow cytometry. *Biotechnol. Prog.*, *16*(1), 31–37.

Vanyushin, B. F., Mazin, A. L., Vasilyev, V. K., & Belozersky, A. N. (1973). The content of 5-methylcytosine in animal DNA: the species and tissue specificity. *Biochim. Biophys. Acta*, *299*(3), 397–403.

Villar-Menéndez, I., Blanch, M., Tyebji, S., Pereira-Veiga, T., Albasanz, J. L., Martín, M., ... Barrachina, M. (2013). Increased 5-methylcytosine and decreased 5-hydroxymethylcytosine levels are associated with reduced striatal A2AR levels in Huntington's disease. *Neuromolecular Med.*, *15*(2), 295–309.

Vu, T. H., Li, T., Nguyen, D., Nguyen, B. T., Yao, X. M., Hu, J. F., & Hoffman, A. R. (2000). Symmetric and asymmetric DNA methylation in the human IGF2-H19 imprinted region. *Genomics*, *64*(2), 132–143.

Wade, P. A., Gegonne, A., Jones, P. L., Ballestar, E., Aubry, F., & Wolffe, A. P. (1999). Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nat. Genet.*, *23*(1), 62–66.

Wakefield, R. I., Smith, B. O., Nan, X., Free, A., Soteriou, A., Uhrin, D., ... Barlow, P. N. (1999). The solution structure of the domain from MeCP2 that binds to methylated DNA. *J. Mol. Biol.*, *291*(5), 1055–1065.

Walavalkar, N. M., Cramer, J. M., Buchwald, W. A., Scarsdale, J. N., & Williams, D. C. (2014). Solution structure and intramolecular exchange of methyl-cytosine binding domain protein 4 (MBD4) on DNA suggests a mechanism to scan for mCpG/TpG mismatches. *Nucleic Acids Res.*, *42*(17), 11218–11232.

Wan, J., Su, Y., Song, Q., Tung, B., Oyinlade, O., Liu, S., ... Xia, S. (2017). Methylated cis-regulatory elements mediate KLF4-dependent gene transactivation and cell migration. *eLife*, *6*, 25.

Wang, D., Hashimoto, H., Zhang, X., Barwick, B. G., Lonial, S., Boise, L. H., ... Cheng, X. (2017). MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma. *Nucleic Acids Res.*, *45*(5), 2396–2407.

Wang, L., Zhang, J., Duan, J., Gao, X., Zhu, W., Lu, X., ... Liu, J. (2014). Programming and inheritance of parental DNA methylomes in mammals. *Cell*, *157*(4), 979–991.

Watson, J. D. & Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737–738.

Wells, J. A., Vasser, M., & Powers, D. B. (1985). Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites. *Gene*, *34*(2-3), 315–323.

Wen, L. & Tang, F. (2014). Genomic distribution and possible functions of DNA hydroxymethylation in the brain. *Genomics*, *104*(5), 341–346.

Wentzel, A., Christmann, A., Adams, T., & Kolmar, H. (2001). Display of Passenger Proteins on the Surface of Escherichia coli K-12 by the Enterohemorrhagic E. coli Intimin EaeA. *J. Bacteriol.*, *183*(24), 7273–7284.

Wescoe, Z. L., Schreiber, J., & Akeson, M. (2014). Nanopores discriminate among five C5-cytosine variants in DNA. *J. Am. Chem. Soc.*, *136*(47), 16582–16587.

Whitmore, L., Miles, A. J., Mavridis, L., Janes, R. W., & Wallace, B. A. (2017). PCDDB: new developments at the Protein Circular Dichroism Data Bank. *Nucleic Acids Res.*, *45*(D1), D303–D307.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag.

Wilding, M., Hong, N., Spence, M., Buckle, A. M., & Jackson, C. J. (2019). Protein engineering: the potential of remote mutations. *Biochem. Soc. Trans.*, *47*(2), 701–711.

Williams, K., Christensen, J., Pedersen, M. T., Johansen, J. V., Cloos, P. A. C., Rappsilber, J., & Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, *473*(7347), 343–348.

Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J., & Campbell, K. H. S. (1997). Viable offspring derived from fetal and adult mammalian cells. *Nature*, *385*(6619), 810–813.

Wilson, D. S. & Keefe, A. D. (2000). Random mutagenesis by PCR. *Curr. Protoc. Mol. Biol.*, *51*(1), 8.3.1-8.3.9.

Wolffgramm, J., Buchmuller, B. C., Palei, S., Muñoz-López, Á., Kanne, J., Janning, P., ... Summerer, D. (2021). Light-activation of DNA-methyltransferases. *Angew. Chem., Int. Ed.*, *60*(24), 13507–13512.

Woodcock, D. M., Lawler, C. B., Linsenmeyer, M. E., Doherty, J. P., & Warren, W. D. (1997). Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. *J. Biol. Chem.*, *272*(12), 7810–7816.

Wu, G., Zhang, X., & Gao, F. (2020). The epigenetic landscape of exercise in cardiac health and disease. *J. Sport Health Sci.*.

Wu, H., Wu, X., Shen, L., & Zhang, Y. (2014). Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.*, *32*(12), 1231–1240.

Wu, H. & Zhang, Y. (2011a). Tet1 and 5-hydroxymethylation: a genome-wide view in mouse embryonic stem cells. *Cell Cycle*, *10*(15), 2428–2436.

— (2011b). Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes Dev.*, *25*(23), 2436–2452.

— (2014). Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell*, *156*(1-2), 45–68.

Wu, X. & Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat. Rev. Genet.*, *18*(9), 517–534.

Wyatt, G. R. (1951). The purine and pyrimidine composition of deoxypentose nucleic acids. *Biochem. J.*, *48*(5), 584–590.

Wyatt, G. R. & Cohen, S. S. (1952). A new pyrimidine base from bacteriophage nucleic acids. *Nature*, *170*(4338), 1072–1073.

Wyman, J. (1948). Heme proteins. *Adv. Protein Chem.*, *4*, 407–531.

— (1964). Linked functions and reciprocal effects in hemoglobin: a second look. *Adv. Protein Chem.*, *19*, 223–286.

Xia, B., Han, D., Lu, X., Sun, Z., Zhou, A., Yin, Q., ... Yi, C. (2015). Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat. Methods*, *12*(11), 1047–1050.

Xiong, J., Zhang, Z., Chen, J., Huang, H., Xu, Y., Ding, X., ... Zhu, B. (2016). Cooperative action between SALL4A and TET proteins in stepwise oxidation of 5-methylcytosine. *Mol. Cell*, *64*(5), 913–925.

Xu, C. & Corces, V. G. (2018). Resolution of the DNA methylation state of single CpG dyads using in silico strand annealing and WGBS data. *Nat. Protocols*, *14*(1), 202–216.

Xu, J., McPartlon, M., & Li, J. (2021). Improved protein structure prediction by deep learning irrespective of coevolution information. *Nat. Mach. Intell.*, *3*(7), 601–609.

Xu, L., Chen, Y. C., Chong, J., Fin, A., McCoy, L. S., Xu, J., ... Wang, D. (2014). Pyrene-based quantitative detection of the 5-formylcytosine loci symmetry in the CpG duplex content during TET-dependent demethylation. *Angew. Chem., Int. Ed.*, *53*(42), 11223–11227.

Xu, T. & Gao, H. (2020). Hydroxymethylation and tumors: can 5- hydroxymethylation be used as a marker for tumor diagnosis and treatment?. *Hum. Genomics*, *14*(15), 1–10.

Yang, J., Horton, J. R., Li, J., Huang, Y., Zhang, X., Blumenthal, R. M., & Cheng, X. (2019). Structural basis for preferential binding of human TCF4 to DNA containing 5-carboxylcytosine. *Nucleic Acids Res.*, *47*(16), 8375–8387.

Yang, J., Horton, J. R., Wang, D., Ren, R., Li, J., Sun, D., ... Cheng, X. (2018). Structural basis for effects of CpA modifications on C/EBPβ binding of DNA. *Nucleic Acids Res.*, *47*(4), gky1264–.

Yang, Y., Kucukkal, T. G., Li, J., Alexov, E., & Cao, W. (2016). Binding analysis of methyl-CpG-binding domain of MeCP2 and Rett syndrome mutations. *ACS Chem. Biol.*, *11*(10), 2706–2715.

Yao, B., Lin, L., Street, R. C., Zalewski, Z. A., Galloway, J. N., Wu, H., ... Jin, P. (2014). Genome-wide alteration of 5-hydroxymethylcytosine in a mouse model of fragile X-associated tremor/ataxia syndrome. *Hum. Mol. Genet.*, *23*(4), 1095–1107.

Yi, L., Gebhard, M. C., Li, Q., Taft, J. M., Georgiou, G., & Iverson, B. L. (2013). Engineering of TEV protease variants by yeast ER sequestration screening (YESS) of combinatorial libraries. *Proc. Natl. Acad. Sci. U.S.A.*, *110*(18), 7229–7234.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., ... Taipale, J. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, *356*(6337).

Yizhak, K., Aguet, F., Kim, J., Hess, J. M., Kübler, K., Grimsby, J., ... Getz, G. (2019). RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*, *364*(6444), eaaw0726.

Yoo, J., Kim, H., Aksimentiev, A., & Ha, T. (2016). Direct evidence for sequence-dependent attraction between double-stranded DNA controlled by methylation. *Nat. Commun.*, *7*(1), 11045–7.

Yu, M., Hon, G. C., Szulwach, K. E., Song, C. X., Zhang, L., Kim, A., ... He, C. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, *149*(6), 1368–1380.

Yu, Y., Blair, S., Gillespie, D., Jensen, R., Myszka, D., Badran, A. H., ... Chagovetz, A. (2010). Direct DNA methylation profiling using methyl binding domain proteins.. *Anal. Chem.*, *82*(12), 5012–5019.

Zeng, T. B., Han, L., Pierce, N., Pfeifer, G. P., & Szabó, P. E. (2019). EHMT2 and SETDB1 protect the maternal pronucleus from 5mC oxidation. *Proc. Natl. Acad. Sci. U.S.A.*, *74*, 201819946.

Zhang, S. C., Wang, M. Y., Feng, J. R., Chang, Y., Ji, S. R., & Wu, Y. (2020). Reversible promoter methylation determines fluctuating expression of acute phase proteins. *eLife*, *9*, e51317.

Zhang, X., Su, J., Jeong, M., Ko, M., Huang, Y., Park, H. J., ... Goodell, M. A. (2016). DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. *Nat. Genet.*, *48*(9), 1014–1023.

Zhang, Y., Liu, L., Guo, S., Song, J., Zhu, C., Yue, Z., ... Yi, C. (2017). Deciphering TAL effectors for 5-methylcytosine and 5-hydroxymethylcytosine recognition. *Nat. Commun.*, *8*(1), 901.

Zhao, L. Y., Song, J., Liu, Y., Song, C. X., & Yi, C. (2020). Mapping the epigenetic modifications of DNA and RNA. *Protein Cell*, 1–17.

Zhou, T., Xiong, J., Wang, M., Yang, N., Wong, J., Zhu, B., & Xu, R. M. (2014). Structural basis for hydroxymethylcytosine recognition by the SRA domain of UHRF2. *Mol. Cell*, *54*(5), 879–886.

Zhu, C., Gao, Y., Guo, H., Xia, B., Song, J., Wu, X., ... Yi, C. (2017). Single-cell 5-formylcytosine landscapes of mammalian early embryos and ESCs at single-base resolution. *Cell Stem Cell*, *20*(5), 720–731.e5.

Zhu, H., Wang, G., & Qian, J. (2016). Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.*, *17*(9), 551–565.

Zhu, J. K. (2009). Active DNA demethylation mediated by DNA glycosylases. *Annu. Rev. Genet.*, *43*(1), 143–166.

Zhu, Q., Stöger, R., & Alberio, R. (2018). A Lexicon of DNA Modifications: their Roles in Embryo Development and the Germline. *Front. Cell. Dev. Biol.*, *6*, 24.

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T. Y., Kohlbacher, O., ... Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*, *500*(7463), 477–481.

Zou, X., Ma, W., Solov'yov, I. A., Chipot, C., & Schulten, K. (2012). Recognition of methylated DNA through methyl-CpG-binding domain proteins. *Nucleic Acids Res.*, *40*(6), 2747–2758.

Zubay, G. & Doty, P. (1959). The isolation and properties of deoxyribonucleoprotein particles containing single nucleic acid molecules. *J. Mol. Biol.*, *1*(1), 1–IN1.

This document was typeset in 11.0 pt
(17.6 pt) TeX Gyre Pagella by
ConTeXt using LuaMetaTeX.

Printed on acid-free paper by
WIRmachenDRUCK GmbH.

WirmachenDruck.de
Sie sparen, wir drucken!

All carbon emissions created by
producing and printing this thesis
were compensated through a
recognized carbon offset project.

Klimaneutral
Druckprodukt
ClimatePartner.com/12518-1907-1001