Understanding the molecular basis of base-specific interaction of Type IV restriction system McrBC

Thesis submitted in partial fulfilment of the requirements of Five Year BS-MS Dual Degree Program at

Indian Institute of Science Education and Research, Pune

Gyana Gourab Mishra

20121101

Biology Division, IISER Pune

Under the supervision of

Dr. Saikrishnan Kayarat

Biology Division, IISER Pune

CERTIFICATE

This is to certify that this dissertation entitled "*Understanding the molecular basis of base specific interaction of Type IV restriction system McrBC*" towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Gyana Gourab Mishra at IISER Pune under the supervision of Dr. Saikrishnan Kayarat, Associate Professor, Biology Division, IISER Pune during the academic year 2016-2017.

Dr. Saikrishnan Kayarat
Biology Division, IISER Pune

Date: 30/03/2017

DECLARATION

I hereby declare that the matter embodied in the report entitled "*Understanding the molecular basis of base specific interaction of Type IV restriction system McrBC*" are the results of the work carried out by me at the Department of Biology, IISER Pune, under the supervision of Dr. Saikrishnan Kayarat and the same has not been submitted elsewhere for any other degree.

Gyana Gourab Mishra

Date: 30/03/2017

Abstract: -

Base specific DNA protein interaction is fundamental to many of the cellular processes. DNA binding proteins are known to employ direct base read out as well as indirect shape read out mechanisms to recognise their target sites. Restriction modification systems present in prokaryotic organisms are one of the excellent models to study base-specific interaction because of the specificity towards their target sites. McrBC which is a modification dependent restriction endonuclease binds to DNA that contains two R^mC (methylated cytosine preceded by a purine) sites and cleaves close to one of the sites. In the crystal structure of N-terminal DNA binding domain of McrB Leucine 68 is the only residue that interacts with the methyl group of the cytosine through hydrophobic interaction. In order to find the significance of this interaction the side chain of this Leucine was extended by mutating to Phenylalanine and Tryptophan. While phenylalanine mutant recognised both methylated and nonmethylated substrates, tryptophan mutation made the protein recognise neither methylated nor non-methylated substrates. These findings provide interesting insights about how alteration in a weak hydrophobic interaction can lead to change in specificity of a protein. It will also have a wide variety of application in the field of epigenetics, genome sequencing etc.

List of Figures: -

Acknowledgement: -

I am extremely grateful to my guide Dr Saikrishnan Kayarat who helped and guided me at every step during this project. His efforts to develop rational thinking in me and maintain a friendly environment in the lab are unvaluable. I am really thankful to him for the amount of time he has invested in me not only during this 5^{th} year but from the beginning itself.

I am also thankful to Dr Gayathri Pananghat for her useful inputs in every step of my project.

I am grateful to work with very good members in my lab. I specially thank Neha Nirwan who taught me all the methods and techniques during the whole journey of this project. Her friendly and caring behaviour always motivated me to do the work in lab without any problems. I would like to extend my thank you note to the lab members of both Sai and Gayathri's lab for their continuous support.

Also, I thank all my friends out here in IISER who have made my stay unforgettable within these 5 years. Finally, I would like to thank my mom without whom I would have never reached anywhere. She has been my emotional support throughout my career.

1 Introduction

Base-specific interaction of proteins with DNA is fundamental to many of the crucial cellular processes such as replication, transcription, signal transduction, etc. It has been reported that 2-3% of the entire prokaryotic genome and 6-7% of the eukaryotic genome encodes for DNA binding proteins [1]. However, how these proteins recognise DNA sequences in a base-specific manner has remained a long-standing puzzle in the field of molecular biology. After the very first high-resolution structure of DNA, people have tried to use crystallographic [2] and computational approaches [3] to come up with a universal theory that can be used to predict the specificity of any protein for its target site. But progress in this field has met with limited success till now. However, with the exponential increase in the crystal structures of protein bound to DNA, it has become clear that interaction of the protein with DNA is dependent on various factors and to build a unifying theory all those factors are required to be considered.

1.1 Mode of DNA interaction by DNA binding proteins

Broadly, factors influencing the specificity of a protein towards any DNA sequence can be categorised into two types of interactions. One is direct base read out which is due to the interaction of amino acids with DNA bases (such as hydrogen bond, stacking interaction, etc.) [4]. This is dependent on the unique chemical identity of each DNA base. The second factor governing DNA-protein interaction is indirect shape read out which is because of the unique shape adopted by DNA depending on the sequence and three-dimensional architecture of protein in the binding region [5]. Using the data set from protein data bank, the propensity of an amino acid to interact with a nucleic acid base has been studied using various computational approaches [6,7]. However, these preferences are not strictly followed. Many proteins which are known to have very high sequence similarity interact with diverse target sequences, while many diverse proteins adopt a similar strategy to recognise a DNA sequence [8]. Hence, understanding base-specific DNA interaction and making rational alterations in a DNA binding protein to change its specificity to another target site holds fundamental importance in the field of molecular biology.

1.2 Bacterial Restriction-Modification systems as a model system

In this project, I have used the bacterial restriction-modification system as a model system to study base-specific DNA-protein interaction. Restriction-modification (RM) systems are one of the defence mechanisms employed by bacteria to protect themselves from invading bacteriophages. As the name suggest, RM system comprises of two components- Restriction and Modification [9]. The restriction component comprises of an endonuclease which binds and cleaves foreign DNA by recognising a specific sequence while the modification component is generally a methyltransferase which binds to the specific sequence in host DNA and methylates it in order to protect self DNA from the restriction activity of the endonuclease component. Based on the mode of DNA Cleavage, cofactor requirement, and a number of subunits/domains, RM systems have been classified into 4 categories [10]i.e. Type I, II, III and IV. Unlike the other RM systems, type IV restriction enzymes recognise and cleave methylated DNA substrate, and hence also lack the methyltransferase component. This system is believed to have evolved against bacteriophages which methylate their own DNA in order to escape the RM system in bacteria [11,12]. The first identified protein of this system was McrBC [13] which was isolated from E coli K12 strain [14]. The other type IV restriction enzymes that have been identified so far include Mrr and McrA [12].

1.3 McrBC: a modification-dependent type IV restriction system

McrBC is composed of two protein subunits McrB and McrC. McrB is a 54 kDa protein which has DNA binding domain at its N-terminal (1-161 residues) and GTPase domain belonging to AAA+ family at its C-terminal [15,16]. McrC is a 40 kDa protein which contains the nuclease domain belonging to the PD-(D/E)xK family of endonucleases (Pingoud*, 2002). In the presence of GTP, McrB forms a heptameric ring, and two such rings come together in the presence of 2 McrC monomers to form the functional McrBC tetradecameric complex [18]. McrBC binds specifically to methylated cytosine preceded by a purine base (5'-R^{m}C-3'). Cleavage occurs when there are at least two recognition sites about 30 bp to 2000 bp apart [19,20].

1.4 Mode of DNA recognition by McrBC: -

Recognition of modified base makes this protein a useful tool to study epigenetic modifications [21] and hence requires a mechanistic understanding into how it has gained specificity for modified base. The mechanism by which it recognises methylated cytosine is through base flipping where it flips the modified base by 180° into its catalytic pocket [22]. This is the same mechanism that is employed by all methyltransferases (both adenine and cytosine MTases) as well as proteins with SRA – domain (EGFR1, SUVH5, etc.) in order to recognise the specific base [23–26]This is an example where proteins of diverse family employ a similar strategy to bind to DNA.

1.5 Crystal structure of McrB N-terminal domain bound to methylated substrate

The crystal structure of N-terminal DNA binding domain of McrB bound to DNA was reported in 2013 [27]. Interaction of DNA bases and backbone phosphates with the amino acids in this crystal structure is shown in figure 1. The flipped-out cytosine is stabilised in the catalytic pocket by three types of interaction. First is the direct hydrogen bond contact made by Ile82, Asp84 and Thr85 with the Watson-crick edge of cytosine, the second is the stacking interaction by Tyr64, Tyr117 and the third being the hydrophobic interaction of Leu68 with the methyl group (figure 1). While the size of the pocket allows only pyrimidine bases to flip, direct read out by Ile, Asp and Thr helps the protein to discriminate cytosine against thymine. The hydrophobic interaction by Leu68 seems to play the key role in making the protein discriminate between a methylated substrate from a non-methylated one.

However, such hydrophobic interaction is not found in other families of protein that recognise methylated cytosines like MTases and SRA-domain proteins. Structures of C5 MTases [28,29]show that a conserved Gly-Pro-Pro-Cys motif surrounds the flipped cytosine from its N4 position to C6 position with Gly making a hydrogen bond with N4 and Cys making covalent interaction with C6 of cytosine. Hence, when a methyl group is added to the C5 position, it stays in the loop created by the two proline residues present in between Gly and Cys. Similarly, in SUVH5 the methyl group is

believed to be stabilised by the cα and cβ of a Glutamine residue [26].

Hence the hydrophobic interaction shown in the case of McrB is first of its kind reported which contributes specifically to discriminate between a methylated and non-methylated base.

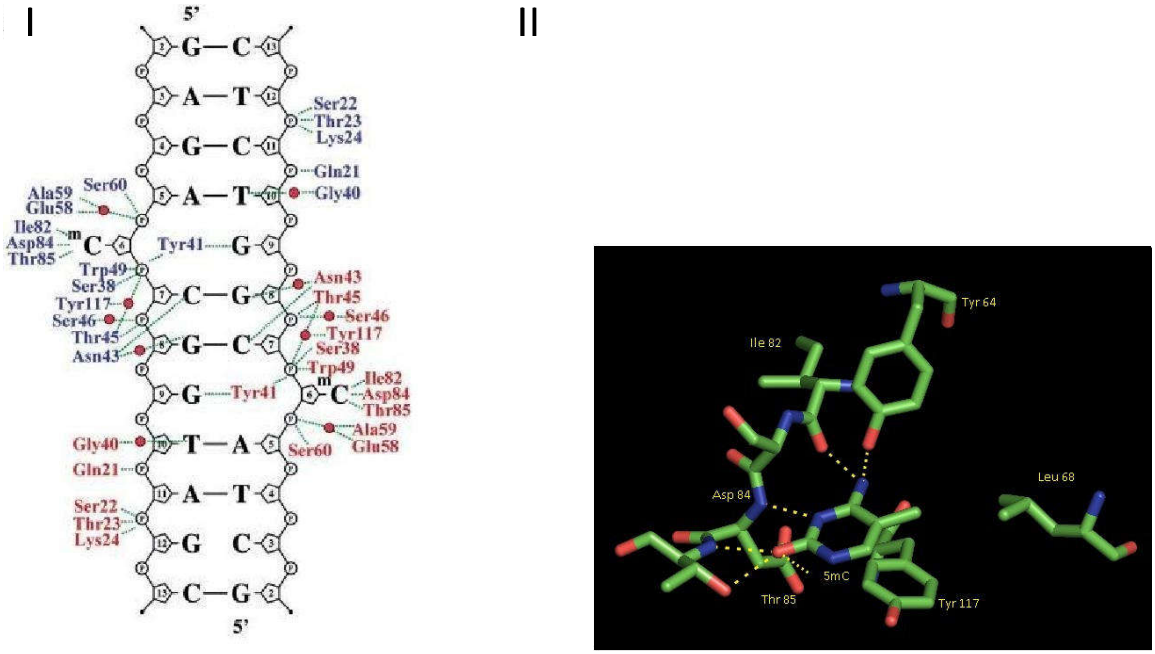


Figure 1: I) Interaction of different amino acids of McrB-N-terminal domain with DNA bases and backbone phosphates (adapted from Sukackaite et al., 2012) II) Interaction of amino acids with flipped cytosine of fig 1(I) is shown where Ile, Asp and Thr are making hydrogen bond contacts shown in yellow dotted lines. Hydrophobicinteraction by Leu and stacking by both Tyr molecules are shown.

As Leucine 68 residue seems to be the only amino acid contributing to the methylation specificity of the protein it was interesting to find if increasing the length of the Leucine side chain to a longer nonpolar residue could make the protein

recognise cytosine and prevent recognition of methylated cytosine. Phenylalanine and Tryptophan seemed two suitable candidates. When modelled in PyMol, conversion of Leucine to Phenylalanine brought the side chain closer to the methyl group while replacement with Tryptophan resulted in a clash (figure B). The idea was to seewhether a protein which has such high specificity towards methylated sites could bechanged into a protein which can now recognise only non-methylated sites. Thiswould not only provide better insights into the way these class of proteins employtheir specificity towards any substrate but also will have a wide variety of applicationin the field of epigenetics, molecular biology, etc.

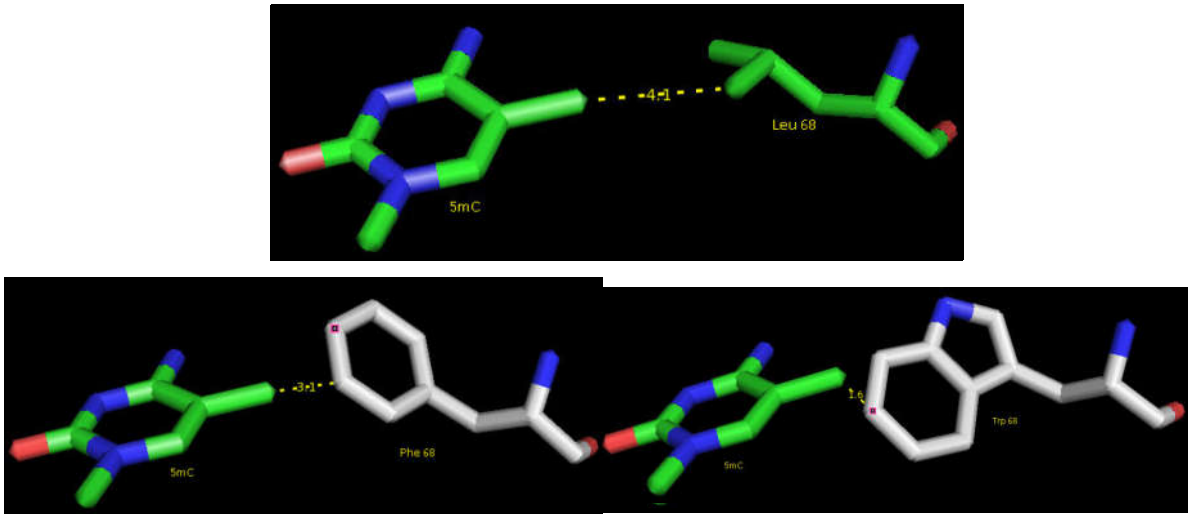


Figure 2: Interaction of mutated residues with methyl group of cytosine PyMol generated structure of McrBL68F and McrBL68W. For comparative analysis, wild-type interaction is shown above. Phenylalanine (Phe) and Tryptophan (Trp) are coloured differently from wild-type Leucine (Leu) residue. 5-methylcytosine is labelled as 5mC.

In the structure of wild type McrB-N-terminal domain with DNA, the shortest distance between Leu68 and the methyl group is 4.1 A° which is ideal to establish hydrophobic interaction. As per the structure generated using PyMol, mutating Leu68 residue to Phenylalanine reduces this distance to 3.1 A whereas mutation with Tryptophan makes it 1.6 A°. Hence Phenylalanine mutant (McrBL68F) seems to provide a moderate steric hindrance to the presence of the methyl group whereas Tryptophan mutant (McrBL68W) would sterically clash with it. These two mutants hence provide a suitable platform to study how tolerable these mutations are for the

protein to functionally recognise a DNA substrate whether it would be a methylated or non-methylated one.

Aims of the project: -

1) Generating and purifying site-directed mutants McrBL68F and McrBL68W

2) Check their GTPase activity to confirm that they are purified in active form

3) Investigating the specificity of the mutants using binding and cleavage assays with methylated as well as nonmethylated substrates

4) Comparing the cleavage pattern of these mutants and the wild type enzyme

2 Materials and Methods: -

2.1 Cloning of mbL68F and mbL68W mutant: -

Using forward primers 5' - CCCGTTATTTTCTATTATAAAGATTTTGATGAG – 3' for mcrbL68F & 5' - CCCGTTATTTGGTATTATAAAGATTTTGATGAG – 3' for mcrbL68W and reverse primer 5' – GATGATGGGATCCCGATGAGTCCCC – 3', both the mutants were PCR amplified using wildtype mcrB gene present in the pHIS17 vector as a template. Both the amplified products were cloned into the pHIS17 vector using restriction-free cloning method. Resulting mcrBL68F-HIS and mcrBL68W-HIS were fully sequenced.

2.2 Purification of McrBL68F and McrBL68W: -

Both the vectors mcrBL68F-pHIS and mcrBL68W-pHIS containing six histidine tag at c-terminus of the gene were overexpressed in E coli BL21 (AI) cells. Cultures were grown in 2 L LB media containing 100 µg/ml of ampicillin in an incubator-shaker at 37 °C until the OD at 600 nm reached 0.6. Both the cultures were then induced with 0.06 % w/v of L-arabinose. After induction, McrBL68F culture was grown at 37 °C for 3 hours whereas McrBL68W culture was grown at 18 °C for 16 hours. Cells were pelleted at 4^0 C and 3315g for 20 minutes. Further the pellets were resuspended in lysis buffer (500 mM NaCl, 50 mM Tris-Cl pH 8, 25 mM imidazole, 5 mM MgCl2, 10% glycerol). 0.04 % of CHAPS was added to the pellet resuspension. Cells were then lysed by sonication at 4^0C for 3 minutes with 1 sec on, three seconds off cycle. This process is repeated one more time after ten mins. The cell lysate was then ultracentrifuged at 4^0C and 159,200g for 40 mins.

Both the mutants were purified by affinity column chromatography followed by anion exchange chromatography using identical strategy. After ultracentrifugation supernatant of the cell, the lysate was loaded onto a 5 ml Ni-NTA column (GE Life Sciences) equilibrated with Buffer A (500 mM NaCl, 50 mM Tris-Cl pH 8, 25 mM imidazole). After complete loading protein was eluted using Buffer A and Buffer B (500 mM NaCl, 50 mM Tris-Cl pH 8, 500

mM imidazole) by a step gradient of 5%, 30%, 50% and 100% Buffer B. 5 mM of DTT is added to all the eluted fractions. The purest fractions were dialysed against 1L of dialysis buffer (50 mM NaCl, 50 mM Tris-Cl pH8, 1 mM EDTA, and 1 mM DTT).

Dialysed McrBL68F and McrBL68W were then centrifuged at 4 °C for 20 mins, and the supernatant is filtered with 0.2 mm filter. It is then loaded onto an 8 ml MonoQ 10/100 GL column (GE Life Sciences) equilibrated with Buffer B50 (50 mM NaCl, 50 mM Tris-Cl pH 8, 1 mM EDTA, 1 mM DTT). Fractions of 1.5ml were collected in 20 column volumes over a linear gradient of 0 to 50% Buffer B1000 (1000 mM NaCl, 50 mM Tris-Cl pH 8, 1 mM EDTA, 1 mM DTT). The pure fractions were then pooled and concentrated using a 2 ml 10 kDa vivaspin2 concentrator (GE Life Sciences). Both the mutants were then washed with storage buffer to remove EDTA (100 mM NaCl, 10 mM Tris-Cl 7.4 and 1 mM DTT) and stored in storage buffer at -80 °C.

McrC used in the experiment was already available in the lab, which was purified similarly. However, MonoS 10/100 GL column (GE Life Sciences) was used instead of MonoQ 10/100 GL column (GE Life Sciences) after dialysis. It was also washed and stored in the same storage buffer at -80 °C.

2.3 GTPase assay: -

To check the relative GTPase activities of the two mutants McrBL68F and McrBL68W as compared to that of wild-type McrB protein malachite green phosphatase assay was performed. The malachite green solution was prepared by adding 44 mg malachite green carbinol base (Sigma-Aldrich) powder to the 36 ml 3N sulphuric acid solution. McrB and McrC were mixed in a ratio of 4:1 and this ratio was maintained for mutants as well. All the assays were performed in triplicates. For each GTPase assay, a master mix containing protein and 1 mM GTP (Jena Bioscience) in hydrolysis buffer (50 mM KCl, 10 mM Tris-Cl pH 8, 5 mM MgCl2, 1 mM DTT) was incubated at 37 °C. At regular time intervals, 20 µl of the reaction mix was withdrawn from the reaction mix. One 20 µl of the reaction mix was withdrawn at 0 minutes which was used as a

blank. The reaction was stopped at each time point by addition of 5 µl of 0.5 M EDTA. After 1 hour, all the samples were transferred to 96 well flat bottom plate. 50 µl of the malachite green mix (800 µl malachite green solution, 200 µl of 7.5 % ammonium molybdate and 16 µl of 11 % Tween 20) was added to each reaction sample and incubated for 10 minutes at room temperature. Absorbance was measured at 630 nm in a Varioscan plate reader.

To measure the amount of Pi released (in moles), a standard curve was plotted using different dilutions of a 2 M aqueous NaH2PO4 solution. 50 µl of the malachite green mix was added to each sample, incubated for 10 minutes at room temperature and absorbance was measured at 630 nm in a Varioscan plate reader.

2.4 DNA binding assay: -

DNA binding studies were carried out with a 60 bp specific DNA (containing a 5'R^{m}C3') and with a 70 bp nonspecific DNA (containing no methylated sites). Specific DNA was obtained by annealing two 60 bp complimentary single strands i.e. 5'-GCCGGGTAACCCGGGTAAGTCCGGGTAAGA^{m}CGGTAGTTCGGTATCGAGGGGTAGGCCGC-3' (MB60MSPI-1F) and 5'GCGGCCTACCCCTCGATACCGAACTA^{m}CGGTCTTACCCGGACTTACCCGGGTTACCCGGC3' (MB60MSPI-1R) (Integrated DNA Technologies, USA). The 70bp nonspecific oligo was obtained by annealing two 70bp ssDNA i.e. 5'TGACCATGATTACGCCAATCAGCAGCTCCAGGTCGTACCTCCAGCTACC AATC CCCGGGTACCGATCTCG3' and 5'CGAGATCGGTACCCGGGGATTGGTAGCTGGAGGTACGACCTGGAGCTG CTG ATTGGCGTAATCATGGTCA3' which was used in the binding assay only after purification using a MonoQ 10/100 GL column.

Binding reactions were carried out in a binding buffer containing 50 mM KCl, 10 mM Tris-Cl pH 8, 15 mM MgCl2, 1 mM DTT, 10% glycerol. 500 nM of specific or nonspecific DNA was incubated with different concentration of protein with 1 mM of GTP. 0.4 mg/ml BSA (NEB) was added to each reaction for enzyme

stability. The reaction is incubated at room temperature for 10 minutes, and 2 µl 6X ST buffer (40 % Sucrose, 0.2 M Tris-Cl pH 7.5) was added before loading onto a 5% native polyacrylamide gel. The native gel was pre-electrophoresed for 30 minutes at 4 °C, 20 mA in 1X TBE (89 mM Tris (pH 7.6), 89 mM boric acid, 2 mM EDTA) buffer. The samples were loaded onto the gel and run for 20 mins at 20 mA and 4 °C in 1X TBE buffer. After the run, the gels were stained in a solution containing 2 µg/ml ethidium bromide for five mins. Gels were imaged using Typhoon TRIO+ variable mode imager at high sensitivity.

2.5 Cleavage assay: -

An 114bp specific DNA substrate was generated by overlap PCR using MB60MSPI-1F and MB60MSPI-2R(5'AGTCAAATTGCATATGCTGGTCTTTCAGCGmCCGGTAATCGTCTTG TGAA GGATCCGCGGC-3') as primers. This was used to check cleavage activity of wild type, McrBL68F and McrBL68W. However, to decipher the cleavage pattern of McrBL68F different DNA substrates were used in the cleavage assay (see result section for the sequence of each substrate). All the substrates were either PCR purified or gel purified.

Cleavage assays were carried out in cleavage buffer (50 mM KCl, 10 mM Tris-Cl pH 8, 15 mM MgCl2, 1 mM DTT). NEB1 buffer (10 mM Bis Tris Propane-HCl pH 7.0, 10 mM $MgCl_2$, 1 mM DTT), NEB2 buffer (50 mM NaCl, 10 mM Tris-HCl pH 7.9, 10 mM $MgCl_2$, 1 mM DTT), NEB3 buffer (100 mM NaCl, 50 mM Tris-HCl pH 7.9, 10 mM $MgCl_2$, 1 mM DTT), NEB4 buffer (50 mM Potassium acetate, 20 mM Tris-acetate pH 7.9, 10 mM Magnesium acetate, 1 mM DTT) and Cutsmart buffer (50 mM Potassium acetate, 20 mM Tris-acetate pH 7.9, 10 mM Magnesium acetate, 100 µg/ml BSA) were also used in some of the cleavage assays. In the reaction mixture 75 nM of DNA substrate was incubated with protein (wild-type or mutants) in the presence or absence of GTP at 37 °C for 30 mins and 2 µl of 6X STE buffer (40% Sucrose, 0.2 M Tris-Cl pH 7.5, 40 mM EDTA, 1 % SDS) was added to each reaction before loading them onto 10 % native PAGE. After loading the gels were run at 180 V in 1X

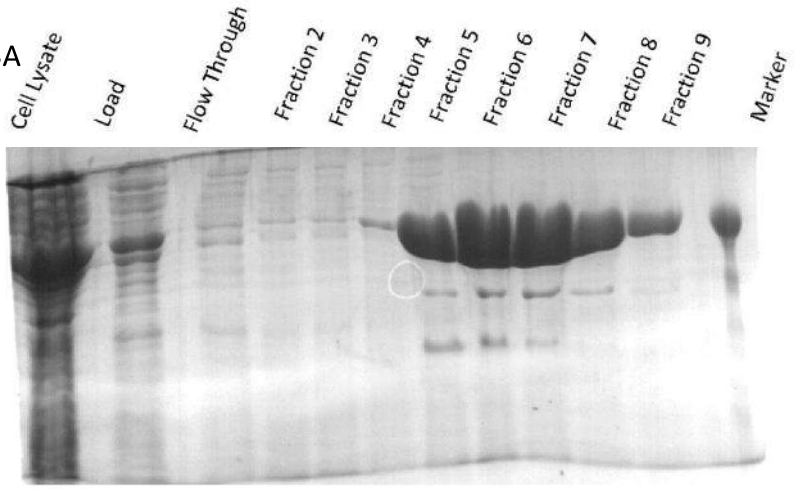
TBE buffer. Duration of running a gel was dependent on the length of the substrate used in that cleavage assay. After the run, gels were stained in a solution containing 2 µg/ml ethidium bromide for five mins. Gels were imaged using Typhoon TRIO+ variable mode imager at high sensitivity.

3 Results: -

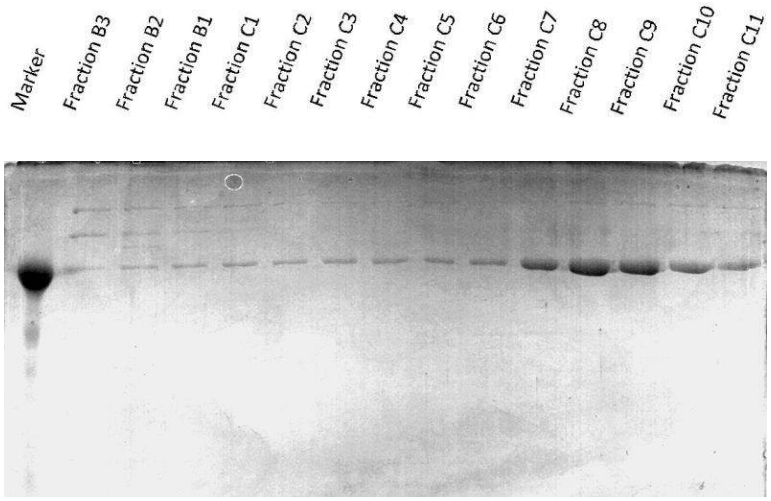
3.1 Purification of McrBL68F and McrBL68W mutant: -

As mentioned in the material and methods section both McrBL68F and an McrBL68W mutants were purified through Ni-NTA column followed by ion exchange chromatography. Figure1 (A, B, C & D) shows the Ni-NTA and MonoQ purification gel of McrBL68F and McrBL68W respectively.

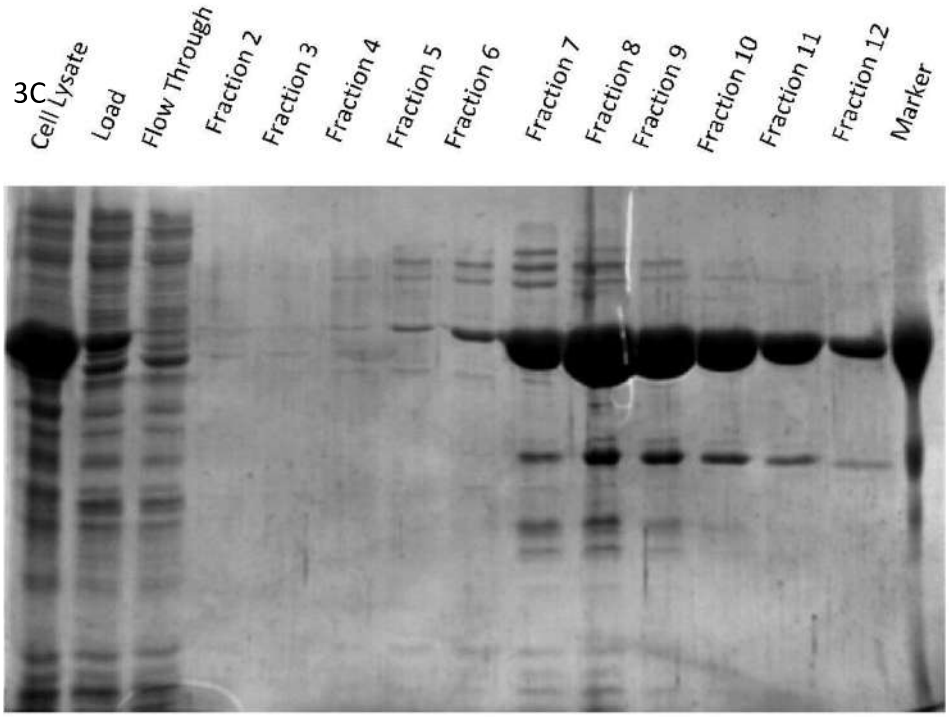
3A



3B



3C



3D

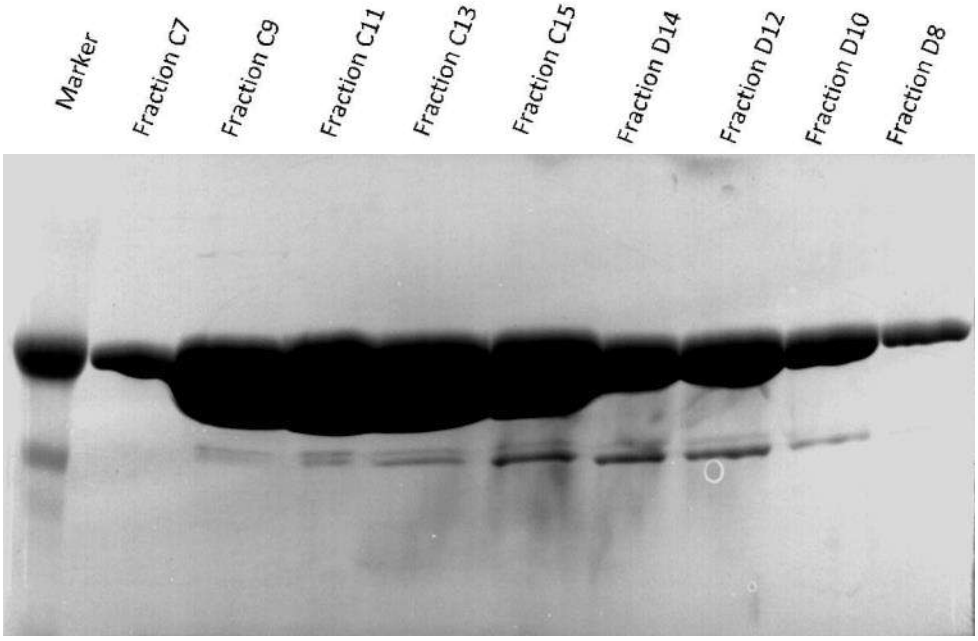


Figure3: Gels showing the purification profile of McrBL68F and McrBL68W mutants. A) McrBL68F Ni-NTA purification gel. Fractions 5-9 were pooled for dialysis. B) McrBL68F MonoQ purification gel. Fractions C7-C11 were pooled and concentrated. C & D) representative gels for McrBL68W Ni-NTA and MonoQ purification respectively. After Ni-NTA fractions 7-11 were dialysed and after MonoQ fractions C9-D14 were pooled and concentrated.

After concentration and washing the protein with storage buffer (see the material section for composition) final yield for McrBL68F was 7.8 mg/ml whereas yield for McrBL68W was 9.2 mg/ml.

3.2 GTPase characterisation of McrBL68F and McrBL68W: -

To check whether the mutants are active or not, their GTPase activity was checked and compared with that of wild-type. McrB shows significant GTPase activity in the presence of McrC. Hence to compare GTPase profile, McrB, McrBL68F and McrBL68W were mixed with McrC in 4:1 ratio & time dependent GTPase assays were performed for all three proteins. A comparative plot showing the GTPase activity of both mutants and wild type is shown in figure 2.

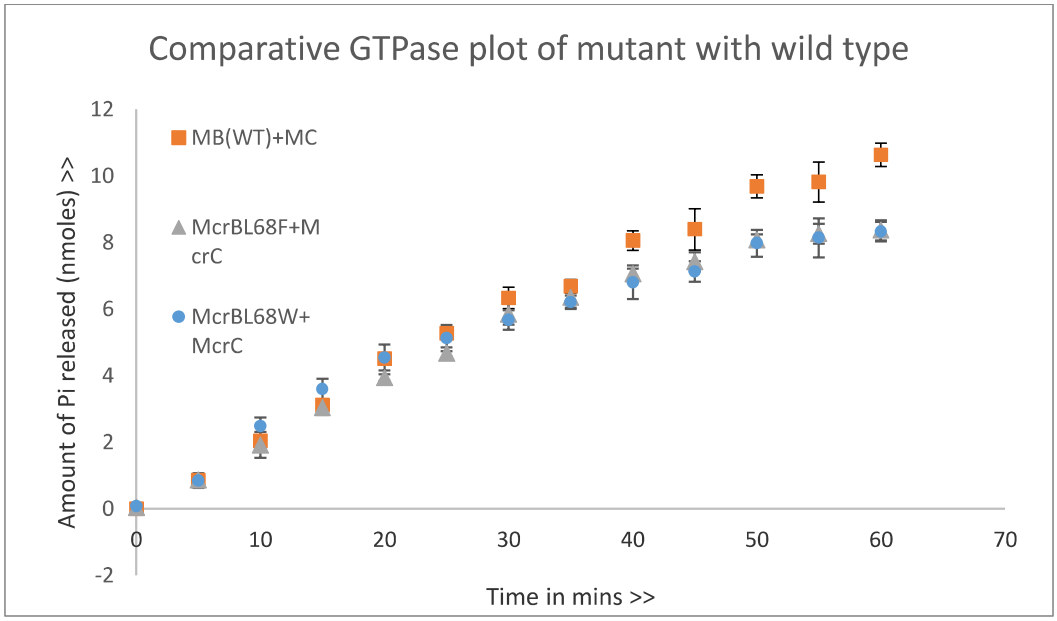


Figure 4: Comparative GTPase profile of McrBL68FC & McrBL68WC with McrBC wild type protein. Phosphate released is measured in nmoles as shown in y – axis.

The plot shows that all of them follow the same pattern and have similar GTPase activity at different time points indicating that these mutants are purified in active form.

3.3 DNA Cleavage test of McrBL68F with plasmid DNA substrate: -

McrBC cleaves DNA containing two 5' – R^mC – 3' sites. Hence all the plasmids extracted from dcm+ cloning strain (used in this study) cannot be cleaved by McrBC wild-type protein, as the cytosines preceded by purine bases are not methylated. To check whether a McrBL68FC (McrBL68F+McrC) can cleave plasmid DNA, two plasmids pUC18 (2.6 kb) and mb-pHIS (4.0 kb) were used as substrates for checking the nucleolytic effect of McrBL68FC. McrBC was used as a control in this reaction.

Cleavage result shown in figure 4 revealed that L68F mutation could cleave plasmid DNA. This indicated that the target site specificity of the mutant either changed or broadened. Wild-type, as expected, did not cleave plasmid DNA.

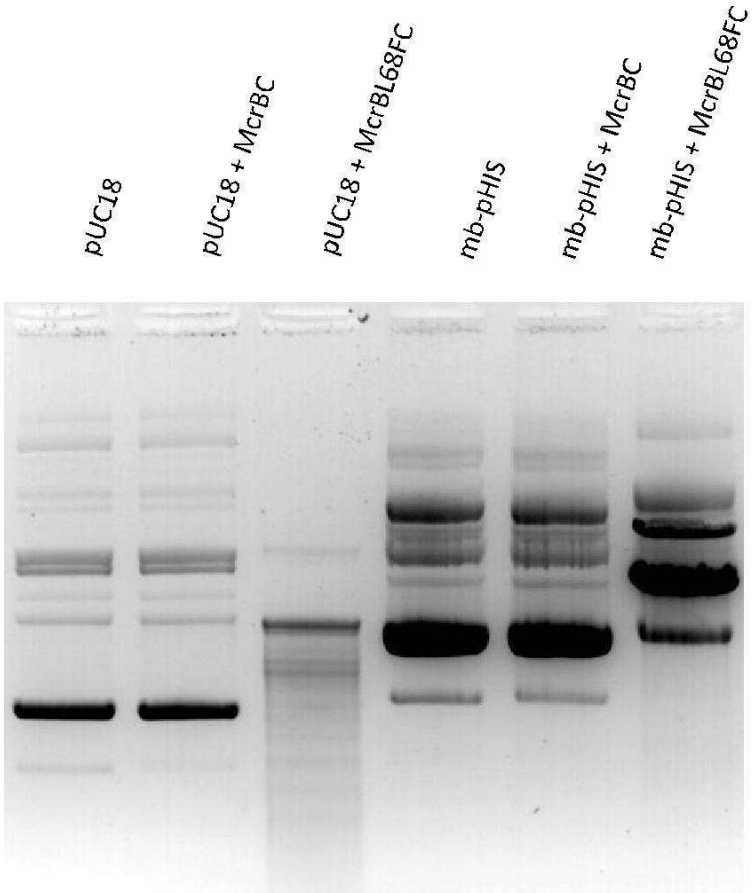
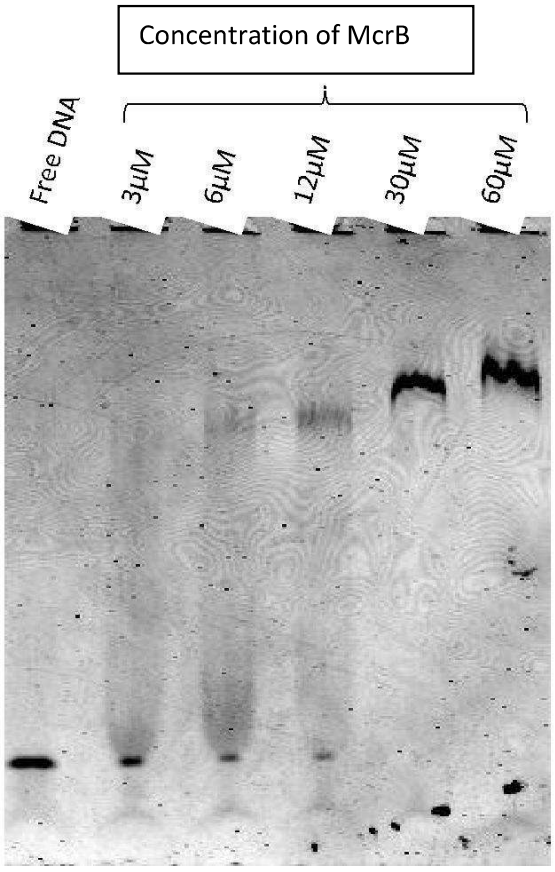


Figure 5: - Cleavage activity of McrBL68FC checked with plasmid substrates pUC18 and mb-pHIS. In both, the cases protein is showing cleavage activity, unlike McrBC wild type which doesn't have nucleolytic activity on these plasmids.

3.4 Comparative binding studies with specific and nonspecific DNA: -

Having established that the mutant protein cleaved with an altered specificity, we proceeded to find how its affinity for specific DNA which contains one 5' – R^mC – 3' site (see the material section for sequence details) varied as compared to that of wild-type. As a control, binding studies with wild type was carried out in a concentration-dependent manner, which was repeated along with mutant protein (Figure 6A & 6B). Also, a 70 bp non-methylated non-specific DNA was also used to see whether the mutant can bind to non-methylated DNA.

6(A)



6(B)

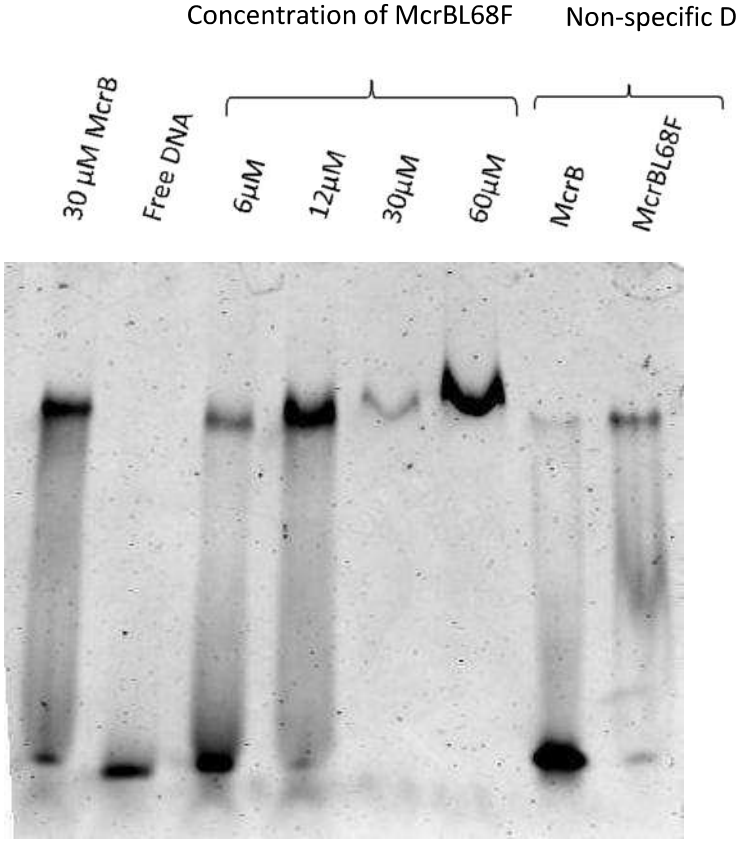


Figure 6: Concentration-dependent binding assay of McrB (6A) and McrBL68F (6B) with 500 nM of specific DNA. The concentration of McrB/McrBL68F is shown above each lane. 4B also shows the binding of non-specific DNA with both McrB wild type and McrBL68F.

The binding assay showed that binding of the McrBL68F mutant was comparable to that of the wild-type protein. However, this does not prove that the mutant bound to methylated cytosine. Additionally, the mutant had a strong binding affinity for nonspecific DNA as well, indicating that the mutant could recognise non-methylated substrates. As mentioned earlier, the affinity of the wild type enzyme for non-methylated DNA was very poor (Figure 4B).

3.5 Concentration-dependent cleavage assay with 114 bp specific DNA: -

To investigate the cleavage pattern of the McrBL68F mutant protein as compared to that of wild-type protein concentration dependent cleavage assay was performed with a 114 bp specific substrate (figure 7) which contains two 5' – R^mC – 3' sites (see material section for more details).

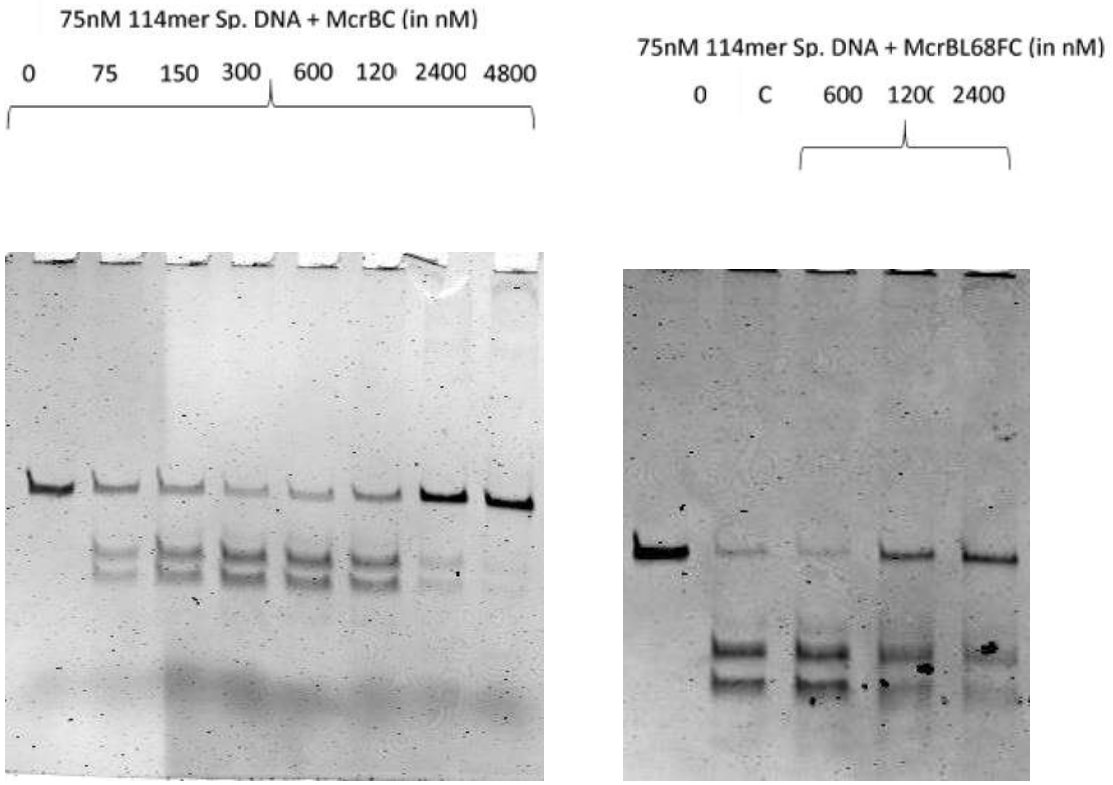


Figure 7: Representative gels for cleavage assay of McrBC and McrBL68FC are shown in 5A and 5B respectively. For the cleavage assay, 75 nM of the 114 bp specific substrate is used, and concentration of McrB/McrBL68F for each reaction is shown above each lane. 'C' in the second lane of fig 5B is controlled i.e. 600 nM of McrBC with 114mer.

The cleavage assay indicated that the mutant McrBL68FC cleaves specific DNA in a similar pattern as that of McrBC. However, whether this pattern is the result of the mutant recognising methylated cytosine or a cytosine site close to the methylated cytosine cannot be deciphered. The 114 bp specific substrate has many cytosines near the methylated cytosine that can act as a binding site for the mutant (see Figure 8).

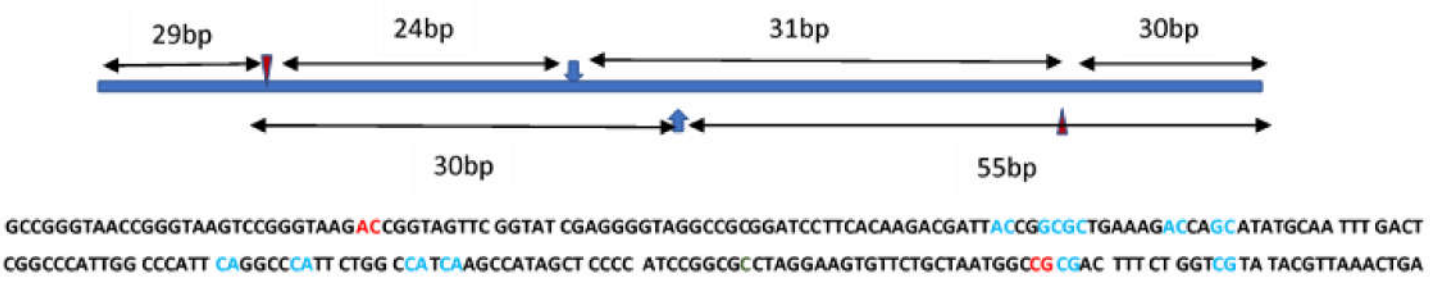


Figure 8: A diagrammatic representation of cleavage pattern of McrBC in 114 bp specific DNA. The two R^{m}C sites are pointed with a red triangle, and the possible cleavage sites are shown in arrow mark. R^{m}C sites in the sequence are coloured red. Also, other RC sites near R^{m}C are highlighted in cyan colour.

Both wildtype and mutant have inhibitory effect at higher concentration. Optimum cleavage is seen when the concentration is around 600 nM – 1200nM. Hence for further cleavage assays 600 nM of protein (600 nM McrB & 150nM of McrC) was used.

3.6 Cleavage check for 114 bp containing only two 5'RC3' sites: -

The mutant was able to bind to non-methylated DNA and also leave 114 bp specific substrate. Hence in order to investigate whether it can recognise non-methylated cytosine preceded by a purine base, a new 114 bp substrate was designed (RC substrate). It had only two 5'RC3' sites located at the same position where the specific substrate had R^{m}C site, and the rest of the sequence did not contain any 5'RC3' (see sequence below for more details). Cleavage activity was also checked with another 114 bp substrate which had no 5'RC3' site at all (NonRC substrate). However, both the substrates were not cleaved by McrBL68F as shown in figure 9.

Sequence of RC substrate: -

GAAGAGAATCTAGAGATAGAGAGGGAAAGACCCTAGAAGAGAAAGGAGAGATAAGGGGAGATCCTTCTCTAA TCC TAT AGGAGCTAGGAAGATCCTCCTT ATCC TA TTAGGAT CTT CTCT TA GATCTC TAT CTCTCCC TTT CTGGGATCTTC TC TTT CCTC T CTATT CCCC TC TAGGAAGAGATTAGGATATCCTCGATCC TT CTAGGAGGAATAGGATAATCCTA

Sequence of NonRC substrate: -

GAAGAGAATCTAGAGATAGAGAGGGAAAGATCCTAGAAGAGAAAGGAGAGATAAGGGGAGATCCTTCTCTAA TCC TAT AGGATCTAGGAAGATCCTCCTT ATCC TA TTAGGAT CTT CTCT TA GATCTC TAT CTCTCCC TTT CTAGGATCTTC TC TTT CCTC T CTATT CCCC TC TAGGAAGAGATTAGGATATCCTAGATCC TT CTAGGAGGAATAGGATAATCCTA

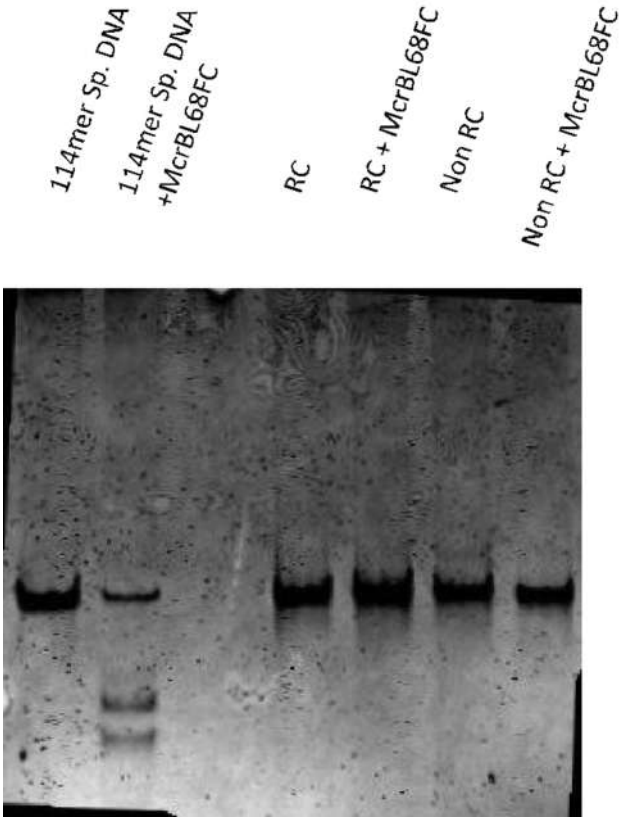


Figure 9: Cleavage assay of McrBL68F with RC and Non-RC substrates. Cleavage of the 114 bp specific substrate is used as a control for the reaction.

The mutant did not cleave RC substrate. This indicates that the protein binding to target site could be dependent on the flanking sequences. Influence of flanking sequences on the affinity of McrBC wild type protein has already been shown[30,31].
It also didn't cleave NonRC substrate which indicated that the mutant could not cleave a substrate that did not have any 5'RC3' sites. One of the other reasons could be that the mutation had completely changed its specificity and the site required for binding was absent in both RC and NonRC sites. It could be that an appropriate base flanking the cytosine was required which may be absent in the RC and NonRC DNA.

3.7 Cleavage check with other non-methylated substrates: -

To check the nucleolytic activity of McrBL68F in non-methylated substrates, DNA of different lengths were used as a substrate for cleavage assay. Longer the substrate length higher is the probability of the existence and recognition of the target site of McrBL68F.

Six DNA substrates (Sub I – VI) having a length range from 150bp to 400bp were taken, and cleavage assays were performed with McrBL68F (figure 10).

8A 8B 8C

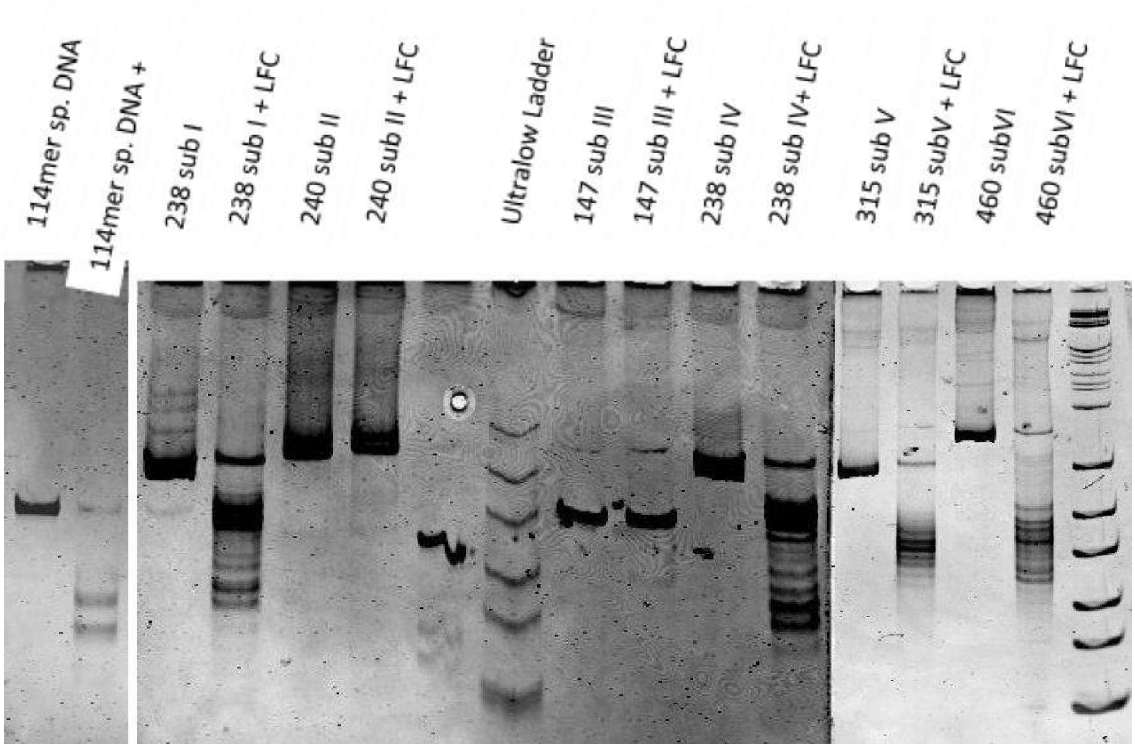


Figure 10: - Cleavage assays for 6 non-methylated DNA substrates were performed. The length of each substrate is mentioned before their substrate number. Cleavage assay shown in a & b were performed together and c was performed at a different time. (The gel images are presented together for the ease of comparison).

Cleavage assays with non-methylated substrates clearly showed that the protein could not cleave all the DNA substrates and hence this mutant had a preference for certain target sites over others. While sub I, IV, V and VI got chopped forming some prominent cleaved bands, sub II and III were not cleaved at all. Though all the substrates have many RC sites, they result in either distinct bands instead of being degraded to very short fragments, or not getting cleaved at all. This poses an interesting question regarding the specificity of this mutant and the factors which influence the cleavage activity of this protein such as flanking sequence, the orientation of binding site, etc.

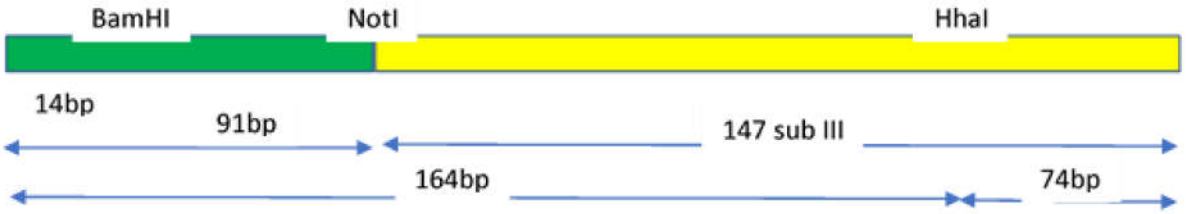
3.8 Analysing cleavage of 238 bp sub IV by McrBL68FC: -

In the cleavage assay shown in figure 10, the sequence of 147 bp sub III which did not get cleaved is a part of 238 bp sub IV (see the figure below). 238 bp sub IV gets cleaved forming several bands with prominent bands being near 150 bp and 75bp region. This clearly indicated that the 147 bp subfragment of 238 bp substrate did not harbour at least two binding sites required for cleavage by the enzyme, while the addition of 91 more base pairs upstream of this provided a suitable condition for the mutant to bind and cleave DNA.

Hence, to narrow down on the target site, 238 bp sub IV was digested with three site-specific Type II restriction enzymes BamHI, NotI and HhaI. While BamHI-cleaved the DNA at 14^{th} base pair (resulting in 14bp and a 224bp), NotI had the site near 91^{st} base pair (producing 90bp and 147bp, which is equivalent to sub III), and HhaI cleaved at 164^{th} base pair (producing 164bp and a 64bp fragment). After restriction digestion, these newly formed substrates were incubated with McrBL68FC to check the cleavage pattern of the mutant, which is shown in figure 11.

11(A)

238 Sub IV:



11(B)

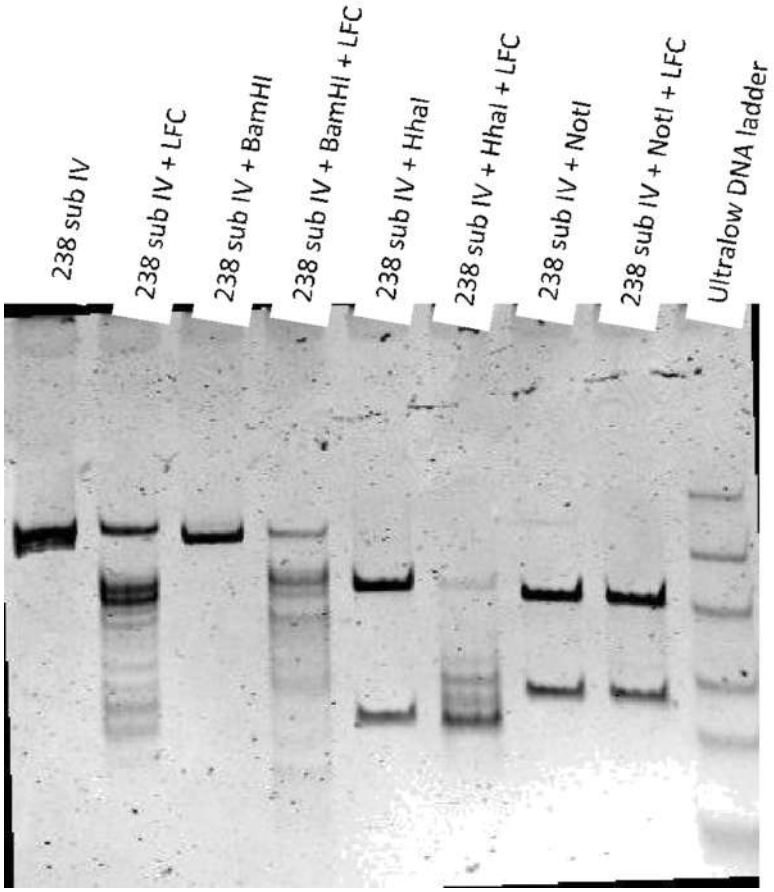
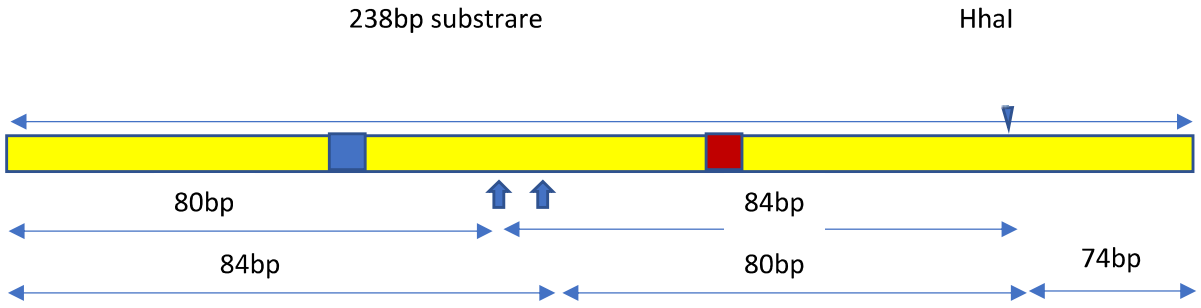


Figure 11: Cleavage with Type II digested 238bp substrate IV. Gel showing the cleavage pattern of 238 bp sub IV digested with BamHI (lane 3,4), HhaI (lane 5,6) and NotI (lane 7,8) with McrBL68FC (LFC). Ultralow DNA ladder was run along with these samples in lane 9.

McrBL68FC cleaves BamHI digested DNA similarly as compared to uncut 238 sub IV indicating that all the binding sites of the protein are after 14bp of this substrate.
NotI digested products didn't get cleaved as expected. This was because NotI cuts this DNA and produces 147bp DNA equivalent to sub III as one of its product, which was not susceptible to cleavage. This also tells us that the mutant required binding sites on both sides of NotI site to cleave DNA.

However, cleavage with HhaI digested 238 sub IV provided interesting insights into the cleavage pattern of the mutant. After digestion, McrBL68FC produced DNA substrates of length 164bp and 74bp as shown in figure 11. After cleavage, 74bp fragment stayed uncleaved whereas 164bp got cleaved into 2 prominent fragments of approximately 80bp and 84bp. This can emerge from two types of cleavage, which are shown in the figure below.



: CACA (37, 47), AGCC (41), CACC (49), AACT (54)

: CACC (105), TGCC (107), GGCC (113), CGCT (116), GACC (125)

Figure 12: a representative diagram of cleavage pattern of HhaI digested DNA products. Arrow marks shown in the diagram represent the possible sites of cleavage which is either 80 bp or 84 bp from the left end. Blue and red coloured squares represent the possible binding site for the protein based on the assumption that the protein binds to 30 bp upstream of one of the cleavage site (like the wild-type). RC sites with the two adjacent bases flanking it located near this region (50th and 110th position) are given below the figure, and the numbers represent their position in the sequence.

Amongst all the probable sites mentioned in the figure, GACC, AGCC and CACA are three sites which are absent in 238 bp sub II or 147 bp sub III (substrates which were not cleaved by the mutant.) (see annexure section for sequence details.)

Out of these three sequences, GACC was first chosen to be tested as to whether it is the binding site for the mutant. This is described in the next section.

3.9 Cleavage check with 114 bp GACC substrate: -

A 114 bp GACC substrate was designed whose sequence was same with that of RC substrate except it had GACC at both ends (in RC substrate it is GACC on one side and AGCT on the other side). Cleavage assay with this substrate is shown in figure 11.

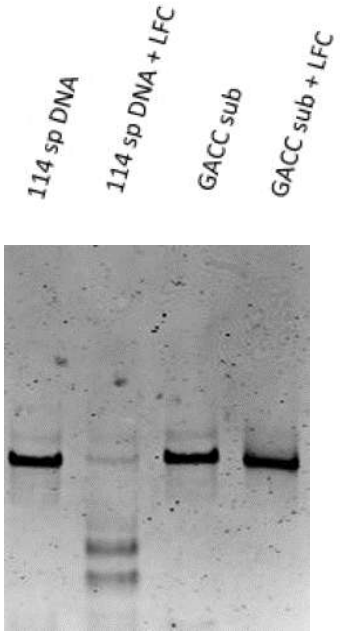


Figure 13: Cleavage assay with GACC substrate is shown. The 114 bp Specific methylated substrate is used as a control. LFC refers to McrBL68FC.
The mutant protein did not cleave this substrate indicating that GACC is not the target site for the mutant or it does not provide a suitable environment for two complexes of McrBL68FC to bind and cleave DNA.

3.10 Cleavage check with Single site methylated CGCC substrate: -

The mutant was cleaving 114 bp specific DNA but not non-methylated DNA. Hence, an intermediate Single site methylated substrate was designed which had the same sequence as that of 114 bp specific sequence except for methylation from the top strand was removed and the GACC site was replaced with CGCC site (identical to the methylated site of lower strand.) GACC was replaced as the protein doesn't seem to recognise that site. Cleavage assay performed with this substrate is shown in figure 14 below.

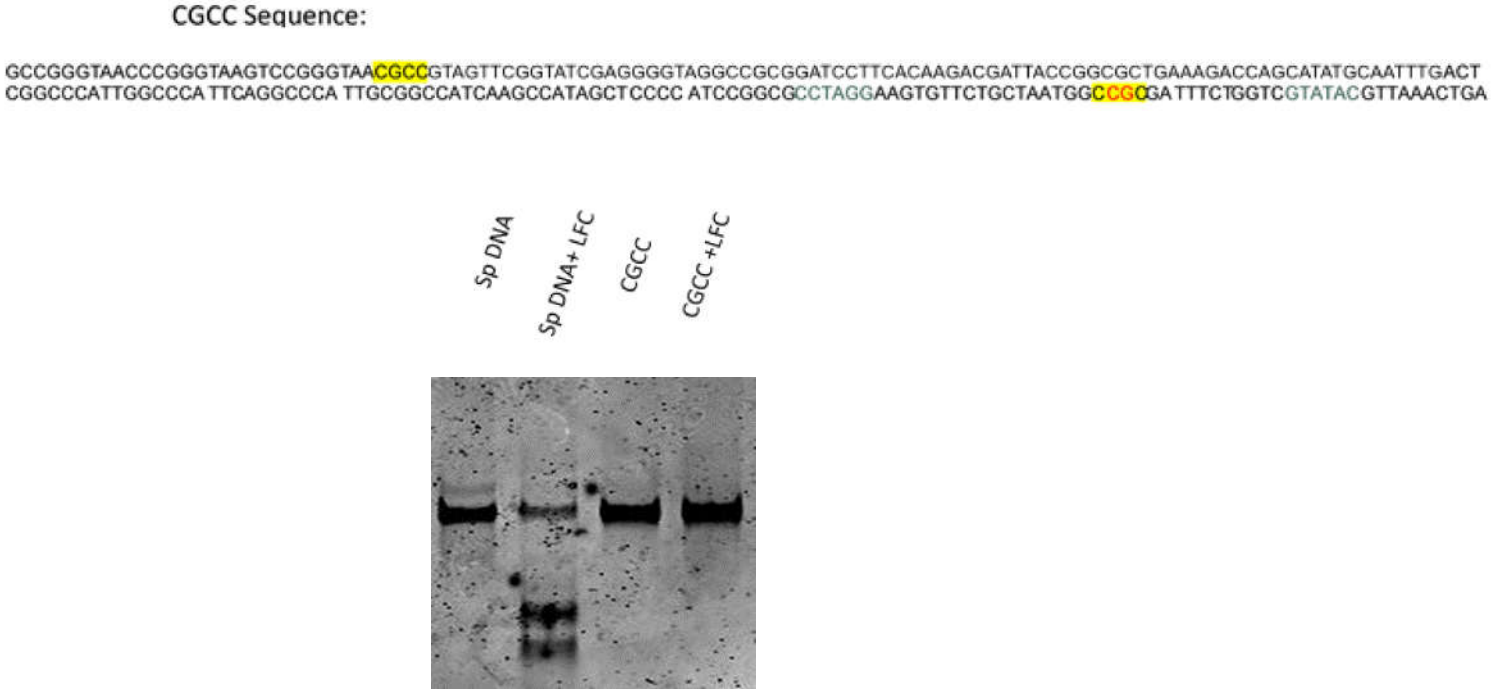


Figure 14: Cleavage assay of McrBL68FC (LFC) with CGCC substrate. The sequence of the substrate is mentioned above the gel.

Protein didn’t cleave CGCC substrate as well which was striking as this DNA had only one modification and the cleavage activity of the protein was not observed. Before trying other possible target sites, we decided to check the activity of the protein with the longer substrates.

3.11 Cleavage check of 900bp substrate VII: -

In order to check whether the protein is active enough to cleave longer substrates into smaller DNA fragments, a 900 bp sub-VII was chosen for cleavage assay. Two batches of this substrate were available. One was agarose gel purified, and the other was PCR purified. Cleavage of identical substrates purified using different methods was surprisingly giving different patterns of cleavage with McrBL68FC as shown in figure 15.

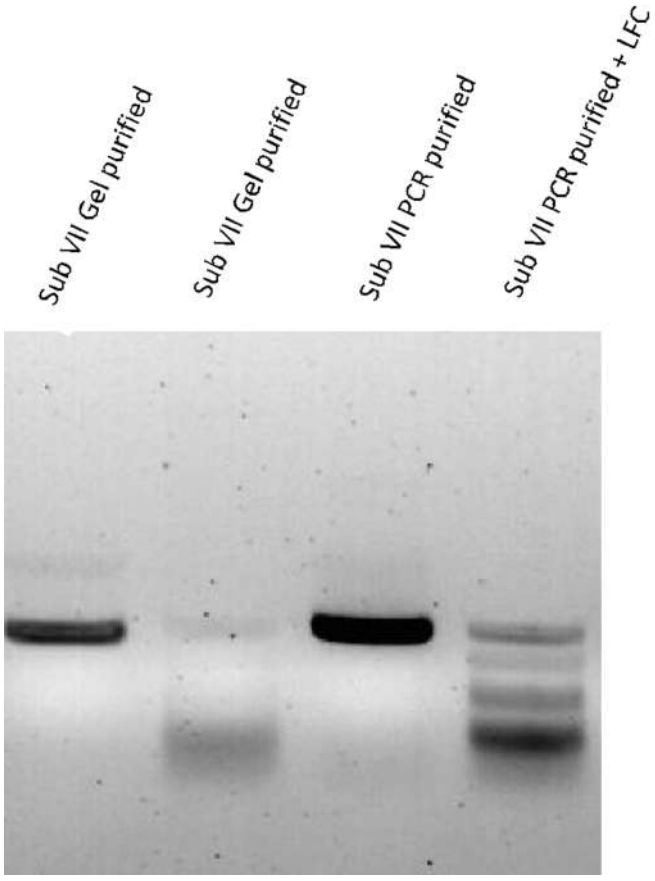


Fig 15: Mutant cleavage assay with gel purified and PCR purified substrate VII.

As seen in figure 15 gel purified DNA is more efficiently cleaved by the mutant as compared to PCR purified substrate. This clearly indicated that the cleavage activity of the mutant enzyme was modulated by the ingredients of the buffer. Gel purified substrates often have contaminants like acetate ions, guanidinium ion, etc. These contaminants could have modulated the enzyme activities. Hence, in order to check the effect of ingredients, the cleavage of activity of the mutant protein was analysed in different buffers.

3.12 Cleavage activity of McrBL68FC under different buffer conditions: -

PCR purified sub VII DNA was chosen as the substrate and cleavage were checked using NEB1, NEB2, NEB3, NEB4 and cutsmart buffers (see material section for their composition) (figure 16).

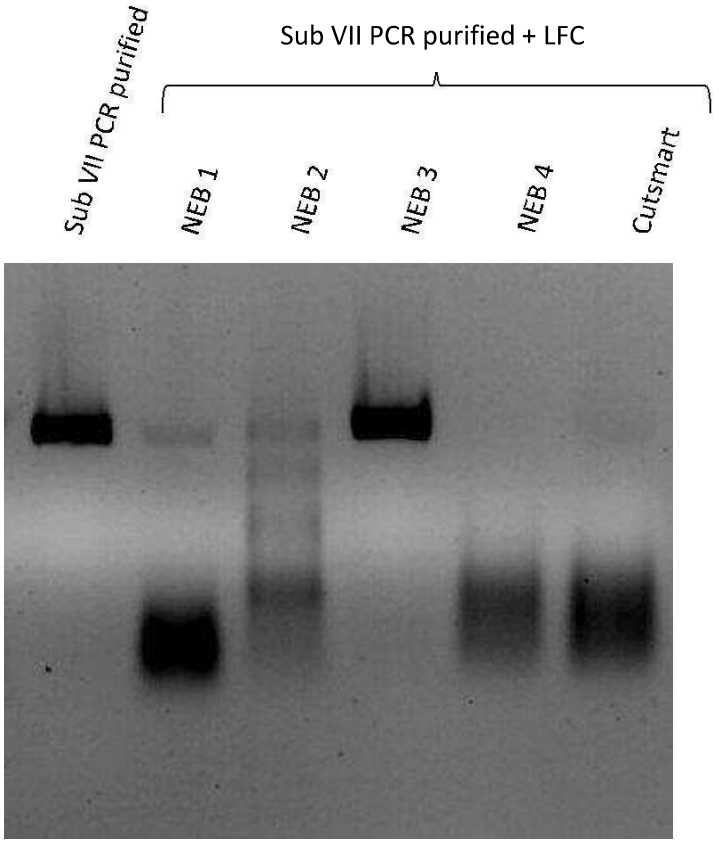


Figure 16: cleavage assay of sub-VII in different buffers as mentioned above each lane. Uncut DNA is represented in the lane I.

Use of different buffers to check the nucleolytic activity of the protein clearly showed that it was buffer dependent. Best cleavage activity was seen with NEB1, NEB4 and cutsmart buffers. NEB2 had a composition similar to that of the original cleavage buffer, and hence it gave a pattern similar to that seen with the original buffer. NEB3 completely inhibited the activity of the protein. Hence, NEB1 and NEB4 buffer were chosen for subsequent experiments to check the cleavage of the 114 bp substrates (specific, RC, GACC & CGCC substrates).

3.13 Cleavage check for 114 bp in NEB1 and NEB4 buffers: -

All the available 114 bp substrates like specific, RC, GACC, CGCC substrates were checked for cleavage activity of the McrBL68FC with NEB1 and NEB4 buffers (figure 17).

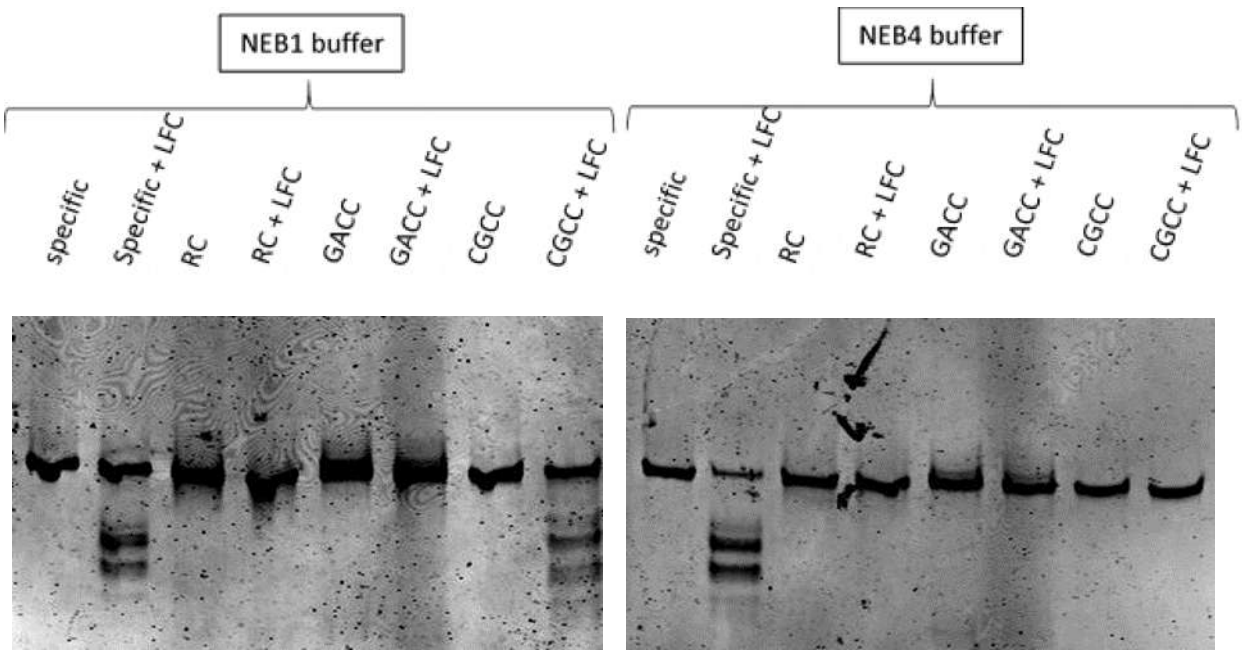


Figure 17: the Nucleolytic activity of LFC with 114 bp substrates (mentioned above each lane) with NEB1 and NEB4 buffers.

With NEB1 buffer the protein cleaved CGCC substrate albeit with lower efficiency. This clearly tells that the protein under optimised buffer condition can cleave a single site methylated DNA containing CG^{m}CC site at one end and CGCC at the other end.

3.15 Cleavage check for MBL68W mutant: -

Now that McrBL68FC was seen to have gained specificity towards the non-methylated substrate, McrBL68W was checked for its cleavage activity. But as shown in figure 16 it neither cleaved specific methylated 114 bp nor 900 bp sub-VII.

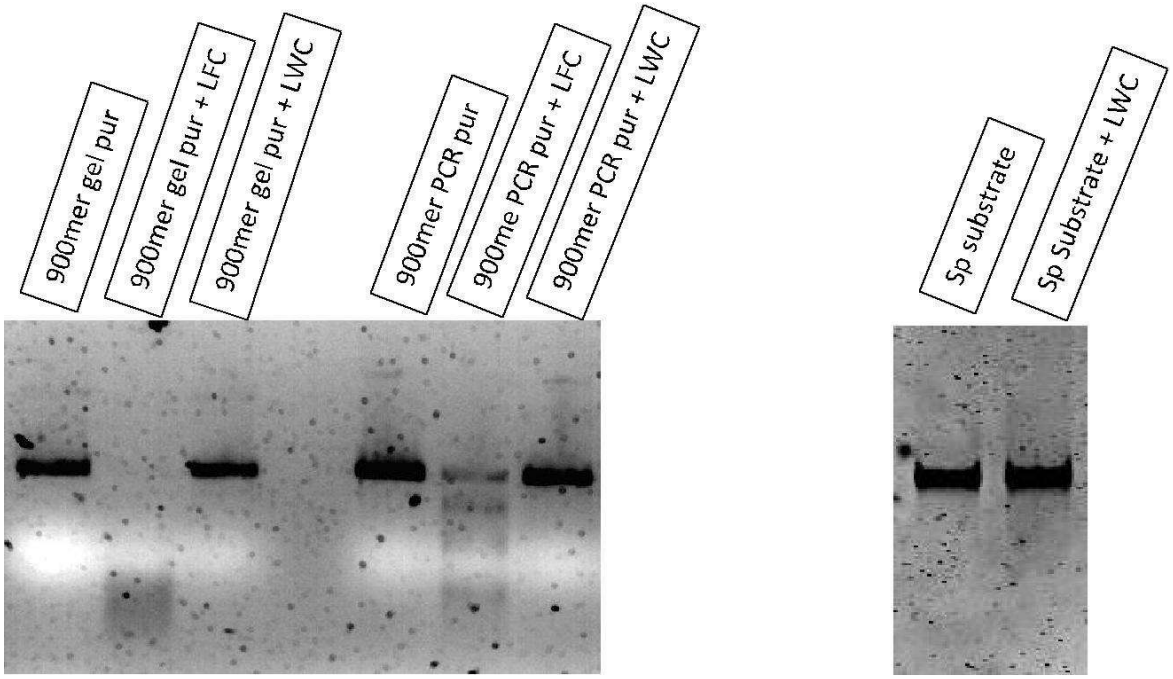


Figure 18: Both gel and PCR purified 900 bp were used in cleavage assay (Left). McrBL68W + McrC is denoted as LWC. Cleavage was also performed with a specific substrate which also didn't give any cleavage result.

This tells us that McrBL68W is not allowing methylated as well as non-methylated cytosine into the catalytic pocket. Hence, this mutation provides an ideal model for optimum length prediction to maintain hydrophobic interaction.

4 Discussion: -

As reported in the crystal structure of McrB-N-terminal bound to DNA, very few amino acids seem to make direct contact with the DNA bases. However, the flipped-out DNA base seems to be perfectly placed inside the catalytic pocket with hydrogen bonds, stacking interaction and hydrophobic interactions as mentioned earlier. One of the most important characteristics of McrBC is to maintain its high specificity towards methylated cytosine containing substrate because promiscuity towards non-methylated substrates would make it recognise self-DNA in a bacterial cell leading to the cleavage of the native genome which would be detrimental to the survival of the cell. However, in the structure, only a Leucine residue seems to be responsible for making this protein discriminate between methylated and non-methylated substrate that too through hydrophobic interaction and poses a great puzzle to how this sole interaction is chosen by this class of protein to maintain its specificity. In this project, I have tried to reduce the distance of this interaction by mutating it to phenylalanine and tryptophan to see whether the protein could still stay active and change its specificity towards non-methylated substrates. While McrBL68W does not seem to recognise either methylated or non-methylated substrates, McrBL68F seems to have gained specificity for non-methylated substrates, while still appearing to recognise methylated DNA in the same way that wild type does.

In order to check whether the purified mutants were active, the GTPase activity was compared with that of the wild type. McrB alone has low intrinsic GTPase activity, which is stimulated in the presence of McrC. Hence, the McrB mutants were also combined with McrC, and time-dependent GTPase assays were performed. Both the mutants showed comparable GTPase activity as that of wildtype McrBC, which indicated that both the mutants had been purified in active form.

Initial cleavage check with pUC18 plasmid and mb-pHIS plasmid confirmed that McrBL68F mutant had gained a specificity, which the wild type enzyme lacked. Because of the absence of 5’ – RmC – 3’ sites in dcm+ extracted plasmids, wild type tends to have no cleavage activity, but McrBL68FC mutant could cleave both the plasmids. This was the first evidence of the mutant showing alteration in the

specificity of wild-type target site, which was investigated further.

When concentration-dependent binding assays were performed in McrBL68F mutant showed similar binding to the methylated substrate as that of wild type suggesting that the protein still may have retained its specificity towards methylated substrates. One of the interesting results which came out of the binding assay was that the mutant, unlike wild type, showed higher affinity for a non-methylated nonspecific 70 bp substrate which suggested the protein gained additional specificity for non-methylated DNA.

Concentration-dependent cleavage assay with methylated specific 114 bp substrate showed similar cleavage pattern for both McrBL68FC and McrBC, where optimum cleavage is seen in a concentration range of 600 nM to 1200 nM and inhibition of nucleolytic activity at higher concentrations. This again strengthens the idea of protein recognising methylated cytosine. However, the presence of other non-methylated RC sites in the vicinity of the methylated cytosine in the specific substrate can result in the mutant binding to those sites and give a similar cleavage pattern. Hence, in order to get rid of this problem RC and Non-RC substrates were designed, which had only one pair of RC sites and no RC sites, respectively. But, the mutant did not cleave either of those substrates. This could mean that the binding of this protein to a substrate is dependent on its sequence or it has a lower affinity for the RC sites chosen to investigate the cleavage pattern.

Hence, to find the preferential binding site of the protein many of the non-methylated substrates with a length range of 140 to 400 bp were chosen for cleavage assays with the mutant. One of the striking outcomes was the protein not being able to cleave 147 bp sub III and 238 bp sub II although these substrates have many RC sites in their sequence. The 147 bp sub III is a part of 238 bp sub IV which was getting cleaved into several bands, and hence these two substrates stand as an ideal model to check the cleavage pattern of this protein. When restriction digestion was performed for sub IV, it gave a hint regarding the possible binding sites of the mutant, and out of all possible binding sites, GACC was chosen first, and a 114 bp substrate that contains GACC at both ends was designed. But this substrate was also not cleaved by the mutant.

In order to check whether a single site methylated substrate will be cleaved, hemi-methylation was incorporated in CGCC-substrate where one site was non-methylated CGCC, and the other site was $CG^{m}CC$. This substrate was also not cleaved by this mutant. However, when methylation is on both sites, the protein tends to cleave it with the same pattern as that of the wild type. x.

As substrate length was emerging as one of the possible crucial factors in influencing the activity of the protein, cleavage was carried out with a 900 bp substrate. While the mutant completely shredded the gel purified substrate, cleavage efficiency was less when the same substrate was PCR purified and used for cleavage. This clearly indicated that the protein activity was influenced by the ingredients of the reaction conditions. The chemical impurities from gel purified DNA could be stabilising the protein and hence increasing its cleavage efficiency. Therefore, cleavage activity was checked under different buffer conditions. Surprisingly, buffers showed different cleavage patterns with NEB1 and NEB4 showing highest efficiency. If one goes with the composition of the buffers, NEB2 has an almost similar composition with the cleavage buffer used in all assays, and as expected that gave similar cleavage pattern as was seen with cleavage buffer. However, NEB3, where the concentration of all the components of NEB2 is increased, inhibits the activity of the protein. The only difference NEB1 and NEB4 have with respect to NEB2 and NEB3 is the lack of sodium salt. Hence It can be argued that sodium could be acting as an inhibitory ion for the protein. But NEB1 was the only buffer in which protein cleaved a 114bp single site methylated CGCC substrate. This requires further investigation into checking the nucleolytic activity of the protein with already used substrates and further optimisation is needed.

McrBL68W mutant, however, didn't cleave methylated or non-methylated substrate indicating that the mutation might not be allowing even non-methylated cytosine to get into the catalytic pocket. This provides a fascinating insight into the optimum size of the nucleotide binding pocket as a mode to recognise its desired base.

In future, longer substrates are required to be designed in order to find the desired binding site of the protein as length seems to be one the possible factor playing a

role in making the protein stabilise on DNA. Also, finding out the environmental factors like buffer composition influencing the nucleolytic activity of the protein is essential. One of the crucial findings of this study is how alteration in weak hydrophobic interaction can lead to dramatic modification in the specificity of protein. Mutating Leu to Phe could be considered as the first step to making McrBC recognise non-methylated substrate, and it is our future goal to design a mutant which would only recognise non-methylated substrate.

This holds significant value in the field of epigenetics, sequencing etc. Designing a mutant of the type IV restriction system that can alter the specificity of the protein by recognising unmodified bases provides fascinating insights into the molecular mode of mechanism the protein employs to recognise its target site. However, if with certain mutations in the neighbouring residues of Phenylalanine that would perfectly position it to block methylated group of cytosine and make the protein only recognise non-methylated bases that could be used as a complimentary system to sequence DNA containing modified bases. Also with the increasing reports of involvement of epigenetic in cellular as well as physiological behaviour in various organisms, having a system of proteins that can recognise modified bases in its wild type form and unmodified bases in mutant form would be a useful technological tool.

Appendix (substrate sequences): -
240bp Sub II

5'CATATGGTGAAGGTAAAGTTCAAGTATAAGGGTGAAGAGAAAGAAGTAGACA
CTTCAAAGATAAAGAAGGTTTGGAGAGTAGGCAAAATGGTGTCCTTTACCTATG
ACGACAATGGTAAGACAGGTAGAGGAGCTGTAAGCGAGAAAGATGCTCCAAAA
GAATTATTAGACATGTTAGCAAGAGCAGAAAGAGAGAAGAAAGGATCCCGTGT
CACGTCACTGGCAATGAATGGCGTCGGC**3'**

147bp sub III

5'CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGC
ACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAGGGGAT
TACTCCCTAGTCTCCAGGCACGTGTCAGATATATACATCCTGT**3'**

238bp sub IV

5'GCCTGCAGGTCCGGGATCCTAATGACCAAGCTAGACGTGAGCCTTCACACC
GAGTTCATCCCTTATGTGATGGACCCTATACGCGGCCGCCCTGGAGAATCCCG
GTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAAACGCAC
GTACGCGCTGTCCCCCGCGTTTTAACCGCCAAGGGGATTACTCCCTAGTCTCC
AGGCACGTGTCAGATATATACATCCTGT**3'**

315bp Sub V

5'GTTTACTTTAAGAAGGAGATATACATATGTCCCTGGACACCCCCAACGAGAA
GCCCGCTGGCAAGGCTCGCGCCCGGAAGGCCCCCGCCTCCAAGGCCGGCGC
CACGAACGCGGCGTCGACCTCTTCCTCCACCAAGGCCATCACCGACACGCTG
CTGACGGTGCTGTCCAGCAACCTGCAGGCCCGCGTGCCCAAGGAGCTGGTCG
GTGAGTCCGGCGTGGAGCTGGCGCACCTGCTCAACCAGGTGCTGGACCAGTT
CGCGGCCTCCGAGCACCGCAAGCATGGATCCCATCATCATCATCATCATTAAA
AGC3'

460bp Sub VI

5'AAGAAGGAGATATACATATGGATGGGCACGCAACTGGTGATGTACGAAGAG
GAGTTCACCAAGATCAACGCCGTTTGCGACCGGCTTACCAAGGACGCGAACG
CGAAGGTGGTCTTCCTCGTCGACAAGAACGGGCAGCTCATCTCCTCCGCGGG
TCAGACGCAGAACATCGACACCACGTCACTGGCCTCGCTGACGGCCGGTAAC
GTGGCCGCGATGGGTGGCCTGGCCAAGCTGATTGGGGAGAACGAGTTCCCCA
ACCAGTTCCACGAGGGGGCCAAGGACTCGCTGTACATGACCATCGTCGGCAG
CCGGGTCGTGCTGGTCGTCATCTTTGACAACCGCACCAGCCTCGGCCTCGTC
CGCCTTCGCATCAAGAAGGCCAGCGACGAGCTCACGAAGATCTTCGAGAGCC
TGGTGAAGAAGACTGACAGTCCTCATCATCATCATCATCATTAAGGATCC**3'**

900bp Sub VII

5'CTGTCTGGAAGATGCGTTAAATGATTTGTTTATCCCTGAAACCACAATAGAGA
CGATACTCAAACGATTAACCATCAAAAAAAATATTATCCTCCAGGGGCCGCCCG
GCGTTGGAAAAACCTTTGTTGCACGCCGTCTGGCTTACTTGCTGACAGGAGAA
AAGGCTCCGCAACGCGTCAATATGGTTCAGTTCCATCAATCTTATAGCTATGAG
GATTTTATACAGGGCTATCGTCCGAATGGCGTCGGCTTCCGACGTAAAGACGG
CATATTTTACAATTTTTGTCAGCAAGCTAAAGAGCAGCCAGAGAAAAAGTATATT
TTTATTATAGATGAAATCAATCGTGCCAATCTCAGTAAAGTATTTGGCGAAGTGA
TGATGTTAATGGAACATGATAAACGAGGTGAAAACTGGTCTGTTCCCCTAACCT
ACTCCGAAAACGATGAAGAACGATTCTATGTCCCGGAGAATGTTTATATCATCG
GTTTAATGAATACTGCCGATCGCTCTCTGGCCGTTGTTGACTATGCCCTACGCA
GACGATTTTCTTTCATAGATATTGAGCCAGGTTTTGATACACCACAGTTCCGGA
ATTTTTTACTGAATAAAAAAGCAGAACCTTCATTTGTTGAGTCTTTATGCCAAAA
AATGAACGAGTTGAACCAGGAAATCAGCAAAGAGGCCACTATCCTTGGGAAAG
GATTCCGCATTGGGCATAGTTACTTCTGCTGTGGGTTGGAAGATGGCACCTCT
CCGGATACGCAATGGCTTAATGAAATTGTGATGACGGATATCGCCCCTTTACTC
GAAGAATATTTCTTTGATGACCCCTATAAACAACAGAAATGGACCAACAAATTAT
TAGGGGACTCATCGGGATCCCATCATCATCATCATCAT**3'**

235bp Sub I

5'GGTGATGTACGAAGAGGAGTTCACCATAATCAACGCCGTTTGCGACCGGCTT
ACCAAGGACGCGAACGCGAAGGTGGTCTTCCTCGTCGACAAGAACGGGCAGC
TCATCTCCTCCGCGGGTCAGACGCAGAACATCGACACCACGTCACTGGCCTC
GCTGACGGCCGGTAACGTGGCCGCGATGGGTGGCCTGGCCAAGCTGATTAG
GGAGAACGAGTTCCCCAACCAGTTCCAC3'

References: -

1. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome Biol.* **1,** REVIEWS001 (2000).
2. Steitz, T. A. Structural studies of protein–nucleic acid interaction: the sources of sequence-specific binding. *Q. Rev. Biophys.* **23,** 205 (1990).
3. Yonezawa, M., Doi, N., Kawahashi, Y., Higashinakagawa, T. & Yanagawa, H. DNA display for in vitro selection of diverse peptide libraries. *Nucleic Acids Res.* **31,** 118e–118 (2003).
4. Seeman, N. C., Rosenberg, J. M. & Rich, a. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. U. S. A.* **73,** 804–808 (1976).
5. Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. *Nature* **461,** 1248–1253 (2009).
6. Sathyapriya, R. & Vishveshwara, S. Interaction of DNA with clusters of amino acids in proteins. *Nucleic Acids Res.* **32,** 4109–4118 (2004).
7. Luscombe, N. M., Laskowski, R. a & Thornton, J. M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* **29,** 2860–2874 (2001).
8. Rohs, R. *et al.* Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **79,** 233–69 (2010).
9. Loenen, W. A. M., Dryden, D. T. F., Raleigh, E. A., Wilson, G. G. & Murrayy, N. E. Highlights of the DNA cutters: A short history of the restriction enzymes. *Nucleic Acids Res.* **42,** (2014).
10. Wilson, G. G. & Murray, N. E. Restriction and modification systems. *Annu. Rev. Genet.* **25,** 585–627 (1991).
11. Bickle, T. A. & Krüger, D. H. Biology of DNA restriction. *Microbiol. Rev.* **57,** 434–50 (1993).
12. Loenen, W. A. M. & Raleigh, E. A. The other face of restriction: Modification-dependent enzymes. *Nucleic Acids Research* **42,** 56–69 (2014).
13. Raleigh, E. A. & Wilson, G. Escherichia coli K-12 restricts DNA containing 5-

methylcytosine. *Proc. Natl. Acad. Sci. U. S. A.* **83,** 9070–4 (1986).

14. Raleigh, E. A. Organization and function of the mcrBC genes of Escherichia coli K-12. *Mol. Microbiol.* **6,** 1079–1086 (1992).
15. Gast, F.-U. *et al.* The Recognition of Methylated DNA by the GTP-Dependent Restriction Endonuclease McrBC Resides in the N-Terminal Domain of Mcr B. *Biol. Chem.* **378,** 975–982 (1997).
16. Pieper, U., Schweitzer, T., Groll, D. H., Gast, F.-U. & Pingoud, A. The GTP-binding Domain of McrB: More Than Just a Variation on a Common Theme? *J. Mol. Biol.* **292,** 547–556 (1999).
17. and, U. P. & Pingoud*, A. A Mutational Analysis of the PD...D/EXK Motif Suggests That McrC Harbors the Catalytic Center for DNA Cleavage by the GTP-Dependent Restriction Enzyme McrBC from Escherichia coli†. (2002). doi:10.1021/BI0156862
18. Panne, D., Müller, S. A., Wirtz, S., Engel, A. & Bickle, T. A. The McrBCx restriction endonuclease assembles into a ring structure in the presence of G nucleotides. *EMBO J.* **20,** 3210–3217 (2001).
19. Sutherland, E., Coe, L. & Raleigh, E. A. McrBC: a multisubunit GTP-dependent restriction endonuclease. *J. Mol. Biol.* **225,** 327–348 (1992).
20. Stewart, F. J. & Raleigh, E. A. Dependence of McrBC cleavage on distance between recognition elements. *Biol. Chem.* **379,** 611–6 (1998).
21. Fouse, S. D., Nagarajan, R. P. & Costello, J. F. Genome-scale DNA methylation analysis. *Epigenomics* **2,** 105–117 (2010).
22. Roberts, R. J. & Cheng, X. Base flipping. *Annu. Rev. Biochem.* **67,** 181–98 (1998).
23. Cheng, X. & Roberts, R. J. AdoMet-dependent methylation, DNA methyltransferases and base flipping. *Nucleic Acids Res.* **29,** 3784–3795 (2001).
24. Cheng, X., Kumar, S., Klimasauskas, S. & Roberts, R. J. Crystal structure of the HhaI DNA methyltransferase. *Cold Spring Harb. Symp. Quant. Biol.* **58,** 331–8 (1993).
25. Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y. & Shirakawa, M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* **455,** 818–821 (2008).
26. Rajakumara, E. *et al.* 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,** 137–52 (2011).

27. Sukackaite, R., Grazulis, S., Tamulaitis, G. & Siksnys, V. The recognition domain of the methyl-specific endonuclease McrBC flips out 5-methylcytosine. *Nucleic Acids Res.* **40,** 7552–7562 (2012).
28. Reinisch, K. M., Chen, L., Verdine, G. L. & Lipscomb, W. N. The crystal structure of HaeIII methyltransferase convalently complexed to DNA: an extrahelical cytosine and rearranged base pairing. *Cell* **82,** 143–53 (1995).
29. Cheng, X., Kumar, S., Posfai, J., Pflugrath, J. W. & Roberts, R. J. Crystal structure of the HhaI DNA methyltransferase complexed with S-adenosyl-L-methionine. *Cell* **74,** 299–307 (1993).
30. Stewart, F. J., Panne, D., Bickle, T. A. & Raleigh, E. A. Methyl-specific DNA binding by McrBC, a modification-dependent restriction enzyme. *J. Mol. Biol.* **298,** 611–622 (2000).
31. Gast, F. U. *et al.* The recognition of methylated DNA by the GTP-dependent restriction endonuclease McrBC resides in the N-terminal domain of McrB. *Biol. Chem.* **378,** 975–82 (1997).