Stony Brook University

Escherichia coli McrA: Construction of Recombinant Forms of McrA and Study of Its Binding to Methylated DNA for Use as a Tool in Epigenetic Studies

A Dissertation Presented

by

Elizabeth A. Mulligan

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Molecular Genetics and Microbiology

Stony Brook University

December 2009

Stony Brook University

The Graduate School

Elizabeth A. Mulligan

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Dr. John J. Dunn, PhD - Dissertation Advisor
Senior Scientist, Biology Department, Brookhaven National Laboratory

Dr. Janet Hearing, PhD - Chairperson of Defense
Associate Professor, Department of Molecular Genetics and Microbiology

Dr. Carl Anderson, PhD
Senior Geneticist, Biology Department, Brookhaven National Laboratory

Dr. Nancy Reich, PhD
Professor, Department of Molecular Genetics and Microbiology

Dr. Wei-Xing Zong, PhD
Assistant Professor, Department of Molecular Genetics and Microbiology

Dr. Peter Gergen, PhD
Department of Biochemistry and Cell Biology, Stony Brook University

This Dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Escherichia coli McrA: Construction of Recombinant Forms of McrA and Study of Its Binding to Methylated DNA for Use as a Tool in Epigenetic Studies

By

Elizabeth A. Mulligan

Doctor of Philosophy

in

Molecular Genetics and Microbiology

Stony Brook University

2009

Epigenetic research has become increasingly important to the understanding of cancer. Changes in the methylation status of cytosines within the CpG islands of tumor suppressor genes and oncogenes have been linked to human cancers. Methods for determining the methylation status of individual cytosines within the CpG islands of genes affecting oncogenesis are becoming an important area of study. Current techniques often involve the use of sodium bisulfate to differentiate between methylated and unmethylted cytosine, however this technique is not practical for genome-wide analysis of cytosine methylation. The *Escherichia coli* McrA protein is a putative m^5C-specific nuclease known to bind HpaII methylated DNA sequences (Cm^5CGG). It is a potential tool for epigenetic studies determining the methylation status of CpG islands, but its precise

recognition sequence has remained undefined. Our research focused on developing recombinant forms of McrA able to enrich methylated DNA fragments in affinity purification. We cloned and characterized a recombinant McrA (rMcrA) protein, which is predicted to contain a Cys_4-Zn^{++} finger and a catalytically important histidine triad in its putative nuclease domain. These features allow rMcrA to bind to several metal chelate resins without addition of a poly-histidine affinity tag. This observation was used to develop an efficient protocol for the rapid purification of nearly homogeneous rMcrA. The native protein is a dimer with a high α-helical content as measured by circular dichroism analysis. To aid in determining McrA's binding specificity we cloned and expressed recombinant McrA with a C-terminal StrepII tag (rMcrA-S) to aid in protein purification and affinity capture of human DNA fragments with m^5C residues. Sequence analysis of a subset of these fragments, and electrophoretic mobility shift assays with model methylated and unmethylated oligonucleotides suggest that the canonical binding site for rMcrA-S is N(Y>R) m^5CGR. In addition to binding symmetrically HpaII methylated double-stranded DNA, rMcrA-S binds DNA containing a single, hemimethylated HpaII site; however, it does not bind mismatches of A, C, T, or U opposite the m^5C residue but does bind if I is opposite the m^5C. These observations may lead to development of rMcrA-S-based m^5C detection assays independent of bisulfite modification.

Dedication

This dissertation is dedicated, in memory of my friend and fellow doctoral student Teresa Haire, whose warm personality, and friendship I miss every day.

Table of Contents

List of Abbreviations ix

Table of Figures xi

Table of Tables xii

Chapter 1: Introduction 1

Cytosine Methylation 2

CpG Islands 4

Methylcytosine Across Taxa 4

DNA Methylation and Disease 5

DNA Hypermethylation and Hypomethylation 6

Loss of Imprinting 7

Detection of Cytosine Methylation 8

Bacterial Restriction 10

McrA 10

Figures 12

Materials and Methods for Chapter 3 17

Materials 17

Cloning of McrA 17

McrA expression in *E. coli* 18

rMcrA purification 19

Analytical Methods 20

Analytical size exclusion chromatography 20

Mass determination 20

Circular dichroism 21

In vitro activity assays 21

In vitro packaging 22

Nuclease Activity of rMcrA 22

McrA Induction Time Course 22
Materials and Methods for Chapter 4 23
Oligonucleotides 23
Enzymes and plasmids 24
rMcrA-S 24
rMcrA-S affinity enrichment of mCGI's 25
Ligation mediated PCR (LM-PCR) 26
Bisulfite Sequencing 26
In vitro Binding Assay 27
Chapter 3: Cloning, Purification and Initial Characterization of *E. coli* McrA 29
Introduction 30
Results and Discussion 33
Expression of recombinant McrA in *E. coli* 33
rMcrA selectively binds HpaII methylated DNA 35
rMcrA does not appear to be a m^5C specific nuclease 37
Chapter 4: Defining the Binding Site of McrA 47
Introduction 48
Results 50
rMcrA-S binds to Methylated Human DNA 50
rMcrA-S binds Cm^5CGG and m^5CGG 53
rMcrA-S selectively binds N(Y>R) m^5CGR 54
rMcrA-S also binds Hemi-Methylated DNA 55
Discussion 56
Tables and Figures 58
Chapter 5: Summary and Future Directions 68
Summary 69
Future Directions 71

rMcrA-S Based Assay for CpG Methylation Identification 72
Nuclease Activity of rMcrA 73
References 75
Appendix 85

List of Abbreviations

m^5C – C^5-methylcytosine

m^5CpG – Methylated Cytosine-Guanosine dinucleotide

CD – Circular Dichroism

CpG – Cytosine-Guanosine dinucleotide

SDS-PAGE – Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis

RT – Room Temperature

ORF – Open Reading Frame

YT – Yeast-Tryptone Medium

PBS – Phosphate Buffered Saline

Kan – Kanamycin

CAM – Chloramphenicol

EMSA – Electrophoretic Mobility Shift Assay

DNMT's – DNA (Cytosine-5) Methyltransferases

DNA – Deoxyribonucleic Acid

CGI – CpG Island

TSS – Transcriptional Start Site

PWS – Prader-Willi Syndrome

AS – Angelmen Syndrome

rgl – Restricts Glucose-Less Phage

Mcr – Methyl Cytosine Restriction

M.HpaII – HpaII Methylase

M.SssI – CpG Methylase

SAM – S-Adenosyl-L-Methionine

TOF-MS – Time of Flight Mass Spectrometry

Table of Figures

Figure 1.1 Cytosine Methylation……..12

Figure 1.2 Cytosine Deamination…….13

Figure 1.3 DNA Methylation Effects……14

Figure 1.4 Bisulfite Modification……...15

Figure 3.1 Purification of McrA……43

Figure 3.2 Analytical Tests of McrA…….44

Figure 3.3 McrA Binding Activity……45

Figure 3.4 Time Course after Induction of McrA……...46

Figure 4.1 CpG ratios in the Human Genome vs. rMcrA-s Pull-downs…….62

Figure 4.2 Bisulfite Sequencing Patterns…….63

Figure 4.3 Binding Assays for rMcrA-S……...64

Figure 4.4 rMcrA-S binds N(Y>R) m5CGR Sequences.....65

Figure 4.5 rMcrA-S binding to Fully- and Hemi-methylated Cm^{5}CGG Sites……66

Figure 4.6 rMcrA-S Binding to DNA Sequences with Mismatch Opposite m^{5}C...67

Figure 5.1 Sequence Specific McrA Detection Diagnostic CpG's……74

Table of Tables

Table 3.1 Secondary Structure Analysis of rMcrA from CD Spectrum……...40

Table 3.2 Biological Properties of rMcrA……41

Table 3.3 Summary of McrA and HpaII in BL21(DE3)……..42

Table 4.1 HpaII sites in Human DNA……..58

Table 4.2 List of Oligonucleotides……59

Table 4.3 Fully Base-paired Oligonucleotide Cassettes……60

Table 4.4 Heteroduplex Cassettes……61

Acknowledgments

First and foremost I would like to thank my dissertation advisor Dr. John Dunn, for his support, patience and guidance throughout my graduate studies. I will always be thankful for his help and support in pointing me in new research direction after my first project crashed and burned. He has given me endless advice which has been invaluable in my research, and has helped to shape the scientist I am today. His intelligence, integrity, and understanding made him an excellent scientist and mentor.

I sincerely want to thank my committee members, Dr. Janet Hearing, Dr. Nancy Reich, Dr. Carl Anderson Dr. Wei-Xing Zong, and Dr. Peter Gergen for all their helpful advice and guidance with my project. I would like to thank my committee for always being supportive of my project and the twist and turns it has taken over the course of the last few years.

To all the members of the Dunn lab, Laurali, Judi, and Barbara for all their technical support and for making the lab an enjoyable place to work. Special thanks to Laurali for her ability to write a paragraph of information legibly on the top of an eppendorf tube. Also thanks go to Dr. Bill Studier and Eileen Matz for all their help with the pREX cloning.

To all my friends at Stony Brook University, thank you for making graduate school enjoyable. I am indebted to Betty for all the hours we spent in the library our first couple of years, especially our marathon sessions in the library the day before a test. I would have never survived the classes and the qualifier without her. I need to thank Mary for all her advice on life, and shoes. Thanks to Lindsay, Erin, Kasey, and Kate for always being there, keeping me in the Stony Brook loop even though I was way out east at BNL, and for all the pie parties.

I must also thank my family, especially my parents. My mom and dad have been there with unending professional, emotional, and financial support. They always believed in me and always made sure I kept the big picture in mind. I cannot thank them enough for all the love and support they gave me.

Chapter 1: Introduction

Gene expression within a cell is not static; rather it is a dynamic process that is influenced by tissue type, cell cycle and the microenvironment of a particular cell. With the human genome sequenced [1] and the number of human genes largely determined, new research is aimed at how gene expression is modulated. One focus of this research is in the field of epigenetics, which a recent consensus definition stated, "an epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence [2]." In addition epigenetic modifications of chromosomes are reversible and epigenetic patterns are reproduced during cellular mitosis and meiosis. These modifications include cytosine methylation, methylation and acetylation of histone proteins. The most common type of epigenetic modification in the mammalian genome is cytosine methylation of CpG dinucleotides which influences the structure of chromatin, inactivation of the X-chromosome in females, genomic imprinting, and affects embryonic development, the expression of genes in a tissue specific manner, and the timing of replication within a cell [3-11].

Cytosine Methylation

DNA methylation is found in both prokaryotic and eukaryotic cells, but while methylation can occur at the N6 position of adenine or the N4 or C5 position of cytosine, only C5 methylation of cytosine is found in mammalian cells. The pattern of CpG methylation is not only different species to species, but also tissue type to tissue type within a given organism. Methylation of the 5 position of the cytosine base at CpG dinucleotides is carried out by DNA (Cytosine-5) Methyltransferases (DNMT's) which catalyze the transfer of a methyl group from S-adenosyl-L-methionine (SAM) to the 5 position of cytosine

(Fig. 1.1). DNMT3a and DNMT3b are the *de novo* methyltransferases while in contrast DNMT1 is the maintenance methyltransferase which methylates the daughter strands of newly replicated DNA [12]. The protein DNMT2 exists in mammals but does not have any detectable methyltransferase activity [13].

Statistically CpG dinucleotides are underrepresented in the human genome and when they occur in coding regions they are normally modified at the cytosine base to 5-methylcytosine (m^5CpG) and become hotspots for mutation [14-17]. Unmethylated cytosine can deaminate to uracil which is an abnormal base in DNA, thus the DNA repair machinery of the cell uses the opposite strand guanine as the template to replace the incorrect uracil base with cytosine. In contrast 5-methylcytosine can spontaneously deaminate to thymine which is problematic for the DNA repair mechanisms since both guanine and thymine, while mismatched, are normal bases in DNA, and it is not readily apparent which base should be used as the template (Fig. 1.2). This T:G mismatch can be repaired by the enzyme Thymine DNA glycosylase (TDG). This enzyme which can also repair U:G mismatches is needed to prevent the mutations that can occur with 5-methylcytosine deamination [18-19]. Even with the existence of a repair enzyme to repair deaminations, 5-methylcytosine seems to be evolutionarily selected against accounting for only 0.75 – 1% of the total bases in the human genome or about 4% of all cytosines and seems to be an evolutionary consequence of this ineffective and mutation-prone DNA repair at spontaneously deaminated 5-methylcytosines [20]. Since 5-methylcytosine deaminates to thymine and can create mutations it is often found in non-coding regions near transcriptional start sites or in endogenous repeats and transposable elements [14-15].

CpG Islands

As stated above, in mammalian cells, CpG dinucleotides are statistically deficient. They are also asymmetrically represented in the genome, being underrepresented in coding regions and overrepresented in or near promoter regions. Clusters of CpG dinucleotides are termed CpG islands (CGIs). CGIs are normally associated with the promoter regions of genes and are unmethylated [21-23]. CpG islands are defined as regions of DNA that are ≥ 500 bp and have greater than 50% G+C base composition and a CpG [observed/expected] of more than 0.6 [22, 24-25]. The human genome contains ~30,000 CGIs which accounts for about 10% of the total DNA and about half of these islands are found near annotated transcriptional start sites (TSS); the remainder being intra- or intergenic (non-TSS). However, as pointed out by Bird and co-workers, several non-TSS CGIs have been shown to coincide with previously unforeseen but functional promoters raising the possibility that all CGIs function as promoters and are therefore TSS-associated [24]. In normal tissues CGIs are usually unmethylated but a subset (10's to 100's) becomes reproducibly methylated in normal cells (imprinting, X-chromosome inactivation, tissue differentiation) or in diseased cells such as cancer cells [26-27]. The majority of CpG dinucleotides, accounting for 70-80% of all CpG's, are outside of these CGIs and are normally methylated [28].

Methylcytosine Across Taxa

The use of DNA methylation to silence genes and as a mechanism to control transposable elements is well established in both plants and vertebrates. The preceding sections detail methylcytosine as a mechanism for gene silencing in mammals. This mechanism is seen in other taxa as well. The loss of DNA

methylation in the regions of gene promoters has been shown to have an effect on apoptosis in both *Xenopus* [29] and mice [30]. Additionally the loss of DNA methylation in mice affects X-chromosome inactivation, and chromosome stability [31], while in *Arabidopsis* it affects overall chromosome organization [32].

The genomes of *D. melanogaster* and *S. cerevisiae,* in contrast to vertebrate genomes, are deficient in m^5C residues [33-34]. In 1999 however, Hung and colleagues reported that the *D. melanogaster* protein DmMT2, in later reports referred to as DNMT2, has high sequence homology to the mammalian methyltransferase DNMT2, and the fly protein DmMTR1 which has epitopes that are related to conserved motifs in the catalytic domain of mammalian DNMT1 and like DNMT1 it interacts with PCNA [35]. Additionally in 2000 it was reported that Drosophila did contain 5-methylcytosine, but at an amount that is 50 times lower than in mammals [36] and that DNA methylation is restricted to early embryonic developmental stages [37]. Lyko and colleagues also reported that cytosine methylation often occurred in CpT, CpA, and CpC in addition to CpG methylation, but again this methylation is restricted to embryonic stages of development [37]. Recently it has been shown that retrotransposon silencing in embryonic *Drosophila* is due to DNMT2-dependant DNA methylation [38].

DNA Methylation and Disease

The hypermethylation of CpG islands proximal to promoters of tumor suppressor genes, is associated with gene silencing, and is common in cancer [6, 15, 39-40]. This hypermethylated state contributes to a closed chromatin state which is inaccessible to transcription factors, leading to transcriptional silencing of these tumor suppressor genes. Additionally many of the non-transcriptional start site CGIs in the human genome are associated with repetitive sequences

which are typically heavily methylated in normal tissue and therefore transcriptionally silent. However, during tumorigenesis, many of these islands become hypomethylated, a state associated with open chromatin, and resulting in their expression which may help drive DNA breakage and genome instability; known hallmarks of cancer. Figure 3.3 summarizes the effects of DNA methylation in normal and tumor cells.

DNA Hypermethylation and Hypomethylation

Inactivation of tumor suppressor genes is an essential component of oncogenesis. Aberrant hypermethylation of CpG islands in tumor suppressor genes is known to cause gene silencing and is a hallmark often found in cancer. The first tumor suppressor gene described to be silenced by this mechanism was the retinoblastoma gene RB [41-44]. Methylation was discovered in a CpG island at the 5′ end of the RB gene in a retinoblastoma tumor that in normal tissue was unmethylated [42]. Numerous other tumor suppressor genes are known to be silenced in cancer cells via hypermethylation of a 5′ CpG island including $p16^{INC}$, APC and BRCA1.

CpG island hypermethylation in cancer cells is not limited to tumor suppressor genes. Genes involved in DNA repair such as MGMT (O^6-methylguanine–DNA methyltransferase) which removes alkyl groups from the O^6 position of guanine, by transferring the alkyl onto one of its own cysteine residues [45], have been shown to be methylated in tumor tissue but not normal tissue [46]. Although there are a number of genes that are repeatedly methylated in different cancer types the complete profile of hypermethylated genes is unique to each tumor type [46]. A list of genes positively identified as being aberrantly methylated in

cancer currently includes 66 genes (www.mdanderson.org: methylation in cancer).

Although hypermethylation of CpG islands has received more attention, hypomethylation of DNA was identified in cancer cells prior to DNA hypermethylation. During the 1980's Feinberg and Vogelstein discovered that significant numbers of CpG's that were methylated in normal tissue were unmethylated in cancer tissue [47]. Subsequent work showed that global hypomethylation occurred early in tumor tissues [48] and this phenomenon was seen across multiple tumor types regardless of the tumor being benign or malignant [49-50]. More recently it has been shown that there are CpG islands that are normally methylated, but become hypomethylated in cancer [51].

Hypomethylation has been seen in kidney, liver, colon, pancreatic, stomach, uterine, cervical, and lung cancers [52-59]. This hypomethylated state can lead to gene activation including the HRAS [60] and cyclin D2 which are activated in this manner in gastric cancer. In cervical cancer, hypomethylation leads to activation of HPV16 and seem to accounts for tumor latency seen in this type of cancer [61-62].

Loss of Imprinting

In addition to cancer, epigenetic modifications are known to play an important role in diseases involving loss of imprinting. The diseases Prader-Willi (PWS) and Angelmen (AS) syndromes can be used to illustrate this phenomenon. PWS is characterized by hypotonia, hypogonadism, polyphagia and mild mental delay, [63], while AS is characterized by microcephaly, seizures, jerky movements, lack of speech development, excessive laughing or smiling and

a generally happy disposition [64]. As summarized by Horsthemke and Wagstaff, the phenotypically distinct PWS and AS both arise from imprinted genes on chromosome 15q11-q13, which have distinct maternal and paternal patterns of CpG methylation, methylation of histone H3 Lys4, histone H3 Lys9 and histone h4 Lys20, and acetylation of histone H3 and H4 [65] such that certain genes within this region are only expressed on the maternal or paternal chromosome. Deletion, imprinting defect, i.e. the paternal chromosome has a maternal imprint, or mutation leads to the lack of expression of the corresponding maternal or paternal genes leading to the syndromes. Broadly, the lack of expression of the paternally imprinted genes in this region give rise to PWS [65-66] while lack of expression of the maternally imprinted genes in the same chromosomal loci results in Angelman syndrome [67].

Detection of Cytosine Methylation

CpG islands have the potential to serve as biomarkers for cancer diagnosis and prognosis and there is concentrated research in the area of detecting aberrantly methylated CpG islands for the purpose of mapping the methylation pattern within these regions [3, 68-71]. A number of different methods are used to determine the methylation status of cytosines. Often these methods utilize sodium bisulfite treatment in order to distinguish between cytosine and methylcytosine.

Bisulfite sequencing in currently the "gold standard" for determining the methylation status of individual cytosines across a region of amplified DNA and this technique is now routinely used when studying the methylation of CGIs [72]. The technique takes advantage of the fact that sodium bisulfite converts cytosine to uracil. Methylcytosines can deaminate to thymine with sodium

bisulfite treatment, but the reaction rate is much slower than that of cytosine conversion to uracil [73-74]. Following modification of the DNA, region(s) of interest are PCR amplified using primers specific to the bisulfite modified sequence. This step converts the uracils to thymine (Fig. 1.4). DNA fragments are subsequently cloned and sequenced and finally the sequence is compared to the unmodified sequence enabling interrogation of the methylation status of the cytosines of individual CpG dinucleotides.

This method, while widely utilized, has several drawbacks. Sodium bisulfite treatment is harsh, and fragments the DNA [75]. It also requires a long incubation time, high temperature and low pH in order to get complete conversion of cytosines to uracil. The conversion itself needs to be checked against an internal control of non CpG cytosines, in which all cytosines are converted to thymine after PCR, to make sure there are no false positives. This method necessitates the sequencing of multiple clones of each DNA fragment since individual CpG sites may not be methylated 100% of the time. Thus percent methylation at individual sites is often used as the parameter when reporting results. Since CGIs by definition contain numerous CpG's, it is difficult to design primers without a CpG which can either be CpG or TpG depending on methylation status of the original cytosine [76]. Additionally primer design can be problematic, especially when checking the methylation status of CGIs because after bisulfite modification few cytosine bases are left in the DNA making finding unique sequences for primer design more difficult. For these reasons new methods for determining cytosine methylation status without the use of sodium bisulfite would be valuable in methylation studies.

Bacterial Restriction

Bacterial restriction-modification (RM) systems consisting of methyltransferase and endonuclease functions can be thought of as acting as a primitive "immune system," in that they allow the bacterial cell to differentiate between "self" DNA from foreign "non-self" DNA. The simplest and most common type of these RM systems is the Type II system. In these systems a methylase acts on the bacterial cell's own DNA methylating it at a given sequence. Either a second enzyme as in the case of Type II, and IIS, or a separate domain or subunit of the same enzyme in the case of Type I, IIG, and III will cleave DNA at the same recognition site as the methylase, if the site is unmethylated. A fourth class of RM systems Type IV recognize methylated DNA as in the case of the E. coli systems McrBC, and Mrr. For a concise overview see [77].

McrA

A number of methylcytosine binding proteins have been identified and studied for their ability to be used as tools for the enrichment of methylated CpG DNA sequences [69, 78].

The *E. coli* protein Methyl Cytosine Restriction A (McrA) specifically binds to methylated m^5CG sequences and *in vivo* restricts incoming 5-methylcytosine, and 5-hydroxymethylcytosine containing DNA [79-81]. It was first described when it was found to restrict T-even phages, which lack glucosylation of their 5-hydroymethylcytosines, and termed rglA (restricts glucose-less phage). Subsequently McrA and rglA were found to be located at the same locus (25 min on the chromosome of E. coli K-12) as part of a defective lambdoid prophage element, e14 [82].

Early work showed that McrA restricted DNA methylated by M.HpaII (C m^5CGG), M.SssI (m^5CG) or M.Eco1831I (C m^5CSGG where S is C or G) [82-83]. However, by nature of their recognition sequences, M.SssI and M.Eco1831I will also methylate HpaII sites and thus the restriction could be due to McrA recognition of Cm^5CG. Previous work by our own lab, with several methylases that do not methylate HpaII sites, showed that McrA did not restrict or bind to these methylated sequences [84].

In 2004 Anton and Raleigh used transposon-mediated scanning mutagenesis to make structural predictions about McrA. They found that C-terminal truncations abolished activity of McrA, as did amino acid insertions between residues 28 and 124 [85]. It is thought that the active site of McrA is in the C-terminal region so the lack of activity from the N-terminal insertions may be because of disruption of the DNA binding domain [85]. Additionally in C-terminal insertions they found partial activity of McrA and they concluded that this was due to a lack DNA damaging activity, and that the first 130 to 149 residues of McrA are required for DNA binding [85]. The C-terminal region of the protein is thought to be the location of the active site and McrA is a putative nuclease, however, to date, the biological mechanism of restriction of methylated phage or DNA, by McrA, remains unknown.

Figures

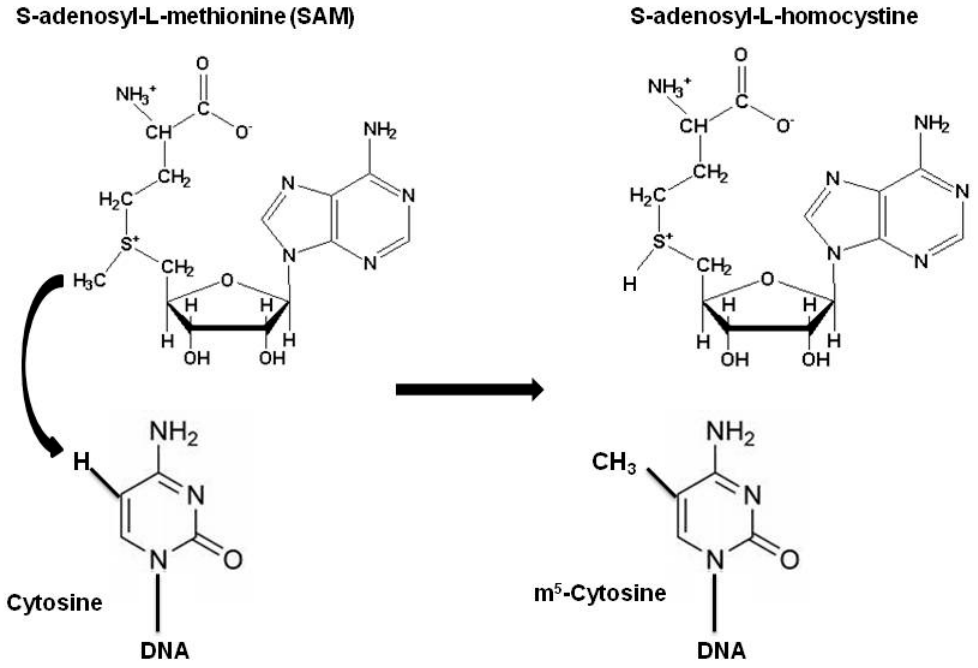


Figure 1.1 Cytosine Methylation. DNMT's catalyzes the transfer of a methyl group from S-adenosyl-L- methionine (SAM) to the 5 position of the cytosine base creating 5-methylcytosine (m^5C) and S-adenosyl-L-homocystine. Figure adapted from [86].

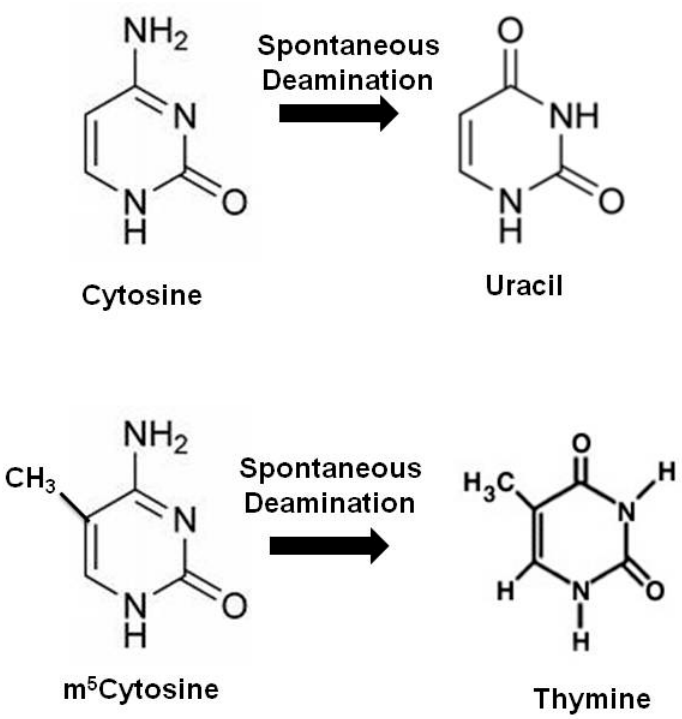


Figure 1.2 Cytosine Deamination. Cytosine deaminates to uracil and methylcytosine deaminates to thymine.

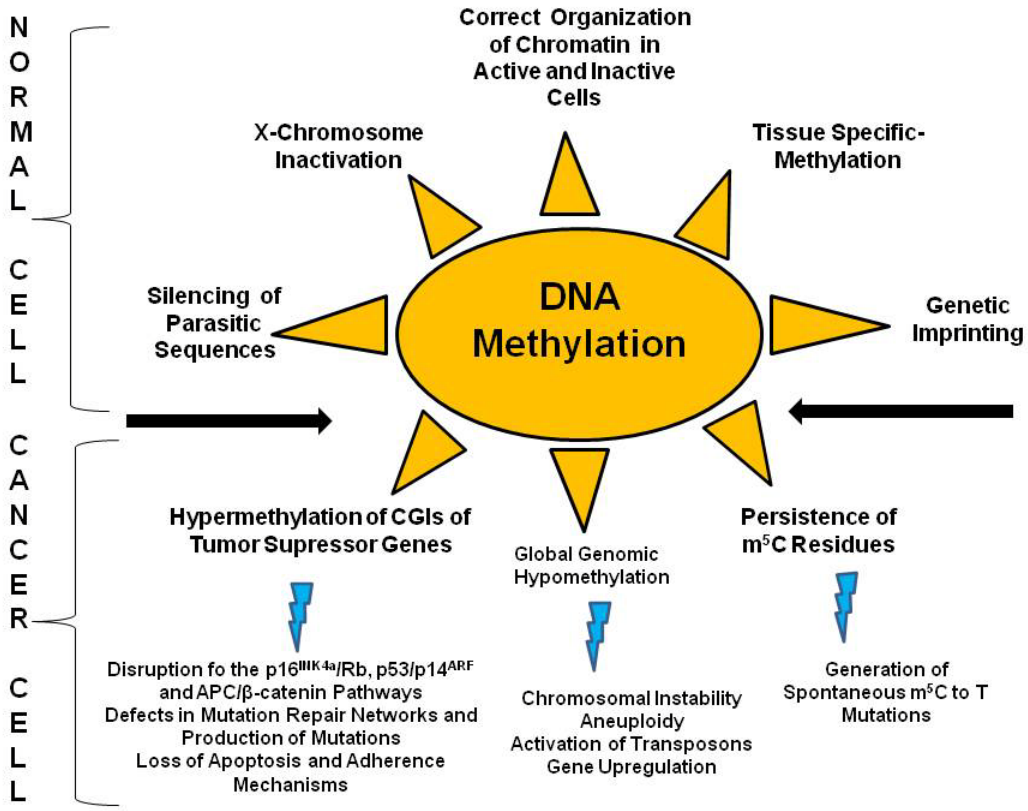


Figure 1.3 DNA Methylation Effects. The effects of DNA methylation in normal and tumor cells. Figure taken and adapted from [15].

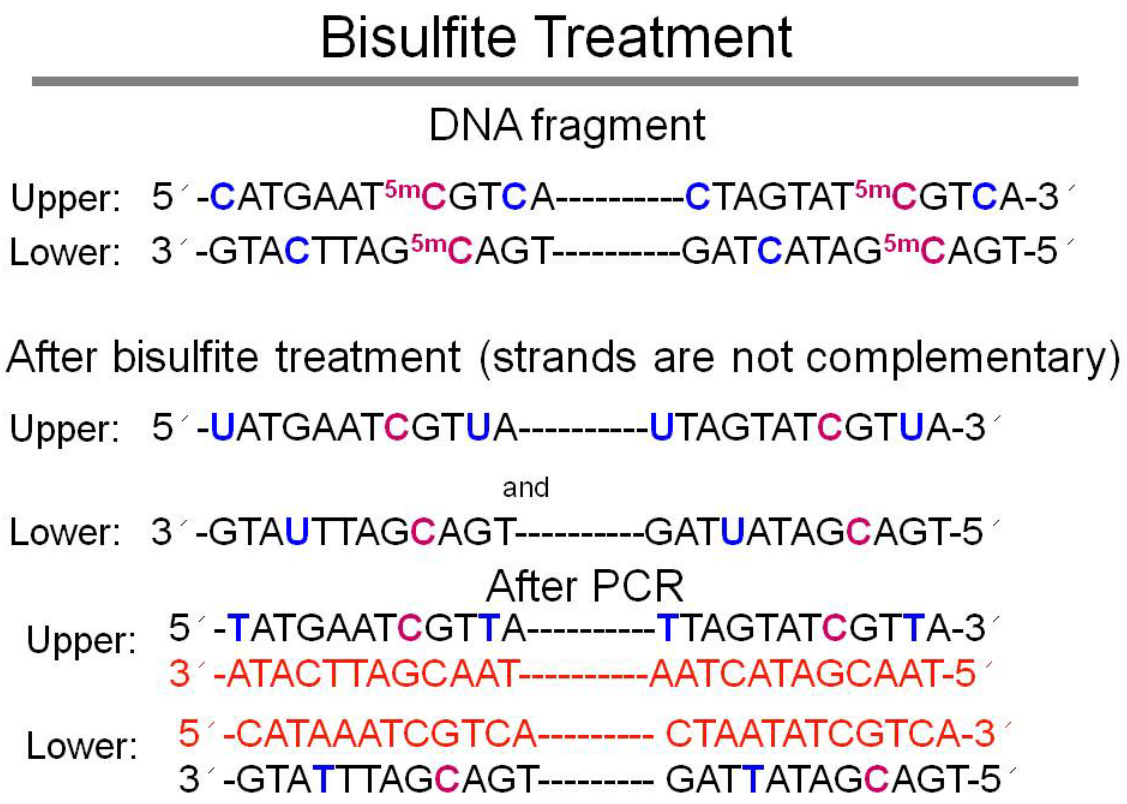


Figure 1.4 Bisulfite Modification. Following DNA modification with sodium bisulfite, the two strands of DNA are no longer complimentary and unmethylated cytosines are deaminated to uracil. During PCR the uracils are amplified as thymines.

Chapter 2: Materials and Methods

Materials and Methods for Chapter 3

Materials

All enzymes are from New England Biolabs (NEB), Ipswich, MA, USA) if not stated otherwise. Colored protein SDS-PAGE markers were purchased from Lonza Rockland, Inc, Rockland, ME. T7 DNA [87] was isolated by phenol extraction of CsCl purified virus grown in *E. coli* Bl21, a *dcm* minus host.

Cloning of McrA

E. coli K12 W3110 (mcrA^{+}) genomic DNA was used as template to PCR amplify the 833 bp [16] McrA coding sequence with McrA-F 5' GA*CGTCTC*CCATGCATGTTTTTGAT- 3' and McrA-R 5'-AGA*GGATCC*CTATTATTTCTGTAATC- 3' as the forward and reverse primers, respectively and PfuTurbo C_x Hotstart DNA polymerase (Stratagene, San Diego, CA). Bases in the primers complementary to the McrA coding sequence are underlined. McrA-F contains a unique *BsmBI* recognition sequence while McrA-R contains a unique site for *BamHI* (shown in italics) which were added to facilitate directional cloning of the digested DNA into the unique *NcoI* and *BamHI* sites of a pET28 expression plasmid. The initial amplicons were purified and blunt end ligated into pZERO-2 cut with EcoRV. Following electroporation into *E. coli* TOP10 (F- mcrA Δ(mrr-hsdRMS-mcrBC) ϕ80lacZΔM15 ΔlacX74 nupG recA1 araD139 Δ(ara-leu)7697 galE15 galK16 rpsL(Str^{R}) endA1 λ^{-}) cells (Stratagene) several pZERO-McrA Kanamyacin resistance (Kan^{R}) plasmids were purified and the correct sequence of the inserts verified by DNA sequencing using standard M13 primers flanking the cloning site. One of the correct recombinant plasmids was digested with *BsmBI* and *BamHI* to release a full-

length McrA fragment having ends compatible with those of *NcoI* + *BamHI* digested pET28. Following gel purification and elution, the McrA containing fragment was directionally cloned into pET28, previously digested with *NcoI* and *BamHI*, using standard ligation conditions and subsequently electroporated into Top10 cells. The resulting pET-rMcrA plasmid DNA was used to transform BL21(DE3) (F^- ompT gal dcm lon $hsdS_B(r_{B^-}\ m_{B^-})$ λ(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5]) cells with and without the chloramphenicol resistant pACYC-RIL tRNA encoding plasmid. All transformants were plated on 2xYT agar plates supplemented with 50 μg/ml Kan or with 50 μg/ml Kan and 25 μg/ml chloramphenicol as needed. The pRIL tRNA plasmid*: argU*(AGA,AGG), *ileY*(AUA) and *leuW*(CUA), was isolated from Stratagene BL21-CodonPlus *McrA*+ cells (Strategies Newsletter 14.2, p.50–53) and then moved into BL21(DE3) to place the plasmid in a *McrA* minus background. This BL21(DE3) mcr^- /$pRIL^+$ strain was kindly provided by F.W. Studier (BNL). See Appendix for pET28-McrA map and sequence.

McrA expression in *E. coli*

Expression of McrA was initially tested in BL21(DE3) with and without the pRIL tRNA plasmid by addition of 0.5 mM IPTG to mid-log phase cells in 2xYT medium or by growth to saturation in ZYM 5052 autoinduction medium at 37 °C and 20 °C [21]. These media contained 100 μg/ml Kan and 25 μg/ml chloramphenicol as needed and were supplemented with 30 μM $ZnS0_4$. Induction and solubility of the expressed McrA protein was followed by SDS-PAGE. Expression of rMcrA was increased ≥ 5-fold (data not shown) by the presence of the pRIL plasmid. This strain was used for all future experiments. Recombinant McrA expressed at 37 °C is insoluble but a sizeable fraction is

soluble if expression is done at 20 °C. We therefore chose to induce protein expression at 20 °C and for ease we also used autoinducing conditions [88]. Shaking cultures (100 to 200 ml in 500 or 1000 ml flasks) were initially started at 37 °C by addition of 1 ml of an overnight culture in 2xYT and moved to 20 °C once growth became visible. Cells were harvested after 48 h by centrifugation at 5,000g for 10min at 4 °C, washed with 1/10 vol. Phosphate Buffered Saline (PBS), recentrifuged and the cell pellets (~1.2 g/50 ml culture) stored frozen at -20 °C.

rMcrA purification

Cell pellets were thawed and resuspended in 10 ml of LEW buffer (50 mM NaPO4, pH 8.0; 300 mM NaCl). Lysozyme (20 mg/ml) was added to a final concentration of 100 μg/ml and the cells were frozen and thawed 3x using a dry ice - ethanol slurry to promote lysis. The extract was sonicated 4-6 times in 30 sec bursts alternated with chilling on ice to reduce viscosity and then centrifuged at 10,000 rpm for 10 min at 4 °C. The pellet was resuspended in 5 ml of LEW and recentrifuged at 10,000 rpm for 10 min at 4 °C. The soluble fractions were pooled and mixed for 30 min at 4 °C with 500 mg of PrepEase High-Yield Ni-chelate resin (USB #78806) pre-equilibrated with LEW. The resin was pelleted by centrifugation (5,000 rpm for 5 min) and batch washed consecutively with 15 ml LEW, LEW + 700 mM NaCl (1 M final NaCl concentration), and LEW + 2 mM imidazole before pouring into a small 1 cm i.d. column. The settled resin was washed with 25 ml of LEW + 2 mM imidazole then with LEW + 100 mM imidazole to elute rMcrA in 2 ml fractions. All chromatography steps were carried out at RT. Peak fractions were pooled, diluted with an equal volume of 10 mM $NaPO_4$, pH 8.0, and passed through a 5-ml bed of SP-Sepharose Fast Flow

(Pharmacia Biotech, Uppsala, Sweden) pre-equilibrated with 0.5 x LEW. The 1 cm i.d. column was washed with ~25 ml 0.5 x LEW and the bound rMcrA eluted with LEW. Peak fractions were pooled and stored at 4 °C. At this stage the protein was ≥99% pure as estimated by Coomassie Blue staining following SDS-PAGE (Fig. 3.1B).

Analytical Methods

Analytical size exclusion chromatography

Analytical size exclusion chromatography (20 μl injection) was performed on a TSK-GEL G3000SW$_{xl}$ (7.8 mm ID cm x 30 cm,TosoHaas) column equilibrated in 25 mM MES, 300 mM NaCl, pH 6.5. The column was calibrated with thyroglobulin (669 kDa), apoferritin (443 kDa), β-amylase (200 kDa), alcohol dehydrogenase (150 kDa), bovine serum albumin (66 kDa), ovalbumin (45 kDa), carbonic anhydrase (29 kDa) and sperm whale myoglobin (17.8 kDa) in the same buffer. The size of rMcrA was determined from its elution time relative to those of the protein standards plotted against the logarithm of their molecular weights.

Mass determination

Time of flight mass spectrometry (TOF-MS) was used to determine the molecular mass of McrA. Samples were desalted and mixed (1:1 dilution) with the matrix sinapinic acid (10 mg/ml in 50% CH_3CN, 0.3% TFA) immediately prior to analysis on a Perseptive Biosystems Voyager-DE Linear Mass Spectrometer.

Circular dichroism

Far UV circular dichroism spectra (CD) of rMcrA (180-280 nm) was measured at NSLS beamline U11 using a quartz cell with a 20 μ path length. rMcrA was 1.6 mg/ml in 50 mM HEPES, 500 mM NaCl, pH 7.2 and the data were corrected by subtraction of a blank spectrum obtained using only buffer. The secondary structure content of rMcrA from the CD spectrum was calculated using the software analysis program CDSSTR from DICHROWEB [89-90].

***In vitro* activity assays**

The ability of purified rMcrA to digest or bind to methyl-cytosine containing DNA was assayed using restriction fragments of unmethylated T7 DNA or DNA incubated with different methyltransferases and S-adenosylmethionine as methyl donor using conditions provided by the supplier. Completeness of the methylation step was confirmed by testing the modified DNA's resistance to digestion with the appropriate restriction enzymes as well as acquired sensitivity, where appropriate, to digestion by McrBC. Digestion was monitored by agarose electrophoresis. These same conditions were used for electrophoretic mobility shift assays (EMSA). EMSA reactions (10 μl) typically contained a mixture of 50 ng NarI digested unmethylated T7 DNA, 50 ng *KpnI* digested *HpaII* methylated T7 DNA, 100 μg/ml BSA and 250 ng sonicated *E. coli* ER2925 as non-specific competitor in various binding buffers with and without added Mg^{2+}. Reactions were started by adding 1 μl of rMcrA diluted in the appropriate binding buffer plus 100 μg/ml BSA. After 15-20 min at RT the samples were loaded on 0.7 % agarose Tris-borate/EDTA (1xTBE; 89 mM Tris, 89 mM boric acid and 2 mM EDTA). gels which were electrophoresed at RT. DNA

was detected by staining with ethidium bromide (0.5 μg/ml) and illumination with UV light.

***In vitro* packaging**

Unmodified and HpaII methylated T7 DNAs were packaged into virions *in vitro* using an extract obtained from Novagen (San Diego, CA) using 1 μg DNA/25 μl extract for 30 min at room temperature. Reactions were stopped by dilution with 2xYT medium and stored at 4 °C. Cells used as indicator were grown overnight at 37 °C in non-inducing 2xYT medium supplemented with antibiotics as required.

Nuclease Activity of rMcrA

BL21(DE3) cells containing a combination of the following plasmisds, pET28-McrA, pACYC-M.HpaII, pET-OspC, and pACYC-RIL were plated onto 2xYT plates supplemented with 50 μg/ml kanamycin and 25 μg/ml chloramphenicol and grown for 16 hr at 37 °C and then colonies were counted.

McrA Induction Time Course

Bl21-AI cells containing the plasmids pREXLS31-McrA and pACYC-M.HpaII were grown in 2xYT supplemented with 50 μg/ml kanamycin and 25 μg/ml chloramphenicol and grown to log phase and then McrA was induced by the addition of 0.01M IPTG and 0.05% arabanose (final concentrations). Cells were collected at 0 min, 15 min, 30 min, 1 hr, 2 hrs, 3 hrs, and 4 hrs post induction. Total DNA was extracted over a column (Qiagen, Valencia, CA) and run on a 1% agarose gel.

Materials and Methods for Chapter 4

Oligonucleotides

All oligonucleotides were purchased from Integrated DNA Technologies, Coralville, Iowa. Individual oligonucleotides, with and without m^5C, were dissolved in 10 mM Tris-0.1 mM EDTA buffer, pH 7.5 and annealed at 36 μM with their complements as needed in 1X One-Phor-All Buffer (10 mM Tris-Acetate pH 7.5, 10 mM Mg- Acetate, 50 mM K-Acetate) (GE Healthcare, Piscataway, NJ) by heating for 2 min at 98ºC followed by slow cooling to room temperature and 10% PAGE analysis to verify their complete conversion to duplexes with minimal amounts (<5%) of residual single-stranded oligonucleotides. Table 4.2 lists the oligonucleotides used for this study and Tables 4.3 and 4.4 gives the sequences of their resulting fully base paired ds-cassettes. Oligonucleotide cassettes used to determine binding of rMcrA-S to ds-cassettes with a mismatched base opposite m^5C are shown in Table 3. Oligonucleotides 1-6 were synthesized with m^5C at various positions, while the remaining oligonucleotides were methylated post-synthesis when needed using the ds-specific CpG methylase (M.SssI) after annealing to form duplexes as described above. A typical reaction (20 μl) included 180 pmol duplex DNA in SssI buffer (10 mM Tris-HCl, 50 mM NaCl, 10 mM $MgCl_2$, and 1 mM dithiothreitol) supplemented with 160 μM SAM. Reactions were started with the addition of 8 units of M.SssI enzyme and allowed to incubate overnight at 37°C, followed by spiking with additional SAM for a final concentration of 320 μM, and incubated at 37°C for an another 2 hrs. M.SssI was inactivated by heating at 65°C for 20 min. To check that the methylation was complete, a duplex containing a single *HpaII* (CCGG) site, and no other CpG dinucleotides, was included as a control. The DNA was then checked for protection against

digestion by *HpaII* (methyl sensitive) but susceptibility to digestion by the methyl insensitive isoschiztomer *MspI*. For each assay, 2 μl of methylation reaction was added to 10 mM Bis-Tris-Propane-HCl, 10 mM $MgCl_2$, 1 mM dithiothreitol supplemented with 100 μg/ml Bovine Serum Albumen (BSA), followed by addition of 10 units of either *HpaII* or *MspI* enzyme. Reactions were incubated at 37°C for 2 hrs before being analyzed by agarose gel electrophoresis. Gels were stained with ethidium bromide and DNA bands visualized under UV light.

Enzymes and plasmids

All enzymes, unless otherwise stated were from New England Biolabs (NEB), Ipswich MA, USA. T7-based pET vectors were from our own collection.

rMcrA-S

The 290 amino acid McrA coding sequence with either an 8 amino acid N-terminal or C-terminal StrepII tag [17] (underlined) (MASWSHPQFEKGA-***start of McrA*** or ***end of McrA***–SAWSHPQFEK, respectively) were constructed by PCR amplification from a wild-type, untagged *mcr*A clone using appropriately designed primers which appended the StrepII tag and unique restriction sites to aid in cloning on the ends of the PCR product. After restriction enzyme digestion and gel purification the amplicons were cloned into pET28. See Appendix for maps and sequences of these constructions. The accuracy of the clones were verified by DNA sequencing using primers flanking the cloning sites and then moved into the expression host BL21(DE3)/pRIL and the recombinant proteins were expressed following autoinduction at 20 °C as previously described except that after SP-Sepharose Fast Flow (Pharmacia Biotech, Uppsala, Sweden)

chromatography the recombinant proteins were further purified by binding to Strep Tactin® SpinPrep™ filters (Novagen, Madison, WI) followed by elution with LEW buffer (50 mM NaPO4, pH 8.0; 300 mM NaCl) containing 10 mM biotin. Purified proteins were stored at 4°C. Preliminary studies (data not shown) indicated that the N-terminal tagged protein (rS-McrA) was less efficient than its C-terminal analogue (rMcrA-S) in binding HpaII methylated DNA fragments to strepavidin coated magnetic beads (Nanolink, Solulink, San Diego, CA); therefore, all remaining experiments utilized rMcrA-S.

rMcrA-S affinity enrichment of mCGI's

Human genomic A549 DNA (lung carcinoma) was exhaustively digested with *MseI* whose recognition sites (TTAA) rarely occur in GC-rich regions thereby leaving most CGI's intact. Digested DNA was phenol extracted, precipitated with ethanol and then dissolved in TE super low EDTA (TEsl) (10 mM Tris-HCl, 0.1 mM EDTA), pH 8.0. Approximately 750 ng of fragmented DNA was incubated at RT with ~7 nMol of rMcrA-S in 200 µl LEW buffer supplemented with 100 µg/ml BSA and 250 ng sonicated *E. coli* ER2925 DNA as carrier for 20 min. Following incubation, 25 µl bed volume of Nanolink magnetic strepavidin beads that were prewashed twice in 1X LEW + 100µg/ml BSA were added and incubated with gentle mixing at RT for 1 hr to capture the rMcrA-S/A549 DNA complexes. The unbound fraction was removed and the beads were washed 3x with 100 µl 1X HEPES + 250 mM NaCl; 3x with 1X HEPES + 700 mM NaCl and 2x with 50 µl 1X HEPES + 250 mM NaCl. The beads with bound DNA fragments were washed in 50 µl 1x Quick Ligase Buffer (66 mM Tris-HCl, 10mM $MgCl_2$, 1mM, 1mM ATP and 7.5% (W/V) polyethylene glycol (PEG 6000) and resuspended in 50 µl 2X Quick Ligase Buffer.

Ligation mediated PCR (LM-PCR)

An *MseI* compatible adaptor DNA cassette was formed by annealing two oligonucleotides: *MseI* Top: 5′– AGCAACTGTGCTATCCGAGGGAT–3′ and *MseI* Bottom: 5′– TAATCCCTCGGA–3′ as described above and then ligated to the MseI compatible ends of the DNA captured by rMcrA-S on the strepavidin beads. 100 pMol of adaptor and 3000 units of T4 DNA ligase were added to the resuspended magnetic beads in 50 µl 1x Quick Ligation Buffer. The reaction was incubated at 16 °C overnight.

The beads were then washed and equilibrated with 100 µl 1X Thermo Pol Buffer (20 mM Tris-HCl pH 8.8, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 2 mM $MgSO_4$, 0.1 % Triton X-100). PCR amplification was done in 50 µl NEB Thermo Pol buffer with Taq polymerase using a single unphosphorylated primer: 5′– AGCAACTGTGCTATCCGAGGGAT– 3′. Reactions were started by incubation at 72°C for 10 min to fill-in the single–stranded regions of the appended cassettes followed by cycling 14 times at 94 °C for 20 sec, 68 °C for 30 sec, and 72 °C for 2 min 30 sec [91 and T. Rauch personal communication]. Amplified products were purified using Qiagen PCR Purification Kit columns **(**Qiagen, Valencia, CA) and resuspended in 20 µl 50 mM HEPES (pH 7.5). Linkered fragments were cloned using a pSMART GC kit (Lucigen, Richmond, CA). Individual recombinant clones were sequenced using standard ABI dideoxy sequencing.

Bisulfite Sequencing

Genomic A459 DNA was bisulfite converted according to manufacturer's instructions (Qiagen, Valencia, CA) followed by whole genome amplification (Qiagen, Valencia, CA) without the initial denaturing steps since the DNA is already single-stranded following bisulfite conversion. Primers were designed

(http://www.urogene.org/methprimer/index1.html) for two regions of bisulfite modified genomic DNA. Region C1 was amplified using the forward and reverse primers C1F3: 5′- TGGGGTGTTTTTTTTGTATT- 3′ and C1R1: 5′-AAAATCCCACCCTAAACC-3′ respectively and region C2 with the forward and reverse primers C2F3: 5′-TGGGTTTTGTATAGGTTAAA-3′ and C2R1: 5′-AACAACCAAAAAATTTTCAC- 3′, respectively. PCR was done following the manufacturer's conditions using NEB Thermo Pol buffer and Taq polymerase supplemented with 5% DMSO (final concentration) and amplified as follows: 94 °C for 2 min, 2 cycles of (94 °C for 30 sec, 55 °C for 30 sec, 72 °C for 2 min) 2 cycles of (94 °C for 20 sec, 54 °C 30 sec, 72 °C for 2 min) 2 cycles of (94°C for 20 sec, 53 °C for 30 sec, 72 °C for 2 min) 2 cycles of (94 °C for 20 sec, 52 °C for 30 sec, 72 °C for 2 min) and 30 cycles of (94 °C for 20 sec, 51 °C for 30 sec, 72 °C for 2 min). PCR products were gel-purified using PCR purification columns (Qiagen, Valencia, CA), ligated into pCR4TOPO vector (Invitrogen, Carlsbad, CA) via the standard manufacturer's protocol, electroporated into *E. coli* Top 10 cells, and plated on 2xYT supplemented with 50 μg/ml Kan. Individual colonies were picked, grown in 2xYT media with 50 μg/ml Kan. Plasmids were isolated using alkaline lysis followed by column purification (Fermentes, Glen Burnie, MD), and sequenced using primers flanking the cloning site. Sequences were edited using Sequencher software (Gene Codes, Ann Arbor, MI) and analyzed using CLUSTLW.

***In vitro* Binding Assay**

Binding of McrA-S to different DNA sequences was assessed using an EMSA. Reactions (10μl) typically contained a mixture of 10-25 pmol of various test DNAs and 250 ng sonicated *E. coli* ER2925 DNA as a non-methylated, non-

specific competitor in 50 mM Tris-HCl, 100 mM NaCl, 1 mM dithiothreitol 10 mM $MgCl_2$ (NEBuffer #3), supplemented with 100 μg/ml BSA. The reaction was started with the addition of 25-175 pmol McrA-S and after 45 min at RT the entire sample was loaded on a 10% acrylamide Tris-Acetate/EDTA gel (1xTAE: 40 mM Tris-Acetate, 1 mM disodium EDTA) which were electrophoresed at RT. Following electrophoresis the gels were stained with ethidium bromide (0.5 μg/ml) at RT and visualized with UV light. Gels were then stained with Coomassie Blue to detect rMcrA-S.

Chapter 3: Cloning, Purification and Initial Characterization of *E. coli* McrA

Introduction

Expression strains of *Escherichia coli* BL21(DE3) overproducing the *E. coli* m^5C McrA restriction protein were produced by cloning the *mcrA* coding sequence behind a T7 promoter. The recombinant *mcrA* minus BL21(DE3) host produces active McrA as evidenced by its acquired ability to selectively restrict the growth of T7 phage containing DNA methylated *in vitro* by HpaII methylase. The *mcrA* coding region contains several non-optimal E. coli triplets. Addition of the pACYC-RIL tRNA encoding plasmid to the BL21(DE3) host increased the yield of recombinant McrA (rMcrA) upon induction about 5 to 10-fold. McrA protein expressed at 37 °C is insoluble but a significant fraction is recovered as soluble protein after autoinduction at 20 °C. rMcrA protein, which is predicted to contain a Cys_4-Zn^{++} finger and a catalytically important histidine triad in its putative nuclease domain, binds to several metal chelate resins without addition of a poly-histidine affinity tag. This feature was used to develop an efficient protocol for the rapid purification of nearly homogeneous rMcrA. The native protein is a dimer with a high α-helical content as measured by circular dichroism analysis. Under all conditions tested purified rMcrA does not have measurable nuclease activity on HpaII methylated (Cm^5CGG) DNA, although the purified protein does specifically bind HpaII methylated DNA. These results have implications for understanding the *in vivo* activity of McrA in "restricting" m^5C-containing DNA and suggest that rMcrA may have utility as a reagent for affinity purification of DNA fragments containing m^5C residues.

Epigenetic factors, and in particular methylation at the 5-position in cytosine residues in CpG dinucleotides to form m^5CpG, regulate the function of vertebrate genomes by controlling gene expression and chromatin folding.

Aberrant hypermethylation of CpG islands near promoters of human tumor suppressor and other genes is now recognized as an important contributing factor in cancer, aging and several other pathological states. Detecting these aberrantly methylated regions and accurately determining their methylation profile is an area of considerable interest primarily because of their potential use as diagnostic and prognostic biomarkers for cancer [68-71]. Reagents, such as m^5CpG binding proteins, which preferentially bind to methylated CpG dinucleotides and enzymes that cleave specifically at methylated CpG dinucleotides, are useful tools for identification and characterization of m^5C-containing DNA regions [69-71, 78, 91-93]. In this study we set out to characterize McrA, an E. coli protein purported to be a nuclease with specificity towards m^5C-containing DNA to evaluate its usefulness as a reagent for the identification of m^5C residues at single nucleotide resolution.

Wild-type *E. coli* K-12 strains possess several restriction systems in addition to the classical EcoK *hsdR/M/S* host-specificity restriction-modification mechanism. One of these, Mar (for *methylated adenine restriction*), is directed only against DNA containing N^6-methyladenine residues while another Mrr (for methylated adenine recognition and restriction) has been reported to restrict DNA containing N^6-methyladenine and also DNA with C^5-methylcytosine residues [94-95]. Neither system restricts DNA methylated by the *E. coli* enzymes encoded by *dam*, which methylates the A residue in the sequence GATC, or by *dcm*, which modifies the internal cytosine in CCWGG (W is A or T) sequences at the C^5 position [81, 94].

DNA containing C^5-methylcytosine (m^5C) is also restricted by the Mcr (for modified cytosine restriction) system which is identical to the previously described Rgl (for restricts glucose-less phage) restriction system that blocks the

growth of T-even phages, but only when they contain 5-hydroxymethyl cytosine in their DNA, i.e., when their 5-hydroxymethylcytosine residues are not glucosylated [96]. Later work further subdivided the Mcr system into two genetically distinct regions: McrA (equal to RglA) on an easily excisable but defective lambdoid prophage element e14 located at 25 min on the *E. coli* K-12 chromosome and McrB (or RglB) at map position 99 min in a region that includes the EcoK restriction/modification and Mrr systems [80, 83]. McrA recognizes DNA containing C^5-methylcytosine or C^5-hydroxymethylcytosine while McrB also recognizes DNA containing N^4-methylcytosine. The *mcrB* locus encodes two polypeptides McrB and C which together function as a nuclease recognizing in *cis* two half sites 5′-G/A 5mC ($N_{40\text{-}3000}$) G/A 5mC-3′. Cleavage requires GTP hydrolysis and occurs at a non-fixed distance between the two methylated half sites [97].

Early studies showed that DNAs methylated by M.HpaII (Cm^5CGG), M.Eco1831I (Cm^5CSGG where S is C or G) and M.SssI (m^5CG) are restricted by the McrA system and further studies demonstrated that clones expressing the McrA open reading frame conferred both McrA and RglA phenotypes on a *mcr* minus host [82-83]. However, since the McrA protein has never been purified its precise sequence preferences and its mode of action remain unclear although it is generally believed to be a member of the ßßα-Me finger superfamily of nucleases acting specifically on m^5C-containing DNA. McrA also contains an H-N-H motif common to homing endonucleases as well as many restriction and DNA repair enzymes. The core ßßα-Me domain of McrA (residues 159 to 272 of the 277 amino acid long polypeptide) was modeled by Bujnicki [98] and coworkers using a protein sequence threading approach. This region contains three histidine residues (H-228, 252, and 256) predicted to coordinate a Mg^{2+} ion, as

well as four cysteine residues (C-207, 210, 248, and 251) which form a putative zinc finger, most likely involved in coordinating Zn^{2+} or some other divalent metal ion to help stabilize the protein's structure.

While McrA is predicted to function as a nuclease this has never been demonstrated and to date the mechanism for biological restriction of modified phage or plasmid DNAs by McrA is not known. Furthermore, although a slightly N-terminal truncated form of the polypeptide has been cloned in an expression vector [99-100], McrA protein has not been purified. Here we report the cloning, expression, purification and initial characterization of full-length, biologically active rMcrA. All attempts to demonstrate that rMcrA is a nuclease acting on m^5C-containing DNA have failed but electrophoretic mobility shift analysis demonstrates that purified rMcrA interacts specifically with DNA fragments containing Cm^5CGG sequences. The production of the recombinant McrA protein in good yield opens up the possibility of obtaining its 3D-structure and will help further investigations into its genuine mode of action *in vivo*.

Results and Discussion

Expression of recombinant McrA in *E. coli*

Prior to this study, expression of full length recombinant McrA (rMcrA) protein has never been reported. Our approach was to place, by PCR, a unique recognition sequence, for BsmBI at an appropriate distance upstream of the *McrA* start codon and another unique site, BamHI, just past the *McrA* stop codon. BsmBI is a Type IIS restriction endonuclease that cleaves the double-stranded DNA outside of its recognition site. Cutting this amplicon with BsmBI and BamHI leaves 4 base cohesive ends complementary to the overhangs produced

by cutting a standard pET28 cloning/T7-based expression vector with NcoI and BamHI. After directional cloning and sequence verification, rMcrA expression conditions were optimized for reproducible purification of soluble protein. Routine expression was carried out by autoinduction in ZYM 5052 medium [88] at 20 °C in a *mcrA*$^{-}$ BL21(DE3) host harboring the pRIL t-RNA plasmid. $ZnSO_4$ (30 μM) was included in the medium as McrA is thought to have a C4-Zn finger motif [98]. Since we wanted to determine whether the cloned *mcrA* gene was active we did not use the similar expression host provided by Strategene as it has a *mcrA*$^{+}$ genotype.

To test for biological activity we titered T7 phage containing normal, non-methylated or *in vitro* HpaII methylated, T7 DNA on BL21(DE3) pRIL cells with and without pET-rMcrA. The presence of pET-rMcrA reduced the efficiency of plating (EOP) of the phage containing methylated DNA approximately a thousand-fold compared to their EOP (see Table 3.2) on this same host containing a similar pET28 vector with a non-related recombinant gene insert at the same position. This reduction in EOP was selective and not seen with *in vitro* packaged T7 phages containing non-methylated DNA. From these results we concluded that rMcrA is biologically active.

McrA contains three histidines in a ßßα-Me domain predicted to co-ordinate a Mg^{2+} ion which is believed to be essential for the phosphodiester bond cleavage [98]. This prediction suggested to us that rMcrA might bind directly to matrices typically used for affinity purification of poly-His tagged proteins. Preliminary experiments demonstrated that fairly high purity rMcrA, as visualized by Coomassie blue stained SDS-PAGE gels, could be obtained by affinity chromatography of crude extracts on Ni, Zn or Co charged NTA beads followed by elution with buffers containing 100-250 mM imidazole. No binding

was observed to uncharged NTA beads. Presumably binding to these metal resins is via one or more histidine residues in the ßßα-Me domain as the rMcrA protein lacks a poly-His tag. Several different metal-ion-charged resins were tried but we chose the IDA resin from USB for routine use because of its ease of preparation, high capacity and purity of the eluted rMcrA protein. Batch chromatography on SP-Sepharose Fast Flow was used to further purify rMcrA (Fig. 3.1 A). This step also removes imidazole from the buffer. The concentration of rMcrA was determined by spectrophotometry assuming a calculated molar extinction coefficient of 35,870 from its amino acid composition (http://encorbio.com/protocols/Prot-MW-Abs.htm) ignoring the contribution of the protein's seven cysteine residues. The absorbance of a 1 mg/ml solution at 280 nm is calculated to be 1.14. Approximately 7 to 8 mg of rMcrA is obtained from 50 ml of autoinduced cells. Size exclusion chromatography (Fig. 3.2A) indicates that rMcrA is a dimer (62 kDa) under native conditions with a high α-helical content (~60 %) as determined from its CD spectrum (Fig. 3.2C) [101]. ESI mass spectrometric analysis (Fig. 3.2D) of non-reduced rMcrA is in good agreement with the calculated mass of the monomeric protein (31,389 expected vs 31,393 observed). Furthermore, the relatively small peak detected at the expected mass position of an rMcrA dimer is consistent with non-covalent associations forming the rMcrA dimer.

rMcrA selectively binds HpaII methylated DNA

The DNA-binding properties of rMcrA were examined using fragments of unmethylated and HpaII methylated T7 DNA by EMSA. The results of a typical EMSA carried out in the absence of added Mg^{2+} are shown in Fig. 3.3 and similar results were obtained if $Mg^{2+}$was present. The addition of rMcrA had no effect

on the mobility of the non-methylated fragments until much higher ratios of moles of protein to CCGG sites in the DNA were used (T7 DNA contains 58 CCGG sites); however, the motilities of the fragments with methylated HpaII sites were significantly retarded in the presence of even small amounts of rMcrA. Furthermore, the addition of rMcrA resulted in the appearance of two distinct shifted products for each of the smaller HpaII-methylated KpnI fragments. Separate experiments demonstrated that binding of rMcrA to HpaII methylated T7 DNA does not modify the m^5cytosine ring as does binding of the Arabidopsis thaliana proteins DEMETER (DME) and repressor of silencing (ROS1) [102-104]. These closely related DNA glycosylase domain containing proteins remove 5-methylcytosine from the DNA backbone and then their lyase activities cleave the resulting abasic site by successive β- and δ-elimination reactions. If McrA has a similar activity it could account for its known restriction of methylated DNA and initiation of a SOS response following *in vivo* induction of HpaII in $mcrA^+$ cells. To rule out this possibility we treated rMcrA reacted DNA with proteinase K, EDTA and SDS followed by phenol/chloroform extraction and demonstrated that the DNA was still resistant to cutting by HpaII but sensitive to cutting with MspI which cleaves the same sequence even when the internal cytosine is 5-methylated.

We next determined by EMSA whether rMcrA binds to other methylated DNA sequences. For these assays we used T7 DNA modified by HpaII and the other six methyltransferases listed in Table 3.2. Most of these enzymes with the exception of dam methyltransferase have more than 50 sites in T7 DNA. Interestingly no shift products were observed for any of these non-HpaII methylation patterns nor did dual methylation by HpaII and Msp I (m^5Cm^5CGG) prevent binding (data not shown). These observations are consistent with and

extend the finding about the specificity of McrA based solely on *in vivo* studies [80-81, 94]. However, they still do not identify conclusively the minimal sequence or number of sites needed for McrA binding.

rMcrA does not appear to be a m^5C specific nuclease

Numerous attempts to find conditions under which rMcrA would digest HpaII methylated T7 DNA or pGEM3 plasmid DNA *in vitro* were unsuccessful. Among the parameters tested were different buffers (NEB #1, #2, #3 and #4; 50 mM K glutamate, pH 6.0; LEW; and 20 mM Hepes, pH 7.2, 300 mM NaCl) and addition of 1.5 or 3 mM Ca^{2+}; Cd^{2+}; Cu^{2+}, Fe^{2+}; Mg^{2+}; Mn^{2+}, Ni^{2+}or Zn^{2+} ions; as well as addition 3 mM ATP or GTP or 10 μM S- adenosylmethionine to the standard NEB restriction buffers. While we cannot rule out the possibility that rMcrA was inactivated during purification this seems unlikely given the ability of the protein to selectively bind HpaII methylated DNA (see above).

Although we were unable to determine whether or not rMcrA had nuclease activity, we were able to compile some preliminary data. Standard cloning protocols were used to introduce the plasmids pET28-McrA which had been previously developed (see above) and an additional pACYC-M.HpaII plasmid, kindly provided by Elizabeth Raleigh (NEB) that constitutively expresses the HpaII methylase protein, into Bl21 (DE3) cells. Additionally pET28-McrA was introduced into BL21 (DE3) cells containing the pACYC-RIL plasmid; the pET-OspC, a vector containing the Borellia OspC protein was introduced into BL21 (DE3) cells containing the plasmid pACYC-M.HpaII. These were used as controls to rule out the effects of having two plasmids in the cell. Cells grown as described in chapter 2. The pET vectors are under the control of a T7 inducible promoter, however this is a leaky promoter and even without induction by IPTG

the combination of pACYC-MHpaII, and pET28-McrA was lethal for the cells (Table 3.3). In contrast the cells containing the combination of either pACYC-M.HpaII and pET-OspC or pACYC-RIL and pET28-McrA were not restricted in their growth. (Table 3.3). There were some rare colonies that were seen in the cells containing pACYC-MHpaII, and pET28-McrA and we were intrigued to know why these colonies survivied. We subsequently amplified the McrA gene from the plasmid in these colonies and then sequenced the resulting PCR product. All of these colonies contained a deletion in the McrA gene.

This led us to develop new plasmids containing rMcrA with more tight control over the T7 inducible promoter. For this work, the vectors pREX-S31 and pREX-LS31, which were developed and kindly provided F.W. Studier, were used. The pREX-S31 vector has a T7 promoter followed downstream by the lacO-S operator. This is a symmetrical lac operator which has been shown to bind lac repressor more tightly than the lac operator present in standard pET vectors [105]. The pREX-LS31 vector has the T7 promoter flanked upstream by the normal lac operator and downstream by the symmetrical lac operator. These plasmids were then introduced into BL21-AI cells, in which the T7 RNA polymerase is under the control of an arabanose promoter and as such has a lower basal and induced expression of the T7 RNA polymerase [88] compared to BL21 (DE3), along with the pACYC-M.HpaII plasmid and grown with appropriate antibiotics. Under these conditions the cells were able to grow with both of the plasmids. When induced using IPTG and arabanose the amount of total DNA recovered over time is reduced compared to an uninduced control (Fig 3.4 A and B). From this point our lab is interested in using these vectors to investigate the putative nuclease activity of McrA.

Presumably, McrA acting simply as a DNA binding protein could interfere with methylated phage development or plasmid maintenance [85]. It is possible that bound McrA interacts with some other E. coli protein(s) to cause cleavage *in vivo* at methylated HpaII sites. Our attempts to demonstrate that rMcrA can act in concert with McrBC to cause cleavage at methylated HpaII sites have been unsuccessful.

In this study, the smallest number of methylated sites observed to cause a mobility shift was seven although preliminary data (EAM) indicates that as few as two sites may be sufficient for efficient binding. Such detailed studies will require the use of synthetic oligonucleotides.

Our data further indicate that since rMcrA does not bind to HhaI methylated DNA it should only interact with a subset of the sequences recognized by the methyl binding domains of the eukaryotic proteins MeCP2 and MBD2 which together with a monoclonal antibody to m^5cytosine are used routinely to affinity purify methylated CpG islands from total genomic sonicates or restriction digests [69-70, 78, 91-92]. Fusing the McrA to an appropriate, presumably non-poly His, affinity tag might permit its use in matrix assisted binding of fragments containing only Cm^5CGG sequences which are frequently found in aberrantly methylated CpG islands having diagnostic value.

In summary, purified rMcrA does not by itself appear to be a nuclease. Although we cannot fully exclude the trivial explanation that the protein was damaged during purification or that it requires a hitherto unknown co-factor or untried reaction conditions, we favor the idea that its role *in vivo* is more elusive and complicated than initially proposed.

Tables and Figures

Table 3.1

Secondary structure analysis of rMcrA from CD spectrum

Secondary structure type	
α-Helix Type 1 – regular (%)	38
α-Helix Type 2 – distorted (%)	22
β-Sheet Type 1 – regular (%)	5
β-Sheet Type 2 – distorted (%)	3
Turns (%)	11
Unordered/random coil (%)	21
Total	100

Table 3.2

Biological properties of rMcrA

Methylase	Sequence	# of sites in T7	rMcrA binding	T7 inhibition[a]
AluI	5'…AGm^5CT…3'	140	NO	ND
dam	5'…Gm^6ATC…3'	6	NO	ND
HaeIII	5'…GGm^5CC…3'	68	NO	ND
HhaI	5'…Gm^5CGC…3'	103	NO	ND
HpaII	5'…Cm^5CGG…3'	58	YES	YES ~ 10^3-fold
MspI	5'…m^5CCGG…3'	58	NO	ND
MspI + HpaII	5'…m^5C m^5CGG…3'	58	YES	ND
TaqI	5'…TCGm^6A…3'	111	NO	ND

[a]determined by plating T7 phage containing *in vitro* methylated DNA on BL21(DE3)/pRIL with and without the pET28-rMcrA plasmid relative to plating of *in vitro* packaged phage without methylated DNA on these same hosts.

Table 3.3

Cell Line	Plasmid(s) (200ng ea) Order of Addition	Resistance	Vol Plated/Dilution	Colonies	CFU/mL
BL21(DE3)	pACYC-MHpaII, pET28-McrA	CAM, Kan	50 µl/NA	0	N/A
BL21(DE3)	pACYC-MHpaII, pET28-McrA	CAM, Kan	100 µl/NA	3	30
BL21(DE3)	pACYC-MHpaII, pET28-McrA	CAM, Kan	200 µl/NA	4	20
BL21(DE3)	pACYC-MHpaII, pET-OspC	CAM, Kan	100 µl/1:100	82	8.2E+04
BL21(DE3)	pACYC-MHpaII, pET-OspC	CAM, Kan	100 µl/1:1000	8	8.0E+04
BL21(DE3)	pACYC-MHpaII, pET-OspC	CAM, Kan	100 µl/1:10000	1	1.0E+05
BL21(DE3)	pACYC-RIL, pET28-McrA	CAM, Kan	100 µl/1:100	TNTC	N/A
BL21(DE3)	pACYC-RIL, pET28-McrA	CAM, Kan	100 µl/1:1000	TNTC	N/A
BL21(DE3)	pACYC-RIL, pET28-McrA	CAM, Kan	100 µl/1:10000	484	4.8E+07
BL21(DE3)	pACYC-RIL, pET-OspC	CAM, Kan	100 µl/1:100	TNTC	N/A
BL21(DE3)	pACYC-RIL, pET-OspC	CAM, Kan	100 µl/1:1000	TNTC	N/A
BL21(DE3)	pACYC-RIL, pET-OspC	CAM, Kan	100 µl/1:10000	380	3.8E+07
BL21(DE3)	pET28-McrA, pACYC-MHpaII	Kan, CAM	100 µl/NA	0	N/A
BL21(DE3)	pET28-McrA, pACYC-MHpaII	Kan, CAM	200 µl/NA	0	N/A

Summary of results of plasmid introductions into BL21 (DE3) cells and plated with appropriate antibiotics under non inducing conditions. The introduction of plasmids containing M.HpaII, and McrA is lethal for the cells. The introduction of either McrA with a pRIL plasmid or M.HpaII with OspC did not have a detrimental effect on the cells.

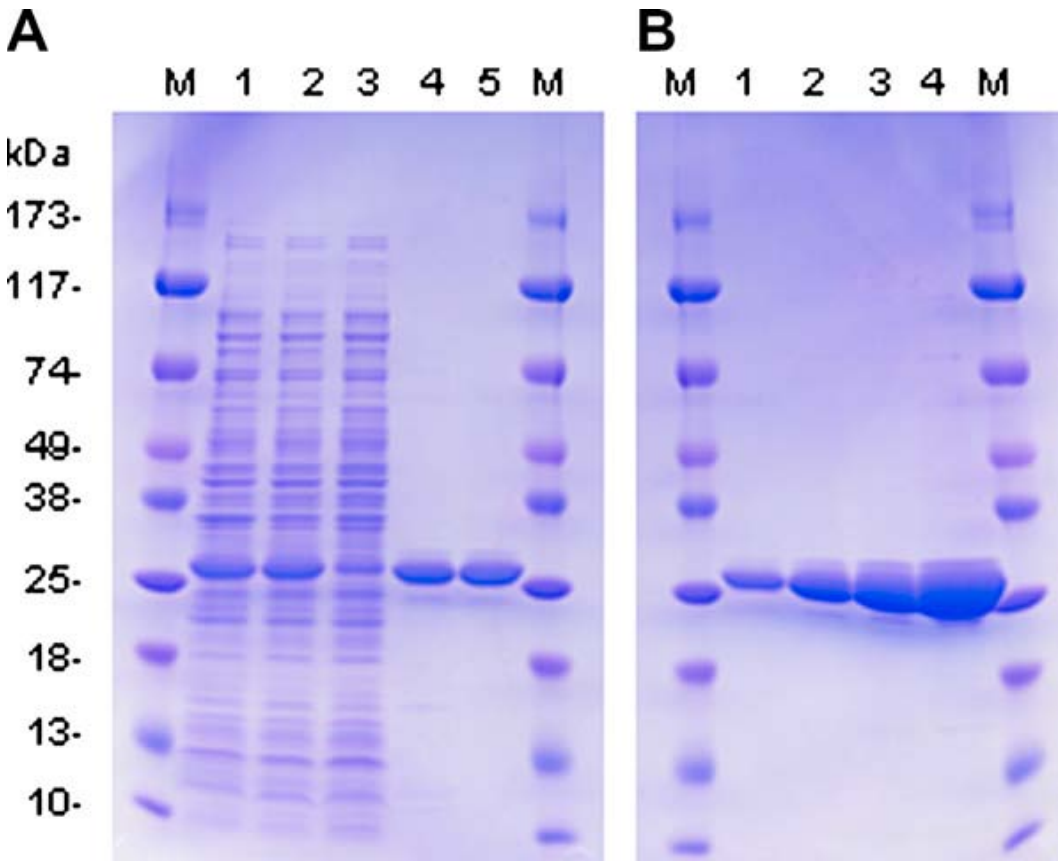


Figure 3.1 Purification of McrA. (A) Purificationof McrA from *E. coli* BL21(DE3)/pRIL. The following samples were loaded onto 4%–20% SDS–PAGE: lane 1, total extract prior to centrifugation; lane 2, supernatant after centrifugation;lane 3, protein not bound to the IDA resin; lane 4, pool of IDA 100 mM imidazole eluant; lane 5, pool of SP-Fast Flow Sepharose eluant. (B) SDS–PAGE of increasing amounts of purified rMcrA: lane 1, 1.5 lg; lane 2, 3.5 lg; lane 3 lg; 7.0 lg; lane 4, 15 lg. M indicates lanes with ProSieve color protein molecular weight markers.

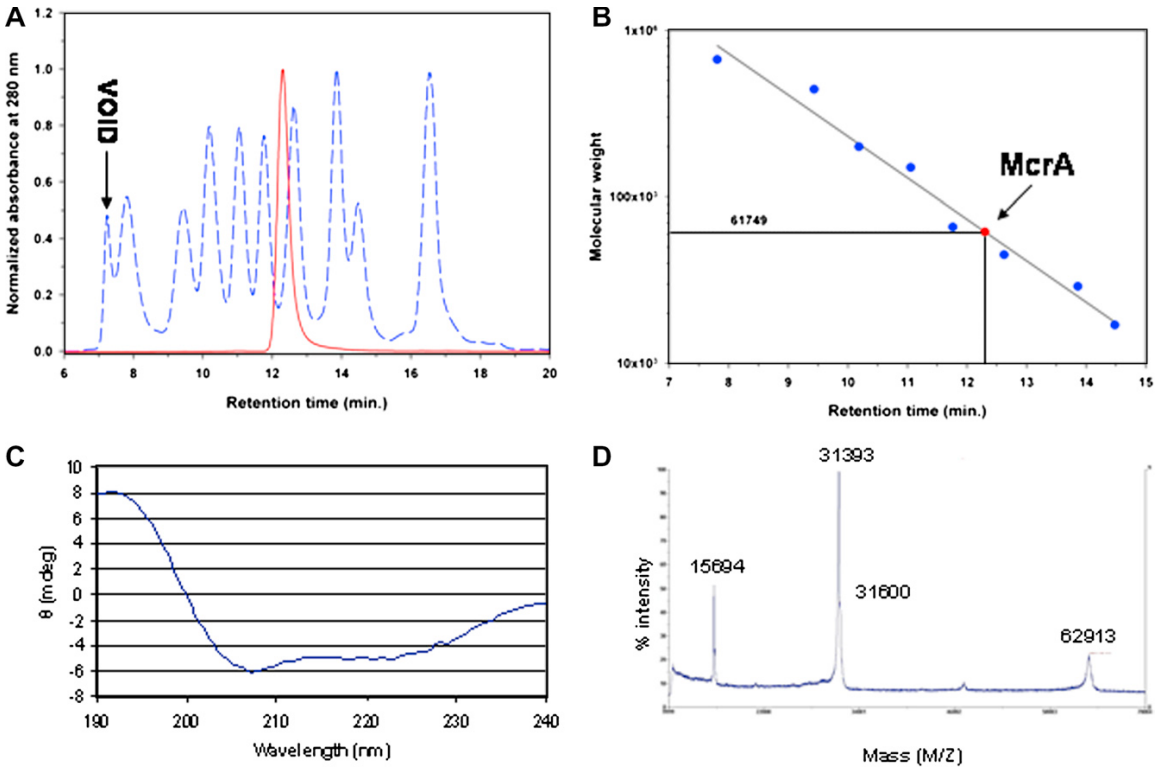


Figure 3.2. (A) Analytical size exclusion chromatography of rMcrA. rMcrA eluted as a single peak (solid red line). The elution peaks for the void and marker proteins are indicated by dashed lines. (B) The mass of native rMcrA calculated relative to its elution time and that of the molecular mass standards (C) CD spectrum of rMcrA. The shape of the spectrum indicates a high content of α-helix in the secondary structure (see Table 3.1). (D) TOF-MS analysis of native (non-reduced) rMcrA.

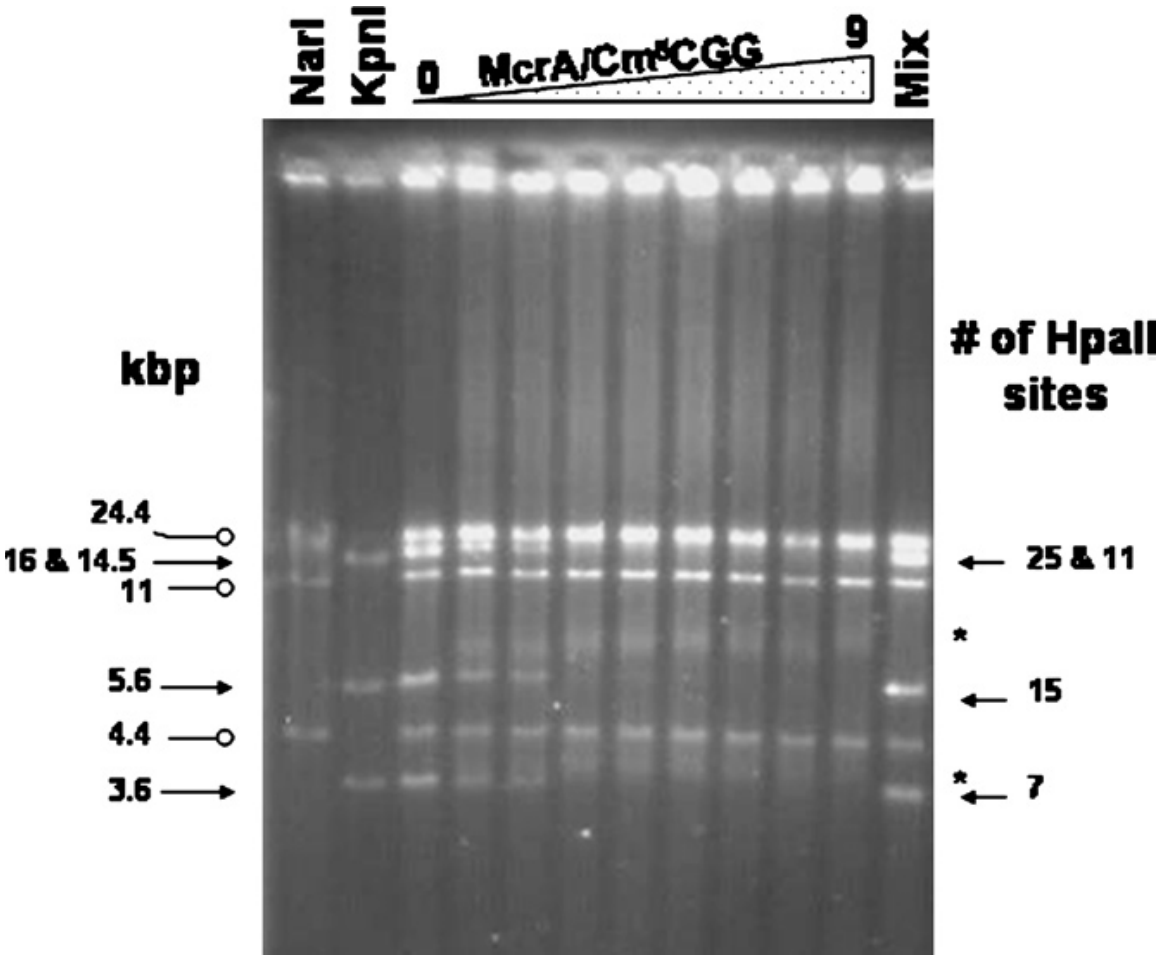


Figure 3.3. Electrophoretic mobility shift assay (EMSA) showing the DNA binding activity of rMcrA to HpaII methylated (solid arrows) but not unmethylated (open circles) T7 DNA fragments. Unmodified and HpaII methylated (Cm^5CGG) T7 DNA was digested with NarI and KpnI respectively. Increasing amounts of rMcrA were added to a 1:1 mixture of the digested T7 DNAs. The molar ratio of added rMcrA to total Cm^5CGG sites is: 0; 0.23; 0.3; 0.45; 0.56; 1.14; 2.3; 4.5 and 9.0. The sizes of the fragments (kbp) and number of Cm^5CGG sites in each is indicated on the left and right, respectively [87]. The positions of the EMSA shifted smaller HpaII methylated KpnI fragments are labeled with asterisks.

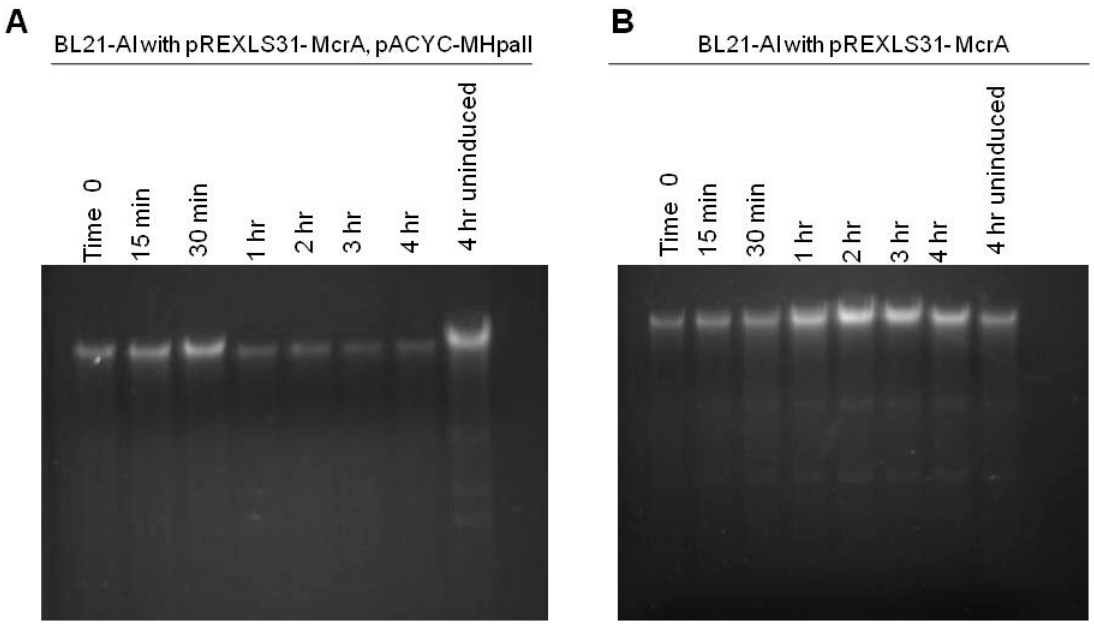


Figure 3.4. Time course after induction of BL21-AI cells with IPTG and arabanose. A)Total DNA from BL21 cells with the plasmids pREXLS31-McrA and pACYC-M.HpaII, at 0 time, 15 min, 30 min, 1hr, 2hr, 3hr, and 4hr post induction and an uninduced control grown in tandem at 4hr timepoint. B) Total DNA from BL21 cells with the plasmid pREXLS31-McrA only at 0 time, 15 min, 30 min, 1hr, 2hr, 3hr, and 4hr post induction and an uninduced control grown in tandem at 4hr timepoint.

Chapter 4: Defining the Binding Site of McrA

Introduction

Epigenetic modification of DNA is emerging as one of the most important links between the environment, life style, and changes in gene expression: particularly significant is the methylation or loss thereof at the C5 position in cytosine in CpG dinucleotides to form m^5CpG or to convert m^5CpG to CpG [21, 106-110]. In the human genome only ~4% of all cytosine residues are methylated because CpG dinucleotides are underrepresented. This dearth might be an evolutionary consequence of ineffective, mutation-prone DNA repair at spontaneously deaminated 5-methylcytosines [20]. Many CpG dinucleotides cluster into what have been termed CpG islands (CGIs). CGIs are identified bioinformatically and experimentally as DNA sequences ≥500 bp with a base composition greater than 50% G+C and a CpG [observed/expected] of more than 0.6 [22, 24-25]. The human genome has ~30,000 CGIs, accounting for about 10% of the total DNA, about half of which are found near annotated transcriptional start sites (TSS-CGIs); the remainder being intra- or intergenic (non-TSS). As pointed out by Bird and co-workers however, several non-TSS CGIs have been shown to coincide with previously unforeseen, but functional promoters raising the possibility that all CGIs function as promoters and are therefore TSS-associated [24]. In normal tissues CGIs are usually unmethylated but a subset (10's to 100's) becomes reproducibly methylated in normal cells (imprinting, X-chromosome inactivation, tissue differentiation) or in diseased cells such as cancer cells [26-27]. Consequently, there is growing interest in determining the global DNA methylation patterns in normal and diseased tissues.

CGI methylation has been shown to induce long term changes in gene expression through direct interference with transcription factor binding and via the action of methyl-CpG-binding proteins that recruit chromatin remodeling

enzymes to form "closed" heterochromatin thereby preventing transcription (gene silencing). The main repeat families, SINEs, LINEs and LTRs contribute 27%, 12%, and 7% of the 28 million CpG dinucleotides in the human genome respectively (hg18, http://genome.ucsc.edu). During tumororigenesis however, many of these non-TSS islands become hypomethylated, which may activate repetitive element expression, fostering DNA breakage and genome instability, known hallmarks of cancer. Currently, 66 genes are positively identified as being aberrantly methylated in cancer as listed in (www.mdanderson.org: methylation in cancer).

The most common approaches for determining DNA methylation patterns can be can be classified into several major categories: immunoprecipitation [111], or m^5CpG affinity-capture using m^5CpG-binding proteins and related methods [24, 69, 78, 112] restriction enzyme-based methods, microarray-based methods or various combinations of these methods [84] and bisulfite-based modification followed by single or multi- locus DNA sequencing or more recently by high-throughput, massively parallel sequencing [98].

Bisulfite sequencing employs an initial chemical modification step followed by PCR amplification wherein nonmethylated cytosines are replaced by T but methylated cytosines are resistant to modification and remain unchanged. After treatment, DNA regions of interest are typically amplified using sets of region-specific primers and the resulting amplicons are analyzed individually using a variety of techniques to identify C to T transitions or unmodified C positions, which correspond, respectively, to unmethylated and methylated cytosines in the native DNA. Unfortunately, these high-throughput approaches are still very expensive and mapping of the resulting short sequence reads to their respective genomic loci is challenging and beyond the scope of most laboratories.

Additionally, detecting methylated- and unmethylated-CpG dinucleotides at single nucleotide resolution requires some expertise and/or costly equipment. In an attempt to moderate these drawbacks we recently demonstrated that the *E. coli* McrA protein (for modified cytosine restriction) forms complexes with symmetrically HpaII-methylated double-stranded DNA (Cm^5CGG), but not unmethylated DNA [84]; however, its precise recognition sequence has remained undefined. This prompted us to design experiments to determine the spectrum of endogenously methylated McrA targets in human DNA. We have successfully used McrA fused to a short 8 amino acid long StrepII tag (rMcrA-S) to affinity capture methylated MseI restriction fragments from total human DNA. Standard sequencing of these fragments in conjunction with bisulfite genome sequencing analysis helped us more fully ascertain the DNA-binding profile of McrA. rMcrA-S was also used in electrophoretic mobility shift assays (EMSA) with symmetrically methylated, hemimethylated and nonmethylated double-stranded DNA probes with a canonical HpaII site and various single base pair permutations flanking the central m^5CpG dinucleotide or opposite the m^5C residue. Together, these data have helped define the minimal recognition sequence and base-pairing requirements for McrA's interaction with DNA. These observations may lead to the development of rMcrA-S-based m^5C detection assays that are independent of bisulfite modification.

Results

rMcrA-S binds to Methylated Human DNA

In the previous chapter we discussed that purified full length recombinant McrA (rMcrA) binds to but lacks detectable nuclease activity at methylated HpaII (Cm^5CGG) sequences [84] and as such, may have utility as a reagent for

affinity purification of human DNA fragments containing m^5C residues. The human genome contains about 2.3 million HpaII sites of which roughly 12% are located in CGIs (Table 4.1). However, since these initial studies were performed using T7 DNA methylated *in vitro* with HpaII methylase we could not ascertain whether McrA can also bind and perhaps be used to affinity purify DNA fragments with related sequences containing m^5CpG's or if McrA could be used to preferentially enrich for a subset of CGI's containing several methylated HpaII sites. A standard technique for converting McrA into an affinity reagent for such studies would be to fuse a $(His)_6$ tag to either end of the protein and then use the recombinant tagged protein to generate an affinity matrix for binding DNA fragments containing m^5CG. However, rMcrA intrinsically binds to Ni-charged NTA supports presumably because of interaction with three suitably positioned histidines in its ßßα-Me domain [84, 98]. Preliminary experiments indicated that binding of rMcrA/DNA complexes to Ni-charged NTA magnetic beads was very inefficient, presumably due to steric hindrance. We therefore decided to add an 8 amino acid long StrepII tag (WSHPQFEK) [113] to either end of the protein to aid in affinity capture of methylated DNA fragments independent of the Ni binding site. In our hands this tag when placed at the C-terminus does not seem to adversely affect the protein's capture of DNA fragments with methylated HpaII sequences whereas its N-terminal tagged counterpart is much less efficient (data not shown).

To resolve the extent of McrA's binding specificity, we initially used rMcrA-S to enrich for methylated sequences from a MseI digest of genomic human DNA since it would contain m^5CpG dinucleotides in all possible contexts and further MseI is known to leave CGIs mostly intact [26]. In addition, we could see if rMcrA-S preferentially captures CGI's with methylated HpaII sites. Accordingly,

we washed bound fragments obtained by rMcrA-S affinity purification with high ionic strength buffer and then LM-PCR-amplified, cloned and sequenced to determine if they all contained HpaII sites and if any originated from chromosome regions defined as being a CGI. We found that most clones from a library of non-size selected amplicons were not from CGIs but from regions containing repetitive sequences; regions known to contain m^5CpGs. Some minor enrichment for CGI fragments was noted when the amplified DNA was gel purified and fragments between 0.5 and 2 Kb (the size range expected for CGIs in the digest) were used to prepare the library. While many clones in both libraries contained one or more HpaII sites some had no HpaII sites although they invariably contained several CpG dinucleotides (Fig. 4.1). Overall, the increase in CpG dinucleotides in these libraries relative to the starting DNA was about 2-3-fold.

After we located all the affinity captured sequences in the human genome using the UCSC Genome Browser (March, 2006 assembly), we chose two unique regions for further study: "C1," within Ch. 8q24.3 containing 2 HpaII sites, one CCG sequence and 5 other CpG dinucleotides; and "C2," within Ch. 18q11.2 lacking HpaII sites but having a single CGG sequence plus 3 other CpGs within the MseI fragment. We next determined the methylation status of these sites *in vivo* via sodium bisulfite sequencing. Briefly, total genomic A549 DNA was bisulfite modified, amplified using a whole genome amplification assay, and the C1 and C2 regions amplified using bisulfite specific primers C1F3, C1R1, and C2F3, C2R1, respectively. Standard sequencing methods were used to deduce the genomic methylation pattern in 12 clones from each region.

Sequences were checked with an internal control. On the C1 clones the first cytosine, which is umethylated, of the HpaII sites was checked for

conversion to thymine. All 11 C1 clones had C to T conversions at the first cytosine in both HpaII sites. For the C2 clones we used a CpT site to check for cytosine to thymine conversion. Only clone C2.12 did not have conversion of cytosine to thymine at this position. We subeequently checked all other non CpG cytosines and found that clone C2.12 had complete conversion at all other non CpG sites. Further all other C2 clones had complete conversion of cytosine to thymine at non CpG sites.

Alignment of these sequences showed that 62.5-100% of all CpG sites in C1 were methylated and all C1 clones had at least one methylated HpaII sequence or m^5CGG sequence (hereafter referred to as a ¾ HpaII site) (Fig. 4.2 a). In the case of the C2 clones, from a region which lacks any HpaII sites, all 12 clones were methylated at the ¾ HpaII site (Fig. 4.2 b). Ten of twelve C2 clones had 100% methylation across all four CpG dinucleotides. These data led us to investigate whether or not the ¾ HpaII site is a minimal binding site for rMcrA-S.

rMcrA-S binds Cm^{5}CGG and m^{5}CGG

To further define a consensus rMcrA-S DNA recognition site we turned to EMSA using 24 bp synthetic oligonucleotide cassettes containing one of the following: (A) three m^{5}CpG's including one in an HpaII site (Cm^{5}CGG); (B) one with a single m^{5}C in a "3/4 HpaII site" (m^{5}CGG); (C) one containing a single m^{5}C in an HpaII site and, as a control, one with no m^{5}Cs (Table 4.3). As shown in Fig. 4.3, rMcrA-S has high affinity for a ds-cassette containing a single symmetrically methylated HpaII site, however, at higher ratios of protein to DNA, binding to the cassette with a single symmetrically methylated ¾ HpaII site becomes evident but no binding is seen to the unmethylated control cassette (Fig. 4.3 a & b) at the protein/DNA ratios used here. In other experiment (data not show)

rMcrA-S seems to have a relatively higher affinity, as judged by the complete shifting of the input DNA cassette containing a single methylated HpaII site compared to the cassette containing three methylated sites. This might be related to steric hindrance since the three methylated sites are closely positioned in tandem. We have also found that the binding is independent of Mg^{++} ions in the binding buffer (data not shown).

rMcrA-S selectively binds N(Y>R) m^{5}CGR

To further define rMcrA-S's binding preference we used complementary synthetic oligonucleotides that were annealed and methylated *in vitro* using M.SssI as described in Materials and Methods. As shown in Table 4.3 these oligonucleotides all contain a single CpG dinucleotide either proceeded by D (A, G or T) or followed by H (A, C or T). The results of gel shift assays with these methylated cassettes are shown in Fig 4.4.

Under our standard EMSA conditions rMcrA-S preferentially shifts cassettes if a purine follows the m^{5}CpG dinucleotide (m^{5}CGR) rather than a pyrimidine (m^{5}CGY) (Fig. 4.4 a). These findings might help to explain why our previous *in vivo* studies found that McrA did not bind T7 DNA with methylated HhaI sites (Gm^{5}CGC) [84].

We also investigated the importance of cytosine preceding m^{5}CGR. Double-stranded cassettes were designed with a single NCGR site, M.SssI methylated *in vitro* and then analyzed for rMcrA-S's binding by EMSA. rMcrA-S bound all 8 cassettes (Fig. 4.4 b) but it seems to have a somewhat higher affinity for duplexes with Ym^{5}CGR (Fig. 4.4 b - lanes 2, 4, 6, and 8) but it also shifts duplex cassettes with a Rm^{5}CGR sequence; results consistent with our initial findings that rMcrA-S was able to affinity purify genomic MseI fragments lacking a methylated HpaII site but containing Am^{5}CGG, a ¾ HapII site.

rMcrA-S also binds Hemi-Methylated DNA

During mammalian DNA replication CpG dinucleotides in the daughter strand are initially unmethylated until methylated by the maintenance methyltransferase, DNMT1 [114]. We were therefore interested to learn if rMcrA-S interacts with a hemimethylated Cm^{5}CGG sequence. For these studies a synthetic oligonucleotide with single m^{5}C added during synthesis was annealed with its methylated or non-methylated complements (Table 4.4). As shown in Fig. 4.5, added rMcrA-S only gel shifts the ds-cassettes with a fully methylated or hemimethylated HpaII site (Cm^{5}CGG); it fails to shift the unmethylated cassette. The band seen at the bottom of the gel in lanes 5-14 is the unbound cassette. In lanes 1-4 where the fully methylated cassette is used we see complete shifting of this bottom band, while in the lanes contiaing the hemimethylated cassette some of this band remains indicating that at the molar ratios of HpaII sites to rMcrA-S tested, we had incomplete binding of the cassette DNA. In contrast lanes 9-12 none of the unmethylated cassette is shifted with any of the cassette:rMcrA-S ratios tested. In lanes 1-4 with the fully methylated cassette and in lanes 5-8 with the hemimethylated cassette we see that there are 2 shifted bands indicated by an asterisk. One explination of these two bands is that they represent the cassttes binding to monomeric rMcrA-S, in the case of the lower of the two bands and to dimeric rMcrA-S in the case of the upper of the two bands.

We next tested rMcrA-S's ability to gel shift ds-cassettes were an A, C, T, U, or I residue is placed opposite the m^{5}C (Fig. 4.6). Interestingly, EMSA shifts were observed only when a G or I is opposite the m^{5}C; no shifted complexes are seen with the others.

Discussion

From these results we can conclude that the rMcrA-S can bind ds-DNA fragments with a single symmetrically methylated Cm^{5}CGG sequence or sites where the methylated central CpG dinucleotide is preceded by Y or followed by R. Interestingly, under the conditions we used here, rMcrA-S does not seem to have a high affinity to human genomic CGIs with methylated *HpaII* sites. Further studies with CGI-specific and other types of microarrays will be needed to determine if rMcrA-S can be used for high-resolution DNA methylation analysis of tiled CGIs. This is in contrast to affinity pull-downs with MBD2b/MBD3L1 complexes used in the MIRA assay which, under similar ionic strength conditions, prefers sites with at least two methylated CpG's within ~50 base pairs [115].

The most significant finding of this study is the discovery that rMcrA-S can gel-shift hemimethylated duplexes. In principle, this ability could be manipulated by designing bisulfite-independent-, solid state-, EMSA-, or microarray-based assays for rapid interrogation of the methylation status of specific, individual *HpaII* sites, and perhaps, NCGR sites in total human genomic DNA. For example, the CGI of PITX2, a homeodomain transcription-factor gene implicated in the progression of breast cancer, has two *HpaII* sites that, when methylated, indicate poor patient prognosis [116]. By capturing this diagnostic region from appropriately digested total genomic DNA by m^{5}CpG affinity purification [69], we might apply it, in combination with rMcrA-S binding, to determine their methylation status by ELISA using Strep-Tactin/enzyme conjugates or monoclonal antibody (StrepMAB-Classic) [112]. The treatments would involve denaturation of the captured DNA, and annealing to tethered complementary oligonucleotides with and without stepwise mismatches

opposite the diagnostic m^5C residues. Alternatively capture oligos and denatured test DNA could be annealed in solution followed by addition of rMcrA-S. Samples could then be analyzed by gel electrophoresis to separate the rMcrA-S/DNA complexes from the capture oligo as shown in Fig. 4.6. In this format the capture oligos would be biotinylated or fluorescently labeled to allow for sensitive detection of the gel shifted complexes.

An rMcrA-S-based approach has several potential advantages over the bisulfite-based methods, such as methylation-specific PCR (MSP) [69] or the more labor-intensive cloning and sequencing of bisulfite-modified DNA regions. It would eliminate all the tedious bisulfite-modification-based steps, concerns about the potential degradation of DNA during the conversion steps, and spurious results from incomplete conversion of cytosine to uracil. Finally, following the bisulfite conversion of unmethylated cytosines, there usually is a lack of C (sense) or G (antisense strand) nucleotides in the PCR products, so that these amplicons have a significantly lower GC content than the initial genomic DNA, thereby confounding the design of appropriate amplification primers. As we envisioned here, the capture oligonucleotide(s), primers, and the template DNA would contain all four bases. Their annealing would be much more specific, thereby greatly reducing the probability of nonspecific capture or random priming, particularly since stringent annealing conditions could be used routinely.

Tables and Figures

Table 4.1. HpaII sites in human male DNA and in CpG islands as defined by the UCSC web site.

Chrmosome	*HpaII Sites*	*UCSC Annotation-Islands*	*UCSC Annotation-HpaII Sites*
Chr 1	194234	2463	24371
Chr 2	166715	1680	17446
Chr 3	121254	1159	11592
Chr 4	103564	1019	10446
Chr 5	112074	1227	11796
Chr 6	110408	1251	11320
Chr 7	126732	1552	14369
Chr 8	98767	1028	10627
Chr 9	103249	1230	12248
Chr 10	110108	1150	12789
Chr 11	108817	1371	13550
Chr 12	103436	1221	11320
Chr 13	57539	605	5576
Chr 14	70014	788	7827
Chr 15	72558	787	8529
Chr 16	100892	1491	13467
Chr 17	112337	1622	16690
Chr 18	50106	508	5556
Chr 19	111044	2544	19265
Chr 20	63814	799	7996
Chr 21	30601	356	3165
Chr 22	57977	716	7430
Chr X	90237	891	8276
Chr Y	15698	181	1089
Chr M	23	0	0
Total	2292198	27639	266740

Table 4.2. List of oligonucleotides used in this study. C indicates a m^5C included in the synthesis of the oligonucleotides.

Oligo	5'-3' Sequence
1	GCCTTCAGCGCCGGCGGATCCAGT
2	ACTGGATCCGCCGGCGCTGAAGGC
3	GCCTTCAGCGCCGGCGGATCCAGT
4	ACTGGATCCGCCGGCGCTGAAGGC
5	GCCTTCAGCGCCGGCGGATCCAGT
6	ACTGGATCCGCCGGCGCTGAAGGC
7	GCCTTCAGCGCCGGCGGATCCAGT
8	ACTGGATCCGCCGGCGCTGAAGGC
9	CCCTCTGACGGAGGAGGCTCCTGC
10	GCAGGAGCCTCCTCCGTCAGAGGG
11	CCCTCTGACGCAGGAGGCTCCTGC
12	GCAGGAGCCTCCTGCGTCAGAGGG
13	CCCTCTGACGAAGGAGGCTCCTGC
14	GCAGGAGCCTCCTTCGTCAGAGGG
15	CCCTCTGACGTAGGAGGCTCCTGC
16	GCAGGAGCCTCCTCAGTCAGAGGG
17	CCCTCTGCCGGAGGAGGATCCTGC
18	GCAGGATCCTCCTCCGGCAGAGGG
19	CCCTCTGTCGGAGGAGGCTCCTGC
20	GCAGGAGCCTCCTCCGACAGAGGG
21	CCCTCTGGCGGAGGAGGCTCCTGC
22	GCAGGAGCCTCCTCCGCCAGAGGG
23	CCCTCTGTCGAAGGAGGCTCCTGC
24	GCAGGAGCCTCCTTCGACAGAGGG
25	CCCTCTGGCGAAGGAGGCTCCTGC
26	GCAGGAGCCTCCTTCGCCAGAGGG
27	CCCTCTGCCGAAGGAGGCTCCTGC
28	GCAGGAGCCTCCTTCGGCAGAGGG
29	ACTGGATCCGCCAGCGCTGAAGGC
30	ACTGGATCCGCCCGCGCTGAAGGC
31	ACTGGATCCGCCTGCGCTGAAGGC
32	ACTGGATCCGCCUGCGCTGAAGGC
33	ACTGGATCCGCCIGCGCTGAAGGC

Table 4.3. Fully base-paired oligonucleotide cassettes used for EMSA studies.

Cassette	Homo Duplex Sequences	rMcrA-S Bound
CG,CCGG, CGG	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCGGCCGCCTAGGTCA-5'	Yes
CCGG	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCGGCCGCCTAGGTCA-5'	Yes
CGG	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCGGCCGCCTAGGTCA-5'	Yes
Unmethylated	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCGGCCGCCTAGGTCA-5'	No
CGG and ACGG	5'-CCCTCTGACGGAGGAGGCTCCTGC-3' 3'-GGGAGACTGCCTCCTCCGAGGACG-5'	Yes
CGC	5'-CCCTCTGACGCAGGAGGCTCCTGC-3' 3'-GGGAGACTGCGTCCTGGCAGGACG-5'	No
CGA and ACGA	5'-CCCTCTGACGAAGGAGGCTCCTGC-3' 3'-GGGAGACTGCTTCCTCCGAGGACG-5'	Yes
CGT	5'-CCCTCTGACGTAGGAGGCTCCTGC-3' 3'-GGGAGACTGCATCCTCCGAGGACG-5'	No
CCGG	5'-CCCTCTGCCGGAGGAGGATCCTGC-3' 3'-GGGAGACGGCCTCCTCCTAGGTGC-5'	Yes
TCGG	5'-CCCTCTGTCGGAGGAGGCTCCTGC-3' 3'-GGGAGACAGCCTCCTCCGAGGACG-5'	Yes
GCGG	5'-CCCTCTGGCGGAGGAGGCTCCTGC-3' 3'-GGGAGACCGCCTCCTCCGAGGACG-5'	Yes
TCGA	5'-CCCTCTGTCGAAGGAGGCTCCTGC-3' 3'-GGGAGACAGCTTCCTCCGAGGACG-5'	Yes
GCGA	5'-CCCTCTGGCGAAGGAGGCTCCTGC-3' 3'-GGGAGACCGCTTCCTCCGAGGACG-5'	Yes
CCGA	5'-CCCTCTGCCGAAGGAGGCTCCTGC-3' 3'-GGGAGACGGCTTCCTCCGAGGACG-5'	Yes

Table 4.4 Oligonucleotide cassettes used to determine the binding of rMcrA-S to ds-cassettes with a G or a mismatched base (A, C, T, U or I in bold) opposite a m^5C (C).

Cassette	*Hetero Duplex Sequences*	*rMcrA-S Bound*
m^5C Control	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCGGCCGCCTAGGTCA-5'	Yes
C/G	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCG**G**CCGCCTAGGTCA-5'	Yes
C/A	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCG**A**CCGCCTAGGTCA-5'	No
C/C	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCG**C**CCGCCTAGGTCT-5'	No
C/T	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCC**T**CCGCCTAGGTCA-5'	No
C/U	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCG**U**CCGCCTAGGTCA-5'	No
C/I	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCG**I**CCGCCTAGGTCA-5'	Yes
Unmethylated Control	5'-GCCTTCAGCGCCGGCGGATCCAGT-3' 3'-CGGAAGTCGCGGCCGCCTAGGTCA-5'	No

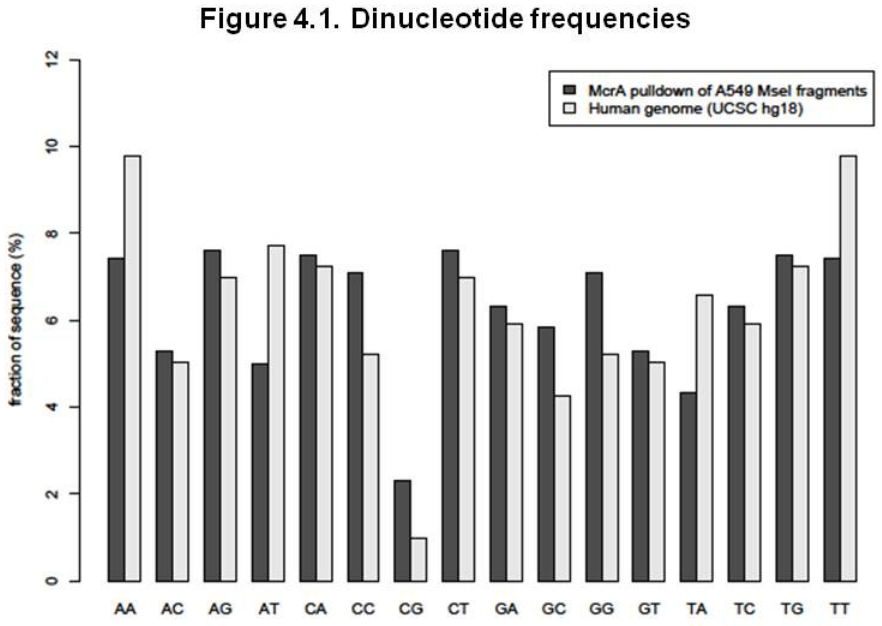


Figure 4.1. CpG ratios in human genome vs. rMcrA-s pull-downs. Dinucleotide frequencies compiled for both strands of 33 unique rMcrA-S pull-down fragments (11400 nt total), compared with total-genome frequencies compiled fromUCSChg18.

(http://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=136554479&chromInfoPage=)

Figure 4.2. Methylation Patterns of DNA fragments bound by rMcrA-S

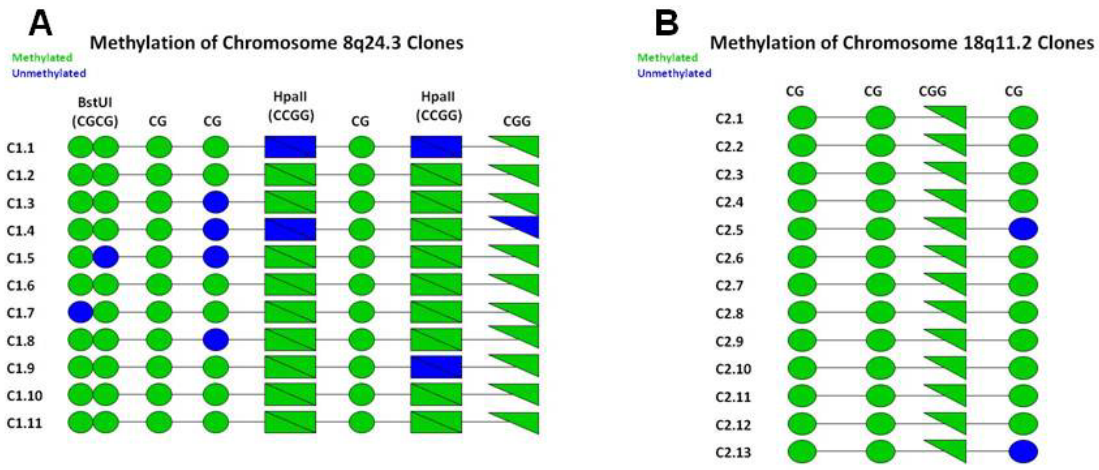


Figure 4.2. Bisulfite sequencing patterns of 2 genomic regions (C1 and C2) recovered after rMcrA-S affinity enrichment from MseI digested A549 DNA. Methylated cytosines in individual clones are Green and unmethylated cytosines are Blue. Individual circles represent CpG's, double circles BstUI sites (CGCG), triangles CGG sites and rectangles HpaII sites (CCGG). Green and blue represent C and m^5C, respectively.

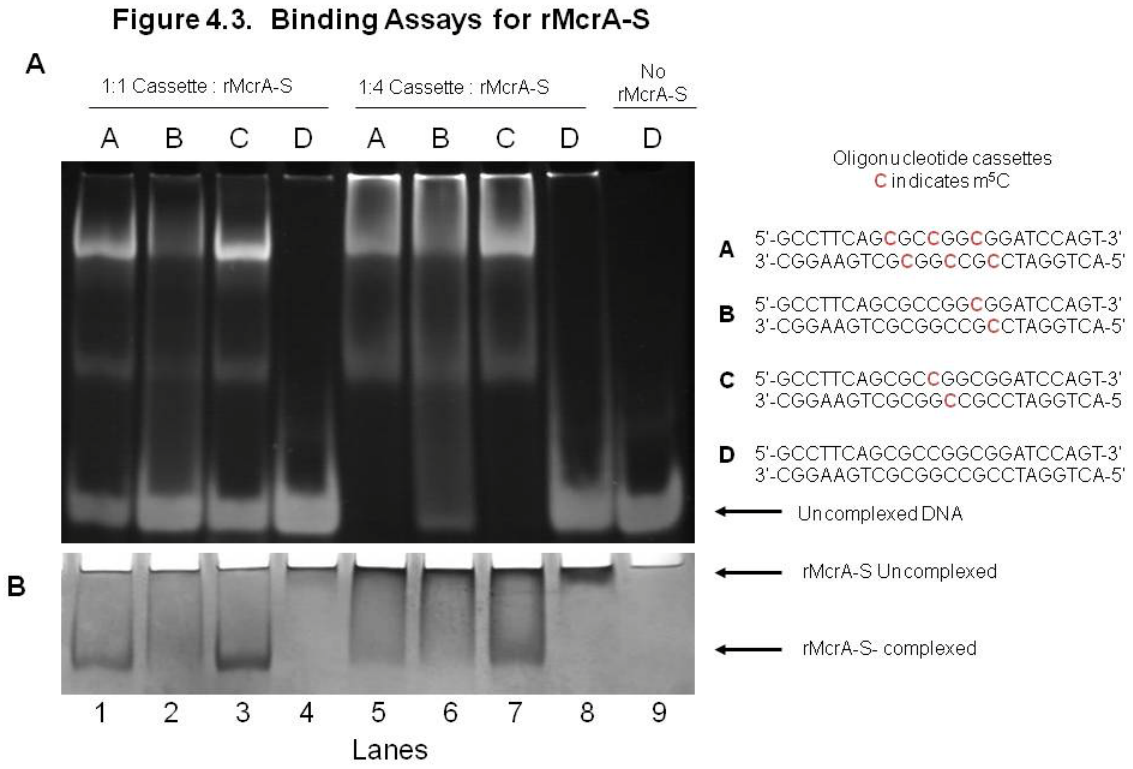


Figure 4.3. Binding Assays for rMcrA-S. (A) Electrophoretic mobility shift assays with different molar ratios (1:1 and 1:4 - DNA: protein) of rMcrA-S binding unmethylated and symmetrically methylated (m^5CpG) containing 24 bp double-stranded oligonucleotides: Cassette A contains three m^5C including one in an HpaII site; cassette B contains a single m^5C in a "3/4 HpaII site"; cassette C contains a single m^5C in an HpaII site; cassette D contains no m^5Cs. Molar ratios based on monomeric McrA [18]. **(B)** Coomassie Blue stain of EMSA (A) gel.

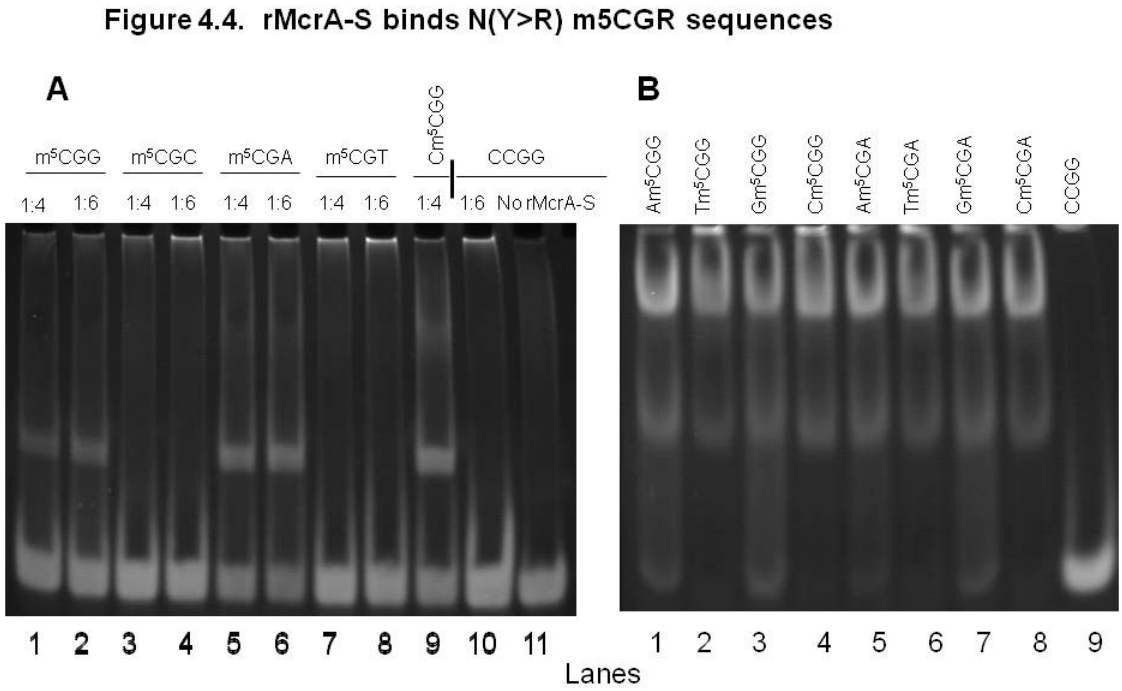


Figure 4.4. rMcrA-S binds N(Y>R) m5CGR sequences. (A) EMSA showing rMcrA-S binding 24bp oligonucleotide cassettes containing a purine following the Cm^{5}CG sequence (lanes 1, 2, 5 and 6) or with a pyrimidine followed the Cm^{5}CG (lanes 3, 4, 7 and 8). Lane 9 is a positive control using a cassette with a single symmetrically methylated HpaII site; lanes 10 & 11 are negative controls. The molar ratio of HpaII sites to rMcrA-S for each cassette is: 1:4, or 1:6 (ratios based on monomeric rMcrA-S [18]). **(B)** EMSA showing rMcrA-S binding 24bp oligonucleotide cassettes containing Nm^{5}CGR sequences. rMcrA-S binds all cassettes with this sequence (lanes 1-8). Lane 9 contains an unmethylated cassette as a negative control. The molar ratio of HpaII sites to rMcrA-S for each reaction is 1:6.

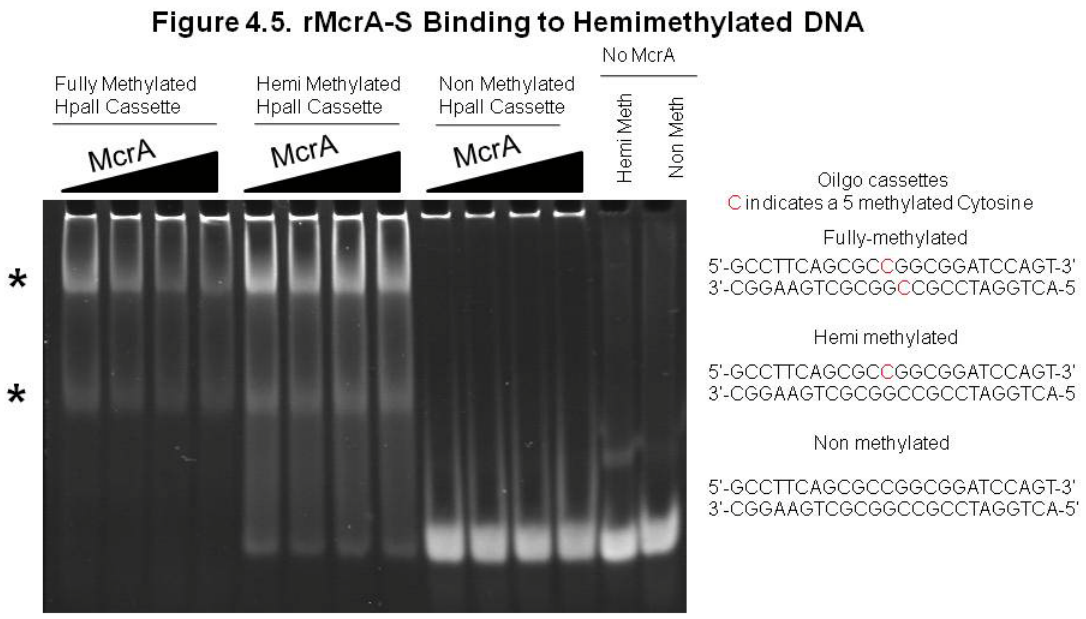


Figure 4.5. EMSA showing rMcrA-S binding to DNA with fully and hemimethylated Cm^{5}CGG sites.

The molar ratio of HpaII sites to rMcrA-S for each cassette is: 1:4, 1:5, 1:6, and 1:7; increasing from left to right. Lanes 1-4 fully methylated DNA, 5-8 hemimethylated DNA, lanes 9-12 non-methylated DNA. Lanes 13 & 14 are negative controls. The locations of the presumed DNA+protein complexes are show by asterisks. The band at the top of the gel is sonicated nonmethylated *E. coli* competitor DNA and the band at the bottom of the gel in lanes 5-14 is the unbound cassette.

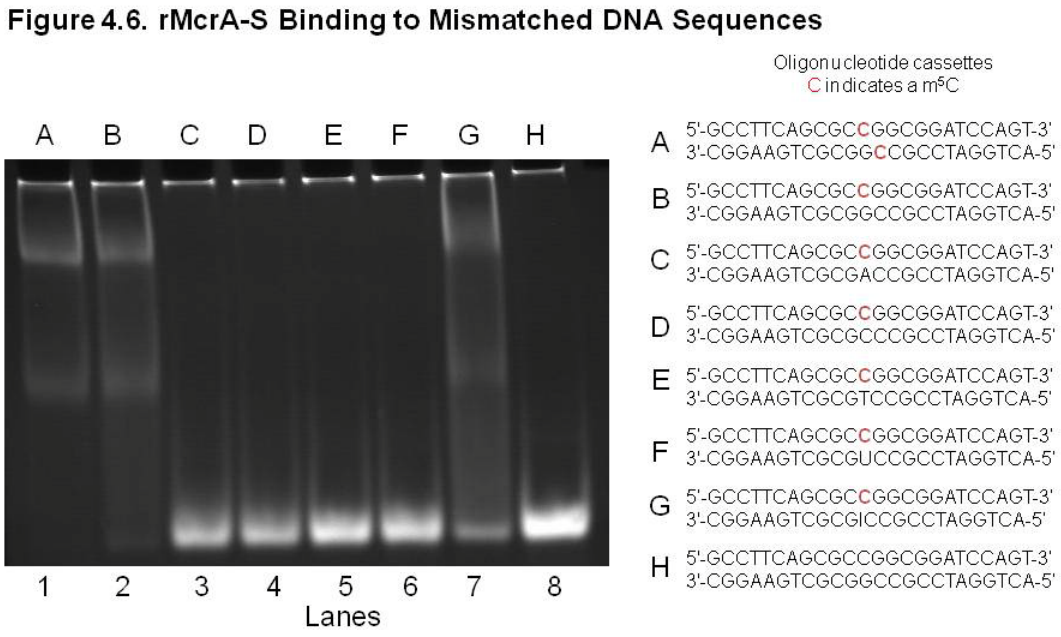


Figure 4.6. rMcrA-S Binding to DNA Sequences with a mismatch opposite m^5C. The letters A-H at the top of the gel correspond to the cassettes to the right of the figure. EMSA showing rMcrA-S binding to a DNA with a symmetrically methylated HpaII site (lane 1), a hemi-methylated HpaII site (lane 2) and a mismatch with inosine across from the m^5C (lane 7). rMcrA-S does not bind if A, C T or U is placed opposite the m^5C (lanes 3-6, respectively). The molar ratio of DNA to rMcrA-S is 1:6.

Chapter 5: Summary and Future Directions

Summary

Over the course of this dissertation we have analyzed whether the *E. coli* protein McrA has the ability to be used as a tool in cytosine methylation studies. We were able to express and purify several recombinant forms of McrA which allowed for ease of purification and study of its binding to m^5CpG containing DNA sequences.

We were successfully able to develop a simple method to purify nearly homogeneous, and biologically active, rMcrA. It was determined that rMcrA bound to methylated HpaII sites *in vitro* and it restricts the growth of incoming HpaII methylated T7 phage in cells expressing the pET28-rMcrA plasmid. Additionally rMcrA did not bind to sequences containing various other methylated sequences, including AGm^5CT (AluI), Gm^6ATC (*Dam*), GGm^5CC (HaeIII), Gm^5CGC (HhaI), and m^5CCGG (MspI). rMcrA preferentially bound HpaII methylated sequences over unmethylated sequences employing EMSA methods. These initial experiments did not include sufficient probes to determine the exact sequence(s) recognized by rMcrA. Also, no nuclease activity was seen for rMcrA, however we could not rule out the inactivation of nuclease activity during purification. We think this is unlikely due to the fact that rMcrA retained the ability to selectively bind HpaII methylated DNA. It is also possible that an unknown co-factor is required for nuclease activity.

With the addition of a C-terminal StrepII tag we were able to develop an efficient method for purification of rMcrA-S that we could then study its efficacy as a tool for enrichment of methylated DNA fragments. This recombinant form of the McrA protein was able to bind methylated HpaII sequences in human DNA cut with the restriction enzyme MseI. Interestingly we found that in addition to fragments containing methylated HpaII sequences Cm^5CGG, rMcrA-

S bound fragments that did not contain any methylated HpaII sites but instead contained Am^{5}CGG sequences. This gave us our first evidence that Cm^{5}CGG was not the only sequence bound by rMcrA-S.

Further investigation into the minimal binding site for rMcrA-S lead us to develop a series of 24 bp methylated oligonucleotide cassettes with which we could interrogate the nucleotides 5′ and 3′ of the m^{5}CpG dinucleotide. Through a series of EMSA experiments we determined the minimal binding site for rMcrA-S to be N(Y>R) m^{5}CGR. Additionally rMcrA-S was found to bind cassettes containing only a single methylated site and that it preferentially bound these sites over unmethylated sequences. Hemimethylated sequences were also bound by rMcrA-S as was a mismatch of the m^{5}C with inosine but not with mismatches of A, C, T, or U. This in particular we think will be helpful in developing an assay for the detection of m^{5}C′s in CGI′s without the use of sodium bisulfite. Since rMcrA-S cannot bind to a hemimethylated mismatch with A, C, T, or U oligonucleotide sequences can be annealed to enriched methylated DNA fragments with one of these bases across all but one possible methylcytosine site in the context of a rMcrA-S binding sequence. If rMcrA-S is still able to bind the resulting cassette it indicates that the one non mismatched site is methylated. Subsequent use of different oligonucleotide sequences will allow for the interrogation of the methylation status of each cytosine within Nm^{5}CGR sites of the enriched fragments. This is discussed in more detail in the future directions section of this chapter.

The data obtained in these experiments is not quantitative with respect to the amount of methylated DNA bound to rMcrA-S. One method for making this type of data more quantitative would be to use gel scanning techniques to ascertain the intensities of the bands on the EMSA gels. There are however

variations in the amount of ethidium bromide that intercalates into the DNA depending on the sequence of the DNA [117-118] and this technique's accuracy is dependent on the ratio of DNA:ethidium bromide [119].

The research detailed in this dissertation supports the hypothesis that rMcrA-S can be used to identify and enrich for methylated DNA fragments. One drawback in using rMcrA-S to enrich for methylated DNA fragments is that many of the m^5CpG containing fragments enriched by rMcrA-S are in repetitive sequences, but this could be mitigated by performing sequential enrichments with other methyl binding proteins. For example rMcrA-S could be used in conjunction with methyl binding domain (MBD) of the rat protein MeCP2. This MBD binds methylated Gm^5CGC sites [69]. By doing successive enrichments with both of these proteins it is conceivable that an increased amount of CGI's would be enriched compared to either of these proteins on their own. Based on the results gained by this dissertation project, continuing research into the development of a sodium bisulfite free method of interrogating individual CpG sites for methylation using rMcrA-S is warranted.

Future Directions

Further work is needed to develop a method for interrogating the methylation status of individual CpG's within a CGI. Also basic McrA *in vivo* complementation can be done to determine whether or not McrA has nuclease activity *in vivo* and if a hitherto unidentified *E. coli* co-factor is needed to cleave DNA.

rMcrA-S Based Assay for CpG Methylation Identification

We have already discussed the idea of using rMcrA-S in conjunction with other methyl binding proteins or domains to enrich for methylated CGI's. This type of assay could be used in the determination of the CGI's that are methylated before and after a particular DNA insult such as low dose ionizing radiation, or drug treatment. The enriched DNA fragments would need to be hybridized to arrays containing a spectrum of sequences containing *HpaII* sites in order to determine which CGI's are methylated after a given insult.

As a tool for interrogating the methylation status of cytosines in CpG dinucleotides, the ability of rMcrA-S to bind hemimethylated DNA sequences can be exploited. Currently our lab is developing an assay in which rMcrA-S is used to determine the methylation status of cytosines within Nm^5CGR sites without the use of a sodium bisulfite modification step. Methylcytosine containing DNA once purified can be denatured and annealed to oligonucleotide sequences containing mismatches (A, C, T, or U) opposite the cytosine of different CpG dinucleotides with Nm^5CGR sequences. These sequences would create a region of ds-DNA in which only one cytosine base in an Nm^5CGR site is not mismatched. If this cytosine is methylated it would create a hemimethylated Nm^5CGR site where rMcrA-S could bind. All the other cytosines within rMcrA-S binding sequences, even if methylated, are mismatched and therefore rMcrA-S is unable to bind these sites. After annealing either S1 nuclease or mung bean nuclease can be used to digest away the single stranded regions of the original fragment DNA that is to either end of the cassette. These ds-cassettes can then be probed with rMcrA-S to look for binding at the individual cytosines that have been left without a mismatch. Binding would indicate methylation at that site. Figure 5.1 illustrates this type of assay using the PITX2 CGI [115] as an example.

Nuclease Activity of rMcrA

Our work thus far has been unable to determine conclusively whether or not rMcrA has nuclease activity. We plan to expand upon the experiments described in chapter 3 in order to solve this question, and use a southern blot technique in order to see if there is DNA degradation.

We would like to make a set of ^{32}P labeled probes to two *E. coli* housekeeping genes katG and recA along with a probe for the rMcrA gene on the pREXLS31 vector as it contains multiple HpaII sites. As in chapter 3 BL21-AI cells contain pACYC-M.HpaII and pREXLS31-rMcrA, and a control of BL21-AI with pREXLS31-rMcrA would be grown to log phase and then induced with IPTG and arabanose. Cells removed from culture at successive time points would be used for total DNA extraction and then run out on 10% acrylamide gels. These gels would be probed with the above described probes to look for a loss of the signal over the time course. This would indicate DNA degradation and therefore nuclease activity of rMcrA.

Figures and Tables

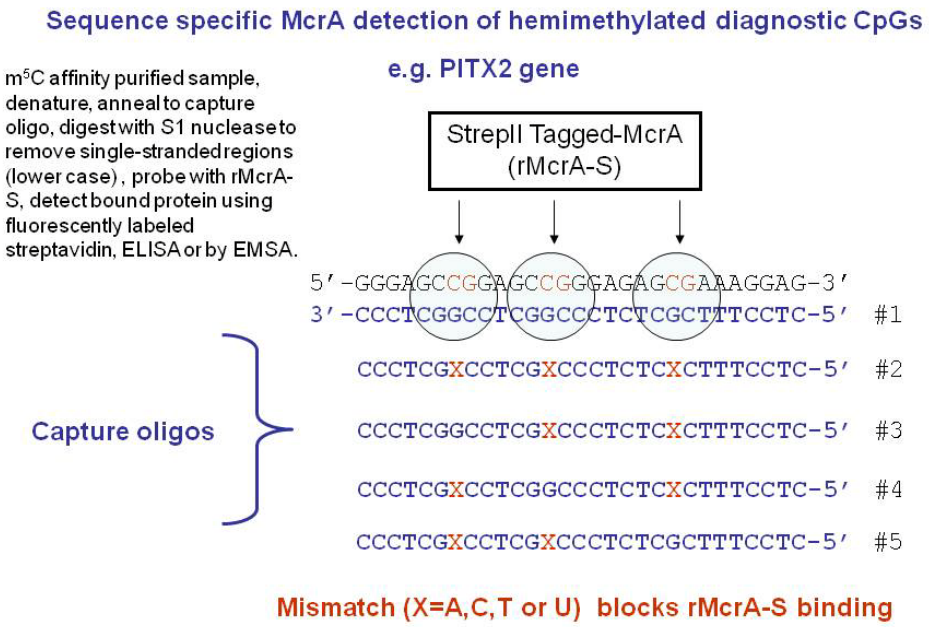


Figure 5.1. rMcrA-S is expected to bind to sites 1 and 2 and possibly 3 if they are methylated in the starting DNA sample and captured by oligonucleotide sequence #1. No binding should be seen if the PITX2 CGI is unmethylated at all three sites. Likewise, no binding should be seen if either methylated or unmethylated genomic DNA is captured by oligonucleotide sequence #2 since mismatches are opposite all three potential m^5C residues. This oligonucleotide sequence serves as a negative control. The methylation status of each individual site can be differentiated using capture oligonucleotide sequences #3, #4 and #5. If all three sites are methylated then oligonucleotide sequences #3, #4 and #5 should give a positive signal. If only site 1 is methylated then a positive signal should be seen only when using oligonucleotide sequences #1 and #3 for capture; if only site 2 is methylated then a positive signal should be seen only when using oligonucleotide sequences #1 and #4 for capture: if only site #3 is methylated then it should only give a positive signal when using oligonucleotide sequence #1 and #5.

References

1. *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
2. Berger, S.L., et al., *An operational definition of epigenetics.* Genes & Development, 2009. **23**: p. 781-783.
3. Baylin, S.B. and J.G. Herman, *DNA hypermethylation in tumorigenesis: epigenetics joins genetics.* Trends Genet, 2000. **16**(4): p. 168-74.
4. Chow, J.C. and C.J. Brown, *Forming facultative heterochromatin: silencing of an X chromosome in mammalian females.* Cell Mol Life Sci, 2003. **60**(12): p. 2586-603.
5. Ehrlich, M., *Expression of various genes is controlled by DNA methylation during mammalian development.* J Cell Biochem, 2003. **88**(5): p. 899-910.
6. Jones, P.A. and P.W. Laird, *Cancer epigenetics comes of age.* Nat Genet, 1999. **21**(2): p. 163-7.
7. Rideout, W.M., 3rd, et al., *5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes.* Science, 1990. **249**(4974): p. 1288-90.
8. Robertson, A.K., et al., *Effects of chromatin structure on the enzymatic and DNA binding functions of DNA methyltransferases DNMT1 and Dnmt3a in vitro.* Biochem Biophys Res Commun, 2004. **322**(1): p. 110-8.
9. Robertson, K.D., *DNA methylation, methyltransferases, and cancer.* Oncogene, 2001. **20**(24): p. 3139-55.
10. Robertson, K.D. and A.P. Wolffe, *DNA methylation in health and disease.* Nat Rev Genet, 2000. **1**(1): p. 11-9.
11. Ting, A.H., et al., *Mammalian DNA methyltransferase 1: inspiration for new directions.* Cell Cycle, 2004. **3**(8): p. 1024-6.
12. Chuang, L.S., et al., *Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1.* Science, 1997. **277**(5334): p. 1996-2000.
13. Okano, M., S. Xie, and E. Li, *Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells.* Nucl. Acids Res., 1998. **26**(11): p. 2536-2540.

14. Ehrlich, M., et al., *Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells.* Nucleic Acids Res, 1982. **10**(8): p. 2709-21.
15. Esteller, M. and J.G. Herman, *Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours.* J Pathol, 2002. **196**(1): p. 1-7.
16. Illingworth, R.S. and A.P. Bird, *CpG islands - [`]A rough guide'.* FEBS Letters, 2009. **583**(11): p. 1713-1720.
17. Larsen, F., et al., *CpG islands as gene markers in the human genome.* Genomics, 1992. **13**(4): p. 1095-1107.
18. Chen, D., et al., *T:G Mismatch-specific Thymine-DNA Glycosylase Potentiates Transcription of Estrogen-regulated Genes through Direct Interaction with Estrogen Receptor {alpha}.* J. Biol. Chem., 2003. **278**(40): p. 38586-38592.
19. Hardeland, U., et al., *Modification of the human thymine-DNA glycosylase by ubiquitin-like proteins facilitates enzymatic turnover.* EMBO J, 2002. **21**(6): p. 1456-64.
20. Adams, R.L. and R. Eason, *Increased G + C content of DNA stabilizes methyl CpG dinucleotides.* Nucleic Acids Res, 1984. **12**(14): p. 5869-77.
21. Bird, A., *DNA methylation patterns and epigenetic memory.* Genes Dev, 2002. **16**(1): p. 6-21.
22. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes.* J Mol Biol, 1987. **196**(2): p. 261-82.
23. Jones, P.A. and S.B. Baylin, *The fundamental role of epigenetic events in cancer.* Nat Rev Genet, 2002. **3**(6): p. 415-28.
24. Illingworth, R., et al., *A novel CpG island set identifies tissue-specific methylation at developmental gene loci.* PLoS Biol, 2008. **6**(1): p. e22.
25. Takai, D. and P.A. Jones, *The CpG island searcher: a new WWW resource.* In Silico Biol, 2003. **3**(3): p. 235-40.
26. Shen, L., et al., *Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters.* PLoS Genet, 2007. **3**(10): p. 2023-36.

27. Weber, M., et al., *Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells.* Nat Genet, 2005. **37**(8): p. 853-62.

28. Deimling, A.v., ed. *Gliomas*. Vol. Recent Results in Cancer Research. 2009, Springer Berlin Heidelberg. 217-239.

29. Stancheva, I., C. Hensey, and R.R. Meehan, *Loss of the maintenance methyltransferase, xDnmt1, induces apoptosis in Xenopus embryos.* EMBO J, 2001. **20**(8): p. 1963-73.

30. Jackson-Grusby, L., et al., *Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation.* Nat Genet, 2001. **27**(1): p. 31-9.

31. Panning, B. and R. Jaenisch, *DNA hypomethylation can activate Xist expression and silence X-linked genes.* Genes Dev, 1996. **10**(16): p. 1991-2002.

32. Soppe, W.J., et al., *DNA methylation controls histone H3 lysine 9 methylation and heterochromatin assembly in Arabidopsis.* EMBO J, 2002. **21**(23): p. 6549-59.

33. Proffitt, J.H., et al., *5-Methylcytosine is not detectable in Saccharomyces cerevisiae DNA.* Mol Cell Biol, 1984. **4**(5): p. 985-8.

34. Urieli-Shoval, S., et al., *The absence of detectable methylated bases in Drosophila melanogaster DNA.* FEBS Lett, 1982. **146**(1): p. 148-52.

35. Hung, M.S., et al., *Drosophila proteins related to vertebrate DNA (5-cytosine) methyltransferases.* Proc Natl Acad Sci U S A, 1999. **96**(21): p. 11940-5.

36. Gowher, H., O. Leismann, and A. Jeltsch, *DNA of Drosophila melanogaster contains 5-methylcytosine.* EMBO J, 2000. **19**(24): p. 6918-23.

37. Lyko, F., B.H. Ramsahoye, and R. Jaenisch, *DNA methylation in Drosophila melanogaster.* Nature, 2000. **408**(6812): p. 538-40.

38. Phalke, S., et al., *Retrotransposon silencing and telomere integrity in somatic cells of Drosophila depends on the cytosine-5 methyltransferase DNMT2.* Nat Genet, 2009. **41**(6): p. 696-702.

39. Bird, A.P., *CpG-rich islands and the function of DNA methylation.* Nature, 1986. **321**(6067): p. 209-213.

40. Rountree, M.R., et al., *DNA methylation, chromatin inheritance, and cancer.* Oncogene, 2001. **20**(24): p. 3156-65.

41. Greger, V., et al., *Frequency and parental origin of hypermethylated RB1 alleles in retinoblastoma.* Hum Genet, 1994. **94**(5): p. 491-6.

42. Greger, V., et al., *Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma.* Hum Genet, 1989. **83**(2): p. 155-8.

43. Ohtani-Fujita, N., et al., *CpG methylation inactivates the promoter activity of the human retinoblastoma tumor-suppressor gene.* Oncogene, 1993. **8**(4): p. 1063-7.

44. Sakai, T., et al., *Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene.* Am J Hum Genet, 1991. **48**(5): p. 880-8.

45. Pegg, A.E., *Repair of O(6)-alkylguanine by alkyltransferases.* Mutat Res, 2000. **462**(2-3): p. 83-100.

46. Esteller, M., et al., *A gene hypermethylation profile of human cancer.* Cancer Res, 2001. **61**(8): p. 3225-9.

47. Feinberg, A.P. and B. Vogelstein, *Hypomethylation distinguishes genes of some human cancers from their normal counterparts.* Nature, 1983. **301**(5895): p. 89-92.

48. Gama-Sosa, M.A., et al., *The 5-methylcytosine content of DNA from human tumors.* Nucleic Acids Res, 1983. **11**(19): p. 6883-94.

49. Feinberg, A.P., et al., *Reduced genomic 5-methylcytosine content in human colonic neoplasia.* Cancer Res, 1988. **48**(5): p. 1159-61.

50. Goelz, S.E., et al., *Hypomethylation of DNA from benign and malignant human colon neoplasms.* Science, 1985. **228**(4696): p. 187-90.

51. Strichman-Almashanu, L.Z., et al., *A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes.* Genome Res, 2002. **12**(4): p. 543-54.

52. Akiyama, Y., et al., *Cell-type-specific repression of the maspin gene is disrupted frequently by demethylation at the promoter region in gastric intestinal metaplasia and cancer cells.* Am J Pathol, 2003. **163**(5): p. 1911-9.

53. Badal, V., et al., *CpG methylation of human papillomavirus type 16 DNA in cervical cancer cell lines and in clinical specimens: genomic hypomethylation correlates with carcinogenic progression.* J Virol, 2003. **77**(11): p. 6227-34.

54. Cho, M., et al., *Hypomethylation of the MN/CA9 promoter and upregulated MN/CA9 expression in human renal cell carcinoma.* Br J Cancer, 2001. **85**(4): p. 563-7.

55. de Capoa, A., et al., *DNA demethylation is directly related to tumour progression: evidence in normal, pre-malignant and malignant cells from uterine cervix samples.* Oncol Rep, 2003. **10**(3): p. 545-9.

56. Iacobuzio-Donahue, C.A., et al., *Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays.* Am J Pathol, 2003. **162**(4): p. 1151-62.

57. Oshimo, Y., et al., *Promoter methylation of cyclin D2 gene in gastric carcinoma.* Int J Oncol, 2003. **23**(6): p. 1663-70.

58. Piyathilake, C.J., et al., *Race- and age-dependent alterations in global methylation of DNA in squamous cell carcinoma of the lung (United States).* Cancer Causes Control, 2003. **14**(1): p. 37-42.

59. Sato, N., et al., *Frequent hypomethylation of multiple genes overexpressed in pancreatic ductal adenocarcinoma.* Cancer Res, 2003. **63**(14): p. 4158-66.

60. Feinberg, A.P. and B. Vogelstein, *Hypomethylation of ras oncogenes in primary human cancers.* Biochem Biophys Res Commun, 1983. **111**(1): p. 47-54.

61. Feinberg, A.P. and B. Tycko, *The history of cancer epigenetics.* Nat Rev Cancer, 2004. **4**(2): p. 143-53.

62. Wilson, A.S., B.E. Power, and P.L. Molloy, *DNA hypomethylation and human diseases.* Biochim Biophys Acta, 2007. **1775**(1): p. 138-62.

63. Goldstone, A.P., *Prader-Willi syndrome: advances in genetics, pathophysiology and treatment.* Trends Endocrinol Metab, 2004. **15**(1): p. 12-20.

64. Buntinx, I.M., et al., *Clinical profile of Angelman syndrome at different ages.* Am J Med Genet, 1995. **56**(2): p. 176-83.

65. Horsthemke, B. and J. Wagstaff, *Mechanisms of imprinting of the Prader-Willi/Angelman region.* Am J Med Genet A, 2008. **146A**(16): p. 2041-52.

66. Sahoo, T., et al., *Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster.* Nat Genet, 2008. **40**(6): p. 719-21.

67. Van Buggenhout, G. and J.-P. Fryns, *Angelman syndrome (AS, MIM 105830).* Eur J Hum Genet, 2009.

68. Baylin, S.B., et al., *Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer.* Hum Mol Genet, 2001. **10**(7): p. 687-92.

69. Cross, S.H., et al., *Purification of CpG islands using a methylated DNA binding column.* Nat Genet, 1994. **6**(3): p. 236-44.

70. Gebhard, C., et al., *Rapid and sensitive detection of CpG-methylation using methyl-binding (MB)-PCR.* Nucl. Acids Res., 2006. **34**(11): p. e82-.

71. Gebhard, C., et al., *Genome-Wide Profiling of CpG Methylation Identifies Novel Targets of Aberrant Hypermethylation in Myeloid Leukemia.* Cancer Res, 2006. **66**(12): p. 6118-6128.

72. Frommer, M., et al., *A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.* Proc Natl Acad Sci U S A, 1992. **89**(5): p. 1827-31.

73. Hayatsu, H., Y. Wataya, and K. Kai, *Addition of sodium bisulfite to uracil and to cytosine.* Journal of the American Chemical Society, 2002. **92**(3): p. 724-726.

74. Shapiro, R., R. Servis, and M. Welcher, *Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite.* Journal of the American Chemical Society, 2002. **92**(2): p. 422-424.

75. Tanaka, K. and A. Okamoto, *Degradation of DNA by bisulfite treatment.* Bioorg Med Chem Lett, 2007. **17**(7): p. 1912-5.

76. Li, L.C., *Designing PCR primer for DNA methylation mapping.* Methods Mol Biol, 2007. **402**: p. 371-84.

77. *New England Biolabs Catalog and Technical Reference*. 2009-2010. p. 294.

78. Rauch, T. and G.P. Pfeifer, *Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer.* Lab Invest, 2005. **85**(9): p. 1172-80.

79. Raleigh, E.A., et al., *Nomenclature relating to restriction of modified DNA in Escherichia coli.* J Bacteriol, 1991. **173**(8): p. 2707-9.

80. Raleigh, E.A., et al., *McrA and McrB restriction phenotypes of some E.coli strains and implications for gene cloning.* Nucl. Acids Res., 1988. **16**(4): p. 1563-1575.

81. Raleigh, E.A. and G. Wilson, *Escherichia coli K-12 restricts DNA containing 5-methylcytosine.* Proceedings of the National Academy of Sciences of the United States of America, 1986. **83**(23): p. 9070-9074.

82. Hiom, K. and S.G. Sedgwick, *Cloning and structural characterization of the mcrA locus of Escherichia coli.* J Bacteriol, 1991. **173**(22): p. 7368-73.

83. Raleigh, E.A., R. Trimarchi, and H. Revel, *Genetic and physical mapping of the mcrA (rglA) and mcrB (rglB) loci of Escherichia coli K-12.* Genetics, 1989. **122**(2): p. 279-96.

84. Mulligan, E.A. and J.J. Dunn, *Cloning, purification and initial characterization of E. coli McrA, a putative 5-methylcytosine-specific nuclease.* Protein Expr Purif, 2008.

85. Anton, B.P. and E.A. Raleigh, *Transposon-mediated linker insertion scanning mutagenesis of the Escherichia coli McrA endonuclease.* J Bacteriol, 2004. **186**(17): p. 5699-707.

86. Turek-Plewa, J. and P.P. Jagodzinski, *The role of mammalian DNA methyltransferases in the regulation of gene expression.* Cell Mol Biol Lett, 2005. **10**(4): p. 631-47.

87. Dunn, J.J. and F.W. Studier, *Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements.* J Mol Biol, 1983. **166**(4): p. 477-535.

88. Studier, F.W., *Protein production by auto-induction in high density shaking cultures.* Protein Expr Purif, 2005. **41**(1): p. 207-34.

89. Whitmore, L. and B.A. Wallace, *DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W668-73.

90. Whitmore, L. and B.A. Wallace, *Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases.* Biopolymers, 2008. **89**(5): p. 392-400.

91. Zilberman, D., et al., *Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription.* Nat Genet, 2007. **39**(1): p. 61-9.

92. Rauch, T., et al., *MIRA-assisted microarray analysis, a new technology for the determination of DNA methylation patterns, identifies frequent methylation of homeodomain-containing genes in lung cancer cells.* Cancer Res, 2006. **66**(16): p. 7939-47.

93. Tarasova, G.V., T.N. Nayakshina, and S.K. Degtyarev, *Substrate specificity of new methyl-directed DNA endonuclease GlaI.* BMC Mol Biol, 2008. **9**: p. 7.

94. Heitman, J. and P. Model, *Site-specific methylases induce the SOS DNA repair response in Escherichia coli.* J. Bacteriol., 1987. **169**(7): p. 3243-3250.

95. Waite-Rees, P.A., et al., *Characterization and expression of the Escherichia coli Mrr restriction system.* J Bacteriol, 1991. **173**(16): p. 5207-19.

96. Fleischman, R.A., J.L. Cambell, and C.C. Richardson, *Modification and restriction of T-even bacteriophages. In vitro degradation of deoxyribonucleic acid containing 5-hydroxymethylctosine.* J Biol Chem, 1976. **251**(6): p. 1561-70.

97. Stewart, F.J. and E.A. Raleigh, *Dependence of McrBC cleavage on distance between recognition elements.* Biol Chem, 1998. **379**(4-5): p. 611-6.

98. Bujnicki, J.M., M. Radlinska, and L. Rychlewski, *Atomic model of the 5-methylcytosine-specific restriction enzyme McrA reveals an atypical zinc finger and structural similarity to betabetaalphaMe endonucleases.* Mol Microbiol, 2000. **37**(5): p. 1280-1.

99. Ramalingam, R., et al., *Molecular cloning and sequencing of mcrA locus and identification of McrA protein in Escherichia coli.* J. Biosci., 1992. **17**: p. 217-232.

100. Shivapriya, R., et al., *Expression of the mcrA gene of escherichia coli is regulated posttranscriptionally, possibly by sequestration of the Shine-Dalgarno region.* Gene, 1995. **157**(1-2): p. 201-207.

101. Sreerama, N. and R.W. Woody, *A self-consistent method for the analysis of protein secondary structure from circular dichroism.* Anal Biochem, 1993. **209**(1): p. 32-44.

102. Agius, F., A. Kapoor, and J.K. Zhu, *Role of the Arabidopsis DNA glycosylase/lyase ROS1 in active DNA demethylation.* Proc Natl Acad Sci U S A, 2006. **103**(31): p. 11796-801.

103. Morales-Ruiz, T., et al., *DEMETER and REPRESSOR OF SILENCING 1 encode 5-methylcytosine DNA glycosylases.* Proc Natl Acad Sci U S A, 2006. **103**(18): p. 6853-8.

104. Zhu, J., et al., *The DNA glycosylase/lyase ROS1 functions in pruning DNA methylation patterns in Arabidopsis.* Curr Biol, 2007. **17**(1): p. 54-9.

105. Cranenburgh, R.M., K.S. Lewis, and J.A. Hanak, *Effect of plasmid copy number and lac operator sequence on antibiotic-free plasmid selection by operator-repressor titration in Escherichia coli.* J Mol Microbiol Biotechnol, 2004. **7**(4): p. 197-203.

106. Fraga, M.F., et al., *Epigenetic differences arise during the lifetime of monozygotic twins.* Proc Natl Acad Sci U S A, 2005. **102**(30): p. 10604-9.

107. Reik, W., *Stability and flexibility of epigenetic gene regulation in mammalian development.* Nature, 2007. **447**(7143): p. 425-32.

108. Walsh, C.P. and T.H. Bestor, *Cytosine methylation and mammalian development.* Genes Dev, 1999. **13**(1): p. 26-34.

109. Weaver, I.C., et al., *Epigenetic programming by maternal behavior.* Nat Neurosci, 2004. **7**(8): p. 847-54.

110. Tost, J., ed. *Epigenetics*. 2008, Caister Academic Press, Book Systems Plus.

111. Herman, J.G., et al., *Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands.* Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9821-6.

112. Pomraning, K.R., K.M. Smith, and M. Freitag, *Genome-wide high throughput analysis of DNA methylation in eukaryotes.* Methods, 2009. **47**(3): p. 142-50.

113. Schmidt, T.G. and A. Skerra, *The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins.* Nat Protoc, 2007. **2**(6): p. 1528-35.

114. Bolden, A.H., et al., *The primary DNA sequence determines in vitro methylation by mammalian DNA methyltransferases.* Prog Nucleic Acid Res Mol Biol, 1986. **33**: p. 231-50.

115. Maier, S., et al., *DNA-methylation of the homeodomain transcription factor PITX2 reliably predicts risk of distant disease recurrence in tamoxifen-treated, node-negative breast cancer patients--Technical and clinical validation in a multi-centre setting in collaboration with the European Organisation for Research and Treatment of Cancer (EORTC) PathoBiology group.* Eur J Cancer, 2007. **43**(11): p. 1679-86.

116. Rauch, T.A., et al., *High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer.* Proc Natl Acad Sci U S A, 2008. **105**(1): p. 252-7.

117. Krugh, T.R. and C.G. Reinhardt, *Evidence for sequence preferences in the intercalative binding of ethidium bromide to dinucleoside monophosphates.* J Mol Biol, 1975. **97**(2): p. 133-62.

118. Krugh, T.R., F.N. Wittlin, and S.P. Cramer, *Ethidium bromide-dinucleotide complexes. Evidence for intercalation and sequence preferences in binding to double-stranded nucleic acids.* Biopolymers, 1975. **14**(1): p. 197-210.

119. Dutton, M.D., R.J. Varhol, and D.G. Dixon, *Technical considerations for the use of ethidium bromide in the quantitative analysis of nucleic acids.* Anal Biochem, 1995. **230**(2): p. 353-5.

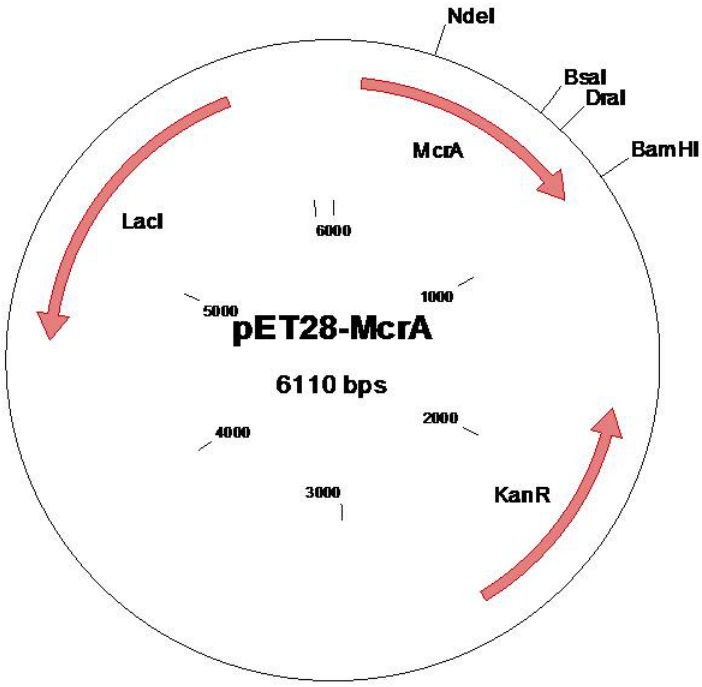
Appendix

Legend:

Red bases = location of PCR primers

Green dotted lines and letters under sequences = translated sequences

Blue Letters above sequences = restriction enzyme recognition sites




```
   1  cccgcgaaat taatacgact cactataggg gaattgtgag cggataacaa
      gggcgcttta attatgctga gtgatatccc cttaacactc gcctattgtt
                          1517

  51  ttcccctcta gaaataattt tgtttaactt taagaaggag atataccatg
      aaggggagat ctttattaaa acaaattgaa attcttcctc tatatggtac
                                                 McrA >>>
                                                       m

 101  catgtttttg ataataatgg aattgaactg aaagctgagt gttcgatagg
      gtacaaaaac tattattacc ttaacttgac tttcgactca caagctatcc
      >.......................McrA.......................>
        h  v  f   d  n  n   g  i  e  l   k  a  e   c  s  i
```

```
151  tgaagaggat ggtgtttatg gtctaatcct tgagtcgtgg gggccgggtg
     acttctccta ccacaaatac cagattagga actcagcacc cccggcccac
     >.......................McrA.......................>
     g  e  e  d   g  v  y   g  l  i   l  e  s  w   g  p  g

  201  acagaaacaa agattacaat atcgctcttg attatatcat tgaacggttg
     tgtctttgtt tctaatgtta tagcgagaac taatatagta acttgccaac
     >.......................McrA.......................>
      d  r  n   k  d  y  n   i  a  l   d  y  i   i  e  r  l

251  gttgattctg gtgtatccca agtcgtagta tatctggcgt catcatcagt
     caactaagac cacatagggt tcagcatcat atagaccgca gtagtagtca
     >.......................McrA.......................>
       v  d  s   g  v  s   q  v  v  v   y  l  a   s  s  s

              NdeI
            --+----
301  cagaaaacat atgcattctt tggatgaaag aaaaatccat cctggtgaat
     gtcttttgta tacgtaagaa acctactttc tttttaggta ggaccactta
     >.......................McrA.......................>
     v  r  k  h   m  h  s   l  d  e   r  k  i  h   p  g  e

351  attttacttt gattggtaat agcccccgcg atatacgctt gaagatgtgt
     taaaatgaaa ctaaccatta tcgggggcgc tatatgcgaa cttctacaca
     >.......................McrA.......................>
      y  f  t   l  i  g  n   s  p  r   d  i  r   l  k  m  c

401  ggttatcagg cttattttag tcgtacgggg agaaaggaaa ttccttccgg
     ccaatagtcc gaataaaatc agcatgcccc tctttccttt aaggaaggcc
     >.......................McrA.......................>
       g  y  q   a  y  f   s  r  t  g   r  k  e   i  p  s

451  caatagaacg aaacgaatat tgataaatgt tccaggtatt tatagtgaca
     gttatcttgc tttgcttata actatttaca aggtccataa atatcactgt
     >.......................McrA.......................>
     g  n  r  t   k  r  i   l  i  n   v  p  g  i   y  s  d

501  gtttttgggc gtctataata cgtggagaac tatcagagct ttcacagcct
     caaaaacccg cagatattat gcacctcttg atagtctcga aagtgtcgga
     >.......................McrA.......................>
      s  f  w   a  s  i  i   r  g  e   l  s  e   l  s  q  p

551  acagatgatg aatcgcttct gaatatgagg gttagtaaat taattaagaa
     tgtctactac ttagcgaaga cttatactcc caatcattta attaattctt
     >.......................McrA.......................>
       t  d  d   e  s  l   l  n  m  r   v  s  k   l  i  k

601  aacgttgagt caacccgagg gctccaggaa accagttgag gtagaaagac
     ttgcaactca gttgggctcc cgaggtcctt tggtcaactc catctttctg
     >.......................McrA.......................>
     k  t  l  s   q  p  e   g  s  r   k  p  v  e   v  e  r

                        BsaI
                       -------
```

```
 651  tacaaaaagt ttatgtccga gacccgatgg taaaagcttg gattttacag
      atgtttttca aatacaggct ctgggctacc attttcgaac ctaaaatgtc
      >.......................McrA.......................>
       l  q  k   v  y  v  r   d  p  m   v  k  a   w  i  l  q


 701  caaagtaaag gtatatgtga aaactgtggt aaaaatgctc cgttttattt
      gtttcatttc catatacact tttgacacca tttttacgag gcaaaataaa
      >.......................McrA.......................>
        q  s  k   g  i  c   e  n  c  g   k  n  a   p  f  y

    DraI
   ---+--
 751  aaatgatgga aacccatatt tggaagtaca tcatgtaatt cccctgtctt
      tttactacct ttgggtataa accttcatgt agtacattaa ggggacagaa
      >.......................McrA.......................>
      l  n  d  g   n  p  y   l  e  v   h  h  v  i   p  l  s

 801  caggtggtgc tgatacaaca gataactgtg ttgccctttg tccgaattgc
      gtccaccacg actatgttgt ctattgacac aacgggaaac aggcttaacg
      >.......................McrA.......................>
       s  g  g   a  d  t  t   d  n  c   v  a  l   c  p  n  c

 851  catagagaat tgcactatag taaaaatgca aaagaactaa tcgagatgct
      gtatctctta acgtgatatc atttttacgt tttcttgatt agctctacga
      >.......................McrA.......................>
        h  r  e   l  h  y   s  k  n  a   k  e  l   i  e  m

                                            BamHI
                                           -+----
 901  ttacgttaat ataaaccgat tacagaaata atagggatcc gaattcgagc
      aatgcaatta tatttggcta atgtctttat tatccctagg cttaagctcg
      >............McrA.............>>
      l  y  v  n   i  n  r   l  q  k   -

 951  tccgtcgaca agcttgcggc cgcactcgag caccaccacc accaccactg
      aggcagctgt tcgaacgccg gcgtgagctc gtggtggtgg tggtggtgac

1001  agatccggct gctaacaaag cccgaaagga agctgagttg gctgctgcca
      tctaggccga cgattgtttc gggctttcct tcgactcaac cgacgacggt

1051  ccgctgagca ataactagca taacccccttg gggcctctaa acgggtcttg
      ggcgactcgt tattgatcgt attggggaac cccggagatt tgcccagaac
              1518

1101  aggggttttt tgctgaaagg aggaactata tccggattgg cgaatgggac
      tccccaaaaa acgactttcc tccttgatat aggcctaacc gcttaccctg

1151  gcgccctgta gcggcgcatt aagcgcggcg ggtgtggtgg ttacgcgcag
      cgcgggacat cgccgcgtaa ttcgcgccgc ccacaccacc aatgcgcgtc

1201  cgtgaccgct acacttgcca gcgccctagc gcccgctcct ttcgctttct
      gcactggcga tgtgaacggt cgcgggatcg cgggcgagga aagcgaaaga

1251  tcccttcctt tctcgccacg ttcgccggct ttccccgtca agctctaaat
      agggaaggaa agagcggtgc aagcggccga aaggggcagt tcgagattta
```

```
1301  cgggggctcc ctttagggtt ccgatttagt gctttacggc acctcgaccc
      gcccccgagg gaaatcccaa ggctaaatca cgaaatgccg tggagctggg

1351  caaaaaactt gattagggtg atggttcacg tagtgggcca tcgccctgat
      gttttttgaa ctaatcccac taccaagtgc atcacccggt agcgggacta

1401  agacggtttt tcgccctttg acgttggagt ccacgttctt taatagtgga
      tctgccaaaa agcgggaaac tgcaacctca ggtgcaagaa attatcacct

1451  ctcttgttcc aaactggaac aacactcaac cctatctcgg tctattcttt
      gagaacaagg tttgaccttg ttgtgagttg ggatagagcc agataagaaa

1501  tgatttataa gggattttgc cgatttcggc ctattggtta aaaaatgagc
      actaaatatt ccctaaaacg gctaaagccg gataaccaat tttttactcg

1551  tgatttaaca aaaatttaac gcgaatttta acaaaatatt aacgtttaca
      actaaattgt ttttaaattg cgcttaaaat tgttttataa ttgcaaatgt

1601  atttcaggtg gcacttttcg gggaaatgtg cgcggaaccc ctatttgttt
      taaagtccac cgtgaaaagc ccctttacac gcgccttggg gataaacaaa

1651  atttttctaa atacattcaa atatgtatcc gctcatgaat taattcttag
      taaaaagatt tatgtaagtt tatacatagg cgagtactta attaagaatc
                                                KanR <<.<
                                                     -

1701  aaaaactcat cgagcatcaa atgaaactgc aatttattca tatcaggatt
      ttttgagta gctcgtagtt tactttgacg ttaaataagt atagtcctaa
      <.......................KanR.......................<
       f  f  e   d  l  m  l   h  f  q   l  k  n   m  d  p  n

1751  atcaatacca tatttttgaa aaagccgttt ctgtaatgaa ggagaaaact
      tagttatggt ataaaaactt tttcggcaaa gacattactt cctcttttga
      <.......................KanR.......................<
        d  i  g   y  k  q   f  l  r  k   q  l  s   p  s  f

1801  caccgaggca gttccatagg atggcaagat cctggtatcg gtctgcgatt
      gtggctccgt caaggtatcc taccgttcta ggaccatagc cagacgctaa
      <.......................KanR.......................<
      e  g  l  c   n  w  l   i  a  l   d  q  y  r   d  a  i

1851  ccgactcgtc caacatcaat acaacctatt aatttcccct cgtcaaaaat
      ggctgagcag gttgtagtta tgttggataa ttaaagggga gcagtttta
      <.......................KanR.......................<
       g  v  r   g  v  d  i   c  g  i   l  k  g   e  d  f  i

1901  aaggttatca agtgagaaat caccatgagt gacgactgaa tccggtgaga
      ttccaatagt tcactcttta gtggtactca ctgctgactt aggccactct
      <.......................KanR.......................<
        l  n  d   l  s  f   d  g  h  t   v  v  s   d  p  s

1951  atggcaaaag tttatgcatt tctttccaga cttgttcaac aggccagcca
      taccgttttc aaatacgtaa agaaaggtct gaacaagttg tccggtcggt
      <.......................KanR.......................<
      f  p  l  l   k  h  m   e  k  w   v  q  e  v   p  w  g
```

```
2001  ttacgctcgt catcaaaatc actcgcatca accaaaccgt tattcattcg
      aatgcgagca gtagttttag tgagcgtagt tggtttggca ataagtaagc
      <.......................KanR.......................<
       n  r  e   d  d  f  d   s  a  d   v  l  g   n  n  m  r


2051  tgattgcgcc tgagcgagac gaaatacgcg atcgctgtta aaaggacaat
      actaacgcgg actcgctctg ctttatgcgc tagcgacaat tttcctgtta
      <.......................KanR.......................<
        s  q  a   q  a  l   r  f  v  r   d  s  n   f  p  c

2101  tacaaacagg aatcgaatgc aaccggcgca ggaacactgc cagcgcatca
      atgtttgtcc ttagcttacg ttggccgcgt ccttgtgacg gtcgcgtagt
      <.......................KanR.......................<
      n  c  v  p   i  s  h   l  r  r   l  f  v  a   l  a  d

2151  acaatatttt cacctgaatc aggatattct tctaatacct ggaatgctgt
      tgttataaaa gtggacttag tcctataaga agattatgga ccttacgaca
      <.......................KanR.......................<
       v  i  n   e  g  s  d   p  y  e   e  l  v   q  f  a  t

2201  tttcccgggg atcgcagtgg tgagtaacca tgcatcatca ggagtacgga
      aaagggcccc tagcgtcacc actcattggt acgtagtagt cctcatgcct
      <.......................KanR.......................<
        k  g  p   i  a  t   t  l  l  w   a  d  d   p  t  r

2251  taaaatgctt gatggtcgga agaggcataa attccgtcag ccagtttagt
      attttacgaa ctaccagcct tctccgtatt taaggcagtc ggtcaaatca
      <.......................KanR.......................<
      i  f  h  k   i  t  p   l  p  m   f  e  t  l   w  n  l

2301  ctgaccatct catctgtaac atcattggca acgctacctt tgccatgttt
      gactggtaga gtagacattg tagtaaccgt tgcgatggaa acggtacaaa
      <.......................KanR.......................<
       r  v  m   e  d  t  v   d  n  a   v  s  g   k  g  h  k

2351  cagaaacaac tctggcgcat cgggcttccc atacaatcga tagattgtcg
      gtctttgttg agaccgcgta gcccgaaggg tatgttagct atctaacagc
      <.......................KanR.......................<
        l  f  l   e  p  a   d  p  k  g   y  l  r   y  i  t

2401  cacctgattg cccgacatta tcgcgagccc atttataccc atataaatca
      gtggactaac gggctgtaat agcgctcggg taaatatggg tatatttagt
      <.......................KanR.......................<
      a  g  s  q   g  v  n   d  r  a   w  k  y  g   y  l  d

2451  gcatccatgt tggaatttaa tcgcggccta gagcaagacg tttcccgttg
      cgtaggtaca accttaaatt agcgccggat ctcgttctgc aaagggcaac
      <.......................KanR.......................<
       a  d  m   n  s  n  l   r  p  r   s  c  s   t  e  r  q


2501  aatatggctc ataacacccc ttgtattact gtttatgtaa gcagacagtt
      ttataccgag tattgtgggg aacataatga caaatacatt cgtctgtcaa
      <...KanR...<<
        i  h  s   m
```

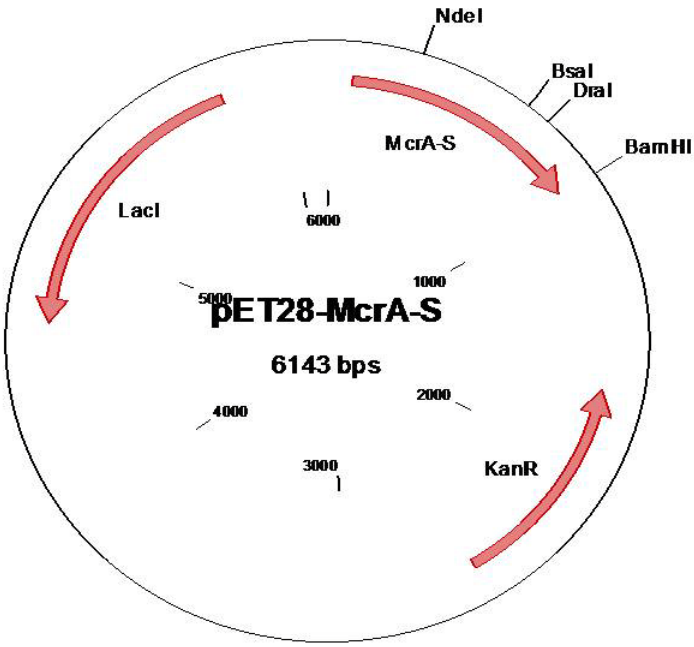
```
2551  ttattgttca tgaccaaaat cccttaacgt gagttttcgt tccactgagc
      aataacaagt actggtttta gggaattgca ctcaaaagca aggtgactcg

2601  gtcagacccc gtagaaaaga tcaaaggatc ttcttgagat ccttttttc
      cagtctgggg catcttttct agtttcctag aagaactcta ggaaaaaaag


2651  tgcgcgtaat ctgctgcttg caaacaaaaa aaccaccgct accagcggtg
      acgcgcatta gacgacgaac gtttgttttt ttggtggcga tggtcgccac

2701  gtttgtttgc cggatcaaga gctaccaact ctttttccga aggtaactgg
      caaacaaacg gcctagttct cgatggttga gaaaaaggct tccattgacc

2751  cttcagcaga gcgcagatac caaatactgt ccttctagtg tagccgtagt
      gaagtcgtct cgcgtctatg gtttatgaca ggaagatcac atcggcatca

2801  taggccacca cttcaagaac tctgtagcac cgcctacata cctcgctctg
      atccggtggt gaagttcttg agacatcgtg gcggatgtat ggagcgagac

2851  ctaatcctgt taccagtggc tgctgccagt ggcgataagt cgtgtcttac
      gattaggaca atggtcaccg acgacggtca ccgctattca gcacagaatg

2901  cgggttggac tcaagacgat agttaccgga taaggcgcag cggtcgggct
      gcccaacctg agttctgcta tcaatggcct attccgcgtc gccagcccga

2951  gaacgggggg ttcgtgcaca cagcccagct tggagcgaac gacctacacc
      cttgcccccc aagcacgtgt gtcgggtcga acctcgcttg ctggatgtgg

3001  gaactgagat acctacagcg tgagctatga gaaagcgcca cgcttcccga
      cttgactcta tggatgtcgc actcgatact ctttcgcggt gcgaagggct

3051  agggagaaag gcggacaggt atccggtaag cggcagggtc ggaacaggag
      tccctctttc cgcctgtcca taggccattc gccgtcccag ccttgtcctc

3101  agcgcacgag ggagcttcca gggggaaacg cctggtatct ttatagtcct
      tcgcgtgctc cctcgaaggt ccccctttgc ggaccataga aatatcagga

3151  gtcgggtttc gccacctctg acttgagcgt cgatttttgt gatgctcgtc
      cagcccaaag cggtggagac tgaactcgca gctaaaaaca ctacgagcag

3201  agggggggcgg agcctatgga aaaacgccag caacgcggcc tttttacggt
      tccccccgcc tcggatacct ttttgcggtc gttgcgccgg aaaaatgcca

3251  tcctggcctt ttgctggcct tttgctcaca tgttctttcc tgcgttatcc
      aggaccggaa aacgaccgga aaacgagtgt acaagaaagg acgcaatagg

3301  cctgattctg tggataaccg tattaccgcc tttgagtgag ctgataccgc
      ggactaagac acctattggc ataatggcgg aaactcactc gactatggcg

3351  tcgccgcagc cgaacgaccg agcgcagcga gtcagtgagc gaggaagcgg
      agcggcgtcg gcttgctggc tcgcgtcgct cagtcactcg ctccttcgcc

3401  aagagcgcct gatgcggtat tttctcctta cgcatctgtg cggtatttca
      ttctcgcgga ctacgccata aaagaggaat gcgtagacac gccataaagt
```

```
3451  caccgcatat atggtgcact ctcagtacaa tctgctctga tgccgcatag
      gtggcgtata taccacgtga gagtcatgtt agacgagact acggcgtatc

3501  ttaagccagt atacactccg ctatcgctac gtgactgggt catggctgcg
      aattcggtca tatgtgaggc gatagcgatg cactgaccca gtaccgacgc

3551  ccccgacacc cgccaacacc cgctgacgcg ccctgacggg cttgtctgct
      ggggctgtgg gcggttgtgg gcgactgcgc gggactgccc gaacagacga

3601  cccggcatcc gcttacagac aagctgtgac cgtctccggg agctgcatgt
      gggccgtagg cgaatgtctg ttcgacactg gcagaggccc tcgacgtaca

3651  gtcagaggtt ttcaccgtca tcaccgaaac gcgcgaggca gctgcggtaa
      cagtctccaa aagtggcagt agtggctttg cgcgctccgt cgacgccatt

3701  agctcatcag cgtggtcgtg aagcgattca cagatgtctg cctgttcatc
      tcgagtagtc gcaccagcac ttcgctaagt gtctacagac ggacaagtag

3751  cgcgtccagc tcgttgagtt tctccagaag cgttaatgtc tggcttctga
      gcgcaggtcg agcaactcaa agaggtcttc gcaattacag accgaagact

3801  taaagcgggc catgttaagg gcggtttttt cctgtttggt cactgatgcc
      atttcgcccg gtacaattcc cgccaaaaaa ggacaaacca gtgactacgg

3851  tccgtgtaag ggggatttct gttcatgggg gtaatgatac cgatgaaacg
      aggcacattc cccctaaaga caagtacccc cattactatg gctactttgc

3901  agagaggatg ctcacgatac gggttactga tgatgaacat gcccggttac
      tctctcctac gagtgctatg cccaatgact actacttgta cgggccaatg

3951  tggaacgttg tgagggtaaa caactggcgg tatggatgcg gcgggaccag
      accttgcaac actcccattt gttgaccgcc atacctacgc cgccctggtc

4001  agaaaaatca ctcagggtca atgccagcgc ttcgttaata cagatgtagg
      tctttttagt gagtcccagt tacggtcgcg aagcaattat gtctacatcc

4051  tgttccacag ggtagccagc agcatcctgc gatgcagatc cggaacataa
      acaaggtgtc ccatcggtcg tcgtaggacg ctacgtctag gccttgtatt

4101  tggtgcaggg cgctgacttc cgcgtttcca gactttacga aacacggaaa
      accacgtccc gcgactgaag gcgcaaaggt ctgaaatgct ttgtgccttt

4151  ccgaagacca ttcatgttgt tgctcaggtc gcagacgttt tgcagcagca
      ggcttctggt aagtacaaca acgagtccag cgtctgcaaa acgtcgtcgt

4201  gtcgcttcac gttcgctcgc gtatcggtga ttcattctgc taaccagtaa
      cagcgaagtg caagcgagcg catagccact aagtaagacg attggtcatt

4251  ggcaaccccg ccagcctagc cgggtcctca acgacaggag cacgatcatg
      ccgttggggc ggtcggatcg gcccaggagt tgctgtcctc gtgctagtac

4301  cgcacccgtg gggccgccat gccggcgata atggcctgct tctcgccgaa
      gcgtgggcac cccggcggta cggccgctat taccggacga agagcggctt

4351  acgtttggtg gcgggaccag tgacgaaggc ttgagcgagg gcgtgcaaga
      tgcaaaccac cgccctggtc actgcttccg aactcgctcc cgcacgttct
```

```
4401  ttccgaatac cgcaagcgac aggccgatca tcgtcgcgct ccagcgaaag
      aaggcttatg gcgttcgctg tccggctagt agcagcgcga ggtcgctttc

4451  cggtcctcgc cgaaaatgac ccagagcgct gccggcacct gtcctacgag
      gccaggagcg gcttttactg ggtctcgcga cggccgtgga caggatgctc

4501  ttgcatgata aagaagacag tcataagtgc ggcgacgata gtcatgcccc
      aacgtactat ttcttctgtc agtattcacg ccgctgctat cagtacgggg

4551  gcgcccaccg gaaggagctg actgggttga aggctctcaa gggcatcggt
      cgcgggtggc cttcctcgac tgacccaact tccgagagtt cccgtagcca

4601  cgagatcccg gtgcctaatg agtgagctaa cttacattaa ttgcgttgcg
      gctctagggc cacggattac tcactcgatt gaatgtaatt aacgcaacgc

4651  ctcactgccc gctttccagt cgggaaacct gtcgtgccag ctgcattaat
      gagtgacggg cgaaaggtca gcccttgga cagcacggtc gacgtaatta
       <<.....................LacI.....................<
         -  q  g   s  e  l   r  s  v   q  r  a  l   q  m  l

4701  gaatcggcca acgcgcgggg agaggcggtt tgcgtattgg gcgccagggt
      cttagccggt tgcgcgcccc tctccgccaa acgcataacc cgcggtccca
      <.....................LacI.....................<
       s  d  a   l  a  r  p   s  a  t   q  t  n   p  a  l  t

4751  ggtttttctt ttcaccagtg agacgggcaa cagctgattg cccttcaccg
      ccaaaaagaa aagtggtcac tctgcccgtt gtcgactaac gggaagtggc
      <.....................LacI.....................<
        t  k  r   k  v  l   s  v  p  l   l  q  n   g  k  v

4801  cctggccctg agagagttgc agcaagcggt ccacgctggt ttgccccagc
      ggaccgggac tctctcaacg tcgttcgcca ggtgcgacca aacggggtcg
      <.....................LacI.....................<
      a  q  g  q   s  l  q   l  l  r   d  v  s  t   q  g  l

4851  aggcgaaaat cctgtttgat ggtggttaac ggcgggatat aacatgagct
      tccgctttta ggacaaacta ccaccaattg ccgccctata ttgtactcga
      <.....................LacI.....................<
       l  r  f   d  q  k  i   t  t  l   p  p  i   y  c  s  s

4901  gtcttcggta tcgtcgtatc ccactaccga gatatccgca ccaacgcgca
      cagaagccat agcagcatag ggtgatggct ctataggcgt ggttgcgcgt
      <.....................LacI.....................<
        d  e  t   d  d  y   g  v  v  s   i  d  a   g  v  r

4951  gcccggactc ggtaatggcg cgcattgcgc ccagcgccat ctgatcgttg
      cgggcctgag ccattaccgc gcgtaacgcg ggtcgcggta gactagcaac
      <.....................LacI.....................<
      l  g  s  e   t  i  a   r  m  a   g  l  a  m   q  d  n

5001  gcaaccagca tcgcagtggg aacgatgccc tcattcagca tttgcatggt
      cgttggtcgt agcgtcaccc ttgctacggg agtaagtcgt aaacgtacca
      <.....................LacI.....................<
       a  v  l   m  a  t  p   v  i  g   e  n  l   m  q  m  t
```

```
5051  ttgttgaaaa ccggacatgg cactccagtc gccttcccgt tccgctatcg
      aacaactttt ggcctgtacc gtgaggtcag cggaagggca aggcgatagc
      <.....................LacI.....................<
        q  q  f   g  s  m   a  s  w  d   g  e  r   e  a  i

5101  gctgaatttg attgcgagtg agatatttat gccagccagc cagacgcaga
      cgacttaaac taacgctcac tctataaata cggtcggtcg gtctgcgtct
      <.....................LacI.....................<
      p  q  i  q   n  r  t   l  y  k   h  w  g  a   l  r  l


5151  cgcgccgaga cagaacttaa tgggcccgct aacagcgcga tttgctggtg
      gcgcggctct gtcttgaatt acccgggcga ttgtcgcgct aaacgaccac
      <.....................LacI.....................<
       r  a  s   v  s  s  l   p  g  a   l  l  a   i  q  q  h

5201  acccaatgcg accagatgct ccacgcccag tcgcgtaccg tcttcatggg
      tgggttacgc tggtctacga ggtgcgggtc agcgcatggc agaagtaccc
      <.....................LacI.....................<
        g  l  a   v  l  h   e  v  g  l   r  t  g   d  e  h

5251  agaaaataat actgttgatg ggtgtctggt cagagacatc aagaaataac
      tcttttatta tgacaactac ccacagacca gtctctgtag ttctttattg
      <.....................LacI.....................<
      s  f  i  i   s  n  i   p  t  q   d  s  v  d   l  f  l

5301  gccggaacat tagtgcaggc agcttccaca gcaatggcat cctggtcatc
      cggccttgta atcacgtccg tcgaaggtgt cgttaccgta ggaccagtag
      <.....................LacI.....................<
       a  p  v   n  t  c  a   a  e  v   a  i  a   d  q  d  d

5351  cagcggatag ttaatgatca gcccactgac gcgttgcgcg agaagattgt
      gtcgcctatc aattactagt cgggtgactg cgcaacgcgc tcttctaaca
      <.....................LacI.....................<
        l  p  y   n  i  i   l  g  s  v   r  q  a   l  l  n

5401  gcaccgccgc tttacaggct tcgacgccgc ttcgttctac catcgacacc
      cgtggcggcg aaatgtccga agctgcggcg aagcaagatg gtagctgtgg
      <.....................LacI.....................<
      h  v  a  a   k  c  a   e  v  g   s  r  e  v   m  s  v

5451  accacgctgg cacccagttg atcggcgcga gatttaatcg ccgcgacaat
      tggtgcgacc gtgggtcaac tagccgcgct ctaaattagc ggcgctgtta
      <.....................LacI.....................<
       v  v  s   a  g  l  q   d  a  r   s  k  i   a  a  v  i

5501  ttgcgacggc gcgtgcaggg ccagactgga ggtggcaacg ccaatcagca
      aacgctgccg cgcacgtccc ggtctgacct ccaccgttgc ggttagtcgt
      <.....................LacI.....................<
        q  s  p   a  h  l   a  l  s  s   t  a  v   g  i  l

5551  acgactgttt gcccgccagt tgttgtgcca cgcggttggg aatgtaattc
      tgctgacaaa cgggcggtca acaacacggt gcgccaaccc ttacattaag
      <.....................LacI.....................<
      l  s  q  k   g  a  l   q  q  a   v  r  n  p   i  y  n
```

```
5601  agctccgcca tcgccgcttc cacttttcc cgcgttttcg cagaaacgtg
      tcgaggcggt agcggcgaag gtgaaaaagg gcgcaaaagc gtctttgcac
      <.......................LacI.......................<
       l  e  a   m  a  a  e   v  k  e   r  t  k   a  s  v  h

5651  gctggcctgg ttcaccacgc gggaaacggt ctgataagag acaccggcat
      cgaccggacc aagtggtgcg ccctttgcca gactattctc tgtggccgta
      <.......................LacI.......................<
        s  a  q   n  v  v   r  s  v  t   q  y  s   v  g  a

5701  actctgcgac atcgtataac gttactggtt tcacattcac caccctgaat
      tgagacgctg tagcatattg caatgaccaa agtgtaagtg gtgggactta
      <.....................LacI.....................<<
      y  e  a  v   d  y  l   t  v  p   k  v  n  v   v

5751  tgactctctt ccgggcgcta tcatgccata ccgcgaaagg ttttgcgcca
      actgagagaa ggcccgcgat agtacggtat ggcgctttcc aaaacgcggt

5801  ttcgatggtg tccgggatct cgacgctctc ccttatgcga ctcctgcatt
      aagctaccac aggccctaga gctgcgagag ggaatacgct gaggacgtaa

5851  aggaagcagc ccagtagtag gttgaggccg ttgagcaccg ccgccgcaag
      tccttcgtcg ggtcatcatc caactccggc aactcgtggc ggcggcgttc

5901  gaatggtgca tgcaaggaga tggcgcccaa cagtcccccg gccacggggc
      cttaccacgt acgttcctct accgcgggtt gtcagggggc cggtgccccg

5951  ctgccaccat acccacgccg aaacaagcgc tcatgagccc gaagtggcga
      gacggtggta tgggtgcggc tttgttcgcg agtactcggg cttcaccgct

6001  gcccgatctt ccccatcggt gatgtcggcg atataggcgc cagcaaccgc
      cgggctagaa ggggtagcca ctacagccgc tatatccgcg gtcgttggcg

6051  acctgtggcg ccggtgatgc cggccacgat gcgtccggcg tagaggatcg
      tggacaccgc ggccactacg gccggtgcta cgcaggccgc atctcctagc

6101  agatctcgat
      tctagagcta
```




```
  1  taatacgact cactataggg gaattgtgag cggataacaa ttcccctcta
     attatgctga gtgatatccc cttaacactc gcctattgtt aaggggagat
              1517

 51  gaaataattt tgtttaactt taagaaggag atataccatg catgtttttg
     ctttattaaa acaaattgaa attcttcctc tatatggtac gtacaaaaac
                                            >>..McrA-S...>
                                              m   h  v  f

101  ataataatgg aattgaactg aaagctgagt gttcgatagg tgaagaggat
     tattattacc ttaacttgac tttcgactca caagctatcc acttctccta
     >.....................McrA-S.....................>
      d  n  n   g  i  e  l   k  a  e   c  s  i   g  e  e  d

151  ggtgtttatg gtctaatcct tgagtcgtgg gggccgggtg acagaaacaa
     ccacaaatac cagattagga actcagcacc cccggcccac tgtctttgtt
     >.....................McrA-S.....................>
       g  v  y   g  l  i   l  e  s  w   g  p  g   d  r  n

201  agattacaat atcgctcttg attatatcat tgaacggttg gttgattctg
     tctaatgtta tagcgagaac taatatagta acttgccaac caactaagac
     >.....................McrA-S.....................>
     k  d  y  n   i  a  l   d  y  i   i  e  r  l   v  d  s

                                                      NdeI
                                                     --+---
251  gtgtatccca agtcgtagta tatctggcgt catcatcagt cagaaaacat
     cacatagggt tcagcatcat atagaccgca gtagtagtca gtcttttgta
     >.....................McrA-S.....................>
      g  v  s   q  v  v  v   y  l  a   s  s  s   v  r  k  h
```

```
301  atgcattctt tggatgaaag aaaaatccat cctggtgaat attttacttt
     tacgtaagaa acctactttc tttttaggta ggaccactta taaaatgaaa
     >.....................McrA-S.....................>
       m  h  s   l  d  e   r  k  i  h   p  g  e   y  f  t

351  gattggtaat agccccccgcg atatacgctt gaagatgtgt ggttatcagg
     ctaaccatta tcgggggcgc tatatgcgaa cttctacaca ccaatagtcc
     >.....................McrA-S.....................>
     l  i  g  n   s  p  r   d  i  r   l  k  m  c   g  y  q

401  cttattttag tcgtacgggg agaaaggaaa ttccttccgg caatagaacg
     gaataaaatc agcatgcccc tctttccttt aaggaaggcc gttatcttgc
     >.....................McrA-S.....................>
      a  y  f   s  r  t  g   r  k  e   i  p  s   g  n  r  t

451  aaacgaatat tgataaatgt tccaggtatt tatagtgaca gtttttgggc
     tttgcttata actatttaca aggtccataa atatcactgt caaaaacccg
     >.....................McrA-S.....................>
       k  r  i   l  i  n   v  p  g  i   y  s  d   s  f  w

501  gtctataata cgtggagaac tatcagagct ttcacagcct acagatgatg
     cagatattat gcacctcttg atagtctcga aagtgtcgga tgtctactac
     >.....................McrA-S.....................>
     a  s  i  i   r  g  e   l  s  e   l  s  q  p   t  d  d

551  aatcgcttct gaatatgagg gttagtaaat taattaagaa aacgttgagt
     ttagcgaaga cttatactcc caatcattta attaattctt ttgcaactca
     >.....................McrA-S.....................>
      e  s  l   l  n  m  r   v  s  k   l  i  k   k  t  l  s

601  caacccgagg gctccaggaa accagttgag gtagaaagac tacaaaaagt
     gttgggctcc cgaggtcctt tggtcaactc catctttctg atgtttttca
     >.....................McrA-S.....................>
       q  p  e   g  s  r   k  p  v  e   v  e  r   l  q  k

              BsaI
            -------
651  ttatgtccga gacccgatgg taaaagcttg gattttacag caaagtaaag
     aatacaggct ctgggctacc attttcgaac ctaaaatgtc gtttcatttc
     >.....................McrA-S.....................>
     v  y  v  r   d  p  m   v  k  a   w  i  l  q   q  s  k

                                                DraI
                                              ----+--
701  gtatatgtga aaactgtggt aaaaatgctc cgttttattt aaatgatgga
     catatacact tttgacacca tttttacgag gcaaaataaa tttactacct
     >.....................McrA-S.....................>
      g  i  c   e  n  c  g   k  n  a   p  f  y   l  n  d  g

751  aacccatatt tggaagtaca tcatgtaatt cccctgtctt caggtggtgc
     ttgggtataa accttcatgt agtacattaa ggggacagaa gtccaccacg
     >.....................McrA-S.....................>
       n  p  y   l  e  v   h  h  v  i   p  l  s   s  g  g
```

```
 801  tgatacaaca gataactgtg ttgccctttg tccgaattgc catagagaat
      actatgttgt ctattgacac aacgggaaac aggcttaacg gtatctctta
      >.....................McrA-S.....................>
      a  d  t  t   d  n  c   v  a  l   c  p  n  c   h  r  e

 851  tgcactatag taaaaatgca aaagaactaa tcgagatgct ttacgttaat
      acgtgatatc atttttacgt tttcttgatt agctctacga aatgcaatta
      >.....................McrA-S.....................>
       l  h  y   s  k  n  a   k  e  l   i  e  m   l  y  v  n

 901  ataaaccgat tacagaaaag cgcttggagc cacccgcagt tcgaaaaata
      tatttggcta atgtcttttc gcgaacctcg gtgggcgtca agctttttat
                            'D27
      >.....................McrA-S.....................>
        i  n  r   l  q  k   s  a  w  s   h  p  q   f  e  k

             BamHI
            -+-----
 951  ataataggga tccgaattcg agctccgtcg acaagcttgc ggccgcactc
      tattatccct aggcttaagc tcgaggcagc tgttcgaacg ccggcgtgag
      > McrA-S
      -

1001  gagcaccacc accaccacca ctgagatccg gctgctaaca aagcccgaaa
      ctcgtggtgg tggtggtggt gactctaggc cgacgattgt ttcgggcttt

1051  ggaagctgag ttggctgctg ccaccgctga gcaataacta gcataacccc
      ccttcgactc aaccgacgac ggtggcgact cgttattgat cgtattgggg
                                      1518

1101  ttggggcctc taaacgggtc ttgaggggtt ttttgctgaa aggaggaact
      aaccccggag atttgcccag aactccccaa aaaacgactt tcctccttga

1151  atatccggat tggcgaatgg gacgcgccct gtagcggcgc attaagcgcg
      tataggccta accgcttacc ctgcgcggga catcgccgcg taattcgcgc

1201  gcgggtgtgg tggttacgcg cagcgtgacc gctacacttg ccagcgccct
      cgcccacacc accaatgcgc gtcgcactgg cgatgtgaac ggtcgcggga

1251  agcgcccgct cctttcgctt tcttcccttc ctttctcgcc acgttcgccg
      tcgcgggcga ggaaagcgaa agaagggaag gaaagagcgg tgcaagcggc

1301  gctttccccg tcaagctcta aatcgggggc tccctttagg gttccgattt
      cgaaaggggc agttcgagat ttagcccccg agggaaatcc caaggctaaa

1351  agtgctttac ggcacctcga ccccaaaaaa cttgattagg gtgatggttc
      tcacgaaatg ccgtggagct ggggtttttt gaactaatcc cactaccaag

1401  acgtagtggg ccatcgccct gatagacggt ttttcgccct ttgacgttgg
      tgcatcaccc ggtagcggga ctatctgcca aaaagcggga aactgcaacc

1451  agtccacgtt ctttaatagt ggactcttgt tccaaactgg aacaacactc
      tcaggtgcaa gaaattatca cctgagaaca aggtttgacc ttgttgtgag

1501  aaccctatct cggtctattc ttttgattta taagggattt tgccgatttc
      ttgggataga gccagataag aaaactaaat attccctaaa acggctaaag
```

```
1551  ggcctattgg ttaaaaaatg agctgattta acaaaaattt aacgcgaatt
      ccggataacc aatttttac tcgactaaat tgtttttaaa ttgcgcttaa

1601  ttaacaaaat attaacgttt acaatttcag gtggcacttt tcggggaaat
      aattgtttta taattgcaaa tgttaaagtc caccgtgaaa agccccttta

1651  gtgcgcggaa cccctatttg tttatttttc taaatacatt caaatatgta
      cacgcgcctt ggggataaac aaataaaaag atttatgtaa gtttatacat

1701  tccgctcatg aattaattct tagaaaaact catcgagcat caaatgaaac
      aggcgagtac ttaattaaga atctttttga gtagctcgta gtttactttg
                        <<.............KanR.............<
                           -  f  f   e  d  l  m   l  h  f

1751  tgcaatttat tcatatcagg attatcaata ccatattttt gaaaaagccg
      acgttaaata agtatagtcc taatagttat ggtataaaaa ctttttcggc
      <.......................KanR.......................<
       q  l  k   n  m  d  p   n  d  i   g  y  k   q  f  l  r

1801  tttctgtaat gaaggagaaa actcaccgag gcagttccat aggatggcaa
      aaagacatta cttcctcttt tgagtggctc cgtcaaggta tcctaccgtt
      <.......................KanR.......................<
        k  q  l   s  p  s   f  e  g  l   c  n  w   l  i  a

1851  gatcctggta tcggtctgcg attccgactc gtccaacatc aatacaacct
      ctaggaccat agccagacgc taaggctgag caggttgtag ttatgttgga
      <.......................KanR.......................<
      l  d  q  y   r  d  a   i  g  v   r  g  v  d   i  c  g

1901  attaatttcc cctcgtcaaa aataaggtta tcaagtgaga aatcaccatg
      taattaaagg ggagcagttt ttattccaat agttcactct ttagtggtac
      <.......................KanR.......................<
       i  l  k   g  e  d  f   i  l  n   d  l  s   f  d  g  h

1951  agtgacgact gaatccggtg agaatggcaa aagtttatgc atttctttcc
      tcactgctga cttaggccac tcttaccgtt ttcaaatacg taaagaaagg
      <.......................KanR.......................<
        t  v  v   s  d  p   s  f  p  l   l  k  h   m  e  k

2001  agacttgttc aacaggccag ccattacgct cgtcatcaaa atcactcgca
      tctgaacaag ttgtccggtc ggtaatgcga gcagtagttt tagtgagcgt
      <.......................KanR.......................<
      w  v  q  e   v  p  w   g  n  r   e  d  d  f   d  s  a

2051  tcaaccaaac cgttattcat tcgtgattgc gcctgagcga gacgaaatac
      agttggtttg gcaataagta agcactaacg cggactcgct ctgctttatg
      <.......................KanR.......................<
       d  v  l   g  n  n  m   r  s  q   a  q  a   l  r  f  v

2101  gcgatcgctg ttaaaaggac aattacaaac aggaatcgaa tgcaaccggc
      cgctagcgac aattttcctg ttaatgtttg tccttagctt acgttggccg
      <.......................KanR.......................<
        r  d  s   n  f  p   c  n  c  v   p  i  s   h  l  r
```

```
2151  gcaggaacac tgccagcgca tcaacaatat tttcacctga atcaggatat
      cgtccttgtg acggtcgcgt agttgttata aaagtggact tagtcctata
      <.......................KanR.......................<
      r  l  f  v   a  l  a   d  v  i   n  e  g  s   d  p  y

2201  tcttctaata cctggaatgc tgttttcccg gggatcgcag tggtgagtaa
      agaagattat ggaccttacg acaaaagggc ccctagcgtc accactcatt
      <.......................KanR.......................<
       e  e  l   v  q  f  a   t  k  g   p  i  a   t  t  l  l

2251  ccatgcatca tcaggagtac ggataaaatg cttgatggtc ggaagaggca
      ggtacgtagt agtcctcatg cctattttac gaactaccag ccttctccgt
      <.......................KanR.......................<
        w  a  d   d  p  t   r  i  f  h   k  i  t   p  l  p

2301  taaattccgt cagccagttt agtctgacca tctcatctgt aacatcattg
      atttaaggca gtcggtcaaa tcagactggt agagtagaca ttgtagtaac
      <.......................KanR.......................<
      m  f  e  t   l  w  n   l  r  v   m  e  d  t   v  d  n

2351  gcaacgctac ctttgccatg tttcagaaac aactctggcg catcgggctt
      cgttgcgatg gaaacggtac aaagtctttg ttgagaccgc gtagcccgaa
      <.......................KanR.......................<
       a  v  s   g  k  g  h   k  l  f   l  e  p   a  d  p  k

2401  cccatacaat cgatagattg tcgcacctga ttgcccgaca ttatcgcgag
      gggtatgtta gctatctaac agcgtggact aacgggctgt aatagcgctc
      <.......................KanR.......................<
        g  y  l   r  y  i   t  a  g  s   q  g  v   n  d  r

2451  cccatttata cccatataaa tcagcatcca tgttggaatt taatcgcggc
      gggtaaatat gggtatattt agtcgtaggt acaaccttaa attagcgccg
      <.......................KanR.......................<
      a  w  k  y   g  y  l   d  a  d   m  n  s  n   l  r  p

2501  ctagagcaag acgtttcccg ttgaatatgg ctcataacac cccttgtatt
      gatctcgttc tgcaaagggc aacttatacc gagtattgtg gggaacataa
      <..............KanR..............<<
       r  s  c   s  t  e  r   q  i  h   s  m

2551  actgtttatg taagcagaca gttttattgt tcatgaccaa aatcccttaa
      tgacaaatac attcgtctgt caaaataaca agtactggtt ttagggaatt

2601  cgtgagtttt cgttccactg agcgtcagac cccgtagaaa agatcaaagg
      gcactcaaaa gcaaggtgac tcgcagtctg gggcatcttt tctagtttcc

2651  atcttcttga gatccttttt ttctgcgcgt aatctgctgc ttgcaaacaa
      tagaagaact ctaggaaaaa aagacgcgca ttagacgacg aacgtttgtt

2701  aaaaaccacc gctaccagcg gtggtttgtt tgccggatca agagctacca
      tttttggtgg cgatggtcgc caccaaacaa acggcctagt tctcgatggt

2751  actctttttc cgaaggtaac tggcttcagc agagcgcaga taccaaatac
      tgagaaaaag gcttccattg accgaagtcg tctcgcgtct atggtttatg

2801  tgtccttcta gtgtagccgt agttaggcca ccacttcaag aactctgtag
      acaggaagat cacatcggca tcaatccggt ggtgaagttc ttgagacatc
```

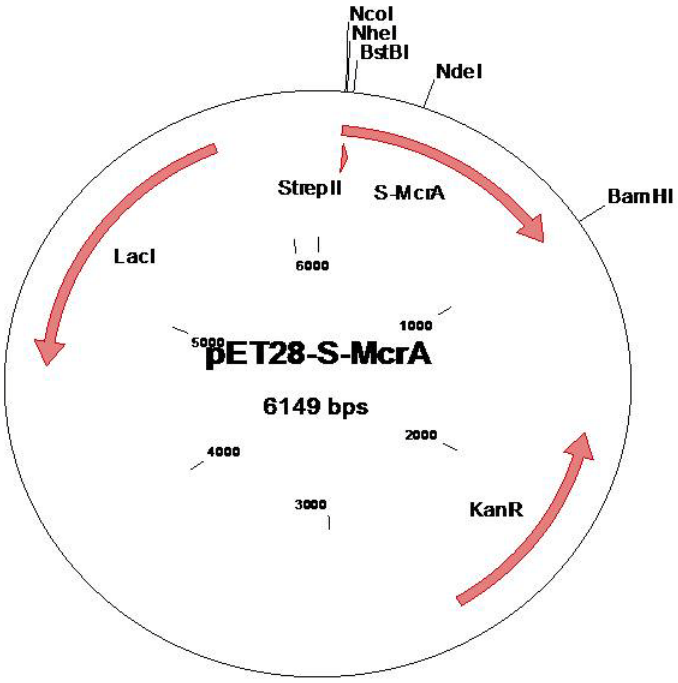
```
2851  caccgcctac atacctcgct ctgctaatcc tgttaccagt ggctgctgcc
      gtggcggatg tatggagcga gacgattagg acaatggtca ccgacgacgg

2901  agtggcgata agtcgtgtct taccgggttg gactcaagac gatagttacc
      tcaccgctat tcagcacaga atggcccaac ctgagttctg ctatcaatgg

2951  ggataaggcg cagcggtcgg gctgaacggg gggttcgtgc acacagccca
      cctattccgc gtcgccagcc cgacttgccc cccaagcacg tgtgtcgggt

3001  gcttggagcg aacgacctac accgaactga gatacctaca gcgtgagcta
      cgaacctcgc ttgctggatg tggcttgact ctatggatgt cgcactcgat

3051  tgagaaagcg ccacgcttcc cgaagggaga aaggcggaca ggtatccggt
      actctttcgc ggtgcgaagg gcttccctct ttccgcctgt ccataggcca

3101  aagcggcagg gtcggaacag gagagcgcac gagggagctt ccaggggaa
      ttcgccgtcc cagccttgtc ctctcgcgtg ctccctcgaa ggtccccctt

3151  acgcctggta tctttatagt cctgtcgggt ttcgccacct ctgacttgag
      tgcggaccat agaaatatca ggacagccca aagcggtgga gactgaactc

3201  cgtcgatttt tgtgatgctc gtcaggggg cggagcctat ggaaaaacgc
      gcagctaaaa acactacgag cagtcccccc gcctcggata ccttttgcg

3251  cagcaacgcg gccttttac ggttcctggc cttttgctgg ccttttgctc
      gtcgttgcgc cggaaaaatg ccaaggaccg gaaaacgacc ggaaaacgag

3301  acatgttctt tcctgcgtta tcccctgatt ctgtggataa ccgtattacc
      tgtacaagaa aggacgcaat aggggactaa gacacctatt ggcataatgg

3351  gcctttgagt gagctgatac cgctcgccgc agccgaacga ccgagcgcag
      cggaaactca ctcgactatg gcgagcggcg tcggcttgct ggctcgcgtc

3401  cgagtcagtg agcgaggaag cggaagagcg cctgatgcgg tattttctcc
      gctcagtcac tcgctccttc gccttctcgc ggactacgcc ataaaagagg

3451  ttacgcatct gtgcggtatt tcacaccgca tatatggtgc actctcagta
      aatgcgtaga cacgccataa agtgtggcgt atataccacg tgagagtcat

3501  caatctgctc tgatgccgca tagttaagcc agtatacact ccgctatcgc
      gttagacgag actacggcgt atcaattcgg tcatatgtga ggcgatagcg

3551  tacgtgactg ggtcatggct gcgccccgac acccgccaac acccgctgac
      atgcactgac ccagtaccga cgcggggctg tgggcggttg tgggcgactg

3601  gcgccctgac gggcttgtct gctcccggca tccgcttaca gacaagctgt
      cgcgggactg cccgaacaga cgagggccgt aggcgaatgt ctgttcgaca

3651  gaccgtctcc gggagctgca tgtgtcagag gttttcaccg tcatcaccga
      ctggcagagg ccctcgacgt acacagtctc caaaagtggc agtagtggct

3701  aacgcgcgag gcagctgcgg taaagctcat cagcgtggtc gtgaagcgat
      ttgcgcgctc cgtcgacgcc atttcgagta gtcgcaccag cacttcgcta

3751  tcacagatgt ctgcctgttc atccgcgtcc agctcgttga gtttctccag
      agtgtctaca gacggacaag taggcgcagg tcgagcaact caaagaggtc
```

```
3801  aagcgttaat gtctggcttc tgataaagcg ggccatgtta agggcggttt
      ttcgcaatta cagaccgaag actatttcgc ccggtacaat tcccgccaaa

3851  tttcctgttt ggtcactgat gcctccgtgt aaggggggatt tctgttcatg
      aaaggacaaa ccagtgacta cggaggcaca ttccccctaa agacaagtac

3901  ggggtaatga taccgatgaa acgagagagg atgctcacga tacgggttac
      ccccattact atggctactt tgctctctcc tacgagtgct atgcccaatg

3951  tgatgatgaa catgcccggt tactggaacg ttgtgagggt aaacaactgg
      actactactt gtacgggcca atgaccttgc aacactccca tttgttgacc

4001  cggtatggat gcggcgggac cagagaaaaa tcactcaggg tcaatgccag
      gccataccta cgccgccctg gtctcttttt agtgagtccc agttacggtc

4051  cgcttcgtta atacagatgt aggtgttcca cagggtagcc agcagcatcc
      gcgaagcaat tatgtctaca tccacaaggt gtcccatcgg tcgtcgtagg

4101  tgcgatgcag atccggaaca taatggtgca gggcgctgac ttccgcgttt
      acgctacgtc taggccttgt attaccacgt cccgcgactg aaggcgcaaa

4151  ccagacttta cgaaacacgg aaaccgaaga ccattcatgt tgttgctcag
      ggtctgaaat gctttgtgcc tttggcttct ggtaagtaca acaacgagtc

4201  gtcgcagacg ttttgcagca gcagtcgctt cacgttcgct cgcgtatcgg
      cagcgtctgc aaaacgtcgt cgtcagcgaa gtgcaagcga gcgcatagcc

4251  tgattcattc tgctaaccag taaggcaacc ccgccagcct agccgggtcc
      actaagtaag acgattggtc attccgttgg ggcggtcgga tcggcccagg

4301  tcaacgacag gagcacgatc atgcgcaccc gtggggccgc catgccggcg
      agttgctgtc ctcgtgctag tacgcgtggg caccccggcg gtacggccgc

4351  ataatggcct gcttctcgcc gaaacgtttg gtggcgggac cagtgacgaa
      tattaccgga cgaagagcgg ctttgcaaac caccgccctg gtcactgctt

4401  ggcttgagcg agggcgtgca agattccgaa taccgcaagc gacaggccga
      ccgaactcgc tcccgcacgt tctaaggctt atggcgttcg ctgtccggct

4451  tcatcgtcgc gctccagcga aagcggtcct cgccgaaaat gacccagagc
      agtagcagcg cgaggtcgct ttcgccagga gcggctttta ctgggtctcg

4501  gctgccggca cctgtcctac gagttgcatg ataaagaaga cagtcataag
      cgacggccgt ggacaggatg ctcaacgtac tatttcttct gtcagtattc

4551  tgcggcgacg atagtcatgc cccgcgccca ccggaaggag ctgactgggt
      acgccgctgc tatcagtacg gggcgcgggt ggccttcctc gactgaccca

4601  tgaaggctct caagggcatc ggtcgagatc ccggtgccta atgagtgagc
      acttccgaga gttcccgtag ccagctctag ggccacggat tactcactcg


4651  taacttacat taattgcgtt gcgctcactg cccgctttcc agtcgggaaa
      attgaatgta attaacgcaa cgcgagtgac gggcgaaagg tcagcccttt
                              <<.........LacI..........<
                                -  q   g  s  e   l  r  s
```

```
4701  cctgtcgtgc cagctgcatt aatgaatcgg ccaacgcgcg gggagaggcg
      ggacagcacg gtcgacgtaa ttacttagcc ggttgcgcgc ccctctccgc
      <.....................LacI.....................<
      v  q  r  a   l  q  m   l  s  d   a  l  a  r   p  s  a

4751  gtttgcgtat tgggcgccag ggtggttttt cttttcacca gtgagacggg
      caaacgcata acccgcggtc ccaccaaaaa gaaaagtggt cactctgccc
      <.....................LacI.....................<
       t  q  t   n  p  a  l   t  t  k   r  k  v   l  s  v  p

4801  caacagctga ttgcccttca ccgcctggcc ctgagagagt tgcagcaagc
      gttgtcgact aacgggaagt ggcggaccgg gactctctca acgtcgttcg
      <.....................LacI.....................<
        l  l  q   n  g  k   v  a  q  g   q  s  l   q  l  l

4851  ggtccacgct ggtttgcccc agcaggcgaa aatcctgttt gatggtggtt
      ccaggtgcga ccaaacgggg tcgtccgctt ttaggacaaa ctaccaccaa
      <.....................LacI.....................<
      r  d  v  s   t  q  g   l  l  r   f  d  q  k   i  t  t

4901  aacggcggga tataacatga gctgtcttcg gtatcgtcgt atcccactac
      ttgccgccct atattgtact cgacagaagc catagcagca tagggtgatg
      <.....................LacI.....................<
       l  p  p   i  y  c  s   s  d  e   t  d  d   y  g  v  v

4951  cgagatatcc gcaccaacgc gcagcccgga ctcggtaatg gcgcgcattg
      gctctatagg cgtggttgcg cgtcgggcct gagccattac cgcgcgtaac
      <.....................LacI.....................<
        s  i  d   a  g  v   r  l  g  s   e  t  i   a  r  m

5001  cgcccagcgc catctgatcg ttggcaacca gcatcgcagt gggaacgatg
      gcgggtcgcg gtagactagc aaccgttggt cgtagcgtca cccttgctac
      <.....................LacI.....................<
      a  g  l  a   m  q  d   n  a  v   l  m  a  t   p  v  i

5051  ccctcattca gcatttgcat ggtttgttga aaaccggaca tggcactcca
      gggagtaagt cgtaaacgta ccaaacaact tttggcctgt accgtgaggt
      <.....................LacI.....................<
       g  e  n   l  m  q  m   t  q  q   f  g  s   m  a  s  w

5101  gtcgccttcc cgttccgcta tcggctgaat ttgattgcga gtgagatatt
      cagcggaagg gcaaggcgat agccgactta aactaacgct cactctataa
      <.....................LacI.....................<
        d  g  e   r  e  a   i  p  q  i   q  n  r   t  l  y

5151  tatgccagcc agccagacgc agacgcgccg agacagaact taatgggccc
      atacggtcgg tcggtctgcg tctgcgcggc tctgtcttga attacccggg
      <.....................LacI.....................<
      k  h  w  g   a  l  r   l  r  a   s  v  s  s   l  p  g



5201  gctaacagcg cgatttgctg gtgacccaat gcgaccagat gctccacgcc
      cgattgtcgc gctaaacgac cactgggtta cgctggtcta cgaggtgcgg
      <.....................LacI.....................<
       a  l  l   a  i  q  q   h  g  l   a  v  l   h  e  v  g
```

```
5251  cagtcgcgta ccgtcttcat gggagaaaat aatactgttg atgggtgtct
      gtcagcgcat ggcagaagta ccctctttta ttatgacaac tacccacaga
      <.......................LacI.......................<
        l  r  t   g  d  e   h  s  f  i   i  s  n   i  p  t


5301  ggtcagagac atcaagaaat aacgccggaa cattagtgca ggcagcttcc
      ccagtctctg tagttcttta ttgcggcctt gtaatcacgt ccgtcgaagg
      <.......................LacI.......................<
      q  d  s  v   d  l  f   l  a  p   v  n  t  c   a  a  e

5351  acagcaatgg catcctggtc atccagcgga tagttaatga tcagcccact
      tgtcgttacc gtaggaccag taggtcgcct atcaattact agtcgggtga
      <.......................LacI.......................<
       v  a  i   a  d  q  d   d  l  p   y  n  i   i  l  g  s

5401  gacgcgttgc gcgagaagat tgtgcaccgc cgctttacag gcttcgacgc
      ctgcgcaacg cgctcttcta acacgtggcg gcgaaatgtc cgaagctgcg
      <.......................LacI.......................<
        v  r  q   a  l  l   n  h  v  a   a  k  c   a  e  v

5451  cgcttcgttc taccatcgac accaccacgc tggcacccag ttgatcggcg
      gcgaagcaag atggtagctg tggtggtgcg accgtgggtc aactagccgc
      <.......................LacI.......................<
      g  s  r  e   v  m  s   v  v  v   s  a  g  l   q  d  a

5501  cgagatttaa tcgccgcgac aatttgcgac ggcgcgtgca gggccagact
      gctctaaatt agcggcgctg ttaaacgctg ccgcgcacgt cccggtctga
      <.......................LacI.......................<
       r  s  k   i  a  a  v   i  q  s   p  a  h   l  a  l  s

5551  ggaggtggca acgccaatca gcaacgactg tttgcccgcc agttgttgtg
      cctccaccgt tgcggttagt cgttgctgac aaacgggcgg tcaacaacac
      <.......................LacI.......................<
        s  t  a   v  g  i   l  l  s  q   k  g  a   l  q  q

5601  ccacgcggtt gggaatgtaa ttcagctccg ccatcgccgc ttccactttt
      ggtgcgccaa cccttacatt aagtcgaggc ggtagcggcg aaggtgaaaa
      <.......................LacI.......................<
      a  v  r  n   p  i  y   n  l  e   a  m  a  a   e  v  k

5651  tcccgcgttt tcgcagaaac gtggctggcc tggttcacca cgcgggaaac
      agggcgcaaa agcgtctttg caccgaccgg accaagtggt gcgccctttg
      <.......................LacI.......................<
       e  r  t   k  a  s  v   h  s  a   q  n  v   v  r  s  v

5701  ggtctgataa gagacaccgg catactctgc gacatcgtat aacgttactg
      ccagactatt ctctgtggcc gtatgagacg ctgtagcata ttgcaatgac
      <.......................LacI.......................<
        t  q  y   s  v  g   a  y  e  a   v  d  y   l  t  v


5751  gtttcacatt caccaccctg aattgactct cttccgggcg ctatcatgcc
      caaagtgtaa gtggtgggac ttaactgaga gaaggcccgc gatagtacgg
      <.....LacI.....<<
      p  k  v  n   v  v
```

```
5801  ataccgcgaa aggttttgcg ccattcgatg gtgtccggga tctcgacgct
      tatggcgctt tccaaaacgc ggtaagctac cacaggccct agagctgcga

5851  ctcccttatg cgactcctgc attaggaagc agcccagtag taggttgagg
      gagggaatac gctgaggacg taatccttcg tcgggtcatc atccaactcc


5901  ccgttgagca ccgccgccgc aaggaatggt gcatgcaagg agatggcgcc
      ggcaactcgt ggcggcggcg ttccttacca cgtacgttcc tctaccgcgg

5951  caacagtccc ccggccacgg ggcctgccac catacccacg ccgaaacaag
      gttgtcaggg ggccggtgcc ccggacggtg gtatgggtgc ggctttgttc

6001  cgctcatgag cccgaagtgg cgagcccgat cttccccatc ggtgatgtcg
      gcgagtactc gggcttcacc gctcgggcta gaaggggtag ccactacagc

6051  gcgatatagg cgccagcaac cgcacctgtg gcgccggtga tgccggccac
      cgctatatcc gcggtcgttg gcgtggacac cgcggccact acggccggtg

6101  gatgcgtccg gcgtagagga tcgagatctc gatcccgcga aat
      ctacgcaggc cgcatctcct agctctagag ctagggcgct tta
```




```
  1  taatacgact cactataggg gaattgtgag cggataacaa ttcccctcta
     attatgctga gtgatatccc cttaacactc gcctattgtt aaggggagat
              1517

                                                    NheI
                                                   -+----
                                                NcoI
                                              -+-----
 51  gaaataattt tgtttaactt taagaaggag atataccatg gctagctgga
     ctttattaaa acaaattgaa attcttcctc tatatggtac cgatcgacct
                                             >>..S-McrA...>
                                               m   a  s  w
                                               StrepII >>.>
                                                       w

                  BstBI
                 --+---
101  gccacccgca gttcgaaaaa ggcgccatgc atgtttttga taataatgga
     cggtgggcgt caagcttttt ccgcggtacg tacaaaaact attattacct
     >.....................S-McrA.....................>
      s  h  p   q  f  e  k   g  a  m   h  v  f   d  n  n  g
     >.......StrepII......>>
      s  h  p   q  f  e  k

151  attgaactga aagctgagtg ttcgataggt gaagaggatg gtgtttatgg
     taacttgact ttcgactcac aagctatcca cttctcctac cacaaatacc
     >.....................S-McrA.....................>
       i  e  l   k  a  e   c  s  i  g   e  e  d   g  v  y
```

```
201  tctaatcctt gagtcgtggg ggccgggtga cagaaacaaa gattacaata
     agattaggaa ctcagcaccc ccggcccact gtctttgttt ctaatgttat
     >.....................S-McrA.....................>
     g  l  i  l   e  s  w   g  p  g   d  r  n  k   d  y  n

251  tcgctcttga ttatatcatt gaacggttgg ttgattctgg tgtatcccaa
     agcgagaact aatatagtaa cttgccaacc aactaagacc acatagggtt
     >.....................S-McrA.....................>
      i  a  l   d  y  i  i   e  r  l   v  d  s   g  v  s  q

                                                NdeI
                                              --+----
301  gtcgtagtat atctggcgtc atcatcagtc agaaaacata tgcattcttt
     cagcatcata tagaccgcag tagtagtcag tcttttgtat acgtaagaaa
     >.....................S-McrA.....................>
       v  v  v   y  l  a   s  s  s  v   r  k  h   m  h  s

351  ggatgaaaga aaaatccatc ctggtgaata ttttactttg attggtaata
     cctactttct ttttaggtag gaccacttat aaaatgaaac taaccattat
     >.....................S-McrA.....................>
     l  d  e  r   k  i  h   p  g  e   y  f  t  l   i  g  n

401  gcccccgcga tatacgcttg aagatgtgtg gttatcaggc ttattttagt
     cgggggcgct atatgcgaac ttctacacac caatagtccg aataaaatca
     >.....................S-McrA.....................>
      s  p  r   d  i  r  l   k  m  c   g  y  q   a  y  f  s

451  cgtacgggga gaaaggaaat tccttccggc aatagaacga aacgaatatt
     gcatgcccct ctttccttta aggaaggccg ttatcttgct ttgcttataa
     >.....................S-McrA.....................>
       r  t  g   r  k  e   i  p  s  g   n  r  t   k  r  i

501  gataaatgtt ccaggtattt atagtgacag tttttgggcg tctataatac
     ctatttacaa ggtccataaa tatcactgtc aaaaacccgc agatattatg
     >.....................S-McrA.....................>
     l  i  n  v   p  g  i   y  s  d   s  f  w  a   s  i  i

551  gtggagaact atcagagctt tcacagccta cagatgatga atcgcttctg
     cacctcttga tagtctcgaa agtgtcggat gtctactact tagcgaagac
     >.....................S-McrA.....................>
      r  g  e   l  s  e  l   s  q  p   t  d  d   e  s  l  l

601  aatatgaggg ttagtaaatt aattaagaaa acgttgagtc aacccgaggg
     ttatactccc aatcatttaa ttaattcttt tgcaactcag ttgggctccc
     >.....................S-McrA.....................>
       n  m  r   v  s  k   l  i  k  k   t  l  s   q  p  e

651  ctccaggaaa ccagttgagg tagaaagact acaaaaagtt tatgtccgag
     gaggtccttt ggtcaactcc atctttctga tgtttttcaa atacaggctc
     >.....................S-McrA.....................>
     g  s  r  k   p  v  e   v  e  r   l  q  k  v   y  v  r

701  acccgatggt aaaagcttgg attttacagc aaagtaaagg tatatgtgaa
     tgggctacca ttttcgaacc taaaatgtcg tttcatttcc atatacactt
     >.....................S-McrA.....................>
      d  p  m   v  k  a  w   i  l  q   q  s  k   g  i  c  e
```

```
 751  aactgtggta aaaatgctcc gttttattta aatgatggaa acccatattt
      ttgacaccat ttttacgagg caaaataaat ttactacctt tgggtataaa
      >.....................S-McrA.....................>
        n  c  g   k  n  a   p  f  y  l   n  d  g   n  p  y

 801  ggaagtacat catgtaattc ccctgtcttc aggtggtgct gatacaacag
      ccttcatgta gtacattaag gggacagaag tccaccacga ctatgttgtc
      >.....................S-McrA.....................>
      l  e  v  h   h  v  i   p  l  s   s  g  g  a   d  t  t

 851  ataactgtgt tgccctttgt ccgaattgcc atagagaatt gcactatagt
      tattgacaca acgggaaaca ggcttaacgg tatctcttaa cgtgatatca
      >.....................S-McrA.....................>
       d  n  c   v  a  l  c   p  n  c   h  r  e   l  h  y  s

 901  aaaaatgcaa aagaactaat cgagatgctt tacgttaata taaaccgatt
      tttttacgtt ttcttgatta gctctacgaa atgcaattat atttggctaa
      >.....................S-McrA.....................>
        k  n  a   k  e  l   i  e  m  l   y  v  n   i  n  r

                   BamHI
                  -+----
 951  acagaaataa tagggatccg aattcgagct ccgtcgacaa gcttgcggcc
      tgtctttatt atccctaggc ttaagctcga ggcagctgtt cgaacgccgg
      >.S-McrA>>
      l  q  k  -

1001  gcactcgagc accaccacca ccaccactga gatccggctg ctaacaaagc
      cgtgagctcg tggtggtggt ggtggtgact ctaggccgac gattgtttcg

1051  ccgaaaggaa gctgagttgg ctgctgccac cgctgagcaa taactagcat
      ggctttcctt cgactcaacc gacgacggtg gcgactcgtt attgatcgta
                                          1518

1101  aacccttgg ggcctctaaa cgggtcttga ggggtttttt gctgaaagga
      ttggggaacc ccggagattt gcccagaact ccccaaaaaa cgactttcct

1151  ggaactatat ccggattggc gaatgggacg cgccctgtag cggcgcatta
      ccttgatata ggcctaaccg cttaccctgc gcgggacatc gccgcgtaat

1201  agcgcggcgg gtgtggtggt tacgcgcagc gtgaccgcta cacttgccag
      tcgcgccgcc cacaccacca atgcgcgtcg cactggcgat gtgaacggtc

1251  cgccctagcg cccgctcctt tcgctttctt cccttccttt ctcgccacgt
      gcgggatcgc gggcgaggaa agcgaaagaa gggaaggaaa gagcggtgca

1301  tcgccggctt tccccgtcaa gctctaaatc gggggctccc tttagggttc
      agcggccgaa aggggcagtt cgagatttag cccccgaggg aaatcccaag

1351  cgatttagtg ctttacggca cctcgacccc aaaaaacttg attagggtga
      gctaaatcac gaaatgccgt ggagctgggg ttttttgaac taatcccact

1401  tggttcacgt agtgggccat cgccctgata gacggttttt cgccctttga
      accaagtgca tcacccggta gcgggactat ctgccaaaaa gcgggaaact
```

```
1451  cgttggagtc cacgttcttt aatagtggac tcttgttcca aactggaaca
      gcaacctcag gtgcaagaaa ttatcacctg agaacaaggt ttgaccttgt

1501  acactcaacc ctatctcggt ctattctttt gatttataag ggattttgcc
      tgtgagttgg gatagagcca gataagaaaa ctaaatattc cctaaaacgg

1551  gatttcggcc tattggttaa aaaatgagct gatttaacaa aaatttaacg
      ctaaagccgg ataaccaatt ttttactcga ctaaattgtt tttaaattgc

1601  cgaattttaa caaaatatta acgtttacaa tttcaggtgg cacttttcgg
      gcttaaaatt gttttataat tgcaaatgtt aaagtccacc gtgaaaagcc

1651  ggaaatgtgc gcggaacccc tatttgttta tttttctaaa tacattcaaa
      cctttacacg cgccttgggg ataaacaaat aaaaagattt atgtaagttt

1701  tatgtatccg ctcatgaatt aattcttaga aaaactcatc gagcatcaaa
      atacataggc gagtacttaa ttaagaatct ttttgagtag ctcgtagttt
                              <<.........KanR...........<
                                -   f  f  e  d   l  m  l

1751  tgaaactgca atttattcat atcaggatta tcaataccat atttttgaaa
      actttgacgt taaataagta tagtcctaat agttatggta taaaaacttt
      <.......................KanR.......................<
       h  f  q   l  k  n  m   d  p  n   d  i  g   y  k  q  f

1801  aagccgtttc tgtaatgaag gagaaaactc accgaggcag ttccatagga
      ttcggcaaag acattacttc ctcttttgag tggctccgtc aaggtatcct
      <.......................KanR.......................<
        l  r  k   q  l  s   p  s  f  e   g  l  c   n  w  l

1851  tggcaagatc ctggtatcgg tctgcgattc cgactcgtcc aacatcaata
      accgttctag gaccatagcc agacgctaag gctgagcagg ttgtagttat
      <.......................KanR.......................<
      i  a  l  d   q  y  r   d  a  i   g  v  r  g   v  d  i

1901  caacctatta atttcccctc gtcaaaaata aggttatcaa gtgagaaatc
      gttggataat taaaggggag cagtttttat tccaatagtt cactctttag
      <.......................KanR.......................<
       c  g  i   l  k  g  e   d  f  i   l  n  d   l  s  f  d

1951  accatgagtg acgactgaat ccggtgagaa tggcaaaagt ttatgcattt
      tggtactcac tgctgactta ggccactctt accgttttca aatacgtaaa
      <.......................KanR.......................<
        g  h  t   v  v  s   d  p  s  f   p  l  l   k  h  m

2001  ctttccagac ttgttcaaca ggccagccat tacgctcgtc atcaaaatca
      gaaaggtctg aacaagttgt ccggtcggta atgcgagcag tagttttagt
      <.......................KanR.......................<
      e  k  w  v   q  e  v   p  w  g   n  r  e  d   d  f  d

2051  ctcgcatcaa ccaaaccgtt attcattcgt gattgcgcct gagcgagacg
      gagcgtagtt ggtttggcaa taagtaagca ctaacgcgga ctcgctctgc
      <.......................KanR.......................<
       s  a  d   v  l  g  n   n  m  r   s  q  a   q  a  l  r
```

```
2101  aaatacgcga tcgctgttaa aaggacaatt acaaacagga atcgaatgca
      tttatgcgct agcgacaatt ttcctgttaa tgtttgtcct tagcttacgt
      <.......................KanR.......................<
        f  v  r   d  s  n   f  p  c  n   c  v  p   i  s  h

2151  accggcgcag gaacactgcc agcgcatcaa caatattttc acctgaatca
      tggccgcgtc cttgtgacgg tcgcgtagtt gttataaaag tggacttagt
      <.......................KanR.......................<
      l  r  r  l   f  v  a   l  a  d   v  i  n  e   g  s  d

2201  ggatattctt ctaatacctg gaatgctgtt ttcccgggga tcgcagtggt
      cctataagaa gattatggac cttacgacaa aagggcccct agcgtcacca
      <.......................KanR.......................<
       p  y  e   e  l  v  q   f  a  t   k  g  p   i  a  t  t

2251  gagtaaccat gcatcatcag gagtacggat aaaatgcttg atggtcggaa
      ctcattggta cgtagtagtc ctcatgccta ttttacgaac taccagcctt
      <.......................KanR.......................<
        l  l  w   a  d  d   p  t  r  i   f  h  k   i  t  p

2301  gaggcataaa ttccgtcagc cagtttagtc tgaccatctc atctgtaaca
      ctccgtattt aaggcagtcg gtcaaatcag actggtagag tagacattgt
      <.......................KanR.......................<
      l  p  m  f   e  t  l   w  n  l   r  v  m  e   d  t  v

2351  tcattggcaa cgctaccttt gccatgtttc agaaacaact ctggcgcatc
      agtaaccgtt gcgatggaaa cggtacaaag tctttgttga gaccgcgtag
      <.......................KanR.......................<
       d  n  a   v  s  g  k   g  h  k   l  f  l   e  p  a  d

2401  gggcttccca tacaatcgat agattgtcgc acctgattgc ccgacattat
      cccgaagggt atgttagcta tctaacagcg tggactaacg ggctgtaata
      <.......................KanR.......................<
        p  k  g   y  l  r   y  i  t  a   g  s  q   g  v  n

2451  cgcgagccca tttataccca tataaatcag catccatgtt ggaatttaat
      gcgctcgggt aaatatgggt atatttagtc gtaggtacaa ccttaaatta
      <.......................KanR.......................<
      d  r  a  w   k  y  g   y  l  d   a  d  m  n   s  n  l

2501  cgcggcctag agcaagacgt ttcccgttga atatggctca taacacccct
      gcgccggatc tcgttctgca aagggcaact tataccgagt attgtgggga
      <...................KanR...................<<
       r  p  r   s  c  s  t   e  r  q   i  h  s   m

2551  tgtattactg tttatgtaag cagacagttt tattgttcat gaccaaaatc
      acataatgac aaatacattc gtctgtcaaa ataacaagta ctggttttag

2601  ccttaacgtg agttttcgtt ccactgagcg tcagaccccg tagaaaagat
      ggaattgcac tcaaaagcaa ggtgactcgc agtctggggc atcttttcta

2651  caaaggatct tcttgagatc ctttttttct gcgcgtaatc tgctgcttgc
      gtttcctaga agaactctag gaaaaaaaga cgcgcattag acgacgaacg

2701  aaacaaaaaa accaccgcta ccagcggtgg tttgtttgcc ggatcaagag
      tttgtttttt tggtggcgat ggtcgccacc aaacaaacgg cctagttctc
```

```
2751  ctaccaactc tttttccgaa ggtaactggc ttcagcagag cgcagatacc
      gatggttgag aaaaaggctt ccattgaccg aagtcgtctc gcgtctatgg

2801  aaatactgtc cttctagtgt agccgtagtt aggccaccac ttcaagaact
      tttatgacag gaagatcaca tcggcatcaa tccggtggtg aagttcttga

2851  ctgtagcacc gcctacatac ctcgctctgc taatcctgtt accagtggct
      gacatcgtgg cggatgtatg gagcgagacg attaggacaa tggtcaccga

2901  gctgccagtg gcgataagtc gtgtcttacc gggttggact caagacgata
      cgacggtcac cgctattcag cacagaatgg cccaacctga gttctgctat

2951  gttaccggat aaggcgcagc ggtcgggctg aacgggggt tcgtgcacac
      caatggccta ttccgcgtcg ccagcccgac ttgccccca agcacgtgtg

3001  agcccagctt ggagcgaacg acctacaccg aactgagata cctacagcgt
      tcgggtcgaa cctcgcttgc tggatgtggc ttgactctat ggatgtcgca

3051  gagctatgag aaagcgccac gcttcccgaa gggagaaagg cggacaggta
      ctcgatactc tttcgcggtg cgaagggctt ccctctttcc gcctgtccat

3101  tccggtaagc ggcagggtcg gaacaggaga gcgcacgagg gagcttccag
      aggccattcg ccgtcccagc cttgtcctct cgcgtgctcc ctcgaaggtc

3151  ggggaaacgc ctggtatctt tatagtcctg tcgggtttcg ccacctctga
      ccccttgcg gaccatagaa atatcaggac agcccaaagc ggtggagact

3201  cttgagcgtc gatttttgtg atgctcgtca ggggggcgga gcctatggaa
      gaactcgcag ctaaaaacac tacgagcagt ccccccgcct cggatacctt

3251  aaacgccagc aacgcggcct ttttacggtt cctggccttt tgctggcctt
      tttgcggtcg ttgcgccgga aaaatgccaa ggaccggaaa acgaccggaa

3301  ttgctcacat gttctttcct gcgttatccc ctgattctgt ggataaccgt
      aacgagtgta caagaaagga cgcaataggg gactaagaca cctattggca

3351  attaccgcct ttgagtgagc tgataccgct cgccgcagcc gaacgaccga
      taatggcgga aactcactcg actatggcga gcggcgtcgg cttgctggct

3401  gcgcagcgag tcagtgagcg aggaagcgga agagcgcctg atgcggtatt
      cgcgtcgctc agtcactcgc tccttcgcct tctcgcggac tacgccataa

3451  ttctccttac gcatctgtgc ggtatttcac accgcatata tggtgcactc
      aagaggaatg cgtagacacg ccataaagtg tggcgtatat accacgtgag

3501  tcagtacaat ctgctctgat gccgcatagt taagccagta tacactccgc
      agtcatgtta gacgagacta cggcgtatca attcggtcat atgtgaggcg

3551  tatcgctacg tgactgggtc atggctgcgc cccgacaccc gccaacaccc
      atagcgatgc actgacccag taccgacgcg gggctgtggg cggttgtggg

3601  gctgacgcgc cctgacgggc ttgtctgctc ccggcatccg cttacagaca
      cgactgcgcg ggactgcccg aacagacgag ggccgtaggc gaatgtctgt

3651  agctgtgacc gtctccggga gctgcatgtg tcagaggttt tcaccgtcat
      tcgacactgg cagaggccct cgacgtacac agtctccaaa agtggcagta
```

```
3701  caccgaaacg cgcgaggcag ctgcggtaaa gctcatcagc gtggtcgtga
      gtggctttgc gcgctccgtc gacgccattt cgagtagtcg caccagcact

3751  agcgattcac agatgtctgc ctgttcatcc gcgtccagct cgttgagttt
      tcgctaagtg tctacagacg gacaagtagg cgcaggtcga gcaactcaaa

3801  ctccagaagc gttaatgtct ggcttctgat aaagcgggcc atgttaaggg
      gaggtcttcg caattacaga ccgaagacta tttcgcccgg tacaattccc

3851  cggttttttc ctgtttggtc actgatgcct ccgtgtaagg gggatttctg
      gccaaaaaag gacaaaccag tgactacgga ggcacattcc ccctaaagac

3901  ttcatggggg taatgatacc gatgaaacga gagaggatgc tcacgatacg
      aagtaccccc attactatgg ctactttgct ctctcctacg agtgctatgc

3951  ggttactgat gatgaacatg cccggttact ggaacgttgt gagggtaaac
      ccaatgacta ctacttgtac gggccaatga ccttgcaaca ctcccatttg

4001  aactggcggt atggatgcgg cgggaccaga gaaaaatcac tcagggtcaa
      ttgaccgcca tacctacgcc gccctggtct ctttttagtg agtcccagtt

4051  tgccagcgct tcgttaatac agatgtaggt gttccacagg gtagccagca
      acggtcgcga agcaattatg tctacatcca caaggtgtcc catcggtcgt

4101  gcatcctgcg atgcagatcc ggaacataat ggtgcagggc gctgacttcc
      cgtaggacgc tacgtctagg ccttgtatta ccacgtcccg cgactgaagg

4151  gcgtttccag actttacgaa acacggaaac cgaagaccat tcatgttgtt
      cgcaaaggtc tgaaatgctt tgtgcctttg gcttctggta agtacaacaa

4201  gctcaggtcg cagacgtttt gcagcagcag tcgcttcacg ttcgctcgcg
      cgagtccagc gtctgcaaaa cgtcgtcgtc agcgaagtgc aagcgagcgc

4251  tatcggtgat tcattctgct aaccagtaag gcaaccccgc cagcctagcc
      atagccacta agtaagacga ttggtcattc cgttggggcg gtcggatcgg

4301  gggtcctcaa cgacaggagc acgatcatgc gcacccgtgg ggccgccatg
      cccaggagtt gctgtcctcg tgctagtacg cgtgggcacc ccggcggtac

4351  ccggcgataa tggcctgctt ctcgccgaaa cgtttggtgg cgggaccagt
      ggccgctatt accggacgaa gagcggcttt gcaaaccacc gccctggtca

4401  gacgaaggct tgagcgaggg cgtgcaagat tccgaatacc gcaagcgaca
      ctgcttccga actcgctccc gcacgttcta aggcttatgg cgttcgctgt

4451  ggccgatcat cgtcgcgctc cagcgaaagc ggtcctcgcc gaaaatgacc
      ccggctagta gcagcgcgag gtcgctttcg ccaggagcgg cttttactgg

4501  cagagcgctg ccggcacctg tcctacgagt tgcatgataa agaagacagt
      gtctcgcgac ggccgtggac aggatgctca acgtactatt tcttctgtca

4551  cataagtgcg gcgacgatag tcatgccccg cgcccaccgg aaggagctga
      gtattcacgc cgctgctatc agtacggggc gcgggtggcc ttcctcgact

4601  ctgggttgaa ggctctcaag ggcatcggtc gagatcccgg tgcctaatga
      gacccaactt ccgagagttc ccgtagccag ctctagggcc acggattact
```

```
4651  gtgagctaac ttacattaat tgcgttgcgc tcactgcccg ctttccagtc
      cactcgattg aatgtaatta acgcaacgcg agtgacgggc gaaaggtcag
                                      <<......LacI........<
                                        -  q  g   s  e  l

4701  gggaaacctg tcgtgccagc tgcattaatg aatcggccaa cgcgcgggga
      ccctttggac agcacggtcg acgtaattac ttagccggtt gcgcgccccт
      <.......................LacI.......................<
      r  s  v  q   r  a  l   q  m  l   s  d  a  l   a  r  p

4751  gaggcggttt gcgtattggg cgccagggtg gtttttcttt tcaccagtga
      ctccgccaaa cgcataaccc gcggtcccac caaaaagaaa agtggtcact
      <.......................LacI.......................<
       s  a  t   q  t  n  p   a  l  t   t  k  r   k  v  l  s

4801  gacgggcaac agctgattgc ccttcaccgc ctggccctga gagagttgca
      ctgcccgttg tcgactaacg ggaagtggcg gaccgggact ctctcaacgt
      <.......................LacI.......................<
        v  p  l   l  q  n   g  k  v  a   q  g  q   s  l  q

4851  gcaagcggtc cacgctggtt tgccccagca ggcgaaaatc ctgtttgatg
      cgttcgccag gtgcgaccaa acggggtcgt ccgcttttag gacaaactac
      <.......................LacI.......................<
      l  l  r  d   v  s  t   q  g  l   l  r  f  d   q  k  i

4901  gtggttaacg gcgggatata acatgagctg tcttcggtat cgtcgtatcc
      caccaattgc cgccctatat tgtactcgac agaagccata gcagcatagg
      <.......................LacI.......................<
       t  t  l   p  p  i  y   c  s  s   d  e  t   d  d  y  g

4951  cactaccgag atatccgcac caacgcgcag cccggactcg gtaatggcgc
      gtgatggctc tataggcgtg gttgcgcgtc gggcctgagc cattaccgcg
      <.......................LacI.......................<
        v  v  s   i  d  a   g  v  r  l   g  s  e   t  i  a

5001  gcattgcgcc cagcgccatc tgatcgttgg caaccagcat cgcagtggga
      cgtaacgcgg gtcgcggtag actagcaacc gttggtcgta gcgtcaccct
      <.......................LacI.......................<
      r  m  a  g   l  a  m   q  d  n   a  v  l  m   a  t  p

5051  acgatgccct cattcagcat ttgcatggtt tgttgaaaac cggacatggc
      tgctacggga gtaagtcgta aacgtaccaa acaacttttg gcctgtaccg
      <.......................LacI.......................<
       v  i  g   e  n  l  m   q  m  t   q  q  f   g  s  m  a

5101  actccagtcg ccttcccgtt ccgctatcgg ctgaatttga ttgcgagtga
      tgaggtcagc ggaagggcaa ggcgatagcc gacttaaact aacgctcact
      <.......................LacI.......................<
        s  w  d   g  e  r   e  a  i  p   q  i  q   n  r  t

5151  gatatttatg ccagccagcc agacgcagac gcgccgagac agaacttaat
      ctataaatac ggtcggtcgg tctgcgtctg cgcggctctg tcttgaatta
      <.......................LacI.......................<
      l  y  k  h   w  g  a   l  r  l   r  a  s  v   s  s  l
```

```
5201  gggcccgcta acagcgcgat ttgctggtga cccaatgcga ccagatgctc
      cccgggcgat tgtcgcgcta aacgaccact gggttacgct ggtctacgag
      <.....................LacI.....................<
       p  g  a   l  l  a  i   q  q  h   g  l  a   v  l  h  e

5251  cacgcccagt cgcgtaccgt cttcatggga gaaaataata ctgttgatgg
      gtgcgggtca gcgcatggca gaagtaccct cttttattat gacaactacc
      <.....................LacI.....................<
        v  g  l   r  t  g   d  e  h  s   f  i  i   s  n  i

5301  gtgtctggtc agagacatca agaaataacg ccggaacatt agtgcaggca
      cacagaccag tctctgtagt tctttattgc ggccttgtaa tcacgtccgt
      <.....................LacI.....................<
      p  t  q  d   s  v  d   l  f  l   a  p  v  n   t  c  a

5351  gcttccacag caatggcatc ctggtcatcc agcggatagt taatgatcag
      cgaaggtgtc gttaccgtag gaccagtagg tcgcctatca attactagtc
      <.....................LacI.....................<
       a  e  v   a  i  a  d   q  d  d   l  p  y   n  i  i  l

5401  cccactgacg cgttgcgcga gaagattgtg caccgccgct ttacaggctt
      gggtgactgc gcaacgcgct cttctaacac gtggcggcga aatgtccgaa
      <.....................LacI.....................<
        g  s  v   r  q  a   l  l  n  h   v  a  a   k  c  a

5451  cgacgccgct tcgttctacc atcgacacca ccacgctggc acccagttga
      gctgcggcga agcaagatgg tagctgtggt ggtgcgaccg tgggtcaact
      <.....................LacI.....................<
      e  v  g  s   r  e  v   m  s  v   v  v  s  a   g  l  q

5501  tcggcgcgag atttaatcgc cgcgacaatt tgcgacggcg cgtgcagggc
      agccgcgctc taaattagcg gcgctgttaa acgctgccgc gcacgtcccg
      <.....................LacI.....................<
       d  a  r   s  k  i  a   a  v  i   q  s  p   a  h  l  a

5551  cagactggag gtggcaacgc caatcagcaa cgactgtttg cccgccagtt
      gtctgacctc caccgttgcg gttagtcgtt gctgacaaac gggcggtcaa
      <.....................LacI.....................<
        l  s  s   t  a  v   g  i  l  l   s  q  k   g  a  l

5601  gttgtgccac gcggttggga atgtaattca gctccgccat cgccgcttcc
      caacacggtg cgccaaccct tacattaagt cgaggcggta gcggcgaagg
      <.....................LacI.....................<
      q  q  a  v   r  n  p   i  y  n   l  e  a  m   a  a  e

5651  acttttccc gcgttttcgc agaaacgtgg ctggcctggt tcaccacgcg
      tgaaaaaggg cgcaaaagcg tctttgcacc gaccggacca agtggtgcgc
      <.....................LacI.....................<
       v  k  e   r  t  k  a   s  v  h   s  a  q   n  v  v  r

5701  ggaaacggtc tgataagaga caccggcata ctctgcgaca tcgtataacg
      cctttgccag actattctct gtggccgtat gagacgctgt agcatattgc
      <.....................LacI.....................<
        s  v  t   q  y  s   v  g  a  y   e  a  v   d  y  l
```

```
5751  ttactggttt cacattcacc accctgaatt gactctcttc cgggcgctat
      aatgaccaaa gtgtaagtgg tgggacttaa ctgagagaag gcccgcgata
      <.........LacI.........<<
      t  v  p  k   v  n  v   v

5801  catgccatac cgcgaaaggt tttgcgccat tcgatggtgt ccgggatctc
      gtacggtatg gcgctttcca aaacgcggta agctaccaca ggccctagag

5851  gacgctctcc cttatgcgac tcctgcatta ggaagcagcc cagtagtagg
      ctgcgagagg gaatacgctg aggacgtaat ccttcgtcgg gtcatcatcc

5901  ttgaggccgt tgagcaccgc cgccgcaagg aatggtgcat gcaaggagat
      aactccggca actcgtggcg gcggcgttcc ttaccacgta cgttcctcta

5951  ggcgcccaac agtcccccgg ccacggggcc tgccaccata cccacgccga
      ccgcgggttg tcaggggggcc ggtgccccgg acggtggtat gggtgcggct

6001  aacaagcgct catgagcccg aagtggcgag cccgatcttc cccatcggtg
      ttgttcgcga gtactcgggc ttcaccgctc gggctagaag gggtagccac

6051  atgtcggcga tataggcgcc agcaaccgca cctgtggcgc cggtgatgcc
      tacagccgct atatccgcgg tcgttggcgt ggacaccgcg gccactacgg

6101  ggccacgatg cgtccggcgt agaggatcga gatctcgatc ccgcgaaat
      ccggtgctac gcaggccgca tctcctagct ctagagctag ggcgcttta
```