Directionality of DNA Mismatch Repair in *Escherichia coli*

A M Mahedi Hasan

Thesis presented for the degree of Doctor of Philosophy
The University of Edinburgh
2014

Declaration

I hereby declare that this thesis was composed by me and that the research presented is my own, unless stated otherwise.

A M Mahedi Hasan

September 2014

Acknowledgement

I would like to thank David Leach for giving me the opportunity to carry out the research for my PhD under his supervision in his lab. His guidance and support were indispensable. It had been four years with great fun with wonderful people around, both past and present. I never became bored while Benura Azeroglu (along with Lisa) was around, stealing my EDTA solution and letting me lend from her as well. A special thanks goes out to Elise Darmon who had spent so much time (even in her busiest time) reviewing a very long and messy draft of my thesis.

I would like to thank all the members of the whole fourth floor of the Darwin building, for the many coffee breaks and cakes. A big thank goes out to Heidrun Interthal for her kind permission to use her printer (free of charge) while that in the Leach lab was out of order. And thanks for all the critical scientific criticises which led me to think twice before I could present any interesting result to people.

I would like to thank the people in the Genepool for all their help with this thesis and putting up with me for the last few years sending so many GeneMapper plates to their facility. And thanks to Ewa Okely for introducing me with the boring protocol (may be the only one that exists) for analysing GeneMapper data with the slowest computer in the world.

I would like to thank the Almighty as he gave me two worried parents who were too concerned about my health and studies, even from thousands of miles away.

The support I got from my loving wife (and a caring mother), Nurun Nahar Fancy, has always been unprecedented. I could not produce a better thesis without her critical suggestions during editing and formatting the final version of it. And life would have been so boring without the little monster, Maahnoor Binte Hasan, who always keeps us busy with all her trouble makings and loads of fun.

I would also like to thank my second supervisor, David Finnegan for his supervision and monthly meetings. And, finally I would like to thank the Darwin Trust of Edinburgh for granting me the scholarship that allowed me to study for my PhD at the University of Edinburgh.

Lay Summary

DNA is the blue-print of life which genetically characterises one organism from another. Copying of DNA from a pre-existing parental DNA template ensures that the genetic information is carried on throughout generations. However, sometimes errors occur during this copying process like misspellings occur during writing. This type of errors is called mismatches. They can lead to genetic disease, cancer and death. The main system that corrects this type of errors is called the DNA mismatch repair (MMR) system. The MMR system is evolutionarily conserved from bacteria to humans and is an example of a remarkable molecular machine. However, the complete molecular mechanism of the MMR system is not elucidated yet. Here I have shown that this process operates in one direction with respect to the chromosome of *Escherichia coli.* The most likely explanation for this directionality of action is that the MMR system is associated with the complex responsible for chromosomal DNA copying. During repair, a portion of DNA is removed along with the mismatch and this gap is filled in with the correct sequence. I suggest that there exists two classes of DNA removal processes: removal of a short sequence that is primarily responsible for MMR and of a long sequence that also leads to the deletion of DNA at a reporter sequence. GATC motifs in the genomic DNA are essential for MMR to occur. I tested whether MMR causes any selection pressure for the distribution of GATC motifs in the *E. coli* chromosome. I have shown that the MMR system has evolved to utilise the natural random distribution of GATC motifs to maintain genomic integrity. To my knowledge, this study is the first proven support of the chromosomal directionality of MMR which shades more light on the association of the DNA copying and correction processes to maintain the genetic integrity of life.

Abstract

Non-canonical base pairs that escape the proof-reading activity of the DNA polymerase emerge from DNA replication as DNA mismatches. To promote genomic integrity, these DNA mismatches are corrected by a secondary protection system, called DNA mismatch repair (MMR). Understanding the details of MMR is important for human health as defects in mismatch repair can result in cancer (*e.g.* hereditary nonpolyposis colorectal cancer, also known as Lynch syndrome).

Being normally stochastic in nature, mismatches can emerge at random locations in a chromosome. Therefore, using a molecular tool to generate substrates for the MMR system at a defined locus has been particularly useful in my study of DNA mismatch repair *in vivo*. In this study, I have used a CTG•CAG repeat array, also called the "TNR array", to generate frequent substrates for the MMR system in *Escherichia coli*. In *E. coli*, the MMR system searches for hemi-methylated GATC motifs around a mismatch to initiate removal of the faulty nascent (un-methylated) strand. Analysing the usage of GATC motifs around the TNR array, I have found that the MMR system preferentially utilizes the GATC motifs on the origin distal side of the TNR array demonstrating that the bidirectionality of MMR *in vitro* is constrained in live cells. My results suggest that *in vivo* MMR operates by searching for the nearest hemimethylated GATC site located between the mismatch and the replication fork and excision of the nascent strand occurs directionally away from the fork towards the mismatch.

Previous *in vitro* studies have established that the excision reaction during MMR terminates at a discrete point about 100 bp beyond a mismatch. However, *in vivo* recombination at a 275 bp tandem repeat, which has been proposed to be mediated by single stranded DNA generated during the excision reaction, has suggested that the end point of the excision reaction in live cells may extend much further from the mismatch than this. I have used this assay for extended excision to determine the influence of GATC sites on excision tracts. In this study, modification of the GATC motifs on the origin proximal side of the TNR has shown that the excision reaction does not stop at a GATC motif on the origin proximal side of the mismatch. In addition, sequential modifications of GATC motifs on the origin distal side of the TNR array, thereby shifting the start point of the excision reaction to a greater distance, have suggested that the length of an excision tract is a function of the distance it covers from the start point rather than from a mismatch.

My observation of directionality with respect to DNA replication in the recognition of GATC sites suggested that MMR and DNA replication might be coupled in some way and that perhaps active (or blocked) MMR might impede the progress of the replication fork. However, no replication intermediates were detected using two-dimensional agarose gel electrophoresis of genomic DNA fragment containing the TNR array upon restriction digestion. I was therefore unable to support the hypothesis that active or blocked MMR led to a slowing down of DNA replication.

Given my observation of a decrease in MMR by separating the mismatch from the closest origin distal GATC site, I set out to test whether MMR caused any selection pressure for the genomic distribution of GATC motifs. To do this, I generated artificial model genomes using a Markovian algorithm based on the nucleotide composition and codon usage in *E. coli*. Strikingly, the comparison of the distribution of GATC motifs in the *E. coli* genome with those from artificial sequences has shown that GATC motifs are distributed randomly in *E. coli* genome, except for a small clustering effect which has been detected for short spaced (0-40 basepairs) GATC motifs. The observed distribution of slightly over-represented GATC motifs in the *E. coli* genome appears to be a function of the total number of GATC motifs and it seems that the DNA mismatch repair system has evolved to utilize the natural distribution of GATC motifs to maintain genomic integrity.

Abbreviations

ANOVA	Analysis of variance
ATP	Adenosine triphosphate
BIR	Break-induced replication
Bp	Base-pair
CFU	Colony forming unit
cm	Centimetre
Cm^R	Chloramphenicol resistance cassette
°C	Degrees Celsius
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded DNA
g	Gram
HR	Homologous recombination
JM	Joint molecule
Kb	Kilo basepair
l	Litre
M	Molar
MCS	Multiple cloning site
µg	Microgram
µl	Microlitre
µm	Micrometre
µM	Micromolar
mg	Milligram
ml	Millilitre
mM	Millimolar
ng	Nanogram
NHEJ	Non-homologous end joining
nm	Nanometre
OD_{600nm}	Optical density at 600 nm

PCR	Polymerase chain reaction
pH	Power of hydrogen
PMGR	Plasmid-mediated gene replacement
RER	Ribonucleotide excision repair
ROS	Reactive oxygen species
RNS	Reactive nitrogen species
ssDNA	Single-stranded DNA
T_m	Melting temperature
UV light	Ultra-violet light
v/v	Volume per unit volume
w/v	Weight per unit volume

Table of contents

Chapter 3. The DNA mismatch repair machinery moves towards the replication fork in its search for a hemimethylated GATC site............123-136

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The philosophical question about “life” has been asked perhaps from time immemorial and is yet to be answered even in this era of scientific advancements. How life is sustained - still amazes the world of science and finding answers to the question seems to be a never ending quest. At the very molecular level, life is maintained with a common set of rules as all living organisms share common genetic building block – sugar-phosphate based nucleic acids – be it DNA or RNA. Life is the cumulative function of maintenance, actions and/or interactions of different components, which actually are coded by the nucleic acid – the blue-print of life. Thus, it is imperative that the nucleic acid is kept safe and maintained accurately throughout generations. However, this delicate chemical compound is under continuous risk of being altered or damaged by different components from both inside and outside its biological environment. Nucleic acid can be damaged as a result of exposure to different chemical and physical agents like benzo[α]pyrene, polychlorinated biphenyls, dioxin, cigarette smoke, asbestos, ultraviolet light, radon *etc* (Li, 2008). Even in its ambient biological environment, it can be damaged by different endogenous reactive metabolites like reactive oxygen and nitrogen species (ROS and RNS) (Li, 2008). On the other hand, nucleic acid sequence can be altered as a result of errors arising during normal and/or aberrant DNA metabolism, which includes replication, transcription, recombination and repair (Li, 2008). In addition, mismatches can arise from the deamination of 5-methylcytosine (G•5-mC), a modified base normally present in the DNA of both prokaryotes and eukaryotes as a mode of gene regulation (Friedberg et al., 2006). The deamination converts a G•5-mC base pair to G•T base pair which is not a Watson-Crick base-pair (Figure 1.1).

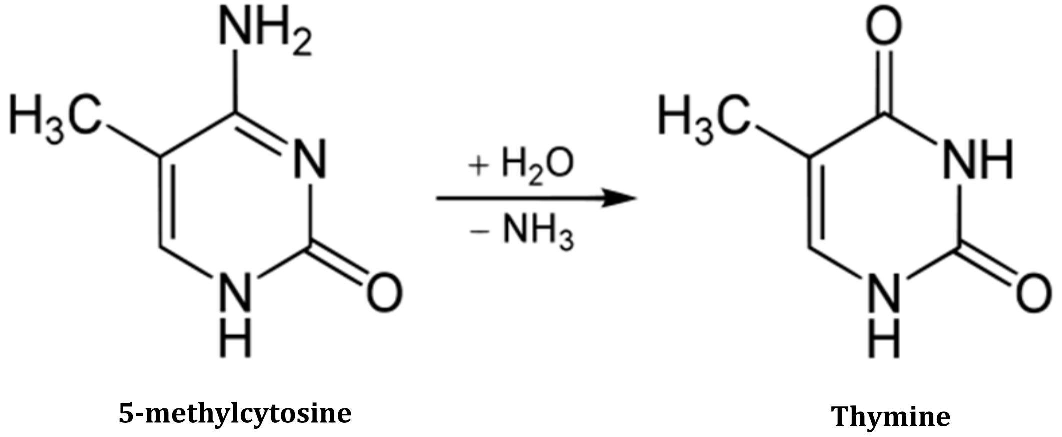


Figure 1.1. Deamination of 5-methylcytosine into thymine. Deamination of 5-methylcytosine into thymine forms a non Watson-Crick base-pair G•T. If not repaired, this mispair eventually can form an A•T base-pair after replication.

Throughout the course of evolution, nucleic acid has been maintained very efficiently with minimal changes. However, the basic rule remains the same: the genetic material of an organism is replicated, as well as, passed on to the next generation and thus the species is sustained. During this process, the genetic blue-print can be altered or damaged by aforementioned ways, in which the main contributor is DNA replication error - that is, incorporating a wrong base while synthesising a new copy of DNA. It has been found that the rate of this type of error is as much as 10^{-5} per base per replication of a genome, even though the replicative polymerase has its own proof-reading capability. This kind of error has usually been defined as a mismatch. If this error is not corrected, it stays permanently in the genome and is inherited by the next generation (Figure 1.2). In addition, a mutation in the coding or regulatory sequence can alter the molecular mechanism, disrupt a cellular function and eventually can lead to deadly diseases (e.g., predisposition to cancer).

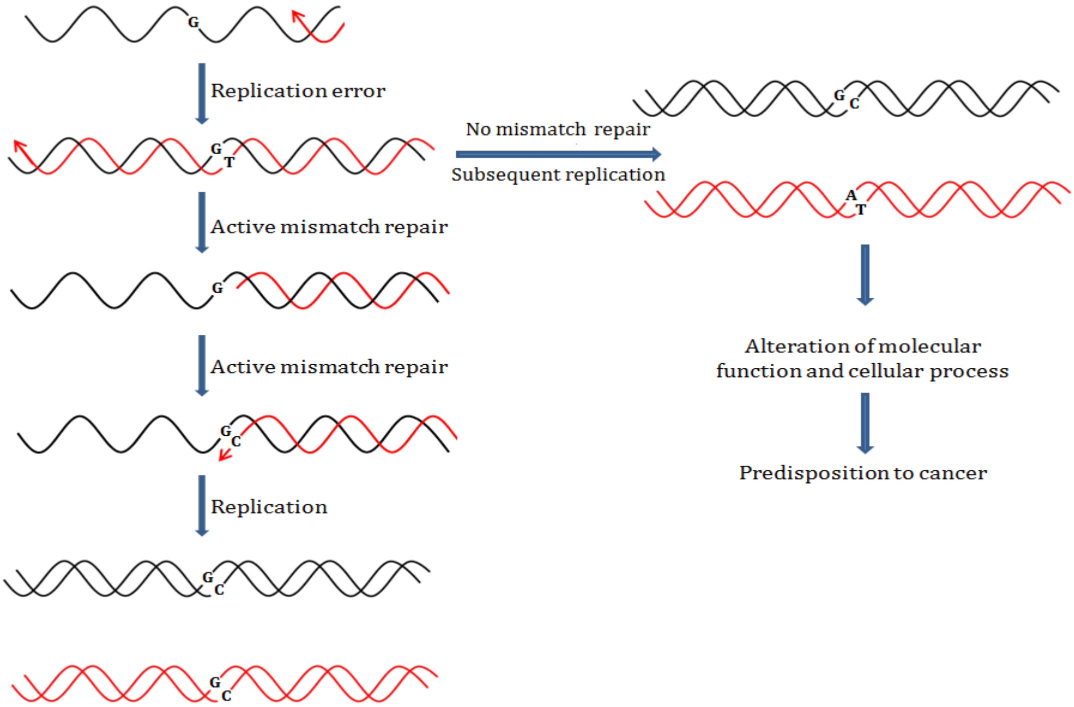


Figure 1.2. An unrepaired mismatch in the genome can lead to a deadly disease. A mismatch arises due to an error in replication. However, a secondary level of protection (after the proof-reading activity of the replication machinery) – the DNA mismatch repair system repairs the mismatch to ensure genomic integrity. However, in the case of faulty mismatch repair, the mismatch is inherited by the next generation and can cause a mutation in the coding or regulatory sequence in the genome, which can alter a molecular function and disrupt a cellular process. In turn, this might lead to a deadly disease, like cancer.

A cell deploys different levels of repair mechanism for those kinds of errors. The mismatch repair (MMR) system is one of the key guardians that ensure genomic integrity of that organism. This system is conserved in almost all the organisms found on earth, with the exception of *Actinobacteria, Mollicutes*, part of archea and some other bacteria (*e.g., Mycobacterium tuberculosis, Helicobacter pylori*)

(Sachadyn, 2010). The existence of the DNA MMR pathway was first reported in the early 1960s to explain the unanticipated segregation of genetic markers in fungi and bacteria. Robin Holliday and Evelyn Witkin independently proposed the existence of the MMR system in 1964 (Holliday, 1964; Whitehouse, 1963; Witkin and Sicurella, 1964).

1.2 Mismatches and the mismatch repair system

The MMR system recognises a single base-pair mismatch on a newly synthesised strand and replaces it with the correct nucleoside tri-phosphate. However, not all mismatches are equally good substrates for the MMR system as the binding affinity of MutS, the first protein of the MMR pathway that is capable of recognising a mismatch, varies from 10 to more than 1500 times compared to its binding affinity to a Watson-Crick base-pair (Gradia et al., 2000; Schofield et al., 2001). In addition, MutS can recognise and repair insertion/deletion loops, also called IDLs, which are usually 3-4 base-long partnerless extrahelical nucleotides formed by the slippage of primers or template strands during replication (Jiricny et al., 1988a). It has been found that the binding affinity varies with the composition of the mismatch and the local sequence context (Jiricny et al., 1988a; Mazurek et al., 2009). Among all eight possible single nucleotide mismatches the highest affinity has been found, by the gel-shift assay, for the G-T mismatch (gives rise to transition mutation) and IDLs, which are the most frequent mis-incorporations/ replication errors (Jiricny et al., 1988b). Actually, this function gave MutS its first acronym – GTBP (G/T-binding protein) (Palombo et al., 1995). A-C (gives rise to transition mutation), A-A and G-G mismatches (give rise to transversion mutation) typically are good substrates for MMR, whereas T-T, T-C or A-G mismatches are recognised with less efficiency. The lowest affinity has been found for the C-C mismatch which is the rarest mis-incorporation (Huang and Crothers, 2008; Nakahara et al., 2000; Su et al., 1988). However, the *Arabidopsis thaliana* MSH2-MSH7, which are homologues of MutS protein in eukaryotes, binds strongly to C-G and G-A substrates, not to the +1 IDLs (Wu et al., 2003). On the other hand, heteroduplexes containing large non-homologies are not processed by this system (Dohet et al., 1987). The relative affinity, which is the measure of recognition of a mismatch by MutS protein, has been determined by gel-shift studies. Therefore, these data should be interpreted with caution as this assay detects preferentially protein/DNA complexes with slow k_{off}. In addition, transiently bound mispairs may persist long enough to trigger the MMR *in vivo* and those were not detected in these studies.

1.3 A brief introduction of the mismatch repair system

In broad terms, the MMR system has to meet two basic criteria (Jiricny, 2006):

1. Recognition of a mismatch or similar substrate (*e.g.* IDL).
2. Directing an arsenal of proteins to take part in MMR on the erroneous nascent strand.

How these two tasks are accomplished was first elucidated by studying different *E. coli* strains which were mutant for the key proteins of the MMR system and eventually by reconstructing the prototypic MMR system (*in vitro*) from different combinations of purified proteins (Lahue et al., 1989). The name of the proteins and their homologues in eukaryotes, along with their functions are mentioned in Table 1.1.

Respective roles for these proteins have been proposed (and now established) during mismatch repair in *E. coli* (Figure 1.3) (Modrich, 1991). When the replisome creates a mismatch by inserting a wrong nucleoside tri-phosphate in the nascent strand, the first component of the MMR system - MutS protein finds the mismatch and binds it. Then, MutS activates and binds the second protein - MutL, as well as, ADP. Next, the MutS-ADP-MutL complex activates the nascent-strand specific endonuclease, MutH. This endonuclease makes an incision on the nascent strand harbouring a mismatch at a hemimethylated GATC motif which could be either on the 3' side or the 5' side within 2 kb of the mismatch. It should be noted that the GATC is usually methylated at the adenine by the Dam-methylase enzyme. However, just after the passage of the replisome, GATC on the nascent strand is not immediately methylated as the Dam-methylase lags behind the replication fork by about 0.5 - 3 minutes and the MMR system utilises this window of time to repair any existing mismatch (Barras and Marinus, 1989). After cleavage of DNA by MutH, a DNA helicase, UvrD unwinds the nascent strand and appropriate exonuclease(s) excise(s) the nascent strand. The remaining single-stranded DNA, corresponding to the parent strand, is bound by the single strand binding proteins (Ssb). The Polymerase III holoenzyme synthesises the nascent strand without any error and the DNA ligase seals the remaining nick at the end of the nascent strand which has just been repair synthesised. The detail of the molecular mechanism of the MMR system is still a matter of discussion/arguments, which will be addressed in the following sections.

Table 1.1. Major proteins of the MMR system in prokaryotes and eukaryotes along with their functions (Kunkel and Erie, 2005)

E. coli	Function	Homologues in eukaryotes	Function
MutS	Binds mismatches and IDLs	MutSα (MSH2, MSH6)	Recognition of base–base mismatches and small IDLs
		MutSβ (MSH2, MSH3)	Recognition of IDLs
MutL	Matchmaker that coordinates multiple steps in MMR	MutLα (MLH1, PMS2)	Forms a ternary complex with mismatch DNA and MutSα; increases discrimination between heteroduplexes and homoduplexes; also functions in meiotic recombination
		MutLβ (MLH1, PMS1)	Unknown
		MutLγ (MLH1, MLH3)	Primary function in meiotic recombination; backup for MutLα in the repair of base–base mismatches and small IDLs
MutH	Nicks unmethylated nascent strand at hemimethylated GATC sites	No homologues have been found yet	N/A
UvrD(MutU)	Loaded onto DNA at nick by MutS and MutL. Unwinds DNA to allow excision of ssDNA	None	N/A

ExoI, ExoX	Perform 3’ to 5’ excision of ssDNA	EXOI (Rth1)	3’-5’ Exonuclease; mismatch excision
ExoVII, RecJ	Perform 5’ to 3’excision of ssDNA (ExoVII also has 3’ to 5’ excision capability)	3’ exo of Pol δ 3’ exo of Pol ε	5’-3’ Exonuclease; mismatch excision
Pol III Holoenzyme	Accurate repair synthesis	DNA pol δ	Accurate repair synthesis
β clamp	Interacts with MutS and MutL. May recruit them to mismatches. Enhances processivity of DNA pol III	PCNA	Interacts with MutS and MutL homologues. Recruits MMR proteins to mismatches. Increases mismatch binding specificity of Msh2-Msh6. Participates in excision. Participates in DNA repair synthesis.
γ complex	Loads β-clamp onto DNA	RFC complex	Loads PCNA onto DNA. Modulates excision polarity
Ssb	Single-stranded DNA-binding protein. Participates in excision and repair synthesis	RPA	Single-stranded DNA-binding protein; repair synthesis
		HMGB1	Accessory protein; stimulated excision
		PARP	Accessory protein; improved mismatch selectivity
DNA ligase	Seals nicks after the completion of DNA (repair) synthesis	DNA ligase	Seals the nick after repair synthesis

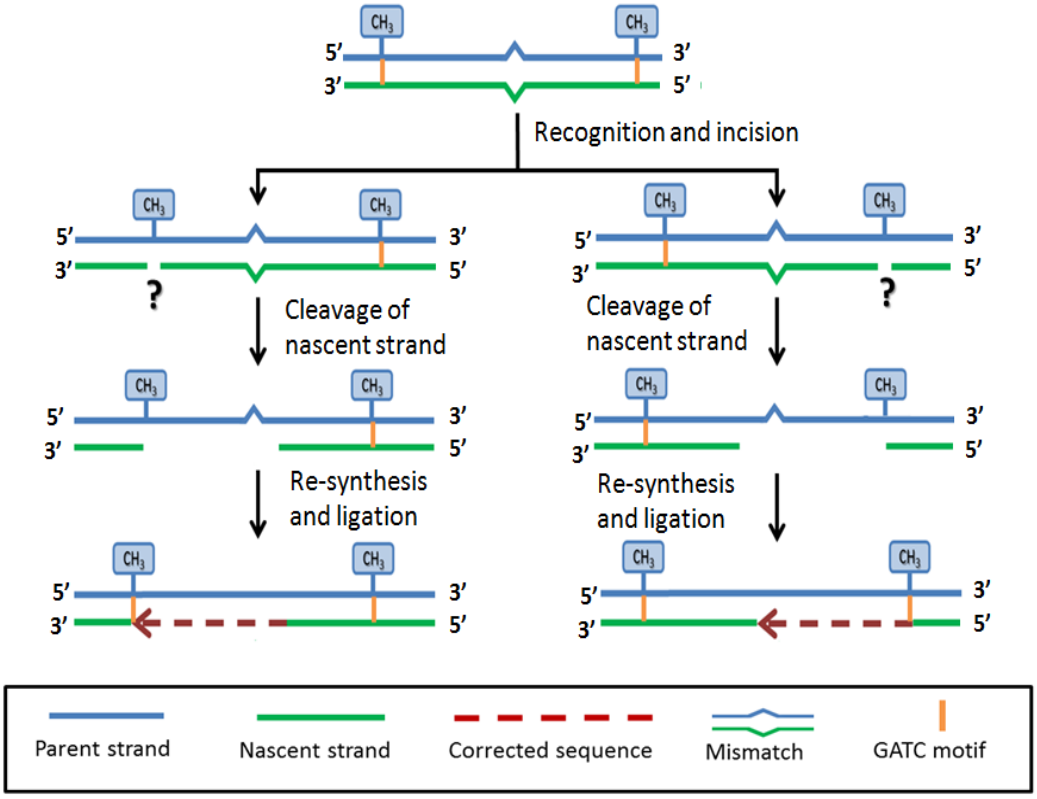


Figure 1.3. Molecular mechanism of the mismatch repair system. When the replisome creates a mismatch by inserting a wrong nucleoside tri-phosphate in the nascent strand, MutS protein finds the mismatch and binds it. MutL and MutH come along while MutH makes an incision on the nascent strand at the hemimethylated GATC motif - either on the 3' side or the 5' side of the mismatch. DNA helicase, UvrD unwinds the nascent strand and appropriate exonuclease(s) excise(s) the nascent strand. The remaining single-stranded parent strand is then bound by the single strand binding proteins (Ssb). The Polymerase III holoenzyme synthesises an error-free nascent strand and the DNA ligase seals the remaining nick at the end of the nascent strand.

1.4 Proteins in the MMR system and their functions in DNA mismatch repair

1.4.1 MutS and its homologues recognise the mismatch

The MutS protein functions as a regional lesion sensor. It is a member of a protein family called ABC proteins (ATP-binding cassette) as it contains a highly conserved ATPase domain at its carboxy-terminal. 300 proteins sequenced from 169 species containing this motif have been divided into 5 groups (Sachadyn, 2010). Only the first group, namely "MutS1", contains proteins that are involved in DNA mismatch repair. Structurally, prokaryotic MutS proteins function as homodimers and the subunits are associated with each other at the carboxyl termini in a way that their ATP-binding domain becomes partially intertwined. However, functionally, MutS shows asymmetry while binding to a mismatch as only one subunit contacts the mismatch via a phenylalanine residue of a highly conserved N-terminal Glu-X-Phe domain (Lamers et al., 2000; Obmolova et al., 2000). MutS binds to the mismatch by inserting the Phe residue into the minor groove of the duplex DNA and thus induces a 60° bend in the DNA (Figure 1.4). This bending is stabilised by the formation of hydrogen bonds between N^7 or N^3 atom of glutamate and purine or pyrimidine respectively (Natrajan et al., 2003). Further asymmetry is observed regarding the way the MutS homodimer binds a nucleotide by the Walker-ATP-binding sites. In *E. coli,* one of the two ATP binding sites binds ADP and the other remains unoccupied while MutS contacts a mismatch. However, it is different in the case of *Thermus aquaticus* (Junop et al., 2001; Lamers et al., 2003, 2004). In solution, free *E. coli* MutS protein has been found to utilise both ATP-binding sites as one binds ADP and the other one binds ATPγS (Baitinger et al., 2003; Monti et al., 2011).

Strikingly, a true asymmetry (both structural and functional) is observed in MutS homologues in eukaryotes, which forms a heterodimer. So far, the only exception found is *Sacchromyces cerevisiae* mitochondrial MutS homologue which forms a homodimer (Chi and Kolodner, 1994). Of the eight eukaryotic MSH (**M**ut**S** **H**omologue) polypeptides found to date, MSH1, MSH6 and MSH7 contain the Glu-X-Phe motif. MSH6 and MSH7 form heterodimers with MSH2 separately to form MutSα and MutSγ respectively. MutSα, which closely resembles the structure of bacterial MutS, recognises mismatches and IDLs of 1-2 nucleotides using the conserved Phe residue of the MSH6 subunit (Warren et al., 2007). Another heterodimer, namely MutSβ (MSH2-MSH3), is involved in repairing a subset of small IDLs and some mismatches in *S. cerevisiae* even though it lacks the conserved Glu-X-Phe motif (Harrington and Kolodner, 2007). The MSH3 in MutSβ interacts with the sugar-phosphate backbone of the IDLs via several basic and polar amino acids (Gupta et al., 2012). MutSβ has also been found to be involved in processing of branched structures, double strand break repair and expansion of tri-nucleotide repeats (Peña-Diaz and Jiricny, 2012; Sugawara et al., 1997; Surtees and Alani, 2006). MSH4 is involved in meiotic recombination and forms a hetero-oligomer with MSH5 (Hirano and Noda, 2004; Winand et al., 1998). MSH1 is thought to be involved in mismatch repair in mitochondria (Reenan and Kolodner, 1992). On the other hand, no biochemical information is yet available for MSH8 (Jiricny, 2013).

1.4.2 MutS and its interaction with ADP/ATP

While the N-terminus of MutS and its homologues are dedicated to recognising mismatch and IDLs, the ATP-binding domain of the C-terminus is devoted towards the next stages of the MMR pathway (Dufner et al., 2000; Lee et al., 2007; Malkov et al., 1997; Sachadyn, 2010; Shell et al., 2007). The occupancy of ATP-binding sites and the ATPase activity has been studied for the last 15-20 years extensively and still lies at the centre of arguments as different models have been proposed for this along with the later stages of the MMR.

Several lines of early experiments led to the proposal that upon binding ADP, MutS and its orthologs form a loose clamp around the DNA and travel along the helix by lateral, corkscrew-like movement (Jiricny, 2013). It has been found that ATP-bound MutS homologue cannot productively recognise a mismatch on DNA (Gradia et al., 1999). However, with only ADP, MutS shows direct dissociation from the DNA substrate as the formation of sliding clamp is ATP-dependent. Another model proposed that a spatial movement through the space led MutS to find a mismatch on the nascent strand and this movement is fuelled by the hydrolysis of ATP (Cho et al., 2012). Upon detecting a mismatch (possibly by detection of perturbed stacking and hydrogen bonding interactions), MutS pauses and bends the DNA, and this conformational change brings about a rapid ADP to ATP exchange and concomitant inhibition of ATP hydrolysis. In the ATP-bound form, the protein retracts the phenylalanine residue out of the helix and diffuses freely along the helix (Cho et al., 2012; Jiricny, 2006). A similar mechanism was proposed for MutSβ in human (Gupta et al., 2012). Those findings mentioned above had led to the proposal of a model named "the praying hands model" (Figure 1.4) (Jiricny, 2000).

From different lines of *in vitro* evidence it is obvious that the mismatch and ATP activated MutS sliding clamp leaves the mismatch and slides along the DNA. This could serve two purposes:

1. It would empty the mismatch site for allowing new MutS loading on it (Jeong et al., 2011).
2. The diffused sliding clamp could interact with the excision-repair synthesis factors either by recruiting them at the incision site or activating those factors that are already bound there.

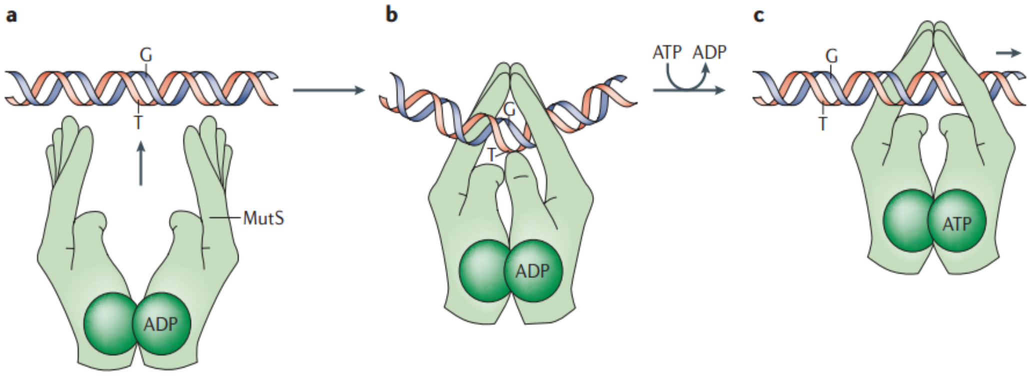


Figure 1.4. The praying hands model of binding to DNA and translocation of MutS. a) In the absence of DNA, the ATP-binding domains of the two MutS subunits become interlocked whereas the remainder of the proteins stays disordered. b) Upon contacting a DNA mismatch, MutS dimer anchors itself on the substrate and wraps around to form a clamp-like stable structure that resembles a pair of praying hands. One of the subunits inserts the "Phe" residue of the "Phe-X-Glu" motif (resembling the thumb of one of the hands) into the minor groove of the helix at the mismatch site. This results in a bending of the DNA by ~60˚. ATP-binding site of the subunit that contacts the mismatch is occupied by an ADP molecule while the ATP-binding site of other subunit stays empty. c) An ADP→ATP exchange converts the static clamp-like MutS dimer into a sliding clamp (also called "thumb-out, fingers-closed" structure) which moves

along the DNA contour. The MutS–ATP complex has been postulated to translocate by forcing the Phe-containing ‘thumb’ out of the DNA through steric interactions, while keeping the ‘fingers’ closed around it. The transformation and translocation require ATP binding but not its hydrolysis (Jiricny, 2006).

The interaction between the ternary complex and excision-repair synthesis factors was also proposed to be accomplished by 3D diffusion of MutS or the MutS-MutL-DNA ternary complex (Wang and Hays, 2003, 2004). However, recent findings support the idea that MutS or MutS-MutL complex translocates along the DNA contour (Cho et al., 2012; Gorman et al., 2007; Pluciennik and Modrich, 2007).

1.4.3 MutL and its homologues

MutL proteins are ATPases of the GHKL (gyrase/Hsp90/histidine-kinase/MutL) family (Dutta and Inouye, 2000). The ATPase domain is situated in the N-terminus and the dimerisation domain is at the C-terminus of the protein. In *E. coli*, the MutL protein works as a homodimer and is recruited by the dimer of ATP-activated MutS which has already recognised a mismatch on the double stranded DNA. This interaction is thought to modulate the ATP hydrolysis dependent turnover of the MutS-ATP-MutL complex and the downstream interactions with other proteins of the MMR system. Four homologues of prokaryotic MutL have been found in eukaryotes, namely – PMS1, PMS2, MLH1 and MLH3. Their respective roles in MMR have been mentioned in Table 1.1.

1.4.4 Interaction between MutS and MutL: the formation of a ternary complex

Though the MMR is best understood by studying *E. coli* model, the initial interactions among MMR proteins (especially MutS and MutL) and DNA have been studied mostly with their homologues in eukaryotes (Harfe and Jinks-Robertson, 2000; Modrich, 1991; Modrich and Lahue, 1996). Among a handful of models on this interaction, four are mostly supported by the authors of different reviews and mainstream experimental literatures –

1. The first model postulates that the MSH complex forms a "sliding clamp" upon addition of ATP and diffuses along DNA, permitting the downstream reactions to take place (Figure 1.5) (Gradia et al., 2000). In addition, (which has been suggested for bacterial MutS protein as well) this diffusion is assisted by loading of a constant flux of new MSH complexes on the freshly unoccupied mismatch. This gives a gradient of MSH complexes heading in a unidirectional manner (Acharya et al., 2003). The MLH complex presumably comes into action by binding the sliding clamps to take part in the signalling process.

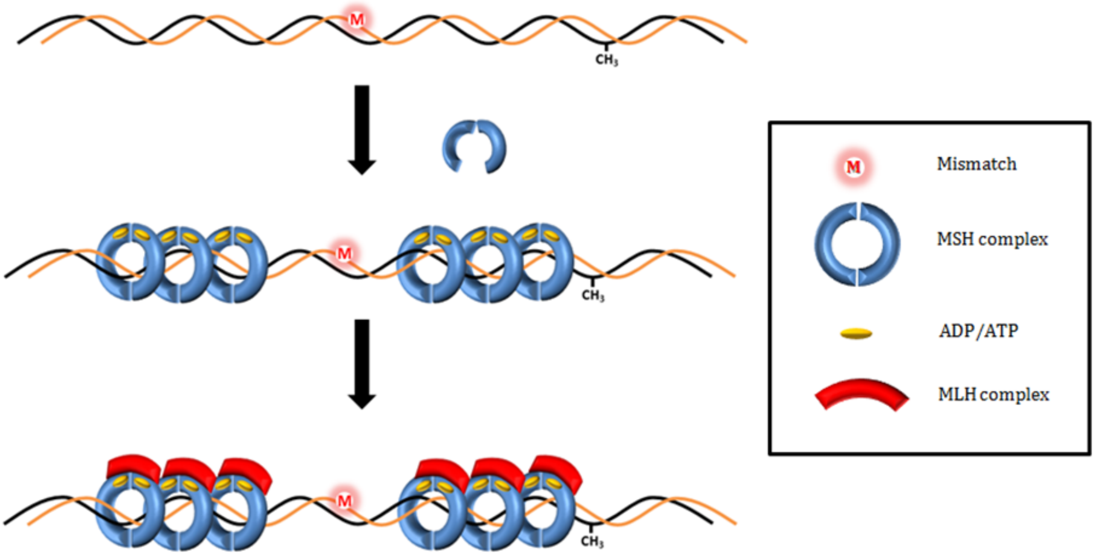


Figure 1.5. MutS complex (or dimer) forms a gradient along DNA substrate. Upon binding a mismatch, in the presence of ATP, the MutS dimer translocates along the DNA substrate. This translocation process generates a gradient of MutS complex away from the mismatch. The MLH complex presumably comes into action by binding the sliding clamps to take part in the signalling process.

2. The second model, which is also called the "active translocation model", states that the MutSα complex binds a mismatch and translocates along the DNA in an ATP hydrolysis dependent manner (Figure 1.6) (Blackwell et al., 1998). This model is analogous to the model postulated for bacterial MutS, which suggests a looping out of DNA as it is spooled through MutS and that other proteins interact with MutS for downstream signalling and initiation of excision reaction (Allen et al., 1997). However, it has also been argued that the translocation is probably ATP hydrolysis independent while the hydrolysis happens at a later stage (Gradia et al., 1997, 1999).

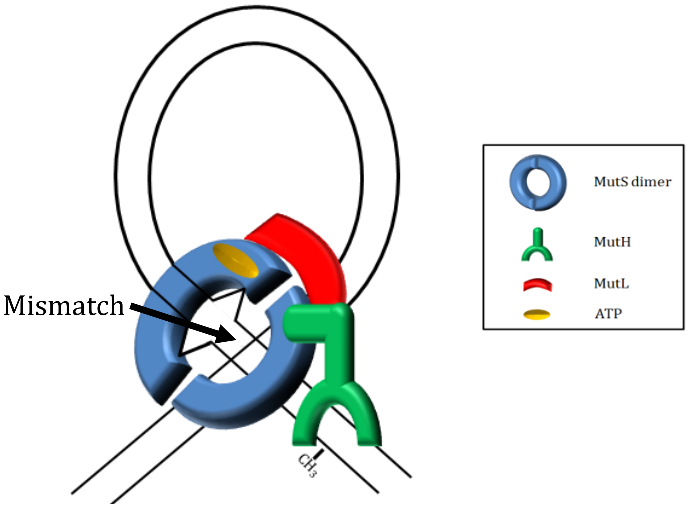


Figure 1.6. The active translocation model. MutS binds a mismatch and translocates along the DNA in an ATP hydrolysis dependent manner. MutS interacts with other downstream proteins (MutL, MutH *etc.*) by forming DNA loop as the DNA is spooled through MutS, which is still bound with the mismatch.

3. In a third model, MutS is proposed to bind the mismatch while other downstream component(s) (especially MutL) polymerises along the DNA contour up to the strand discrimination signal (Figure 1.7). In principle, the polymerisation event can be either dependent or independent of ATP hydrolysis. In support of this model, a study using atomic force microscopy has found that the MLH1-PMS1 binds to duplex DNA in a cooperative manner (Drotschmann et al., 2002). However, this model is receiving little interest in recent literatures.

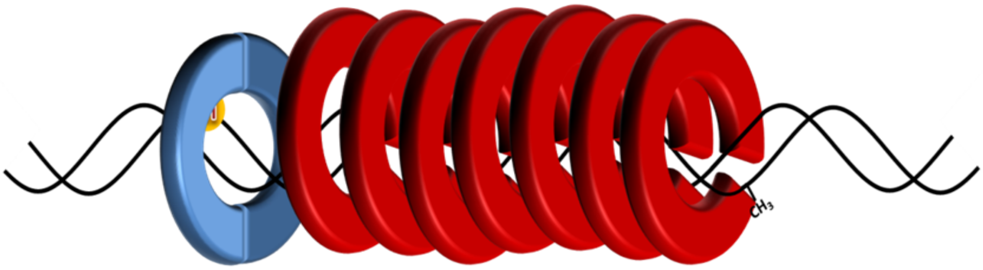


Figure 1.7. Cooperative binding of MutL proteins along DNA contour. MutS dimer (denoted by two blue half-circles) binds a mismatch on the DNA. Upon binding to MutS, MutL (depicted in red) also binds to the DNA contour in a cooperative manner in search of the strand discrimination signal where it can interact with other downstream proteins (not shown here) of the MMR pathway.

4. A different model for mismatch binding and subsequent signalling to later reaction has been proposed, which is also known as the "static transactivation model". This model has been postulated mainly from studies of bacterial MMR proteins. In this model, MutL stabilises the MutS dimer at the mismatch and this static complex interacts in *trans* with other downstream proteins at different sites to initiate the excision and repair synthesis reactions (Selmane et al., 2003).

Early *in vitro* studies showed some pioneering findings about the formation of a ternary complex (MutS-ATP-MutL bound to a mismatch containing DNA) and its interaction with downstream proteins of MMR, while some results made the whole scenario a bit puzzling. Contrary to the second model, the MutS-MutL complex has been found to translocate along the DNA substrate in either way *in vivo,* which spares the need of loop formation (Smith et al., 2001). In addition, the human MMR proteins - MutSα and MutLα can form relatively stable ATP-dependent ternary complex on free-ended oligonucleotide substrates (Blackwell et al., 2001; Plotz, 2002; Räschle et al., 2002). This finding supports for a free translocating ternary complex along a mismatch containing DNA contour. Moreover, surface plasmon resonance studies have showed that the human and yeast MutSα-MutLα complex can slide along the DNA contour like the MutSα sliding clamp (Blackwell et al., 2001; Mendillo et al., 2005). On the contrary, the finding of a successful mismatch-provoked degradation of a nicked strand that contains a physical obstacle between the mismatch and the incision-initiation site argues in favour of the presence of a looped DNA during MMR while the MutS-MutL complex remains on the mismatch (Wang and Hays, 2003, 2004). In addition, MutS dependent loop formation of DNA substrate in an *in vitro* experiment favours the second model (Allen et al., 1997). However, previous studies on the MutS-MutL-DNA ternary complex used gel shift assay and only observed an efficient ternary complex in the presence of ATP and Mg^{2+} (Blackwell et al., 2001; Bowers et al., 2000, 2001; Kijas et al., 2003; Räschle et al., 2002). Some of these studies did not verify the requirement of a mismatch in the formation of the ternary complex (Bowers et al., 2000, 2001; Räschle et al., 2002).

Plotz and collaborators found that the ternary complex formation is independent of the mismatch, while another group found that a MutS protein incapable of hydrolysing ATP cannot form the ternary complex (Kijas et al., 2003; Plotz, 2002). Similar results were found for MSH2-MSH6 and MLH1-PMS1 forming a ternary complex in the presence of ATP and Mg^{2+} (Mendillo et al., 2005). As this ternary complex was found to bind both homo- and heteroduplex DNA with equal efficiency, it has been proposed that the binding was due to end binding of a linear substrate (Blackwell et al., 2001). However, a real time binding experiment did not find any end binding effect for the ternary complex (Blackwell et al., 2001). When MLH1-PMS1 was incubated with MSH2-MSH6 in the presence of ATPγS, less ternary complex formation was observed on the mismatched substrate (Gradia, 2000; Mendillo et al., 2005). The reduced rate of association in the presence of ATPγS has led to the suggestion that the hydrolysis of previously bound ATP to the free sliding clamp forms an ADP-bound MutS homologue which in turn binds to the mismatch. In addition, MLH1-PMS1 heterodimer does not bind to either homo- or hetero-duplex DNA in the absence of MSH2-MSH6 complex (Mendillo et al., 2005). So, it has been postulated that the ADP-bound MSH2-MSH6 binds the mismatch first and the MLH1-PMS1 heterodimer comes later to form a ternary complex. A study with MutS that has been preloaded onto the mispaired substrate has shown that the ternary complex is formed with the addition of MutL and ATP (Selmane et al., 2003). In addition, an ATPase dead MutS protein bound to a mispair allowed MutL to interact with it after addition of ATP (Mendillo et al., 2005). Therefore, ATP binding, not hydrolysis, by MutS is needed for the formation of ternary complex with MutL.

Therefore MLH1-PMS1 can assemble onto MSH2-MSH6 at a mispair to form a ternary complex under either of the conditions:

1. When MutS and MutL proteins are incubated together with the substrate (a mismatch containing DNA) and ATP, or

2. When MSH2-MSH6 is preloaded onto the mispair in the presence of ADP and MLH1-PMS1 is then added along with ATP or ATPγS.

1.4.5 Strand discrimination and start of incision-excision reaction: the role of MutH

The next step of the MMR pathway is to distinguish the nascent strand that is harbouring a mismatch and this task is accomplished by searching and identifying a genomic signal along the DNA contour. This signal appears different in prokaryotes and eukaryotes and, as a result, so is the signal detection mechanism. Most of the information regarding this mechanism in prokaryotes has come from the studies on the *E. coli* MMR system (Lahue et al., 1987). The signal in the *E. coli* MMR is a hemimethylated GATC motif in the chromosome in which the adenine is methylated at the N^6 position on one strand (just after the replisome has passed, it is the parent strand) whereas the other strand remains unmethylated (the nascent strand under same condition). In *E. coli*, this hemimethylated GATC is recognised and incised on the nascent strand by the cryptic endonuclease, MutH (Lee et al., 2005). The MutL protein is believed to mediate the interaction between the mismatch-activated MutS and the MutH endonuclease. Interestingly, MutH interacts with the carboxyl terminus of MLH1 (the domain that houses the endonuclease in other MLHs), which makes it possible that the overall geometry of the strand incision complex is similar in all organisms (Au et al., 1992). In principle, the GATC site should be within 2kb from the mismatch for being detected by the MMR system in *E. coli* (Bruni et al., 1988; Lahue et al., 1987).

The MutH endonuclease belongs to a protein family containing several restriction enzymes carrying a conserved PD-D/E(X)K motif (Kinch et al., 2005). The purified MutH protein has a molecular weight of 25000. MutH has an extremely weak endonuclease activity (less than 1 scission/h/MutH monomer equivalent) that can cleave the nascent DNA strand 5' to the GATC. A symmetrically methylated GATC is resistant to the endonuclease activity of MutH. On the other hand, a hemimethylated GATC gets incised on the unmethylated GATC strand and DNA with completely non-methylated d(GATC) is incised on one of the DNA strands (Welsh et al., 1987). Though, MutH endonuclease activity was found independent of a mismatch *in vitro*, it is not the case *in vivo* (Au et al., 1992; Welsh et al., 1987).

However, the total scenario in eukaryotes is different as there is no homologue of MutH found to date. Therefore the way that the MMR system in eukaryotes differentiates the nascent strand to perform the incision-excision reaction repairing a mismatch remained a mystery. In 1986, Claverys and Lacks hypothesised that MMR in most organisms is directed by strand discontinuities (Claverys and Lacks, 1986). However, this hypothesis only supports mismatch repair in the lagging strand. Evidence in support of this hypothesis came from different *in vitro* studies on DNA heteroduplex containing a nick near the mismatch (Holmes et al., 1990; Thomas et al., 1991). In addition, *in vivo* studies in *S. cerevisiae* showed that the lagging strand was more efficiently repaired by MMR possibly due to the availability of termini of each Okazaki fragment for the excision repair factors (helicase and exonucleases) to load on to the DNA (Nick McElhinny et al., 2010a; Pavlov et al., 2003). Still, all these evidences fail to explain postreplicative MMR on the leading strand as it is devoid of free DNA terminus within the vicinity of a mismatch. However, reconstituting human MMR system by administering MutSα, MutLα, EXO1 (a unidirectional, 3’ → 5’, exonuclease), PCNA (Proliferative Cell Nuclear Antigen, a homologue of prokaryotic β clamp), RFC (the clamp loader) and RPA (single-strand binding protein) resulted in successful bidirectional excision reaction of mismatch containing DNA strand (Dzantiev et al., 2004; Genschel et al., 2002). Close analysis of this result revealed that the MutLα harbours a cryptic endonuclease activity which gets activated while interacting with MutSα, a mismatch containing heteroduplex DNA and PCNA (Kadyrov et al., 2006, 2007).

However, the evidence mentioned above does not explain the strand discrimination capability which is a crucial factor in MMR. This puzzle might be solved by an indirect hypothesis which requires spatial association of different proteins with the DNA substrate. It was found that MutLα is activated when associated with PCNA (Pluciennik et al., 2010). PCNA is loaded at the double- and single-strand boundary facing the terminus and provides MutSα-MutLα-DNA ternary complex with a fixed geometry while interacting with it on a DNA substrate (McNally et al., 2010). In this fashion, the ternary complex will be able to slide along the DNA contour without flipping round. As the endonuclease activity is encoded in PMS2 of the MutLα, only one strand is thought to be hydrolysed at the phosphodiester backbone, with correct orientation (Kadyrov et al., 2006).

Perhaps the most important finding regarding strand discrimination by eukaryotes has come from two recent studies that the incorporation of different ribonucleotides (as errors) during replication could act as a genomic marker akin to the hemimethylated GATC motifs in *E. coli* (Ghodgaonkar et al., 2013; Lujan et al., 2013). More than a million ribonucleotide monophosphates (rNMPs) are found to be incorporated during replication of mouse chromosomes and a similar scenario is reported for *Saccharomyces cerevisiae* (Hiller et al., 2012; Reijns et al., 2012). However, the incorporation of such ribonucleotides differs from the leading strand to lagging strand while four times more incorporation is observed in the leading strand (1 rNMP/1,250 dNMPs in leading strand versus 1/5,000 in lagging strand) (Nick McElhinny et al., 2010a, 2010b). In addition, using strand-specific probes of DNA fragments generated by alkaline hydrolysis, Lujan and collaborators have demonstrated that ribonucleotides are preferentially incorporated into the nascent leading strand in *S. pombe* (Lujan et al., 2012). In agreement with that, Lujan and collaborators have shown that the leading strand polymerase (Pol-ε) conserves a leucine (Leu612) instead of methionine that renders it more prone to incorporate ribonucleotides during replication (Lujan et al., 2013). These ribonucleotides are present only transiently as they are efficiently repaired by the ribonucleotide excision repair (RER) system which is initiated by the nick created in the DNA backbone by the RNase H2 and then removed by the flap endonuclease FEN1 and/or EXO1 (Eder and Walder, 1991; Eder et al., 1993; Nick McElhinny et al., 2010a; Rydberg and Game, 2002; Sparks et al., 2012). Ghodgaonkar and collaborators have confirmed the participation of RNase H2 in

DNA mismatch repair system in an *in vitro* experiment using human and mouse (cell free) nuclear extract and a heteroduplex substrate containing a single rCMP residue (Ghodgaonkar et al., 2013). They have also shown that knocking down of RNase H2 lowers the efficiency of MMR on the nascent leading strand (Ghodgaonkar et al., 2013). On the other hand, using an RNase H2 mutant and an impaired leading strand polymerase (pol ε) strain that incorporates 16-fold more ribonucleotides into DNA during replication than in wild type cells, Lujan and collaborators have found reduced *in vivo* MMR on nascent leading strand (Lujan et al., 2013). They have proposed that the misincorporated ribonucleotides on the leading strand might be the signals for strand discrimination as RNase H2 might make a nick at the ribonucleotide (mimicking the MutH in *E. coli*), thereby creating an entry point for EXO 1 to excise the mismatch containing nascent strand or by DNA synthesis to displace the mismatch containing nascent strand followed by cleavage of the 5’ DNA flap (Kadyrov et al., 2006; Lujan et al., 2013). However, Ghodgaonkar and collaborators have proposed that the contribution of this pathway to the total fidelity of the MMR could be small as the distance between a misincorporated ribonucleotide and a mismatch has to be less than 1 kb (Ghodgaonkar et al., 2013).

1.4.6 Unwinding of nascent strand by UvrD helicase

Once the MMR system differentiates the nascent DNA strand from its parent strand (and MutH creates a nick at a hemimethylated GATC motif in case of *E. coli*), UvrD helicase is loaded on the DNA contour to unwind the mismatch containing nascent DNA strand. The UvrD helicase, showing a modest processivity, usually unwinds DNA duplex of 40-50 bp (Ali and Lohman, 1997; Dessinges et al., 2004; Maluf, 2003). However, the processivity of UvrD as a translocase is significantly higher: 2400±600 bp which is still longer than the usual length of the DNA to be unwound in a MMR reaction: 1-2 kb (Fischer et al., 2004). In an *in vitro* experiment, a circular DNA heteroduplex underwent UvrD mediated helicase activity from a nick towards the mismatch in the presence of MutS and MutL (Yamaguchi et al., 1998). For some reasons yet to be understood, MutL modulates the UvrD mediated unwinding of DNA (Matson and Robertson, 2006). The first indication that MutL acts to load UvrD onto DNA came from DNA binding studies showing that the addition of MutL increases the affinity of UvrD for DNA (Matson and Robertson, 2006). Electrophoretic gel mobility shift experiments have revealed that UvrD, in the presence of AMP-PNP, forms a weak complex with ssDNA that dissociates in the course of electrophoresis and is difficult to detect. However, a supershifted MutL–UvrD–ssDNA complex is formed that is more stable than the UvrD–ssDNA complex indicating that MutL and UvrD formed a specific complex with a greater affinity for ssDNA than UvrD alone (Mechanic et al., 2000). It has been proposed that MutL loads UvrD onto the DNA in an iterative process but does not form a clamp on the DNA during the unwinding reaction as the processivity of UvrD as a helicase is limited (Mechanic et al., 2000).

The increased activity of UvrD in the presence of MutL could be mediated by either or both of the following possibilities:

1. MutL might increase the rate of UvrD association with the DNA.
2. MutL might decrease the rate of UvrD dissociation from ssDNA.

In support of the first possibility, an increased association of UvrD with a 20 bp partially duplex DNA substrate was found in the presence of MutL (Matson and Robertson, 2006). On the other hand, the second possibility would be true if the MutL functioned as a clamp to increase the processivity of UvrD – which is not the case according to the following experiments. Firstly, in the presence of MutL, no increased unwinding of DNA by UvrD has been observed at a 148 bp blunt duplex DNA substrate using a ssDNA trap (Matson and Robertson, 2006). Secondly, in a single turn-over assays with a 92 bp partial duplex DNA, MutL is not found to alter the processivity of UvrD compared to a 20 bp partial duplex DNA (Matson and Robertson, 2006). Under identical conditions, a smaller fraction of 92 bp partial duplex molecules have been unwound in comparison to 20 bp molecules (Matson and Robertson, 2006). If MutL were acting to increase the processivity of UvrD, the same fraction of substrate should have been unwound in each case. On the other hand, an increased loading of UvrD by MutL has also been investigated using long DNA substrates to find out the *in vivo* processivity of UvrD (Matson and Robertson, 2006).

Several turn-over helicase reactions with a 750 bp blunt duplex DNA substrate and an 851 bp partial duplex DNA substrate support the notion that MutL loads UvrD onto DNA in an iterative process (Matson and Robertson, 2006). Unwinding of the two long substrates was generally described by a burst phase (unwinding of ssDNA by preloaded UvrD) followed by a steady-state phase (by the newly recruited UvrD molecules). In the absence of MutL, no burst phase of unwinding has been found using a 750 bp blunt duplex DNA substrate (Ali and Lohman, 1997). However, based on the reported processivity for UvrD (40–50 bp), unwinding of a longer substrate and detection of the significant burst phase requires multiple binding events by UvrD. A model has been proposed to explain the MutL stimulated DNA unwinding of UvrD (Figure 1.8). Here, the first step would be to load UvrD onto the DNA. This loading is enhanced as the affinity of the UvrD–MutL complex for DNA is increased (due to the presence of MutL). Once it is loaded, the UvrD begins to unwind the duplex and additional UvrD molecules are loaded behind the leading UvrD. Therefore, a high concentration of UvrD may increase the overall rate of UvrD-catalysed unwinding. Eventually, the leading molecule of UvrD dissociates from the DNA since UvrD translocates through duplex DNA with an average of ten steps (as the reported processivity of UvrD is 40–50 bp) before dissociating (Ali and Lohman, 1997). In case the of a UvrD alone (in an inefficient process), a partially unwound duplex DNA can re-anneal when the leading molecule dissociates from it and the whole process must start over (Figure 1.8A). On the other hand, in the presence of MutL, multiple UvrD molecules are loaded onto the DNA duplex and dissociation of a single UvrD molecule does not result in re-annealing of partially unwound DNA duplex. In that case additional UvrD molecules translocate along the ssDNA template and continue unwinding the duplex DNA (Figure 1.8B). This is consistent with the observation that UvrD is considerably more processive as a translocase moving along ssDNA than as a DNA helicase (Fischer et al., 2004).

It was proposed that the unwinding reaction of UvrD might be fuelled by MutL mediated ATP hydrolysis (Robertson et al., 2006). However, using a MutL point mutant (MutL-E29A) that binds, but does not hydrolyse ATP, Robertson and collaborators have demonstrated that MutL-catalysed ATP hydrolysis is not required for MutL-dependent stimulation of the UvrD unwinding reaction (Robertson et al., 2006). Importantly, it is the ATP-bound form of MutL that is specifically responsible for stimulating UvrD as a second MutL point mutant (MutL-D58A), which does not bind ATP, does not stimulate the unwinding reaction catalysed by UvrD (Robertson et al., 2006).

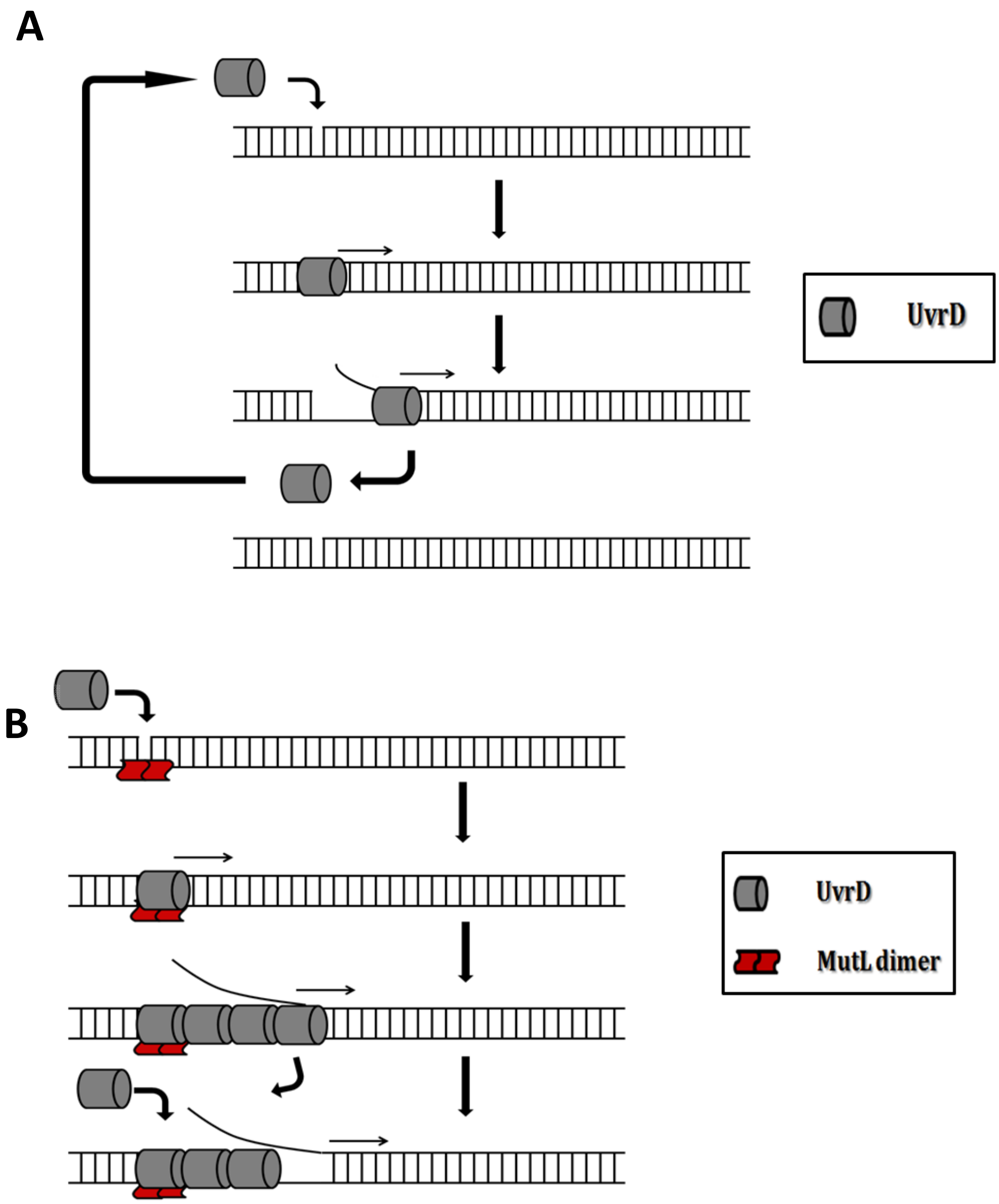


Figure 1.8. A model for the MutL assisted UvrD loading, translocation and unwinding of ssDNA in MMR pathway. A) Loading of a single UvrD molecule results in an inefficient process as the partially unwound duplex DNA can re-anneal when the UvrD molecule dissociates from it. B) In the presence of MutL, multiple UvrD molecules load onto the DNA duplex. Therefore, dissociation of a single UvrD molecule does not result in re-annealing of a partially unwound DNA duplex. Additional UvrD molecules translocate along the ssDNA template and continue unwinding the duplex DNA.

1.4.7 The excision reaction: the role of different exonucleases

The next step of the MMR pathway is to remove the error containing segment of nascent strand and to ensure a second chance to synthesise the correct strand of DNA. To date, at least four single-strand exonucleases, namely ExoI, ExoVII, RecJ and ExoX, have been identified to be involved in this excision reaction in *E. coli* (Burdett et al., 2001; Kunkel and Erie, 2005). Unlike RecJ, ExoI and ExoVII, which are processive exonucleases, ExoX is a distributive exonuclease that hydrolyses only one or a few nucleotides before releasing its substrate (Viswanathan et al., 2001). The *in vitro* MMR excision reaction is bi-directional as a GATC site can be situated on the 5’ or 3’ side of the mismatch and this arsenal of exonucleases can cleave either in 5’→3’ or 3’→5’ direction (Cooper et al., 1993; Grilley et al., 1993). Both RecJ and ExoVII have 5’→3’exonuclease activity and both have previously been shown to be functionally redundant for the repair of a 5’ heteroduplex where a hemi-methylated GATC motif resides on the 5’ side of the mismatch (Chase and Richardson, 1974; Cooper et al., 1993; Lovett and Kolodner, 1989). When either or both of ExoI and ExoX are mutated, no change in 3′→5′ hydrolytic activity has been found for repairing a 5′ G-T heteroduplex (Viswanathan et al., 2001). ExoI deficiency reduces correction of a 3′ heteroduplex (where a hemi-methylated GATC motif resides on the 3’ side of the mismatch) by 50%, and deficiency of both ExoI and ExoX reduces mismatch repair by 63% (Viswanathan et al., 2001). This residual 3′ repair activity in ExoI$^-$ExoX$^-$ background is attributable to the 3′→5′ activity of ExoVII as ExoI$^-$ ExoVII$^-$ ExoX$^-$ triple mutant background abolishes all repair activity on a 3′ heteroduplex (Chase and Richardson, 1974; Viswanathan et al., 2001). These findings indicate that either ExoVII or RecJ is sufficient to meet the exonuclease requirement for repairing 5′ heteroduplexes in *E. coli* extracts, whereas a single exonuclease among ExoI, ExoX and ExoVII can contribute in a redundant fashion to the repair of the 3′ heteroduplexes.

recJ and *xseA* (ExoVII large subunit) orthologues can be found in most of the bacterial genomes sequenced to date, while other exonucleases in MMR are restricted to few bacteria (Viswanathan et al., 2001). Thus a different collection of exonucleases is likely to be functional for the MMR excision reaction in different species (Viswanathan et al., 2001). The functional redundancy of exonucleases in *E. coli* may be due to their specialised roles in other DNA metabolic pathways. For instance, RecJ mediates recombination via the RecF pathway and may process stalled replication forks (Courcelle and Hanawalt, 1999; Lovett and Clark, 1984). RecJ and ExoI are involved in recombination via the RecBCD pathway (Friedman-Ohana and Cohen, 1998; Miesel and Roth, 1996; Razavy et al., 1996; Viswanathan and Lovett, 1998). ExoI and ExoVII play important roles in avoiding misalignment errors during replication, such as frameshifts and quasi-palindrome templated mutations (Viswanathan and Lovett, 1998; Viswanathan et al., 2000).

The components of the excision reaction in eukaryotes are not similar to those of prokaryotes. In yeast, genetic studies have implicated at least four nucleases in MMR – Exonuclease 1, Rad27, the intrinsic 3’ exonuclease activity of Polδ and Polε. The identification of the exact function of these MMR related yeast nucleases must await further study because each of these four nucleases also contributes to genome stability by participating in other DNA metabolisms, such as proofreading, flap removal during processing of Okazaki fragments, and cell cycle checkpoint control (Datta et al., 2000; Tishkoff et al., 1997). It has been observed that *EXO1*$^{-/-}$mouse cells have a lower mutation rate than that of *MSH*$^{-/-}$ cells, which is thought to be an indication of the existence of other nuclease(s) participating in MMR excision reaction in mouse (Wei et al., 2003). However, in a different *in vitro* study, extracts from EXO1 deficient mouse cells were found defective in MMR activity on either 5' or 3' nick containing heteroduplexes (Wei et al., 2003). Surprisingly, purified human EXO1 has been found to initiate excision at either a 5' or a 3' nick, though it has a 5' → 3' polarity in cleaving DNA (Genschel et al., 2002). To explain this result, it has been proposed that the EXO1 might have a cryptic 3' exonuclease activity or it might take part in assembling a repair complex (Amin et al., 2001; Genschel et al., 2002). In support of the former possibility, the catalytic Asp173 of the EXO1 was replaced with Ala and the resulting purified protein showed a defect in both 5' and 3' excision reactions (Dzantiev et al., 2004).

Extracts from different mutant backgrounds and purified proteins have revealed the participation of other MMR proteins in the MMR excision reaction. In a reconstituted reaction with purified proteins, MutSα activates and confers high processivity to EXO1 in a 5' mismatch excision (Genschel and Modrich, 2003). Once the mismatch is excised, termination of the excision reaction is conferred by suppressing EXO1 activity via MutS*α* and MutL*α,* and displaced by RPA proteins (Genschel and Modrich, 2003). PCNA is found to be essential for the 3' but not for the 5' excision reaction in studies using cell extracts. In a purified system, the 5' excision reaction requires only MutSα, EXO1 and RPA, while the 3' excision reaction also requires MutLα, PCNA and RFC (Dzantiev et al., 2004; Fang and Modrich, 1993; Genschel and Modrich, 2003; Genschel et al., 2002; Guo et al., 2004). RFC not only loads PCNA onto DNA, it is required to suppress nonproductive 5' excision away from the mismatch, an effect that depends on the integrity of the ligase homology domain of the largest subunit of the five-protein RFC complex (Dzantiev et al., 2004). This suppression may be due to the ability of this ligase homology domain to bind to recessed 5' phosphoryl termini. Because PCNA is loaded onto primer templates in an orientation-dependent manner, the different protein requirements for 5' and 3' excision and the ability of PCNA to interact with MutSα, MutLα and EXO1 have led to a model, wherein "an orientation-dependent encounter of PCNA at a strand discontinuity by the mobile MutSα-MutLα complex results in differential hydrolytic responses according to the 3' or 5' placement of the discontinuity" (Dzantiev et al., 2004; Tsurimoto, 1999).

1.4.8 Repair synthesis of the excised single stranded DNA

Once the mismatch containing nascent DNA strand has been unwound by the helicase II (UvrD) and excised by exonuclease(s), the single-strand DNA-binding proteins (Ssb) rapidly bind the single stranded parent DNA to prevent nuclease attack (Ramilo et al., 2002). In prokaryotes, the repair synthesis of the correct nascent strand is mediated by the DNA polymerase III holoenzyme (Lahue et al., 1989). The involvement of other DNA polymerases, like DNA polymerase I, T7 DNA polymerase, T4 DNA polymerase and AMV reverse transcriptase has been ruled out in several experiments using purified systems (Lahue et al., 1989; Modrich and Lahue, 1996). On the other hand, in eukaryotes, the repair synthesis is catalysed by an aphidocolin-sensitive polymerase, likely DNA polymerase δ with the assistance of RFC (Replication Factor C, also known as the clamp loader) and PCNA (homolog of β clamp) (Longley et al., 1997; Wood and Shivji, 1997). Once repair synthesis is completed, DNA ligase seals the remaining nick to restore the covalent continuity of the repaired nascent strand (Burdett et al., 2001; Lahue et al., 1989).

1.5 The mismatch repair system in hereditary nonpolyposis colorectal cancer

HNPCC (**H**ereditary **n**on**p**olyposis **c**olorectal **c**ancer), also known as Lynch syndrome II or Muir-Torre syndrome, is one of the most common cancer predisposition diseases in human. In the USA, the total annual incidence of colon cancer is about 150000, of which a small percentage (10-15%) of patients developes colon cancer due to inherited mutations. In addition, 5-6% of the total patients falls under a category of life time risk of having this cancer (Lynch and de la Chapelle, 2003). Usually, HNPCC kindreds are defined as those in which at least three relatives in two generations have colorectal cancer, with one of the relatives having been diagnosed at younger than 50 years of age. A striking feature of HNPCC is that it falls under the same cancer category of having instabilities of simple repeated sequences - microsatellite instability (Fishel et al., 1993; Leach et al., 1993; Parsons et al., 1993). In addition, studies on MMR in *E. coli* and *S. cerevisiae* has led to the prediction that the phenotype observed in HNPCC tumours might be caused by a mutation in any gene involved in the MMR pathway (Aaltonen et al., 1993; Ionov et al., 1993; Peltomäki et al., 1993; Thibodeau et al., 1993). Genetic linkage analysis on two large HNPCC kindreds mapped a microsatellite polymorphism marker on chromosome 2p, which had a C→T transition in one copy of the *MSH2* gene on the same chromosome (Fishel et al., 1993; Leach et al., 1993; Peltomäki et al., 1993). For one kindred, this transition changed a highly conserved amino acid Pro to a Leu and in another, it caused a nonsense mutation (Leach et al., 1993). A second HNPCC locus was mapped on chromosome 3p21, which was linked to a mutation in the *MLH1* gene and found in several families with colorectal cancer. Later, mutations in the *PMS1* gene (on chromosome 2) and *PMS2* (on chromosome 7) were found in the germ line in a HNPCC patient (Bronner et al., 1994; Lindblom et al., 1993; Nyström-Lahti et al., 1994; Papadopoulos et al., 1994). A germ line mutation in the *MSH6* gene was identified in a Japanese patient with an atypical clinical presentation of HNPCC with a later onset of the disease (Akiyama et al., 1997; Berends et al., 2002; Miyaki et al., 1997; Wijnen et al., 1999).

Later, mutations in MLH3 and EXO1 genes had been found to be associated with HNPCC (Wu et al., 2001a, 2001b). Another interesting finding came with the analysis of a colorectal cancer cell line, termed H6 (Parsons et al., 1993). These H6 cells are defective in strand specificity during MMR repair of all eight possible mismatches and in the repair of 2-, 3- and 4-nucleotide insertion-deletion mismatches (Papadopoulos et al., 1994). It had been found that the H6 cells were devoid of a wild-type MLH1 allele (Papadopoulos et al., 1994). It is remarkable that the HNPCC individuals are heterozygous for these mutations, which means that they have a mutated version of these genes along with a wild-type copy. This heterozygosity prevails up to the point of preneoplastic tissue converting into tumour tissue (Lynch et al., 1993). So, how this heterozygosity is ultimately responsible for the cancer?

One reasonable model is that the cells of HNPCC carry out MMR with less efficiency, which might lead to mutation(s) in some genes involved in cell viability and cell growth – *i.e.*, mutation in genes causing uncontrolled cell growth or genes with repeated sequences in the exon regions. Therefore, the MMR might become more inefficient that gives rise to more mutations. A mononucleotide tract in the *TGFBR2* gene, which regulates the growth of the colon epithelial cells, gets mutated in 90% of the HNPCC colon tumours and plays a stage-specific role in multistep colon tumorigeneses (Grady et al., 1998; Jacob and Praz, 2002; Lynch and de la Chapelle, 2003; Wang et al., 1999).

Another example of the association of gene-specific nucleotide repeat with HNPCC is the sequence GCA-GAA-ATA-AAA-GAA in the coding region of the APC gene, which gets mutated to GCA-GAA-AAA-AAA-GAA in most of the individuals having familial adenomatous polyposis. This mutation does not affect the normal function of the APC gene (tumor supression), but creates a template for greater instability (Aoki and Taketo, 2007; Laken et al., 1997). On the other hand, some of the MMR genes, like *MSH2*, *MSH3* and *MSH6* themselves have repeat tracts in their coding regions that can be mutated in the cells with initial inherited subtle defects in MMR (Duval et al., 2001; Guerrette et al., 1998; Jacob and Praz, 2002; Lynch and de la Chapelle, 2003; Ohmiya et al., 2001; Yin et al., 1997). Consequently, acquiring further mutation(s) in the copy of the gene that was previously wild-type results in MMR double-mutant cells could render a dramatic effect in the tumour progression and genomic stability.

1.6 The mismatch repair system in other DNA metabolic pathways

1.6.1 The mismatch repair system in somatic hypermutation and class switch recombination

The common notion of MMR proteins is that they are the repair proteins involved in repairing replication errors. Contrary to this idea, MMR proteins increase diversity to the antibody repertoires by actively promoting mutations. As a means of protection from infections and foreign substances, vertebrates have evolved adaptive humoral immunity, which provides antibodies that circulate throughout the body and into secretions. These antibodies bind strongly and specifically to the invading foreign bodies to dispose of them in different ways (Maizels, 2005). The genomes of vertebrates lack sufficient capacity to accommodate all of the necessary immunoglobulin (Ig) gene variants. Amazingly, a small amount of genetic material can generate a highly diverse antibody repertoire that is sufficient to deal with all possible antigens (Diaz and Flajnik, 1998). Human pre- and pro-B cells in the bone marrow constantly produce a highly diverse repertoire of antigen-binding sites by forming heavy (H) and light (L) chain in the immunoglobulins (Ig) through rearrangements of germ line Ig variable (V), diversity (D) and joining (J) elements even prior to antigen (Ag) exposure (Li et al., 2004a; Longerich et al., 2006; Maizels, 2005; Di Noia and Neuberger, 2007). Upon exposure to an Ag, the mature B cells become stimulated to proliferate, differentiate and migrate to the dark zone of the germinal centres (GC) in the secondary lymphoid organs where they become centroblasts and express a large amount of activation-induced cytidine deaminase (AID) (Kelsoe, 1996; MacLennan, 2005). AID initiates somatic hypermutation (SHM) of the antibody V regions that encode the specific antigen-binding sites (Muramatsu et al., 1999, 2000) and class switch recombination (CSR) – unique type of intrachromosomal deletion recombination within a special G-rich tandem repeated DNA sequence (the S region) (Stavnezer et al., 2008). These mutations and recombinations result in the amino acid replacements in the H and L chains that are responsible for the affinity maturation and fine tuning of antibody specificity. In addition to that, switching the Ig isotype IgM to IgG, IgE or IgA are also aided by this process (Casali et al., 2006; Li et al., 2004a; Longerich et al., 2006; Maizels, 2005; Di Noia and Neuberger, 2007; Teng and Papavasiliou, 2007). MMR proteins play an essential role in these processes. The AID protein, a homologue of an RNA editing enzyme, deaminates dCs to dUs sequentially in the ssDNA in the transcription bubble of the V and S regions of Ig genes (Bransteitter et al., 2003; Chaudhuri et al., 2003; Muramatsu et al., 2000; Pham et al., 2003; Storb, 1998). However, this deamination accounts for more than half of the mutations that are necessary for creating the variety of antibodies (Peled et al., 2008).

MSH2 and MSH6 defective cells demonstrated the involvement of these two proteins in generating mutations during SHM and CSR. Both MSH2 and MSH6 mutant mice showed about 5-fold lower level of mutations in V region during SHM and a decreased frequency of CSR than wild type mice (Li et al., 2004b;

Rada et al., 1998; Wiesendanger et al., 2000). On the other hand, MSH3 does not show any role in either SHM or CSR, which indicates that only MutSα takes part in those processes (Li et al., 2004b; Wiesendanger et al., 2000). In addition, MLH1, PMS2 and MLH3 (but not PMS1) are involved in SHM and CSR (Schrader et al., 2005; Wu et al., 2006). It is suggested that mutations during SHM accumulate in two steps (Figure 1.9). In the first step, a G:C base pair is mutated into a G:U by AID. In the second step, MutSα recognises the mismatch and interacts with MLH1 and PMS2 (Li et al., 2004a). The DNA harbouring the G:U mismatch is cleaved by an unknown endonuclease, then digested by EXO1 and repair synthesis is accomplished by error-prone polymerases (including translesional polymerase Polη) (Figure1.9) (Bardwell et al., 2004; Delbos et al., 2005; Martomo et al., 2005; Masuda et al., 2005; Pavlov et al., 2002; Rogozin et al., 2001; Zan et al., 2005; Zeng et al., 2001). This process introduces mutations with A/T bias which accounts for more than half of the total mutations in the V and J regions *in vivo* (Rada et al., 1998).

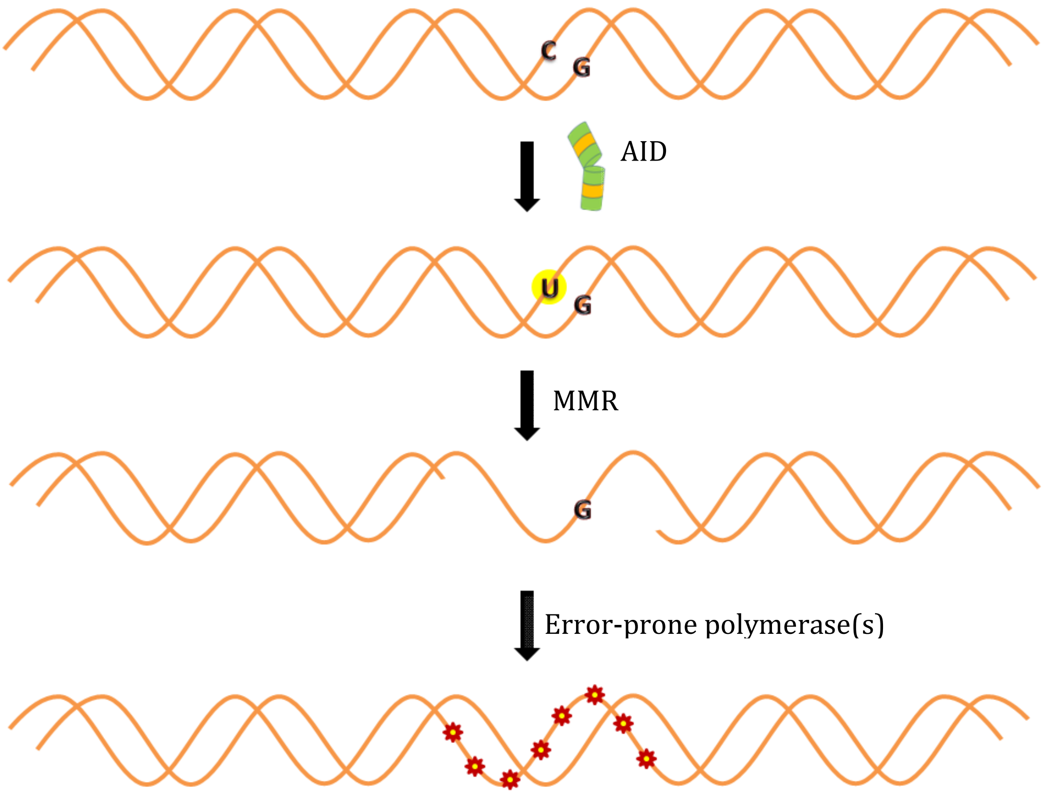


Figure 1.9. Involvement of MMR during somatic hypermutation. AID deaminates a cytidine residue and creates a uridine:guanosine (U:G) mismatch that is recognised by the Mismatch repair system . The U-bearing strand is excised and, at loci that undergo SHM, monoubiquitylated PCNA (proliferating cell nuclear antigen) recruits error-prone polymerases to fill the gap, leading to transition and transversion mutations (denoted by the red stars ✹). These mutations account for about 60% of the total mismatches during somatic hypermutation.

On the other hand, the exact role of MMR proteins in CSR is yet to be elaborated, even though their involvement in this process has long been found. It has been suggested that MutSα, MLH1 and PMS2 proteins are involved in DNA strand break after G:U mismatches or at least have a role in the processing of the breaks (Li et al., 2004b; Schrader et al., 2002).

1.6.2 Mismatch repair proteins in cell-cycle arrest and apoptosis

MMR proteins also take part in cell-cycle arrest and apoptosis. This involvement is evident in cells resistant against different chemotherapeutic agents which have already lost the MMR capability. It been found that bacterial and mammalian cells devoid of MMR reaction due to mutations in MMR proteins are about 100-fold resistant to alkylating agents, 10-fold tolerant of 6-thioguanine and also to methylating agents, cisplatin and UV radiation (Branch et al., 1993; Fink et al., 1996; Swann et al., 1996; Waters and Swann, 1997). Explaining these findings, two models have been proposed (Wang and Edelmann, 2006). The first model takes the general notion of MMR proteins as repair proteins where cells, with the MMR system, are engaged in futile cycles of repair removing mismatches repeatedly. In any case, two mismatches that occur in close proximity might initiate a double strand break as two MMR excision reactions coincide and as a result the cell to initiate apoptosis. Loss of MMR capability prevents this futile cycle at a cost of increased mutations in the presence of anticancer drugs and thus the cell stays alive. In the second model, which is called the "damage sensor" model, MutS and MutL homologues are proposed to initiate a cell-cycle delay and apoptosis, which are independent of their common ability as MMR proteins. Supporting this model, it has been found that to restore G2 arrest in the MLH1 deficient cell, almost 100% more expression of MLH1 protein from an inducible system relative to the wild type level is required. On the other hand, only 10% expression of MLH1 is sufficient for restoration of MMR reactions. Thus, it has been suggested that these two responses are independent (Cejka et al., 2003). In addition, MutS missense mutants were later identified retaining normal apoptotic capability to DNA damaging agents (Lin et al., 2004; Yang et al., 2004). According to any of these models, MMR protein(s) recruit(s) ATR (a kinase) to the damaged site and activates ATR. ATR has been found to interact with the human MutL homologue (Cannavo et al., 2007). Activation of ATR in turn phosphorylates Chk1 to activate cell-cycle check points and this downstream signalling eventually leads to cell apoptosis (Yoshioka et al., 2006).

1.6.3 Trinucleotide repeat instability and MMR

Repeat instability is a unique kind of mutation system which does not follow Mendelian genetics and has been found responsible for more than 40 neurological, neurodegenerative and neuromuscular diseases (Pearson et al., 2005). Unlike static mutations, this particular type of mutation is dynamic – that is, this mutation arises from continuous integrational expansion of shorter TNRs and it continues to mutate and pass through generations (Pearson et al., 2005). Multiple processes (individually or combined) are assumed to be involved for repeat instability that also shows complex patterns between and within tissues that vary with developmental, epigenetic, proliferative and possibly environmental cues (Pearson et al., 2005). Among different disease linked repeats, expandable trinucleotide repeats (TNRs) such as $(CGG)_n\cdot(CCG)_n$, $(CAG)_n\cdot(CTG)_n$, $(GAA)_n\cdot(TTC)_n$, $(GCN)_n\cdot(NGC)_n$ are the most common. Other repeats responsible for disease are tetranucleotides (in dystrophia myotonica type 2), pentanucleotides (in spinocerebellar ataxia type 10), minisatellites (in epilepsy, progressive myoclonic 1), megasatellites (in facioscapulohumeral muscular dystrophy 1A), *etc* (Pearson et al., 2005).

Studies on MMR mutant mouse models manifesting symptoms of Huntington's disease and myotonic dystrophy showed stable repeat length (van den Broek et al., 2002; Kovtun and McMurray, 2001; Manley et al., 1999; Savouret et al., 2003). Manley and collaborators have found that the expansion of (CAG)n repeats is reduced in all *msh2*$^{-/-}$ mouse tissues compared to *msh2*$^{+/+}$ tissues (Manley et al., 1999). In addition, Kovtun and collaborators have found that the absence of Msh2 protein in the progeny completely abolishes germline expansion and age-dependent somatic expansion (Kovtun and McMurray, 2001). Usually, MMR system recognises and repairs single mismatch repair, thus stabilises TNR array. However, in this case, it recognises lots of mismatches in different unusual structures formed by repeat tracts. A dimer of MutS proteins (in prokaryotes) or the MSH2-MSH3 complex (in eukaryotes) binds these mismatches and sequesters, as well as stabilises, the slipped-stranded structures without repairing them (Kovtun and McMurray, 2001). As MMR operates on nascent strands, expansion of repeat tracts is favoured due to sequestration and stabilisation of hairpin structures formed on the nascent strands. However, in prokaryotes, the reported natural propensity of such repeat tracts is contraction which is orientation dependent relative to the origin of replication (Jackson et al., 2014; Kang et al., 1995).

1.7 Studying DNA mismatch repair in *Escherichia coli*

1.7.1 Tri-nucleotide repeat (TNR) array as a source of substrate for the DNA mismatch repair (MMR) system

Mismatches arise mainly from errors during the DNA replication and can occur at any random locus in the genome. However, it is imperative that a mismatch occurs at a defined locus to be studied. In addition, the system should provide quantitative values regarding the efficiency of the MMR system so that certain molecular aspects of this system can be analysed. These perquisites have been attained in this study by using a 294 bp (98 unit) long CTG•CAG repeat array. It has been shown that the length of TNR array varies more frequently at a single unit level in MMR deficient cells than in MMR proficient wild type cells (Blackwood et al., 2010). When the replication machinery replicates a repeated array like a micro-satellite, it increases the chance either to add an extra unit of the repeat (Figure 1.10B) or to miss a repeat unit in the nascent strand (Figure 1.10C) (Blackwood et al., 2010). This phenomenon leads to having a change in the repeat length by single unit – either minus or plus respectively.

It has already been described that the MMR system not only recognises a single nucleotide mismatch in the genome, but also detects IDLs (3-4 base partnerless extrahelical nucleotides) with great efficiency (Jiricny et al., 1988b). Therefore, a TNR array can be used as a molecular tool to generate frequent substrates for the MMR system at a defined locus. In this study, the CTG•CAG repeated array used by Blackwood and collaborators has been utilised where the CTG repeats are on the leading strand (Blackwood et al., 2010). This trinucleotide repeat array is inserted in the *lacZ* gene in the *E. coli* genome. In addition, the variation of length of the repeat array at the level of a single unit has been defined simply as “instability” which has become a quantitative phenotype to detect the directionality and the efficiency of the DNA mismatch repair system.

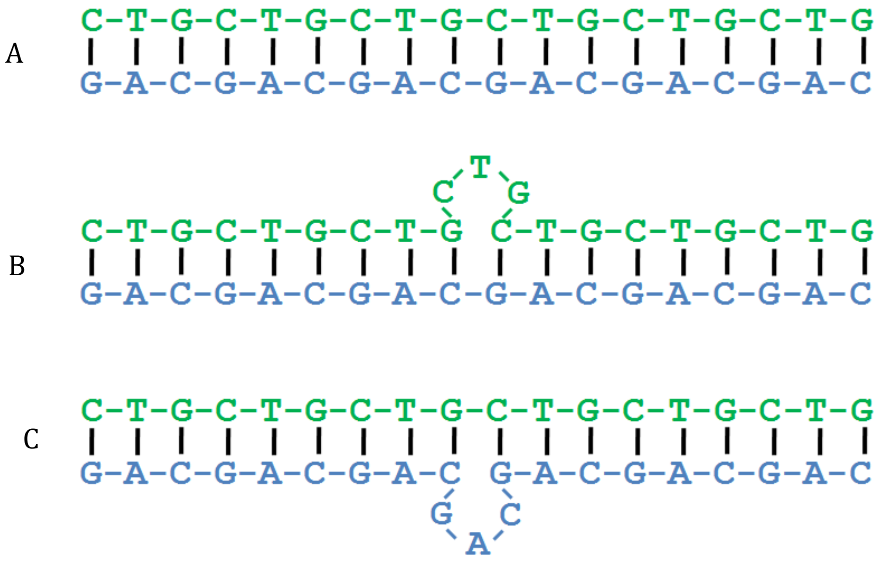

A
C-T-G-C-T-G-C-T-G-C-T-G-C-T-G-C-T-G
G-A-C-G-A-C-G-A-C-G-A-C-G-A-C-G-A-C
B
T
C G
C-T-G-C-T-G-C-T-G C-T-G-C-T-G-C-T-G
G-A-C-G-A-C-G-A-C-G-A-C-G-A-C-G-A-C
C
C-T-G-C-T-G-C-T-G-C-T-G-C-T-G-C-T-G
G-A-C-G-A-C-G-A-C G-A-C-G-A-C-G-A-C
G C
A

Figure 1.10. A model representing microsatellite instability at the level of a single repeat unit. This phenomenon is exemplified by using a CTG•CAG repeat array where CTG repeats are on the parent strand (green in colour) and CAG repeats are on the nascent strand (blue in colour) (A). During replication, occasionally, the replication machinery in an organism (either prokaryotic or eukaryotic) increases its tendency either to miss a single unit of the repeat which results in an extra-helical loop in the parent strand, therefore the nascent strand becomes shortened by a single unit of repeat (B). On the other hand, occasionally the replication machinery adds a single unit of repeat which results in an extra-helical loop with an extra unit of repeated sequence in the nascent strand (C).

1.7.2 A 275 bp tandem repeat as a tool to detect the excision reaction events during MMR

A sequential endeavour of MutS, MutL and MutH ranges from detection of a mismatch, distinguishing the error-containing DNA strand to making an incision at a hemi-methylated GATC on that strand (Jiricny, 2006). Then, an orchestrated activity of the DNA helicase II (UvrD) and the single strand dependent exonuclease(s) (one or more of ExoI, ExoVII, RecJ and ExoX) constitutes the excision reaction commencing from that incision and migrates towards the mismatch which is to be repaired. Grilley and collaborators have found in an *in vitro* experiment that the excision reaction terminates at different discrete sites within a 100 nucleotides region beyond a mismatch (Grilley et al., 1993). However, Blackwood and collaborators have shown that an effect of the DNA mismatch repair system can be detected at a distance of 6.3kb in *E. coli* suggesting the existence of longer excision tracts beyond the mismatch *in vivo* (Blackwood et al., 2010). Using a CTG•CAG trinucleotide repeat array, they have shown that the recombination of a 275 bp tandem repeat, which is placed on the origin proximal side of the TNR array, depends on the DNA mismatch repair system and the rate of recombination is a function of the distance between the TNR array and the tandem repeat (Figure 1.11). When the tandem repeat (also known as the Zeocin tandem repeat) is placed at a distance of 6.3kb from the TNR array on the origin proximal side, the rate of recombination becomes 2 fold higher than the wild type background which does not harbour the TNR array (Blackwood et al., 2010). However, the rate of recombination of the Zeocin tandem repeat increases 6 fold over the wild type background when the tandem repeat is placed at a distance of only 0.5kb from the TNR array on the origin proximal side (Figure 1.11).

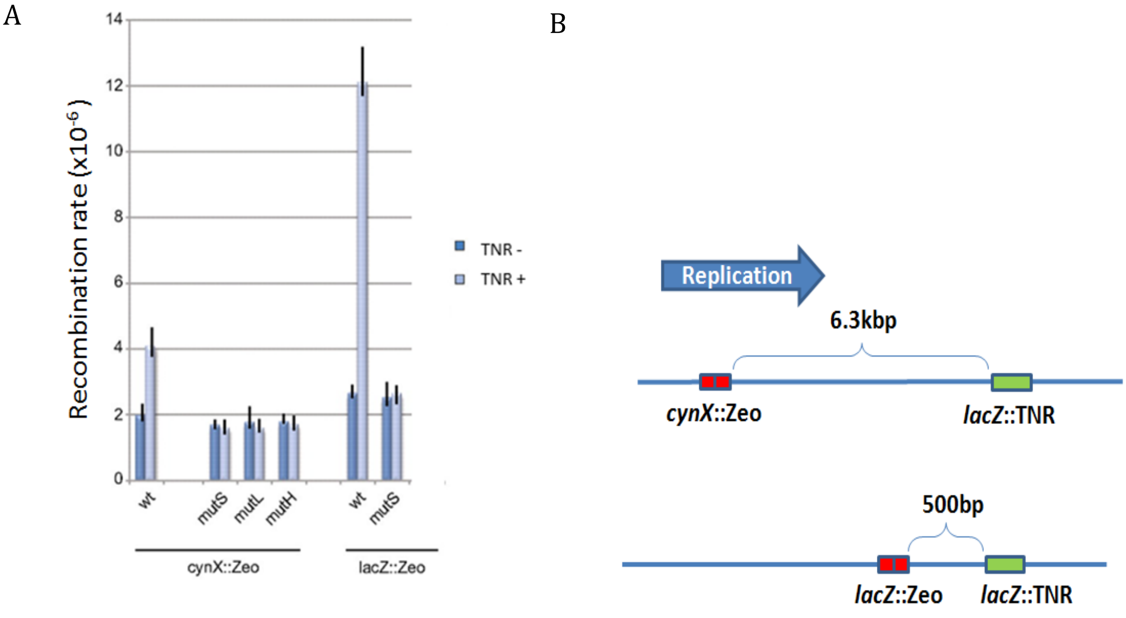


Figure 1.11. CTG•CAG TNR array stimulates the recombination at a 275 bp tandem repeat (Zeo) inserted on the origin proximal side of the TNR array. The recombination is dependent on functional DNA mismatch repair system as the mutant backgrounds show very low levels of recombination and so do the backgrounds without the TNR array (A). Moreover, the rate of recombination is a function of distance between the TNR array and the 275 bp tandem repeat. The rate of recombination at the 275 bp tandem repeat increases two fold compared to the wild type when the tandem repeat is inserted at 6.3kb away in the *cynX* gene on the origin proximal side of the TNR array (A & B). Surprisingly, the rate of recombination increases 12 fold compared to the wild type when the tandem repeat is inserted at only 500 bp away in the *lacZ* gene on the same side of the TNR array (A & B). Figure (A) collected from (Blackwood et al., 2010). In Figure (B), *E. coli* chromosome is shown in blue horizontal line, the TNR array is shown in green rectangle and the Zeocin tandem array is shown two tandem red boxes along with the distance from the TNR array.

In this study, the Zeocin tandem repeat, which is inserted on the origin proximal side of the TNR array, has been used to detect an excision reaction event in the course of DNA mismatch repair. The tandem repeat is composed of two defective ORFs of 275 bps (Figure 1.12). The first ORF has an intact initiation codon, but lacks the sequence for the C-terminus of the coded protein. On the other hand, the second ORF has an intact sequence except for the initiation codon. Upon recombination, which is mediated by replication slippage, the tandem repeat becomes a functional single ORF which codes for a protein that confers resistance against the antibiotic Zeocin (Blackwood et al., 2010; Eykelenboom et al., 2008).

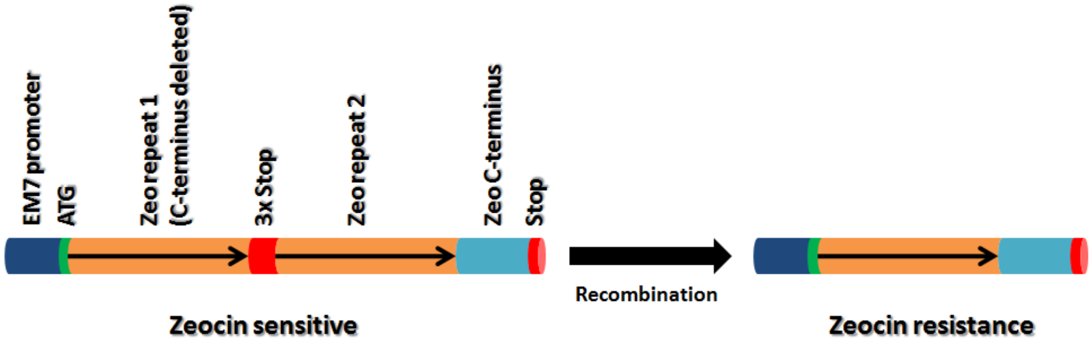


Figure 1.12. Structure of the 275 bp tandem repeat. The first part of the repeat contains an initiation codon but a deletion in the C-terminus of the protein it codes. The second contains an intact ORF except for the initiation codon. These two parts are separated by stop codons (3x Stop) at every frame to stop any translation starting from the initiation codon. Upon recombination between the incomplete repeats, a complete protein coding sequence emerges which codes for a protein that confers resistance against the antibiotic Zeocin.

1.7.3 Using native two-dimensional agarose gel electrophoresis to investigate any MMR dependent interruption of the progression of DNA replication

Nonlinear duplex DNA migrates anomalously in an agarose gel compared to the migration of linear molecule with equal molecular mass and this migration can be exaggerated by increasing either voltage or agarose concentration (Bell and Byers, 1983; Oppenheim, 1981). Therefore, two-dimensional (2D) gel electrophoresis has been shown to differentiate various non-linear DNA molecules from their counterpart linear duplex DNA of equal mass. In 1981, Ariella Oppenheim first introduced this technique to separate closed circular DNA plasmid from its linear DNA and nicked circular counterpart (Oppenheim, 1981). Then, this technique was modified and coupled with Southern blotting to analyse different replication and recombination intermediates (Bell and Byers, 1983). In 1987, Brewer and Fangman proved that the autonomous replication sequence (ARS) in yeast is an origin of replication upon isolating replication bubble structure by restriction digestion, followed by native 2D gel electrophoresis using a recombinant plasmid containing yeast ARS1 (Brewer and Fangman, 1987). Recombinant plasmid based 2D gel electrophoresis experiments became popular in the early period. In 1991, Martinez and collaborators were able to see a unidirectional replication by 2D gel electrophoresis using a bacterial plasmid based system (Martín-Parras et al., 1991). Later, other implementations of 2D gel electrophoresis became familiar to different groups of scientists. Krasilnikova and Mirkin used this technique to identify a pausing of the replication fork due to the difficulties of replicating a repeated array using both recombinant bacterial and yeast based plasmids

(Krasilnikova and Mirkin, 2004). In addition, this method was widely used to map replication fork barriers, replication termini and has been reviewed several times (Brewer and Fangman, 1988; Dijkwel and Hamlin, 1997; Friedman and Brewer, 1995; Pohlhaus and Kreuzer, 2006; Wellauer et al., 1976). In time, the 2D gel electrophoresis method became the means of direct study on the chromosomal DNA for studying DNA replication and recombination. Barre and collaborators used this technique to show that the resolution of *E. coli* chromosome dimers is coupled to cell division (Barre et al., 2000). 2D agarose gel electrophoresis has even been used to analyse the physical interaction of homologous chromosome in the process of generating heteroduplex DNA and to analyse telomeric DNA (Allers and Lichten, 2001; Makovets, 2009).

In the first dimension of 2D agarose gel electrophoresis, DNA fragments are separated based on molecular weight at a low voltage and 4°C to maintain a native condition. On the other hand, in the second dimension, high voltage and a DNA intercalating agent – ethidium bromide (0.3 µg/l) are applied to maximise the separation of the DNA fragments based on molecular shape of different DNA species. The particular pattern of DNA migration during the second dimension leads to unequivocal evidence of an origin of replication. 2D gel electrophoresis was basically developed to capture the replication intermediates by proper selection of restriction sites and visualise them after gel electrophoresis followed by Southern blotting and hybridisation with radio-labelled probes. For example, a duplex DNA with molecular mass “n” is under the process of replication and separated out from the rest of the DNA in the chromosome of an organism. As the replication fork is in the process of passing the duplex DNA of interest, there are different replication intermediates (RIs) ranging from unreplicated DNA with the molecular mass of “n” to fully replicated DNA with the molecular mass of “2n”. As noted before, the first dimension of the 2D gel electrophoresis will separate the RIs according to their molecular mass from slightly more than “n” to slightly less than “2n”. On the other hand, the DNA molecules will migrate along the agarose gel based on their molecular 3D shape where the half replicated RIs, having three branches of equal length (a “Y-shaped” DNA with three equal arms), deviates most extensively in its 3D shape from the simple linear DNA molecule and will migrate most slowly. As replication of the DNA fragment closes to completion, the unreplicated branch becomes progressively shorter in comparison to the two replicated branches and contributes less to the 3D shape. Thus the branched Rls would produce a continuous arc that begins at the position of linear molecules with the molecular mass “n”, has an inflection point for the molecules that are half-replicated, and ends at the position of linear molecules with the molecular mass “2n” on two-dimensional agarose gels, as shown in Figure 1.13A. This arc, formed due to the migration of “Y-shaped” molecules, is called the “Y-arc”, and is indicative that the DNA molecule is in the process of being replicated.

The DNA fragment containing an origin of replication at the centre will form a bubble (or eye form) and the replication fork will migrate along the DNA on both directions. In the first dimension the DNA molecules would separate between the molecular mass “n” and “2n” as expected. However, their properties would differ from that of those of simple Y-shaped DNAs and thus would not form the typical Y-arc. As the replication progresses and the molecular mass increases, the unreplicated linear arms decrease and the circle increases in diameter. So the relationship between the extent of replication and the 3D shape of the bubble structures becomes complex as the replication progresses. As a result, a characteristic shape is obtained after the completion of 2D gel electrophoresis for the migration of the bubble structure. Unfortunately, it is difficult to distinguish an intermediate created due to this type of fragment (indicated by the bubble-arc) from that created by a Y-shaped intermediate (Figure 1.13C), if the origin of replication stays at the centre of the fragment. This problem can be avoided by choosing the right combination of restriction sites so that the origin of replication falls asymmetrically in the DNA fragment. The replication fork closest to the end of the fragment will become Y-shaped from the initial bubble created at the origin. Further replication of the fragment led by one remaining fork would generate a series of large Y-shaped RIs, yielding a two-dimensional gel pattern with a discontinuity. This is called the "bubble-to-Y-transition" and is highly indicative of an origin of replication.

On the other hand, double Y-shaped DNA structures created by the approach of the two replication meeting at the DNA fragment will create an "X" shaped DNA. Thus the RIs generated from the beginning of the invading Y-shaped DNA structures to just before the merging of those replication fork (X-shape) will create RIs with molecular mass just above "n" to just below "2n" and will migrate in the first dimension as expected. However, it would not create a Y-arc upon completion of second dimensional gel electrophoresis because of its different topology in 3D structure than the simple Y-shaped DNA. In the case of double Y-shaped DNAs, all the four branches increase in length in direct proportion to the extent of replication while moving to each other. Ultimately, after the completion of migration on both dimensions in the 2D gel electrophoresis, DNA molecules with two replication forks moving towards each other generate the characteristic structure shown (Figure 1.13).

Therefore, getting a discontinuous pattern (rather than a “Y-arc”) is more informative in at least two ways –

Firstly, only the replication bubble and “X” structures can form such pattern. Moreover, those patterns can be distinguished from each other.

Secondly, the molecular mass of the RIs at the discontinuity reflects the position of the corresponding replication fork at the moment the other fork passed the restriction site. Estimating the distance between the point and the restriction site helps to map the origin of replication in the fragment.

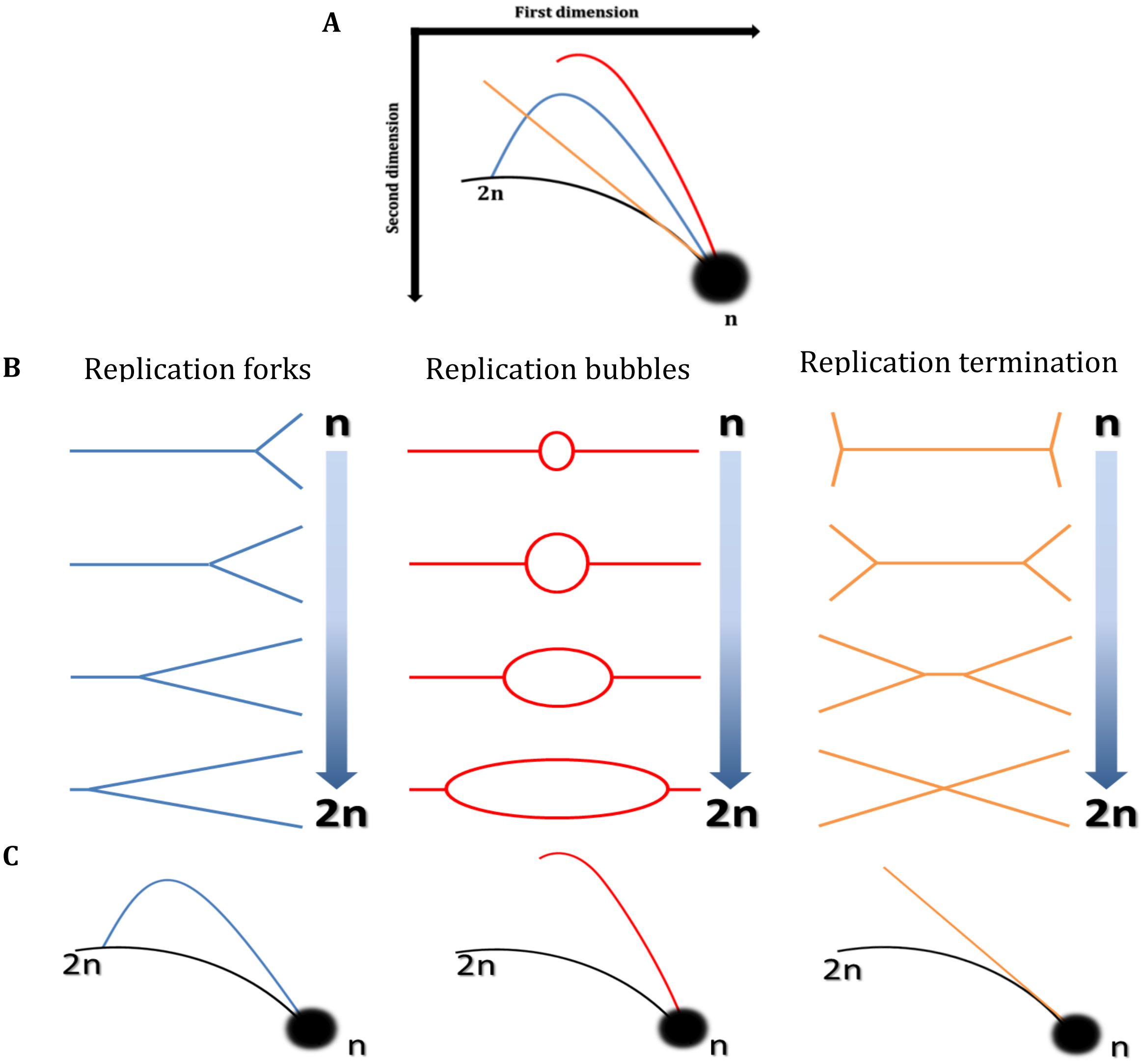


Figure 1.13. Overview of the 2-D agarose gel electrophoresis migration patterns. The large spots designated "n" indicate the positions of the abundant linear species of the restriction fragments. "2n" indicates the location of this linear species just prior to completion of replication. Accumulation of a particular structure, such as a blocked replication fork at a specific location along a restriction fragment, generates a spot along the relevant migration line. (A) Overview of the most common migration patterns observed. The arc of linear DNA is represented by the thin black line, which runs through n and 2n. (B) Breakdown of the different molecular shapes placed above the migration pattern they generate for replication forks, replication bubbles and replication termination. Figure modified from (Friedman and Brewer, 1995).

1.7.4 An *in silico* approach to generate artificial sequences and compare the distribution of GATC motifs with that of the *E. coli* genome

To test whether MMR caused any selection pressure on the genomic distribution of the GATC motifs in *E. coli*, two statistical methods have been applied to determine over- or under- representation of GATC motifs in *E. coli*.

1.7.4.1 The Rho(ρ) statistic for determining over- or under-representation of a motif

There are several ways to calculate whether a motif of particular sequence length is over- or under-represented than it is expected by chance. One way is to calculate the Rho(ρ) statistic. The Rho(ρ) statistic has been described by Karlin and Cardon to compute the over- or under-representation of any particular dinucleotide among the 16 possible combinations of dinucleotides (Karlin and Cardon, 1994). For a 2-nucleotide motif, the Rho(ρ) statistic is calculated as:

$$\rho_{\text{xy}} = \frac{fxy}{fx \times fy}$$

Where, f_{xy} = frequency of dinucleotide “xy”,

f_x = frequency of nucleotide “x” and

f_y = frequency of nucleotide “y”.

The idea behind the ρ statistic is that – if a DNA sequence has the frequency f_x and f_y of two different nucleotides “x” and “y” respectively, the expected frequency of the dinucleotide “xy” will be the product of their individual frequencies, that is (f_x×f_y). On the other hand, if the real frequency of the dinucleotide “xy” is found to be f_{xy}, it is expected to be equal to the product of the individual frequencies of the nucleotides that compose it. If it were true, the ρ statistic would be equal to 1. Therefore, ρ will be greater than 1 if the dinucleotide is more common in the sequence than expected or it is said to be "over-represented". On the other hand, ρ will be less than 1 if the dinucleotide is less common in the sequence than expected or it is said to be "under-represented".

This Rho(ρ) statistic can be adopted for a motif of any length in a DNA sequence.

Therefore,

$$\rho_{\mathrm{GATC}} = \frac{fGATC}{fG \times fA \times fT \times fC}$$

where, f_{GATC} = frequency of tetra-nucleotide "GATC",

f_G = frequency of nucleotide "G",

f_A = frequency of nucleotide "A",

f_T = frequency of nucleotide "T" and

f_C = frequency of nucleotide "C".

1.7.4.2 A Markovian model gives a weighted value with the over- or under-representativeness of a motif

The Rho(ρ) statistic basically describes the property of over- or under-representation of a DNA motif in the genome rather than providing a weighted value to it. To get a better understanding of the over-representativeness of the GATC motif, a probabilistic model (Markovian model) has been applied using the statistical program called R'MES. A Markov chain model of order *m* fits the observed counts of all oligonucleotides of length 1 up to (m+1) of the observed sequence.

Let's consider the following notations:

n is the genome length,

A is the four letter DNA alphabet,

X = $X_1X_2 \ldots \ldots \ldots X_n$ is a random sequence of letters from A (model Mm),

w is a word of length ℓ of A,

N denotes the count,

Y_i(**w**) is 1 if w occurs at position i in **X**, and 0 otherwise.

The number of occurrences of **w** in **X** can be written like

$$\mathrm{N}(\mathbf{w}) = \sum_{i=1}^{n-\ell+1} \mathrm{Yi}(\mathbf{w})$$

And its expectation is EN(**w**) = (n − ℓ + 1)P(**w** at i). The probability P(**w** at i)can be easily expressed with respect to the transition probabilities.

In Markovian approximation, a motif of size ℓ can be only analysed in $M0$ up to $M(\ell - 2)$ because higher models would fit the motif count itself (the motif will then be expected by definition). So the expected number of the GATC motifs can be calculated up to second order. Both Gaussian approximation and compound Poisson approximation method have been applied to calculate the expected occurrence of GATC motif. The Gaussian approximation is not good for the count of expectedly rare words, while compound Poisson approximation gives an accurate *p*-value. However, the Gaussian distribution is well adapted when the estimated expected counts are far from 100 (which is the case for GATC motif in *E. coli*). On the other hand, the Poisson approximation is satisfactory for the count of a non-overlapping word, not for overlapping words. R'MES converts the *p*-values into scores of exceptionality. Here, exceptionally frequent motifs will have high positive scores, whereas exceptionally rare motifs will have high negative scores.

1.7.4.3 An *in silico* approach to generate different artificial sequences

To determine whether the distribution of GATC motifs in *E. coli* is purely expected by chance or shaped by any selection pressure imposed by MMR, it is essential to generate different artificial sequences. There are several mathematical processes to generate an artificial sequence. The simplest approach would be using "Multinomial sequence model" which assumes that in the process of evolution the DNA sequence is produced by a random process that randomly selects one of the four nucleotides at each position in the sequence. According to this model, the probability of choosing any one of the four nucleotides depends on a predetermined probability distribution. Thus, a multinomial model has only four parameters: the predetermined probability of the four nucleotides p_A, p_C, p_G, and p_T respectively where $p_A + p_C + p_G + p_T = 1$. However, the main drawback of this process is that it is not an accurate image of how sequences evolved. The model assumes that each part of the sequence has the same frequency of each nucleotide, that is $p_A = p_C = p_G = p_T$. For example, the first 100 bases have the same composition of each nucleotide although the total composition of the four nucleotides follows the predetermined probabilities. Thus, this assumption may not be true for a particular sequence if there are considerable differences in the nucleotide frequencies in the same part of the sequence. Another crucial setback of this model is that the probability of a nucleotide at a position depends only on the predetermined probability of that nucleotide, not on the nucleotides around it in the sequence. For many sequences this particular drawback may not have any impact. However, for a biologically meaningful sequence it is crucial that a nucleotide at a given position should depend on the adjacent nucleotide composition and often has an evolutionary pressure for that particular nucleotide due to the adjacent compositions.

On the other hand, a more accurate representation of the evolution of a sequence is the “Markov sequence model”. This model assumes that the sequence is produced by a process that chooses any of the four nucleotides in the sequence and the probability of choosing any nucleotide at a particular position depends on the nucleotide chosen for the previous position. Details on this model will be discussed in Chapter 2 and 6.

1.8 Work in this thesis

Work presented in this thesis aimed to investigate the *in vivo* directionality of the DNA mismatch repair system in *Escherichia coli*. Most of the previous studies regarding MMR were based on *in vitro* experimentation using linear or plasmid based heteroduplex DNA in defined systems (Blackwell et al., 1998; Dzantiev et al., 2004; Thomas et al., 1991). Being stochastic in nature, a mismatch can occur at any locus in the chromosome. Therefore, a defined system that generates substrates for the MMR system at a defined locus is crucial. An incredible system of a trinucleotide repeat (TNR) array has been used in this study as such a system that addresses different aspects of MMR in an *in vivo* system (in *E. coli*) and redefines the MMR pathway.

In Chapter 3, the *in vivo* directionality of so far thought bidirectional MMR system in *E. coli* has been determined by calculating the frequency of instability of a CTG•CAG repeat array as a result of sequential modification of GATC motifs around the repeat array. The frequency of instability has been analysed by determining the length variation of the repeat array using ABI 3730 Genetic Analyser followed by GeneMapper® microsatellite analysis.

A 275 bp tandem repeat has been found to recombine due to ssDNA generated during the excision repair reaction of MMR and the frequency of recombination is a function of the distance from the TNR array (Blackwood et al., 2010). Chapter 4 discusses the short and long distance excision reaction during MMR using this tandem repeat system.

The MMR machinery has been proposed to be functionally coupled with the replication machinery (López de Saro et al., 2006; Warbrick, 2000, 2006). In Chapter 5, the hypothesis that "A functional MMR impedes the progression of the DNA replication machinery" has been tested using native two-dimensional agarose gel electrophoresis of a genomic DNA fragment of *E. coli* containing the TNR array upon restriction digestion with a view to accumulate DNA replication intermediates.

In Chapter 6, I have used an *in silico* approach to test whether MMR imposes any selection pressure on the distribution of the GATC motifs in *E. coli* genome as the methylation state of this motif is the genomic signature for the MMR system to select which DNA strand to repair and thereby, to maintain genomic integrity.

Chapter 2

Materials and methods

2.1 Materials

2.1.1 Buffers and stock solutions

All buffers and solutions were stored at room temperature. Chemicals were dissolved in double distilled (ddH_2O), sterile water (Milli-Q®), unless stated otherwise.

5 X Tris-borate (TBE)	**1 Liter**
0.89 M Tris Base	53 g
0.89 M Boric acid	27.5 g
10 mM EDTA	3.4 g
pH adjusted to 8.0 with HCl	

20 X SSC	**1Litre**
3 M NaCl	175.2 g
300 mM Tri-sodium citrate	77.4 g
pH adjusted to 7.0 with HCl	

20 X SSPE	**500 ml**
3 M NaCl	87 g
200 mM NaH_2PO_4	12 g
20 mM EDTA	20 ml of 0.5 M solution at pH 8.0

50 X Tris-acetate (TAE)	**1 Litre**
2 M Tris-base	242 g
0.95 M glacial acetic acid	57.1 ml
0.05 M EDTA	14.6 g

Alkaline transfer buffer	**1 Litre**
0.5 M NaOH	20 g
10 X SSC	500 ml of 20 X SSC

Church-Gilbert buffer	**20 ml**
7% SDS	14 ml of a 10% stock solution
0.5 M NaH_2PO_4	5 ml of a 2 M solution at pH 7.2
1 mM EDTA	40 µl of a 0.5 M solution at pH 8.0
1% BSA	0.2 g

The NaH_2PO_4, EDTA, BSA, and 960 µl of distilled water were mixed until the BSA was completely dissolved. The SDS was then added and the whole was heated to allow for easy mixing. The warm solution was filter-sterilised using a filter with a pore size of 0.4 µm.

Depurination solution	**500 ml**
0.25 M HCl	12.5 ml of a 37% solution

Detection buffer	**500 ml**
0.1 M Tris-HCl	6.055 g
0.1 M NaCl	2.922 g

The pH was adjusted to 9.5 with HCl

EDTA	**1 L**
0.5 M EDTA	186.12 g

The pH was adjusted to 8.0 with NaOH

High stringency buffer	**500 ml**
0.1 X SSC	2.5 ml of 20 X SSC
0.1% SDS	5 ml of 10% SDS

Low stringency buffer	**500 ml**
2 X SSC	50 ml of 20 X SSC
0.1% SDS	5 ml of 10% stock SDS solution

NDS buffer	**500 ml**
0.5 M Na_2.EDTA	93 g

10 mM Tris-base	0.6 g
0.6 mM NaOH	11 g
34 mM N-lauroylsarcosine	5 g

The Na_2.EDTA, Tris-Base, and NaOH were dissolved in 350 ml of distilled water. Separately, the N-lauroylsarcosine was completely dissolved in 50 ml of distilled water and then added to the main solution. The pH was adjusted to 8.0 with NaOH and the total volume was brought up to 500 ml.

Stringency washing solution	**500 ml**
0.5 X SSC	12.5 ml of 20 X SSC
0.1% SDS	5 ml of 10% SDS stock solution

Stripping buffer	**50 ml**
50% formamide	25 ml of 100% formamide
5 X SSPE	12.5 ml of 20 X SSPE

TE buffer	**1 Litre**
10 mM Tris	1.2 g
1 mM EDTA	0.3 g

pH adjusted to 7.4 with HCl

TEN buffer	**1 Litre**
50 mM Tris	6.1 g
50 mM EDTA	14.6 g
100 mM NaCl	5.8 g
pH adjusted to 8.0 with HCl	

Washing solution	**500 ml**
2 X SSC	50 ml of 20 X SSC
0.1% SDS	5 ml of 10% SDS

20% (w/v) Arabinose

Made up to 20% (w/v) in ddH_2O and autoclaved.

20% (w/v) Glucose

Made up to 20% (w/v) in ddH_2O and autoclaved.

20% (w/v) Sucrose

Made up to 20% (w/v) in ddH_2O and autoclaved.

80% (v/v) Glycerol

Made up to 80% (v/v) in sterile Milli-Q® water and autoclaved.

2.5 M $CaCl_2$

Made up to 2.5 M in sterile Milli-Q® water; Sterilised using a 0.2 μm syringe filter and used at 2.5 mM and 0.1 M.

Ethidium bromide (EtBr) 10 mg/ml (FlukaBiochemika)

Stored in the dark at room temperature; diluted to 0.5 µg /ml in sterile Milli-Q® water when used for staining of electrophoresis gels for the visualisation of DNA under a UV light source, unless otherwise stated.

2.1.2 Culture media

All liquid growth media were made up to the required volume in distilled water and autoclaved. These were stored at room temperature except during experiments when they were pre-warmed to the desired temperature. M9-minimal medium was filter sterilised prior to use. To obtain solid media, agar was added to liquid media prior to autoclaving except for M9-minimal agar where 2% agar-distilled water was first autoclaved, melted, and allowed to cool prior to the addition of salts and sugar. Melted agar was stored at 55°C and allowed to cool prior to the addition of antibiotics and inducers (when required).

Table 2.1. Bacteria growth media. Concentrations indicated as percentages are weight/volume (w/v).

Media	**Composition**
L Broth	1% bacto-tryptone, 0.5% yeast extract, 1% NaCl; pH adjusted to 7.5 with NaOH
LB Agar	1% bacto-tryptone, 0.5% yeast extract, 1% NaCl, 1.5% bactoagar; pH adjusted to 7.5 with NaOH
M9 Salts (4x)	28 g Na_2HPO_4, 12 g KH_2PO_4, 2 g NaCl, 4 g NH_4Cl; made up to 1 L with double-distilled sterile water
M9-Minimal Media	1 x M9 salts, 0.2% glycerol, 1 mM $MgSO_4$, 5µM $CaCl_2$

M9-Minimal Agar	1 x M9 salts, 0.2% glycerol, 1 mM $MgSO_4$, 5 µM $CaCl_2$, 1.5% bactoagar
Low salt LB Agar	10 g tryptone, 5 g NaCl, 5 g yeast extract dissolved in 950ml ddH2O; pH adjusted to 8.0 with 5 N NaOH; made up to 1 l with deionised H_2O. For solid medium, autoclaved for 20 minutes at 15 psi (1.05 kg/cm^2) on liquid cycle.

2.1.3 Antibiotics

All antibiotics (Table 2.2) were dissolved in Milli-Q® water, unless otherwise stated and stored at -20°C.

Table 2.2. List of antibiotics

Antibiotics	**Solvent**	**Stock concentration**	**Working concentration**
Ampicillin (Amp)	Water	100 mg ml^{-1}	100 µg ml^{-1}
Chloramphenicol (Cm)	100% ethanol	50 mg ml^{-1}	50 µg ml^{-1}
Kanamycin (Km)	Water	50 mg ml^{-1}	50 µg ml^{-1}
Rifampicin (Rif)	Methanol	50 mg ml^{-1}	50 µg ml^{-1}
Tetracyclin (Tc)	50% ethanol	15 mg ml^{-1}	3 µg ml^{-1}
Zeocin (Zeo)	Water	100 mg ml^{-1}	25-50 µg ml^{-1}

2.1.4 Enzymes

DNA polymerase enzymes used were Taq polymerase purchased from Roche. GoTaq® and Pfu DNA polymerase were purchased from Promega. All restriction enzymes were purchased from New England Biolabs (NEB). All enzymes were used according to the manufacturer's guidelines.

2.1.5 *E. coli* strains

Strains that were used in this study are listed in Table 2.3. XL1-Blue was used for all cloning procedures and the propagation of plasmid DNA, whereas derivatives of MG1655 were used for experiments.

2.1.6 Plasmids

Plasmids used in this study are listed in Table 2.4. The *E. coli* strain XL1-Blue was used for all cloning procedures and the propagation of plasmid DNA.

2.1.7 Oligonucleotides

Oligonucleotides were manufactured by MWG Biotech where they were synthesised, HPSF purified and lyophilized. 100 mM stock solutions were made by dissolving in sterile water and stored at -20°C. Working solutions were made up to 5 mM in sterile water and stored at -20°C. Primers were designed using the internet-based tool Primer3Plus and Primer-BLAST (URL are http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi and http://www.ncbi.nlm.nih.gov/tools/primer-blast respectively). Oligonucleotides used in this study are shown in Table 2.5.

Table 2.3. ***E. coli*** **strains used in this study**

Strain	Background	Genotype	Source
DL-1719	XL1-Blue	*recA1, endA1, gyrA96, thi-1, hsdR17, supE44, relA1, lac, [F' proAB, lacI*^{q}*, lacZΔM15, Tn10]*	Stratagene
DL-4150	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$	Ewa Okely
DL-4151	MG1655	*lacZ::zeo*	Ewa Okely
DL-4954	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ + 1st GATC site in the range of 365575-367799 bp has been modified.	This study
DL-4955	MG1655	*lacZ::zeo* + 1st GATC site in the range of 365575-367799 bp has been modified.	This study
DL-4987	MG1655	*lacZ::zeo lacZ::CTG*$_{9\,8}$+ 1st and 2nd GATC sites in the range of 365575-367799 bp have been modified.	This study
DL-4988	MG1655	*lacZ::zeo* + 1st and 2nd GATC sites in the range of 365575-367799 bp have been modified.	This study
DL-5022	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ + 1st to 3rd GATC sites in the range of 365575-367799 bp have been modified.	This study
DL-5023	MG1655	*lacZ::zeo* + 1st to 3rd GATC sites in the range of 365575-367799 bp have been modified.	This study
DL-5078	MG1655	XL1-Blue + pTOF-Dn3	This study
DL-5079	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ *ΔpriB*(partial deletion)	This study

DL-5080	MG1655	*lacZ::zeo ΔpriB* (partial deletion)	This study
DL-5081	MG1655	XL1-Blue + pTOF-Dn567	This study
DL-5086	MG1655	XL1-Blue + pTOF-Dn4	This study
DL-5089	MG1655	*lacZ::zeo lacZ::(CTG)$_{98}$* + 1st to 4th GATC sites in the range of 365575-367799 bp have been modified.	This study
DL-5090	MG1655	*lacZ::zeo* + 1st to 4th GATC sites in the range of 365575-367799 bp have been modified.	This study
DL-5151	MG1655	*lacZ::zeo mutS::Tc ΔpriB*	This study
DL-5152	MG1655	*lacZ::zeo lacZ::(CTG)$_{98}$ mutS::Tc ΔpriB*	This study
DL-5153	MG1655	*lacZ::(CAG)$_{84}$ lacZ::zeo ΔpriB*	This study
DL-5155	MG1655	*lacZ::zeo lacZ::(CTG)$_{98}$ΔpriB ΔmutS*	This study
DL-5156	MG1655	*lacZ::zeo ΔpriB ΔmutS*	This study
DL-5157	MG1655	*lacZ::zeo lacZ::(CTG)$_{98}$* + 365575-367799 bp deleted	This study
DL-5158	MG1655	*lacZ::zeo* + 365575-367799 bp deleted	This study
DL-5159	MG1655	*lacZ::zeo* + All GATC motifs within 365575-367799 bp have been modified	This study
DL-5160	MG1655	*lacZ::zeo* + 1st to 4th and 7th GATC motifs have been modified within 365575-367799 bp	This study

DL-5161	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ + 1st to 6th GATC motifs have been modified within 365575-367799 bp	This study
DL-5162	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ + 1st to 4th and 7th GATC motifs have been modified within 365575-367799 bp	This study
DL-5171	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ *ΔpriB ΔmutS*	This study
DL-5262	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ + All GATC motifs have been modified within 365575-367799 bp	This study
DL-5308	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ + 1st to 8th GATC motifs have been modified within 365575-368575 bp	This study
DL-5309	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ + 1st to 8th GATC motifs have been modified within 365575-368575 bp	This study
DL-5310	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ + 1st to 9th GATC motifs have been modified within 365575-368575 bp	This study
DL-5311	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ + 1st to 9th GATC motifs have been modified within 365575-368575 bp	This study
DL-5357	MG1655	*lacZ::zeo ΔpriB* + All GATC motifs have been modified within 365575-367799 bp	This study
DL-5358	MG1655	*lacZ::zeo ΔmutS* + All GATC motifs have been modified within 365575-367799 bp	This study
DL-5359	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ *ΔpriB* + All GATC motifs have been modified within 365575-367799 bp	This study
DL-5360	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ *ΔmutS* + All GATC motifs have been modified within 365575-367799 bp	This study
DL-5361	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ *ΔpriB* + 1st to 8th GATC motifs have been modified within 365575-368575 bp	This study
DL-5362	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ *ΔmutS* + 1st to 8th GATC motifs have been modified within 365575-368575 bp	This study

DL-5363	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ *ΔpriB* + 1st to 9th GATC motifs have been modified within 365575-368575 bp	This study
DL-5364	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$*ΔmutS* + 1st to 9th GATC motifs have been modified within 365575-368575 bp	This study
DL-5365	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ + 1st to 8th GATC motifs have been modified within 365575-368575 bp + GATC changed @180 O/D of TNR	This study
DL-5366	MG1655	*lacZ::zeo lacZ::(CTG)*$_{98}$ + 1st to 9th GATC motifs have been modified within 365575-368575 bp + GATC changed @180 O/D of TNR	This study

Table 2.4. Plasmids used in this study

Strain	**Plasmid**	**Brief description**	**Source**
DL-2715	pKO-*mutS*	For knocking out *mutS* gene.	Ewa Okely
DL-2765	pKO-recQ	For knocking out *recQ* gene.	Ewa Okely
DL-4822	pKO-priB	For knocking out *priB* gene.	Benura Azeroglu
DL-5078	pDn3	To modify the 3rd GATC motif between 365575-367799 bp in *E. coli* genome.	This study
DL-5081	pDn5-6	To modify the 5th and the 6th GATC motif between 365575-367799 bp in *E. coli* genome.	This study
DL-5086	pDn4	To modify the 4th GATC motif between 365575-367799 bp in *E. coli* genome.	This study

DL-5154	pKO-Dn	To delete DNA sequence from 365575-367799 bp in *E. coli* genome.	This study
DL-5711	pSpDn	To introduce a 2 kb GATC motif free *Schizosaccharomyces pombe* sequence at 365725 bp.	This study
DL-5719	pDn1	To modify the 1st GATC motif between 365575-367799 bp in *E. coli* genome.	This study
DL-5720	pDn2	To modify the 2nd GATC motif between 365575-367799 bp in *E. coli* genome.	This study
DL-5721	pUp1-12	To modify all GATC motifs between 363500-365575 bp in *E. coli* genome.	This study
DL-5727	pKO-Up	To delete the DNA sequence between 363500-365575 bp in *E. coli* genome.	This study

Table 2.5. Oligonucleotides used in this study

Name of the oligonucleotides	Sequence (5' to 3')	Purpose
Dn_GATC_1.1.F	AAA AAA CTC GAG TAA AGT GTA AAG CCT GGG	Used to generate a 800 bp DNA fragment to clone in pTOF24 plasmid to modify an origin distal GATC motif at 472 bp from the TNR array
Dn_GATC_1.1.R	GCC ATC TGA CCG TTG GCA AC	
Dn_GATC_1.2.F	AAA AAG TCG ACC TTC TCG CGC AAC GCG TCA G	
Dn_GATC_1.2.R	GTT GCC AAC GGT CAG ATG GC	

Dn_GATC_2.1.F	AAA AAC TCG AGC AGC ATC GCA GTG GGA AC	Used to generate a 800 bp DNA fragment to clone in pTOF24 plasmid to modify an origin distal GATC motif at 845 bp from the TNR array
Dn_GATC_2.1.R	CAG TGG GCT AAT CAT TAA CTA TCC GC	
Dn_GATC_2.2.F	GCG GAT AGT TAA TGA TTA GCC CAC TG	
Dn_GATC_2.2.R	AAA AAG TCG ACC CAG TAA CGT TAT ACG ATG TCG	
Dn_GATC_3.1.F	AAA AAC TCG AGG GCA CTC CAG TCG CCT TCC	Used to generate a 800 bp DNA fragment to clone in pTOF24 plasmid to modify an origin distal GATC motif at 949 bp from the TNR array
Dn_GATC_3.1.R	CTC GCG CCG GTC AAC TGG	
Dn_GATC_3.2.F	CCA GTT GAC CGG CGC GAG	
Dn_GATC_3.2.R	AAA AAG TCG ACG AAG GGG TTG AAT CGC AGG C	
Dn_GATC_4.1.F	AAA AAC TCG AGC GCG AGA TTT AAT CGC CGC	Used to generate a 800 bp DNA fragment to clone in pTOF24 plasmid to modify an origin distal GATC motif at 1356 bp from the TNR array
Dn_GATC_4.1.R	GTA GCA AAA CAT ATC GAA GAA GGG GTT G	
Dn_GATC_4.2.F	CAA CCC CTT CTT CGA TAT GTT TTG CTA C	

Dn_GATC_4.2.R	AAA AAG TCG ACG AAA CCA CTC ACC GCC ATC GCC	
Dn_GATC_5-6.1.F	AAA AAC TCG AGT AAT TCA GCT CCG CCA TCG CC	Used to generate a 800 bp DNA fragment to clone in pTOF24 plasmid to modify two origin distal GATC motifs at 1495 bp and 1506 bp from the TNR array
Dn_GATC_5-6.1.R	CCG CGG CTG GGA CCA GGA GGA GAA GAT TGC CTC TAT CGC C	
Dn_GATC_5-6.2.F	GGC GAT AGA GGC AAT CTT CTC CTC CTG GTC CCA GCC GCG G	
Dn_GATC_5-6.2.R	AAA AAG TCG ACG TAG CCC CTC CGA TGA TAG	
Dn_GATC_7.1.F	AAA AAC TCG AGT GCC TGT TCA ATG GTC ATT G	Used to generate a 800 bp DNA fragment to clone in pTOF24 plasmid to modify an origin distal GATC motif at 1826 bp from the TNR array
Dn_GATC_7.1.R	CTG CTG GGC GGT CTG TTG C	
Dn_GATC_7.2.F	GCA ACA GAC CGC CCA GCA G	
Dn_GATC_7.2.R	AAA AAG TCG ACG TTA AAA ACA TGT AAA TG	
Up1.F	AAA AAC TCG AGT CAA ATT CAG ACG GCA AAC	Used to generate a 500 bp DNA fragment to clone in pTOF24 plasmid to modify an origin proximal GATC motif at 126 bp from the TNR array
Up1.R	AAA AAG TCG ACC AAC GTC GTG ACT GGG AAA AC	

Up2.F	AAA AAC TCG AGT CAA ATT CAG ACG GCA AAC	Used to generate a 500 bp DNA fragment to clone in pTOF24 plasmid to modify an origin proximal GATC motif at 225 bp from the TNR array
Up2.R	AAA AAG TCG ACC AAC GTC GTG ACT GGG AAA AC	
Up3.F	CCG CCA CAT ATC CTG GTC TTC CAG ATA ACT G	Used to generate DNA fragment to modify an origin proximal GATC motif at 590 bp from the TNR array
Up3.R	CAG TTA TCT GGA AGA CCA GGA TAT GTG GCG G	
Up4.F	CGT AGT GTG ACG CGG TCG GCA TAA CCA CC	Used to generate DNA fragment to modify an origin proximal GATC motif at 854 bp from the TNR array
Up4.R	GGT GGT TAT GCC GAC CGC GTC ACA CTA CG	
Up5.F	GGT AGC CAG CGC GGG TCA TCG GTC AGA CG	Used to generate DNA fragment to modify an origin proximal GATC motif at 1280 bp from the TNR array
Up5.R	CGT CTG ACC GAT GAC CCG CGC TGG CTA CC	
Up6.F	CAG ATA ATC ACA CTC GGG TGA TTA CGG TCG CGC TGC ACC ATT C	Used to generate DNA fragment to modify five origin proximal GATC motifs from1335 bp to 1430 bp from the TNR array
Up7.F	CCC AGC GAC CAG ATA ATC ACA CTC GGG TGA TTA C	
Up7.R	GTA ATC ACC CGA GTG TGA TTA TCT GGT CGC TGG G	

Up8-9.F	CAC CGG GCG GGA AGG GTC GAC AGA TTT AAT CCA GCG ATA CAG	
Up8-9.R	GCT GTA TCG CTG GAT TAA ATC TGT CGA CCC TTC CCG CCC GGT G	
Up10.F	GGC GTA TTC GCA AAG AAT CAG CGG GCG CGT C	Used to generate DNA fragment to modify an origin proximal GATC motif at 1594 bp from the TNR array
Up10.R	GAC GCG CCC GCT GAT TCT TTG CGA ATA CGC C	
Up11.F	TTT AAT CAG CGA CTG GTC CAC CCA GTC CCA G	Used to generate DNA fragment to modify an origin proximal GATC motif at 1690 bp from the TNR array
Up11.R	CTG GGA CTG GGT GGA CCA GTC GCT GAT TAA A	
Up12.F	CAT ACA GAA CTG GCA ATC GTT CGG CGT ATC	Used to generate DNA fragment to modify an origin proximal GATC motif at 1787 bp from the TNR array
Up12.R	GAT ACG CCG AAC GAT TGC CAG TTC TGT ATG	
End-Up.F	AAAAACTGCAGGGTCGGGATAGTTTTCTTGCG	Used to amplify a 2 kb fragment to modify all origin proximal GATC motifs within 2 kb from the TNR array
End-Up.R	AAAAAGGATCCCAACGTCGTGACTGGGAAAAC	
seqA_KO.F1	AAA AAC TCG AGC CGG GTA ATT CAT TAC CGG ATC	Used to generate a 800 bp DNA fragment to clone in a pTOF24 plasmid to knock

seqA_KO.F2	GTT GAT GAT GAA CTC TAC AGC GAG AAG GTT TGC GGA ACT ATC	out *seqA* gene
seqA_KO.R1	GAT AGT TCC GCA AAC CTT CTC GCT GTA GAG TTC ATC ATC AAC	
seqA_KO.R2	AAA AAG TCG ACC GGG GTG AAG CCA TTG TTT TC	
Sp.R1	GTT TCA TCT GTG GTG GTT ATT TAC ATC TAA TTG AG	To amplify a GATC motif-free 2 kb DNA sequence from *Schizosaccharomyces pombe* and clone in a pTOF24 plasmid and ultimately insert into *E. coli* chromosome
Sp.R2	GTT TCA TCT GTG GTG ACA CCT TTC CAG CAC CAT AGG	
Sp-3.F	GCC GTA ACC GAC CCA GCG CCC GTT GGG TTG CAT GCT TGG CGT TTA	
Sp-3.R	TAA CTC GGC GTT TCA TCT GTG GTG AAA AGC AGG GTT CCG GAC AT	
Sp-4.F	GCC GTA ACC GAC CCA GCG CCC GTT GGA GGT GCA CTT GCC TAT GGT	
Sp-4/5.R	TAA CTC GGC GTT TCA TCT GTG GTG AAA TAG ACC GCC GAG GCA AA	
Sp-5.F	GCC GTA ACC GAC CCA GCG CCC GTT GAC GTC ACT TGG AGA AAC GCA	
Sp-Ec2 joint.F1	CTC AAT TAG ATG TAA ATA CAC CAC AGA TGA AAC GCC GAG	

Sp-Ec2 joint.F2	CTA TGG TGC TGG AAA GGT GCA CCA CAG ATG AAA CGC C	
Zeo-CF1	AAA AAG TCG ACT ATC AAA CAC TCG CCT GGT G	To amply *lacZ*::zeocin tandem repeat containing DNA sequence.
Zeo-CR2	AAA AAC TGC AGT TAT TGG CGC GGG TAG TAT C	
Ex-test_F	TTA TGC TTC CGG CTC GTA TG	To amply *lacZ*::TNR array containing DNA sequence.
Ex-test_R	GGC GAT TAA GTT GGG TAA CG	
LacZ.dist F	ATC GTC GTA TCC CAC TAC CG	To amplify a DNA sequence to make radio-labelled probe
LacZ.dist R	TTT CCA TGC GAG GTT AAA GG	

2.2 Methods

2.2.1 Bacterial methods

2.2.1.1 Bacterial stocks stored at -80°C

From an overnight culture, 0.75 ml was mixed with 0.75 ml of 80% (v/v) glycerol and placed in a 1.5 ml Eppendorf tube. The tube was vortexed, sealed with a strip of parafilm, and stored at -80°C.

2.2.1.2 Overnight cultures

Five ml of L-broth with the required additives were inoculated with a single colony from an LB agar plate which was previously prepared from streaking the -80°C stock. Cultures were incubated overnight at 37°C with shaking (120 rpm).

2.2.1.3 Transformation of *E. coli* by $CaCl_2$ treatment followed by a heat shock

The bacterial strain to be transformed was cultured overnight. On the next day, the culture was diluted 50 times in 25 ml L-broth and put at 37°C with shaking (120 rpm) for 2 hours. Then, 2 ml of culture was spun down in a table-top centrifuge at maximum speed for 1 minute. The supernatant was removed and 500 µl of fresh pre-chilled 0.1 M $CaCl_2$ was added. The cell pellet was re-suspended and the suspension was put on ice for 30 minutes. The culture was centrifuged as described before and the supernatant was discarded. Then, 100 µl of fresh pre-chilled 0.1 M $CaCl_2$ was added and the pellet was re-suspended. After that, 0.5 µl of DNA (when using a plasmid isolated by mini-prep) or 10 µl of DNA (when transforming with a newly ligated plasmid obtained from an *in vitro* ligation mix) was added to the mixture and put on ice for another 30 minutes. The cells were then heat shocked in a table-top heat block at 37°C for 5 minutes, placed back on ice for about 5 minutes, and then mixed with 400 µl of fresh L-broth without any selection. The cells were allowed to recover at 37°C (30°C when a temperature sensitive plasmid was used for the transformation) with shaking (120 rpm) for 1 hour. After that, the cells were plated on LB agar plates with the appropriate selection added. For every transformation, a control (without addition of DNA) was prepared in parallel following the same procedure.

2.2.1.4 Plasmid mediated gene replacement

In order to modify the chromosome (either base alteration or gene knockout), a plasmid mediated recombineering based chromosome alteration technique was followed. This technique involved two regions flanking the desired alteration that will recombine twice independently to make the targeted alteration in the chromosome at the targeted locus. This technique, plasmid-mediated gene replacement (PMGR), was first described by Link and collaborators (Link et al., 1997) (Fig. 2.1). The plasmid, pTOF24, has a temperature sensitive replication initiator protein (*repA*101$_{\mathrm{TS}}$), two positive selection markers in the form of a chloramphenicol resistance gene (*cat*) and a kanamycin resistance gene (*aph*), and a gene encoding for a levansucrase (*sacB*) that can act as negative selection as *E. coli* strains expressing this gene become sensitive to sucrose (Merlin et al., 2002). The kanamycin resistance gene is flanked by XhoI/SalI restriction sites, which allows for the replacement of the *aph* gene with a cloning fragment of interest. Targeted gene alteration(s), flanked by about 400 bp of homology to the surrounding sequence was/were amplified by cross-over PCR. The amplified DNA sequence was cloned into pTOF24 using the Xho/SalI restriction sites. The strain to be modified was transformed, using the $CaCl_2$ mediated method described earlier, with pTOF24-derivatives using chloramphenicol resistance as a selectable marker. Following the transformation, transformants were recovered by incubating plates at 30°C so as to ensure autonomous replication of the plasmid. Successful transformants were then streaked on fresh LB-agar + chloramphenicol and placed at 42°C to select for strains in which the plasmid had integrated into the chromosome at one of the homology arms. This was repeated a second time to ensure purity of the integrants. In order to complete PMGR, the plasmid had to be excised from the chromosome. This was done by culturing individual integrants in 5 ml LB with no selection, overnight, at 30°C with shaking. The culture obtained, containing a mixture of cells that had either lost or retained the plasmid, was serially diluted from 10^{-1} to 10^{-6} in L-broth media. Then, 100 µl of dilutions 10^{-4}, 10^{-5} and 10^{-6} were plated onto LB-agar containing 5% sucrose to select against all cells that had retained the plasmid. Colonies were checked for chloramphenicol sensitivity to confirm the loss of the plasmid. Sucrose resistant/chloramphenicol sensitive colonies were checked for integration of the cloning fragment of interest by boiled colony PCR. It was expected that about 50% of the colonies would retain the wild type DNA sequence and 50% would acquire the DNA alteration. Once colonies that generated the expected PCR product size were identified, the chromosomal modification was confirmed by sequencing.

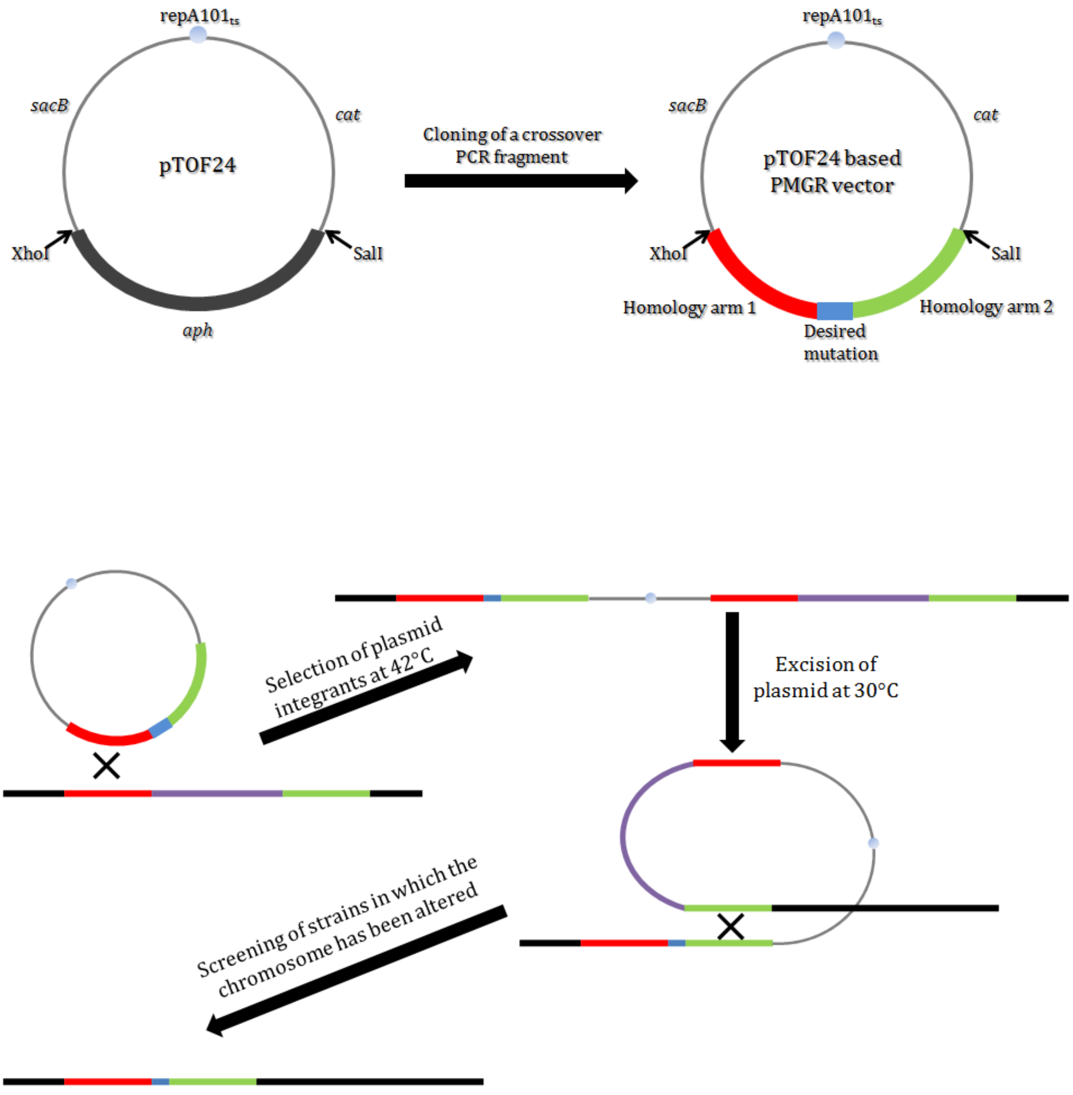

repA101ts
sacB
cat
pTOF24
XhoI
SalI
aph
Cloning of a crossover PCR fragment
repA101ts
sacB
cat
pTOF24 based PMGR vector
XhoI
SalI
Homology arm 1
Homology arm 2
Desired mutation
Selection of plasmid integrants at 42°C
Excision of plasmid at 30°C
Screening of strains in which the chromosome has been altered

Figure 2.1. Plasmid mediated gene replacement (PMGR). (A) Construction of pTOF24-derivatives containing products from a crossover PCR that are cloned into the PstI/SalI locus of the plasmid. *repA101$_{TS}$* codes for a temperature sensitive replication initiator protein (the permissive temperature being 30°C and the non-permissive temperature being 42°C). *aph* and code for kanamycin and chloramphenicol resistance, respectively. *sacB* codes for a levansucrase, which converts sucrose into a toxic product for *E. coli,* and is used as a negative selection marker. (B) Utilisation of these pTOF24-derivatives for targeted chromosomal modifications in *E. coli*. At 42°C the plasmid can only be replicated

if it integrates into the chromosome. pTOF24-derivatives are designed to contain two regions of homology to the chromosome. Integration at 42°C will occur by RecA-mediated homologous recombination between one of the two regions of homology and the same region in the chromosome. This will result in integration of the entire plasmid sequence. Growing in the absence of selection, the integrant at 30°C in liquid culture will allow the plasmid to excise from the chromosome. This also occurs by RecA-mediated homologous recombination. If integration occurs at the first region of homology (red) and excision occurs at the second (green), the wild type region of the chromosome is replaced with the modified DNA insert from the plasmid.

2.2.2 Molecular biology methods

2.2.2.1 Plasmid purification

To isolate plasmid DNA, the *E. coli* strain harbouring the desired plasmid was first grown overnight with appropriate selection in L broth at either 37°C, or 30°C (if the plasmid was temperature sensitive). Plasmids were extracted from these cells using a QIAprep® Spin Miniprep Kit (Qiagen) according to the manufacturer's guidelines. Following Qiagen's recommendations, low copy number plasmids (pTOF24 and its derivatives) were isolated from 5 ml of overnight culture as opposed to 1 ml of culture for high copy number plasmids. Plasmid DNA was stored in double distilled sterile water at -20°C.

2.2.2.2 Genomic DNA extraction

2.2.2.2.1 Boiled cell method

For the majority of polymerase chain reaction (PCR) applications, boiling cells produced genomic DNA of adequate quality to act as a template. A single colony was picked from a plate and cells were boiled by suspending in 30 µl of sterile water and heating to 99.9°C for 10 min before centrifuging for 2 minutes in a table-top microcentrifuge. 2 µl of the resulting supernatant was used as a template for a 25 µl PCR.

2.2.2.2.2 Kit method Genomic DNA extraction for PCR (Promega kit)

When the DNA amplified by PCR was needed for cloning, the genomic DNA used as a template was isolated using a Wizard® Genomic DNA purification kit (Promega) according to the manufacturer's instructions. This genomic DNA was rehydrated at 65°C for 1 hour or at 4°C overnight in double distilled water and stored at -20°C.

2.2.2.3 Plasmid DNA preparation for PCR (QIAGEN kit)

The QIAGEN QIAprep® Spin Miniprep Kit was used following the manufacturer's instructions. DNA was eluted in 30 µl of MQ-water and stored at -20°C.

2.2.2.4 Polymerase chain reactions (PCR)

2.2.2.4.1 Normal PCR reaction

FinnzymesPhusion® High-Fidelity DNA polymerase Cat.No. F-530 is a highly processive and extremely accurate DNA polymerase (with an error rate of 4.4 X 10^{-7}). Because of these qualities it was chosen as the polymerase for all PCRs when the product was required for cloning or as a template for labelling with ^{32}P. When PCRs were carried out for checking fragment sizes, for example following PMGR, PromegaGoTaq® Flexi DNA Polymerase, Cat. No. M829, was used instead.

Reactions were carried out using a PeqLab Biotechnologie GmbH peqSTAR 96 Universal Gradient PCR machine. Primer annealing temperatures were dependent on primer sequence and were altered accordingly. Typically, the annealing temperature increases as the primer length and %GC content increase. The extension time was determined by the polymerase of choice and the length of the template to be amplified. Typically, the longer the template, the longer the extension time needed. Phusion® High-Fidelity DNA Polymerase can extend a 1Kb fragment in 30 seconds, while GoTaq® Flexi DNA Polymerase takes twice as long, requiring 1 minute per 1Kb of template.

A typical cycle programme was as follows:

Program	**Temperature**	**Time**	
Initial template denaturation	95°C	5 min	
Template denaturation	95°C	30 sec	35 cycles
Primer annealing	50-65°C	30 sec	
Extension	72°C	30 sec-3 min	
Final extension	72°C	10 min	
Storage	8°C	indefinite	

2.2.2.4.2 Crossover PCR

Crossover PCR (Fig. 2.2) was used in order to join two separate fragments of DNA without the need for restriction and ligation. For this technique, four primers were required, two for amplifying the first DNA fragment and two for amplifying the second DNA fragment. Internal primers, the reverse primer for the first DNA fragment and the forward primer for the second DNA fragment, were designed to have 20-25 base pairs of homology to each other. Initially, the two DNA fragments were amplified separately, creating two products that at one extremity contained 20-25 base pairs of homology to each other. Finally, to join the two DNA products, a crossover PCR was set up where the two products from the initial PCRs were used as template. During melting and cooling of the templates, the region of homology would bring the two DNA fragments together to create a single new fragment for amplification using the external primers, the forward primer for the first DNA fragment and the reverse primer for the second DNA fragment. The resulting PCR product would be a fusion of the two DNA fragments of interest.

2.2.2.5 Sequencing of DNA (Applied Biosystems kit)

Sequencing was carried out using the Applied BiosystemsBigDye® Terminator v3.1 Cycle-Sequencing Kit following the manufacturer's instructions. Template DNA consisted of a purified PCR product. Sequencing reactions were analysed by the Genepool Sequencing Service (now a part of Edinburgh Genomics), Ashworth Laboratories, University of Edinburgh, using an ABI PRISM® 3100-Avant Genetic Analyser.

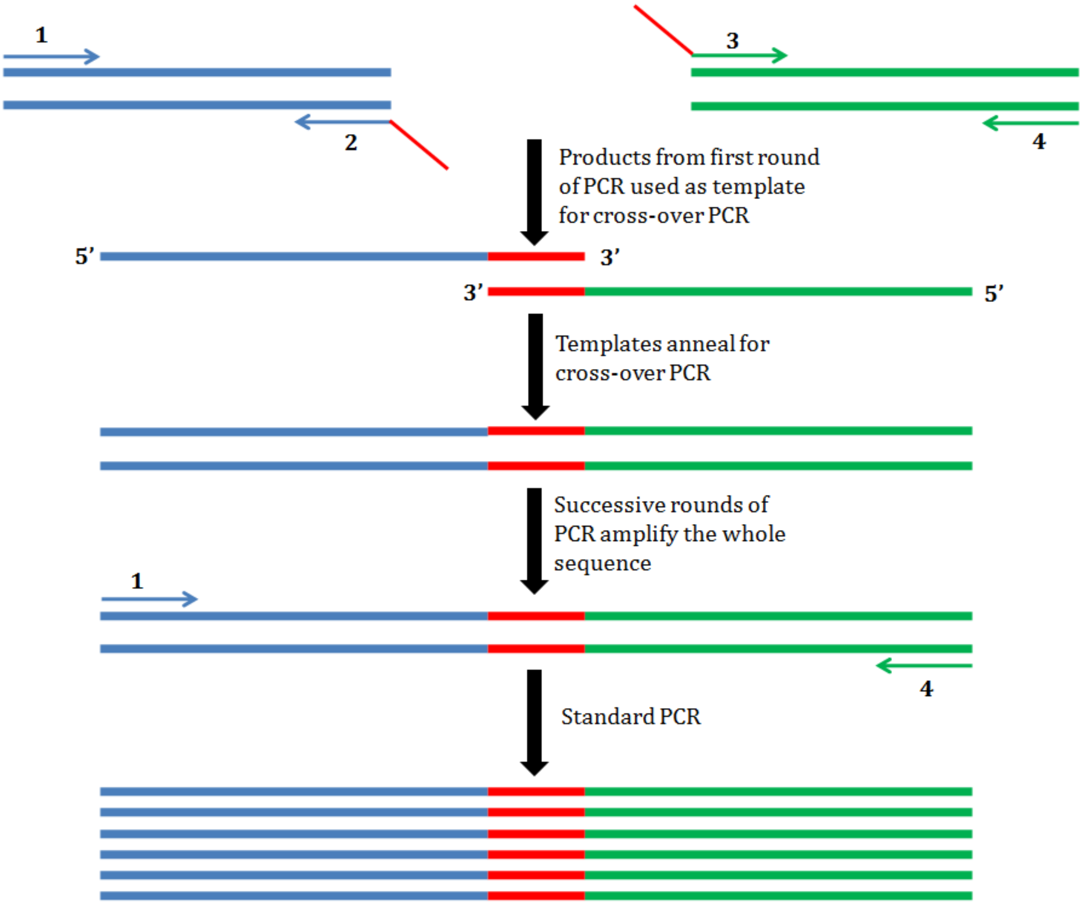

1
2
3
4
Products from first round of PCR used as template for cross-over PCR
5'
3'
3'
5'
Templates anneal for cross-over PCR
Successive rounds of PCR amplify the whole sequence
1
4
Standard PCR

Figure 2.2. Cross-over PCR. Initially, the two DNA fragments (shown in blue and green) are amplified separately using primers 1-4. Primers 2 and 3 are designed to have a 20-25 bp region of homology to each other. In the crossover PCR, the products from the first PCR reaction are used as template. When these melt and re-anneal, the region of homology between them will bring both strands together while 3' ends of each strand at that region will serve as a site for primer extension. This will create a new, single, template for amplification using primers 1 and 4.

2.2.2.6 PCR product purification for cloning (QIAGEN kit)

The QIAGEN QIAquick® PCR purification kit (Cat. No. 28104) or the QIAquick® Gel Extraction Kit (Cat. No. 28704) was used for cleaning DNA fragments for cloning. The manufacturer's instructions were followed. DNA was eluted in 30 µl of MQ-water and stored at -20°C.

2.2.2.7 Restriction digestion of PCR purified DNA

Restriction enzymes and buffers were obtained from New England Biolabs (NEB). 30 µl of either PCR purified DNA or plasmid DNA were digested following the manufacturer's instructions. Samples were incubated at the optimum digestion temperature (37°C for all enzymes used in this work) for 2-4 hours or overnight.

2.2.2.8 Ligation of DNA fragments

To ligate fragments of DNA, the New England Biolabs Quick Ligation™ Kit was used following the manufacturer's instructions. Ligation reactions were performed in a total of 20 µl per reaction volume.

2.2.2.9 Agarose gel electrophoresis of PCR products or plasmid DNA

DNA fragments from PCR reactions or digested plasmids were separated on a 1% (w/v) agarose gel. The appropriate amount of agarose (MELFORD agarose electrophoresis grade Cat. No. MB1200) was dissolved in 1 X TAE and allowed to cool to 55°C. Safeview (NBS Biologicals Ltd, SafeView, Cat. No. NBS-SV1; 5 µl in 100 ml of liquid agarose) was added to visualise the DNA under UV light. Gels were run from 90-140 V for up to 2 hours and DNA was visualised using a UV box (BioRad). The size of fragments was checked using DNA ladders (New England Biolabs, 1 Kb DNA Ladder Cat. No. N3232, 100 bp DNA Ladder Cat. No. N3231). When necessary, DNA was quantified using a Nanodrop (ND-1000 v3.5).

2.2.2.10 Agarose gel electrophoresis of chromosomal DNA

2.2.2.10.1 Preparing *E. coli* DNA in agarose plugs (for 2-D gel electrophoresis)

An overnight culture of the desired strain was diluted 50 times in L-broth and allowed to grow at 37°C with shaking to an OD_{600nm} of 0.4 – 0.8. Cultures were harvested and washed twice in TEN buffer by spinning and re-suspending. After the final wash, cells were re-suspended in TEN buffer to give an OD_{600nm} of 80. The cell suspension was briefly warmed to 37°C and then mixed with an equal volume of agarose (Invitrogen UltraPure™ LMP Agarose Cat. no. 16520-050) at 0.8% in 1X TEN for 2-D gel samples, giving a final agarose concentration of 0.4%. The cell and agarose mix was then poured into plug moulds (Bio-Rad Cat. no. 1703706), and allowed to set at 4°C for 30 minutes. Once set, all 10 plugs were extruded from the moulds into a Falcon tube and 1 ml of NDS + proteinase K (1 mg ml^{-1}) was added per plug of sample. Tubes were left overnight at 37°C with gentle rocking. The buffer was replaced with fresh 1 ml NDS + proteinase K (1 mg ml^{-1}) per plug of sample for a second night of incubation at 37°C with gentle rocking. Plugs were stored at 4°C in fresh NDS without added proteinase K.

2.2.2.10.2 Digestion of DNA set in agarose plugs

In order to remove any remaining proteinase K and NDS solution, which would inhibit the restriction enzyme, plugs were washed thoroughly in 1.5 ml per plug of 1 X restriction buffer (without BSA and DTT) for 6 hours, replacing old 1 X restriction buffer every hour with fresh buffer. Once washed, the plugs were digested in 500 µl 1 X restriction buffer + BSA, using 500 U of enzyme per plug.

Digestions were left overnight at 37°C with gentle rocking. Plugs were briefly cooled to 4°C before quickly washing them in 1.5 ml TE just prior to loading onto the gels.

2.2.2.10.3 Native two dimensional agarose gel electrophoresis

Native two dimensional agarose gel electrophoresis (2-D agarose gel) was used to separate branched DNA structures from their linear counterpart. Initially, the chromosomal DNA was separated in conditions that minimised the structural differences between fragments but maximised the difference in size. Thereafter, the lane containing the DNA was sliced out and turned 90°, placing the wells to the left, and run in a second dimension in the presence of 0.3 μg ml^{-1} EtBr, so as to maximise the differences in shape between different structures. For separation in the first dimension, chromosomal DNA was prepared in agarose plugs and digested with the relevant restriction enzyme. Plugs were then attached to the gel comb, making sure to leave at least 1 lane gap between samples, using 10 μl of liquid agarose (0.4% (w/v) MELFORD agarose electrophoresis grade (Cat. No. MB1200) in 1 X TBE). This was allowed to set at 4°C for 30 minutes. The remainder of the agarose was then poured around the plugs and allowed to set at 4°C for 30-60 minutes. 2.5 μg of NEB 1 Kb DNA Ladder (Cat. No. N3232) was loaded onto the gel, which was run in 1 X TBE at 1 V/cm and 4°C for 26 hours. The Marker lane was cut out of the gel and stained with 0.5 μg ml^{-1} EtBr for viewing. Intermediates run at a higher molecular weight than their linear counterparts and therefore the lane in which the DNA was run was sliced out and cut 1 cm below where the linear species of the DNA of interest was expected to have migrated to. This gel slice was turned 90°, placing the wells to the left, and placed into the casting tray for the second dimension. The second dimension agarose (1% (w/v) MELFORD agarose electrophoresis grade Cat. No. MB1200 in 1 X TBE + 0.3 µg ml^{-1}EtBr) was then poured over the first dimension slices so as to cover them completely. The whole was allowed to set at 4°C for 30-60 minutes and was then run for 10 hours in 1 X TBE + 0.3 µg ml^{-1} EtBr, at 6 V/cm and 4°C. During this time the running buffer was re-circulated using a pump so as to maintain a constant concentration of EtBr across the gel and tank. The gel was then exposed to UV light to view the DNA arcs and then processed for Southern blotting.

2.2.2.11 Southern blotting of DNA

2.2.2.11.1 Alkaline transfer of DNA to a positively charged nylon membrane

Once viewed under UV light, the gel was washed in depurination solution for 30 minutes to fragment the DNA for easier transfer (pulsed-field gels were not depurinated as this process was previously shown to reduce the transfer of DNA separated using this technique). The gel was then washed in alkaline transfer buffer for 1 hour and the transfer stack was set up and the DNA was allowed to transfer overnight (Fig. 2.3). After transfer, the nylon membrane was allowed to dry, completely, at room temperature, and then was UV cross-linked, using a Stratagene UV Stratalinker™ 1800. Membranes obtained from Pulsed-field gel electrophoresis were crosslinked using 1200 Jm^{-2} while membranes derived from 2-D agarose gel electrophoresis were cross-linked using 1000 Jm^{-2}. Once cross-linked, membranes were washed in distilled water, dried, and sealed in a plastic envelope between two pieces of Whatmann paper for storing at 4°C.

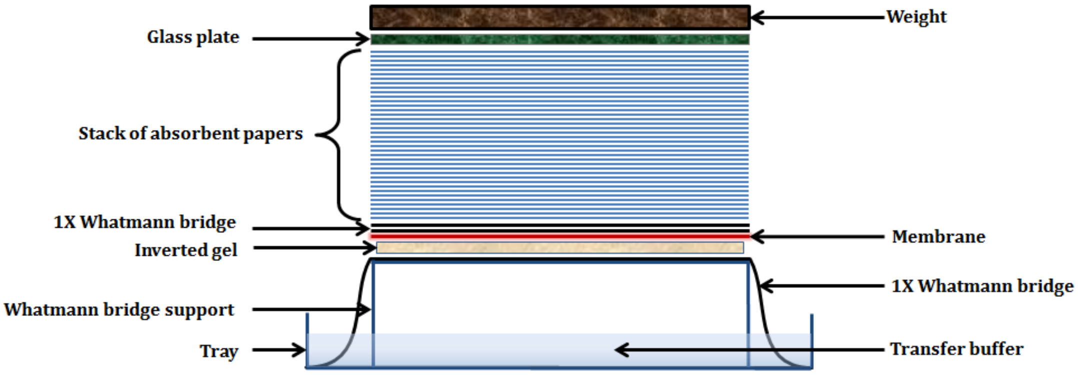


Figure 2.3. Southern blot transfer stack. A glass plate was placed over a plastic container filled with transfer buffer and a strip of Whatmann paper was placed over this like a bridge, making sure both ends of the strip came into contact with the buffer. A piece of Whatmann paper, the size of the gel, was placed on top of the Whatmann bridge and the inverted gel was stacked next, followed by the membrane, pre-wetted in transfer buffer for 5 minutes, and two more pieces of Whatmann paper. Finally a stack of absorbent paper was placed on top followed by another glass plate. The whole stack was placed under a 1 Kg weight. The DNA transferred from the gel to the membrane overnight by capillary action.

2.2.2.11.2 Labelling of a probe with ^{32}P

A Stratagene Prime-It II random primer labelling kit (Cat. No. 300385) was used to radioactively label probes with ^{32}P α-dATP following the manufacturer's guidance. Templates for the reaction were obtained by PCR from chromosomal DNA as previously described. Labelling reactions were allowed to proceed at 37°C for 30 minutes. To clean probes, GE Healthcare illustraMicrospin™ G-25 columns were used (Ca. No. 27-5325-01) following the manufacturer's guidance.

2.2.2.11.3 Detection of a ^{32}P labelled probe

The cross-linked membrane was placed into a hybridisation bottle and re-hydrated in 2 X SSC. 10-15 ml (depending on the size of the membrane) of Church-Gilbert buffer, pre-warmed to 65°C, was added and the bottle was put into a hybridisation oven at 65°C for 2-6 hours to pre-hybridise. Once pre-hybridisation was over, 20 µl or 100 µl of probe were diluted with 100 µl or 20 µl sterile MQ-water, respectively, and boiled in a thermocycler for 5 minutes. In the meantime, the 10 ml of prehybridisation Church-Gilbert buffer were replaced with 10 ml of fresh, pre-warmed at 65°C, Church-Gilbert buffer. Once fully denatured, the probe was immediately added to the hybridisation bottle, which was placed back into the oven to hybridise overnight. The hybridisation buffer + probe was removed and a 15 minute wash in 200 ml of washing solution was performed to remove excess, unbound, probe, followed by a 30 minute wash in 200 ml of stringency washing solution, to remove unspecific hybridisation of the probe. The membrane was then wrapped in cling film and exposed to GE healthcare storage phosphor screens (Cat. No. 63-0034-86 and 63-0034-79). Exposure lasted between 1 and 7 days. Screens were scanned using a Molecular Dynamics Storm 860 phosphor imager scanner. When needed, membranes were stripped by washing for 1 hour in 50 ml of stripping buffer and then washed for 30 minutes in 200 ml of stringency washing solution in order to wash away all the formamide. Membranes were re-exposed to check for the degree of stripping, when needed, and pre-hybridised again for additional probing.

2.2.3 Genetic assays

2.2.3.1 A genetic assay for detecting instabilities of the length of a CTG•CAG repeat array.

CTG•CAG TNR arrays were inserted into the bacterial chromosome in two different orientations relative to the origin of replication (Zahra et al., 2007). When an array was inserted in such an orientation, that CTG repeats were present on the leading strand of the duplex DNA, and CAG repeats were therefore present on the lagging strand, that was termed the CTG orientation. Analysis of the instability of TNR arrays in the *E. coli* chromosome was conducted using a genetic assay (Figure 2.4A) designed in the Leach laboratory by Rabaab Zahra (Zahra et al., 2007).

2.2.3.2 Instability assay

Strains on which the instability assay was to be performed were streaked out from the –80°C stock onto LB plates to produce single colonies. Sixty single colonies were then selected for each strain and each colony was inoculated into 5 ml of L-broth and grown overnight in a shaking incubator at 37°C. The instability of the TNR array at this stage of growth was assessed. A 10^{-6} dilution of the overnight cultures was produced in L broth, and 100µl of this dilution was plated onto LB agar plates. Plates were incubated overnight in a 37°C incubator to allow growth of the colonies. Eight colonies were then selected from each plate and PCR carried out to determine the length of the TNR array in each. The PCR using Ex-Test primers was designed to amplify the region of the *lacZ* gene in which the TNR array was inserted (Table 2.5). For each PCR, the reverse primer Ex-Test R and a modified version of the forward primer Ex-Test F which was labelled with a fluorescent 5'-FAM (Fam-Ex-Test F), were used.

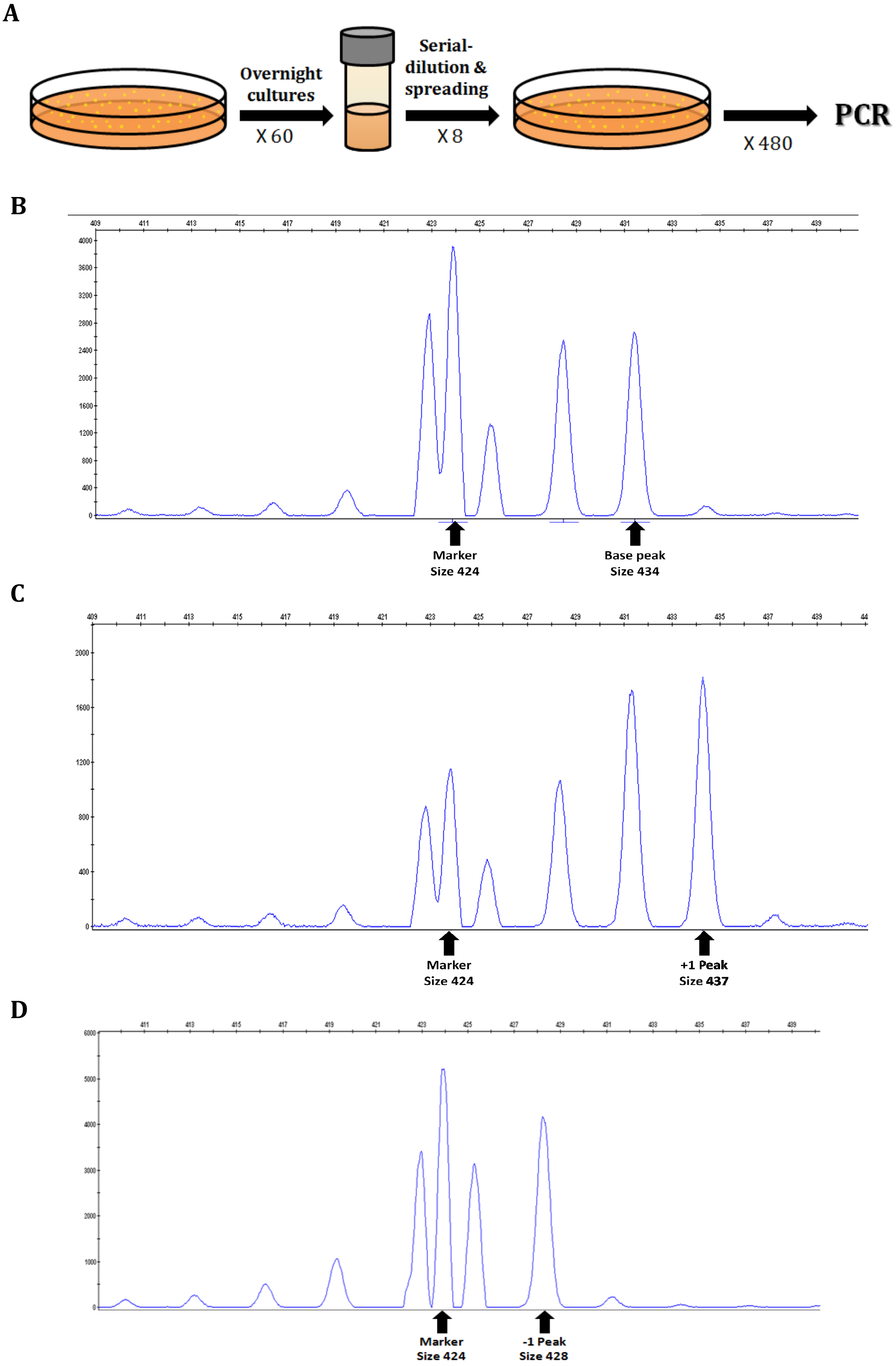

A
Overnight cultures
X 60
Serial-dilution & spreading
X 8
X 480
PCR
B
409
411
413
415
417
419
421
423
425
427
429
431
433
435
437
439
4000
3600
3200
2800
2400
2000
1600
1200
800
400
0
Marker
Size 424
Base peak
Size 434
C
2000
1600
1200
800
400
0
Marker
Size 424
+1 Peak
Size 437
D
6000
5000
4000
3000
2000
1000
0
Marker
Size 424
-1 Peak
Size 428

Figure 2.4. A schematic representation of the TNR instability assay and GeneMapper® analysis. A. The strain to be tested was streaked out onto an LB plate (without any selection) from -80°C stock culture. Sixty overnight cultures were setup and grown at 37°C for 14-16 hours. 10^{-6} dilutions were made and plated onto LB plates from each overnight culture. Eight colonies from each plate (altogether 480 samples per strain to be tested) were then tested for repeat length by PCR. B. The peak represents the parental length of $(CTG)_{98}$ TNR array, 434 bp ((3bp x 98) + 140 bp = 434 bp). C-D. The region amplified outside the repeat array is 140 bp. Outputs displaying both +1 and -1 TNR unit peaks were considered to be result of addition and deletion events, respectively, occurring during the growth of the colony on the agar plate.

2.2.3.3 GeneMapper® analysis

Products of the PCR reaction were diluted 1:25 in sterile water. 1 ml of HiDi™ Formamide (Applied Biosystems®, Cat. No. 4311320) was then mixed with 5 µl of GeneScan™-1200 LIZ™ size standard (Applied Biosystems®, Cat. No. 4379950), and 9 µl of this solution was mixed with 1 µl of the diluted PCR products in barcoded 96 well plates. Plates were sent to The Gene Pool (now a part of the Edinburgh of Genomics), The University of Edinburgh, where they were run on an ABI 3730 Genetic Analyzer (Applied Biosystems®, Cat. No. 3730S), which separates DNA fragments through capillary electrophoresis. Data produced were analysed using the software GeneMapper v3.7®. The data relates the time taken for the DNA fragments in the PCR products to pass a laser (x-axis). The software compares this time to that of the size standard and converts time to size, allowing sizing of the DNA fragments. Using this method the size of TNR arrays could be determined to ±5bp. Figure 2.4(B) shows examples of GeneMapper results obtained by this method.

Figure 2.4B shows a base peak surrounded by a number of 'stutter' peaks, a typical output of the GeneMapper program. These 'stutter' peaks are *in vitro* artefacts from the PCR reaction and not real instability products. The size of the TNR array can be calculated from the fragment size detected by subtracting 140 bp (the size of the region outside the repeat array amplified by the Ex-Test primers) and dividing by 3, to give the number of TNR units. Colonies producing results such as displayed in Figure 2.4B were considered to have not undergone an instability event, as the base length of the TNR array was indicated. Colonies in which expansion of a single repeat unit has occurred in the overnight culture, extended the size of the TNR array by 3 bp (Figure 2.4C) while those in which deletion by a single TNR unit has occurred in the overnight culture has been shown in Figure 2.4D.

2.2.3.4 Statistical methods

Graphs and statistics were produced using the software Microsoft Excel 2010. The proportion of instability events resulting in a deletion of the TNR array, and the proportion of instability events resulting in an expansion of the array, were calculated for each of 60 overnight cultures produced from a particular *E. coli* strain being investigated. The proportion of instability events calculated for each culture was treated as sample data, the mean of the samples was then taken to determine the population mean for the strain.

The instability assay data was considered as a binary response, whether an instability event had occurred or not, and was scored accordingly: 1 = instability event, 0 = no event. A logistic regression model was applied to the data to compare the instability in mutant strains to the corresponding wild type strain for that orientation of TNR array. Logistic regression is used in situations in which the dependent variable produces binary responses - in this case whether an instability event had occurred (=1) or not (=0). The proportion of occurrence(s) of instability from 8 replicates of each (among 60 overnight cultures) sample culture has been calculated. For example, if 1 colony from randomly selected 8 colonies (from the plate after serial dilution in figure 2.4A) is found to have instability (either a single unit TNR insertion or deletion), then the proportion of instability, P = (1+0+0+0+0+0+0+0)/8 = 0.125. Likewise, the proportion of instability is calculated from each of the 60 initial overnight cultures and the "Frequency of instability" is calculated as the mean proportion of using the following equation:

$$\frac{\sum_{i=0}^{N} P}{N}$$

where, N is the sample size

Standard deviation within the population was then calculated, and from that the standard error of the mean was calculated using the formula:

$$\text{Standard Error of Mean} = \frac{StandardDeviation}{\sqrt{N}}$$

Where, N is the sample size. Population means were plotted for each strain investigated and error bars drawn showing the standard error of the mean for each.

The non-parametric Kruskal-Wallis test applied to test for differences in instabilities produced between strains. The Kruskal-Wallis test is a non-parametric equivalent of the parametric test ANOVA (analysis of variants) and is used to test whether population medians between different groups are equivalent or not. Unlike ANOVA analysis it does not assume that the populations are normally distributed and does not require a large sample size. The significance of the p-value has been tested at 95% confidence level. If the p-value becomes greater than 0.05, the null hypothesis (H_0 = There is no difference between the data sets, meaning that, the strains compared are phenotypically same) is accepted and vice versa. However, there is always a chance of the occurrence of a type II error (failing to reject an invalid null hypothesis) while comparing two strains Kruskal-Wallis test does not make any normality assumption about the population distribution (otherwise parametric ANOVA would have been a better choice). However, the statistical analysis has not been corrected for multiple testing – testing whether the significant results appeared by chance.

2.2.3.5 Fluctuation assay:

36 overnight cultures were grown from isolated colonies at 37°C in low-salt LB broth (0.5 g l^{-1} NaCl). Recombinant colonies were selected on low-salt LB agar plates containing 35 mg ml^{-1} of zeocin. Fluctuation test tables was used (Spell

and Jinks-Robertson, 2004) to calculate the rate of recombination (events/cell/generation) and 95% confidence intervals. An example of the calculation table is shown in Appendix D.

2.2.4 Calculating the relative efficiency of the MMR and of recombination:

The relative efficiency of the DNA mismatch repair system has been calculated in the following the way:

The relative efficiency of the MMR system in the absence of specific GATC site(s)

$$E_{(\Delta GATC)} = C_{(\Delta GATC)} \,/\, C_{(+)}$$

where the frequency of corrected mismatches in the presence of all GATC sites

$$C_{(+)} = F_{(-)} - F_{(+)}$$

where the $F_{(+)}$ represents the frequency of mutations in the presence of MMR and $F_{(-)}$the frequency of mutations in the absence of MMR and where the frequency of corrected mismatches in the absence of specific GATC sites

$$C_{(\Delta GATC)} = F_{(-)} - F_{(\Delta GATC)}$$

where $F_{(\Delta GATC)}$represents the frequency of mutations in the absence specific GATC sites.

The relative efficiency of MMR in a wild type situation ($E_{(+)}$) wasset to 1

$$E_{(+)} = C_{(+)} / C_{(+)} = 1$$

Similarly, the relative rate of recombination of the zeocin cassette has also been calculated:

The relative efficiency of recombination of the zeocin cassette in the absence of specific GATC site(s)

$$\boldsymbol{Z}_{(\Delta\mathrm{GATC})} = \boldsymbol{R}_{(\Delta\mathrm{GATC})} \,/\, \boldsymbol{R}_{(+)}$$

where rate of recombination in wildtype that accompanies mismatch repair events

$$\boldsymbol{R}_{(+)} = \boldsymbol{Q}_{(+)} - \boldsymbol{Q}_{(-)}$$

where $\boldsymbol{Q}_{(+)}$ represents the rate of recombination in the wild type background in the presence of TNR array and MMR and $\boldsymbol{Q}_{(-)}$ the rate of recombination in the presence of TNR array, but in the absence of MMR and where the rate of recombination (that accompanies mismatch repair events) in the absence of specific GATC sites (in the presence of TNR array)

$$\boldsymbol{R}_{(\Delta\mathrm{GATC})} = \boldsymbol{Q}_{(\Delta\mathrm{GATC})}-\boldsymbol{Q}_{(-)}$$

where $\boldsymbol{Q}_{(\Delta\mathrm{GATC})}$ represents the rate of recombination in the absence specific GATC sites (in the presence of TNR array).

The relative efficiency of recombination in a wild type situation ($Z_{(\mathrm{wt})}$) was set to 1.

$$\boldsymbol{Z}_{(\mathrm{wt})} = \boldsymbol{R}_{(+)} / \boldsymbol{R}_{(+)} = \mathbf{1}$$

2.2.5 An *in silico* approach to generate different artificial sequences

In an approach to generate artificial sequences, the Markovian model has been applied. This model assumes that a sequence is produced by a process that chooses any of the four nucleotides in the sequence and the probability of choosing any nucleotide at a particular position depends on the nucleotide chosen for the previous position. For example, if "C" has already been chosen at the previous position, the probability of choosing any of the four nucleotides at the current position depends on the predetermined probability distribution p_{CA}, p_{CC}, p_{CG}, and p_{CT} which is the probability of choosing nucleotides "A", "C", "G" and "T" respectively after "C". These four parameters are called the "transition probabilities" that serves the basic purpose of the Markov model along with four initial probabilities (p_A, p_C, p_G, and p_T). This model is also called the first order Markov chain as the current state actually depends on another single state (the previous state). Thus, for example, a sequence "ATTCGCA" will depend on the initial probability p_A for choosing "A" at the first position and the following transition probabilities: p_{AT}, p_{TT}, p_{TC}, p_{CG}, p_{GC} and p_{CA} respectively (Figure 2.5).

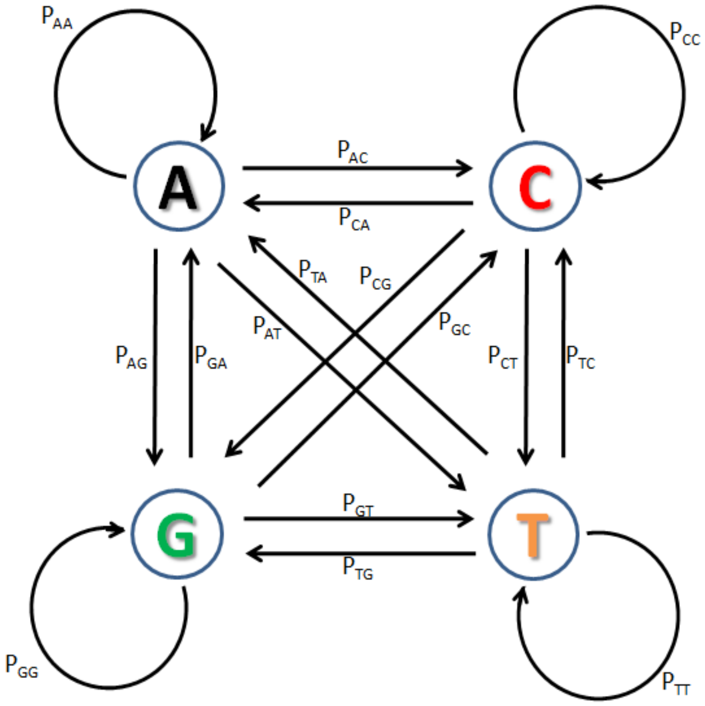


Figure 2.5. A schematic representation of transition probabilities for choosing a nucleotide at a particular position in a sequence. The respective transition probabilities (P_{AA} , P_{AC} , P_{AG} , P_{AT}, ... , P_{TT}) are shown along the arrows directing the transition.

Three model sequences were generated –

1. Random sequence (Rand).
2. Dinucleotide based sequence (DN).
3. Dicodon based sequence (DC).

2.2.5.1 The random sequence (Rand)

Following the Markov sequence model, the initial probabilities where set in such a way that individual nucleotides have the equal probability to be chosen for the first position. Thus, p_A = p_C= p_G = p_T = 0.25. After occupying the first position by a single nucleotide, the next position has been chosen for any of the four nucleotides as randomly as possible. Trying so, the transition probabilities had been defined as:

$p_{AA} = p_{AC} = p_{AG} = p_{AT} = 0.25$ (thus, $p_{AA} + p_{AC} + p_{AG} + p_{AT} = 1$)

$p_{CA} = p_{CC} = p_{CG} = p_{CT} = 0.25$ (thus, $p_{CA} + p_{CC} + p_{CG} + p_{CT} = 1$)

$p_{GA} = p_{GC} = p_{GG} = p_{GT} = 0.25$ (thus, $p_{GA} + p_{GC} + p_{GG} + p_{GT} = 1$)

$p_{TA} = p_{TC} = p_{TG} = p_{TT} = 0.25$ (thus, $p_{TA} + p_{TC} + p_{TG} + p_{TT} = 1$)

Ten random sequences were generated with the same length of the *E. coli* genome, which is 4639675 bp. In addition, three particular transition probabilities – p_{GA}, p_{AT} and p_{TC} had been modified and the process was optimized in a trial and error basis so that the average number of GATC attained the range of 19120±1% to generate another ten random sequences.

2.2.5.2 The dinucleotide sequence (DN)

One biological constraint had been introduced in this method. For this attempt, the first nucleotide had been chosen according to the overall composition of the nucleotides in the whole *E. coli* chromosome. The initial probabilities had been calculated from the proportion of each nucleotide in the *E. coli* genome using package “seqinr” in statistical package R and shown in Figure 2.6.

$p_A = 0.246187$

$p_C = 0.254232$

$p_G = 0.253665$

$p_T = 0.245916$

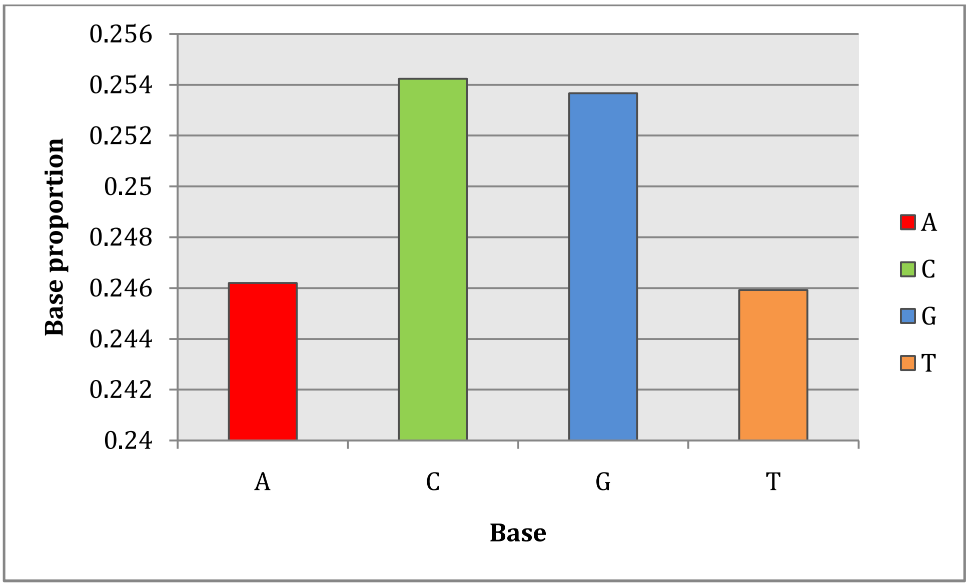


Figure 2.6. Proportion of each nucleotide in the *E. coli* genome based on which the initial probabilities had been calculated. The red bar represents the proportion of "Adenine", the green bar represents the proportion of "guanine", the blue bar represents the proportion of "cytosine" and the orange bar represents that for "thymine" that are plotted on a scale showing the respective base proportions along the *y*-axis.

The next task was to define the (4 × 4) transition matrix with more essence of biology. As about 85% of the *E. coli* genome corresponded to coding sequences, the transition matrix had been built on the composition of the coding sequence of *E. coli*. The frequencies of all the 16 possible dinucleotides had been calculated from the coding sequence concatenated into a single sequence (Figure 2.7) using Perl script, statistical package R and MATLAB®.

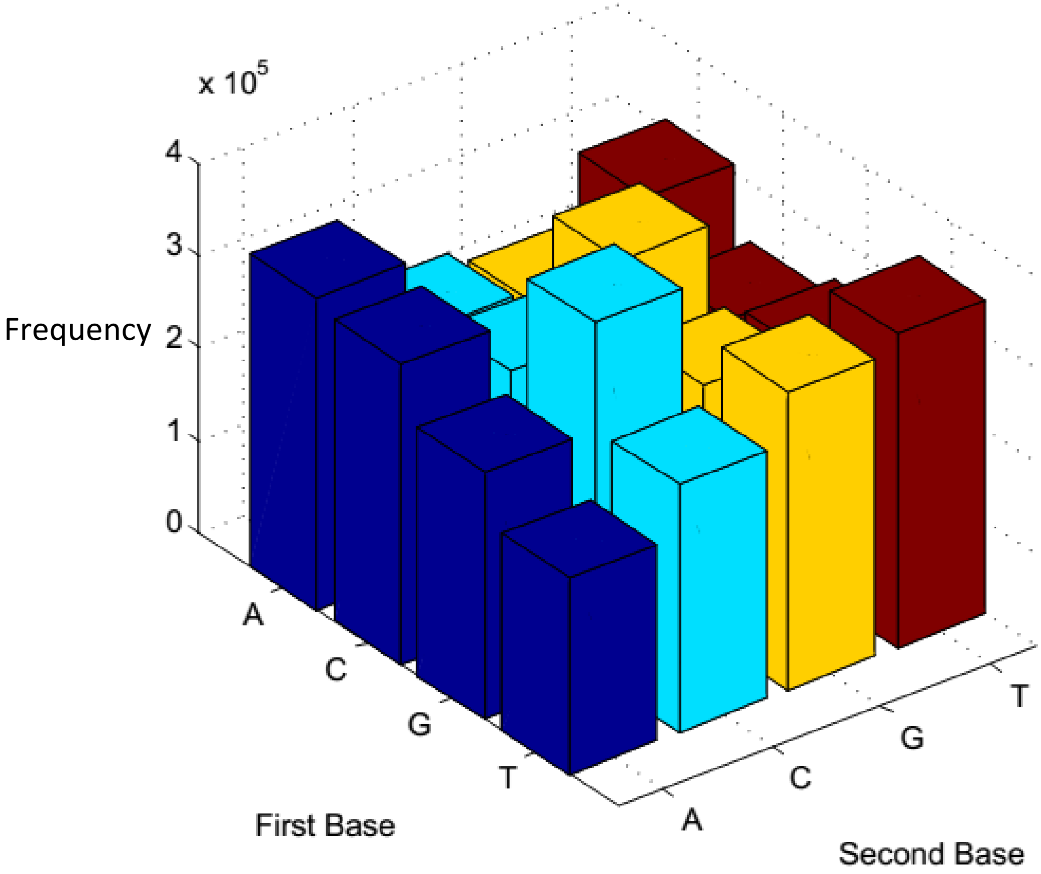


Figure 2.7. The frequency distribution of 16 different dinucleotides in the *E. coli* genome presented in a side by side fashion. Statistical software MATLAB® has been used to count and plot this frequency distribution.

The names of sequences generated from this approach have been coined as the "dinucleotide sequences". One drawback of this approach was that the average number of GATC motifs per sequence generated became about 16000, which was far less than the total number of GATC motifs in *E. coli* genome. Yet again, three particular transition probabilities – p_{GA}, p_{AT} and p_{TC} have been modified in an attempt to bring the number of GATC motifs to similar level (19120±1%) to that in the *E. coli* genome. The process had been optimized by trial and error basis. Ten sequences of 4639675 bp length had been generated by this approach.

2.2.5.3 The dicodon sequences (DC)

Further attempts were made to generate more biologically relevant artificial sequences. As mentioned before, the *E. coli* genome is composed of about 85% of coding sequence and the codon usage has been taken into account to generate artificial sequences as has been done in Section 2.2.5.2. The principle remained the same as for the dinucleotide sequences; only the codons were taken into consideration in place of nucleotides in such a way that the total length of the sequence remained 4639675 bp. In this approach, which had been called the dicodon approach, the first three positions of the sequence were occupied by a codon with a predefined initial probability (Figure 2.8). The initial probabilities have been defined by the codon frequency in the coding region of the *E. coli* genome and calculated from the proportion of each codon in the coding region of the *E. coli* genome (Figure 2.8).

Like the dinucleotide approach, the hexamers in the coding region of the *E. coli* genome have been calculated using the statistical package R from the concatenated coding sequences (created using in-lab Perl script). An 8 × 8 matrix has been created to define the transition matrix. This approach has been named the "dicodon approach" because of the usage of hexamers (two codons). At first ten artificial sequences were generated with the length of 4639675 bp where the average number of GATC motifs was over 2300 per genome. Yet again, 16 transition probabilities were modified in an attempt to bring the number of GATC motifs to a similar level to that of the *E. coli* genome (19120). After several trial and errors, ten artificial sequences were generated in this approach with an average of 19120±1% GATC motifs per sequence generated.

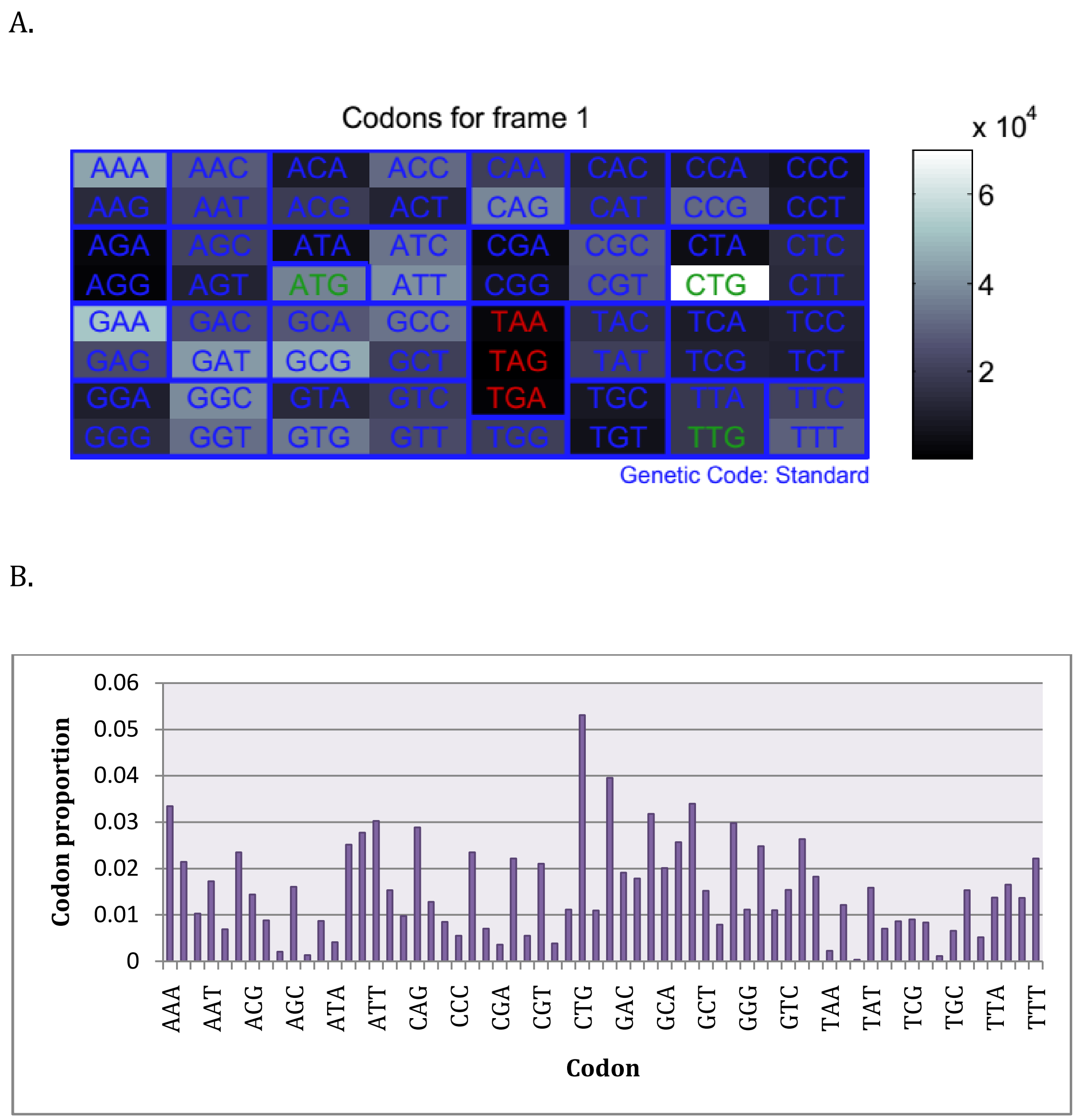


Figure 2.8. The codon usage of the coding regions of the *E. coli* genome. (A) The relative abundances have been shown with the colour gradient using statistical software MATLAB and (B) the weighted values (as the codon proportion) have been plotted for each codon.

Chapter 3

The DNA mismatch repair machinery moves towards the replication fork in its search for a hemimethylated GATC site

3.1 Introduction

The *E. coli* DNA mismatch repair system shows bidirectionality in *in vitro* experiments (Cooper et al., 1993). However, Blackwood and collaborators found that MMR-stimulated recombination at a 275 bp tandem repeat only occurs when the tandem repeat is placed on the origin proximal side of a system that generates substrate for the MMR system suggesting that MMR might be directional *in vivo* (Blackwood et al., 2010). In this study, I have investigated the directionality of MMR in *E. coli*.

Genomes of all organisms are scattered with simple repeats and tandem repeats occur in the form of iterations of repeat units ranging from a single base pair to even thousands of base pairs (Ellegren, 2004). Perfect or near-perfect tandem iterations of short sequence motifs (mono-, di-, tri- and tetranucleotide repeats) are extremely common in eukaryotic genomes (Ellegren, 2004). These short repeats are popularly known as microsatellites which are mainly composed of 1-6 nucleotide long simple sequence repeats (SSR) (Miah et al., 2013). Microsatellites are highly polymorphic and the length instability at a single unit level is very prominent in MMR deficient cells (Blackwood et al., 2010). In addition, the MMR system is known for repairing small extrahelical loop-outs of 3-4 nucleotides created due to replication slippage along with single base mismatches (Jiricny, 1998). Therefore, a microsatellite can be used as a source of frequent mismatches at a defined locus for studying MMR *in vivo*. In this study, a 98 unit (294 base-pairs) long CTG•CAG repeated array has been used where the CTG repeats are located on the leading strand. This trinucleotide repeat array is inserted in the *lacZ* gene in the *E. coli* genome. In addition, the variation of length of the repeat array at the level of a single unit has been defined as "instability" which is considered as a quantitative phenotype to detect the efficiency of the DNA mismatch repair system in this study. This study focused on the usage of GATC sites by the DNA MMR system on both sides of the trinucleotide repeat array in order to understand the potential usage of GATC motifs on both sides of a mismatch occurred at an anonymous locus in the genome. The significance of the difference of instabilities among different bacterial populations has been calculated using Kruskal-Wallis analysis (described in Chapter 2).

Since the early 1970s, many studies have been carried out to elucidate the molecular mechanism of DNA mismatch repair in different organisms (Friedberg et al., 2006). The current established molecular mechanism of the DNA mismatch repair system has mostly been established from *in vitro* studies. Interestingly, in *E. coli* the methylation state of the DNA has been found to affect the strand discrimination as the unmethylated nascent DNA strand undergoes almost 100% repair of a mismatch while the fully methylated parent DNA strand does not (Wagner and Meselson, 1976). The methylation of a DNA strand is mediated by the DNA adenine methylase (Dam) enzyme at the adenine base of a GATC motif (Geier and Modrich, 1979; Marinus and Morris, 1974). Until the late 1980s, a GATC motif was considered as a beacon for strand discrimination by the DNA mismatch repair system in *E. coli*. However, the molecular mechanism of strand discrimination became clear after the finding of an enzyme encoded by the *mutH* gene (Welsh et al., 1987). MutH protein is a cryptic endonuclease, which utilises the methylation state of a GATC motif to incise only a hemimethylated GATC motif. This incision provides the starting point for an excision reaction to cleave the faulty nascent strand and is followed by a repair synthesis reaction.

Several *in vitro* studies using synthetic heteroduplexes and bacteriophage λ, M13 or f1MR1 have shown that the DNA mismatch repair system repairs a mismatch by using a GATC motif within the vicinity of the mismatch (Bruni et al., 1988; Grilley et al., 1993; Pukkila et al., 1983). Moreover, these *in vitro* studies have established that the DNA mismatch repair system can repair a mismatch if there is a GATC motif within 2 kb distance from the mismatch, although an early finding claimed a distance of only 1 kb (Bruni et al., 1988; Lahue et al., 1987). Until now, this distance of 2kb has been considered the 'rule of thumb' in different studies and *in vitro* experiments have been so designed that the nearest GATC motif from a mismatch stays within the 2 kb range. In this study, the *in vitro* range of 2 kb distance has also been experimentally re-examined for the *in vivo* system (in *E. coli*). As mentioned earlier, I have used a 294 bp (98 unit) long CTG•CAG repeat array in this study, which has been inserted between the coordinates of 365502 and 365522 bp in the *lacZ* gene of the genome of *E. coli* K-12 MG1655 (NC_000913.1). For simplicity, the CTG•CAG repeat will be mentioned as the TNR (tri-nucleotide repeat) array for the rest of this chapter. The CTG repeat tract is on the leading strand of the right replicore as depicted in the Figure 3.1.

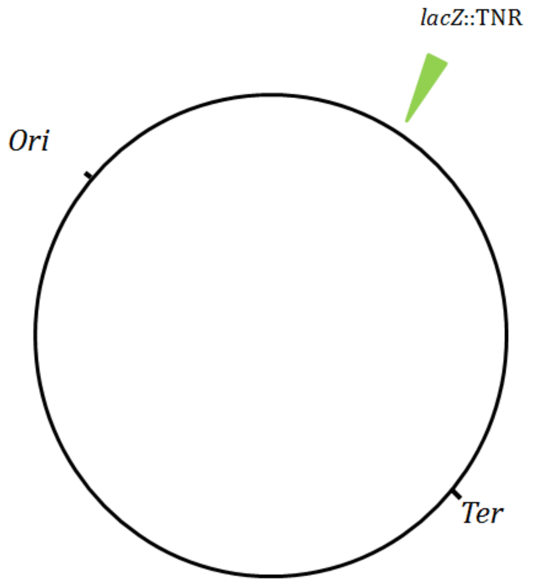


Figure 3.1. A schematic representation of the *E. coli* genome indicating the relative positions of the origin of replication (*Ori*), the terminus locus (Ter) and the *lacZ*::TNR.

There are several natural GATC motifs around the TNR array. Both sides of the TNR array have been defined with regards to the origin of replication of the *E. coli* chromosome as the origin proximal side and the origin distal side of the TNR array. Keeping in mind that a GATC motif should be present within 2 kb distance from a mismatch *in vitro*, for this study the primary region of interest has been fixed to 2 kb on both sides of the TNR array *in vivo* that gives a sequence of 4.3 kb. The region of interest has 7 GATC motifs on the origin distal side of the TNR array while the origin proximal side has 12 (Figure 3.2). The GATC motifs situated on the origin distal side of the TNR array have been named D1 to D7 in a fashion that D1 is the nearest to the array. For those of on the origin proximal side of the TNR array, P1 is the nearest and P12 is the furthest from the array.

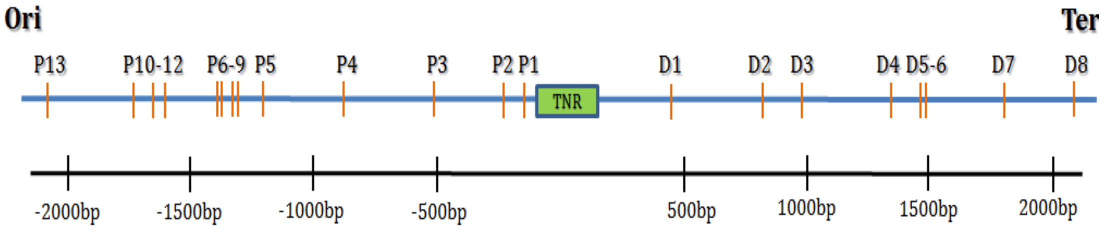


Figure 3.2. A schematic representation of the relative positions of GATC motifs around the TNR array at the region of interest. The GATC motifs are depicted by orange vertical lines around the TNR array (green box) at the region of interest in the *E. coli* genome (depicted by the blue horizontal line). A relative scale has been drawn in the unit of base-pairs centering the TNR array.

3.2 Summary of the strategic approach to process the raw data of the instability assay

As it has been described in the Section 2.2.3.2, the proportion of instability of the TNR array at the single-unit level is calculated for each of the sixty overnight cultures produced from a particular *E. coli* strain being investigated. After the peak calling by the GeneMapper v3.7® (depicted in Figure 2.4), the data for the instability assay is measured as a binary response – whether an instability event had occurred (counted as 1) or not (counted as 0). The proportion of instability from eight replicates of each (among sixty overnight cultures) sample culture is calculated. The strategic procedure (for data processing) is shown here with only three representative samples derived from the same bacterial strain (Table 3.1). For example, if one colony among eight randomly selected colonies (from the plate after serial dilution in figure 2.4A) is found to have an instability (either a single unit TNR deletion, in case of sample X or a single unit insertion, in case of sample Y), the proportion of instability, $P_{(X\ or\ Y)}$ = (1+0+0+0+0+0+0+0)/8 = 0.125. On the other hand, if there is no detectable instability at a single-unit level among eight randomly selected colonies (in case of sample Z), the proportion of instability, $P_{(Z)}$ = (0+0+0+0+0+0+0+0)/8 = 0. Finally, the frequency of instability is calculated to be 0.0834 and the standard error of mean is calculated to be 0.0417 following the equations in the Section 2.2.3.4.

Table 3.1. Representative raw data of the instability assay and the processing strategy.

In case of -1 TNR instability					
	Representative raw sample	**Peak for the marker**	**Base peak**	**Peak for single-unit TNR instability**	**Proportion of instability**
1	Sample X.fsa	424	431	-	0.125
2	Sample X.fsa	424	431	-	
3	Sample X.fsa	424	431	-	
4	Sample X.fsa	424	431	-	
5	Sample X.fsa	424	431	-	
6	Sample X.fsa	424	431	-	
7	Sample X.fsa	424	-	428	
8	Sample X.fsa	424	431	-	
In case of +1 TNR instability					
	Representative raw sample	**Peak for the marker**	**Base peak**	**Peak for single-unit TNR instability**	**Proportion of instability**
1	Sample Y.fsa	424	431	-	0.125
2	Sample Y.fsa	424	431	-	
3	Sample Y.fsa	424	431	-	
4	Sample Y.fsa	424	431	-	
5	Sample Y.fsa	424	431	-	
6	Sample Y.fsa	424	431	-	
7	Sample Y.fsa	424	431	-	
8	Sample Y.fsa	424	-	434	
In case of No TNR instability					
	Representative raw sample	**Peak for the marker**	**Base peak**	**Peak for single-unit TNR instability**	**Proportion of instability**
1	Sample Z.fsa	424	431	-	0

2	Sample Z.fsa	424	431	-	
3	Sample Z.fsa	424	431	-	
4	Sample Z.fsa	424	431	-	
5	Sample Z.fsa	424	431	-	
6	Sample Z.fsa	424	431	-	
7	Sample Z.fsa	424	431	-	
8	Sample Z.fsa	424	431	-	
Calculation					
Proportions of instability:			0.125		
			0.125		
			0		
frequency of instability			0.0834		
Standard deviation			0.072168784		
Standard error			0.041666667		

3.3 The usage of GATC motifs on the origin distal side of the TNR array

Initially, the usage of the GATC motifs on the origin distal side of the TNR array has been studied. During DNA mismatch repair, a GATC motif is the recognition site of the MutH protein, which makes an incision at the 5' position of the G in the 5'-GATC-3' motif that serves as the starting point of the excision reaction (Welsh et al., 1987). To determine how efficient is the DNA mismatch repair system at recognising GATC motif(s), primarily these motifs have been modified sequentially away from the TNR array on the origin distal side. In this study, GATC motifs have been modified by introducing a synonymous mutation such that the protein coding sequence (which is encoded by the *lacI* gene in this case) remains unchanged. Importantly, GATC motifs on the origin proximal side have not been modified for this particular experiment. GeneMapper® (Applied Biosystems) microsatellite analysis (Section 2.2.3.2 and 2.2.3.3) has been used to measure the instability of the length of the TNR array which in turn measures the efficiency of the DNA mismatch repair system in different genetic backgrounds of *E. coli* - the lower the frequency of instability, the more efficient is the DNA mismatch repair system and vice versa.

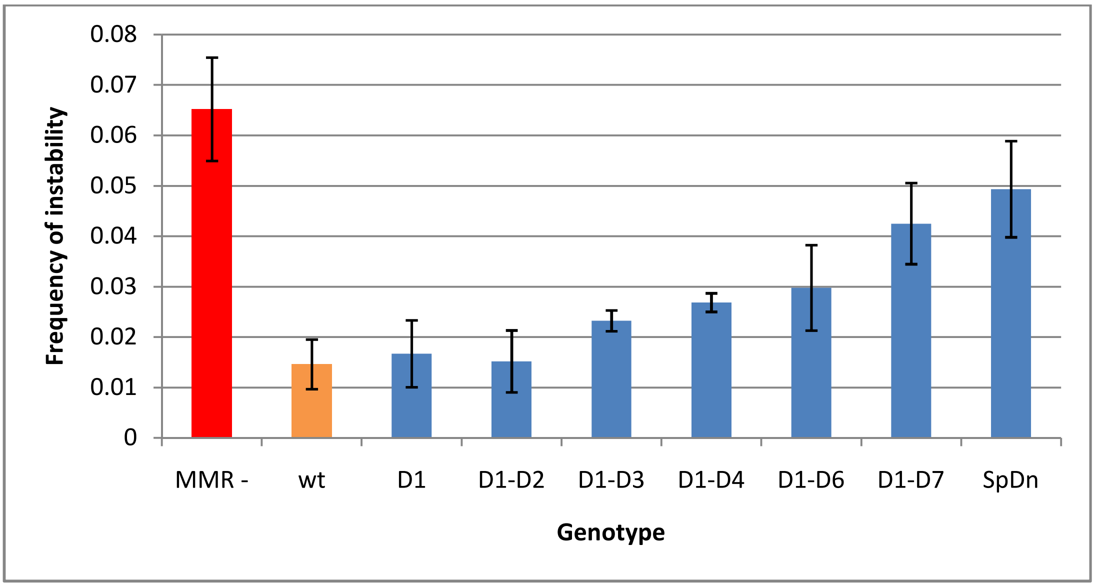


Figure 3.3. The frequency of instability of the TNR array in different genetic backgrounds. In this bar plot, each background has been plotted along the *x*-axis and the height of the corresponding columns shows the respective frequency of instabilities along the *y*-axis. Error-bars represent the standard errors of means. MMR$^-$ is the DNA mismatch repair deficient background and wt is the DNA mismatch proficient background. The sequential modifications of GATC motifs are depicted as D1 to D1-D7 backgrounds. SpDn depicts the strain which has an amplified a 2 kb GATC free *Schizosaccharomyces pombe* genomic DNA inserted at about 200 bp origin distal side of the TNR array.

In the strain where the mismatch repair system has been inactivated by mutating the gene of (*mutS*) the first component of the DNA mismatch repair system, MutS (MMR$^-$ background), the frequency of instability of the TNR array rises about 4.5 fold over the level observed for the wild type cells (wt) where the DNA mismatch repair system is fully functional (Figure 3.3). The basal level of the instability of the TNR array observed for the wild type cells (wt) is assumed to correspond to a small fraction of undetected or unrepaired mismatches and/or secondary mismatches occurring during the repair synthesis by the DNA mismatch repair system.

When the first GATC motif on the origin distal side of the TNR array (472 bp away from the array) has been modified to a non-GATC sequence (D1) and the next available GATC motif is at 844 bp away on the same side, the frequency of instability remains at a similar level of that for the wild type. In addition, the same level of instability of the TNR array is maintained even after the modification of the GATC motif situated at the distance of 844 bp (D1-D2) while the next available GATC motif is 949 bp away on the origin distal side of the TNR array. The first noticeable increment of the level of instability of the TNR array has been observed when the GATC motif at 949 bp has been modified (D1-D3) and the next available GATC motif is situated 1356 bp away from the TNR array on the original distal side. Findings from an early *in vitro* study proposed that the maximum distance between a mismatch and the nearest available GATC motif should be less than 1 kb (Bruni et al., 1988). In this *in vivo* study the level of instability is still very low compared to that in DNA mismatch repair deficient cells (MMR$^-$ background). Moreover, the frequency of instability is not significantly higher than that found in the wild type cells where the DNA mismatch repair system is fully functional at 95% confidence level (*p*-value 0.0889). GATC motifs on the origin proximal side could play a role in this regard by providing a GATC motif needed for starting the cleavage reaction and thus modification of GATC motifs on the origin distal side would then have no noticeable effect regarding the level of instability at the TNR array. An alternative molecular mechanism has been proposed and will be discussed later in this chapter.

Subsequent modifications of GATC motifs, thereby pushing the next available GATC motif further away from the TNR array on the origin distal side resulted in a gradual increase of the frequency of instability of TNR array. The modification of the next GATC motif at a distance of 1356 bp from the TNR array (D1-D4) resulted in a slight increase in the frequency of instability while the next available GATC motif resides at a distance of 1494 bp from the TNR array. The fifth and sixth GATC motifs on the original distal side, situated at a distance of 1494 bp and 1508 bp respectively, have been modified together (D1-D6) considering that their close proximity should result in the same effect on the instability of the TNR array than modifying them separately. This modification results in a slight increase of the frequency of instability from what has been found in the previous modification (background D1-D4).

A visible effect has been observed after modifying all the available GATC motifs within a 2 kb region on the original distal side of the TNR array (background D1-D7). The GATC motif at a distance of 1826 bp from the TNR array has been modified and the next available GATC motif is at a distance of 2370 bp from the TNR array on the origin distal side which is beyond the *in vitro* boundary for the availability of at least one GATC motif from the mismatch (Grilley et al., 1993; Modrich, 1991; Pukkila et al., 1983). In this case, the level of instability of the length of the TNR array rises higher, but still does not reach the level of a MMR$^-$ background. The *p*-value of the difference of frequency of instabilities between MMR$^-$ background and D1-D7 background is 0.0317, which is statistically significant at a 95% confidence level.

At this point, the matter of argument would be the involvement of the GATC motifs on the origin proximal sides, where the first two are quite near to the TNR array (only 110 bp and 206 bp respectively away from the TNR array) which might act as the starting point of the excision reaction during MMR. There are 2 things to be considered: i) whether a GATC on the origin proximal side can contribute during MMR and ii) how efficient is this contribution? If GATC motifs on the origin proximal side of the TNR array were equally efficient as a starting point for the cleavage reaction as those are on the origin distal side of the TNR array, there would have been no change at all in the level of instability of the TNR array after sequential modification of the GATC motifs on the origin distal side only. However, there is a possibility that GATC motifs on the origin proximal side could contribute to the efficiency of the DNA mismatch repair system to a minimal level and this is the reason for the level of instability of the TNR array in D1-D7 background did not raise up to the level of the MMR$^-$ background, even if there is no GATC motifs within 2 kb region on the origin distal side of the TNR array. This possibility will be discussed in a later section.

The next step is to redefine the boundary or the range of the nearest available GATC motif from a mismatch for a successful DNA mismatch repair. As there are several GATC motifs beyond 2 kb boundary on the origin distal side of the TNR array in the *E. coli* genome, rather than modifying those GATC motifs sequentially, a more efficient experiment has been devised to extend the GATC-free region by inserting a non-*E. coli* DNA sequence in the strain D1-D7. A 2 kb long GATC motif-free DNA sequence has been isolated from the fission yeast *Schizosaccharomyces pombe* and inserted at 200 bp away from the TNR array on the origin distal side (background SpD). The rationale behind choosing a non-*E. coli* sequence is to avoid any intra-sequence recombination during strain preparation and performing the assay. However, the potential of altered DNA metabolism after inserting a foreign DNA sequence due to different intrinsic properties of the sequence (different DNA composition *etc*) has not been thoroughly examined in this case. In that strain, the total length of a GATC motif-free sequence reaches up to 4.3 kb. The frequency of the instability of the TNR array has been measured in SpD strain and this time the frequency becomes very high compared to the level for the wild type cells (*p*-value 1.7154e-06, which is statistically significant at a 95% confidence level) where the DNA mismatch repair system is fully functional (Figure 3.3). In addition, the frequency rises very close to that for the DNA mismatch repair deficient background (MMR$^-$) and cannot be differentiated from the MMR$^-$ background (*p*-value is 0.9451 at 95% confidence level) (Figure 3.3). No significant difference has been found between the overall proportions of single TNR unit contraction or expansion by performing a χ^2 test, therefore, single TNR unit contraction events is not favoured over the proportion of single TNR unit expansion, and *vice versa* (Appendix A). However, a 5’-CTG repeat array in the leading strand template is relatively stable than a 5’-CAG repeat array

(Delagoutte et al., 2008). Therefore, it could be difficult to measure relative abundance of either contraction or expansion events in overall rare occurrences of instabilities in this experiment.

An obvious question arises at this point: Is the severe decline of the efficiency of the DNA mismatch repair system in the SpD strain a local phenomenon (due to the increased distance between a mismatch and the nearest GATC motif) or has it become a global phenomenon for the cell (any of the genes in the DNA mismatch repair pathway in the cell might have acquired mutation(s) over the course of genetic modification in the laboratory)? To address this question, the overall frequency of mutation in this strain has been investigated by fluctuation analysis based on point mutation(s) ariseing in the gene *rpoB* which eventually confers resistance against the antibiotic Rifampicin. *rpoB* codes for the β-subunit of RNA polymerase and point mutations in a limited 69 bp long region (rifampicin resistance-determining region) is responsible for 96% of the events resulting in rifampicin resistance (Ford et al., 2013; Mariam et al., 2004; Nicholson and Maughan, 2002)

The rate of mutation (events/base/generation) in the *rpoB* gene increases 46 fold in DNA mismatch repair mutant background (MMR$^-$) compared to the wild type background (Figure 3.4). The rate of mutation in case of the SpD background, which has a 4.3 kb long region free of GATC motifs, has been found to be at the similar level to that of the wild type (95% confidence interval). Therefore, the DNA mismatch repair system is fully functional in the SpD background and the higher level of instability of the TNR array observed for this background (Figure 3.3) is due to the lack of the GATC motifs over a region of 4.3 kb on the origin distal side of the TNR array.

Therefore, it clearly shows that the DNA mismatch repair system still repairs mismatches (with a lower efficiency though), and is not completely inefficient in repairing mismatches when the next available GATC motif is beyond 2 kb while the origin proximal GATC motifs are still there. The relative efficiency of MMR with respect to the availability of origin distal GATC motif will be discussed in a later section.

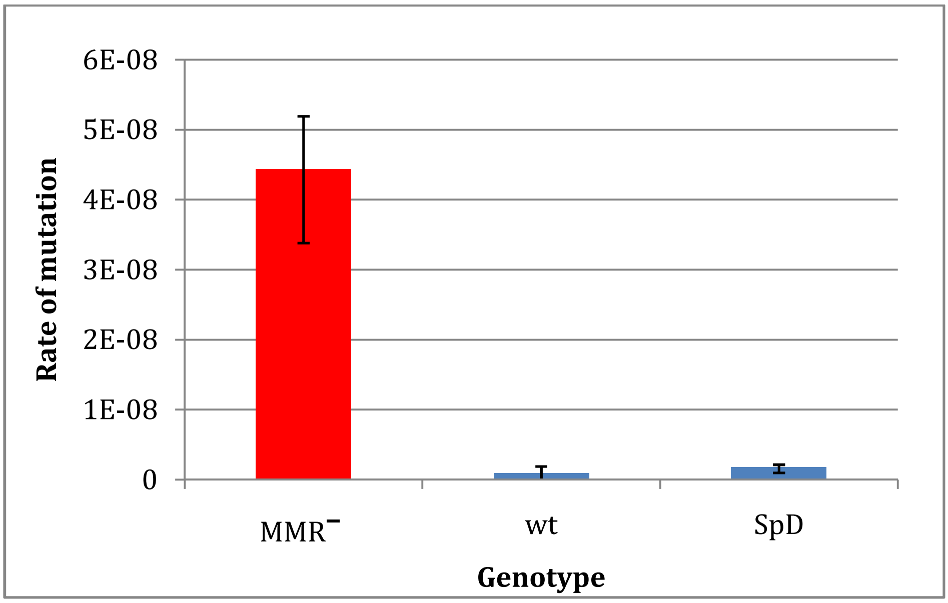


Figure 3.4. The frequency of mutation(s) in the *rpoB* gene under different genetic backgrounds as a measure of the efficiency of the DNA mismatch repair system. In this bar plot, each genetic background has been plotted along the *x*-axis and the height of the corresponding bars shows the respective frequency of mutation along the *y*-axis. The error-bars represent the upper and lower limits of 95% confidence intervals. MMR$^-$ is the DNA mismatch repair deficient background and wt is the MMR proficient background. SpD depicts the genetic background where a 2 kb long GATC motif-free sequence of *Schizosaccharomyces pombe* has been inserted which gives in total of 4.3 kb of GATC motif-free sequence on the origin distal side of the TNR array.

3.4 The usage of GATC motifs on the origin proximal side of the TNR array

The usage of GATC motifs on the origin proximal side of the TNR array has been investigated by modifying them, in the same way described earlier, in the direction away from the TNR array while the GATC motifs on the origin distal side have not been modified. After the modification of the first two GATC motifs (at 110 bp and 206 bp respectively from the TNR array) on the origin proximal side (background P1-P2), the frequency of instability of the TNR array seems to decrease slightly compared to that of the wild type background (the *p*-value is 0.0105 at 95% confidence level) (Figure 3.5). However, when all GATC motifs within a 2 kb region on the origin proximal side of the TNR array have been modified (background P1-P12) and the next available GATC site is at a distance of 2321 bp from the TNR array on the origin proximal side, the frequency of instability is similar to that observed for wild type cells (*p*-value 0.6076 at 95% confidence level) (Figure 3.5). On the other hand, this level does not rise up to the level observed for the mismatch repair deficient background (background MMR$^-$). Moreover, this difference is statistically highly significant (*p*-value 1.1903e-06 at 95% confidence level) (Figure 3.5).

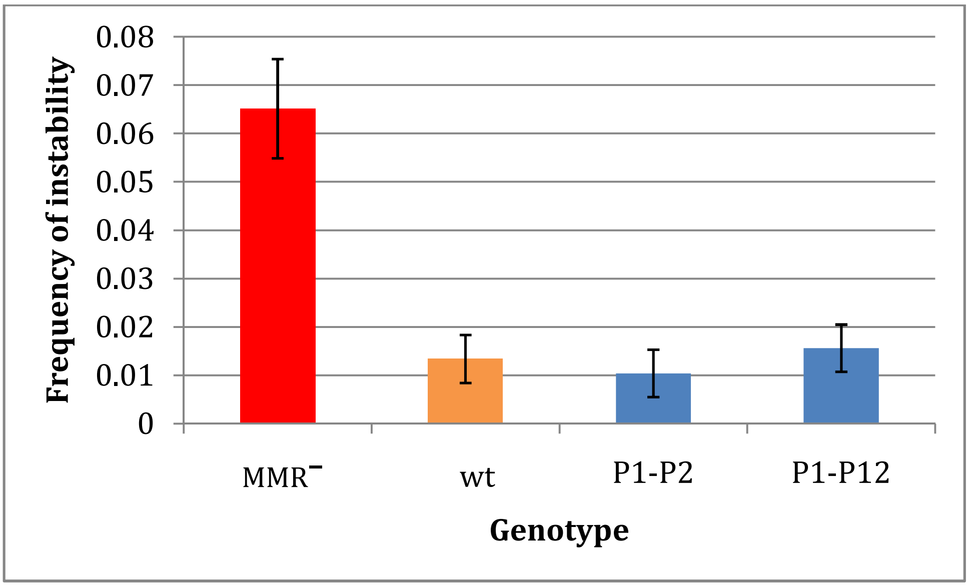


Figure 3.5. Comparing the frequency of the instability of the TNR array in different genetic backgrounds. In the bar plot, each background has been plotted along the *x*-axis and the height of the corresponding bars shows the respective frequency of instabilities along the *y*-axis. Error-bars have been calculated from the standard errors of the mean of the experimental samples. MMR$^-$ is the DNA mismatch repair deficient background (*mutS*$^-$) and wt is the DNA mismatch proficient background. The modification of GATC motifs on the origin proximal side of the TNR array are indicated as P1-P2 and P1-P12 backgrounds.

From the result, it has become clear that GATC motifs on the origin proximal side of the TNR array do not contribute as starting points of the excision reaction during DNA mismatch repair when the origin distal GATC motifs are intact. However, one cannot rule out the possibility that the origin proximal GATC motifs might be used only if there is no origin distal GATC motif (of a mismatch) in a long stre*tc*h of DNA sequence. Therefore, in this case, the GATC motifs on the origin distal side of the TNR array are recognised by MutH protein and are the sole contributors to the start site of the excision reaction during DNA mismatch repair.

3.5 The usage of GATC motif(s): single GATC motif vs. an array of GATC motifs

Different short DNA sequences or DNA motifs have been identified which bear biological significance by getting recognised by respective DNA binding proteins and are thereby involved in different biological processes. In addition, some DNA motifs act better in cluster than solitary as it enhances the probability of their recognition by their respective DNA binding proteins. For example, biochemical studies have established that SeqA, which is not directly involved in the DNA mismatch repair system, binds to GATC motifs situated at the origin of replication of the *E. coli* genome to impede immature start of replication (Brendler and Austin, 1999; Brendler et al., 1995; Kang et al., 1999; Slater et al., 1995). Interestingly, SeqA has a preference for binding to a cluster of GATC motifs as shown in a study where this sequence has been inserted at different loci (in *srlA* and *tnaA* gene sequences) (Waldminghaus et al., 2012). In addition, the frequency is high for closely spaced GATC motifs in the *E. coli* genome (will be discussed in Chapter 6). Therefore, it is worth studying the *in vivo* preference of the DNA mismatch repair proteins for a single or a cluster GATC motif(s).

In vitro experiments have been designed with heteroduplex DNA where usually a single GATC motif resides within a 2 kb region on either side of a mismatch (Grilley et al., 1993; Modrich, 1991; Pukkila et al., 1983). An early *in vitro* study on the utilisation of GATC motif(s) demonstrated an elevated efficiency of the DNA mismatch repair system in a covalently closed circular DNA in the presence of four closely spaced GATC motifs compared to a GATC motif-free homeologous DNA substrate (Lahue et al., 1987). In that experiment, the efficiencies of DNA

MMR for a homeologous substrate was only 10% and 93% with one and two GATC motifs respectively. However, the author did not correlate between the efficiency of DNA mismatch repair system and the number of GATC motif(s) present in the substrate. On the other hand, to my knowledge, no *in vivo* experiments have been yet devised to investigate the number of GATC motif(s) required in the *E. coli* genome for a successful DNA mismatch repair in a natural cellular environment.

As previously shown, the DNA repair system is less efficient in repairing a mismatch when there is no GATC motif in a distance of 2 kb from the TNR array on the origin distal side (for background D1-D7, Figure 3.3) which leads to an increase of the level of instability. A single ectopic GATC motif and an array of 3 GATC motifs separated by 4 nucleotides have been inserted at a distance of 567 bp from the TNR array on the origin distal side in the D1-D7 strain. The modified strains have been named as 1G and 3G, respectively and their level of instability has been calculated. After introducing a single ectopic GATC motif or an array of GATC motifs, the frequency of instability decreases compared to that has been observed in the D1-D7 strain. Moreover, the frequency of instability of TNR array decreases even below to the frequency observed in wild type cells. However, the experimental error bars overlap with that in wild type level with a *p*-value of 0.1150. So, the level of instability observed in the 1G and 3G strains has been considered as the same level observed in wild type cells at 95% confidence level.

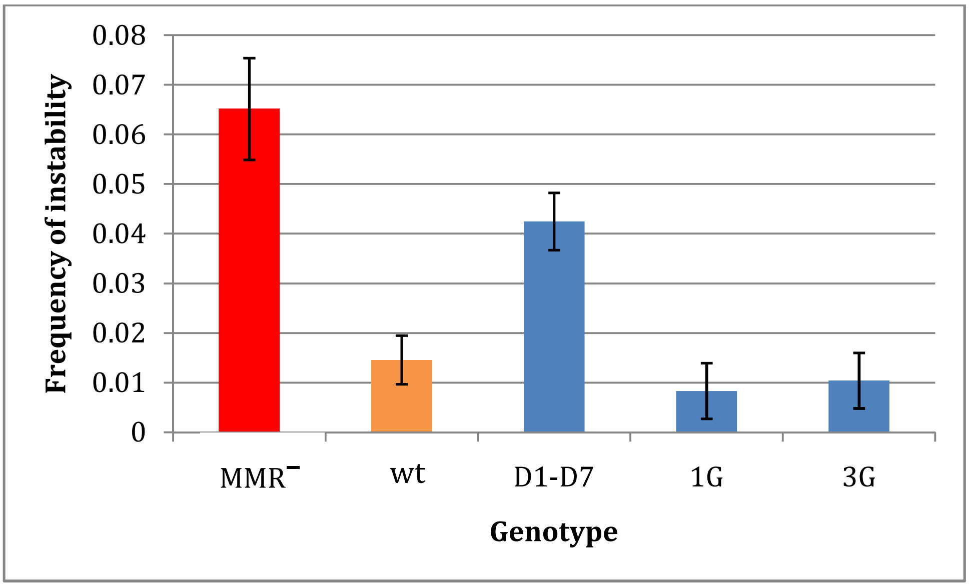


Figure 3.6. The DNA mismatch repair system does not prefer an array of GATC motif(s) over a single GATC motif of repairing a mismatch. In the bar plot, each background has been plotted along the *x*-axis and the height of the corresponding bars shows the respective frequency of instabilities along the *y*-axis. The error-bars have been calculated from the standard errors of the mean of experimental samples. MMR$^-$ is the DNA mismatch repair deficient background (by knocking out the *mutS* gene) and wt is the DNA mismatch proficient background. The D1-D7 background has a 2 kb DNA sequence free of GATC motifs on the origin distal side of the TNR array. 1G and 3G background has an ectopic single GATC and an array of 3 GATC motifs respectively at 567 bp from the TNR array on the origin distal side.

Thus, the DNA mismatch repair system seems to be quite efficient at finding a GATC in the vicinity of a mismatch and an extra GATC nearby, which is found frequently in the *E. coli* genome (will be discussed in the Chapter 6), is not necessary for increasing the efficiency of the DNA mismatch repair system in this case. However, inserting a single GATC motif and an array of GATC motifs (as has been used in the assay described above) to a distant site of the TNR array would make it possible to compare the true impact of having more GATC motifs on the efficiency of MMR.

3.6 Discussion

3.6.1 The DNA mismatch repair system has a distinct directionality in relation to the DNA replication

In vitro studies have shown that an incision of the nascent strand can occur either on 3' or 5' of a mismatch during DNA mismatch repair, depending on the position of the hemimethylated GATC motif (Au et al., 1992; Modrich, 1989). Using electron microscopy and an end-labelling method, Grilley and collaborators showed that the single stranded DNA region created after an excision reaction spans the shortest path between the mismatch and the nearby GATC motif (Grilley et al., 1993). In both directions, the excision reaction uses the same helicase II (UvrD), but the requirement for single-strand specific exonucleases is different depending on the orientation of the heteroduplex substrate. A hemimethylated GATC motif residing on the 3' side of the mismatch requires Exo I and/or Exo X, while in the other direction (5' → 3' cleavage) Exo VII and/or RecJ is/are required for cleaving the single stranded DNA. Therefore, the *in vitro* cleavage reaction of the DNA mismatch repair system is bi-directional and covers the shortest length between the mismatch and the nearest hemimethylated GATC motif. However, in this *in vivo* study, I have clearly shown that the DNA mismatch repair system has a preference for GATC motifs on the origin distal side. This difference between the *in vitro* and *in vivo* studies can be explained if we consider the mode of replication of *E. coli* genome (Figure 3.7).

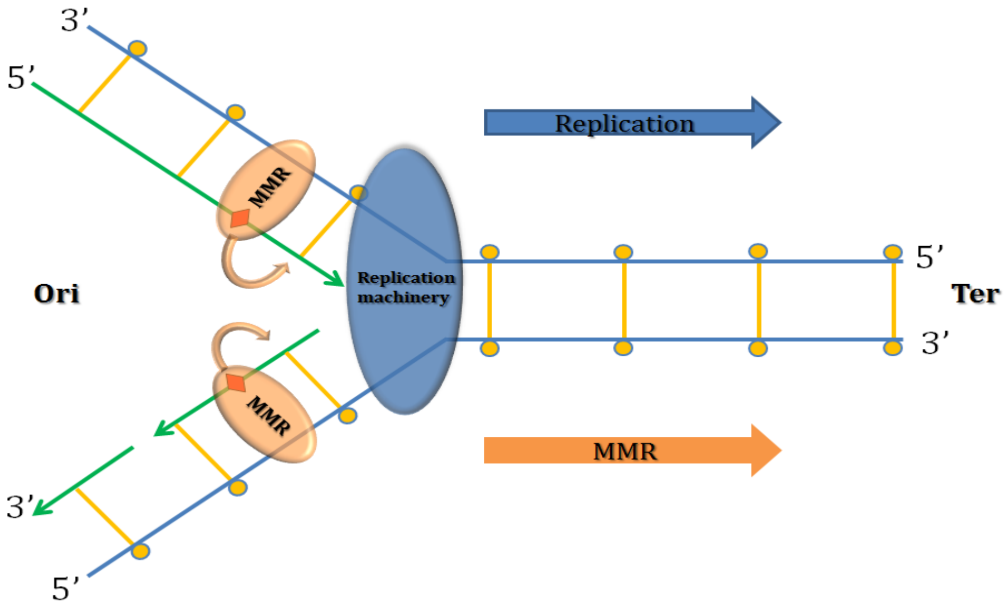


Figure 3.7. A cartoon representation of a replication fork and directionality of MMR at any particular time. Two denatured DNA strands serve as templates (parent strands, shown as blue lines) for the synthesis of a nascent continuous (leading) strand and a nascent discontinuous (lagging) strand (both shown as green lines). At any time point, the replication machinery stays at the junction of the denatured single strands and the parent double strand. Sometimes, the replication machinery leaves a mismatch (depicted by a red diamond) on the nascent strand (either leading or lagging strand). At this moment, the parent strands are methylated (methyl groups are depicted with amber circles) at GATC motifs (depicted with small amber lines) while the nascent strands are yet to be methylated.

During the replication of the *E. coli* genome, the denatured DNA strands are replicated: the leading strand is a continuous nascent strand, while the lagging strand is a discontinuous nascent strand. At this moment, GATC motifs remain unmethylated in both types of nascent DNA strands while GATC motifs on the parent DNA strands are methylated. During DNA mismatch repair, these hemimethylated GATC motifs serve as recognition sites for MutH and as the start point of the excision reaction. On the other hand, different single stranded exonucleases operating in both directions are found to accomplish the excision reaction which, in a narrow perspective, suggests that the DNA mismatch repair system is bidirectional in utilising GATC motifs near a mismatch. However, this *in vivo* study shows that only GATC motifs on the origin distal side take part in DNA mismatch repair. Figure 3.7 demonstrates that to repair a mismatch on the leading nascent strand utilising the origin distal GATC motif, a 3' → 5' exonuclease is needed. This task can be assigned to ExoI or Exo X nuclease. On the other hand, during the repair of a mismatch on the lagging nascent strand and utilising the same GATC motif on the origin distal side, a 5' → 3'exonuclease such as Exo VII or RecJ is required. In both cases the DNA mismatch repair system is utilising a GATC motif which is on the origin distal side and based on the directionality of the single strand dependent exonucleases, the DNA mismatch repair system is travelling away from the origin of replication which is basically the same direction of the replication machinery.

3.6.2 The efficiency of the DNA mismatch repair system

This *in vivo* study has opened a broader perspective of the directionality of the DNA mismatch repair system and redefined the minimal distance required between a mismatch and a GATC motif in the context of a living cell. In addition, it has provided a relative level of efficiency of the DNA mismatch repair as a function of the distance between a mismatch and the nearest GATC motif on the origin distal side of the TNR array (as origin proximal GATC motifs do not seem to contribute to the efficiency of mismatch repair). Sequential modifications of GATC motifs on the origin distal side result in a gradient of efficiency of the DNA mismatch repair system. Relative efficiencies of the DNA mismatch repair system have been calculated using the formula described in Section 2.2.4 of Chapter 2. The relative efficiencies have been plotted in Figure 3.8 as a function of the distances between the next available GATC motifs and the TNR array.

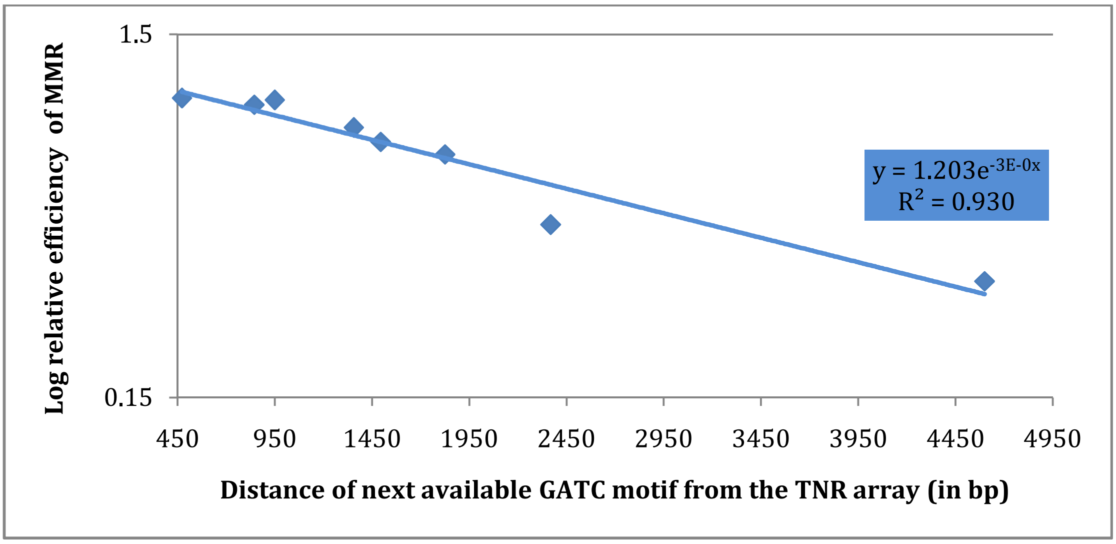


Figure 3.8. The relative efficiency of MMR as a function to the distance between the TNR array and the next available GATC motif. The distances between the TNR array and the next available GATC motif on the origin-distal side are plotted along the *x*-axis while the relative efficiencies of the DNA mismatch repair system are plotted along the *y*-axis. The exponential trendline has been drawn using built in module of Microsoft excel and the R^2 value is shown in the graph.

The efficiency of the MMR system in a wild type cell has been defined as 1 while that in a *mutS* mutant (MMR^-) has been set to 0. Sequential modifications of GATC motifs on the origin-distal side results in the loss of efficiency of MMR depicted in Figure 3.8. Fitting an exponential trendline (R^2 = 0.930) to the data reveals that the MMR system is 50% efficient when the distance between the TNR array and the next available GATC motif is 2.8 kb.

Chapter 4

During DNA mismatch repair, the length of the excision reaction tract depends on the start point; not on the position of the mismatch

4.1 Introduction

Upon detection of a mismatch in the DNA duplex, following a replication error, MutS dimer binds the mismatch (Parker and Marinus, 1992; Su and Modrich, 1986). Then, MutL proteins bind the MutS dimer in an ATP-dependent fashion and assemble into a ternary complex (Allen et al., 1997; Galio et al., 1999; Grilley et al., 1989). This ternary complex in turn recruits a d(GATC) endonuclease - MutH, which recognises a hemimethylated d(GATC) motif to discriminate the nascent strand from its parent strand (Au et al., 1992s). MutH makes an incision on the nascent strand to initiate the excision repair reaction in the course of DNA mismatch repair (Au et al., 1992). The orchestrated activity of the DNA helicase II (UvrD) and the single strand dependent exonuclease(s) (one or more of ExoI, ExoVII, RecJ and ExoX) constitutes the excision reaction of DNA mismatch repair commencing from the MutH mediated incision on the faulty nascent strand and migrates towards the mismatch which is to be repaired. The first step of the excision reaction is mediated by DNA helicase II, the product of the gene *uvrD* or *mutU* which unwinds the nascent DNA strand starting from the incision created by MutH endonuclease (Dao and Modrich, 1998). Using a 6.4 kilobase circular substrate harbouring a G-T heteroduplex and a single-strand incision on either side of the heteroduplex, Modrich and collaborators have showed in an *in vitro* experiment that unwinding of the nascent strand by UvrD favours the shortest path between the incision and the mismatch (Dao and Modrich, 1998). Therefore, the *in vivo* excision reaction might have directionality towards the mismatch along the chromosome. This unwinding reaction is accompanied by degradation of the recently unwound faulty nascent strand by one or more of the four exonucleases that have been identified to date: ExoI, ExoVII, RecJ and ExoX (Viswanathan et al., 2001). *In vitro,* the excision reaction (thereby, the requirement for specific exonucleases) is directed by the location of an incised d(GATC) motif either 5' or 3' of the mismatch on the unmethylated nascent strand (Cooper et al., 1993; Viswanathan et al., 2001). If the incision is made on 5' of the mismatch, the ExoVII and/or RecJ is/are recruited as they possess 5'→3' hydrolytic activity (Chase and Richardson, 1974; Lovett and Kolodner, 1989). On the other hand, if an incision made on the 3' side of the mismatch, ExoI and/or ExoX with 3'→5' hydrolytic activity are recruited (Lehman and Nussbaum, 1964; Viswanathan et al., 2001).

Grilley and collaborators have found that the *in vitro* excision reaction starts from a hemimethylated GATC motif and terminates at different discrete sites within a 100 nucleotides region on the other side of a mismatch (Grilley et al., 1993). However, this finding came from *in vitro* studies of the DNA mismatch repair system and the scenario may be different *in vivo*. From Chapter 3, it has been established that in *E. coli* the excision reaction starts from an incision at the hemimethylated GATC motif on the origin distal side of a mismatch. Using a CTG•CAG trinucleotide repeat array (TNR array) and a recombination reporter (Zeo) on the original proximal side of the TNR, Blackwood and collaborators have shown that recombination of the reporter is stimulated by DNA mismatch repair and the rate of recombination is a function of the distance between the TNR array and the reporter (Blackwood et al., 2010) (Figure 4.1A). When the reporter was placed at a distance of 6.3 kb from the TNR array on the origin proximal side, the rate of recombination was 2 fold higher than the wild type background which does not harbour the TNR array (Blackwood et al., 2010). Interestingly, the rate of recombination of the zeocin tandem repeat increased 6 fold compared to the wild type background when the tandem repeat was placed at a distance of only 0.5 kb from the TNR array on the origin proximal side (Figure 4.1). This *in vivo* finding appears to be inconsistent with the traditional *in vitro* view about the length of the excision reaction tract during mismatch repair. In this current study, I have set out to investigate the *in vivo* length of the excision reaction tract. It was not clear whether the end point depends on the position of the mismatch or the GATC motif. In this study, the positions of the end points regarding the positions of both mismatch and the GATC motifs have been tested.

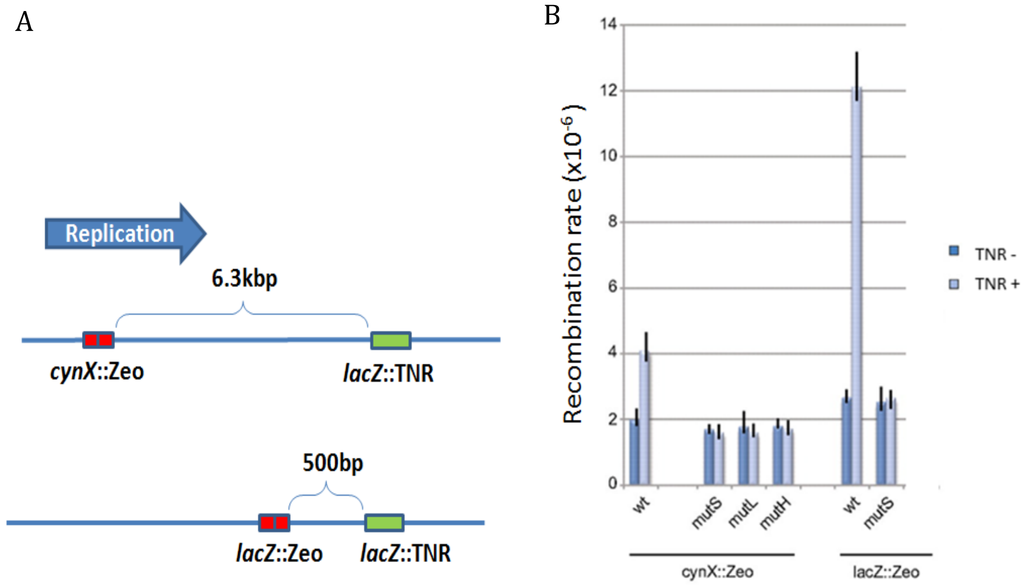


Figure 4.1. A CTG•CAG TNR array stimulates recombination at a reporter – 275 bp tandem repeat (zeo) inserted on the origin proximal side of the TNR array. (A) The *E. coli* chromosome is shown as blue horizontal lines, the TNR array is shown as a green rectangle and the zeocin tandem array is shown as two tandem red boxes at different distances from the TNR array. (B) The recombination rates of the zeocin tandem repeats (recombination reporter) in different genetic backgrounds and with or without the TNR array have been shown here. Figure (B) is from (Blackwood et al., 2010).

In this study, the same recombination reporter – zeocin tandem repeat has been used to determine the end point of the excision reaction tract in the course of DNA mismatch repair. The tandem repeat is composed of two defective ORFs of 275 bps (Figure 4.2). The first ORF has an intact initiation codon, but lacks the sequence for the C-terminus of the coded protein. On the other hand, the second ORF has an intact sequence except for the initiation codon. Upon recombination, which is stimulated by the DNA mismatch repair system, the tandem repeat becomes a functional single ORF which codes for a protein that confers resistance against the antibiotic zeocin (Blackwood et al., 2010). Blackwood and collaborators proposed that recombination is mediated by DNA resynthesis following the formation of a single stranded sequence harbouring the tandem repeat during the excision reaction in the course of DNA mismatch repair (Blackwood et al., 2010; Eykelenboom et al., 2008).

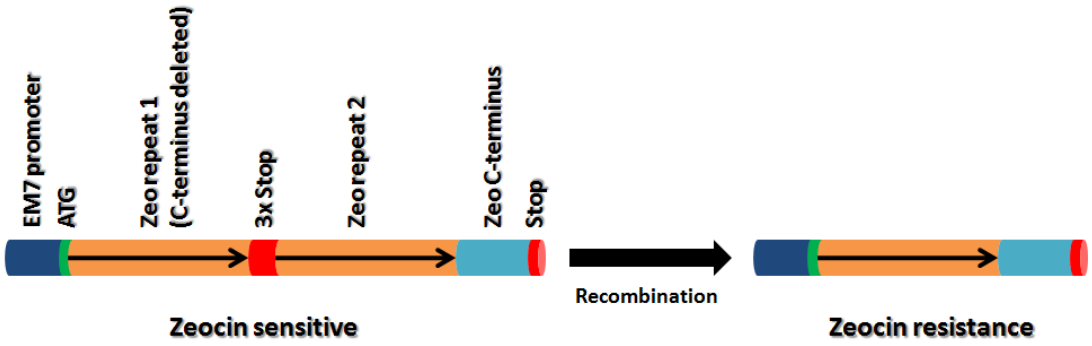


Figure 4.2. Structure of the 275 bp tandem repeat. The first repeat contains an initiation codon but the C-ternimus of the encoded protein is deleted. The second repeat contains an intact ORF except for the initiation codon. These two parts are separated by three stop codons (on each frame) to stop any translation starting from the initiation codon. Upon recombination between these incomplete repeats, a complete protein coding sequence emerges which codes for a protein that confers resistance against the antibiotic zeocin.

4.2 The excision reaction of the DNA mismatch repair system does not end at a GATC motif on the origin proximal side of a mismatch

In a model of the molecular mechanism of DNA mismatch repair, Su and collaborators proposed that the excision reaction might extend from a hemimethylated GATC motif on one side of the mismatch to another hemimethylated GATC motif on the opposite side of the mismatch as the MutH makes incision on both GATC motifs (Figure 4.3) (Su et al., 1989). Therefore, the orchestrated action of the UvrD helicase and respective exonuclease(s) would create a single stranded region from one GATC motif on one side of the mismatch to another GATC motif on the other side of the mismatch.

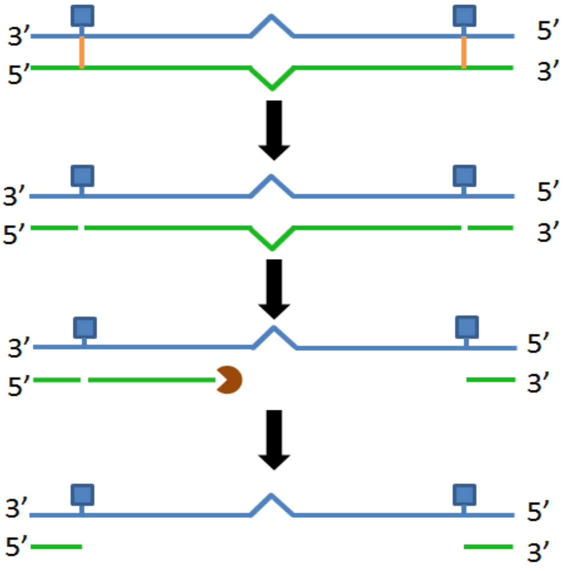


Figure 4.3. A model describing the excision reaction and the formation of a single stranded DNA region during the mismatch repair. After recognition of a mismatch by MutS and formation of a ternary complex along with MutL, MutH (recruited by the ternary complex) incises the nascent strand at the hemimethylated GATC motifs on both sides (3’ and 5’) of the mismatch. The orchestrated activity of the UvrD helicase and respective exonuclease(s) unwinds and degrades the nascent strand spanning both inscisions. Under these circumstances, the excision reaction starts from a GATC motif on one side and ends at another GATC motif on the other side of the mismatch. The parent strands are shown as blue lines and the nascent strands as green lines. GATC motifs are shown as small vertical amber lines and methyl groups are shown as blue boxes.

In Chapter 3, it has been demonstrated that the DNA mismatch repair system utilises the GATC motifs on the origin distal side of a mismatch. Based on the aforementioned model, the excision reaction would start at a GATC motif on the origin distal side and end at a GATC motif on the origin proximal side of a mismatch. Inserting the zeocin tandem repeat at a position on the origin proximal side of a mismatch, thereby studying the frequency of recombination of the repeat, can provide us with a way to test whether the excision reaction terminates at a GATC motif on the origin proximal side. In this *in vivo* experiment, the zeocin tandem repeat has been inserted at 500 bp on the origin proximal side of the TNR array (Figure 4.4). There are two GATC motifs (P1 and P2) in close proximity between the TNR array and the zeocin tandem repeat array. If the DNA mismatch repair system started the excision reaction from a GATC motif on the origin distal side of the TNR array and if it incised at a GATC motif on the origin proximal side (where the excision reaction would have terminated as well), modifying P1 and P2 would increase the rate of recombination of the zeocin tandem repeat as the excision tract would have a higher probability of reaching the next GATC motif. The rate of recombination of the zeocin tandem repeat has been investigated following the modification of P1 and P2 GATC motifs. The significance of the difference of the recombination rates between two genetic backgrounds has been calculated with 95% confidence interval based on a method developed by Spell and Jinks-Robertson (described in Section 2.2.3.5) (Spell and Jinks-Robertson, 2004).

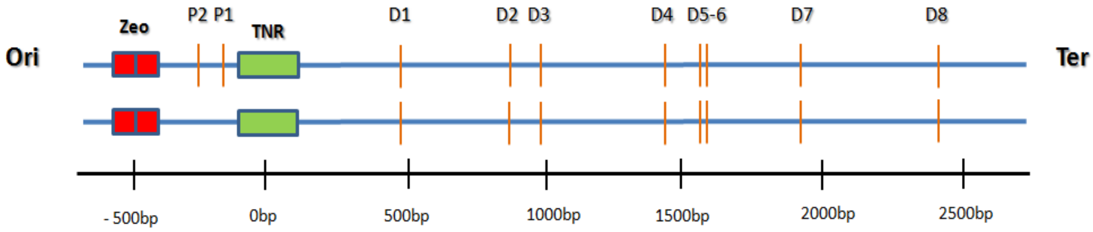


Figure 4.4. A schematic representation of the locus of interest and the experimental approach to investigate the influence of GATC sites on the excision reaction. There are several GATC motifs on both sides of the CTG•CAG trinucleotide repeat (TNR) array. On the origin proximal side of the TNR array, two GATC motifs (P1 and P2) are situated between the TNR array and zeocin tandem repeat. They have been modified to determine their influence on the excision reaction during DNA mismatch repair. The *E. coli* chromosome is shown as blue horizontal lines, the TNR array is shown as green rectangles and the zeocin tandem array is shown as two tandem red boxes. GATC motifs are shown as small vertical amber lines with respective names based on their positions.

The recombination rate did not change significantly after modifying the P1 and P2 GATC motifs between the TNR array and the zeocin tandem repeat (Figure 4.5). In the absence of the TNR repeat array (in all the TNR$^-$ backgrounds), the rate of recombination stays at a minimum level as there were no frequent mismatches to deal with. In a mismatch repair deficient background, even in the presence of the TNR array (the MMR$^-$ TNR$^+$ background), the rate of recombination was still low as there was no excision reaction mediated single stranded DNA and the zeocin tandem repeat did not show elevated recombination. In the presence of the TNR array, the mismatch proficient wild type background (wt TNR$^+$) showed about 7 fold more recombination compared to the TNR$^-$ backgrounds and mismatch repair deficient background. In the mismatch repair proficient cells with modified P1 and P2 GATC motifs, the rate of recombination of the zeocin tandem repeat did not increase, rather it remained at a similar level to that observed for the wt TNR$^+$ background and the difference in the rate of recombination between the wt TNR$^+$ background and P1-P2 TNR$^+$ background was not statistically significant (Spell and Jinks-Robertson, 2004). Therefore, it can be concluded that the excision reaction of the DNA mismatch repair system is not influenced by the presence of GATC motifs on the origin proximal side of a mismatch.

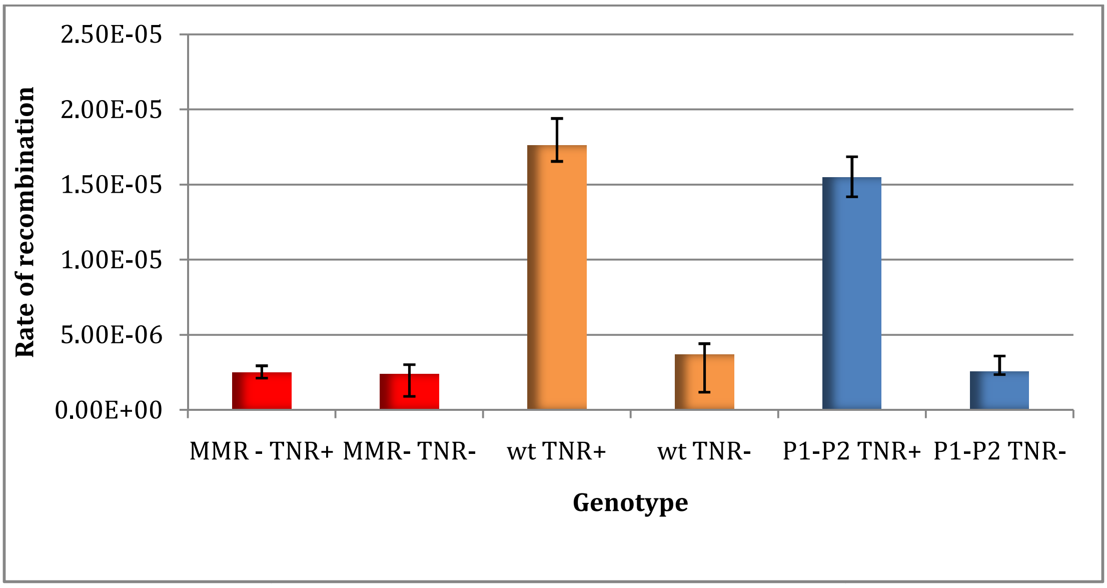


Figure 4.5. The excision reaction during DNA mismatch repair does not end at a GATC motif on the origin proximal side of a mismatch. A comparative distribution of the rate of recombination of the zeocin tandem array under different genetic backgrounds have been shown here. The rate of recombination is shown along the *y*-axis and different genotypes are put along the *x*-axis. The error-bars represent the upper and lower limits of 95% confidence intervals.

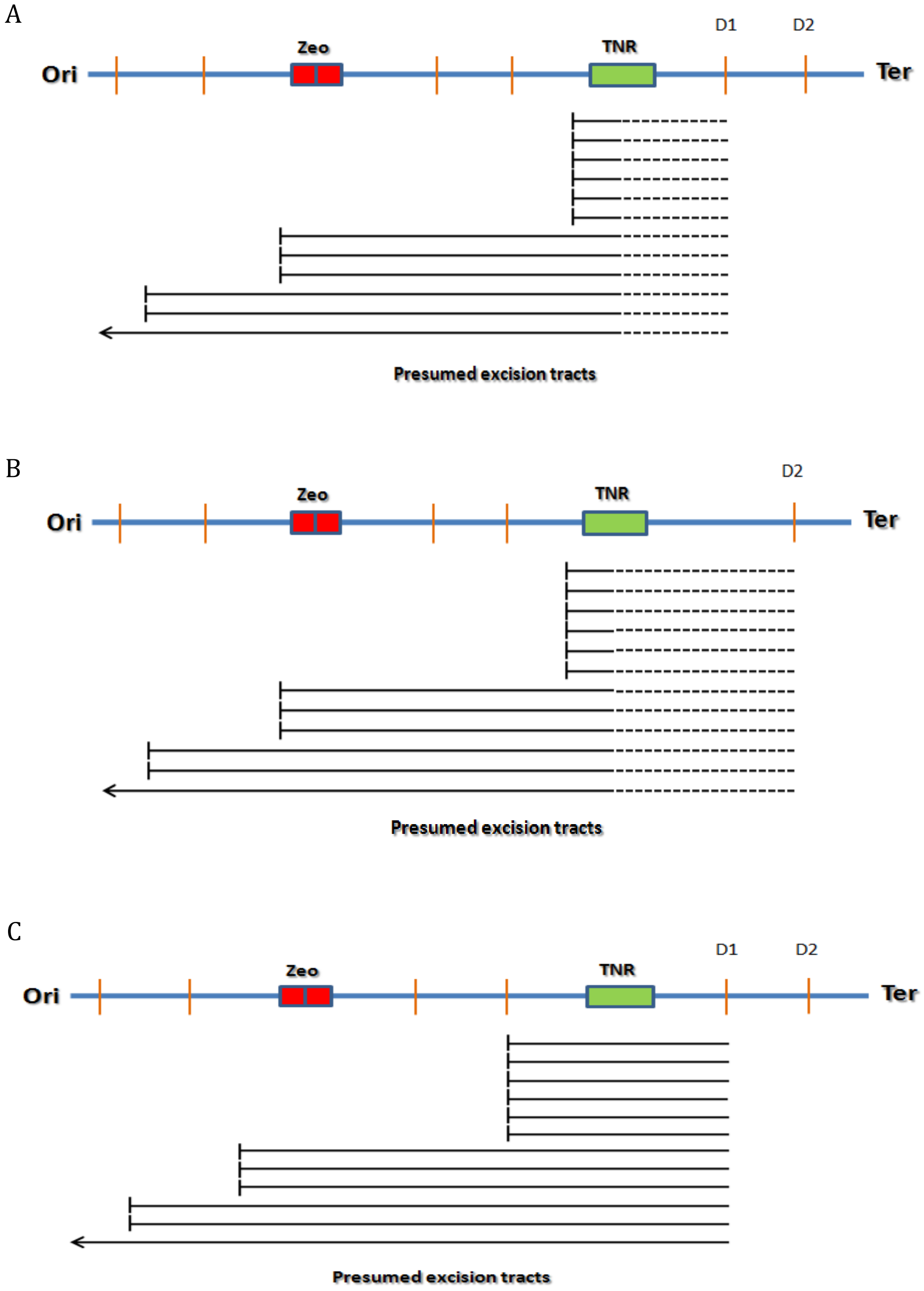
4.3 The distance of an excision tract depends on its start point, rather than the position of a mismatch

According to the previously mentioned study by Blackwood and collaborators, the effect of the excision reaction during DNA mismatch repair (recombination of the zeocin tandem repeat) can still be noticed at a distance of 6.3 kb on the origin proximal side of the TNR array and the rate of recombination is a function of distance between the TNR array and the zeocin tandem repeat (Blackwood et al., 2010). On the other hand, Grilley and collaborators observed that the *in vitro* excision reaction terminates at different discrete sites at a region of 100 bp from the mismatch (Grilley et al., 1993). It is possible that a fraction of the total events of the *in vivo* excision reaction reaches further. However most of them clearly terminate at different discrete points in a "preferred region" near the mismatch on the origin proximal side of the mismatch.

We consider two features of MMR that might determine the length of the excision reaction –

1. The end point of an excision tract depends on the position of the mismatch (irrespective of its start point, Figure 4.6A & B).
2. The end point of an excision tract depends on its start point (a hemimethylated GATC on the origin distal side of the mismatch) (Figure 4.6C & D).

Again, the usage of zeocin tandem repeat and TNR array can be used to identify the feature from which the length of the excision reaction tracts is measured. Inserting the zeocin tandem repeat at different positions on the origin proximal side of the TNR array can provide information about the distribution of excision tract lengths during DNA mismatch repair (Figure 4.7A).



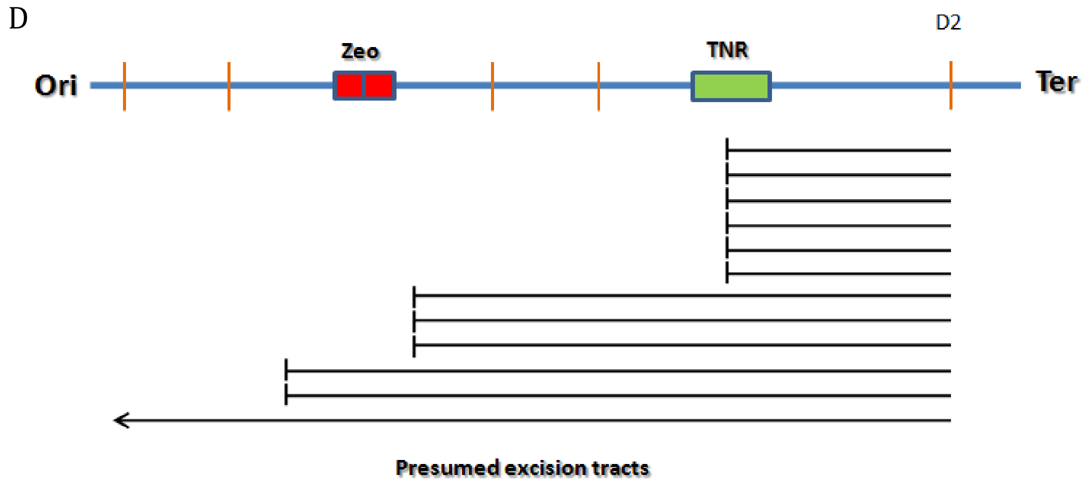


Figure 4.6. Model diagrams describing the excision reaction tract during DNA mismatch repair and its underlying features. Based on Chapter 3, the excision reaction starts at the first hemimethylated GATC motif (D1) on the origin distal side of the TNR array and the excision tracts moves towards the TNR array (or a mismatch in natural conditions). The excision tract can terminate at different discrete points on the origin proximal side with variable lengths. The chromosome of *E. coli* is shown as a blue line, the TNR array as a green box. The zeocin tandem repeat is depicted by two red boxes and the GATC motif (positioned arbitrarily) by small amber vertical lines. The presumed excision tracts are shown in black lines (both solid and dashed). The functional lengths (denoted as solid lines and the contributing feature of the excision tract length) are shown in solid lines. (A & B) The end point of the excision reaction might depend on the position of the TNR array (where the end point of the excision reaction would remain the same upon modification of D1) or (C & D) on the position of the starting GATC motif (where modification of D1 would affect the end point of the excision reaction).

Alternatively, using a fixed position for the zeocin tandem repeat and shifting the next available GATC motif on the origin distal side of the TNR array, thereby shifting the start point of the excision reaction, can be used to identify whether the start position of the excision reaction affects the end point (Figure 4.7B).

The second approach had been implemented in this study by inserting the zeocin tandem repeat at 500 bp away on the origin proximal side of the TNR array. If the termination point really depends on the position of a mismatch (or the TNR array in this study), the rate of recombination of the zeocin tandem repeat will remain the same despite any change of the start point of the excision tract (Figure 4.6A & B). On the other hand, if the length of the excision tract depends on the start point, sequential modification of the origin distal GATC motifs will result in differences in the recombination rate of the zeocin tandem repeat (Figure 4.6C & D).

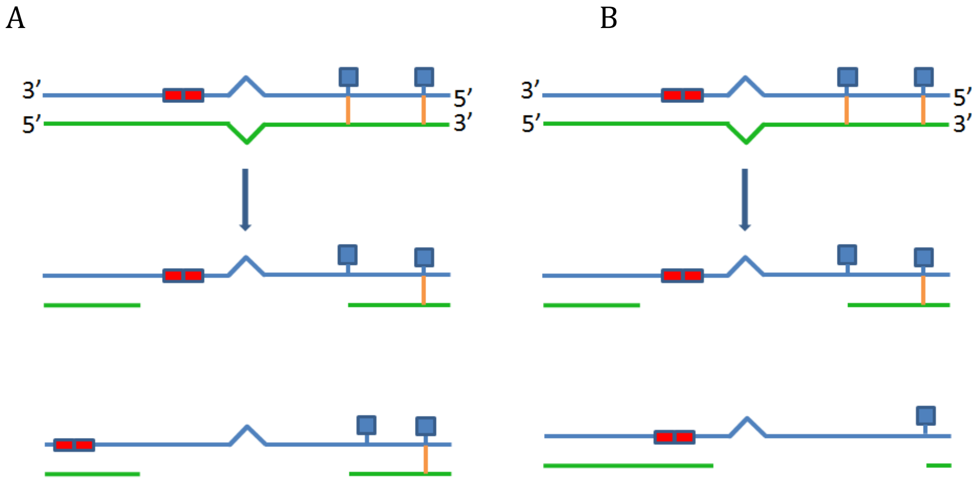


Figure 4.7. Two experimental approaches for investigating the distributions of excision tracts during MMR. Approach (A) shows placing the zeocin tandem repeat at different positions to determine the features that determine the end point of the excision reaction tract, while approach (B) shows using a fixed position of the zeocin tandem repeat and modifying the GATC motifs to shift the start point of the excision reaction to do the same. The parent strands are shown as blue lines and the nascent are as green lines. GATC motifs are shown as small vertical amber lines and the methyl groups are shown as blue boxes. The zeocin tandem array is shown by two tandem red boxes.

A mismatch proficient strain harbouring the *lacZ::*TNR array and the zeocin tandem repeat at 500 bp on the origin proximal side of the TNR array was subjected to sequential modification of GATC motifs that are origin distal to the TNR array (Figure 4.8). The GATC motifs on the origin distal side are denoted as D1, D2 ... *etc* (Figure 4.8).

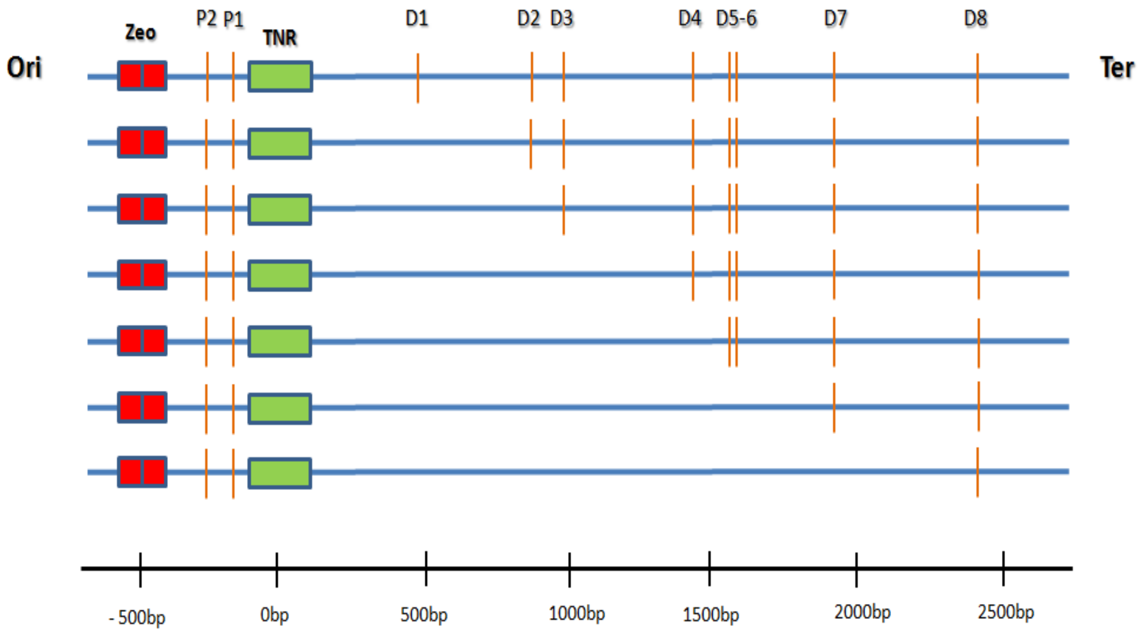


Figure 4.8. A schematic representation of the experimental approach for sequential modification of GATC motifs on the origin distal side of the TNR array to determine the length covered by the excision reaction during repair of a mismatch. Sequential modification of these GATC motifs will shift the starting point of the excision reaction. The *E. coli* chromosome is shown as blue horizontal lines, the TNR array is shown as a green rectangle and the zeocin tandem array is shown as two tandem red boxes. GATC motifs are shown as small vertical amber line with their respective names based on their position (origin distal GATC motifs are named Ds and origin proximal GATC motifs as Ps).

In the absence of the TNR array, the rate of recombination at the zeocin tandem repeat was very low as there was no frequent mismatch that the DNA mismatch repair could deal with and as a consequence, there was no induced single stranded DNA region for recombination (all TNR$^-$ backgrounds in Figure 4.9). In addition, even in the presence of the TNR array, a DNA mismatch repair defective background (*mutS*$^-$) did not show any increased recombination at the zeocin tandem repeat as there was no excision reaction taking place (MMR$^-$ TNR$^+$ background in Figure 4.9). In the wild type strains (wt TNR$^+$) where the DNA mismatch repair system was functional, the rate of recombination increased more than 7 fold compared to the DNA mismatch repair deficient background (MMR$^-$ TNR$^+$). When the first GATC motif on the origin distal side of the TNR array (D1 at a distance of 472 bp) was modified and the next GATC motif was at a distance of 844 bp, the rate of recombination stayed basically the same. A similar rate of recombination prevailed until the next available GATC motif was at a distance of 949 bp (D3) on the origin distal side. However, the rate of recombination started to decrease when the third GATC motif (D3) was modified and the next available GATC motif was 1356 bp away from the TNR array on the origin distal side. For the strain D1-D3 TNR$^+$, the decrease was about 15%, but was not statistically significantly different from the rate of recombination observed in the wild type background (wt TNR$^+$). A gradual decrease in the rate of recombination of the zeocin tandem repeat was observed upon further sequential modification of GATC motifs on the origin distal side of the TNR array. A 19% decrease of the rate of recombination from the wt TNR$^+$ background was observed when the next available GATC motif was 1826 bp away on the origin distal side of the TNR array and this decrease was statistically significant at 95% confidence interval. Surprisingly, the rate of recombination did not decrease to the level observed in backgrounds that were devoid of the TNR array or as the DNA mismatch repair deficient background (MMR$^-$ TNR$^+$) even after the modification of the seventh GATC motif (D7) on the origin distal side of the TNR array, where the next available GATC motif was beyond 2 kb away. In addition, the rate of recombination remained at a similar level even when the next available GATC motif was pushed 4.3 kb away from the TNR array (by inserting a 2 kb GATC motif-free sequence from *Schizosaccharomyces pombe* on the origin distal side of the TNR array).

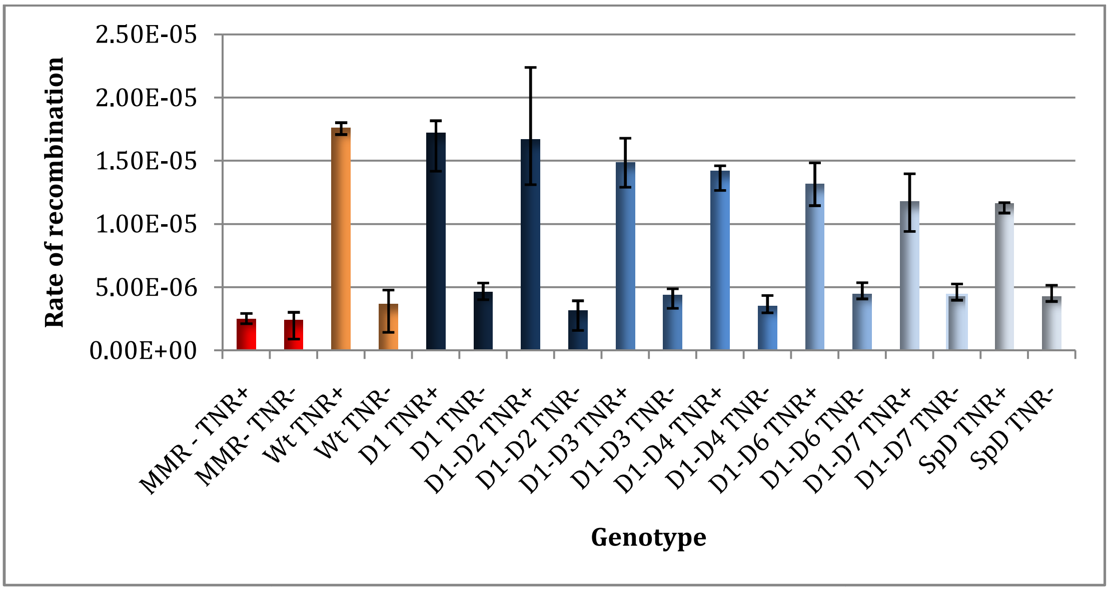


Figure 4.9. The distance of an excision tract depends on its start point, rather than the position of a mismatch. In this bar plot, the rate of recombination at the zeocin tandem array is shown along the *y*-axis and corresponding genetic backgrounds are shown along the *x*-axis**.** The error-bars represent the upper and lower limits of the 95% confidence interval.

In order to quantify the influence of the distance between the mismatch and the GATC site recognised on the origin-distal side of the TNR array, I have plotted the relative recombination efficiency (following respective equations in Section 2.2.4) as has been done previously for the relative efficiency of MMR (Figure 3.9). Here I have defined the relative recombination efficiency as 1 for the wt TNR$^+$ strain with wild type GATC sites and MMR$^-$ TNR$^+$ strain to have a relative recombination efficiency of 0. As can be seen in Figure 4.10, the recombination efficiency is sensitive to the distance between the first available GATC site and the TNR array (or the zeocin cassette). However, the slope of the exponential fitted curve is approximately twice as shallow as that for MMR (compared to Figure 3.9). A 50% recombination efficiency is reached after 6 kb of separation between first available GATC site and the TNR array (6.5 kb to the zeocin cassette). From Chapter 3, it has been established that the DNA mismatch repair system has a preference for GATC motifs on the origin distal side of a mismatch and it loses its efficiency as the distance from the TNR array to the next available GATC motif increases. Here, the rate of recombination decreases as the next available GATC motif moves further away from the TNR array. Therefore, the sequential decrease of the rate of recombination at the zeocin tandem repeat is actually a reflection of the position of the start point of the excision reaction, rather than the position of the mismatch itself. Initiation of excision is, therefore, the contributing feature determining the termination point of the reaction.

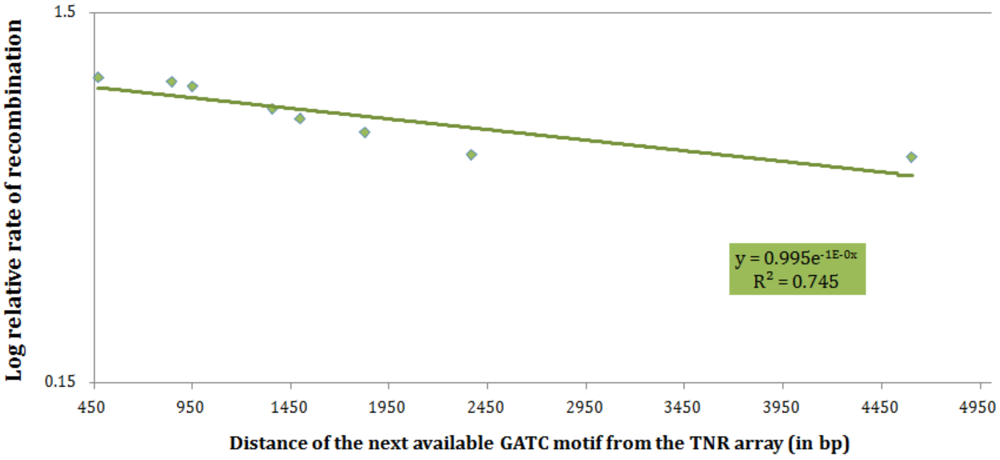


Figure 4.10. The influence of the distance between a mismatch and the GATC site recognised on the origin-distal side of the TNR array. The distances between the TNR array and the next available GATC motif on the origin-distal side are plotted along the *x*-axis while the relative efficiencies of the DNA mismatch repair system are plotted along the *y*-axis. The exponential trendline has been drawn using built in module of microsoft excel and the R^2 value is shown in the graph.

The rate of recombination of the zeocin tandem repeat in the strain SpD TNR^+ did not decrease to the level observed for the mismatch repair deficient backgrounds (all MMR^- backgrounds) even though the efficiency of DNA mismatch repair decreased near to the level of a MMR^- background (Chapter 3). This suggested that the enhanced recombination at the zeocin tandem repeat was not the product of a normal excision tract. I therefore considered whether the reaction might be caused by an MMR-coupled helicase-based molecular event responsible for a longer ssDNA tract. RecQ, a non-MMR helicase with 3'→5' directionality, might be responsible for such events (Bachrati and Hickson, 2003). DNA helicases unwind complementary strands of nucleic acids in a reaction coupled to NTP hydrolysis. The RecQ helicase in *E. coli* falls under a family of conserved enzymes required for maintaining genomic integrity. It has been found in bacteria, fungi, animals and plants, and copy number of the gene ranges from one in *E. coli* and *S. cerevisiae* up to seven in *Arabidopsis thaliana*. All RecQ helicases purified and characterized to date unwind duplex DNA with a 3'→5' directionality based on the DNA strand they bind. *E. coli* RecQ has a broad DNA substrate specificity: from DNA duplexes containing blunt or forked termini to 3- or 4-way Holliday junctions (Harmon and Kowalczykowski, 1998). Therefore, upon excision of a faulty nascent strand by the orthodox MMR machinery, the RecQ helicase might bind the nascent strand at the ssDNA-dsDNA junction and unwind it further in a 3'→5' direction.

To determine whether the RecQ helicase associates with the MMR system to create a longer single stranded region, the *recQ* gene was knocked out in the SpD strain that harbours a 4.3 kb GATC motif-free region on the origin distal side of the TNR array. In this strain, the distance between the next available GATC motif and the zeocin tandem repeat is 4.8 kb. If RecQ were responsible for the longer ssDNA tract, the rate of recombination would be lower in the absence of a functional RecQ protein as there would be no long ssDNA tracts reaching the zeocin tandem repeat.

However, upon analysis, the rate of recombination was at a similar level that was observed for the SpD TNR$^+$ background (Figure 4.11). In addition, its derivative control background that did not harbour the TNR array (SpD *recQ*$^-$ TNR$^-$) showed a similar level of recombination to that was observed in TNR$^-$ backgrounds. Therefore, RecQ helicase is not coupled to a sub-population of highly processive excision tract that was thought to affect the recombination at a distant zeocin tandem repeat during MMR.

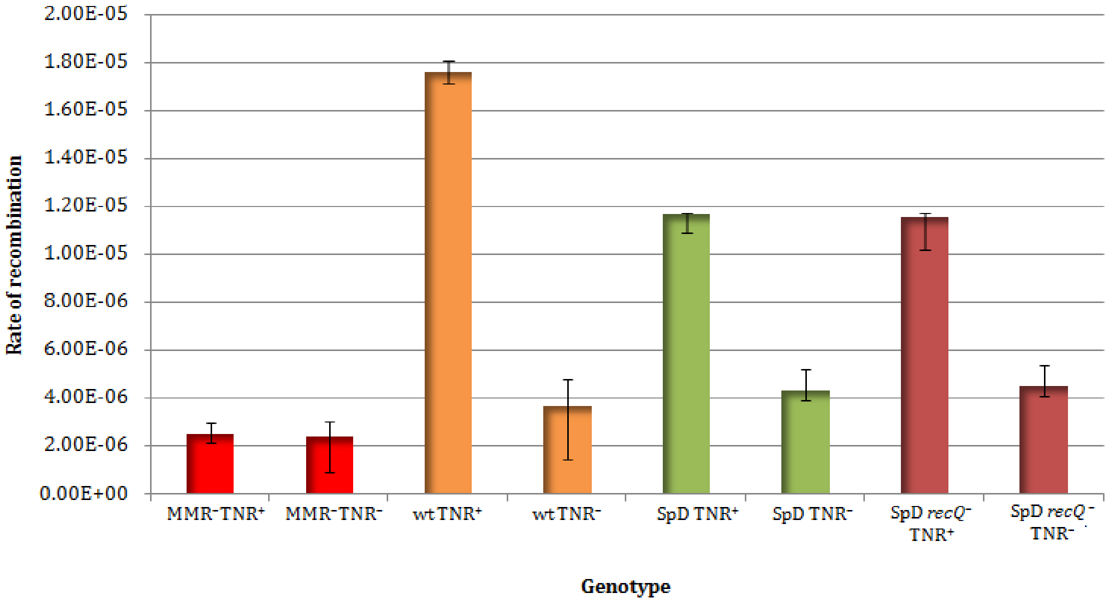


Figure 4.11. RecQ helicase is not responsible for the long ssDNA tract generated during the excision reaction of the mismatch repair. In this bar plot, the rate of recombination at the zeocin tandem array is shown along the *y*-axis and corresponding genetic backgrounds are shown along the *x*-axis**.** The error-bars represent the upper and lower limits of 95% confidence intervals.

4.4 Discussion

The DNA mismatch repair system is a well-conserved molecular process in almost all organisms that maintains genomic integrity (Jun et al., 2006). Though the mode of strand discrimination varies from prokaryotes to eukaryotes and even within prokaryotes, the excision reaction is well-conserved in all organisms (Claverys and Lacks, 1986; Kadyrov et al., 2006, 2007). In this study, we have shown that in *E. coli* the excision reaction starts from a GATC motif on the origin distal side of a mismatch and the efficiency of the DNA mismatch repair decreases as the available GATC motif moves further from the mismatch on this side (Chapter 3). For an energy efficient event of mismatch repair (as well as to avoid futile MMR events), the MMR system should excise the faulty nascent strand beyond the mismatch and synthesise an error-free one. From this chapter, it has been found that the excision reaction is not influenced by a hemimethylated GATC motif on the origin proximal side of a mismatch. *In vitro* experimentations have shown that the excision reaction terminates at different discrete sites within 100 nucleotides beyond the mismatch (Grilley et al., 1993). Interestingly, the *in vivo* effect of mismatch repair can be detected as far as 6.3 kb (on the origin proximal side) by the recombination of a 275 bp tandem repeat that is proposed to be mediated by repair synthesis of nascent strand which is complementary to the single stranded DNA formed during the excision reaction (Blackwood et al., 2010). Even in this current study, the SpD TNR^{+} strains, where the next available GATC motif was 4.8 kb away from the zeocin tandem repeat, showed about 5 fold more recombination relative to a mismatch repair deficient background. Given that there is no reason to expect there to be two different modes of recognition of the same origin-distal GATC site, I hypothesise that two classes of excision tracts exist. MMR is primarily mediated via short excision tracts that need not extend far beyond the mismatch and long excision tracts are responsible for recombination at a greater distance beyond the initiating mismatch. It was hypothesised that there might be a MMR-coupled helicase based system contributing to this event. In this study, the association of RecQ helicase in that process has been examined and found to be unresponsive.

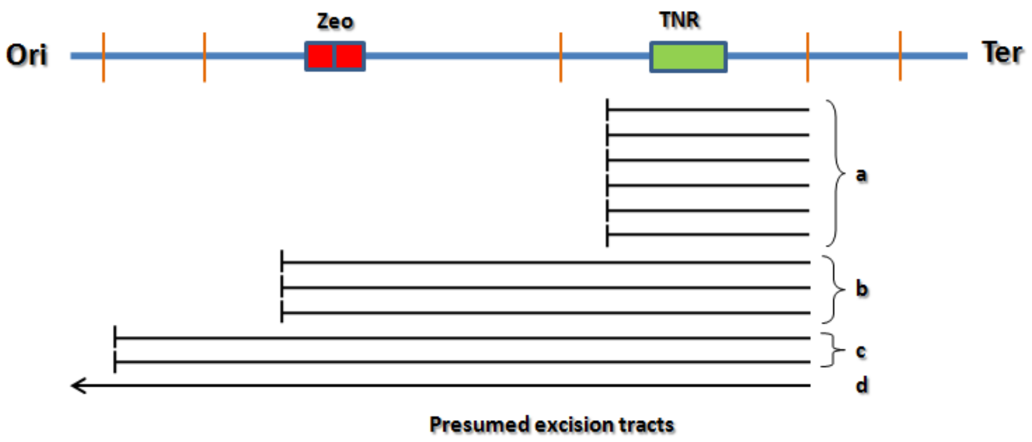


Figure 4.12. A model diagram showing the existence of longer single stranded DNA tracts along with smaller excision tracts generated during DNA mismatch repair. The excision reaction starts at the first hemimethylated GATC motif on the origin distal side of the TNR. Presumably, a large proportion of the excision tracts terminate at a short distance beyond the TNR array (a & b), some extend further (c); even a very small fraction of all excision tracts extends even further away (d). The chromosome of *E. coli* is shown as blue lines, the TNR array as a green box, the zeocin tandem repeat as two red boxes and the GATC motifs (positioned arbitrarily) are shown as small amber vertical lines. The presumed excision tracts are shown in black lines with a vertical line depicting the termination points (termination points at close regions are shown to be the same).

One cannot rule out an additional, but currently unknown, molecular mechanism that would be coupled to the standard DNA mismatch repair system to achieve this reaction. The fact that two different reactions, which are initiated at the same GATC sites, are differentially sensitive to the distance of the first origin-distal GATC site from the TNR array argues strongly that the recognition of the GATC site is not sensitive to distance from the mismatch but instead that the lengths of the two different classes of excision tract determine the ability of a distant GATC to stimulate MMR or recombination. There could be a preferred region for the end points of the excision reaction. However, in this study it is difficult to distinguish (if they are different at all) the most preferred distance of an excision reaction tract (Figure 4.12a) from others (Figure 4.12b, c, d). Grilley and collaborators had designed an *in vitro* experiment using electron microscopy for detecting the length of the single stranded DNA generated during excision reaction (Grilley et al., 1993). A clever optimization of that experiment for an *in vivo* system and thereby generating a frequency distribution of the lengths of the single stranded DNA would answer the questions that remained unanswered in this study.

Chapter 5

DNA mismatch repair does not decelerate a progressing replication machinery in *Escherichia coli*

5.1 Introduction

DNA mismatch repair is a rapid and efficient molecular process. There is a small window of time for the DNA mismatch repair (MMR) system to repair a mismatch left by the DNA replication machinery before GATC motifs get methylated by the Dam methylase (Barras and Marinus, 1989; Urig et al., 2002). In addition, as has been shown in Chapter 3, the MMR system is directional in relation to DNA replication. In addition, some components of the DNA mismatch repair system interact with the β clamp in bacteria or its eukaryotic equivalent – the Proliferating Cell Nuclear Antigen (PCNA) (López de Saro et al., 2006; Warbrick, 2000, 2006). Both β clamp and PCNA are ring shaped components of the DNA replication machinery that encircle the DNA as a clamp and ensure the processivity of the DNA replication machinery, and are therefore generally referred to as "sliding clamps" (López de Saro et al., 2006; Stukenberg et al., 1991). Though various proteins involved in DNA metabolism bind to the sliding clamps, they all share a common aspect in interacting with the β clamp or PCNA (Dalrymple et al., 2001; Warbrick, 2000). They all bind to the clamp via N- or C-terminal flexible extensions containing a short motif and the interaction takes place at a hydrophobic pocket near the C-terminus of the clamp. This interaction occurs in a complex fashion as there is more than one binding surface involved and in the case of PCNA, post-translational modifications regulate the interaction (Bunting et al., 2003; Friedberg et al., 2005; Hoege et al., 2002; Indiani et al., 2005). Moreover, the β clamp is dimeric and the PCNA is a trimeric protein. Therefore, it has been presumed that their homo-oligomeric entity can accommodate more than one ligand simultaneously and can serve as a mobile platform to coordinate sequential actions of multiple enzymes on DNA in a single molecular process (Fujii and Fuchs, 2004; Indiani et al., 2005). DNA polymerase III and IV have been found to establish such two point interactions with the β clamp (Bunting et al., 2003; Chapados et al., 2004; Dohrmann and McHenry, 2005; Gomes and Burgers, 2000). The sliding clamp (as a component of DNA replication machinery) is hypothesized to carry and load component(s) of DNA mismatch repair system in the same manner during early steps of mismatch repair as it interacts with MutS and MutL proteins (Pluciennik et al., 2009).

Interestingly, MutS protein can bind the β clamp via two points – one is at the N-terminus and the other is at the C-terminus, which has the stronger affinity. Deleting the C-terminal motif (^{812}QMSLL^{816}) abolishes the *in vitro* interaction between MutS and the β clamp in the absence of DNA, but does not confer the hypermutability that was found for *mutS* mutants *in vivo* (López de Saro et al., 2006). In contrast, mutating an alanine residue in the (^{15}QQYLRL^{20}) motif in the N-terminus does not affect the interaction between MutS and β clamp, but conferred strong *in vivo* hypermutability in *E. coli* (Pluciennik et al., 2009). Therefore, it has been proposed that the N-terminus of the DNA-bound MutS is the most important point of interaction with the β clamp. On the other hand, MutL binds the β clamp in the presence of single-stranded DNA and the contact point is located on a loop in the N-terminal ATP-binding domain of MutL (Pluciennik et al., 2009). Mutations of two residues, which retained the MutL ATPase activity and its interaction with helicase II, reduces the interaction with the β clamp and abolishes mismatch repair *in vivo*(Pluciennik et al., 2009). In addition, expression of this mutated MutL does not complement a *mutL* defective strain of *E. coli* which indicates that the interaction of MutL and the β clamp is essential for DNA mismatch repair (Pluciennik et al., 2009).

The precise role of these processivity clamps in DNA mismatch repair is yet to be deciphered. The MSH2-MSH6 hetero-dimer has been found to bind PCNA in the absence of a mismatch, but not in the presence of a mismatch which suggests that PCNA could help the MSH complex locating mismatches on the DNA (Flores-Rozas et al., 2000; Lau and Kolodner, 2003). In addition, the clamp binds to DNA at the junction of a single and a double strand DNA in a distinct orientation (Jiricny, 1998; Kolodner, 1996). Based on the aforementioned physical interactions between components of these two molecular machineries and their probable roles in DNA mismatch repair, the replication machinery is presumed to load the component(s) of the DNA mismatch repair system.

We have considered the possibility that, if the DNA replication machinery loads MutS on to the mismatch, the MutS-β interaction might drag the replicase backward by pulling on to the β clamp, as well as, on the DNA polymerase holo-enzyme (Figure 5.1A). Alternatively, simultaneous occupation of one binding site of the β clamp by MutL and the other one by the DNA polymerase III might drag the polymerase III backward (towards the mismatch) by forming a loop structure (Figure 5.1B). This loop structure would be necessary for the early DNA mismatch repair component(s) (*e.g.* MutS, MutL and MutH) to interact with other components at the excision reaction start point as a hemimethylated

GATC motif might be located kilo-bases away from the mismatch (Guarné et al., 2004; Jiricny, 1998; Wang and Hays, 2004). These two scenarios might arrest the DNA replication machinery for a brief time, which would consequently result in the accumulation of DNA replication intermediate structures in the vicinity of a mismatch. I have set out to find whether the DNA replication machinery slows down its normal pace due to the repair of a mismatch that is generated as a result of a replication error.

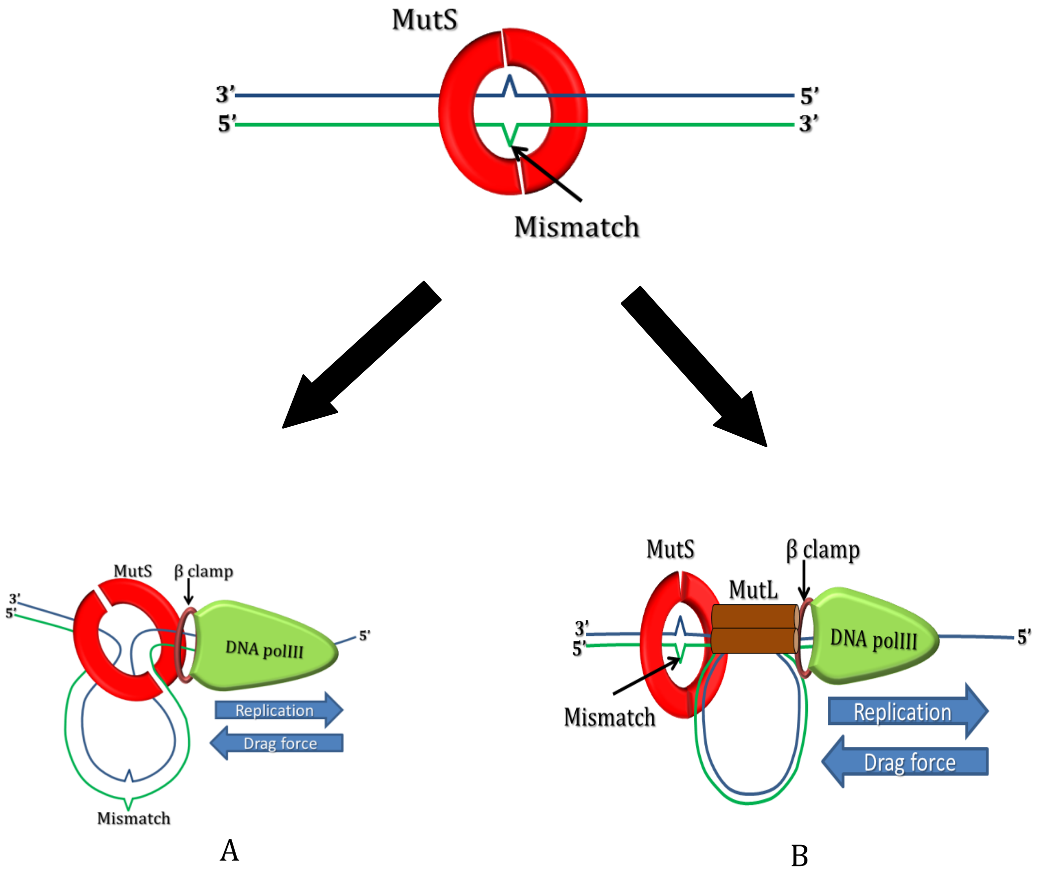


Figure 5.1. A model for proposed back-tracking of the DNA replication machinery by the interaction between MMR proteins and the β clamp. Once the replication machinery leaves a mismatch behind, the first component of the DNA mismatch repair system, MutS binds the mismatch. A) The interaction between MutS and the β clamp might exert a drag force on the replicase in the opposite direction of the DNA synthesis and forms a DNA loop. B) Simultaneous occupation of one binding site of the β clamp by MutL and the

other by the DNA polymerase III might result in the polymerase III back-tracking to the point of the mismatch by forming a loop structure. This schematic representation is adopted from (López de Saro and O'Donnell, 2001; López de Saro et al., 2006)

5.2 *priB* mutation to increase replication intermediates that can be captured by 2-D agarose gel electrophoresis

Genome duplication is a rapid, accurate and highly processive process (Cadman et al., 2005). However, the replication machinery may stall or even derail in all organisms as it is sensitive to DNA damage including single-stranded nicks, gaps, double-stranded breaks and modified bases (Cox, 2001; Cox et al., 2000). Therefore, it is imperative that the replication machinery must be able to reassemble back onto the chromosome to resume the replication and maintain genomic integrity (Kuzminov, 1999). Since origin-dependent loading of the replisome is well-regulated, the reassembly of replisome at non-origin sequence should also require a controlled set of mechanisms (Lopper et al., 2007). There are multiple overlapping mechanisms for reassembling a dissociated replication fork in *E. coli* (McCool et al., 2004; Sandler, 2000). These reassembly pathways are initiated either by PriC or by PriA helicase, PriB and DnaT. However, the latter is considered to be the most important pathway (Heller and Marians, 2005a, 2005b; Lee and Kornberg, 1991; Nurse et al., 1991; Sandler et al., 1996). PriA helicase binds preferentially to three-way branched DNA and D-loop structures at some stalled replication forks and at recombination intermediates, and unwinds the DNA in the 3'→5' direction (Figure 5.2) (McGlynn et al., 1997; Nurse et al., 1999). In the case of a stalled fork, PriA unwinds the lagging strand to produce a single-stranded DNA that will ultimately be utilized as the loading site for the replicative helicase DnaB as it cannot bind a SSB coated single stranded DNA substrate (Jones and Nakai, 1999; LeBowitz and McMacken, 1986; Xu and Marians, 2003). Reloading DnaB is considered to be a crucial step in restarting replication by allowing the reassembly of both leading and lagging strand polymerases on to the respective DNA strands and the resumption of priming of lagging strand synthesis by the primase – DnaG (Yuzhakov et al., 1996). However, PriA does not facilitate DnaB reloading directly, rather it paves the way for that. PriA binds the leading strand at a branched DNA structures (stalled forks or D-loops) followed by recruitment of PriB and subsequently of DnaT (Liu et al., 1996; Xu and Marians, 2003). This PriA-PriB-DnaT complex then promotes reloading of DnaB onto the lagging strand template (Heller and Marians, 2005a).

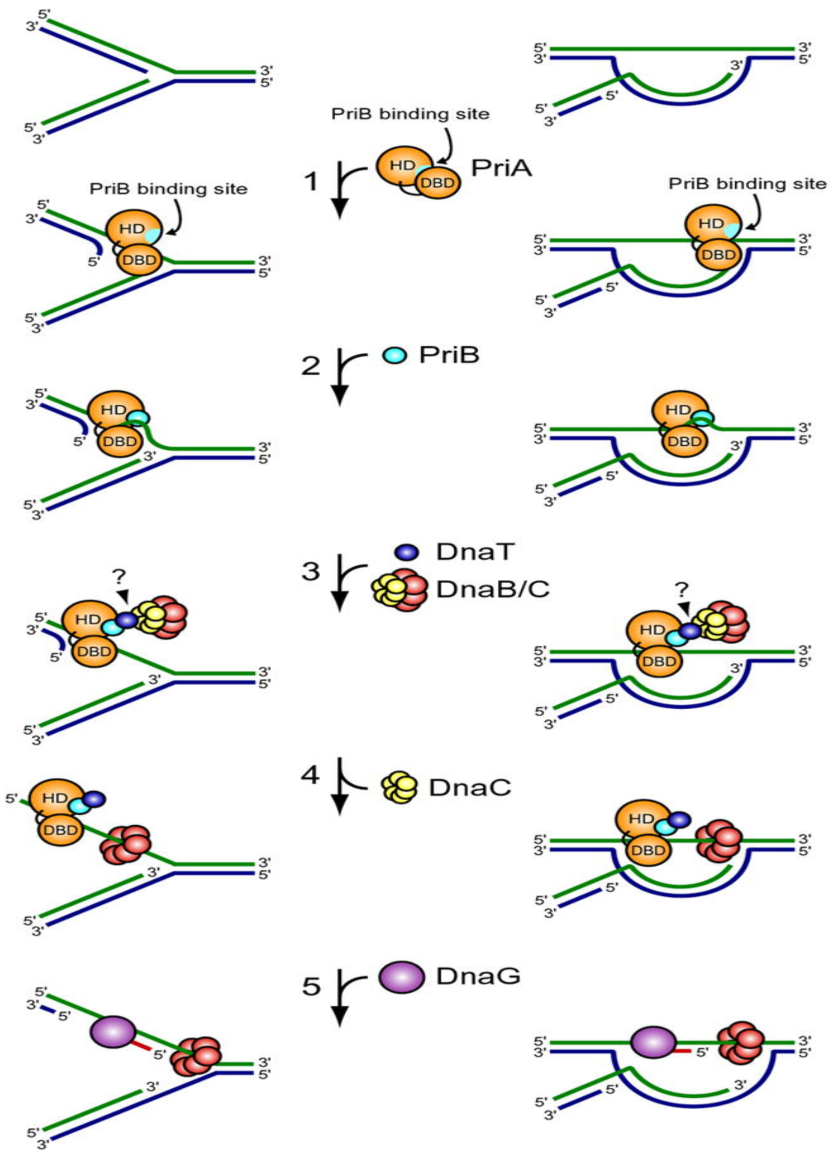


Figure 5.2. Model for primosome assembly. (1) PriA binds a repaired DNA replication fork or D-loop and undergoes a conformational change, exposing the PriB binding site on the PriA HD. (2) PriB binds the PriA:DNA nucleoprotein complex. The PriA:PriB:DNA ternary interaction stabilizes PriA on the DNA and enhances its helicase activity and facilitates the unwinding of nascent lagging strand (if one is present). (3) DnaT is recruited to the PriA:PriB:DNA ternary complex and binds PriB. This interaction causes release of ssDNA by PriB. The DnaB/C complex is recruited to the primosome, perhaps through direct contacts with DnaT. (4) DnaB is loaded from the DnaB/C complex onto ssDNA on the lagging strand template. (5) Recruitment of DnaG allows RNA primer synthesis from which the polymerase III holoenzyme can synthesize a nascent lagging strand. Figure is from (Lopper et al., 2007).

PriB interacts with naked ssDNA or with ssDNA bound by SSB (Allen and Kornberg, 1993) and modulates PriA helicase function (Lee and Marians, 1989). PriB facilitates the interaction between PriA and DnaT by stabilising PriA and DnaT binding on the DNA. Lopper and collaborators proposed a model for the reassembly of replication machinery, which would start with the binding of PriA helicase to a stalled replication fork or D-loop (Lopper et al., 2007). This binding would impose a conformational change in those DNA structures exposing the PriB binding site on the helicase domain (HD) of the PriA. PriB binding to the PriA HD and ssDNA (either pre-existing or created by the helicase activity of PriA) would form a ternary complex. Binding of this ternary complex by the DnaT would release a ssDNA region which would then be bound by the replicative helicase, DnaB (with the helicase-loader DnaC) and the replication process could resume. In this study, *priB* knock-out mutants have been used to maximise the visualization of DNA replication intermediates as the pathway to restart a stalled replication is impaired.

5.3 Native two-dimensional agarose gel electrophoresis (2-D agarose gel) to capture DNA replication intermediates near the TNR array

A native two-dimensional agarose gel electrophoresis (2-D agarose gel) can acquire direct evidence of DNA replication intermediate structures using the right selection of restriction sites around the region of interest in a chromosomal DNA. Unlike fully replicated or fully unreplicated DNA, these intermediate structures form in the process of DNA replication and are non-linear. For example, a partially replicated DNA fragment created by restriction digestion, that does not contain an origin of replication, will take on a variety of "Y-shaped" conformations (Figure 5.3B). Similarly, a DNA fragment containing an origin of replication will start out as a bubble immediately after the initiation of replication and will convert to a Y-shape when replication machinery passes one restriction site (Figure 5.3A).

In the first dimension of native 2-D agarose gel electrophoresis, DNA fragments are separated based on molecular weight at a low voltage and 4°C to maintain a native condition. In the second dimension, high voltage and a DNA intercalating agent – ethidium bromide (0.3μg/l) are applied to maximise the separation of the DNA fragments based on molecular shape of different DNA species. The pattern of DNA migration during the second dimension leads to unequivocal evidence of an origin of replication. Non-linear DNA intermediates of the same sized restriction fragments that are formed during replication have very unusual three-dimensional structures. These fragments will migrate differently in the second dimension and form an arc-shaped structure (Figure 5.4). This arc is formed by the migration of "Y-shaped" molecules and is called the "Y-arc". It is indicative that the DNA molecule is in the process of being replicated (Figure 5.4). A Y-shaped molecule that has three equal length arms will migrate the most slowly and will be at the top of the arc of non-linear DNA (blue line in Figure 5.4A). In contrast, a Y-shaped molecule with two very short replicated arms or a large replicated region will migrate very similarly to the unreplicated version of the same DNA fragment. Importantly, the Y-shaped molecule of an almost completely replicated fragment has a similar shape to a linear molecule, but is almost double in size of the unreplicated fragment. Therefore, these fragments will reside at the other corner of the arc. A DNA fragment containing an origin of replication will form bubble-shaped replication intermediates, which migrate slower than Y-shaped molecules. Unfortunately, it is difficult to distinguish an intermediate created by this type of fragment (indicated by the bubble-arc) from that is created by a Y-shaped intermediate if the origin of replication is at the centre of the fragment (Figure 5.4). Choosing the right combination of restriction sites, so that the origin of replication is not in the centre of the fragment, will create a discontinuity in the arc. This so called “bubble-to-Y-transition” occurs as the replication fork that is close to the end of the fragment will become Y-shaped from the initial bubble formed at the origin. It is highly indicative of an origin of replication. This two-stepwise separation of DNA fragments results in a distinct pattern of spot and arc upon Southern blotting using radio-labelled DNA probes against the DNA fragment of interest (Figure 5.4).

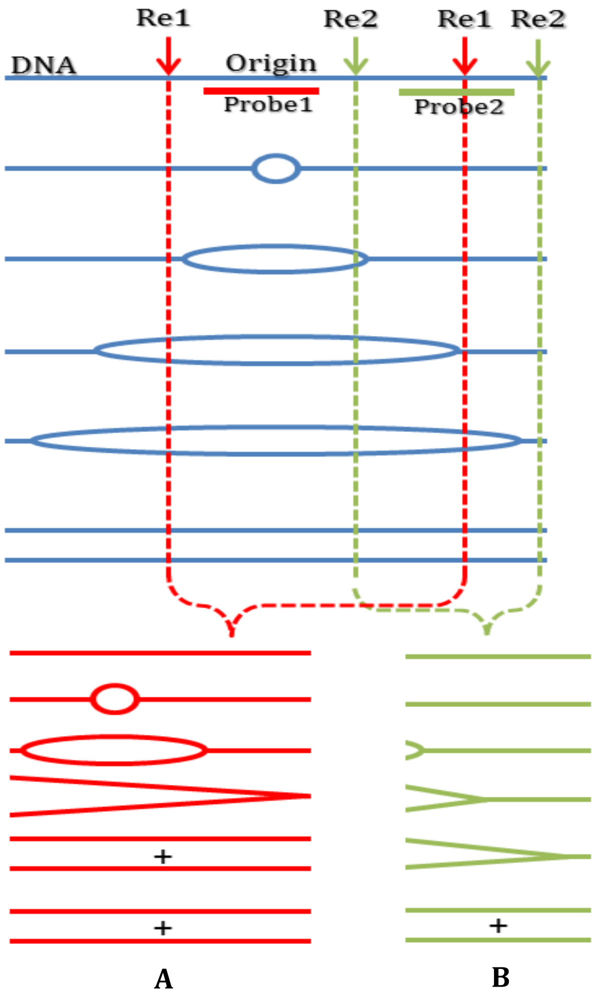


Figure 5.3. Capturing unusual DNA intermediate structures during the process of replication using restriction digestions. This illustration shows the expansion of a "replication bubble" (created by two replication forks progressing away from an origin of replication). Restriction digestion of those replication intermediates followed by migration on a 2-D agarose gel, transfer of the DNA on a membrane and detection by hybridization with the indicated labelled DNA probes will permit the isolation of distinct molecular structures of DNA. (A) If the red restriction enzyme (Re1) is used and only the fragments that hybridise to the red DNA probe are examined, the DNA replication intermediates indicated in red colour will be generated. (B) If the green restriction enzyme (Re2) and the green DNA probe are used to detect the resulting DNA fragments, the DNA replication intermediates indicated in green colour will be observed. The left-hand pattern (red) starts with a DNA fragment containing a "bubble" and eventually ends with "Y-shaped" molecules. The right hand pattern (green) never has a "bubble" and generates a full variety of "Y-shaped" intermediates.

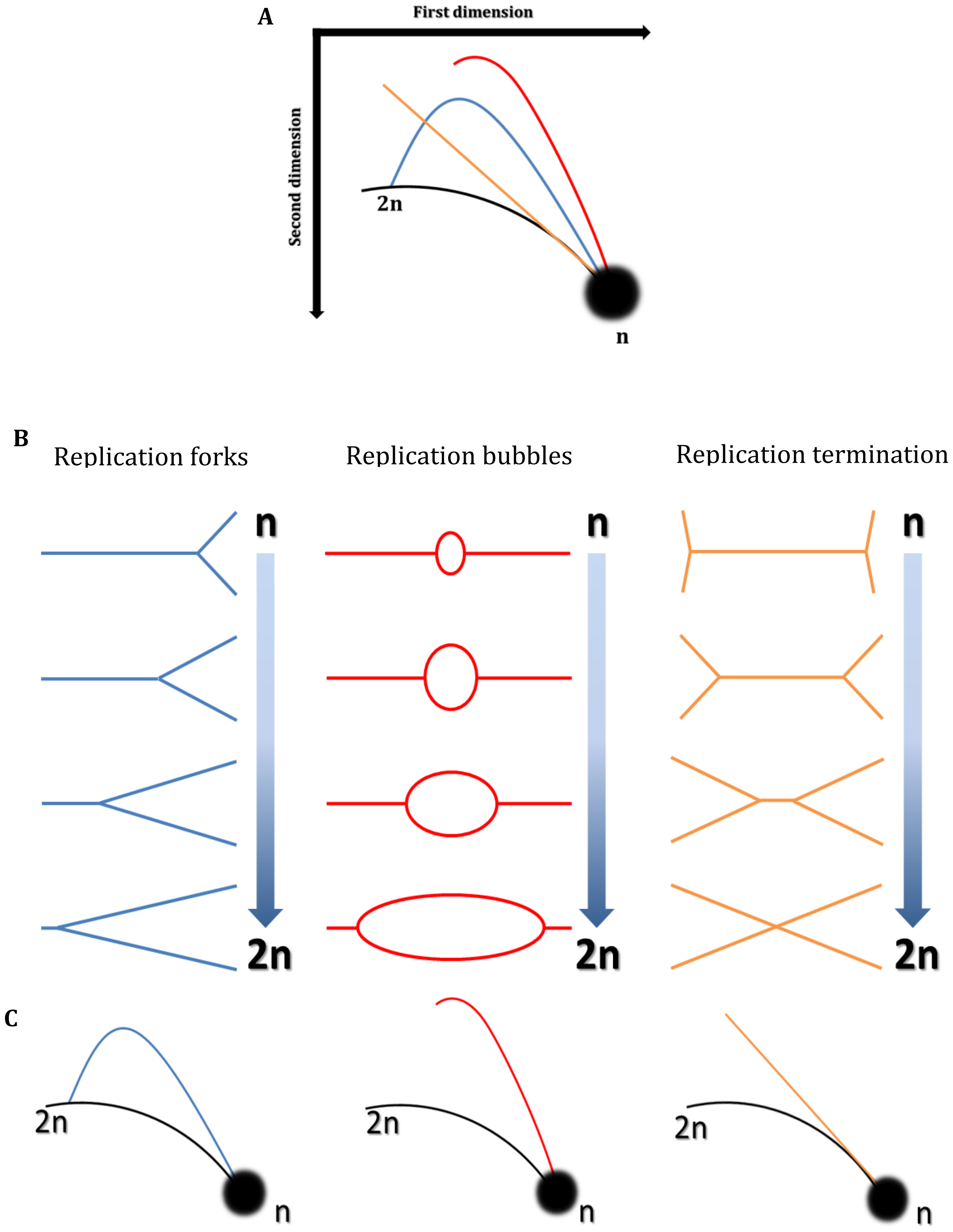


Figure 5.4. Overview of 2-D agarose gel electrophoresis migration patterns. The large spots designated "n" indicate the positions of the abundant linear species within the restriction fragments. "2n" indicates the location of these near-linear species just prior to completion of replication. Accumulation of a particular structure, such as a blocked replication fork at a specific location

along a restriction fragment, generates a spot along the relevant migration line. (A) Overview of the most common migration patterns observed. The arc of linear DNA is represented by the thin black line, which runs through n and 2n. (B) Breakdown of the different molecular shapes placed above the migration pattern they generate for replication forks (blue colour), replication bubbles (red colour) and replication termination (amber colour). Figure adapted from (Friedman and Brewer, 1995).

5.4 No DNA intermediate structures are detected at or near the *lacZ::*TNR array during mismatch repair

Upon detecting a mismatch, MutS binds to it and recruits the MutL protein at that site to form a ternary complex. This ternary complex in turn recruits MutH that searches for a hemimethylated GATC motif on the origin distal side of the mismatch (Chapter 3). However, hemimethylated GATC sites are only generated just after the replication machinery passes those GATC sites while synthesising a nascent DNA strand. Therefore, if a drag force was applied on the DNA replication machinery due to an interaction with the MMR machinery, a replication pause might occur at a site distant from the TNR array on the origin distal side (Figure 5.1). Based on the physical interactions of β clamp with MMR components (MutS and MutL) and its probable role in DNA mismatch repair, I predicted that the replication machinery would feel a drag force towards origin of replication while MutH is searching for the hemimethylated GATC motifs and MutS and/or MutL is still physically attached to the β clamp (Pluciennik et al., 2009).

As described before, *priB*$^{-}$ knock-out mutants were used in this study to increase the amount of DNA replication intermediates at stalled replication forks as the pathway to restart a stalled replication is impaired. Halted replications should be more frequent in the strain harbouring the *lacZ*::TNR array with an efficient MMR system. Therefore, DNA replication intermediates should be easier to detect in the strain *lacZ::*TNR^{+}*priB*$^{-}$*mutS*$^{+}$ than in the strains without the *lacZ::*TNR array *(lacZ::*TNR^{-}*priB*$^{-}$*mutS*$^{-}$and *lacZ::*TNR^{-}*priB*$^{-}$*mutS*$^{+}$*)* or with an impaired DNA mismatch repair system (*lacZ::*TNR^{+}*priB*$^{-}$*mutS*$^{-}$) using 2-D agarose gel electrophoresis and Southern blotting followed by hybridising with radio-labelled probes.

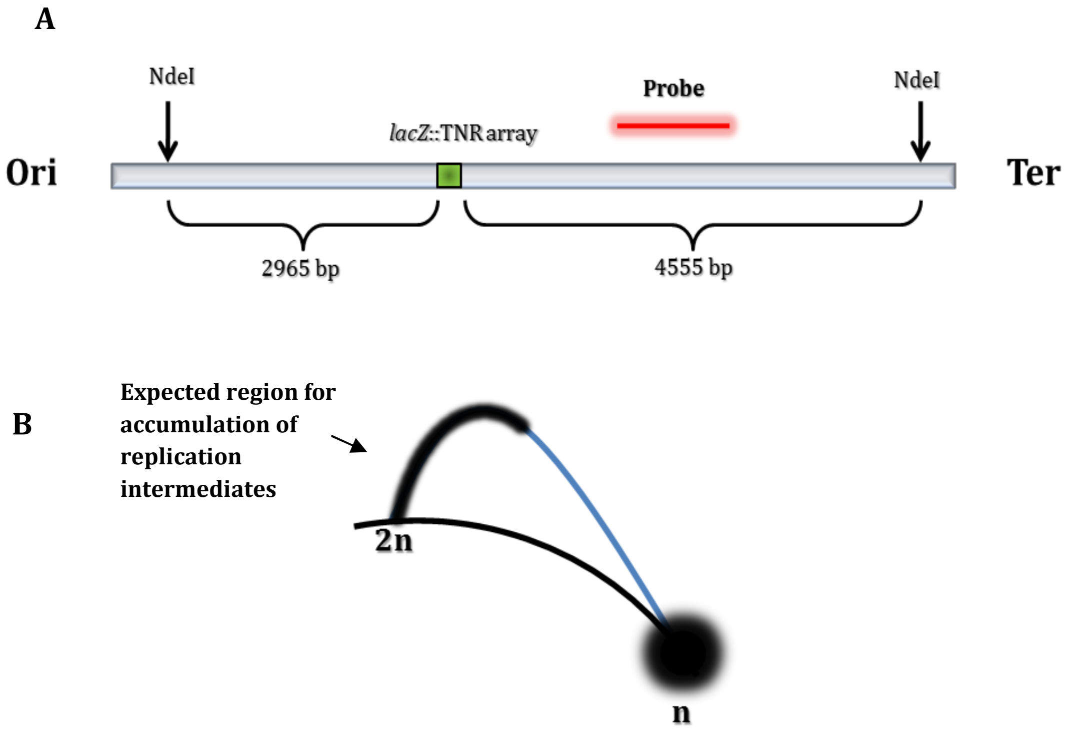

A
NdeI
Probe
NdeI
lacZ::TNR array
Ori
Ter
2965 bp
4555 bp
B
Expected region for accumulation of replication intermediates
2n
n

Figure 5.5. A schematic presentation of the hypothetical scenario where an advancing replication machinery could pause due to mismatch repair. A DNA fragment of the *E. coli* chromosome containing the *lacZ*::TNR array was obtained by NdeI restriction digestion that was analyzed by 2-D agarose gel electrophoresis and Southern bloting. (A) The map of the restriction sites (denoted by vertical arrows) and the probe binding site (the probe is shown as a red horizontal line) relative to the TNR array (as green box) are shown. (B) Expected migration patterns of DNA fragments in 2-D agarose gel electrophoresis as a result of a replication pause near the TNR array due to the interaction between the MMR machinery and the replisome.

Figure 5.5A shows a schematic representation of the locus of interest in the *E. coli* chromosome which has been subjected to 2-D agarose gel electrophoresis after digestion by the NdeI restriction enzyme. The *lacZ::*TNR array is situated at a 1:1.5 distance from the origin distal side of the digested DNA fragment. Therefore, if the replication fork slows down due to the interaction with MutS and/or MutL during repairing a mismatch at the *lacZ*::TNR array, a range of intense signal (corresponding to the section between the *lacZ*::TNR array and 2n of the “Y”-arc) would be observed on the “Y”-arc upon 2-D agarose gel electrophoresis and Southern blotting followed by hybridising with radio-labelled probes (black region on the “Y”-arc in Figure 5.5B).

A radio-labelled probe had been used to detect the restriction digested DNA fragment after Southern transfer of the DNA to a nitrocellulose membrane and radio-labelling. Surprisingly, no region in the “Y”-arc was found to have intense signal in the strain *lacZ::*TNR^{+}*priB*$^{-}$*mutS*$^{+}$ compared to other control strains (*lacZ::*TNR^{+}*priB*$^{-}$*mutS*$^{-}$,*lacZ::*TNR^{-}*priB*$^{-}$*mutS*$^{+}$ and*lacZ::*TNR^{+}*priB*$^{-}$*mutS*$^{-}$) (Figure 5.6). Therefore, no paused or stalled replication machinery has been detected at the CAG•CTG repeat array while synthesising a nascent strand in *E. coli* even in the presence of mismatches that are being repaired.

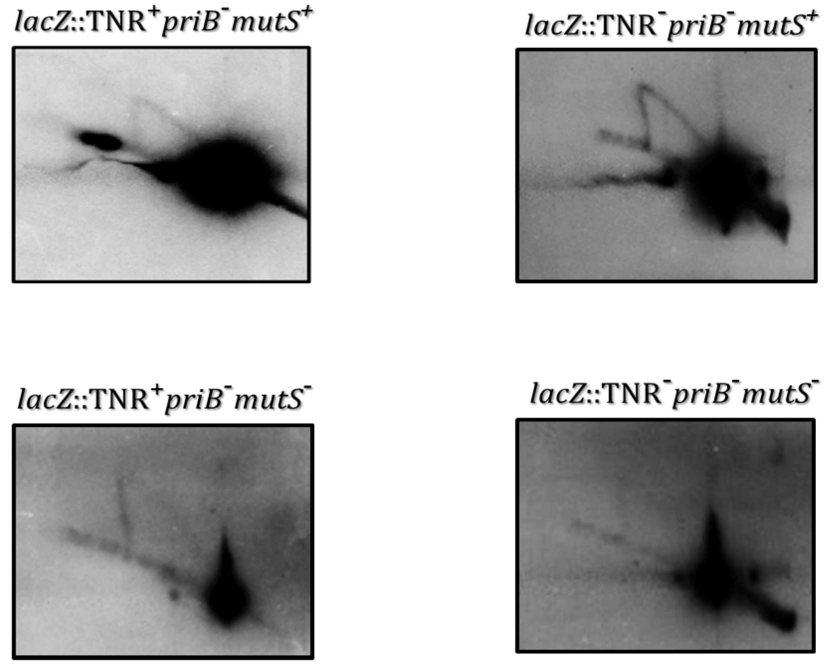

lacZ::TNR$^+$*priB*$^-$*mutS*$^+$
lacZ::TNR$^-$*priB*$^-$*mutS*$^+$
lacZ::TNR$^+$*priB*$^-$*mutS*$^-$
lacZ::TNR$^-$*priB*$^-$*mutS*$^-$

Figure 5.6. DNA replication is efficient in replicating a CAG•CTG repeat array without any pausing detected in the *E. coli* chromosome. A DNA fragment of the *E. coli* chromosome containing the TNR array was obtained by NdeI restriction digestion and analyzed by 2-D agarose gel electrophoresis and Southern bloting followed by hybridising with radio-labelled probes. 2-D agarose gel electrophoresis for strains - *lacZ::*TNR$^+$*priB*$^-$*mutS*$^+$*, lacZ::*TNR$^-$*priB*$^-$*mutS*$^-$*, lacZ::*TNR$^+$*priB*$^-$*mutS*$^+$*, lacZ::*TNR$^-$*priB*$^-$*mutS*$^-$.

In addition, an *E. coli* strain harbouring a long GATC motif free sequence on the origin distal side of a mismatch should accumulate more DNA replication intermediates while MutH would keep searching for a hemimethylated GATC motif than a strain with intact GATC sites. Therefore, the strain *lacZ::*TNR$^+$*priB*$^-$*mutS*$^+$*D1-D7* should accumulate more intermediates than its MMR$^-$ derivative strain (*lacZ::*TNR$^+$*priB*$^-$*mutS*$^-$*D1-D7*). The interaction between the MMR machinery and the replisome, which might cause a drag force, could occur in close proximity of the protein components by forming a DNA loop over a distance between the mismatch and the GATC site recognised. As the next available GATC motif is 2.3 kb away from the TNR array in the NdeI restriction digested fragment, accumulated DNA replication intermediates in the strain *lacZ::*TNR^{+}*priB*$^{-}$*mutS*$^{+}$*D1-D7* would likely have an even longer replicated arms and shorter unreplicated arm than what would be found in the strain *lacZ::*TNR^{+}*priB*$^{-}$*mutS*$^{-}$*D1-D7*. Therefore, an intense signal might be detected near the 2n end of the "Y-arc" after 2-D agarose gel electrophoresis and Southern blotting followed by hybridisation using radio-labelled probe. Again, no accumulation of replicative DNA intermediates were observed along the Y-arc in the strain *lacZ::*TNR^{+}*mutS*$^{+}$*priB*$^{-}$*D1-D7*, rather it looked similar to the control *lacZ::*TNR^{+}*mutS*$^{-}$*priB*$^{-}$*D1-D7* strain (Figure 5.7). Therefore, no drag force (towards origin of replication) was detected on the replisome exerted by the MMR machinery during MMR.

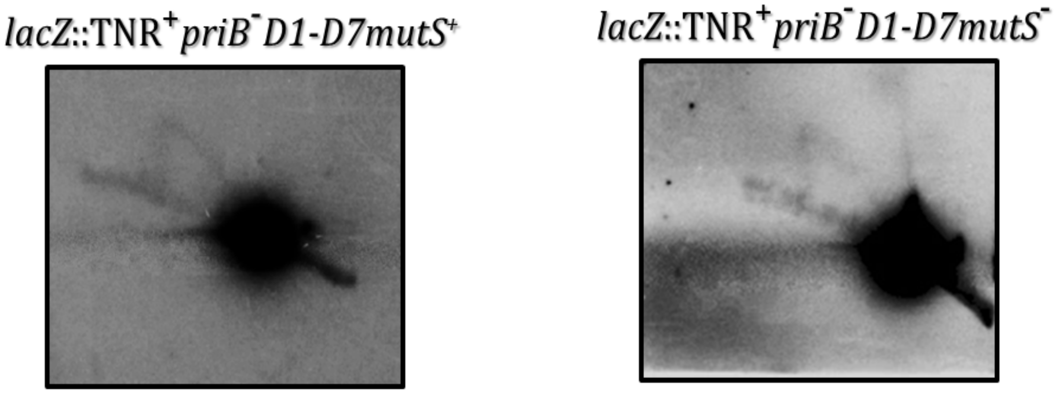

lacZ::TNR^{+}*priB*$^{-}$*D1-D7mutS*$^{+}$
lacZ::TNR^{+}*priB*$^{-}$*D1-D7mutS*$^{-}$

Figure 5.7. Effect of mismatch repair on DNA replication fork at a region free of GATC motifs on the origin distal side of the TNR array. 2-D agarose gel electrophoresis for strains *lacZ::*TNR^{+}*priB*$^{-}$*D1-D7mutS*$^{+}$and*lacZ::*TNR^{+}*priB*$^{-}$*D1-D7mutS*$^{-}$.

5.5 Discussion

It has now been well established that MutS and MutL proteins have physical interactions with the β clamp or PCNA (Pluciennik et al., 2009). In addition, simultaneous two point contacts have been observed at the β clamp interacting with different proteins, like polymerases III and IV (Bunting et al., 2003; Chapados et al., 2004; Dohrmann and McHenry, 2005; Gomes and Burgers, 2000). In this regard, binding the β clamp by N-terminal clamp-interactive residues in MutS or MutL presumably forms an articulated and hinge-like complex that allows β clamp binding and yet permits their rapid recruitment during mismatch repair. Therefore, I set out to test whether the functional and/or physical interaction between the β clamp and components of mismatch repair machinery creates a drag force on the replisome in the opposite direction of replication during mismatch repair. Impairing the pathway that resumes stalled replisomes by knocking out the *priB* gene meant that distinctive DNA replication intermediates could be visualised, providing a background of stalled replication forks above which any effect of MMR might be detected. In this study, a CAG•CTG repeat array has been used as a source of substrates for MutS binding and a stalled replisome has been searched for near the TNR array using 2-D agarose gel electrophoresis and Southern blotting followed by hybridisation with radio-labelled probes. However, from figure 3.3, it has been found that only 6.5% of the DNA molecule in *E. coli* undergoes MMR after incorporating mismatches. This minor event might be hard to detect within the range of sensitivity of Southern blotting followed by hybridisation with radio-labelled probes.

Again, upon 2-D gel electrophoresis and Southern blotting, "Y"- arcs were observed in every strain under study. It confirms the occurrences of replication pause during the normal course of DNA replication in *E. coli* as the chromosome is sensitive to DNA damages including single-stranded nicks, gaps, double-stranded breaks, and modified bases (Cox, 2001; Cox et al., 2000). In a *priB*$^{-}$ background, all events of paused replication would not resume instantly if they were stopped during mismatch repair as PriA-PriB-DnaT mediated pathway is the major contributor for resuming stalled or paused replication (Heller and Marians, 2005a, 2005b; Lee and Kornberg, 1991; Nurse et al., 1991; Sandler et al., 1996). As a result, they would add to the total population of paused replication events, as well as, subsequently be detected at a defined area of the "Y" arc. However, the lack for sensitivity to detect such events using the Southern blotting could be a reason for not getting a conclusive signal even though there are such events occurring at that locus.

In addition, Krasilnikova and Mirkin have shown that the replication machinery stalls at a 70 unit long GGC repeat cloned in a plasmid in *E. coli* and at a 40 unit long GGC repeat cloned in a plasmid in *Saccharomyces cerevisiae* using 2-D agarose gel electrophoresis and Southern blot (Krasilnikova and Mirkin, 2004). However, in this current study, no replication stalling has been detected at 98 unit long CTG•CAG repeat array in the *E. coli* chromosome even though the PriB mediated replication restart pathway has been impaired. GGC repeats can form stable hairpin, triplex and even intrastrand quadruplex structures (Ji et al., 1996). On the other hand, the favourable secondary structure of CTG repeats is a hairpin (Mirkin, 2007). More stable complex structures of GGC repeats could impose more obstacles for the replication machinery to replicate through a long GGC repeat in *E. coli*. In addition, there is a difference between the replication of a plasmid and that of the chromosome (Kovtun and McMurray, 2008).

Chapter 6

No selection pressure, exerted by the DNA mismatch repair system, has been found on the genomic distribution of GATC motifs

6.1 Introduction

Being a widely studied prokaryotic model organism, *Escherichia coli* has an extraordinary position in the study of genetics, molecular biology and biotechnology. Therefore it was one of the earliest organisms whose genomes was sequenced (Blattner et al., 1997). Its publicly accessible genome sequence has stimulated further research towards a more complete understanding of many biological processes in *E. coli,* which have also contributed to the discovery of relevant processes in higher eukaryotes. Regulation of these processes is basically mediated by the recognition of different DNA motifs that are non-randomly distributed throughout the genome of an organism. DNA motifs are defined as short, repeated sequences that have biological functions (Touzain et al., 2011). By and large, motifs are classified into two categories (D'haeseleer, 2006).

1. *cis*-acting regulatory motifs that are the binding sites for the transcription regulators or sigma factors.
2. Motifs that are involved in the maintenance of the chromosome and bind nucleoid-associated proteins such as integration host factor (IHF), Fis, H-NS *etc* (Dillon and Dorman, 2010).

A list of some important *E. coli* specific motifs is given in Table 6.1.

This chapter focuses on the distribution of GATC motifs in the *E. coli* genome and addresses whether there is any selection pressure on their distribution imposed by the DNA mismatch repair system that utilises these motifs to maintain the integrity of the genome.

Table 6.1. Overview* of some important motifs in *E. coli* genome.

Name of the motif	Motif sequence	Characteristics of the sequence	Interacting protein	Distribution	Biological function
DnaA box	TTATNCACA	Non-palindromic nanomer	DnaA protein	Clustered at the origin of replication. Also at secondary sites 107 found in total	Controls the initiation of replication
GATC motif	GATC	Palindromic tetramer	Dam methylase, SeqA, MutH	Clustered at the origin of replication in *E. coli* Total 19120 in *E. coli* genome	Control of chromosome replication, nucleoid segregation, DNA mismatch repair, regulation of transcription *etc*.
Ter motif	GN(A/G)NGTTGT AA(C/T)(T/G)A	23 bp non-palindromic with 14 bp core and directional	Tus protein	10 strong and 4 weaker motifs in *E. coli* genome	Termination of replication
Chi (crossover hot spot instigator)	GCTGGTGG	Non-palindromic octamer	RecBCD protein	1008	Repair of DNA double strand break

* Adapted from (Touzain et al., 2011).

6.2 Biological processes that involve GATC motifs

6.2.1 GATC motifs in the DNA mismatch repair system

Though the replicative DNA polymerase (PolIII) has its own proof reading activity, sometimes it incorporates a wrong nucleoside tri-phosphate in the nascent DNA strand. The wrong nucleotide (a mismatch) in the newly synthesized DNA is recognized by the DNA mismatch repair system (Schofield and Hsieh, 2003). As the replication machinery has just passed that site, the adenine of the GATC motif on the nascent strand is yet to be methylated by the Dam-methylase protein (Bruni et al., 1988; Lahue et al., 1987). In *E. coli*, the DNA mismatch repair system discriminates the nascent strand that contains the mismatch from the parent DNA strand by sensing the methylation state of the GATC motif(s) around the mismatch. Moreover, the GATC motif acts as the starting point of the excision reaction in the course of the DNA mismatch repair mechanism (Schofield and Hsieh, 2003). This reaction removes the faulty nascent strand and recruits the replicative polymerase to synthesise the correct sequence. Thus, the GATC motif acts as a crucial signal to maintain genomic integrity in *E. coli*.

6.2.2 GATC motifs in DNA replication

Replication of the *E. coli* genome is initiated at a genetically unique origin sequence, *oriC*. Although rapidly growing cells contain multiple copies of the origin of replication, the initiation of replication occurs synchronously – once and only once per cell cycle. Secondary initiations are prevented by a sequestration process at the newly replicated origins mediated by the SeqA protein (Lu et al., 1994; Slater et al., 1995). The SeqA protein acts as a negative modulator of replication initiation by binding hemimethylated GATC motifs at *oriC* and the *dnaA* promoter. SeqA is responsible for the binding of the cell membrane to the hemimethylated GATC motifs in the *oriC* sequence. Complete methylation of GATC motifs in the *oriC* sequence is associated with resumption of the initiation of the next round of replication. Though SeqA interacts with the fully methylated GATC motif, its binding to a hemimethylated motif is stronger (Lu et al., 1994; Slater et al., 1995). The later interaction is thought to be the most important one in this respect.

6.2.3 Effect of GATC motifs in transcription control mediated by temperature shift

Transcriptional control mediated by temperature shift has been postulated to be affected by the methylation of GATC motifs. It has been reported that the melting temperature of a 10 bp oligonucleotide decreases up to 13°C upon methylation of GATC motif(s) in the sequence (Hénaut et al., 1996). Therefore, hemimethylated or unmethylated GATC sites, particularly in clusters, can change the stability of the double helix and transcription of a locus might be lowered or impeded at low temperature (relative to a sequence with fully methylated GATC motifs). This decrease in temperature would presumable occur when *E. coli* comes out from the host environment (high temperature and osmolarity, but low concentration of oxygen) to an external environment (low temperature and osmolarity, but high concentration of oxygen) during its life cycle.

6.2.4 Regulatory property of GATC motifs

An unusual distribution of GATC motifs has been found in recognition sites of the upstream regulatory sequences of genes (*fnr*, *crp etc*) that are involved in respiratory regulation and DNA replication (Hénaut et al., 1996). It has been reported that global regulators like CRP, Fnr and IHF restrict Dam-mediated methylation of many GATC motifs (Tavazoie and Church, 1998; Wang and Church, 1992). This shows us that the recognition sites of these global regulators coincide with the Dam-recognition sites (GATC motifs). Therefore, it has been proposed that the methylation state of the GATC motifs modulates the transcription of different genes by limiting the access of the global regulators that bind in their upstream regulatory sequences (Oshima et al., 2002). This process is well documented for the control of the expression of *pili* gene as the methylation state of two GATC motifs within the regulatory region of the pyelonephritis-associated pilus (*pap*) operon controls *pap* transcription by limiting the binding of two regulatory proteins: leucine-responsive regulatory protein (Lrp) and pap-encoded co-regulatory protein (PapI) to it (Blyn et al., 1990; Braaten et al., 1994; van der Woude et al., 1998).

Involvement in the aforementioned biological processes has made genomic distribution of the GATC motif particularly interesting. This study has focused on the distribution of GATC motif in *E. coli* to determine whether these biological roles have shaped the evolution of the genomic distribution.

6.3 The distribution of GATC in the *E. coli* genome

GATC is one of the few reported DNA motifs that have high biological significance. At first the distribution of the GATC motifs in the *E. coli* genome has been investigated (Figure 6.1). The number of GATC motifs per 10 kb in the *E. coli* chromosome is shown by dark-grey bars in the Figure 6.1. Some loci of the genome are found to harbour more GATC motifs than others. Some clusters of GATC motifs are found - especially at least 5 loci are found to have clusters of 6 GATC motifs within only 100 bp. An important cluster of 11 GATC motifs is present at the *oriC*, the origin of replication of the *E. coli* chromosome. However, some other loci do not have any GATC motif for a greater length. Small red circles in the Figure 6.1 show the longest seven GATC free regions (> 3.2 kb). The sequences of these regions will be discussed in greater details in the later part (Section 6.7) in this chapter.

6.4 GATC motifs are over-represented in the *E. coli* genome

Considering the important biological impacts of GATC motifs in the genome of different prokaryotic organisms, it is compelling to look closely into their chromosomal distribution. Being a tetra-nucleotide motif, theoretically, GATC motif should appear once in 4^4 = 256 nucleotides in a genome with randomly distributed nucleotides. In that sense, there should be about 18120 GATC motifs in the 4.6 Mb long genome of the *E. coli*. In reality, the *E. coli* genome has about 1000 more instances of GATC motifs (19120 GATC motifs in *E. coli* str. K-12 substr. MG1655, NCBI reference sequence: NC_000913.2) than what would be expected by chance in a genome of the same length (4639675 bp) with randomly distributed nucleotides.

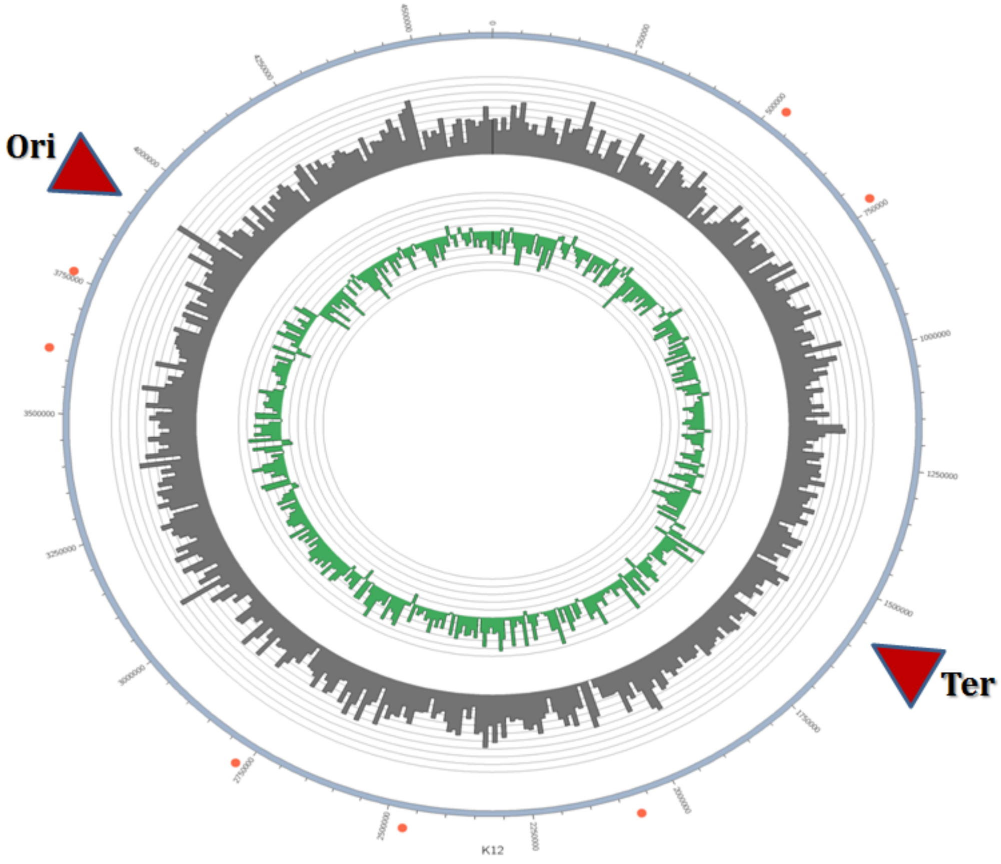


Figure 6.1. The distribution of GATC motifs in the *E. coli* genome. In this CIRCOS graph, the light blue circle represents the 4.6 Mb circular chromosome of *E. coli* with small scale marks indicating 50 kb each. The origin of replication "*Ori*" and the terminus site "Ter" are shown at their approximate loci. Dark grey bars represent the number of GATC per 10 kb region (the axis imposed here spans from 0 to 80 and divided into 10 circular grids) and green bars represent the GC skew in the *E. coli* genome (the axis imposed here spans from -0.10672986 to 0.106270358 and divided into 10 circular grids). GC skew completely changes its magnitude at origin or replication and terminus in *E. coli*. Seven small red circles represent the longest region of GATC motif-free sequences (greater than 2.8 kb) in the *E. coli* genome. The MMR system is 50% efficient when the distance between the TNR array and the next available GATC motif is 2.8 kb (Chapter 3). The positions of GATC and GC skew have been calculated using in-lab Perl scripts (Appendix B).

Now the obvious question arises – how significant (biologically) is the abundance of the GATC motifs in *E. coli* genome? At first glance, it is only about 5.5% above the predicted number. A χ^2 test has been performed on this data to measure its significance under the hypotheses:

H_0 = There is no difference between expected number and the observed number of GATC motifs in *E. coli* genome

H_1 = There is a difference between expected number and the observed number of GATC motifs in *E. coli* genome

As $\chi^2 > \chi^2_{k-1,1-\alpha}$,the null hypothesis is rejected, k - 1 = Degrees of freedom and α=0.05.

The whole genome of *E. coli* has been divided into equal 100 segments and based on the expected number of GATC motifs ($\frac{1 \times 46396.75}{256}$) in each segment randomly by chance, the χ^2 test has been computed. The test statistic appears to be 313.1655 which is far greater than the critical value (123.2252) under 95% confidence level. So, the null hypothesis is rejected with 95% certainty (degree of freedom = 99). Therefore, the observed number of GATC motif in the *E. coli* genome is very significant compared to what is expected by chance. However, to investigate the biological significance of this difference, it is important to understand the sequence composition of the whole genome.

6.4.1 The Rho(ρ) statistic of GATC

There are several ways to calculate whether a motif of a particular sequence-length is over- or under-represented than expected by chance. One way is to calculate the Rho(ρ) statistic. The Rho(ρ) statistic has been described by Karlin and Cardon to compute the over- or under-representation of any particular dinucleotide among their 16 possible combinations (Karlin and Cardon, 1994). For a 2-nucleotide motif, the Rho(ρ) statistic is calculated as:

$$\rho\ \mathrm{xy} = \frac{fxy}{fx \times fy}$$

Where, f_{xy} = frequency of dinucleotide “xy”,

f_x = frequency of nucleotide “x” and

f_y = frequency of nucleotide “y”.

The rationale behind the ρ statistic is that if a DNA sequence has the frequency f_x and f_y of two different nucleotides “x” and “y” respectively, the expected frequency of the dinucleotide “xy” will be the product of their individual frequencies, that is ($f_x \times f_y$). On the other hand, if the real frequency of the dinucleotide “xy” is found to be f_{xy}, it is expected to be equal to the product of the individual frequencies of the nucleotides that compose it. Therefore, in a perfectly random sequence, the ρ statistic would be equal to 1. The ρ statistic will be greater than 1 if the dinucleotide motif is more common in the sequence than expected or it is said to be “over-represented”. On the other hand, ρ will be less than 1 if the dinucleotide motif is less common in the sequence than expected or it is said to be "under-represented".

For example, the Rho(ρ) statistic of a dinucleotide AC would be –

$$\rho_{\mathrm{AC}} = \frac{fAC}{fA \times fC}$$

Where, f_{AC} = frequency of dinucleotide "AC",

f_A = frequency of nucleotide "A" and

f_C = frequency of nucleotide "C" in a DNA sequence.

$f_A = \frac{\text{number of the nucleotide "A"}}{\sum \text{number of all fournucleotides in the DNA sequence}}$ and so on.

This Rho(ρ) statistic can be adopted for a motif of any length in a DNA sequence.

Therefore, ρ statistic of the GATC motif would be –

$$\rho_{\mathrm{GATC}} = \frac{fGATC}{fG \times fA \times fT \times fC}$$

where, f_{GATC} = frequency of tetra-nucleotide "GATC",

f_G = frequency of nucleotide "G",

f_A = frequency of nucleotide "A",

f_T = frequency of nucleotide "T" and

f_C = frequency of nucleotide "C".

$$f_G = \frac{1176923}{1142228 + 1179554 + 1176923 + 1140970} = 0.254$$

$$f_A = \frac{1142228}{1142228 + 1179554 + 1176923 + 1140970} = 0.246$$

$$f_T = \frac{1140970}{1142228 + 1179554 + 1176923 + 1140970} = 0.246$$

$$f_C = \frac{1179554}{1142228 + 1179554 + 1176923 + 1140970} = 0.254$$

$$f_{GATC} = \frac{19120}{4639675} = 0.00412$$

Therefore, $\rho_{GATC} = \frac{0.00412}{0.254 \times 0.246 \times 0.246 \times 0.254} = 1.0555$

Thus, the GATC motif is indeed over-represented ($\rho_{\mathrm{GATC}} > 1$) in the *E. coli* chromosome. However, the value is slightly higher than 1. If we plot the Rho(ρ) statistic of all possible tetranucleotide motifs in *E. coli* genome along the *x*-axis in incremental fashion and compare the Rho(ρ) statistic of GATC motif with that of others (Figure 6.2), we see the CTAG motif is at the lower end and the TTTT motif is at the upper end of the distribution. The GATC motif falls in the middle of the distribution. However, the value barely crosses the threshold of 1. Therefore, the GATC motifs are only slightly over-represented in the *E. coli* genome.

6.4.2 The Markovian model

The Rho(ρ) statistic basically depicts the property of over- or under-representation of a DNA motif in the genome rather than providing a weighted value associated with it. To get a better understanding of the over-representativeness of the GATC motif, a probabilistic model (Markovian model) has been applied using the statistical program called R'MES. In Markovian approximation, a motif of size ℓ can be only analyzed in $M0$ up to $M(\ell - 2)$ orders because higher models would fit the motif count itself (the motif will then be expected by definition). So the expected number of the GATC motifs has been calculated up to the second order. Both a Gaussian approximation and a compound Poisson approximation method have been applied to calculate the expected occurrence of the GATC motif.

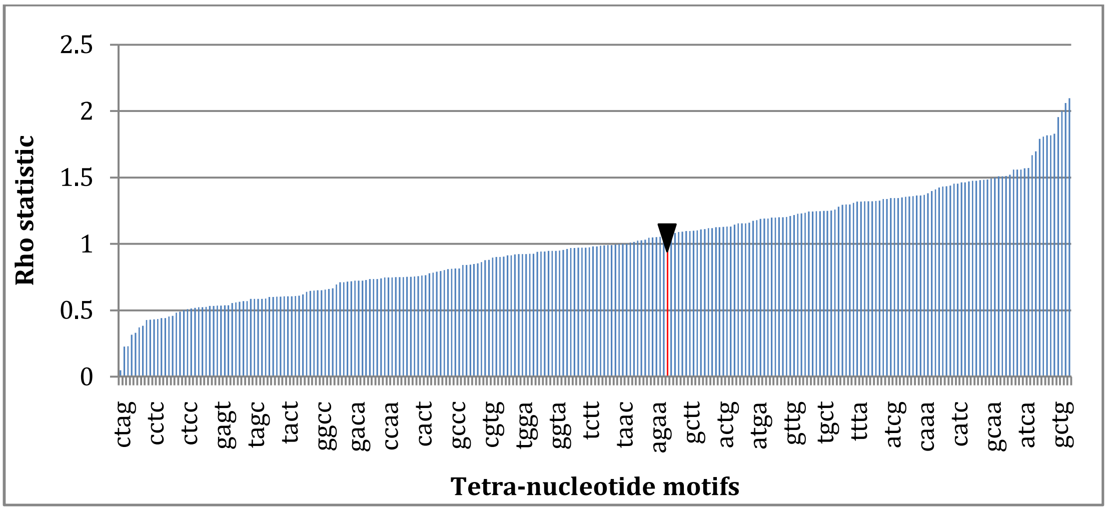


Figure 6.2. The distribution of Rho(ρ) statistic of all possible tetra-nucleotide motifs in the *E. coli* genome. The “seqinr” module in the statistical package R has been used to calculate the Rho(ρ) statistic for all possible tetra-nucleotides in the *E. coli* genome. The “red” bar (marked with a black triangle) among all “blue” bars correspond to the Rho(ρ) statistic for GATC motif. The tetra-nucleotides are plotted along the *x*-axis while the scale of Rho(ρ) statistic has been plotted along the *y*-axis. (Not all the names of motifs are visible due to the space constrains)

As the order differs in different methods, the number of the GATC motifs expected by chance differs as well. Both the Gaussian approximation and the compound Poisson approximation led to predict the same number of the GATC motifs expected by chance under the same order of Markov model. For a zero-order Markov model "M0" (Table 6.2), where the model did not consider the composition of the sequence that surrounded the GATC motif, both approximations estimated the expected number of GATC motif by chance to be 18114.662 and the scores differed slightly. The scores calculated from both approximations clearly showed that the GATC motif is over-represented in the *E. coli* genome (above the threshold value of 0). The rank (from least represented to most represented) was calculated among all the possible 256 tetra-nucleotides.

Table 6.2. The weight profile of the representation of the GATC motifs in the *E. coli* genome

Model	Algorithm	Fit	Count	Expected	Score	Rank
M0	Gauss	Bases	19120	18114.66	7.5739	148
M1	Gauss	Dinucleotides	19120	16981.39	17.823	182
M2	Gauss	Trinucleotides	19120	24160.71	-44.9941	8
M0	Compound Poisson	Bases	19120	18114.66	7.3992	141
M1	Compound Poisson	Dinucleotides	19120	16981.39	16.0815	175
M2	Compound Poisson	Trinucleotides	19120	24160.71	-33.6639	10

For the first-order Markov model (M1), the expected number was estimated to be 16981.393 from both approximations. By definition, the first-order Markov model calculates the expected number of a motif based on the nucleotides occupying the previous position. In this case the expected number had become lower that that observed from the zero order Markov model. The underlying reason was that the GATC motif itself was composed of three rare dinucleotides in the *E. coli* genome – GA, AT and TC which were ranked 6, 10 and 7 respectively (from least represented to most represented) among 16 possible dinucleotides. Therefore, the scores had become even higher in both approximations, which supported the over-representation of GATC motifs in the *E. coli* genome.

Interestingly, in the case of the second-order Markov model (M2), the expected number of GATC motifs was estimated to be 24160.719 in both Gaussian and Compound Poisson approximations, which contradicted the property of over-representation that was observed for other orders (M0 and M1). Under this model, the GATC motif was found to be under-represented relative to the expectation by chance and the scores turned out to be higher negative in both approximations. This phenomenon can be explained by the order of the Markov model used in this analysis. In the second order, the Markov model considered the nucleotides of two previous positions, which was analogous to looking at every possible tri-nucleotide in the *E. coli* genome. The GATC motif itself is composed of two highly frequent trinucleotides in the *E. coli* genome GAT and ATC which were ranked 52 and 51 respectively among the 64 possible trinucleotides and only a second-order Markov model considered this fact. This may, in fact, be the reason behind the higher than expected number of GATC motifs in the *E. coli* genome. Therefore, from this analysis, the GATC cannot be considered a highly over-represented tetranucleotide motif in *E. coli* genome. It is indeed under-represented with respect its component trinucleotides.

6.5 The association of the GATC motif content with the local GC content in *E. coli* genome:

The GATC motif content per 10000 bp to the GC content in the same region has been computed using in-lab Perl script. Association between these two data sets has been analyzed by covariance analysis using MATLAB®. However, no effective correlation between these two data sets has been established (Figure 6.3)

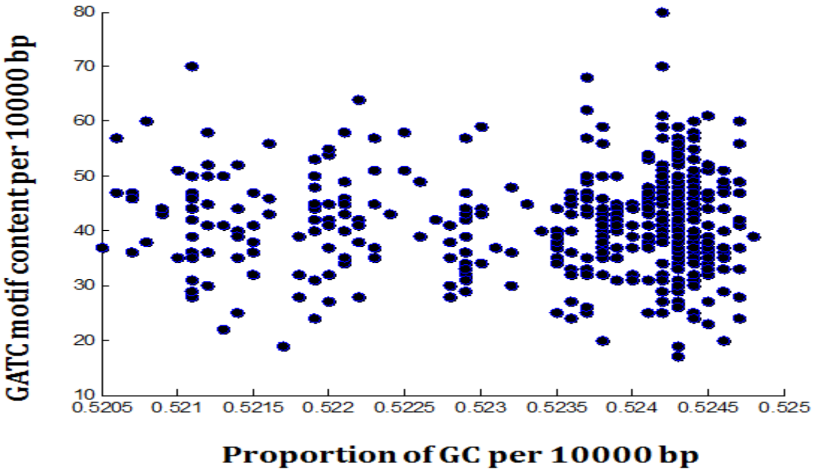


Figure 6.3. Correlation between local GATC motif content and GC content. The Scatter plot shows the correlation between GATC motif content and GC content per 10000 bp. The proportion of GC per 10000 bp is plotted along *x*-axis and the GATC motif content per 10000 bp is plotted along *y*-axis.

6.6 A model for the distribution of GATC motifs throughout the *E. coli* genome

The measure of over- or under-representation of a motif in a genome tells us very little about the biological significance of any particular motif that is unevenly distributed. Therefore, analysing the distribution of GATC motifs in the *E. coli* genome was considered imperative. At first, I have developed a Perl script (Appendix B) to calculate the distances between consecutive GATC motifs in a way that only the lengths spanning two consecutive GATC motifs are measured. These distances are mentioned as inter-GATC motif distances in this chapter. The inter-GATC motif distances ranged from 0 bp to 4836 bp. A frequency distribution of the inter-GATC motif distances was generated (Figure 6.4). It was observed that the distribution had an extreme positive skew (skewed to the right) – meaning that the closely spaced GATC motifs were highly frequent in the E*. coli* genome. This posed the question, whether this distribution was or was not expected by chance. Analysing the distribution of closely spaced GATC motifs revealed an extreme fluctuation of data points. The fluctuation revealed a previously reported periodicity of 3 basepairs for small inter-GATC motif distances (Figure 6.5). However, this periodicity is not observed after randomising the base composition of the *E. coli* genome (Figure 6.5). This periodicity, which has previously been reported for most of the di-, tri- and tetra-nucleotides in the genome of *E. coli*, the coding sequences of both *Saccharomyces cerevisiae* and human genome, suggested that there were indeed biologically important information determining the very short scale distribution of motifs. Graphs for the periodicity observed in the coding sequences of both *Saccharomyces cerevisiae* and human genome are given in appendix C.

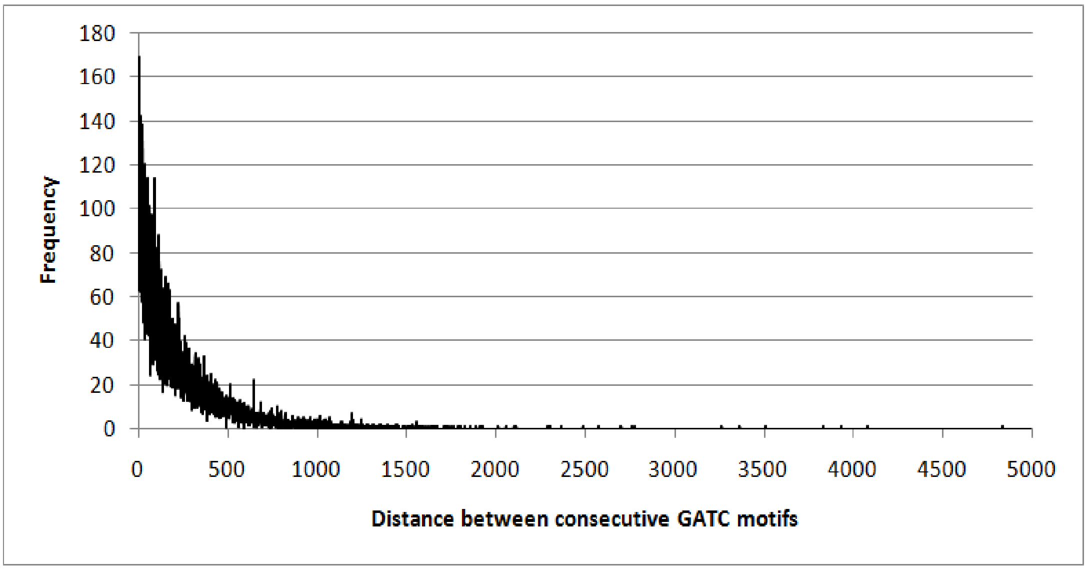


Figure 6.4. Frequency distribution of inter-GATC motif distances in the *E. coli* genome. The distances between consecutive GATC motifs had been plotted along the *x*-axis and the frequencies of these distances had been plotted along the *y*-axis.

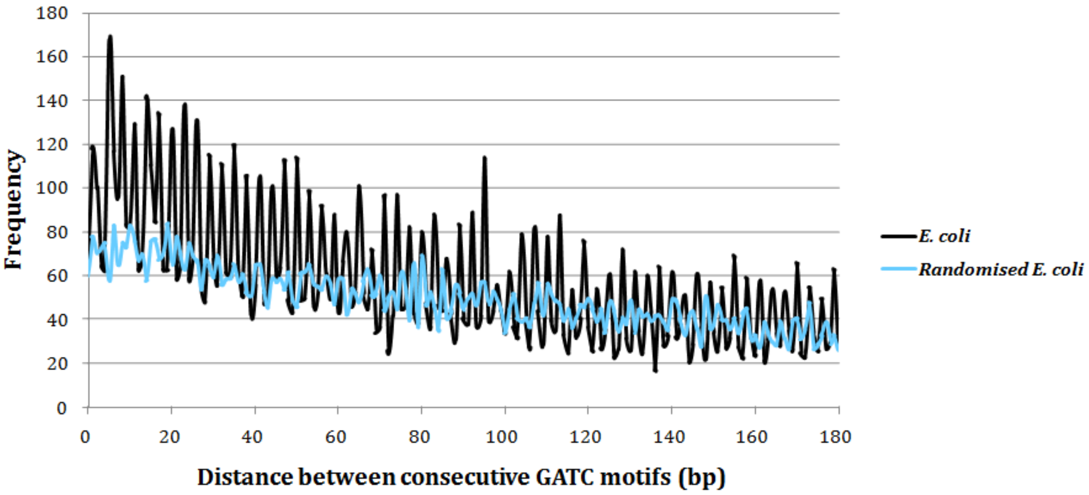


Figure 6.5. *E. coli* genome shows periodicity of 3 basepairs in the frequency distribution of inter GATC motif distances. The 3 basepair periodicity is absent in a randomized *E. coli* genome. The distances between consecutive GATC motifs had been plotted along the *x*-axis and the frequencies of these distances had been plotted along the *y*-axis.

A graph with the frequency distribution of inter-GATC motif distances plotted on a logarithmic scale reveals an expected straight line relationship (Figure 6.6). The logarithm of numerical zero (0) is undefined. Therefore, the distances (along the *x*-axis) for which the GATC motif count was zero, resulted in undefined values along the *y*-axis of the graph and were not plotted. Moreover, there were few (GATC motif free) larger inter-GATC motif distances in the *E. coli* genome. These regions are not clustered at any specific locus in the genome of *E. coli*, rather they are dispersed throughout the genome (red circles in Figure 6.1). Interestingly, for two instances, GATC motifs separated by more than 4 kb (4078 bp and 4836 bp) which exceeds the threshold of *in vitro* boundary regarding the DNA mismatch repair system. To determine whether the distribution of inter-GATC motifs distances has resulted from selection to avoid large separation, I have compared the frequency distribution of inter-GATC motif distances in the *E. coli* genome with that in a genome containing randomly distributed (and under few biological constraints) GATC motifs.

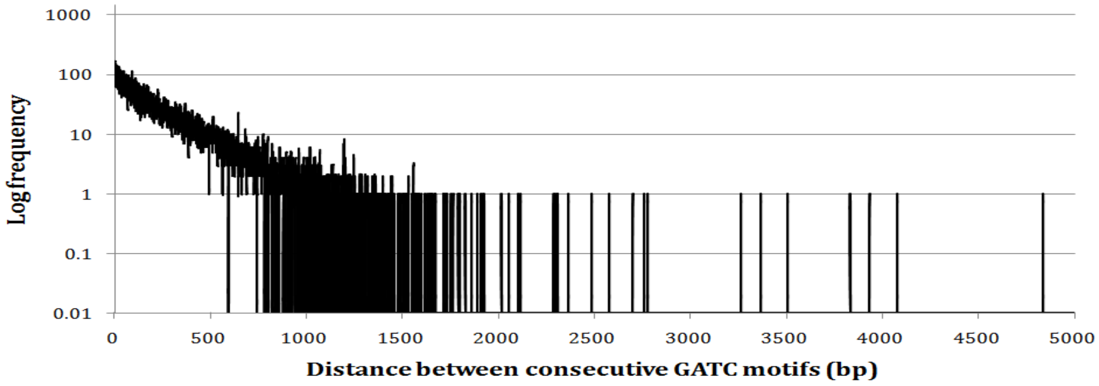


Figure 6.6. Frequency distribution of inter-GATC motif distances in the *E. coli* genome (frequencies in logarithmic scale). The distances between consecutive GATC motifs had been plotted along the *x*-axis and the logarithms of frequencies of these distances had been plotted along the *y*-axis.

6.7 Comparing the frequency distribution of inter-GATC distances in the *E. coli* genome with that in artificial model genomes

At this point, it became necessary to compare the distribution of GATC motifs present in the *E. coli* genome with that in model genomes, which should be random in nature. Three model sequences of the same length as the *E. coli* genome were generated (10 sequences of each model) using Markov sequence models–

1. Random sequence (Rand).
2. Dinucleotide based sequence (DN).
3. Dicodon based sequence (DC).

The rationale behind each model sequence had been discussed in Chapter 2. It is worth mentioning here that the periodicity observed for the closely spaced GATC motifs in the distribution of inter-GATC motif distances in the *E. coli* genome were absent in these artificial genome. The frequency distribution of inter-GATC distances had been generated for all artificial model sequences along with the *E. coli* genome. In the case of artificial model sequences, the average of the frequencies of inter-GATC motif distances was calculated using the 10 model genomes. A non-sliding window of 10 bp was used along the *x*-axis to generate the frequency distribution to minimise fluctuation, yet to retain the significant features in the distribution of the real genome sequences (Figure 6.7).

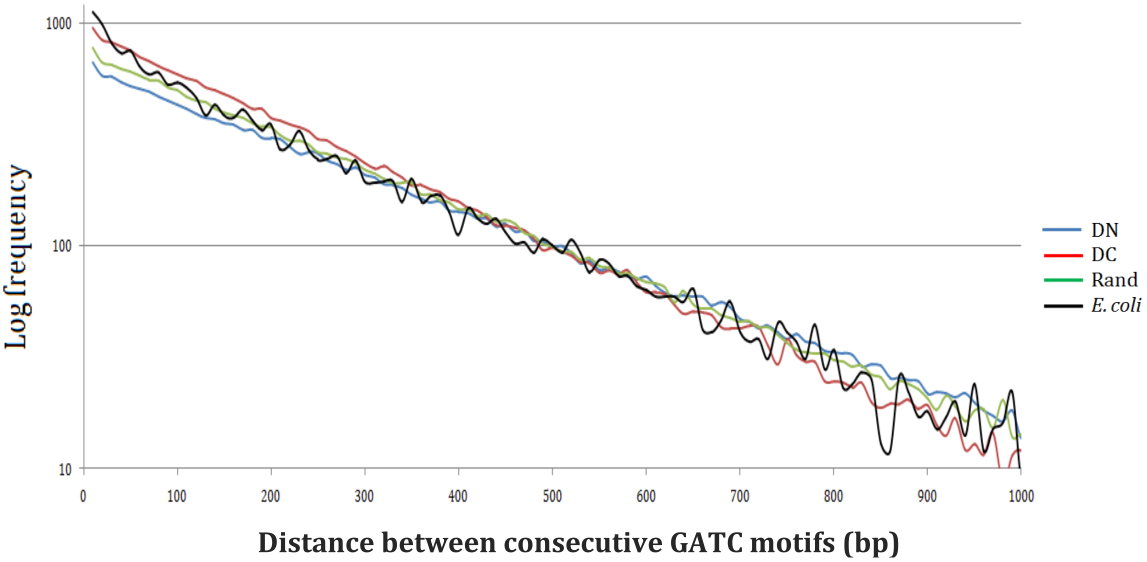


Figure 6.7. Comparative frequency distribution of inter-GATC motif distances in the *E. coli* genome and different artificial model sequences. The distances between consecutive GATC motifs were plotted along the *x*-axis and the logarithms of frequencies of the distances were plotted along the *y*-axis. The blue line represents the frequency distribution of inter-GATC motif distances in the dinucleotide sequences (DN), the red line represents that in the dicodon sequences (DC), the green line represents that in the random sequences (Rand) and the black line is that in the *E. coli* genome. A non-sliding window size of 10 basepairs was used to minimise the noise (highly fluctuating data points) in the graph. Average frequencies of the inter-GATC motif distances from 10 sequences generated for each type of artificial model sequences were used in this analysis.

From the comparison of the frequency distributions, it was observed that the *E. coli* genome has a higher frequency of closely spaced GATC motifs and lower frequency of long-spaced GATC motifs than the random model sequences (Rand). This pattern was consistent with the possibility of a selection pressure in favour of GATC motifs, specifically for the closely spaced GATC motifs in the *E. coli* genome. When the nucleotide composition of the *E. coli* genome was taken into consideration in order to generate a random sequence with biological information (the dinucleotide based model sequences, DN), we found even less closely spaced and more distantly spaced GATC motifs in the "DN" sequences than in the *E. coli* genome. This gave further support for the existence of a selection pressure for closely spaced GATC motifs in the *E. coli* genome. However, when another biological constraint was introduced by considering the composition of amino acid coding sequences, which is about 85% of the *E. coli* genome, the total scenario changed and the closely spaced GATC motifs in the *E. coli* genome became limited than in the complex model sequence (DC) except for the very short spaced GATC motifs.

Interestingly, the differences of slopes for the different model sequences actually reflected their intrinsic properties of harbouring different numbers of GATC motifs. The dicodon based model sequences had an average of about 23000 GATC motifs per genome, the Random sequences had an average of about 18000 GATC motifs per genome and for dinucleotide based model sequences, this number was about 16000. In an attempt to accommodate similar number of GATC motifs (19120 in the *E. coli* genome) in all the artificial model sequences, respective transition probabilities were modified manually to attain an average of 19120±1% GATC motifs per sequence. Then, a new comparative distribution of inter-GATC motif had been generated from the modified artificial model sequences and the *E. coli* genome (Figure 6.8).

This time, the frequency distribution of inter-GATC motif distances in the real genome and those of the artificial model sequences followed similar patterns (Figure 6.8). Two slight deviations from the model genomes were observed. First, the real genome contained a few more large-spaced GATC motifs than expected (Figure 6.8).

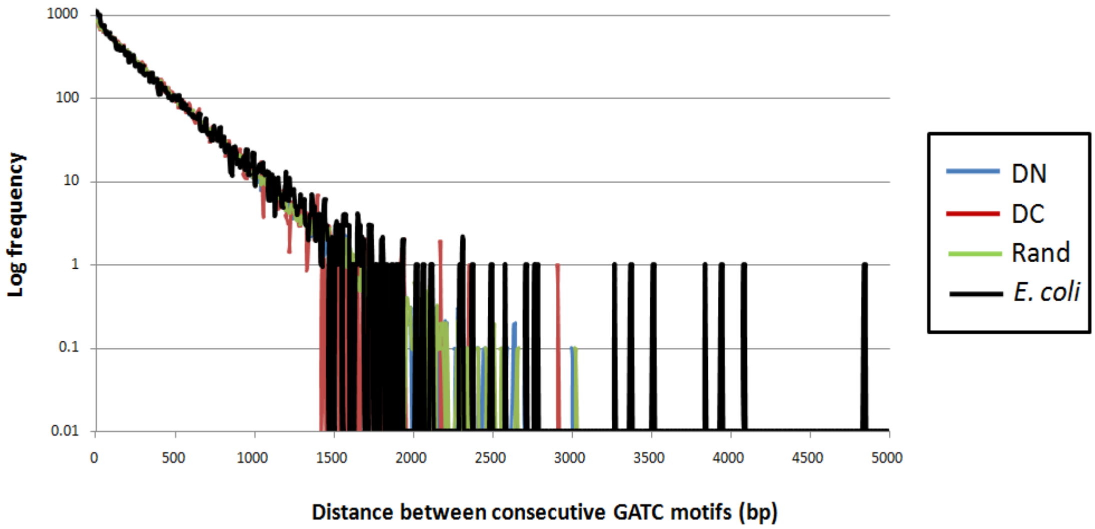


Figure 6.8. Comparative frequency distribution of inter-GATC motif distances in the *E. coli* genome and modified artificial model sequences. The distances between consecutive GATC motifs had been plotted along the *x*-axis and the logarithms of frequencies of the distances have been plotted along the *y*-axis. The blue line represents the modified Dinucleotide sequence (DN), the red line represents the modified Dicodon sequence (DC), the green line represents the Random sequence (Rand) and the black line represents the *E. coli* genome in the frequency distribution of inter-GATC motif distances. A non-sliding window size of 10 basepairs had been used to minimise noise (highly fluctuating data points). Average frequencies of GATC motifs from 10 sequences generated for each type of artificial model sequence had been used in this analysis.

Second, when observed at a higher resolution, the frequency distribution of the inter-GATC distances of the *E. coli* genome showed higher frequency from 0 bp to 40 bp inter-GATC motif distance than that of the artificial model sequences (Figure 6.9). This particular behaviour was indicative of a clustering effect of GATC motifs in the *E. coli* genome. This phenomenon can be explained by a schematic representation in Figure 6.10. The frequency distribution of inter-motif distances of a randomly distributed motif in a circular genome would follow the blue straight line in the Figure 6.10. When the same number of motifs in that circular genome begins to form clusters, the frequency of the short distances would become higher and as an inevitable effect the frequency of the long distances would rise as well. To compensate for these effects (while maintaining a constant number of total GATC sites), the frequency of intermediate distances would decrease and the overall distribution would follow a similar pattern of the red dashed line in the Figure 6.10. Figure 6.8 clearly showed that the frequency distribution corresponding to the *E. coli* genome displayed some distinct peaks for very large-spaced GATC motifs, which were indicative of the comparatively higher frequency of large-spaced motifs in the red dashed line in the model (Figure 6.10).

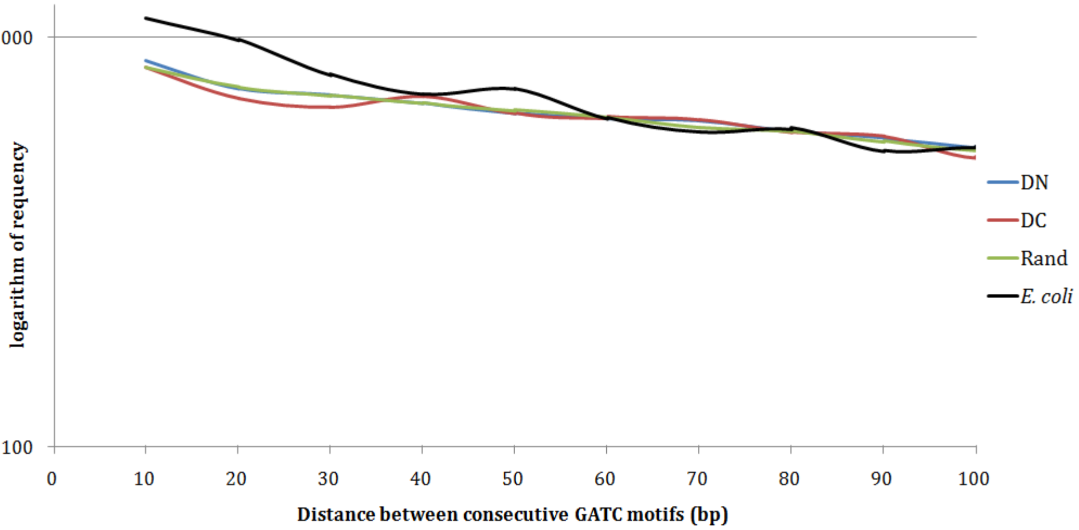


Figure 6.9. A selective representation (from distance of 0 bp to 100 bp) of the frequency distribution of inter-GATC motif distances in the *E. coli* genome and different artificial sequences. The distances between consecutive GATC motifs have been plotted along the *x*-axis and the logarithms of frequencies of the distances have been plotted along the *y*-axis. The blue line represents the modified Dinucleotide sequence (DN), the red line represents the modified Dicodon sequence (DC), the green line represents the modified Random sequence (Rand) and the black line represents the *E. coli* genome in the frequency distribution of inter-GATC motif distances. A non-sliding window of 10 bp had been applied along the *x*-axis to minimise the noise (fluctuating data points) in the distribution.

To visualise the very starting point of the distribution (corresponding to inter-GATC distances from 0 bp to 40 bp), the graph of this region representing the *E. coli* genome was compared to the three graphs corresponding to different artificial sequences (Figure 6.9). The *E. coli* genome contained more of these very closely spaced GATC motifs than the artificial genomes with the same total numbers of GATC sites. However, due to fluctuations of the frequency in the rest of the distribution, it was difficult to detect the fall of frequencies compensating the rise of frequencies at both ends on this scale of plot.

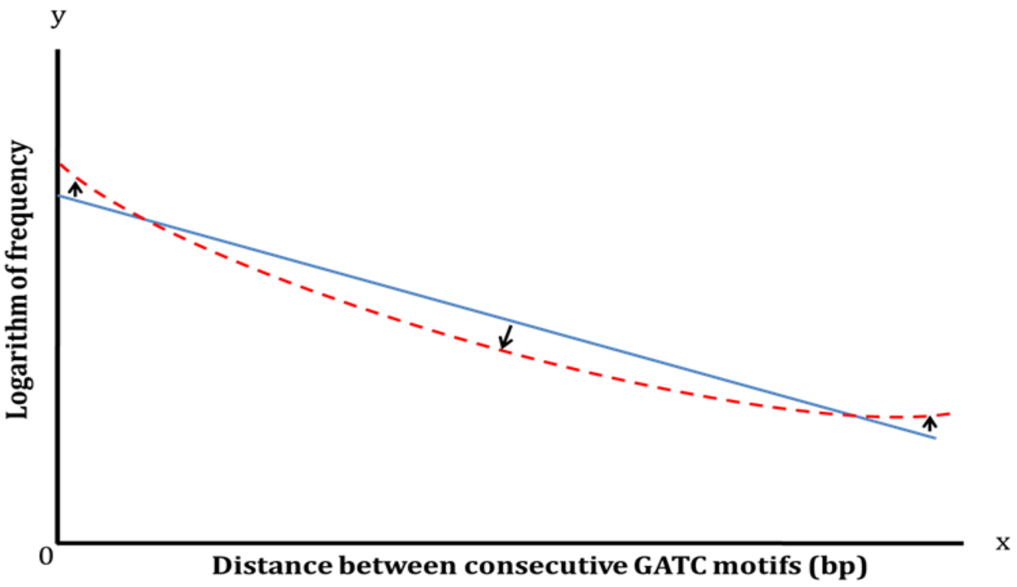


Figure 6.10. A schematic representation describing the clustering effect of a motif in a circular genome. A randomly distributed motif in a circular genome would follow a pattern similar to the solid blue line. However, if this motif was distributed in clusters in the genome, the frequency of the closely spaced motifs in that circular genome would rise and as a consequence the frequency of the distantly spaced motifs would rise as well. On the other hand, to compensate the increment of the frequency at both ends of the distribution, the frequency of the intermediate distances would decrease and the overall distribution of the clustered motif would follow line similar to the red dashed line.

These closely spaced GATC which are distributed throughout the genome, were extracted and annotated. As regulatory sequences did not have clear boundaries and Riva and collaborators did not find any correlation between the distribution of GATC motifs and regulation of transcription of the genes having clusters of GATC motifs in their regulatory regions, only coding sequences were taken into account in this respect (Riva et al., 2004a). About 19% of all GATC motifs (19120) in the *E. coli* genome were found of to have inter-GATC distances of 0 – 40 bp. On the other hand, 1515 coding sequences (more than one third of all reported coding sequences in the *E. coli* genome) had been found to harbour such GATC motifs. Of them, 83% were mapped to different pathways in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database on annotation using Blast2GO (Conesa et al., 2005). KEGG is a database resource that integrates genomic, chemical and systemic functional information (Kanehisa et al., 2014). The genes containing closely space GATC sites were representated in all of the *E. coli* specific pathways (111 of them) found in the KEGG database rather than being exclusive to any single or few pathways involved in a or a set of definite biological functions. Although some gene-sets covered all the *E. coli* specific genes involved in a certain pathway (like – "Chlorocyclohexane and chlorobenzene degradation pathway", "Fluorobenzoate degradation pathway" and "Toluene degradation pathway"), their exclusiveness was not taken into account due to the fact that *E. coli* specific genes covered only a small percentage of all genes involved in those pathways. However, some pathways like the "Starch and sucrose metabolism pathway", "D-Glutamine and D-glutamate metabolism pathway", "Geraniol degradation pathway", "C5-Branched dibasic acid metabolism pathway" and "Cysteine and methionine metabolism pathway" had over 75% representative *E. coli* genes which harbour such short spaced GATC motifs in them. These GATC motifs could regulate the transcription of these essential genes upon methylation by the Dam methylase and lower the melting temperature of the ambient sequence during transcription (Hénaut et al., 1996). Though this analysis identified the well-established 11 GATC motifs at *oriC*, the origin of replication of the *E. coli* genome, it did not reveal any obvious rationale in favour of clustered GATC motifs in the *E. coli* genome.

6.8 Distribution of GATC motifs in the backbone- and variable-segment: mapping of large-spaced GATC motifs in the *E. coli* genome

As GATC motifs play a crucial role in DNA mismatch repair in *E. coli* and an MMR system is conserved in almost all organisms, it has been hypothesised in this study that the DNA sequences that are conserved in the course of evolution in different strains of *E. coli* are under a selection pressure that might favour closely spaced GATC motifs. Alternatively, DNA sequences that are not conserved in the course of evolution might have less selection pressure for closely spaced GATC motifs in them. To analyse the conserved and non-conserved sequences in the *E. coli* genome, which are also called the backbone– and the variable-segment respectively, DNA sequences were obtained from the online database MOSAIC (http://genome.jouy.inra.fr/mosaic/) (Chiapello et al., 2005, 2008). MOSAIC is a relational database that has an interactive web interface to compare closely related bacterial genomes. In MOSAIC, genomes of different strains of the same species have been aligned using Multiple Genome Aligner software or the MAUVE to determine almost identical regions constituting the backbone segment (Darling et al., 2010; Höhl et al., 2002). The backbone segments are interrupted by strain specific inserted regions, which are termed variable segments. In this study, the backbone– and the variable-segments were extracted from MOSAIC release 5.1 (2009-11-26) based on the analysis on five well studied strains of *E. coli* that have had their genome sequenced and are available in public genome sequence databases (Table 6.3).

Table 6.3. An overview[*] of five closely related strains of *E. coli*

Strains of *E. coli*	**Identifier**	**Genome length**	**Percent of backbone segment**
E. coli strain K-12 MG1655	NC_000913.3	4639675 bp	78.90%
E. coli O157:H7 str. Sakai	BA000007.2	5498450 bp	66.58%
E. coli IAI1 / 08	CU928160.2	4700560 bp	77.88%
E. coli UMN026 / ExPEC/ O17:K52:H18	CU928163.2	5202090 bp	70.37%
E. coli CFT073/UPEC/O6:H1/ATCC /CFT073	AE014075.1	5231428 bp	69.97%

[*] Data to generate this table were obtained from the MOSAIC database.

Large portions of their genomes belong to the conserved backbone segment which can be observed from the Figure 6.11. On the other hand, less than 22% to 35% of their genome falls under the variable segment.

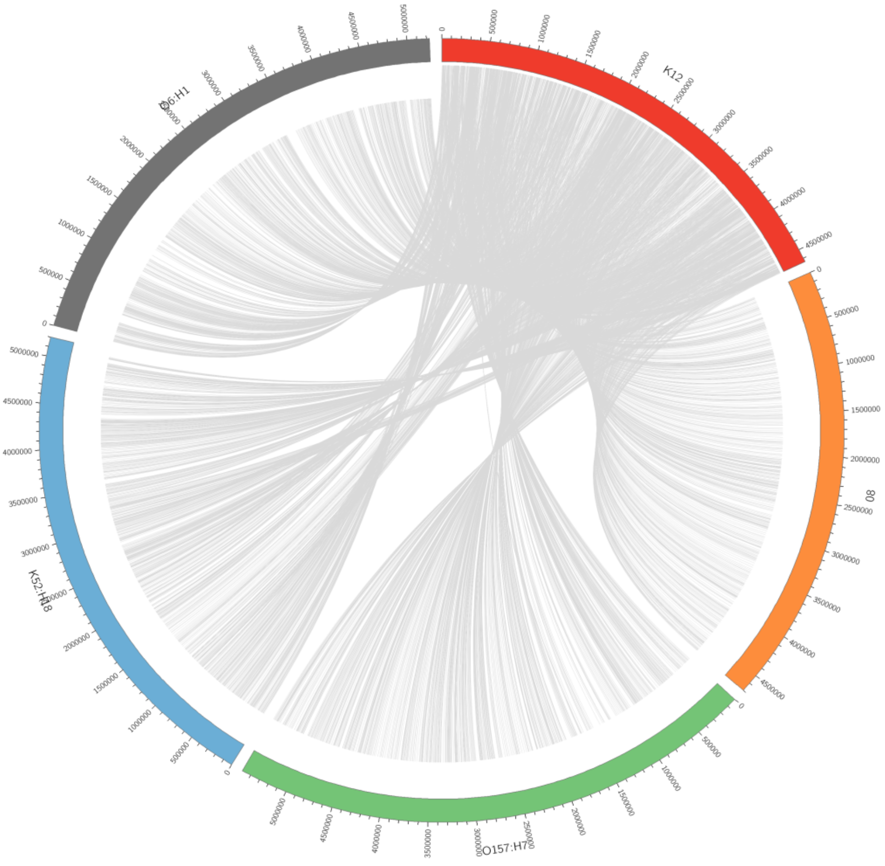


Figure 6.11. A CIRCOS graph representing the common sequences (backbone-segment) of the genomes of all five strains of *E. coli*. This CIRCOS graph had been drawn by analysing data available in the MOSAIC database. The red band corresponds to the genome of *E. coli* strain K-12 MG1655, the orange band corresponds to the genome of *E. coli* IAI1/08, the green band corresponds to the genome of *E. coli* O157:H7 str. Sakai, the blue band corresponds to the genome of *E. coli* UMN026/ExPEC/O17:K52:H18 and the grey band corresponds to the genome of *E. coli* CFT073/UPEC/O6:H1/ATCC/CFT073. The light grey lines connecting different genomes correspond to the homologous sequences of *E. coli* strain K-12 MG1655 found in the four other strains.

The backbone segment of *Escherichia coli* K-12 MG1655 comprised 78.90% of the sequence (Figure 6.11 and 6.12). This segment harbours 83.25% of all GATC motifs found in the whole genome. Therefore, I reasoned that analysing the distribution of GATC motifs in the backbone and the variable segment of the *E.*

coli could give me some insight about the existence of any selection pressure that had shaped its distribution.

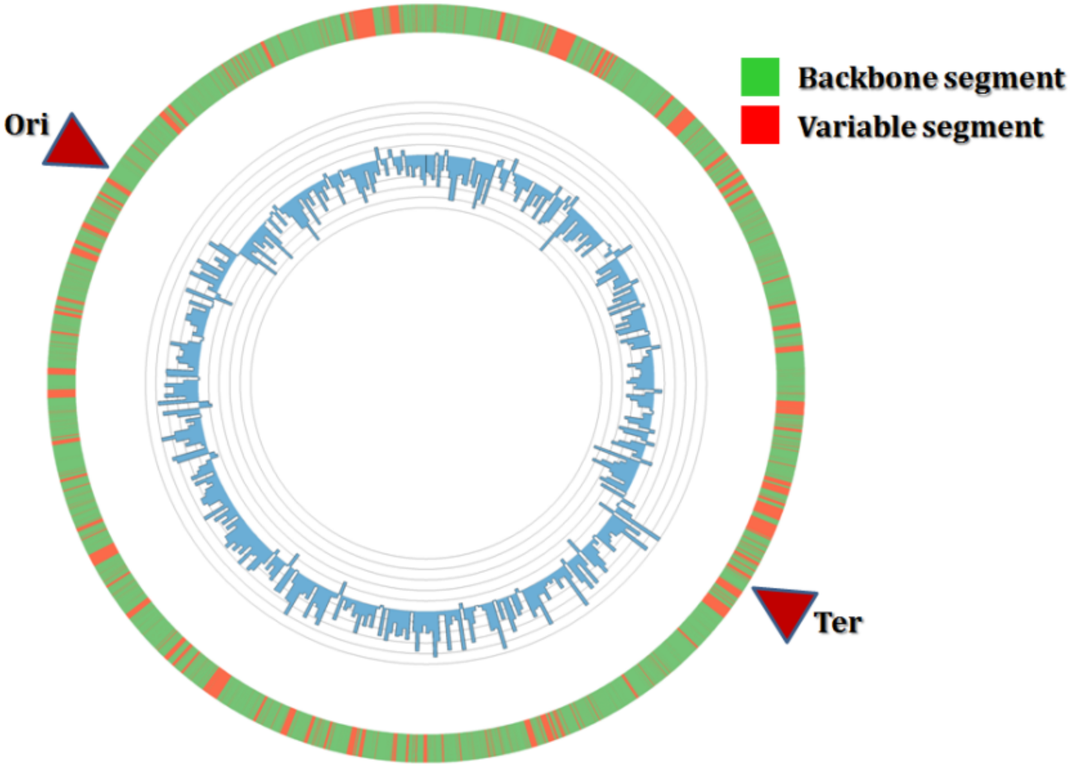


Figure 6.12. A CIRCOS representation of the relative distribution of the backbone- and variable segment of the *E. coli* strain K-12 MG1655. This CIRCOS graph has been drawn by analysing the data available in the MOSAIC database. The green regions correspond to the backbone segment and the red regions correspond to the variable segment. The inner circle shows the GC-skew in the genome of *E. coli* strain K-12 MG1655, which helps to mark the origin of replication and terminus region in the genome and the axis imposed here spans from -0.10672986 to 0.106270358 and divided into 10 circular grids.

As the backbone and variable segment sequences were found scattered throughout the *E. coli* genome, it might seem sensible to concatenate the sequences to make a single contigs for each group. However, concatenating different sequences led to a phenomenon called the “edge effect” which resulted in two problems. The first one was that new GATC motifs emerged at some sequence junctions. The next consequence was that new inter-GATC motif distances are generated at sequence junctions. Concatenating all the sequences of backbone or variable segments separately could contribute a lot to the edge effect. For example, an average of 76 new GATC motifs (0.503% of the total GATC motifs in the backbone) emerged after shuffling and concatenating the backbone sequences several times. In addition to that, new distances at the junctions of the backbone sequences were generated. Therefore, inter-GATC motif distances were calculated by analysing individual backbone and variable segment using an in-lab Perl script (Appendix B). After normalising against corresponding total length of each type of sequence, a frequency distribution of inter-GATC motif distances was generated. The graph representing the backbone sequences followed the graph of the *E. coli* genomic sequence almost perfectly for the short spaced GATC motifs, while the graph representing the variable sequences did not (Figure 6.13). However, the distribution of inter-GATC motif distances based on backbone segments fails to provide further information regarding the distantly spaced GATC motifs as the data points corresponding to the backbone segments did not map with the peaks found for such GATC motifs in the *E. coli* genome (Figure 6.13). On the other hand, the graph representing the variable segments did not follow the graph of the *E. coli* genomic sequence for the short spaced GATC motifs. Although, the variable segments had only 0.1% less GATC motifs per base-pair than the backbone sequences, the deviation observed in the graph indicated a lack of short spaced

GATC motifs in *E. coli* variable segments. Therefore, the *E. coli* backbone segments were found to have a preference for short spaced GATC motifs than the variable segments. In addition, most of the peaks that correspond to the inter-GATC motif distances larger than 3.2 kb in the *E. coli* genome were well mapped in the variable segments. Therefore, the variable segments (or the strain specific sequence regions) harboure the larger GATC free regions in the *E. coli* genome. The peaks that belong to the distantly spaced GATC motifs and were not mapped in the variable segments fell at the junctions of backbone and variable segments.

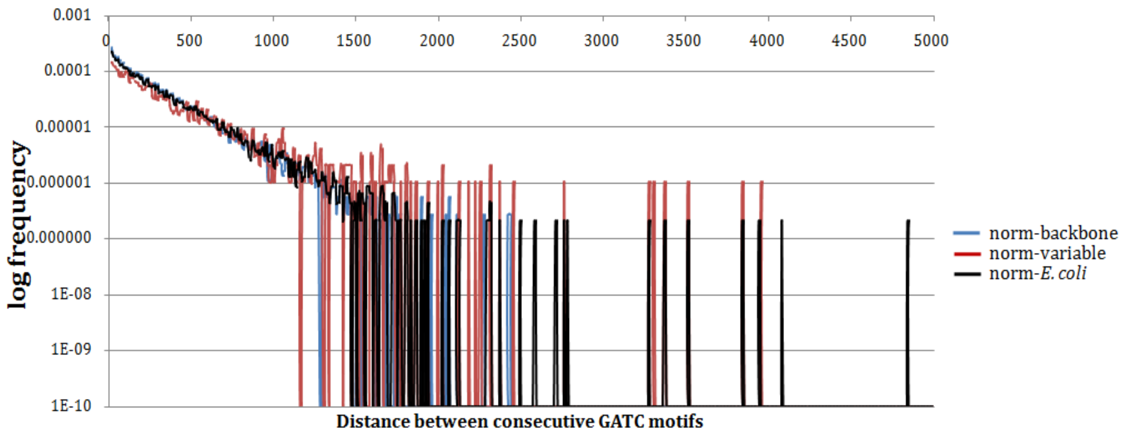


Figure 6.13. Comparative frequency distribution of inter-GATC motif distances of backbone segments, variable segments and the total genome of *E. coli*. After normalising by their respective (total) lengths, the frequencies were plotted on a logarithmic scale along the *y*-axis and the distances between consecutive GATC motifs were plotted along the *x*-axis. The black line corresponds to the distribution of inter-GATC motif distances in the *E. coli* genome, the blue line corresponds to the distribution calculated on each segment of the backbone sequences and the red line correspond to the distribution calculated on each segment of the variable sequences. The "norm" prefix denotes the corresponding frequencies normalized against their respective (total) lengths.

At this point, a closer look at those large GATC motifs free sequences that corresponded to the seven peaks in the frequency distribution of inter-GATC motif distance revealed some interesting information (Table 6.4). A vast proportion of those sequences corresponded to prophage DNA and to members of the *rhs* (rearrangement hot spots) gene family. It is worth mentioning that those GATC-free DNA regions are dispersed throughout the *E. coli* genome rather than being clustered together (Figure 6.1).

The *rhs* genes are found to be involved in contact dependent inhibition (CDI) of cell growth of different bacteria (Poole et al., 2011). The large GATC motif-free sequences in *E. coli* harboured four of the eight *rhs* genes found in the *E. coli* genome (Wang et al., 1998). The sequence composition of *rhs* genes explained the lack of GATC motifs in them. All the members of *rhs* family in *E. coli* share a 3.7 kb long GC rich core ORF which codes for about 1200 amino acids with multiple tandem repeated copies of a YD-repeat domain (Edwards et al., 1998; Feulner et al., 1990; Minet et al., 1999). There is an AT rich highly divergent extension of the core region (ext-a1, 130-177 amino acid long), which is followed by another AT rich region (dsORF-a1) (Aggarwal and Lee, 2011). Therefore, the sequence composition of the *rhs* genes explains the large inter-GATC motif distances and there is no evidence of a selection pressure on GATC motifs in these sequences.

On the other hand, the prophage sequences found in the *E. coli* genome had a different attribute. In the course of the evolution of bacterial genomes, temperate bacteriophages have parasitized their hosts by integrating their genomes into the host genetic material via a process of horizontal gene transfer (Brüssow et al., 2004; Filée et al., 2003). These integrated sequences, which are now called "prophages", sometimes encode various novel abilities, like niche adaptation and virulence factors (Canchaya et al., 2003; Filée et al., 2003). However, *in silico* analysis of the bacterial genome has suggested that most of the prophages are defective in their function and apparently are in a process of mutational decay (Canchaya et al., 2003).

One of the basic assumptions in evolutionary biology is that a gene can reside in a population if it confers a selectable function or selective advantage (Canchaya et al., 2003). Prophages belong to the old parts of bacterial genomes and they are not shared by all of the strains of *E. coli* (Canchaya et al., 2003). Even closely related *E. coliO157* strains share prophage elements with substantially different prophage contents (Canchaya et al., 2003). Therefore accumulating GATC motifs in prophage sequences in the course of evolution would not exert any effect on the total metabolism of the host *E. coli* genome, other than maintaining genomic integrity of that region. However, this is surely not the case here. The evidence mentioned above leads to the conclusion that the prophage sequences are not under any selection pressure for generating GATC motifs that would be used by DNA mismatch repair system to ensure the genomic integrity of *E. coli*.

Table 6.4. An overview of the sequence annotation of the seven largest GATC-free regions in the genome of *E. coli* strain K-12 MG1655

Peak length	Annotation
4836 bp	*ybbP* (Putative inner membrane protein)
	rhsD (*rhsD* element protein)
4078 bp	*rhsA* (*rhsA* element protein)
	yibF (Putative gluthione S-transferase)
3934 bp	*ybfA* (hypothetical protein)
	rhsC (*rhsC* element protein)
3832 bp	*rhsB* (*rhsB* element protein)
3507 bp	*yfdH* (CPS-53 (KpLE1) prophage; bactoprenol glucosyl transferase)
	yfdI (CPS-53 (KpLE1) prophage; putative inner membrane protein)

	yfdK (CPS-53 (KpLE1) prophage protein)
3366 bp	*isrC* (Novel sRNA, function unknown, CP4-44; putative prophage remnant)
	flu (CP4-44 prophage, antigen 43 (Ag43) phase-variable biofilm formation autotransporter)
3265 bp	*yfjT* (CPS-5 prophage protein)
	yfjS (CPS-5 prophage protein)
	yfjK (CPS-5 prophage; putative inner membrane protein)
	yfjR (CPS-5 prophage; putative DNA-binding transcriptional regulator)
	yfjQ (CPS-5 prophage protein)
	yfjP (CPS-5 prophage; putative GTP-binding protein)

6.9 Discussion

Different approaches have been used to understand whether the distribution of GATC sites in the genomes of different bacteria relate to their biological functions (Riva et al., 2004a, 2004b). The involvement of GATC motifs in the DNA mismatch repair system and the regulation of bacterial replication are well established to date. These two systems require certain patterns of local distribution of GATC motifs in the genome. On the other hand, transcriptional control by GATC motif is more controversial. Based on a transcriptome analysis in Dam$^+$ and Dam$^-$ cells, Oshima and collaborators proposed that the methylation state of the GATC motifs at regulatory regions affects protein-DNA interactions that regulate the transcription of certain genes (Oshima et al., 2002). They proposed that the genes whose transcription levels are affected by the *dam* genotype, have an elevated number of GATC motifs in their 500 bp upstream sequences. However, a computational analysis by Riva and collaborators has argued that there is no such transcriptional control by GATC motifs (Riva et al., 2004a). They have also shown that out of 349 *dam*-sensitive genes found in the study of Oshima and collaborators only 3 contain upstream *Fnr* consensus sequences that share GATC motifs and a majority of the genes containing *Fnr* consensus sequences in the *E. coli* genome are not sensitive to the *dam* genotype. Moreover, the positions, frequency and distribution of GATC motifs and the difference in gene expression (found in the transcriptome analysis done by Oshima and collaborators) are shown not to be correlated (Riva et al., 2004a). On the other hand, methylation of GATC motifs in the coding regions of some genes are proposed to affect their transcription (Hénaut et al., 1996). In addition, a major proportion of the *E. coli* genome (about 85%) is coding. Therefore, coding regions of the *E. coli* genome have been taken under consideration in this study for generating artificial model sequences and comparative analysis of the distribution of GATC motifs in the *E. coli* genome.

There are several mathematical processes that could be used to generate an artificial sequence for a comparative study. The simplest approach would be to use a "Multinomial sequence model" which assumes that in the process of evolution the DNA sequence is produced by a random process, meaning that such a model randomly chooses any of the four nucleotides at each position in the sequence. According to this model, the probability of choosing any one of the four nucleotides depends on a predetermined probability distribution. Thus, a multinomial model has only four parameters: the predetermined probability of each of the four nucleotides p_A, p_C, p_G, and p_T, where $p_A + p_C + p_G + p_T= 1$. However, the main drawback of this process is that it is not an accurate representation of how a sequence has evolved. More specifically, the problem of this model is that the probability of a nucleotide at a position depends only on the predetermined probability of that nucleotide, not on the surrounding nucleotides. However, in a biologically meaningful sequence, a nucleotide at a given position can depend on the adjacent nucleotide composition and there is often an evolutionary pressure on this nucleotide from adjacent nucleotides. Considering these drawbacks, the multinomial model that is computationally easier to handle, has been avoided and Markov sequence models have been selected.

Previous studies have analyzed the distribution of GATC motifs and have used this information to interpret the role of GATC motifs in the regulation of gene expression (Hénaut et al., 1996; Riva et al., 2004a). The main goal of this *in silico* study was to detect any selection pressure for GATC in the *E. coli* genome, particularly in respect to DNA mismatch repair. Different approaches were implemented to analyse the distribution of inter-GATC distances in the *E. coli* genome to find an answer. In this study, clusters of GATC motifs were observed in the *E. coli* genome that are more frequent than the number expected purely by chance, and some of these are associated with known biological attributes. Probably the most important cluster of GATC motifs is situated at the origin of replication and contributes to the synchronous initiation of DNA replication (Lu et al., 1994; Slater et al., 1995). The majority of the very closely spaced GATC motifs (0 bp to 40 bp apart) were not noticeably clustered at particular loci. Instead, they were dispersed throughout the *E. coli* genome. They were present in the genes that belonged to all the *E. coli* specific pathways in the KEGG database rather involved in any specific pathway(s) in *E. coli*.

Moreover, a few GATC motif-free regions in the *E. coli* genome were found corresponding to the DNA sequences that became integrated into the *E. coli* genome by horizontal gene transfer process. These regions mainly consist of members of the *rhs* gene family which are originated outside of *E. coli* and the prophage sequences (mainly integrated defective sequences of bacteriophages) (Brüssow et al., 2004; Filée et al., 2003; Jackson et al., 2009). This study has led us to conclude that, apart from the very closely spaced and very distantly spaced

GATC motifs, the distribution of inter-GATC motif distances seems to follow a random distribution in the *E. coli* genome compared to that in artificial genomes generated in this study. Therefore, the DNA mismatch repair system and other GATC interacting pathways have evolved to utilise this distribution of GATC motifs to maintain genomic integrity and perform various other cellular functions.

Chapter 7

Conclusions and Further Work

7.1 Conclusions

My *in vivo* work has investigated the directionality of DNA mismatch repair and my *in silico* modelling has studied potential evolutionary pressures exerted by the MMR system on the genomic distribution of GATC motifs in *E. coli*. During DNA replication, a mismatch can be generated at an anonymous locus in the genome and therefore, most of the previous studies on MMR have been carried out using plasmids or bacteriophage DNA substrates where synthetic heteroduplexes can be constructed (Blackwell et al., 1998; Dzantiev et al., 2004; Thomas et al., 1991). In this *in vivo* study, a 294 bp long CTG•CAG repeat tract has been used to generate frequent locus-specific substrates for the MMR system. The replication and metabolism of plasmids differ from those of genomic DNA (Kovtun and McMurray, 2008). Moreover, among all eight possible mismatches the highest affinity has been found, by the gel-shift assay, for the G-T mismatch (gives rise to transition mutation) and IDLs, which are the most frequent mis-incorporations/ replication errors (Jiricny et al., 1988b). Therefore, TNR array generates the perfect substrate for the *in vivo* MMR reactions and usage of the TNR array has permitted for the first time investigation of directionality by using a "real" chromosomal assay. This study also avoids the artificial nature of re-construction of the DNA mismatch repair system using purified proteins. I believe that the most important finding of this PhD work has been the observation of a distinct directionality of the MMR system in relation to DNA replication.

Blackwood and collaborators found that MMR-stimulated recombination at a 275 bp tandem repeat only occurred when the tandem repeat was placed on the origin proximal side of the CTG•CAG repeat array, which generated the substrates for the MMR system, suggesting that MMR might be directional *in vivo* (Blackwood et al., 2010). This hypothesis was tested in Chapter 3 by analysing the usage of GATC motifs on both origin proximal and origin distal sides of the TNR array by the MMR system on the length instability of the TNR array. By sequential modification of GATC motifs on both origin proximal and origin distal sides of the TNR array, thereby shifting the start site of the excision reaction during MMR, I have shown that MMR is indeed directional in relation to the movement of the replication fork and the efficiency of MMR decreases upon shifting of the next available GATC motif beyond approximately 2 kb in an origin distal direction. At this point, an obvious question arises – "What causes the MMR system to be directional?" Two possibilities can be envisaged in the following scenarios:

1. The Dam methylase methylates hemimethylated GATC sites in the nascent strand as it follows the replication fork, thereby making the GATC motifs that are adjacent to replication fork most available to the MMR system (Figure 7.1). However, it has been reported that the Dam methylase lags behind the replication fork by about 0.5 - 3 minutes (Barras and Marinus, 1989). In this window of time, the replication machinery would replicate 120 kb of DNA (at an average speed of 1000 bp/sec) and generate a lot of hemimethylated GATC sites. If a mismatch arises due to a replication-error in this region, the MMR system would find several hemimethylated GATC sites both on origin proximal side and on origin distal side of the mismatch. Therefore, the Dam methylase

protein alone cannot account for the directionality of the MMR machinery.

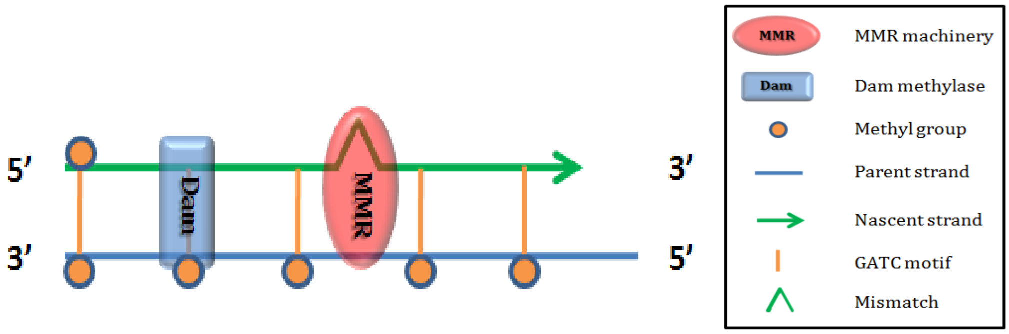


Figure 7.1. A schematic representation of how Dam mediated methylation might govern the direction of the MMR machinery.

2. An energetically active force might drive the MMR system to follow a single direction (Figure 7.2). This energetically active force could be a combination of different molecular processes that determine the usage of newly generated hemimethylated GATC sites adjacent to a progressing replication fork. Interaction of the components of MMR machinery (MutS and/or MutL) with the β clamp might contribute to the recruitment of MutS and/or MutL to the mismatch by the replication machinery, which might provide a molecular basis for this hypothesis (Pluciennik et al., 2009). This process alone or along with other processes might confer directionality on the MMR reaction whether the mismatch is created in a nascent leading strand (Figure 7.2A) or in a nascent lagging strand (Figure 7.2B). Furthermore, this hypothesis allows the MMR machinery to use the replication machinery to drive the search for a

hemimethylated GATC site. In this way, MMR hitch-hikes a ride with the replisome as it searches for a hemimethylated GATC site.

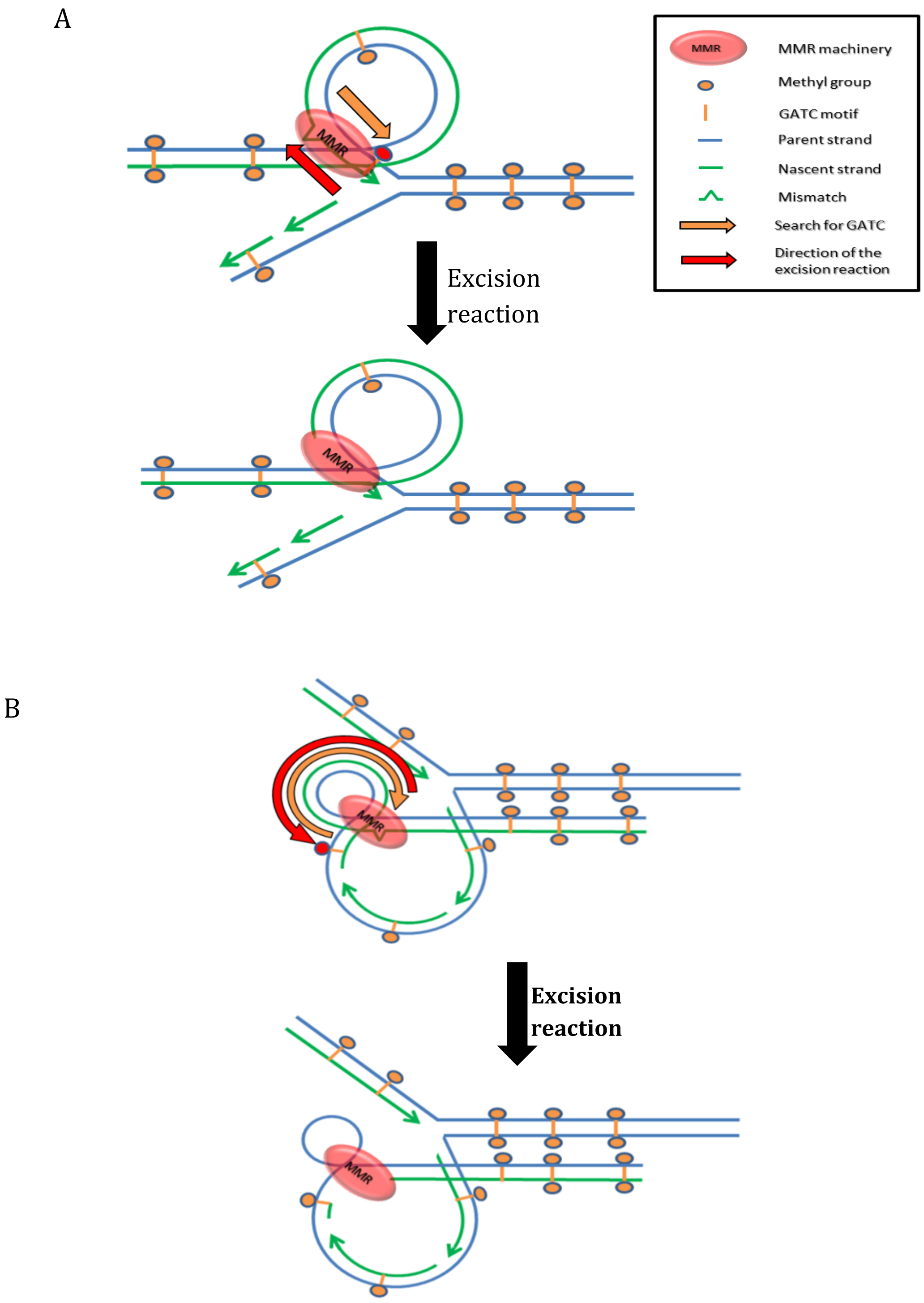


Figure 7.2. A schematic representation of an energetically active force directing the MMR machinery to the replisome. (A) depicts the scenario when a mismatch occurs in the leading nascent strand and (B) depicts the case of the lagging nascent strand. Formation of looped DNA structure can bring a newly replicated hemimethylated GATC site near to the MMR machinery attached to a mismatch during MMR. A hemimethylated GATC motif, which is used as a start point of the excision reaction, is depicted in red.

Lahue and collaborators have shown that having two closely spaced GATC motifs near a mismatch increases the efficiency of mismatch repair by 83% compared to that with a single GATC motif in a covalently closed circular DNA (Lahue et al., 1987). However, in this *in vivo* study, no elevated mismatch repair was observed in the presence of a cluster of 3 closely spaced origin distal GATC motifs compared to that with a single GATC motif near the TNR array. Therefore, a single origin distal GATC motif near a mismatch is sufficient for an efficient MMR. The termination point of the extended excision reaction during MMR detected by the stimulation of recombination at the zeocin resistance cassette was found in this study to depend on the starting hemimethylated GATC motif rather than on the position of the TNR array (Chapter 4). Grillery and collaborators have shown that the *in vitro* excision reaction terminates at a set of discrete sites within a 100 nucleotides region from a mismatch (Grilley et al., 1993). On the other hand, Blackwood and collaborators have found that MMR-stimulated recombination at a 275 bp tandem repeat (by DNA resynthesis following the formation of a single stranded sequence harbouring the tandem repeat) occurs at a distance of 6.3 kb on the origin proximal side of the TNR array (Blackwood et al., 2010). RecQ helicase was found not responsible for the formation of highly processive excision reaction tract that could be a subset of standard excision tract during MMR.

Interaction between MMR components and the β clamp led me to test whether a functional MMR impedes the progression of the DNA replication machinery. In this study, the MMR system was found not to decelerate or to make the replication machinery pause due to a mismatch in strains with a region having intact native GATC motifs or free of GATC motifs. On the other hand, Samadashwily and collaborators have detected a replication pause at a CTG•CAG repeat array in a plasmid system in *E. coli,* but have not characterised whether this is dependent on MMR (Samadashwily et al., 1997). The metabolism of secondary structures that formed at the TNR array (as per their assumption) in the plasmid might be different to when the TNR array is in the chromosome and the replication machinery is more efficient replicating the TNR array. Moreover, they used antibiotic chloramphenicol to minimise protein binding to the secondary structure that formed at the TNR array. This is not a natural situation and *in vivo* metabolism will have been affected. Notably, in the presence of chloramphenicol plasmid replication continues using DNA polymerase I, which is less sensitive to the drug than is the replisome comprising DNA polymerase III.

An *in silico* study in Chapter 6 found a close to random distribution of GATC motifs in the *E. coli* genome with some clusters at different parts of the genome. There are also a few regions that are free of GATC motif for comparatively long distances, which are accounted for by the base composition of local gene sequences. No clear evidence was detected for a selection pressure, exerted by the MMR system on the genomic distribution of GATC sites. Instead, we hypothesise that the MMR system has evolved to utilise the expected distribution of GATC motifs in the *E. coli* genome.

7.2 Further work

The molecular mechanism of the directionality of MMR (in terms of the usage of GATC motifs) has not been determined yet. GATC motifs are shared by different molecular pathways and more than one protein compete for a GATC motif at a given time. A str*etc*h of hemimethylated GATC motif is generated following a replication fork and those sites become the targets for at least three different proteins – MutH, SeqA and Dam methyltransferase (Campbell and Kleckner, 1990; Waldminghaus et al., 2012). Dam methyltransferase methylates the unmethylated adenine of the GATC motif and thus converts a hemimethylated GATC site into a fully methylated one. In that case, the GATC motif becomes unavailable to both MutH and SeqA proteins.

SeqA protein is well-known for its participation in the process of origin sequestration in *E. coli* so that an immature origin is prevented from firing (Waldminghaus et al., 2012). However, recent evidence suggests that SeqA oligomerises to form a higher order structure by binding the hemimethylated GATC motifs (Guarné et al., 2005; Odsbu et al., 2005). A genome-wide study of SeqA binding to DNA shows that the SeqA complex follows the replication fork in a treadmilling fashion – growing at the leading end and diminishing at the tailing end (Waldminghaus et al., 2012). On the other hand, *in vitro* studies have shown that the SeqA protein is not actively dissociated by the Dam methylase (Kang et al., 1999). Moreover, over-expression of Dam methylase does not change the genome-wide SeqA binding pattern that is observed in the wild-type cells (Waldminghaus et al., 2012). Therefore, it has been suggested that the Dam methyltransferase follows the SeqA oligomer and methylates the hemimethylated GATC motifs that become available upon random dissociation of SeqA proteins at the end of the oligomer (Waldminghaus et al., 2012). This is consistent with the reported delay of Dam methyltranferase to methylate a recently generated hemimethylated GATC motif in the close vicinity of the replication fork. Two possible scenarios are worth investigating as an extension of the current study –

1. A SeqA oligomer that is formed on the origin proximal side of a mismatch might restrain the MutH protein searching for a hemimethylated GATC motif on that side of that mismatch and facilitate the unique direction of the MMR system. Therefore, in a *seqA* mutant cell, the MMR would lose its directionality but would still be efficient finding a hemimethylated GATC motif on either side of a mismatch.
2. Alternatively, if the delay of the Dam methylase is due to the formation of oligomeric SeqA structure, Dam would find an unrestricted str*etc*h of DNA where it can methylate all the hemimethylated GATC sites ahead (towards the replication fork). In that case, the efficiency of the MMR system would decrease due to lack of hemimethylated GATC motif near a mismatch in a *seqA* mutant cell.

Therefore, the combinatorial functions of SeqA and Dam methylase on the molecular mechanism of the directionality of the MMR system could be a fascinating topic for the future research.

References

Aaltonen, L.A., Peltomäki, P., Leach, F.S., Sistonen, P., Pylkkänen, L., Mecklin, J.P., Järvinen, H., Powell, S.M., Jen, J., and Hamilton, S.R. (1993). Clues to the pathogenesis of familial colorectal cancer. Science *260*, 812–816.

Acharya, S., Foster, P.L., Brooks, P., and Fishel, R. (2003). The coordinated functions of the *E. coli* MutS and MutL proteins in mismatch repair. Mol. Cell *12*, 233–246.

Aggarwal, K., and Lee, K.H. (2011). Overexpression of cloned RhsA sequences perturbs the cellular translational machinery in *Escherichia coli*. J. Bacteriol. *193*, 4869–4880.

Akiyama, Y., Sato, H., Yamada, T., Nagasaki, H., Tsuchiya, A., Abe, R., and Yuasa, Y. (1997). Germ-line mutation of the hMSH6/GTBP gene in an atypical hereditary nonpolyposis colorectal cancer kindred. Cancer Res. *57*, 3920–3923.

Ali, J.A., and Lohman, T.M. (1997). Kinetic measurement of the step size of DNA unwinding by *Escherichia coli* UvrD helicase. Science. *275*, 377–380.

Allen, G.C., and Kornberg, A. (1993). Assembly of the primosome of DNA replication in *Escherichia coli*. J. Biol. Chem. *268*, 19204–19209.

Allen, D.J., Makhov, A., Grilley, M., Taylor, J., Thresher, R., Modrich, P., and Griffith, J.D. (1997). MutS mediates heteroduplex loop formation by a translocation mechanism. EMBO J. *16*, 4467–4476.

Allers, T., and Lichten, M. (2001). Intermediates of yeast meiotic recombination contain heteroduplex DNA. Mol. Cell *8*, 225–231.

Amin, N.S., Nguyen, M.N., Oh, S., and Kolodner, R.D. (2001). exo1-Dependent mutator mutations: model system for studying functional interactions in mismatch repair. Mol. Cell. Biol. *21*, 5142–5155.

Aoki, K., and Taketo, M.M. (2007). Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene. J. Cell Sci. *120*, 3327–3335.

Au, K.G., Welsh, K., and Modrich, P. (1992). Initiation of methyl-directed mismatch repair. J. Biol. Chem. *267*, 12142–12148.

Bachrati, C.Z., and Hickson, I.D. (2003). RecQ helicases: suppressors of tumorigenesis and premature aging. Biochem. J. *374*, 577–606.

Baitinger, C., Burdett, V., and Modrich, P. (2003). Hydrolytically deficient MutS E694A is defective in the MutL-dependent activation of MutH and in the mismatch-dependent assembly of the MutS.MutL.heteroduplex complex. J. Biol. Chem. *278*, 49505–49511.

Bardwell, P.D., Woo, C.J., Wei, K., Li, Z., Martin, A., Sack, S.Z., Parris, T., Edelmann, W., and Scharff, M.D. (2004). Altered somatic hypermutation and reduced class-switch recombination in exonuclease 1-mutant mice. Nat. Immunol. *5*, 224–229.

Barras, F., and Marinus, M.G. (1989). The great GATC: DNA methylation in *E. coli*. Trends Genet. *5*, 139–143.

Barre, F.X., Aroyo, M., Colloms, S.D., Helfrich, A., Cornet, F., and Sherratt, D.J. (2000). FtsK functions in the processing of a Holliday junction intermediate during bacterial chromosome segregation. Genes Dev. *14*, 2976–2988.

Bell, L., and Byers, B. (1983). Separation of branched from linear DNA by two-dimensional gel electrophoresis. Anal. Biochem. *130*, 527–535.

Berends, M.J.W., Wu, Y., Sijmons, R.H., Mensink, R.G.J., van der Sluis, T., Hordijk-Hos, J.M., de Vries, E.G.E., Hollema, H., Karrenbeld, A., Buys, C.H.C.M., et al. (2002). Molecular and clinical characteristics of MSH6 variants: an analysis of 25 index carriers of a germline variant. Am. J. Hum. Genet. *70*, 26–37.

Blackwell, L.J., Bjornson, K.P., and Modrich, P. (1998). DNA-dependent Activation of the hMutSalpha ATPase. J. Biol. Chem. *273*, 32049–32054.

Blackwell, L.J., Wang, S., and Modrich, P. (2001). DNA chain length dependence of formation and dynamics of hMutSalpha.hMutLalpha.heteroduplex complexes. J. Biol. Chem. *276*, 33233–33240.

Blackwood, J.K., Okely, E.A., Zahra, R., Eykelenboom, J.K., and Leach, D.R.F. (2010). DNA tandem repeat instability in the *Escherichia coli* chromosome is stimulated by mismatch repair at an adjacent CAG·CTG trinucleotide repeat. Proc. Natl. Acad. Sci. U. S. A. *107*, 22582–22586.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. (1997). The Complete Genome Sequence of *Escherichia coli* K-12. Science. *277*, 1453–1462.

Blyn, L.B., Braaten, B.A., and Low, D.A. (1990). Regulation of pap pilin phase variation by a mechanism involving differential dam methylation states. EMBO J. *9*, 4045–4054.

Bowers, J., Tran, P.T., Liskay, R.M., and Alani, E. (2000). Analysis of yeast MSH2-MSH6 suggests that the initiation of mismatch repair can be separated into discrete steps. J. Mol. Biol. *302*, 327–338.

Bowers, J., Tran, P.T., Joshi, a, Liskay, R.M., and Alani, E. (2001). MSH-MLH complexes formed at a DNA mismatch are disrupted by the PCNA sliding clamp. J. Mol. Biol. *306*, 957–968.

Braaten, B.A., Nou, X., Kaltenbach, L.S., and Low, D.A. (1994). Methylation patterns in pap regulatory DNA control pyelonephritis-associated pili phase variation in *E. coli*. Cell *76*, 577–588.

Branch, P., Aquilina, G., Bignami, M., and Karran, P. (1993). Defective mismatch binding and a mutator phenotype in cells tolerant to DNA damage. Nature *362*, 652–654.

Bransteitter, R., Pham, P., Scharff, M.D., and Goodman, M.F. (2003). Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. Proc. Natl. Acad. Sci. U. S. A. *100*, 4102–4107.

Brendler, T., and Austin, S. (1999). Binding of SeqA protein to DNA requires interaction between two or more complexes bound to separate hemimethylated GATC sequences. EMBO J. *18*, 2304–2310.

Brendler, T., Abeles, A., and Austin, S. (1995). A protein that binds to the P1 origin core and the oriC 13mer region in a methylation-specific fashion is the product of the host seqA gene. EMBO J. *14*, 4083–4089.

Brewer, B.J., and Fangman, W.L. (1987). The localization of replication origins on ARS plasmids in S. cerevisiae. Cell *51*, 463–471.

Brewer, B.J., and Fangman, W.L. (1988). A replication fork barrier at the 3' end of yeast ribosomal RNA genes. Cell *55*, 637–643.

Van den Broek, W.J.A.A., Nelen, M.R., Wansink, D.G., Coerwinkel, M.M., te Riele, H., Groenen, P.J.T.A., and Wieringa, B. (2002). Somatic expansion behaviour of the (CTG)n repeat in myotonic dystrophy knock-in mice is differentially affected by Msh3 and Msh6 mismatch-repair proteins. Hum. Mol. Genet. *11*, 191–198.

Bronner, C.E., Baker, S.M., Morrison, P.T., Warren, G., Smith, L.G., Lescoe, M.K., Kane, M., Earabino, C., Lipford, J., and Lindblom, A. (1994). Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. Nature *368*, 258–261.

Bruni, R., Martin, D., and Jiricny, J. (1988). d(GATC) sequences influence *Escherichia coli* mismatch repair in a distance-dependent manner from positions both upstream and downstream of the mismatch. Nucleic Acids Res. *16*, 4875–4890.

Brüssow, H., Canchaya, C., and Hardt, W.-D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol. Mol. Biol. Rev. *68*, 560–602.

Bunting, K.A., Roe, S.M., and Pearl, L.H. (2003). Structural basis for recruitment of translesion DNA polymerase Pol IV/DinB to the beta-clamp. EMBO J. *22*, 5883–5892.

Burdett, V., Baitinger, C., Viswanathan, M., Lovett, S.T., and Modrich, P. (2001). In vivo requirement for RecJ, ExoVII, ExoI, and ExoX in methyl-directed mismatch repair. Proc. Natl. Acad. Sci. U. S. A. *98*, 6765–6770.

Cadman, C.J., Lopper, M., Moon, P.B., Keck, J.L., and McGlynn, P. (2005). PriB stimulates PriA helicase via an interaction with single-stranded DNA. J. Biol. Chem. *280*, 39693–39700.

Campbell, J.L., and Kleckner, N. (1990). *E. coli* oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork. Cell *62*, 967–979.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brüssow, H. (2003). Prophage genomics. Microbiol. Mol. Biol. Rev. *67*, 238–276.

Cannavo, E., Gerrits, B., Marra, G., Schlapbach, R., and Jiricny, J. (2007). Characterization of the interactome of the human MutL homologues MLH1, PMS1, and PMS2. J. Biol. Chem. *282*, 2976–2986.

Casali, P., Pal, Z., Xu, Z., and Zan, H. (2006). DNA repair in antibody somatic hypermutation. Trends Immunol. *27*, 313–321.

Cejka, P., Stojic, L., Mojas, N., Russell, A.M., Heinimann, K., Cannavó, E., di Pietro, M., Marra, G., and Jiricny, J. (2003). Methylation-induced G(2)/M arrest requires a full complement of the mismatch repair protein hMLH1. EMBO J. *22*, 2245–2254.

Chapados, B.R., Hosfield, D.J., Han, S., Qiu, J., Yelent, B., Shen, B., and Tainer, J.A. (2004). Structural basis for FEN-1 substrate specificity and PCNA-mediated activation in DNA replication and repair. Cell *116*, 39–50.

Chase, J.W., and Richardson, C.C. (1974). Exonuclease VII of *Escherichia coli*. Mechanism of action. J. Biol. Chem. *249*, 4553–4561.

Chaudhuri, J., Tian, M., Khuong, C., Chua, K., Pinaud, E., and Alt, F.W. (2003). Transcription-targeted DNA deamination by the AID antibody diversification enzyme. Nature *422*, 726–730.

Chi, N.W., and Kolodner, R.D. (1994). The effect of DNA mismatches on the ATPase activity of MSH1, a protein in yeast mitochondria that recognizes DNA mismatches. J. Biol. Chem. *269*, 29993–29997.

Chiapello, H., Bourgait, I., Sourivong, F., Heuclin, G., Gendrault-Jacquemard, A., Petit, M.-A., and El Karoui, M. (2005). Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. BMC Bioinformatics *6*, 171.

Chiapello, H., Gendrault, A., Caron, C., Blum, J., Petit, M.-A., and El Karoui, M. (2008). MOSAIC: an online database dedicated to the comparative genomics of bacterial strains at the intra-species level. BMC Bioinformatics *9*, 498.

Cho, W.-K., Jeong, C., Kim, D., Chang, M., Song, K.-M., Hanne, J., Ban, C., Fishel, R., and Lee, J.-B. (2012). ATP alters the diffusion mechanics of MutS on mismatched DNA. Structure *20*, 1264–1274.

Claverys, J.P., and Lacks, S.A. (1986). Heteroduplex deoxyribonucleic acid base mismatch repair in bacteria. Microbiol. Rev. *50*, 133–165.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics *21*, 3674–3676.

Cooper, D.L., Lahue, R.S., and Modrich, P. (1993). Methyl-directed mismatch repair is bidirectional. J. Biol. Chem. *268*, 11823–11829.

Courcelle, J., and Hanawalt, P.C. (1999). RecQ and RecJ process blocked replication forks prior to the resumption of replication in UV-irradiated *Escherichia coli*. Mol. Gen. Genet. *262*, 543–551.

Cox, M.M. (2001). Recombinational DNA repair of damaged replication forks in *Escherichia coli*: questions. Annu. Rev. Genet. *35*, 53–82.

Cox, M.M., Goodman, M.F., Kreuzer, K.N., Sherratt, D.J., Sandler, S.J., and Marians, K.J. (2000). The importance of repairing stalled replication forks. Nature *404*, 37–41.

D'haeseleer, P. (2006). What are DNA sequence motifs? Nat. Biotechnol. *24*, 423–425.

Dalrymple, B.P., Kongsuwan, K., Wijffels, G., Dixon, N.E., and Jennings, P.A. (2001). A universal protein-protein interaction motif in the eubacterial DNA replication and repair systems. Proc. Natl. Acad. Sci. U. S. A. *98*, 11627–11632.

Dao, V., and Modrich, P. (1998). Mismatch-, MutS-, MutL-, and Helicase II-dependent Unwinding from the Single-strand Break of an Incised Heteroduplex. J. Biol. Chem. *273*, 9202–9207.

Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One *5*, e11147.

Datta, A., Schmeits, J.L., Amin, N.S., Lau, P.J., Myung, K., and Kolodner, R.D. (2000). Checkpoint-dependent activation of mutagenic repair in *Saccharomyces cerevisiae* pol3-01 mutants. Mol. Cell *6*, 593–603.

Delagoutte, E., Goellner, G.M., Guo, J., Baldacci, G., and McMurray, C.T. (2008). Single-stranded DNA-binding protein in vitro eliminates the orientation-dependent impediment to polymerase passage on CAG/CTG repeats. J. Biol. Chem. *283*, 13341–13356.

Delbos, F., De Smet, A., Faili, A., Aoufouchi, S., Weill, J.-C., and Reynaud, C.-A. (2005). Contribution of DNA polymerase eta to immunoglobulin gene hypermutation in the mouse. J. Exp. Med. *201*, 1191–1196.

Dessinges, M.-N., Lionnet, T., Xi, X.G., Bensimon, D., and Croquette, V. (2004). Single-molecule assay reveals strand switching and enhanced processivity of UvrD. Proc. Natl. Acad. Sci. *101*, 6439–6444.

Diaz, M., and Flajnik, M.F. (1998). Evolution of somatic hypermutation and gene conversion in adaptive immunity. Immunol. Rev. *162*, 13–24.

Dijkwel, P.A., and Hamlin, J.L. (1997). Mapping replication origins by neutral/neutral two-dimensional gel electrophoresis. Methods *13*, 235–245.

Dillon, S.C., and Dorman, C.J. (2010). Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat. Rev. Microbiol. *8*, 185–195.

Dohet, C., Dzidić, S., Wagner, R., and Radman, M. (1987). Large non-homology in heteroduplex DNA is processed differently than single base pair mismatches. Mol. Gen. Genet. *206*, 181–184.

Dohrmann, P.R., and McHenry, C.S. (2005). A bipartite polymerase-processivity factor interaction: only the internal beta binding site of the alpha subunit is required for processive replication by the DNA polymerase III holoenzyme. J. Mol. Biol. *350*, 228–239.

Drotschmann, K., Hall, M.C., Shcherbakova, P. V, Wang, H., Erie, D.A., Brownewell, F.R., Kool, E.T., and Kunkel, T.A. (2002). DNA binding properties of the yeast Msh2-Msh6 and Mlh1-Pms1 heterodimers. Biol. Chem. *383*, 969–975.

Dufner, P., Marra, G., Räschle, M., and Jiricny, J. (2000). Mismatch recognition and DNA-dependent stimulation of the ATPase activity of hMutSalpha is abolished by a single mutation in the hMSH6 subunit. J. Biol. Chem. *275*, 36550–36555.

Dutta, R., and Inouye, M. (2000). GHKL, an emergent ATPase/kinase superfamily. Trends Biochem. Sci. *25*, 24–28.

Duval, A., Rolland, S., Compoint, A., Tubacher, E., Iacopetta, B., Thomas, G., and Hamelin, R. (2001). Evolution of instability at coding and non-coding repeat sequences in human MSI-H colorectal cancers. Hum. Mol. Genet. *10*, 513–518.

Dzantiev, L., Constantin, N., Genschel, J., Iyer, R.R., Burgers, P.M., and Modrich, P. (2004). A defined human system that supports bidirectional mismatch-provoked excision. Mol. Cell *15*, 31–41.

Eder, P.S., and Walder, J.A. (1991). Ribonuclease H from K562 human erythroleukemia cells. Purification, characterization, and substrate specificity. J. Biol. Chem. *266*, 6472–6479.

Eder, P.S., Walder, R.Y., and Walder, J.A. (1993). Substrate specificity of human RNase H1 and its role in excision repair of ribose residues misincorporated in DNA. Biochimie *75*, 123–126.

Edwards, R.A., Keller, L.H., and Schifferli, D.M. (1998). Improved allelic exchange vectors and their use to analyze 987P fimbria gene expression. Gene *207*, 149–157.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. *5*, 435–445.

Eykelenboom, J.K., Blackwood, J.K., Okely, E., and Leach, D.R.F. (2008). SbcCD causes a double-strand break at a DNA palindrome in the *Escherichia coli* chromosome. Mol. Cell *29*, 644–651.

Fang, W.H., and Modrich, P. (1993). Human strand-specific mismatch repair occurs by a bidirectional mechanism similar to that of the bacterial reaction. J. Biol. Chem. *268*, 11838–11844.

Feulner, G., Gray, J.A., Kirschman, J.A., Lehner, A.F., Sadosky, A.B., Vlazny, D.A., Zhang, J., Zhao, S., and Hill, C.W. (1990). Structure of the rhsA locus from *Escherichia coli* K-12 and comparison of rhsA with other members of the rhs multigene family. J. Bacteriol. *172*, 446–456.

Filée, J., Forterre, P., and Laurent, J. (2003). The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. Res. Microbiol. *154*, 237–243.

Fink, D., Nebel, S., Aebi, S., Zheng, H., Cenni, B., Nehmé, A., Christen, R.D., and Howell, S.B. (1996). The role of DNA mismatch repair in platinum drug resistance. Cancer Res. *56*, 4881–4886.

Fischer, C.J., Maluf, N.K., and Lohman, T.M. (2004). Mechanism of ATP-dependent translocation of E.coli UvrD monomers along single-stranded DNA. J. Mol. Biol. *344*, 1287–1309.

Fishel, R., Lescoe, M.K., Rao, M.R., Copeland, N.G., Jenkins, N.A., Garber, J., Kane, M., and Kolodner, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. Cell *75*, 1027–1038.

Flores-Rozas, H., Clark, D., and Kolodner, R.D. (2000). Proliferating cell nuclear antigen and Msh2p-Msh6p interact to form an active mispair recognition complex. Nat. Genet. *26*, 375–378.

Ford, C.B., Shah, R.R., Maeda, M.K., Gagneux, S., Murray, M.B., Cohen, T., Johnston, J.C., Gardy, J., Lipsitch, M., and Fortune, S.M. (2013). Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat. Genet. *45*, 784–790.

Friedberg, E.C., Lehmann, A.R., and Fuchs, R.P.P. (2005). Trading places: how do DNA polymerases switch during translesion DNA synthesis? Mol. Cell *18*, 499–505.

Friedberg, E.C., Walker, G.C., Siede, W., Wood, R.D., Schultz, R.A., and Ellenburger, T. (2006). DNA Repair and Mutagenesis.

Friedman, K.L., and Brewer, B.J. (1995). Analysis of replication intermediates by two-dimensional agarose gel electrophoresis. Methods Enzymol. *262*, 613–627.

Friedman-Ohana, R., and Cohen, A. (1998). Heteroduplex joint formation in *Escherichia coli* recombination is initiated by pairing of a 3'-ending strand. Proc. Natl. Acad. Sci. U. S. A. *95*, 6909–6914.

Fujii, S., and Fuchs, R.P. (2004). Defining the position of the switches between replicative and bypass DNA polymerases. EMBO J. *23*, 4342–4352.

Galio, L., Bouquet, C., and Brooks, P. (1999). ATP hydrolysis-dependent formation of a dynamic ternary nucleoprotein complex with MutS and MutL. Nucleic Acids Res. *27*, 2325–2331.

Geier, G.E., and Modrich, P. (1979). Recognition sequence of the dam methylase of *Escherichia coli* K12 and mode of cleavage of Dpn I endonuclease. J. Biol. Chem. *254*, 1408–1413.

Genschel, J., and Modrich, P. (2003). Mechanism of 5’-directed excision in human mismatch repair. Mol. Cell *12*, 1077–1086.

Genschel, J., Bazemore, L.R., and Modrich, P. (2002). Human exonuclease I is required for 5’ and 3' mismatch repair. J. Biol. Chem. *277*, 13302–13311.

Ghodgaonkar, M.M., Lazzaro, F., Olivera-Pimentel, M., Artola-Borán, M., Cejka, P., Reijns, M.A., Jackson, A.P., Plevani, P., Muzi-Falconi, M., and Jiricny, J. (2013). Ribonucleotides misincorporated into DNA act as strand-discrimination signals in eukaryotic mismatch repair. Mol. Cell *50*, 323–332.

Gomes, X. V, and Burgers, P.M. (2000). Two modes of FEN1 binding to PCNA regulated by DNA. EMBO J. *19*, 3811–3821.

Gorman, J., Chowdhury, A., Surtees, J.A., Shimada, J., Reichman, D.R., Alani, E., and Greene, E.C. (2007). Dynamic basis for one-dimensional DNA scanning by the mismatch repair complex Msh2-Msh6. Mol. Cell *28*, 359–370.

Gradia, S. (2000). The Role of Mismatched Nucleotides in Activating the hMSH2-hMSH6 Molecular Switch. J. Biol. Chem. *275*, 3922–3930.

Gradia, S., Acharya, S., and Fishel, R. (1997). The human mismatch recognition complex hMSH2-hMSH6 functions as a novel molecular switch. Cell *91*, 995–1005.

Gradia, S., Subramanian, D., Wilson, T., Acharya, S., Makhov, A., Griffith, J., and Fishel, R. (1999). hMSH2–hMSH6 Forms a Hydrolysis-Independent Sliding Clamp on Mismatched DNA. Mol. Cell *3*, 255–261.

Gradia, S., Acharya, S., and Fishel, R. (2000). The Role of Mismatched Nucleotides in Activating the hMSH2-hMSH6 Molecular Switch. J. Biol. Chem. *275*, 3922–3930.

Grady, W.M., Rajput, A., Myeroff, L., Liu, D.F., Kwon, K., Willis, J., and Markowitz, S. (1998). Mutation of the type II transforming growth factor-beta receptor is coincident with the transformation of human colon adenomas to malignant carcinomas. Cancer Res. *58*, 3101–3104.

Grilley, M., Welsh, K.M., Su, S.S., and Modrich, P. (1989). Isolation and characterization of the *Escherichia coli* mutL gene product. J. Biol. Chem. *264*, 1000–1004.

Grilley, M., Griffith, J., and Modrich, P. (1993). Bidirectional excision in methyl-directed mismatch repair. J. Biol. Chem. *268*, 11830–11837.

Guarné, A., Ramon-Maiques, S., Wolff, E.M., Ghirlando, R., Hu, X., Miller, J.H., and Yang, W. (2004). Structure of the MutL C-terminal domain: a model of intact MutL and its roles in mismatch repair. EMBO J. *23*, 4134–4145.

Guarné, A., Brendler, T., Zhao, Q., Ghirlando, R., Austin, S., and Yang, W. (2005). Crystal structure of a SeqA-N filament: implications for DNA replication and chromosome organization. EMBO J. *24*, 1502–1511.

Guerrette, S., Wilson, T., Gradia, S., and Fishel, R. (1998). Interactions of human hMSH2 with hMSH3 and hMSH2 with hMSH6: examination of mutations found in hereditary nonpolyposis colorectal cancer. Mol. Cell. Biol. *18*, 6616–6623.

Guo, S., Presnell, S.R., Yuan, F., Zhang, Y., Gu, L., and Li, G.-M. (2004). Differential requirement for proliferating cell nuclear antigen in 5' and 3' nick-directed excision in human mismatch repair. J. Biol. Chem. *279*, 16912–16917.

Gupta, S., Gellert, M., and Yang, W. (2012). Mechanism of mismatch recognition revealed by human MutSβ bound to unpaired DNA loops. Nat. Struct. Mol. Biol. *19*, 72–78.

Harfe, B.D., and Jinks-Robertson, S. (2000). DNA mismatch repair and genetic instability. Annu. Rev. Genet. *34*, 359–399.

Harmon, F.G., and Kowalczykowski, S.C. (1998). RecQ helicase, in concert with RecA and SSB proteins, initiates and disrupts DNA recombination. Genes Dev. *12*, 1134–1144.

Harrington, J.M., and Kolodner, R.D. (2007). *Saccharomyces cerevisiae* Msh2-Msh3 acts in repair of base-base mispairs. Mol. Cell. Biol. *27*, 6546–6554.

Heller, R.C., and Marians, K.J. (2005a). The disposition of nascent strands at stalled replication forks dictates the pathway of replisome loading during restart. Mol. Cell *17*, 733–743.

Heller, R.C., and Marians, K.J. (2005b). Unwinding of the nascent lagging strand by Rep and PriA enables the direct restart of stalled replication forks. J. Biol. Chem. *280*, 34143–34151.

Hénaut, A., Rouxel, T., Gleizes, A., Moszer, I., and Danchin, A. (1996). Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. J. Mol. Biol. *257*, 574–585.

Hiller, B., Achleitner, M., Glage, S., Naumann, R., Behrendt, R., and Roers, A. (2012). Mammalian RNase H2 removes ribonucleotides from DNA to maintain genome integrity. J. Exp. Med. *209*, 1419–1426.

Hirano, M., and Noda, T. (2004). Genomic organization of the mouse Msh4 gene producing bicistronic, chimeric and antisense mRNA. Gene *342*, 165–177.

Hoege, C., Pfander, B., Moldovan, G.-L., Pyrowolakis, G., and Jentsch, S. (2002). RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. Nature *419*, 135–141.

Höhl, M., Kurtz, S., and Ohlebusch, E. (2002). Efficient multiple genome alignment. Bioinformatics *18 Suppl 1*, S312–S320.

Holliday, R. (1964). A mechanism for gene conversion in fungi. Genet. Res *5*, 282–304.

Holmes, J., Clark, S., and Modrich, P. (1990). Strand-specific mismatch correction in nuclear extracts of human and *Drosophila melanogaster* cell lines. Proc. Natl. Acad. Sci. U. S. A. *87*, 5837–5841.

Huang, S.N., and Crothers, D.M. (2008). The role of nucleotide cofactor binding in cooperativity and specificity of MutS recognition. J. Mol. Biol. *384*, 31–47.

Indiani, C., McInerney, P., Georgescu, R., Goodman, M.F., and O'Donnell, M. (2005). A sliding-clamp toolbelt binds high- and low-fidelity DNA polymerases simultaneously. Mol. Cell *19*, 805–815.

Ionov, Y., Peinado, M.A., Malkhosyan, S., Shibata, D., and Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature *363*, 558–561.

Jackson, A., Okely, E.A., and Leach, D.R.F. (2014). Expansion of CAG Repeats in *Escherichia coli* Is Controlled by Single-Strand DNA Exonucleases of Both Polarities. Genetics.

Jackson, A.P., Thomas, G.H., Parkhill, J., and Thomson, N.R. (2009). Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement. BMC Genomics *10*, 584.

Jacob, S., and Praz, F. (2002). DNA mismatch repair defects: role in colorectal carcinogenesis. Biochimie *84*, 27–47.

Jeong, C., Cho, W.-K., Song, K.-M., Cook, C., Yoon, T.-Y., Ban, C., Fishel, R., and Lee, J.-B. (2011). MutS switches between two fundamentally distinct clamps during mismatch repair. Nat. Struct. Mol. Biol. *18*, 379–385.

Ji, J., Clegg, N.J., Peterson, K.R., Jackson, A.L., Laird, C.D., and Loeb, L.A. (1996). In vitro expansion of GGC:GCC repeats: identification of the preferred strand of expansion. Nucleic Acids Res. *24*, 2835–2840.

Jiricny, J. (1998). Replication errors: cha(lle)nging the genome. EMBO J. *17*, 6427–6436.

Jiricny, J. (2000). Mismatch repair: the praying hands of fidelity. Curr. Biol. *10*, R788–R790.

Jiricny, J. (2006). The multifaceted mismatch-repair system. Nat. Rev. Mol. Cell Biol. *7*, 335–346.

Jiricny, J. (2013). Postreplicative mismatch repair. Cold Spring Harb. Perspect. Biol. *5*, a012633.

Jiricny, J., Su, S.S., Wood, S.G., and Modrich, P. (1988a). Mismatch-containing oligonucleotide duplexes bound by the *E. coli* mutS-encoded protein. Nucleic Acids Res. *16*, 7843–7853.

Jiricny, J., Hughes, M., Corman, N., and Rudkin, B.B. (1988b). A human 200-kDa protein binds selectively to DNA fragments containing G.T mismatches. Proc. Natl. Acad. Sci. U. S. A. *85*, 8860–8864.

Jones, J.M., and Nakai, H. (1999). Duplex opening by primosome protein PriA for replisome assembly on a recombination intermediate. J. Mol. Biol. *289*, 503–516.

Jun, S.-H., Kim, T.G., and Ban, C. (2006). DNA mismatch repair system. Classical and fresh roles. FEBS J. *273*, 1609–1619.

Junop, M.S., Obmolova, G., Rausch, K., Hsieh, P., and Yang, W. (2001). Composite active site of an ABC ATPase: MutS uses ATP to verify mismatch recognition and authorize DNA repair. Mol. Cell *7*, 1–12.

Kadyrov, F. a, Dzantiev, L., Constantin, N., and Modrich, P. (2006). Endonucleolytic function of MutLalpha in human mismatch repair. Cell *126*, 297–308.

Kadyrov, F.A., Holmes, S.F., Arana, M.E., Lukianova, O.A., O'Donnell, M., Kunkel, T.A., and Modrich, P. (2007). *Saccharomyces cerevisiae* MutLalpha is a mismatch repair endonuclease. J. Biol. Chem. *282*, 37181–37190.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. *42*, D199–D205.

Kang, S., Jaworski, A., Ohshima, K., and Wells, R.D. (1995). Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in *E. coli*. Nat. Genet. *10*, 213–218.

Kang, S., Lee, H., Han, J.S., and Hwang, D.S. (1999). Interaction of SeqA and Dam methylase on the hemimethylated origin of *Escherichia coli* chromosomal DNA replication. J. Biol. Chem. *274*, 11463–11468.

Karlin, S., and Cardon, L.R. (1994). Computational DNA sequence analysis. Annu. Rev. Microbiol. *48*, 619–654.

Kelsoe, G. (1996). The germinal center: a crucible for lymphocyte selection. Semin. Immunol. *8*, 179–184.

Kijas, A.W., Studamire, B., and Alani, E. (2003). Msh2 separation of function mutations confer defects in the initiation steps of mismatch repair. J. Mol. Biol. *331*, 123–138.

Kinch, L.N., Ginalski, K., Rychlewski, L., and Grishin, N. V (2005). Identification of novel restriction endonuclease-like fold families among hypothetical proteins. Nucleic Acids Res. *33*, 3598–3605.

Kolodner, R. (1996). Biochemistry and genetics of eukaryotic mismatch repair. Genes Dev. *10*, 1433–1442.

Kovtun, I. V, and McMurray, C.T. (2001). Trinucleotide expansion in haploid germ cells by gap repair. Nat. Genet. *27*, 407–411.

Kovtun, I. V, and McMurray, C.T. (2008). Features of trinucleotide repeat instability in vivo. Cell Res. *18*, 198–213.

Krasilnikova, M.M., and Mirkin, S.M. (2004). Analysis of triplet repeat replication by two-dimensional gel electrophoresis. Methods Mol. Biol. *277*, 19–28.

Kunkel, T. a, and Erie, D.A. (2005). DNA mismatch repair. Annu. Rev. Biochem. *74*, 681–710.

Kuzminov, A. (1999). Recombinational repair of DNA damage in *Escherichia coli* and bacteriophage lambda. Microbiol. Mol. Biol. Rev. *63*, 751–813.

Lahue, R.S., Su, S.S., and Modrich, P. (1987). Requirement for d(GATC) sequences in *Escherichia coli* mutHLS mismatch correction. Proc. Natl. Acad. Sci. U. S. A. *84*, 1482–1486.

Lahue, R.S., Au, K.G., and Modrich, P. (1989). DNA mismatch correction in a defined system. Science. *245*, 160–164.

Laken, S.J., Petersen, G.M., Gruber, S.B., Oddoux, C., Ostrer, H., Giardiello, F.M., Hamilton, S.R., Hampel, H., Markowitz, A., Klimstra, D., et al. (1997). Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. Nat. Genet. *17*, 79–83.

Lamers, M.H., Perrakis, A., Enzlin, J.H., Winterwerp, H.H., de Wind, N., and Sixma, T.K. (2000). The crystal structure of DNA mismatch repair protein MutS binding to a G x T mismatch. Nature *407*, 711–717.

Lamers, M.H., Winterwerp, H.H.K., and Sixma, T.K. (2003). The alternating ATPase domains of MutS control DNA mismatch repair. EMBO J. *22*, 746–756.

Lamers, M.H., Georgijevic, D., Lebbink, J.H., Winterwerp, H.H.K., Agianian, B., de Wind, N., and Sixma, T.K. (2004). ATP increases the affinity between MutS ATPase domains. Implications for ATP hydrolysis and conformational changes. J. Biol. Chem. *279*, 43879–43885.

Lau, P.J., and Kolodner, R.D. (2003). Transfer of the MSH2.MSH6 complex from proliferating cell nuclear antigen to mispaired bases in DNA. J. Biol. Chem. *278*, 14–17.

Leach, F.S., Nicolaides, N.C., Papadopoulos, N., Liu, B., Jen, J., Parsons, R., Peltomäki, P., Sistonen, P., Aaltonen, L.A., and Nyström-Lahti, M. (1993). Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. Cell *75*, 1215–1225.

LeBowitz, J.H., and McMacken, R. (1986). The *Escherichia coli* dnaB replication protein is a DNA helicase. J. Biol. Chem. *261*, 4738–4748.

Lee, E.H., and Kornberg, A. (1991). Replication deficiencies in priA mutants of *Escherichia coli* lacking the primosomal replication n' protein. Proc. Natl. Acad. Sci. U. S. A. *88*, 3029–3032.

Lee, M.S., and Marians, K.J. (1989). The *Escherichia coli* primosome can translocate actively in either direction along a DNA strand. J. Biol. Chem. *264*, 14531–14542.

Lee, J.Y., Chang, J., Joseph, N., Ghirlando, R., Rao, D.N., and Yang, W. (2005). MutH complexed with hemi- and unmethylated DNAs: coupling base recognition and DNA cleavage. Mol. Cell *20*, 155–166.

Lee, S.D., Surtees, J.A., and Alani, E. (2007). *Saccharomyces cerevisiae* MSH2-MSH3 and MSH2-MSH6 complexes display distinct requirements for DNA binding domain I in mismatch recognition. J. Mol. Biol. *366*, 53–66.

Lehman, I.R., and Nussbaum, A.L. (1964). THE DEOXYRIBONUCLEASES OF *ESCHERICHIA COLI*. V. ON THE SPECIFICITY OF EXONUCLEASE I (PHOSPHODIESTERASE). J. Biol. Chem. *239*, 2628–2636.

Li, G.-M. (2008). Mechanisms and functions of DNA mismatch repair. Cell Res. *18*, 85–98.

Li, Z., Woo, C.J., Iglesias-Ussel, M.D., Ronai, D., and Scharff, M.D. (2004a). The generation of antibody diversity through somatic hypermutation and class switch recombination. Genes Dev. *18*, 1–11.

Li, Z., Scherer, S.J., Ronai, D., Iglesias-Ussel, M.D., Peled, J.U., Bardwell, P.D., Zhuang, M., Lee, K., Martin, A., Edelmann, W., et al. (2004b). Examination of Msh6- and Msh3-deficient mice in class switching reveals overlapping and distinct roles of MutS homologues in antibody diversification. J. Exp. Med. *200*, 47–59.

Lin, D.P., Wang, Y., Scherer, S.J., Clark, A.B., Yang, K., Avdievich, E., Jin, B., Werling, U., Parris, T., Kurihara, N., et al. (2004). An Msh2 Point Mutation Uncouples DNA Mismatch Repair and Apoptosis. Cancer Res. *64*, 517–522.

Lindblom, A., Tannergård, P., Werelius, B., and Nordenskjöld, M. (1993). Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. Nat. Genet. *5*, 279–282.

Link, A.J., Phillips, D., and Church, G.M. (1997). Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: application to open reading frame characterization. J. Bacteriol. *179*, 6228–6237.

Liu, J., Nurse, P., and Marians, K.J. (1996). The ordered assembly of the phiX174-type primosome. III. PriB facilitates complex formation between PriA and DnaT. J. Biol. Chem. *271*, 15656–15661.

Longerich, S., Basu, U., Alt, F., and Storb, U. (2006). AID in somatic hypermutation and class switch recombination. Curr. Opin. Immunol. *18*, 164–174.

Longley, M.J., Pierce, A.J., and Modrich, P. (1997). DNA polymerase delta is required for human mismatch repair in vitro. J. Biol. Chem. *272*, 10917–10921.

López de Saro, F.J., and O'Donnell, M. (2001). Interaction of the beta sliding clamp with MutS, ligase, and DNA polymerase I. Proc. Natl. Acad. Sci. U. S. A. *98*, 8376–8380.

López de Saro, F.J., Marinus, M.G., Modrich, P., and O'Donnell, M. (2006). The beta sliding clamp binds to multiple sites within MutL and MutS. J. Biol. Chem. *281*, 14340–14349.

Lopper, M., Boonsombat, R., Sandler, S.J., and Keck, J.L. (2007). A hand-off mechanism for primosome assembly in replication restart. Mol. Cell *26*, 781–793.

Lovett, S.T., and Clark, A.J. (1984). Genetic analysis of the recJ gene of *Escherichia coli* K-12. J. Bacteriol. *157*, 190–196.

Lovett, S.T., and Kolodner, R.D. (1989). Identification and purification of a single-stranded-DNA-specific exonuclease encoded by the recJ gene of *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. *86*, 2627–2631.

Lu, M., Campbell, J.L., Boye, E., and Kleckner, N. (1994). SeqA: a negative modulator of replication initiation in *E. coli*. Cell *77*, 413–426.

Lujan, S.A., Williams, J.S., Pursell, Z.F., Abdulovic-Cui, A.A., Clark, A.B., Nick McElhinny, S.A., and Kunkel, T.A. (2012). Mismatch repair balances leading and lagging strand DNA replication fidelity. PLoS Genet. *8*, e1003016.

Lujan, S.A., Williams, J.S., Clausen, A.R., Clark, A.B., and Kunkel, T.A. (2013). Ribonucleotides are signals for mismatch repair of leading-strand replication errors. Mol. Cell *50*, 437–443.

Lynch, H.T., and de la Chapelle, A. (2003). Hereditary colorectal cancer. N. Engl. J. Med. *348*, 919–932.

Lynch, H.T., Smyrk, T.C., Watson, P., Lanspa, S.J., Lynch, J.F., Lynch, P.M., Cavalieri, R.J., and Boland, C.R. (1993). Genetics, natural history, tumor spectrum, and pathology of hereditary nonpolyposis colorectal cancer: an updated review. Gastroenterology *104*, 1535–1549.

MacLennan, I.C.M. (2005). Germinal centers still hold secrets. Immunity *22*, 656–657.

Maizels, N. (2005). Immunoglobulin gene diversification. Annu. Rev. Genet. *39*, 23–46.

Makovets, S. (2009). Analysis of telomeric DNA replication using neutral-alkaline two-dimensional gel electrophoresis. Methods Mol. Biol. *521*, 169–190.

Malkov, V.A., Biswas, I., Camerini-Otero, R.D., and Hsieh, P. (1997). Photocross-linking of the NH2-terminal region of Taq MutS protein to the major groove of a heteroduplex DNA. J. Biol. Chem. *272*, 23811–23817.

Maluf, N.K. (2003). Kinetic Mechanism for Formation of the Active, Dimeric UvrD Helicase-DNA Complex. J. Biol. Chem. *278*, 31930–31940.

Manley, K., Shirley, T.L., Flaherty, L., and Messer, a (1999). Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. Nat. Genet. *23*, 471–473.

Mariam, D.H., Mengistu, Y., Hoffner, S.E., and Andersson, D.I. (2004). Effect of rpoB Mutations Conferring Rifampin Resistance on Fitness of Mycobacterium tuberculosis. Antimicrob. Agents Chemother. *48*, 1289–1294.

Marinus, M.G., and Morris, N.R. (1974). Biological function for 6-methyladenine residues in the DNA of *Escherichia coli* K12. J. Mol. Biol. *85*, 309–322.

Martín-Parras, L., Hernández, P., Martínez-Robles, M.L., and Schvartzman, J.B. (1991). Unidirectional replication as visualized by two-dimensional agarose gel electrophoresis. J. Mol. Biol. *220*, 843–853.

Martomo, S.A., Yang, W.W., Wersto, R.P., Ohkumo, T., Kondo, Y., Yokoi, M., Masutani, C., Hanaoka, F., and Gearhart, P.J. (2005). Different mutation signatures in DNA polymerase eta- and MSH6-deficient mice suggest separate roles in antibody diversification. Proc. Natl. Acad. Sci. U. S. A. *102*, 8656–8661.

Masuda, K., Ouchida, R., Takeuchi, A., Saito, T., Koseki, H., Kawamura, K., Tagawa, M., Tokuhisa, T., Azuma, T., and O-Wang, J. (2005). DNA polymerase theta contributes to the generation of C/G mutations during somatic hypermutation of Ig genes. Proc. Natl. Acad. Sci. U. S. A. *102*, 13986–13991.

Matson, S.W.S.S.W., and Robertson, A.A.B. (2006). The UvrD helicase and its modulation by the mismatch repair protein MutL. Nucleic Acids Res. *34*, 4089–4097.

Mazurek, A., Johnson, C.N., Germann, M.W., and Fishel, R. (2009). Sequence context effect for hMSH2-hMSH6 mismatch-dependent activation. Proc. Natl. Acad. Sci. U. S. A. *106*, 4177–4182.

McCool, J.D., Ford, C.C., and Sandler, S.J. (2004). A dnaT mutant with phenotypes similar to those of a priA2::kan mutant in *Escherichia coli* K-12. Genetics *167*, 569–578.

McGlynn, P., Al-Deib, A.A., Liu, J., Marians, K.J., and Lloyd, R.G. (1997). The DNA replication protein PriA and the recombination protein RecG bind D-loops. J. Mol. Biol. *270*, 212–221.

McNally, R., Bowman, G.D., Goedken, E.R., O'Donnell, M., and Kuriyan, J. (2010). Analysis of the role of PCNA-DNA contacts during clamp loading. BMC Struct. Biol. *10*, 3.

Mechanic, L.E., Frankel, B.A., and Matson, S.W. (2000). *Escherichia coli* MutL loads DNA helicase II onto DNA. J. Biol. Chem. *275*, 38337–38346.

Mendillo, M.L., Mazur, D.J., and Kolodner, R.D. (2005). Analysis of the interaction between the *Saccharomyces cerevisiae* MSH2-MSH6 and MLH1-PMS1 complexes with DNA using a reversible DNA end-blocking system. J. Biol. Chem. *280*, 22245–22257.

Merlin, C., McAteer, S., and Masters, M. (2002). Tools for characterization of *Escherichia coli* genes of unknown function. J. Bacteriol. *184*, 4573–4581.

Miah, G., Rafii, M.Y., Ismail, M.R., Puteh, A.B., Rahim, H.A., Islam, K.N., and Latif, M.A. (2013). A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. Int. J. Mol. Sci. *14*, 22499–22528.

Miesel, L., and Roth, J.R. (1996). Evidence that SbcB and RecF pathway functions contribute to RecBCD-dependent transductional recombination. J. Bacteriol. *178*, 3146–3155.

Minet, A.D., Rubin, B.P., Tucker, R.P., Baumgartner, S., and Chiquet-Ehrismann, R. (1999). Teneurin-1, a vertebrate homologue of the Drosophila pair-rule gene ten-m, is a neuronal protein with a novel type of heparin-binding domain. J. Cell Sci. *112 (Pt 1*, 2019–2032.

Mirkin, S.M. (2007). Expandable DNA repeats and human disease. Nature *447*, 932–940.

Miyaki, M., Nishio, J., Konishi, M., Kikuchi-Yanoshita, R., Tanaka, K., Muraoka, M., Nagato, M., Chong, J.M., Koike, M., Terada, T., et al. (1997). Drastic genetic instability of tumors and normal tissues in Turcot syndrome. Oncogene *15*, 2877–2881.

Modrich, P. (1989). Methyl-directed DNA mismatch correction. J. Biol. Chem. *264*, 6597–6600.

Modrich, P. (1991). Mechanisms and biological effects of mismatch repair. Annu. Rev. Genet. *25*, 229–253.

Modrich, P., and Lahue, R. (1996). Mismatch repair in replication fidelity, genetic recombination, and cancer biology. Annu. Rev. Biochem. *65*, 101–133.

Monti, M.C., Cohen, S.X., Fish, A., Winterwerp, H.H.K., Barendregt, A., Friedhoff, P., Perrakis, A., Heck, A.J.R., Sixma, T.K., van den Heuvel, R.H.H., et al. (2011). Native mass spectrometry provides direct evidence for DNA mismatch-induced regulation of asymmetric nucleotide binding in mismatch repair protein MutS. Nucleic Acids Res. *39*, 8052–8064.

Muramatsu, M., Sankaranand, V.S., Anant, S., Sugai, M., Kinoshita, K., Davidson, N.O., and Honjo, T. (1999). Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. J. Biol. Chem. *274*, 18470–18476.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell *102*, 553–563.

Nakahara, T., Zhang, Q.M., Hashiguchi, K., and Yonei, S. (2000). Identification of proteins of *Escherichia coli* and *Saccharomyces cerevisiae* that specifically bind to C/C mismatches in DNA. Nucleic Acids Res. *28*, 2551–2556.

Natrajan, G., Lamers, M.H., Enzlin, J.H., Winterwerp, H.H.K., Perrakis, A., and Sixma, T.K. (2003). Structures of *Escherichia coli* DNA mismatch repair enzyme MutS in complex with different mismatches: a common recognition mode for diverse substrates. Nucleic Acids Res. *31*, 4814–4821.

Nicholson, W.L., and Maughan, H. (2002). The Spectrum of Spontaneous Rifampin Resistance Mutations in the rpoB Gene of Bacillussubtilis 168 Spores Differs from That of Vegetative Cells and Resembles That of Mycobacterium tuberculosis. J. Bacteriol. *184*, 4936–4940.

Nick McElhinny, S.A., Kissling, G.E., and Kunkel, T.A. (2010a). Differential correction of lagging-strand replication errors made by DNA polymerases {alpha} and {delta}. Proc. Natl. Acad. Sci. U. S. A. *107*, 21070–21075.

Nick McElhinny, S.A., Kumar, D., Clark, A.B., Watt, D.L., Watts, B.E., Lundström, E.-B., Johansson, E., Chabes, A., and Kunkel, T.A. (2010b). Genome instability due to ribonucleotide incorporation into DNA. Nat. Chem. Biol. *6*, 774–781.

Di Noia, J.M., and Neuberger, M.S. (2007). Molecular mechanisms of antibody somatic hypermutation. Annu. Rev. Biochem. *76*, 1–22.

Nurse, P., Zavitz, K.H., and Marians, K.J. (1991). Inactivation of the *Escherichia coli* priA DNA replication protein induces the SOS response. J. Bacteriol. *173*, 6686–6693.

Nurse, P., Liu, J., and Marians, K.J. (1999). Two modes of PriA binding to DNA. J. Biol. Chem. *274*, 25026–25032.

Nyström-Lahti, M., Sistonen, P., Mecklin, J.P., Pylkkänen, L., Aaltonen, L.A., Järvinen, H., Weissenbach, J., de la Chapelle, A., and Peltomäki, P. (1994). Close linkage to chromosome 3p and conservation of ancestral founding haplotype in hereditary nonpolyposis colorectal cancer families. Proc. Natl. Acad. Sci. U. S. A. *91*, 6054–6058.

Obmolova, G., Ban, C., Hsieh, P., and Yang, W. (2000). Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. Nature *407*, 703–710.

Odsbu, I., Klungsøyr, H.K., Fossum, S., and Skarstad, K. (2005). Specific N-terminal interactions of the *Escherichia coli* SeqA protein are required to form multimers that restrain negative supercoils and form foci. Genes Cells *10*, 1039–1049.

Ohmiya, N., Matsumoto, S., Yamamoto, H., Baranovskaya, S., Malkhosyan, S.R., and Perucho, M. (2001). Germline and somatic mutations in hMSH6 and hMSH3 in gastrointestinal cancers of the microsatellite mutator phenotype. Gene *272*, 301–313.

Oppenheim, A. (1981). Separation of closed circular DNA from linear DNA by electrophoresis in two dimensions in agarose gels. Nucleic Acids Res. *9*, 6805–6812.

Oshima, T., Wada, C., Kawagoe, Y., Ara, T., Maeda, M., Masuda, Y., Hiraga, S., and Mori, H. (2002). Genome-wide analysis of deoxyadenosine methyltransferase-mediated control of gene expression in *Escherichia coli*. Mol. Microbiol. *45*, 673–695.

Palombo, F., Gallinari, P., Iaccarino, I., Lettieri, T., Hughes, M., D'Arrigo, A., Truong, O., Hsuan, J.J., and Jiricny, J. (1995). GTBP, a 160-kilodalton protein essential for mismatch-binding activity in human cells. Science. *268*, 1912–1914.

Papadopoulos, N., Nicolaides, N.C., Wei, Y.F., Ruben, S.M., Carter, K.C., Rosen, C.A., Haseltine, W.A., Fleischmann, R.D., Fraser, C.M., and Adams, M.D. (1994). Mutation of a mutL homolog in hereditary colon cancer. Science *263*, 1625–1629.

Parker, B.O., and Marinus, M.G. (1992). Repair of DNA heteroduplexes containing small heterologous sequences in *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. *89*, 1730–1734.

Parsons, R., Li, G.M., Longley, M.J., Fang, W.H., Papadopoulos, N., Jen, J., de la Chapelle, A., Kinzler, K.W., Vogelstein, B., and Modrich, P. (1993). Hypermutability and mismatch repair deficiency in RER+ tumor cells. Cell *75*, 1227–1236.

Pavlov, Y.I., Rogozin, I.B., Galkin, A.P., Aksenova, A.Y., Hanaoka, F., Rada, C., and Kunkel, T.A. (2002). Correlation of somatic hypermutation specificity and A-T base pair substitution errors by DNA polymerase eta during copying of a mouse immunoglobulin kappa light chain transgene. Proc. Natl. Acad. Sci. U. S. A. *99*, 9954–9959.

Pavlov, Y.I., Mian, I.M., and Kunkel, T.A. (2003). Evidence for preferential mismatch repair of lagging strand DNA replication errors in yeast. Curr. Biol. *13*, 744–748.

Pearson, C.E., Nichol Edamura, K., and Cleary, J.D. (2005). Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. *6*, 729–742.

Peled, J.U., Kuang, F.L., Iglesias-Ussel, M.D., Roa, S., Kalis, S.L., Goodman, M.F., and Scharff, M.D. (2008). The biochemistry of somatic hypermutation. Annu. Rev. Immunol. *26*, 481–511.

Peltomäki, P., Aaltonen, L.A., Sistonen, P., Pylkkänen, L., Mecklin, J.P., Järvinen, H., Green, J.S., Jass, J.R., Weber, J.L., and Leach, F.S. (1993). Genetic mapping of a locus predisposing to human colorectal cancer. Science *260*, 810–812.

Peña-Diaz, J., and Jiricny, J. (2012). Mammalian mismatch repair: error-free or error-prone? Trends Biochem. Sci. *37*, 206–214.

Pham, P., Bransteitter, R., Petruska, J., and Goodman, M.F. (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature *424*, 103–107.

Plotz, G. (2002). hMutSalpha forms an ATP-dependent complex with hMutLalpha and hMutLbeta on DNA. Nucleic Acids Res. *30*, 711–718.

Pluciennik, A., and Modrich, P. (2007). Protein roadblocks and helix discontinuities are barriers to the initiation of mismatch repair. Proc. Natl. Acad. Sci. U. S. A. *104*, 12709–12713.

Pluciennik, A., Burdett, V., Lukianova, O., O'Donnell, M., and Modrich, P. (2009). Involvement of the beta clamp in methyl-directed mismatch repair in vitro. J. Biol. Chem. *284*, 32782–32791.

Pluciennik, A., Dzantiev, L., Iyer, R.R., Constantin, N., Kadyrov, F.A., and Modrich, P. (2010). PCNA function in the activation and strand direction of MutLα endonuclease in mismatch repair. Proc. Natl. Acad. Sci. U. S. A. *107*, 16066–16071.

Pohlhaus, J.R., and Kreuzer, K.N. (2006). Formation and processing of stalled replication forks--utility of two-dimensional agarose gels. Methods Enzymol. *409*, 477–493.

Poole, S.J., Diner, E.J., Aoki, S.K., Braaten, B.A., t'Kint de Roodenbeke, C., Low, D.A., and Hayes, C.S. (2011). Identification of functional toxin/immunity genes linked to contact-dependent growth inhibition (CDI) and rearrangement hotspot (Rhs) systems. PLoS Genet. *7*, e1002217.

Pukkila, P.J., Peterson, J., Herman, G., Modrich, P., and Meselson, M. (1983). Effects of high levels of DNA adenine methylation on methyl-directed mismatch repair in *Escherichia coli*. Genetics *104*, 571–582.

Rada, C., Ehrenstein, M.R., Neuberger, M.S., and Milstein, C. (1998). Hot spot focusing of somatic hypermutation in MSH2-deficient mice suggests two stages of mutational targeting. Immunity *9*, 135–141.

Ramilo, C., Gu, L., Guo, S., Zhang, X., Patrick, S.M., Turchi, J.J., and Li, G.-M.G.-M. (2002). Partial reconstitution of human DNA mismatch repair in vitro: characterization of the role of human replication protein A. Mol. Cell. Biol. *22*, 2037–2046.

Räschle, M., Dufner, P., Marra, G., and Jiricny, J. (2002). Mutations within the hMLH1 and hPMS2 subunits of the human MutLalpha mismatch repair factor affect its ATPase activity, but not its ability to interact with hMutSalpha. J. Biol. Chem. *277*, 21810–21820.

Razavy, H., Szigety, S.K., and Rosenberg, S.M. (1996). Evidence for both 3' and 5' single-strand DNA ends in intermediates in chi-stimulated recombination in vivo. Genetics *142*, 333–339.

Reenan, R.A., and Kolodner, R.D. (1992). Isolation and characterization of two *Saccharomyces cerevisiae* genes encoding homologs of the bacterial HexA and MutS mismatch repair proteins. Genetics *132*, 963–973.

Reijns, M.A.M., Rabe, B., Rigby, R.E., Mill, P., Astell, K.R., Lettice, L.A., Boyle, S., Leitch, A., Keighren, M., Kilanowski, F., et al. (2012). Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development. Cell *149*, 1008–1022.

Riva, A., Delorme, M.-O., Chevalier, T., Guilhot, N., Hénaut, C., and Hénaut, A. (2004a). The difficult interpretation of transcriptome data: the case of the GATC regulatory network. Comput. Biol. Chem. *28*, 109–118.

Riva, A., Delorme, M.-O., Chevalier, T., Guilhot, N., Hénaut, C., and Hénaut, A. (2004b). Characterization of the GATC regulatory network in *E. coli*. BMC Genomics *5*, 48.

Robertson, A.B., Pattishall, S.R., Gibbons, E.A., and Matson, S.W. (2006). MutL-catalyzed ATP hydrolysis is required at a post-UvrD loading step in methyl-directed mismatch repair. J. Biol. Chem. *281*, 19949–19959.

Rogozin, I.B., Pavlov, Y.I., Bebenek, K., Matsuda, T., and Kunkel, T.A. (2001). Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. Nat. Immunol. *2*, 530–536.

Rydberg, B., and Game, J. (2002). Excision of misincorporated ribonucleotides in DNA by RNase H (type 2) and FEN-1 in cell-free extracts. Proc. Natl. Acad. Sci. U. S. A. *99*, 16654–16659.

Sachadyn, P. (2010). Conservation and diversity of MutS proteins. Mutat. Res. *694*, 20–30.

Samadashwily, G.M., Raca, G., and Mirkin, S.M. (1997). Trinucleotide repeats affect DNA replication in vivo. Nat. Genet. *17*, 298–304.

Sandler, S.J. (2000). Multiple genetic pathways for restarting DNA replication forks in *Escherichia coli* K-12. Genetics *155*, 487–497.

Sandler, S.J., Samra, H.S., and Clark, A.J. (1996). Differential suppression of priA2::kan phenotypes in *Escherichia coli* K-12 by mutations in priA, lexA, and dnaC. Genetics *143*, 5–13.

Savouret, C., Brisson, E., Essers, J., Kanaar, R., Pastink, A., te Riele, H., Junien, C., and Gourdon, G. (2003). CTG repeat instability and size variation timing in DNA repair-deficient mice. EMBO J. *22*, 2264–2273.

Schofield, M.J., and Hsieh, P. (2003). DNA mismatch repair: molecular mechanisms and biological function. Annu. Rev. Microbiol. *57*, 579–608.

Schofield, M.J., Brownewell, F.E., Nayak, S., Du, C., Kool, E.T., and Hsieh, P. (2001). The Phe-X-Glu DNA binding motif of MutS. The role of hydrogen bonding in mismatch recognition. J. Biol. Chem. *276*, 45505–45508.

Schrader, C.E., Vardo, J., and Stavnezer, J. (2002). Role for mismatch repair proteins Msh2, Mlh1, and Pms2 in immunoglobulin class switching shown by sequence analysis of recombination junctions. J. Exp. Med. *195*, 367–373.

Schrader, C.E., Linehan, E.K., Mochegova, S.N., Woodland, R.T., and Stavnezer, J. (2005). Inducible DNA breaks in Ig S regions are dependent on AID and UNG. J. Exp. Med. *202*, 561–568.

Selmane, T., Schofield, M.J., Nayak, S., Du, C., and Hsieh, P. (2003). Formation of a DNA mismatch repair complex mediated by ATP. J. Mol. Biol. *334*, 949–965.

Shell, S.S., Putnam, C.D., and Kolodner, R.D. (2007). Chimeric *Saccharomyces cerevisiae* Msh6 protein with an Msh3 mispair-binding domain combines properties of both proteins. Proc. Natl. Acad. Sci. U. S. A. *104*, 10956–10961.

Slater, S., Wold, S., Lu, M., Boye, E., Skarstad, K., and Kleckner, N. (1995). *E. coli* SeqA protein binds oriC in two different methyl-modulated reactions appropriate to its roles in DNA replication initiation and origin sequestration. Cell *82*, 927–936.

Smith, B.T., Grossman, A.D., and Walker, G.C. (2001). Visualization of mismatch repair in bacterial cells. Mol. Cell *8*, 1197–1206.

Sparks, J.L., Chon, H., Cerritelli, S.M., Kunkel, T.A., Johansson, E., Crouch, R.J., and Burgers, P.M. (2012). RNase H2-initiated ribonucleotide excision repair. Mol. Cell *47*, 980–986.

Spell, R.M., and Jinks-Robertson, S. (2004). Determination of mitotic recombination rates by fluctuation analysis in *Saccharomyces cerevisiae*. Methods Mol. Biol. *262*, 3–12.

Stavnezer, J., Guikema, J.E.J., and Schrader, C.E. (2008). Mechanism and regulation of class switch recombination. Annu. Rev. Immunol. *26*, 261–292.

Storb, U. (1998). Progress in understanding the mechanism and consequences of somatic hypermutation. Immunol. Rev. *162*, 5–11.

Stukenberg, P.T., Studwell-Vaughan, P.S., and O'Donnell, M. (1991). Mechanism of the sliding beta-clamp of DNA polymerase III holoenzyme. J. Biol. Chem. *266*, 11328–11334.

Su, S.S., and Modrich, P. (1986). *Escherichia coli* mutS-encoded protein binds to mismatched DNA base pairs. Proc. Natl. Acad. Sci. U. S. A. *83*, 5057–5061.

Su, S.S., Lahue, R.S., Au, K.G., and Modrich, P. (1988). Mispair specificity of methyl-directed DNA mismatch correction in vitro [published erratum appears in J Biol Chem 1988 Aug 5;263(22):11015]. J. Biol. Chem. *263*, 6829–6835.

Su, S.S., Grilley, M., Thresher, R., Griffith, J., and Modrich, P. (1989). Gap formation is associated with methyl-directed mismatch correction under conditions of restricted DNA synthesis. Genome *31*, 104–111.

Sugawara, N., Pâques, F., Colaiácovo, M., and Haber, J.E. (1997). Role of *Saccharomyces cerevisiae* Msh2 and Msh3 repair proteins in double-strand break-induced recombination. Proc. Natl. Acad. Sci. U. S. A. *94*, 9214–9219.

Surtees, J.A., and Alani, E. (2006). Mismatch repair factor MSH2-MSH3 binds and alters the conformation of branched DNA structures predicted to form during genetic recombination. J. Mol. Biol. *360*, 523–536.

Swann, P.F., Waters, T.R., Moulton, D.C., Xu, Y.Z., Zheng, Q., Edwards, M., and Mace, R. (1996). Role of postreplicative DNA mismatch repair in the cytotoxic action of thioguanine. Science *273*, 1109–1111.

Tavazoie, S., and Church, G.M. (1998). Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in *E. coli*. Nat. Biotechnol. *16*, 566–571.

Teng, G., and Papavasiliou, F.N. (2007). Immunoglobulin somatic hypermutation. Annu. Rev. Genet. *41*, 107–120.

Thibodeau, S.N., Bren, G., and Schaid, D. (1993). Microsatellite instability in cancer of the proximal colon. Science *260*, 816–819.

Thomas, D.C., Roberts, J.D., and Kunkel, T. a (1991). Heteroduplex repair in extracts of human HeLa cells. J. Biol. Chem. *266*, 3744–3751.

Tishkoff, D.X., Filosi, N., Gaida, G.M., and Kolodner, R.D. (1997). A novel mutation avoidance mechanism dependent on S. cerevisiae RAD27 is distinct from DNA mismatch repair. Cell *88*, 253–263.

Touzain, F., Petit, M.-A., Schbath, S., and El Karoui, M. (2011). DNA motifs that sculpt the bacterial chromosome. Nat. Rev. Microbiol. *9*, 15–26.

Tsurimoto, T. (1999). PCNA binding proteins. Front. Biosci. *4*, D849–D858.

Urig, S., Gowher, H., Hermann, A., Beck, C., Fatemi, M., Humeny, A., and Jeltsch, A. (2002). The *Escherichia coli* dam DNA methyltransferase modifies DNA in a highly processive reaction. J. Mol. Biol. *319*, 1085–1096.

Viswanathan, M., and Lovett, S.T. (1998). Single-strand DNA-specific exonucleases in *Escherichia coli*. Roles in repair and mutation avoidance. Genetics *149*, 7–16.

Viswanathan, M., Lacirignola, J.J., Hurley, R.L., and Lovett, S.T. (2000). A novel mutational hotspot in a natural quasipalindrome in *Escherichia coli*. J. Mol. Biol. *302*, 553–564.

Viswanathan, M., Burdett, V., Baitinger, C., Modrich, P., and Lovett, S.T. (2001). Redundant exonuclease involvement in *Escherichia coli* methyl-directed mismatch repair. J. Biol. Chem. *276*, 31053–31058.

Wagner, R., and Meselson, M. (1976). Repair tracts in mismatched DNA heteroduplexes. Proc. Natl. Acad. Sci. U. S. A. *73*, 4135–4139.

Waldminghaus, T., Weigel, C., and Skarstad, K. (2012). Replication fork movement and methylation govern SeqA binding to the *Escherichia coli* chromosome. Nucleic Acids Res. *40*, 5465–5476.

Wang, H., and Hays, J.B. (2003). Mismatch repair in human nuclear extracts: effects of internal DNA-hairpin structures between mismatches and excision-initiation nicks on mismatch correction and mismatch-provoked excision. J. Biol. Chem. *278*, 28686–28693.

Wang, H., and Hays, J.B. (2004). Signaling from DNA mispairs to mismatch-repair excision sites despite intervening blockades. EMBO J. *23*, 2126–2133.

Wang, J.Y.J., and Edelmann, W. (2006). Mismatch repair proteins as sensors of alkylation DNA damage. Cancer Cell *9*, 417–418.

Wang, M.X., and Church, G.M. (1992). A whole genome approach to in vivo DNA-protein interactions in *E. coli*. Nature *360*, 606–610.

Wang, Q., Lasset, C., Desseigne, F., Frappaz, D., Bergeron, C., Navarro, C., Ruano, E., and Puisieux, A. (1999). Neurofibromatosis and early onset of cancers in hMLH1-deficient children. Cancer Res. *59*, 294–297.

Wang, Y.D., Zhao, S., and Hill, C.W. (1998). Rhs elements comprise three subfamilies which diverged prior to acquisition by *Escherichia coli*. J. Bacteriol. *180*, 4102–4110.

Warbrick, E. (2000). The puzzle of PCNA's many partners. Bioessays *22*, 997–1006.

Warbrick, E. (2006). A functional analysis of PCNA-binding peptides derived from protein sequence, interaction screening and rational design. Oncogene *25*, 2850–2859.

Warren, J.J., Pohlhaus, T.J., Changela, A., Iyer, R.R., Modrich, P.L., and Beese, L.S. (2007). Structure of the human MutSalpha DNA lesion recognition complex. Mol. Cell *26*, 579–592.

Waters, T.R., and Swann, P.F. (1997). Cytotoxic mechanism of 6-thioguanine: hMutSalpha, the human mismatch binding heterodimer, binds to DNA containing S6-methylthioguanine. Biochemistry *36*, 2501–2506.

Wei, K., Clark, A.B., Wong, E., Kane, M.F., Mazur, D.J., Parris, T., Kolas, N.K., Russell, R., Hou, H., Kneitz, B., et al. (2003). Inactivation of Exonuclease 1 in mice results in DNA mismatch repair defects, increased cancer susceptibility, and male and female sterility. Genes Dev. *17*, 603–614.

Wellauer, P.K., Dawid, I.B., Brown, D.D., and Reeder, R.H. (1976). The molecular basis for length heterogeneity in ribosomal DNA from Xenopus laevis. J. Mol. Biol. *105*, 461–486.

Welsh, K.M., Lu, A.L., Clark, S., and Modrich, P. (1987). Isolation and characterization of the *Escherichia coli* mutH gene product. J. Biol. Chem. *262*, 15624–15629.

Whitehouse, H.L.K. (1963). A Theory of Crossing-Over by Means of Hybrid Deoxyribonucleic Acid. Nature *199*, 1034–1040.

Wiesendanger, M., Kneitz, B., Edelmann, W., and Scharff, M.D. (2000). Somatic hypermutation in MutS homologue (MSH)3-, MSH6-, and MSH3/MSH6-deficient mice reveals a role for the MSH2-MSH6 heterodimer in modulating the base substitution pattern. J. Exp. Med. *191*, 579–584.

Wijnen, J., de Leeuw, W., Vasen, H., van der Klift, H., Møller, P., Stormorken, A., Meijers-Heijboer, H., Lindhout, D., Menko, F., Vossen, S., et al. (1999). Familial endometrial cancer in female carriers of MSH6 germline mutations. Nat. Genet. *23*, 142–144.

Winand, N.J., Panzer, J.A., and Kolodner, R.D. (1998). Cloning and characterization of the human and *Caenorhabditis elegans* homologs of the *Saccharomyces cerevisiae MSH5* gene. Genomics *53*, 69–80.

Witkin, E.M., and Sicurella, N.A. (1964). PURE CLONES OF LACTOSE-NEGATIVE MUTANTS OBTAINED IN *ESCHERICHIA COLI* AFTER TREATMENT WITH 5-BROMOURACIL. J. Mol. Biol. *8*, 610–613.

Wood, R.D., and Shivji, M.K. (1997). Which DNA polymerases are used for DNA-repair in eukaryotes? Carcinogenesis *18*, 605–610.

Van der Woude, M., Hale, W.B., and Low, D.A. (1998). Formation of DNA methylation patterns: nonmethylated GATC sequences in gut and pap operons. J. Bacteriol. *180*, 5913–5920.

Wu, S.-Y., Culligan, K., Lamers, M., and Hays, J. (2003). Dissimilar mispair-recognition spectra of Arabidopsis DNA-mismatch-repair proteins MSH2*MSH6 (MutSalpha) and MSH2*MSH7 (MutSgamma). Nucleic Acids Res. *31*, 6027–6034.

Wu, X., Tsai, C.Y., Patam, M.B., Zan, H., Chen, J.P., Lipkin, S.M., and Casali, P. (2006). A role for the MutL mismatch repair Mlh3 protein in immunoglobulin class switch DNA recombination and somatic hypermutation. J. Immunol. *176*, 5426–5437.

Wu, Y., Berends, M.J., Post, J.G., Mensink, R.G., Verlind, E., Van Der Sluis, T., Kempinga, C., Sijmons, R.H., van der Zee, A.G., Hollema, H., et al. (2001a). Germline mutations of EXO1 gene in patients with hereditary nonpolyposis colorectal cancer (HNPCC) and atypical HNPCC forms. Gastroenterology *120*, 1580–1587.

Wu, Y., Berends, M.J., Sijmons, R.H., Mensink, R.G., Verlind, E., Kooi, K.A., van der Sluis, T., Kempinga, C., van dDer Zee, A.G., Hollema, H., et al. (2001b). A role for MLH3 in hereditary nonpolyposis colorectal cancer. Nat. Genet. *29*, 137–138.

Xu, L., and Marians, K.J. (2003). PriA mediates DNA replication pathway choice at recombination intermediates. Mol. Cell *11*, 817–826.

Yamaguchi, M., Dao, V., and Modrich, P. (1998). MutS and MutL activate DNA helicase II in a mismatch-dependent manner. J. Biol. Chem. *273*, 9197–9201.

Yang, G., Scherer, S.J., Shell, S.S., Yang, K., Kim, M., Lipkin, M., Kucherlapati, R., Kolodner, R.D., and Edelmann, W. (2004). Dominant effects of an Msh6 missense mutation on DNA repair and cancer susceptibility. Cancer Cell *6*, 139–150.

Yin, J., Kong, D., Wang, S., Zou, T.T., Souza, R.F., Smolinski, K.N., Lynch, P.M., Hamilton, S.R., Sugimura, H., Powell, S.M., et al. (1997). Mutation of hMSH3 and hMSH6 mismatch repair genes in genetically unstable human colorectal and gastric carcinomas. Hum. Mutat. *10*, 474–478.

Yoshioka, K., Yoshioka, Y., and Hsieh, P. (2006). ATR kinase activation mediated by MutSalpha and MutLalpha in response to cytotoxic O6-methylguanine adducts. Mol. Cell *22*, 501–510.

Yuzhakov, A., Turner, J., and O'Donnell, M. (1996). Replisome assembly reveals the basis for asymmetric function in leading and lagging strand replication. Cell *86*, 877–886.

Zahra, R., Blackwood, J.K., Sales, J., and Leach, D.R.F. (2007). Proofreading and secondary structure processing determine the orientation dependence of CAG x CTG trinucleotide repeat instability in *Escherichia coli*. Genetics *176*, 27–41.

Zan, H., Shima, N., Xu, Z., Al-Qahtani, A., Evinger Iii, A.J., Zhong, Y., Schimenti, J.C., and Casali, P. (2005). The translesion DNA polymerase theta plays a dominant role in immunoglobulin gene somatic hypermutation. EMBO J. *24*, 3757–3769.

Zeng, X., Winter, D.B., Kasmer, C., Kraemer, K.H., Lehmann, A.R., and Gearhart, P.J. (2001). DNA polymerase eta is an A-T mutator in somatic hypermutation of immunoglobulin variable genes. Nat. Immunol. *2*, 537–541.

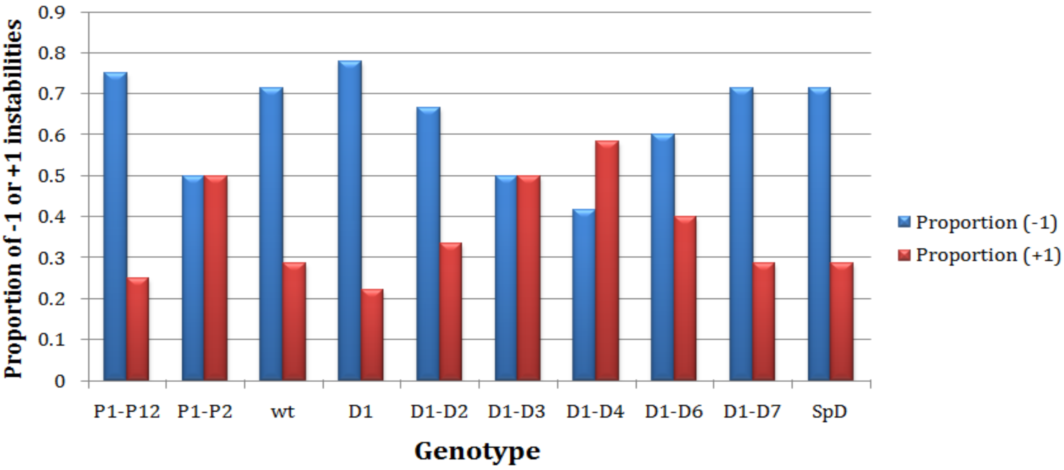


Figure A1: Proportion of deletion and expansion at single TNR unit level in different genetic backgrounds. The proportion of Single TNR unit deletion (-1) and expansion (+1) of each genetic background is shown in this bar plot. Different genetic backgrounds are plotted along the *x*-axis and the proportion of instabilities at the level of single TNR unit is shown along the *y*-axis. The blue bars depict the proportions of Single TNR unit deletion, while the red bars depict those for expansion.

	Expected	observed	O-E	SQ(O-E)	SQ(O-E)/E
P1-P12	0.5	0.75	0.25	0.0625	0.125
P1-P2	0.5	0.5	0	0	0
wt	0.5	0.714286	0.214286	0.045918367	0.091837
D1	0.5	0.777778	0.277778	0.077160494	0.154321
D1-D2	0.5	0.666667	0.166667	0.027777778	0.055556
D1-D3	0.5	0.5	0	0	0
D1-D4	0.5	0.416667	-0.08333	0.006944444	0.013889
D1-D6	0.5	0.6	0.1	0.01	0.02
D1-D7	0.5	0.714286	0.214286	0.045918367	0.091837
SpD	0.5	0.714286	0.214286	0.045918367	0.091837
				Test_statistic	0.644276
	Critical_value at 95% confidence level				16.91898

H_0 = There is no difference between expected number and the observed proportion of deletion
H_1 = There is a difference between expected number and the observed proportion of deletion
As $\chi^2 < \chi^2_{k-1,1-\alpha}$,the null hypothesis is accepted.
where, k - 1 = Degrees of freedom
$\alpha = 0.05$
Therefore, the observed difference of deletion events is not significant.

Appendix B

1. Perl script for finding out the positions of a motif

```
#!/usr/bin/perl

my $DNA_file="fasta_sequence_file";
open(DNA, $file) or die "Cannot open file $file: $!";
my @seq=<DNA>;
close DNA;

my$genome = join('',@seq);
$genome =~ s/\s//g;

my $motif = 'GATC';
my $offset = 0;
my $result = index($genome, $motif, $offset);
while ($result != -1)
{
     my $x = $result+1;
     print "$x\n";
     my $offset = $result + 4;
     my $result = index($genome, $motif, $offset);
}
```

1. Perl script for finding out the inter-GATC motif distances in each backbone- or variable-segment.

```
#!/usr/bin/perl

my $file = "backbone_segment_files";
open(DNA, $file) or die "Cannot open file $file: $!";
$/ = ">";
my $junk = <IN>;
my $i=0;
while ( my $record = <IN> )
{
     chomp $record;
     (my $defLine, my @seqLines) = split /\n/, $record;
     my $sequence = join('',@seqLines);

     my $array[$i]=$sequence;
     $i++;
}
close (IN);

my $motif = 'GATC';

my $length_array = @array;
```

```
for(my $j=0;$j<$length_array;$j++)
{
     my $offset = 0;
     print"backbone: $j\tdistance\n";
     my $result = index($array[$j], $motif, $offset);

     while ($result != -1)
     {
          my $x = $result+1;

######################################################
     #print "\tpositions: $x\n";# will return the
position of the motifs in each backbone segment along
with the respective order number.
######################################################

          push(my @position,$x);
          $offset = $result + 4;
          $result = index($array[$j], $motif, $offset);

     }
     my $length_position = @position;
     for(my $k=0;$k<($length_position-1);$k++)
     {
          my $distance = (($position[$k+1]-
$position[$k])-4);
          print"\t$distance\n";
     }

     undef(@position);# for empting the array.
}
```

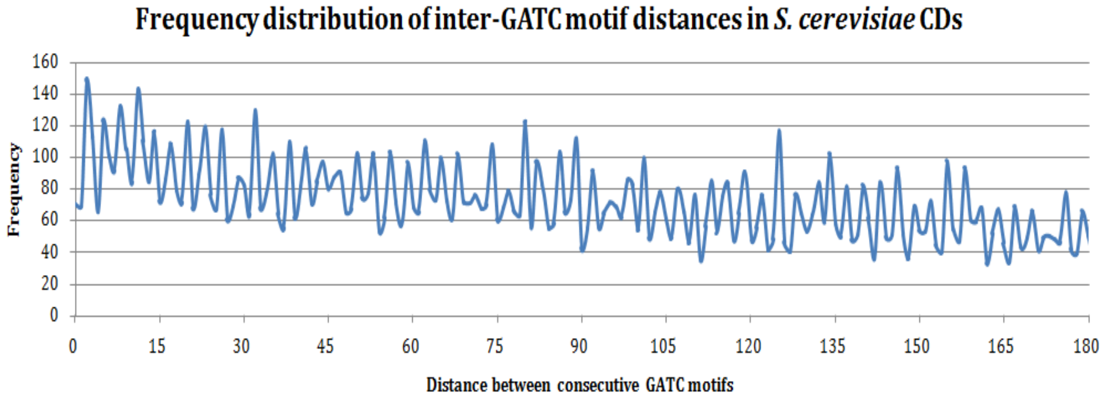


Figure C1. Frequency distribution of inter-GATC motif distances in the CDs of the *Sacchromyces cerevisiae* genome. The distances between consecutive GATC motifs had been plotted along the *x*-axis and the frequencies of these distances had been plotted along the *y*-axis.

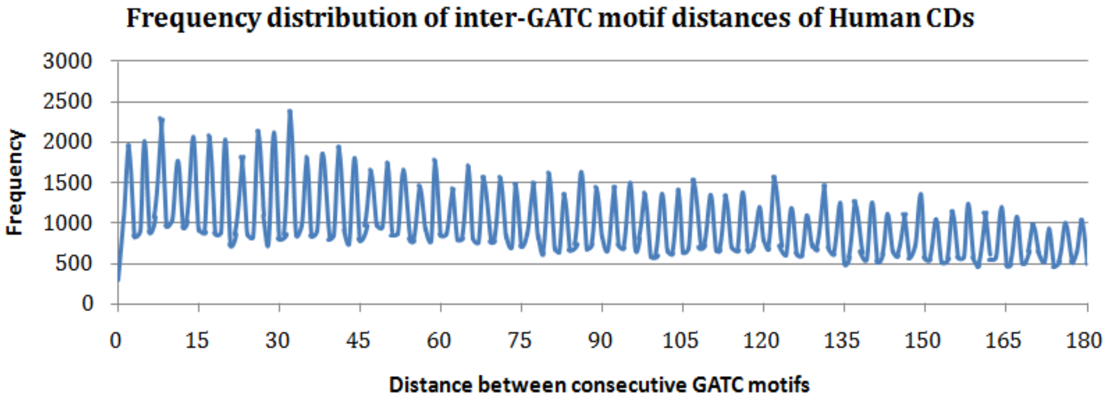


Figure C2. Frequency distribution of inter-GATC motif distances in the CDs of the human genome. The distances between consecutive GATC motifs had been plotted along the *x*-axis and the frequencies of these distances had been plotted along the *y*-axis.

The Fluctuation assay table (calculation for strain D1-D2 TNR$^+$):

D1-D2 TNR$^+$		
Dilution Factor:	1000000	100
Fraction Plated:	0.1	0.1
C.F.U.	**Rank**	**His$^+$**
93	17	326
78	16	314
100	12	272
110	23	391
122	30	507
141	29	485
94	11	270
113	32	531
114	13	275
117	27	464
122	5	244
106	19	349
97	9	260
123	22	390
130	28	475
97	14	288
90	5	244
123	15	295
102	1	133
157	25	437
95	2	145
129	8	252
100	5	244
182	35	715
194	26	456
205	31	523
102	21	361
150	24	421
136	3	225
97	10	269
95	20	351
86	17	326
105	33	589
197	36	780
135	34	652
126	3	225

121.2	**Median =**	337.5
	FREQUENCY =	2.78E-4
	RATE =	**1.67E-5**
	Culture #	
low	9	1.32E-5
high	27	2.25E-5