DNA DAMAGE, REPAIR & REPLICATION USING
E. COLI MODEL SYSTEMS

A dissertation submitted in satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

MICROBIOLOGY AND ENVIRONMENTAL TOXICOLOGY

by

Christopher J. Troll

September 2013

The Dissertation of Christopher J. Troll is
approved:

Professor Donald Smith

Professor Karen Ottemann

Professor Manel Camps

Professor Todd Lowe

Tyrus Miller
Vice provost and Dean of Graduate studies
Copyright © by
Christopher J. Troll
2013
TABLE OF CONTENTS

List of Figures .......................................................... iv
List of Tables ............................................................ vi
Abstract ...................................................................... vii
Abstract: References ......................................................... x
Dedications ..................................................................... xi
Acknowledgments ........................................................... xii
Chapter 1: Introduction to DNA damage, repair and replication .......... 1
Chapter 1: References ........................................................ 15
Chapter 2: Characterization of O\textsuperscript{6}-meG repair by naturally-occurring MGMT mutants provides insight into MGMT structure and function .......... 20
Chapter 2: References ........................................................ 44
Chapter 3: Experimental demonstration of a tradeoff between toxicity and repair capacity that likely drove the evolution of 3MeA glycosylase inducible expression ......................................................... 49
Chapter 3: References ........................................................ 84
Chapter 4: Mutagenic footprint of low-fidelity Pol I replication in \textit{E. coli} reveals an extensive interplay between Pol I and Pol III during ColE1 plasmid replication ................................................................. 88
Chapter 4: References ........................................................ 127
Chapter 5: Conclusions and future directions .......................... 130
Chapter 5: References ........................................................ 137
# LIST OF FIGURES

## Chapter 1

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Methylation of DNA</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Responses to DNA methylating agents</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>Illustration of leading strand synthesis of a ColE1 plasmid</td>
<td>12</td>
</tr>
</tbody>
</table>

## Chapter 2

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Location of non-synonymous MGMT SNP and variants</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>Inability to alkyl transfer O^6^-meG lesions leads to MNNG sensitivity while deficient transactivation leads to MMS sensitivity</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>Protection profile of MGMT mutants to MNNG</td>
<td>37</td>
</tr>
</tbody>
</table>

## Chapter 3

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Protection profile of heterologous glycosylase expression against DNA alkylating agents</td>
<td>62</td>
</tr>
<tr>
<td>2</td>
<td>Viability and protection profile of yeast MAG glycosylases</td>
<td>64</td>
</tr>
<tr>
<td>3</td>
<td>MAG viability under an inducible promoter</td>
<td>66</td>
</tr>
<tr>
<td>4</td>
<td>Viability and protection profile of fitness selected MAG1 mutants</td>
<td>71</td>
</tr>
<tr>
<td>5</td>
<td>Repair of individual lesions by MAG1</td>
<td>72</td>
</tr>
<tr>
<td>6</td>
<td>Protection profile and viability of 3MeA repair selected MAG1 mutants</td>
<td>74</td>
</tr>
<tr>
<td>7</td>
<td>Viability of deconstructed K126N E130G N175D mutant of spMAG1</td>
<td>80</td>
</tr>
<tr>
<td>8</td>
<td>Crystal structure of spMAG1</td>
<td>81</td>
</tr>
</tbody>
</table>
Chapter 4

Figure 1. Identification of mutation hotspots through clustering analysis.....100

Figure 2. Distribution of hotspot indices..................................................101

Figure 3. Identification of putative Okazaki primer processing sites.........102

Figure 4. Sequence coverage.......................................................................107

Figure 5. Comparison of leading-strand vs. lagging-strand marker mutations.........................................................................................108

Figure 6. Mutation spectrum for error-prone Pol I.....................................113

Figure 7. Rationale for determining strand specificity based on frequency of complementary mutations.................................................................114

Figure 8. Concordance between mutation spectra......................................117

Figure 9. Footprint of LF-Pol I plasmid replication....................................125
LIST OF TABLES

Chapter 1

Table 1. Lesion summary.................................................................5

Chapter 2

Table 1. MGMT SNP and their prevalence........................................29
Table 2. Summary of the work performed on the ada strains and genotyping.................................................................30

Chapter 3

Table 1. spMAG1 selection results..................................................69
Table 2. SIFT substitution for mutants..............................................79

Chapter 4

Table 1. Metrics for the libraries included in this study.........................99
Abstract

DNA DAMAGE, REPAIR & REPLICATION USING
E. COLI MODEL SYSTEMS

by

Christopher J. Troll

DNA encodes the genetic material that makes life possible but it is semi labile and prone to damage. In part, what makes DNA such a reliable genetic storage material is that organisms have developed extensive methods in order to repair damaged DNA and to preserve its integrity. Paradoxically though, genetic instability is essential for survival and one of the major causes of genetic instability is DNA damage sustained through endogenous and exogenous means. Without genetic instability, speciation, evolution, adaptation, and some aspects of aging would not occur. Unfortunately genetic instability does have negative aspects as well such as genetic diseases, cancer, and death. Therefore the interplay between DNA damage and repair is truly a double-edged sword.

*E. coli* is a valuable model organism for studying DNA damage, repair and replication due to its fast generation time and genetic malleability. Chapters 2-4 of my dissertation widely differ in their scope and aims but all the chapters take advantage of different *E. coli* genetic and complementation systems in order to answer various questions about DNA damage, DNA repair proteins, or DNA replication.

One particular form of DNA damage used extensively by chemotherapeutic agents is known as reactive DNA methylation. Reactive DNA methylation is defined, in its simplest form, as the non-enzymatic addition of methyl groups (-CH3) to
various positions on DNA nucleotides (2). Reactive DNA methylation lesions can either be innocuous, cytotoxic, or mutagenic depending on exactly where on the nucleobase the lesion occurs. O$_6$-MeG (methyl-guanine) lesions are particularly cytotoxic to humans and the only protein in the entire human body that directly repairs these lesions is called O$_6$-Methyl Guanine Methyl Transferase (MGMT) (3). In Chapter 2 of my thesis I take advantage of an *E. coli* genetic system that is deficient in its ability to repair O$_6$-MeG lesions and determine the repair abilities of all ten non-synonymous MGMT mutants detected in the human population. The results display the O$_6$-MeG repair profile of all ten non-synonymous MGMT mutants as well as give new insight into the structure of the MGMT protein.

During the first step of a DNA repair process known as Base Excision Repair (BER) DNA repair enzymes known as glycosylases recognize and remove damaged nucleotides. Base excision repair is not a perfect repair process because at high concentrations glycosylases have been known to remove undamaged bases and, in addition, the resulting abasic site can become problematic if not processed quickly by other enzymes (4). Glycosylases can be divided into two categories: those that are constitutively expressed and those that are inducible. In Chapter 3 of my thesis I utilize an *E. coli* strain that is devoid of reactive methylation specific glycosylases to understand why organisms have evolved both constitutively expressed and inducible glycosylases. The results show that the activity levels towards substrates in constitutively expressed glycosylases are limited so as not to have a negative effect on the fitness of the organism.
DNA polymerase I (Pol I) is one of two replicative polymerases in *E. coli*. ColE1 plasmids are used extensively in molecular biology. During ColE1 plasmid replication Pol I initiates leading strand synthesis and also processes Okazaki fragments, Pol III replicates the rest of the plasmid. In Chapter 4 of my thesis I utilize an *E. coli* strain with a temperature sensitive endogenous Pol I complemented with an error-prone Pol I in order to determine, with high resolution, where on ColE1 plasmids Pol I is replicating (1). The results show exactly where the switch between Pol I and Pol III occurs as well as the sites of Okazaki processing on the lagging strand. In addition the results show that a small fraction of plasmids are replicated exclusively by Pol I and that Pol I plays a far greater role in replication termination than previously thought. These latter results support earlier reports showing a functional redundancy between Pol I and Pol III and bring up interesting questions regarding the role of Pol I during chromosomal replication.
References:


DEDICATIONS

To my family, who is always interested in and excited by my research

To my friends, whose constant distractions keep me sane and excited about life

And especially to A.H. who convinced me that buying a boat while in graduate school

was impractical
ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Manel Camps for initially taking a chance on a graduate student who had very little molecular biology experience. Without his help I would not be the scientist I am today. I constantly appreciate his ability to see data in new aspects and his undeterred optimism. I would like to thank Dr. David Alexander and Jennifer Allen for teaching me much of the molecular and microbiological techniques I know. I would like to thank Dr. Brandt Eichman and the entire Eichman lab, especially Suraj Adhikary and Claire Cato, for letting me spend a summer with them learning about biochemistry and crystallography. I would like to thank my dissertation committee for their guidance and support. Finally I would like to thank the entire Microbiology and Environmental Toxicology Department for their resources and helping me complete my academic goals.
Chapter 1:

Introduction to DNA damage, repair & replication
DNA damage

DNA damage is inherent to every living organism, and chemicals and metabolites that damage our DNA are ubiquitous in our environment. Alkylating agents constitute one of the major classes of DNA damaging agents and are employed in many chemotherapeutic drugs as well as endogenously produced metabolites (17). Many alkylating agents function by modifying DNA bases with a methyl group at several positions on a given nucleotide. Depending on where the lesion occurs, reactive DNA methylation can be either cytotoxic or mutagenic.

Reactive DNA methylation, the addition of a methyl group to DNA by non-enzymatic means, occurs on all four bases including all of the exocyclic oxygens and most of the ring nitrogens (Fig. 1) (31). This is in contrast to epigenetic methylation, which refers to the enzymatic addition of a methyl group to DNA for regulation of gene expression (11), that is specific to position 5 of cytosine residues in mammalian cells (11,37,46) and to position 6 of adenine residues in prokaryotic cells (7). In reactive DNA methylation, the methyl group is transferred from an electrophilic molecule to a nucleophilic site. Sources of electrophile methyl donors (methylating agents) may be exogenous or endogenous and their frequency of attack depends on the nucleophilicity of the target site and on the strength of methylating agent. Endogenous methylating agents (byproducts of metabolism) tend to be weak and preferentially react with nitrogen moieties located within the rings of the nucleic acids, in an $S_N2$ type of reaction (44). By contrast, exogenous agents (nitrites,
chemotherapeutic agents, cigarette smoke), tend to be strong, preferentially reacting with the oxygen atoms on DNA bases and the phosphate backbone in an $S_N1$ type of reaction (43,45).

Reactive methylation has a variety of toxic effects including blocking replication and transcription (which is cytotoxic), or altering Watson-Crick base pairing (which is mutagenic) (13,31,43). The toxicity of individual lesions varies depending on their location in the DNA molecule. The most abundant lesion, $N^7$-meG is relatively innocuous although it is prone to spontaneous depurination and can become mutagenic if not acted upon in a timely manner (17). $N^1$-meA and $N^3$-meC lesions are exclusively caused by weak ($S_N2$) methylating agents. These lesions interfere with Watson-Crick base pairing and cause DNA mis-pairing and mutations. Both $N^3$-meA and $O^6$-meG lesions are highly cytotoxic, and therefore agents that produce a significant amount of these lesions are used extensively in alkylating chemotherapy. $N^3$-meA lesions protrude into the minor groove of the DNA double helix and prevent DNA polymerases from replicating past the lesion (3,4,36,41,42). $O^6$-meG lesions cause significant cytotoxicity through the initiation of a futile mismatch repair cycles that eventually leads to cell cycle arrest, chromosomal aberrations and apoptosis (25,27). The most common lesions and their biological effects are summarized in Table 1 and shown in Fig. 1.
**Fig. 1. Methylation of DNA.** Blue arrows indicate sites in DNA that are methylated by $S_N1$ agents. Red arrows indicate sites that are methylated by $S_N2$ agents. Green arrows represent sites of common methylation between the two agents (43).
<table>
<thead>
<tr>
<th>Agent</th>
<th>Lesion</th>
<th>Effect</th>
<th>Repair Enzymes</th>
<th>E. Coli Name</th>
<th>Relevant Homolog (organism)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{N1}$ &amp; $S_{N2}$</td>
<td>$N^7$-meG, $N^3$-meA</td>
<td>Mutagenic/Cytotoxic (stalls replication)</td>
<td>3-meA Glycosylases</td>
<td>AlkA</td>
<td>spMAG1 ($S. pombe$)</td>
<td>(3,4,15,17,36,41,42)</td>
</tr>
<tr>
<td>$S_{N1}$</td>
<td>$O^6$-meG</td>
<td>Mutagenic/Cytotoxic (causes $GC\rightarrow AT$ transitions)</td>
<td>DNA Alkyltransferases</td>
<td>Ada</td>
<td>MGMT or AGT (Human)</td>
<td>(25,27,38,39)</td>
</tr>
<tr>
<td></td>
<td>$O^4$-meT</td>
<td>Mutagenic (causes $TA\rightarrow CG$ transitions)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_{N2}$</td>
<td>$N^1$-meA</td>
<td>Cytotoxic/Mutagenic</td>
<td>Oxidative Demethylases</td>
<td>AlkB</td>
<td>N/A</td>
<td>(16,28)</td>
</tr>
<tr>
<td></td>
<td>$N^3$-meC</td>
<td>Cytotoxic/Mutagenic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1. Lesion summary.** The most common methyl lesions associated with $S_{N1}$ agents, $S_{N2}$ agents or both, their effects, and the enzyme responsible for repairing the lesion.
DNA repair

When *Escherichia coli* strains are exposed to DNA alkylating agents a protein known as Ada responds to this insult by upregulating its own expression and that of two other proteins involved in the repair of specific methyl lesions, AlkB and AlkA respectively (43). This coordinated response to reactive DNA alkylation is known as the adaptive response pathway and appears to be restricted to prokaryotes (Fig. 2A). The AlkB protein repairs N₁-meA and N³-meC lesions by a repair process known as oxidative demethylation (12,28). The AlkA protein is a type of repair protein known as a glycosylase and it is responsible for removing, among other substrates, the highly cytotoxic N³-meA lesion, which is generated by both S_N2 and S_N1 alkylating agents (15,29). In addition to being a transcriptional activator the Ada protein repairs O⁶-meG lesions in a 1:1 stoichiometric ratio by irreversibly transferring the methyl lesion onto a nucleophilic cysteine residue located within the active site of the protein (38,39). Together these DNA repair proteins make up the front line of defense against alkylation damage and although the coordinated upregulation of this pathway appears to be restricted to prokaryotes the repair proteins themselves are conserved across all domains of life. Since specific chapters of my thesis focus on the Ada homolog methyl guanine methyl transferase (MGMT) and the AlkA homolog spMAG1, respectively, I will now discuss each of these repair proteins in more detail.

MGMT is the sole O⁶-meG direct repair protein found in humans. Similar to *E. coli* Ada, it functions to remove O⁶-meG lesions by irreversibly transferring the deleterious alkyl group from the oxygen molecule on the guanine nucleotide to a
nucleophilic cysteine residue located within the active site of the protein but unlike Ada it does not transactive additional DNA repair genes (Fig. 2B). Since the transfer of the methyl group is irreversible, methyl transferases are degraded after the transfer is complete. In this respect methyl transferases are not true enzymes but commonly referred to as suicide enzymes. Expression levels of MGMT are important biomarkers because levels of active MGMT protein are proportional to the resistance of cells, tissues and tumors to alkylating agents (5,14,18,40). It is well established that patients with high tumor expression levels of active MGMT protein have poor response to \( S_{N1} \) chemotherapeutic alkylating agents, such as temozolomide and BCNU (8,9,21,24). To date, two polymorphisms and six variants (non-synonymous mutations with a prevalence of less than 1%) of MGMT have been identified in the human population, four of which have not been previously characterized with respect to their abilities to repair \( O^6 \)-meG lesions (10). In the second chapter of my dissertation I exploit the \( E. \ coli \) adaptive response and use two different complementation models to determine the levels of \( O^6 \)-mG repair for all reported non-synonymous MGMT variants and polymorphisms as well as investigate the toxicity of \( O^6 \)-meG lesions to prokaryotes.

\( N^3 \)-meA DNA glycosylases are the first step in a multi-step DNA repair process known as base excision repair (BER). After a glycosylase removes an alkylated base from the phosphate backbone of DNA by cleaving the glycosylic bonds, downstream enzymes that cleave that sugar phosphate backbone (endo/exo nuclease), replace the damaged base (DNA polymerase), and finally ligate the DNA
sugar phosphate backbone (DNA ligase) complete the repair process (Fig. 2C) (34,43). BER is the principal mechanism by which methyl adducts are removed from the genome but it is a delicate DNA repair pathway that requires the coordinated effort of several enzymes working in succession. If any enzyme becomes limiting or overexpressed then the resulting repair intermediate can become even more toxic than the initial lesion itself (19). N³-meA glycosylases can be divided into two categories: those that are constitutively expressed and those that are inducible. These two categories differ in their catalytic efficiency (low for constitutive, high for inducible) and in their substrate specificity (narrow for constitutive and broad for inducible).

The budding yeast *Saccharomyces pombe* contains a constitutive 3MeA glycosylase known as spMAG1 (33), which plays a relatively minor role in the alkylation resistance profile of *S. pombe* (2,26,35). By contrast, its *S. cerevisiae* homolog (scMAG) is inducible and repairs a broader set of alkylation lesions than spMAG1, much akin to *E. coli*’s alkA (1). In the third chapter of my dissertation I again exploit the adaptive response pathway of *E. coli* and use a heterologous complementation system to explore the physiological benefits and constraints of expressing constitutively active verses inducible glycosylases and give evolutionary evidence of why each type of glycosylase possibly evolved.
Fig. 2. Responses to DNA methylating agents. A. The *E. coli* adaptive response operon. B. Alkyltransferase removal of O^6^-meG lesions. C. The multi-step repair pathway of base excision repair.
DNA Replication

Plasmids with a ColE1 origin of replication are used extensively in molecular biology and are the basis for most bacterial recombinant expression vectors. During ColE1 replication initiation, RNA polymerase generates an RNA primer known as RNA II about 600 nucleotides upstream of the ColE1 origin site (23). After RNA II has been synthesized, DNA polymerase I (Pol I) initiates leading strand synthesis (22). *E. coli* contains two replicative polymerases, Pol I and DNA polymerase III (Pol III). Pol I functions independent of additional proteins and its role in ColE1 plasmid replication is confined to leading strand initiation and Okazaki fragment processing (20,47), while Pol III is the major replicative polymerase of *E. coli*. After Pol I initiates leading strand synthesis, the transcript is passed on to replisomal associated Pol III, which performs coupled high-speed replication of the two strands (Fig. 3) (30,32). When Pol III reaches the end of an Okazaki fragment on the lagging strand, Pol I fills in the remaining gap.

A number of questions regarding the role of Pol I during ColE1 plasmid replication *in vivo* remain unsolved. Specifically it is unclear how long the initial leading strand extension product produced by Pol I is, how sharp the transition from Pol I to Pol III is after the initial extension, and how extensive Pol I processing of Okazaki fragments is *in vivo*. Previously my advisor, Dr. Manel Camps, generated a low fidelity version of Pol I by using site directed mutagenesis to disrupt the proofreading domain and widen the active site pocket of the Pol I enzyme to readily allow misincorporations (6). Expression of this variant of Pol I in JS200 *E. coli* cells
(a Pol I temperature sensitive strain) at 37°C leads to the introduction of random mutations by the low fidelity Pol I during ColE1 plasmid replication. We have used extensive sequencing data to determine the spectrum of mutations and the frequency of each mutation introduced by the low fidelity Pol I variant. In the fourth chapter of my thesis I utilize the mutational footprint left by the low fidelity Pol I complemented into the JS200 strain of *E. coli* in order to resolve the questions regarding the extent to which Pol I participates in ColE1 plasmid replication.
Fig. 3. Illustration of leading strand synthesis of a ColE1 plasmid. First RNA polymerase synthesizes the RNA II primer then Pol I performs initiates leading strand synthesis. After Pol I initiation, leading strand synthesis is passed on to repisomal associated Pol III.
Objectives: Hypotheses and Specific Aims

Chapters 2-4 of my dissertation will address the following hypotheses and specific aims:

Chapter 2: Characterization of O\textsuperscript{6}-meG repair by naturally occurring MGMT mutants provides insight into MGMT structure and function

Hypothesis: Non-synonymous polymorphisms and variants of MGMT affect the ability of MGMT to repair O\textsuperscript{6}-meG lesions

Specific Aims:

1. To determine if the naturally occurring MGMT mutants vary in their capability to repair O\textsuperscript{6}-meG lesions
2. To determine the relative toxicity of O\textsuperscript{6}-meG lesions versus N\textsuperscript{3}-meA lesions to \textit{E. coli} during exposure to S\textsubscript{N}1 alkylating agents

Chapter 3: Experimental demonstration of a tradeoff between toxicity and repair capacity that likely drove the evolution of 3MeA glycosylase inducible expression

Hypothesis: Inducible 3MeA glycosylases arose as a way to bypass the negative effects of constitutive glycosylase activity on organismal fitness.
Specific Aims:

1. To determine the fitness effects of heterologous expression of a constitutively active versus inducible glycosylase in *E. coli*
2. To use directed evolution to select mutant glycosylases with increased fitness
3. To show that there is a fitness tradeoff between glycosylase activity and cellular viability

Chapter 4: Mutagenic footprint of low-fidelity Pol I replication in *E. coli* reveals an extensive interplay between Pol I and Pol III during ColE1 plasmid replication

Hypothesis: DNA Pol I plays a more extensive role in ColE1 plasmid replication than previously thought

Specific Aims:

1. To determine how long the initial transcript synthesized by Pol I is during leading strand initiation
2. To determine how sharp the transition from Pol I to Pol III is after the initial extension
3. To determine how extensive Pol I processing of Okazaki fragments is *in vivo*
References


Chapter 2:

Characterization of O\textsuperscript{6}-meG repair by naturally-occurring MGMT mutants provides insight into MGMT structure and function
Abstract

The cytotoxic O$^6$-methyl guanine (O$^6$-meG) lesion is generated by strong alkyl donors, such as methylationg chemotherapeutic agents. Methyl-guanine methyl transferase (MGMT) repairs O$^6$-meG lesions by direct transfer of the lesion to a nucleophilic cysteine residue located within the active site of the protein. Therefore its activity has a strong impact on tumor sensitivity to strong alkylating agents. To date, two non-synonymous polymorphisms and six non-synonymous variants of MGMT have been reported in human populations, four of which have not been previously characterized. In this study, we determined the levels of O$^6$-meG repair for all reported non-synonymous MGMT variants and polymorphisms by measuring the ability of E. coli cells expressing the MGMT mutants to grow on gradients of the strong methylating agent N-methyl-N$'$/nitro-N-nitrosoguanidine (MNNG). As hosts, we used two E. coli complementation models differing in their ability to repair additional cytotoxic lesions: an ada deletion strain, deficient in both transactivation of the adaptive response pathway and alkyl transfer abilities, and an ada point mutant strain specifically deficient in alkyl transfer ability. Most of the MGMT mutants we tested retained near wild-type levels of O$^6$-meG repair in both strains. This included R128Q, a non-synonymous variant located at the arginine finger residue of MGMT. By contrast, the W65C variant had severely crippled O$^6$-meG repair ability and the E30K and L84F mutants, both located adjacent to Zn-coordinating residues, exhibited moderate decreases in levels of O$^6$-meG repair. Our results also highlight the critical
importance of alkyltransferases in mediating the lethal effects of strong alkylating agents in *E. coli*. This observation calls for further investigation into the mechanism underlying O\(^6\)-meG cytotoxicity in prokaryotes.
Introduction

Exogenous chemicals react with DNA causing cytotoxic and mutagenic DNA lesions. One of the most deleterious lesions formed by these agents occurs at the O^6 position on guanine nucleotides (O^6-meG)(16,23,35). This lesion is primarily mutagenic, due to its ability to mismatch with adenine, leading to G to T mutations. In mammalian cells, however, O^6-meG lesions cause significant cytotoxicity through the initiation of a futile mismatch repair cycles that eventually leads to cell cycle arrest, chromosomal aberrations and apoptosis (25,27).

Methyl-guanine methyl transferase (MGMT) is the sole O^6-meG direct repair protein found in humans. MGMT functions by transferring the deleterious alkyl group from the oxygen molecule on the guanine nucleotide to a nucleophilic cysteine residue located within the active site of the protein (reviewed in (36,37)). Repair of O^6-meG lesions occurs in a 1:1 stoichiometric ratio. Therefore MGMT protection from alkylation treatment is proportional to the levels of active protein itself (1,10,17,40). Mutations that affect O^6-meG repair by altering the active site or the thermo-stability of the protein have been described (6,7,11,26).

To date, two polymorphisms and six variants (non-synonymous mutations with a prevalence of less than 1%) of MGMT have been identified in the human population (Table 1) (Fig. 1) (5). At least one of these variants, W65C (rs2282164), is less stable than wild-type MGMT at 37°C and MGMT-deficient cells expressing this variant exhibit increased sensitivity to O^6-meG adducts generated by exposure to MNNG (N-methyl-N'-nitro-N-nitrosoguanidine) (15,21). It is well established that
alterations in the levels of MGMT expression and activity have an impact on resistance to chemotherapeutic alkylating agents such as BCNU and temozolomide (3,4,20,24). Also, although the evidence as to the role that MGMT polymorphisms may play in cancer risk is conflicting, it is clear that sporadic cancer incidences can be influenced by polymorphisms in DNA repair genes (2,18).

Here, we use *E. coli* complementation models to test the O\(^6\)-meG repair proficiency of ten non-synonymous MGMT mutants detected in the human population. This panel includes four variants that to our knowledge have not been previously characterized. The *E. coli* complementation systems for MGMT are based on the inactivation of Ada, the *E. coli* homolog to MGMT. In addition to an alkyltransferase domain that is functionally homologous to MGMT, *E. coli* Ada also contains a transactivation domain that is responsible for upregulating additional proteins involved in the repair of reactive DNA methylation (reviewed in (41)). The GW7101 strain, which is the strain of reference for MGMT complementation, contains a fully deleted *ada* and therefore is deficient in both alkyltransferase and transactivation activities. In an effort to increase the accuracy of our complementation assays, here we utilized an additional strain of *E. coli*. This strain (PJ5), originally isolated by Penny Jeggo *et al.* (22), is deficient in O\(^6\)-meG repair while retaining transactivation activity (30,38,44).

The results of our mutant MGMT characterization are consistent between the two complementation systems and also with previous reports. We found that all mutants tested exhibit near wild-type levels of activity except for the severely
crippled W65C variant and the slightly crippled L84F and E30K mutants. The latter two mutants are immediately adjacent to Zn-coordinating positions, consistent with biochemical studies showing a role of Zn in modulating alkyl-transfer. The R128Q variant, which has never been characterized before, shows robust O⁶-meG repair, despite the fact that a glutamine has replaced the arginine finger residue.

The observed concordance between the two *E. coli* complementation systems despite their differences in Ada functionality as well small difference in MNNG sensitivity is remarkable. These results imply that O⁶-meG may be more cytotoxic to *E. coli* than previously thought and argues against a primarily mutagenic effect of O⁶-meG in prokaryotes. However, the fact that the PJ5 strain is noticeably more resistant to MNNG than the GW7101 strain, confirms that other cytotoxic lesions are also significant players in the lethality of strong (S_N1) alkylating agents to prokaryotes.

**Materials & Methods**

*Bacterial strains*

A summary of the bacterial strains used in this study is provided in table 2. The PJ1, PJ3, PJ5, PJ6 and BS21 were provided to us courtesy of Dr. Geoff Margison, University of Manchester, UK. The PJ1, PJ3, PJ5, PJ6 and BS21 strains were sequenced by amplifying up the *ada* gene using the primers F: 

AAAACTCGAGATGAAAAAAGCCACATGC and R:
AAAAAGCTTTTAGTGGTGATGGTGATGATGCCTCTCCTCATTTTCAGC and then sequenced by Sequetech Corporation (Mountain View, CA).

Plasmid constructs

Human MGMT cDNA (RefSeq NM_002412) was cloned into the pLitmus 28i vector (N.E.B.) using the XhoI and HindIII cloning sites and primers F:
AAAACTCGAGATGGACAAGGATTGTGAA and R:
AAAAAGCTTTCTAGTGGTGATGGTGATGATGGTTTCGGCCAGGCGG. Then site directed mutagenesis and megapriming cloning technique was used to generate the mutant MGMT variants (31). The MGMT mutants tested in this study are E30K (rs2020893) P58S (rs2308322) W65C (rs2282164) L84F (rs12917) R128Q (rs3750824) I143V (rs2308321) G160R (rs2308318) E166D (rs2308320) and K178R (rs2308327).

MNNG gradients

pLitmus 28i vectors containing MGMT variants were transformed into competent GW7101 and PJ5 cells and grown up in the presence of 100µg/ml carbenicillin. Monocultures of all the E. coli strains listed above containing and excluding MGMT variants were grown up overnight to saturation and then the following morning diluted down and allowed to grow up to exponential phase. Once reaching
...exponential phase 50\(\mu\)l of culture mixed with 2ml of top agar were stamped onto a MNNG containing gradient using the thin side of a sterile glass microscope slide. For a reference on how gradients are poured see (45) but in brief square petri dishes (Fisher Scientific) are placed on an slope of 10° while a mixture of MNNG plus 25ml LB agar is poured onto the plate. Once dried the petri dish is set flat and a top layer of 25ml LB agar without MNNG is poured on top. 1M MNNG stock solution is prepared by suspending MNNG powder into 100% DMSO. For reference a 0.02% MNNG gradient is made by mixing 5\(\mu\)l of 1M MNNG with 25ml LB agar. Given the labiality of MNNG, plates need to be made fresh before the experiment and needed to be allowed to cool off individually (no stacking). Gradients are gown at 37°C for 30 hours. Gradient growth is measured in centimeters and the highest point of growth is marked at the end of continuous culture growth and the beginning of growth is marked where the culture starts growing. Length of growth on the gradient is proportional of protection against MNNG toxicity. To confirm all cultures were at similar levels of viability, cultures were plated at a 10\(^{6}\) dilution on petri dishes containing carbenicillin and colonies were counted. All experiments were done in triplicate and the error bars represent the standard deviation.

Statistics

Data are expressed as mean+/- standard deviation (SD). Data were analyzed using Dunnett’s test to test specific hypotheses comparing the performance of various

27
mutants versus an appropriate control, using JMP software (Version 10.0, 2012, SAS Institute). A P-value of <0.05 was considered statistically significant.
Fig. 1. Location of non-synonymous MGMT SNP and variants. Crystal structure of MGMT showing the location of the non-synonymous MGMT mutants as well as other important residues/domains. Figure was constructed off of the the PDB submission by (7) and modeled using PyMOL.
<table>
<thead>
<tr>
<th>MGMT Variants</th>
<th>dbSNP</th>
<th>Prevalence</th>
<th>Previously tested in E. coli models</th>
</tr>
</thead>
<tbody>
<tr>
<td>E30K</td>
<td>rs2020893</td>
<td>rare</td>
<td>No</td>
</tr>
<tr>
<td>P58S</td>
<td>rs2308322</td>
<td>rare</td>
<td>No</td>
</tr>
<tr>
<td>W65C</td>
<td>rs2282164</td>
<td>rare</td>
<td>Yes</td>
</tr>
<tr>
<td>L84F</td>
<td>rs12917</td>
<td>common</td>
<td>Yes</td>
</tr>
<tr>
<td>R128Q</td>
<td>rs3750824</td>
<td>rare</td>
<td>No</td>
</tr>
<tr>
<td>I143V K178R</td>
<td>rs2308321</td>
<td>common</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>rs2308327</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G160R</td>
<td>rs2308318</td>
<td>rare</td>
<td>No</td>
</tr>
<tr>
<td>E166D</td>
<td>rs2308320</td>
<td>rare</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 1. MGMT SNP and their prevalence. Rare means <1% population prevalence. Common means 20-25%. The R128Q variant comes from (19). I143V and K178R are genetically linked.
Table 2. Summary of work performed on the ada strains and genotyping. * Indicates data inferred by our experiments. We report PJ6 confers a slightly higher resistance to MNNG than originally reported in (22).
<table>
<thead>
<tr>
<th>Cell Line</th>
<th>Genotype</th>
<th>ada sequence</th>
<th>Ada protein levels (29)</th>
<th>Transactivation activity (27)</th>
<th>O6mG transferase activity (28)</th>
<th>MNNG resistance</th>
<th>Overall assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB1157</td>
<td>WT</td>
<td>WT</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>Wild-type E. coli</td>
</tr>
<tr>
<td>BJ-1</td>
<td>alkB-</td>
<td>WT</td>
<td>WT*</td>
<td>+++*</td>
<td>+++*</td>
<td>+++</td>
<td>alkB-</td>
</tr>
<tr>
<td>GW7101</td>
<td>ada- alkB-</td>
<td>Δ</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>+</td>
<td>ada-, deficient in both transactivation and alkyl transfer</td>
</tr>
<tr>
<td>MV1932</td>
<td>ada- alkA-</td>
<td>Δ</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>-</td>
<td>Similar to GW7107 but tag1-</td>
</tr>
<tr>
<td>PJ-1</td>
<td>ada</td>
<td>W335stop</td>
<td>minimal</td>
<td>Unknown</td>
<td>-</td>
<td>+ (26)</td>
<td>Very low ada expression</td>
</tr>
<tr>
<td>PJ-3</td>
<td>ada</td>
<td>A310V</td>
<td>WT</td>
<td>+++</td>
<td>+++</td>
<td>+++ (26)</td>
<td>Low alkyl transfer activity</td>
</tr>
<tr>
<td>PJ-5</td>
<td>ada</td>
<td>R322W</td>
<td>WT</td>
<td>+++</td>
<td>-</td>
<td>+ (26)</td>
<td>Deficient in alkyl transfer activity</td>
</tr>
<tr>
<td>PJ-6</td>
<td>ada</td>
<td>R19H</td>
<td>minimal</td>
<td>+</td>
<td>-</td>
<td>+++</td>
<td>Decreased ada expression</td>
</tr>
<tr>
<td>BS21</td>
<td>ada</td>
<td>E74D, A78P, Q79N, E96K, M125I, A219D (26)</td>
<td>Unknown</td>
<td>Unknown</td>
<td>++++</td>
<td>++++</td>
<td>Constitutively expressed ada</td>
</tr>
</tbody>
</table>
Results

Identification of an ada mutant strain specifically deficient in alkyl-transfer

When exposed to a methylating agent such as MNNG or methylmethane sulfonate (MMS), Ada upregulates its own expression and that of two other proteins involved in the repair of specific methyl lesions, AlkB and AlkA (41). This coordinated response to reactive DNA alkylation is known as the adaptive response pathway and appears to be restricted to prokaryotes. The AlkB protein repairs N1-meA and N3-meC lesions by direct repair. These lesions are exclusively produced by SN2 (weak) alkylating agents, represented in this study by MMS (9,28). The AlkA protein is a glycosylase responsible for removing, among other substrates, the highly cytotoxic N3-meA lesion, which is generated by both SN2 and SN1 alkylating agents (MMS and MNNG, respectively) (14,29).

Several *E. coli* ada mutants have been isolated and characterized over the years with respect to their Ada protein expression levels, ability to induce transactivation, ability to transfer O6-meG lesions to their active site, and ability to confer protection against MNNG-induced toxicity and mutagenicity (22,30,38,43,44). Table 2 shows a summary of these previous findings. This table also lists the ada sequence for the PJ strains and for the BS21 strain, a strain expressing a constitutively-active ada protein (22). We found at least one non-synonymous
MGMT mutation for each of the PJ strains, which is consistent with the reported ada phenotypes for these strains.

In order to determine the PJ strain best suited for our study (i.e. the strain with a selective defect in alkyl transferase activity) we determined the resistance levels of all the PJ strains to MNNG and to MMS in vivo. Strains selectively deficient in alkyl-transferase activity would be expected to have wild-type levels of MMS resistance and a dramatic sensitization to MNNG. The results are shown in Fig. 1. The PJ1 strain has a substantial defect in transactivation as well as alkyltransferase activity, consistent with an earlier report indicating this strain has very low levels of ada expression (44). The PJ3 and PJ5 strains both have substantial defects in alkyltransferase activity consistent with the presence of point mutations in their alkyltransferase domains (Table 2). The PJ6 strain has an alkyltransferase defect as well, possibly due to low levels of protein expression (30). Of these strains, the PJ5 strain stands out for its low resistance to MNNG, its high resistance to MMS, and its normal levels of protein expression. Therefore we selected this strain as our alternative, transactivation-competent complementation strain.
Fig. 2. Inability to alkyl transfer O\textsuperscript{6}-meG lesions leads to MNNG sensitivity while deficient transactivation leads to MMS sensitivity. Gradient measurements show the sensitivity of all *E. coli* strains used in this study to both MMS and MNNG induced lesions. Results are reported as centimeters of growth on a 0.02% gradient of the given drug, with 5 cm representing the top of the gradient.
Characterizing the MGMT non-synonymous polymorphisms and variants

We wanted to determine whether the known non-synonymous SNPs and variants of MGMT have an effect on O\(^6\)-meG repair. Of the ten mutants tested in this study, I143V/K178R (which are genetically linked), L84F, and W65C have been previously tested in *E. coli* models (21,32) and G160R has been previously tested biochemically (12). The uncharacterized MGMT variants included in our study are: E30K, P58S, R128Q, and E166D. For comparison with previous complementation studies, which complemented ada- strains (21,32), we used GW7101 cells. The results, presented in Fig. 2a, show a dramatic decrease in MNNG protection for W65C (consistent with previous reports), and much more moderate decreases for E30K, L84F, and E166D. All remaining characterized and untested mutants show close to wild-type levels of resistance.

As mentioned above, ada- strains have lost both transactivation and alkyl-transfer activities. We reasoned that the absence of transactivation activity might affect the ada complementation system by allowing the build-up of other cytotoxic lesions (particularly N\(^3\)-meA), making these lesions limiting for survival at higher concentrations of MNNG. Therefore we tested the same panel of MGMT mutants in our PJ5 strain, which retains transactivation activity. PJ5 cells are overall moderately more resistant to MNNG, consistent with their ability to repair other cytotoxic lesions induced by MNNG exposure. In these cells, the results of our mutant MGMT panel characterization (shown in Fig. 2b) were very similar to those obtained using the ada
deletion strain. The only exception was the E166D variant, which exhibited higher viability in the PJ5 strain.
Fig. 3. Protection profile of MGMT mutants to MNNG. (A) Growth of MGMT mutants on a gradient of 0.02% MNNG complemented into the GW7101 strain of *E. coli*, expressed as centimeters. (B) Growth of MGMT mutants on a gradient of 0.05% MNNG complemented into the PJ5 strain of *E. coli*. Results are reported as centimeters of growth, with 8 cm representing the top of the gradient. Error bars represent the standard deviation of three separate experiments. * indicates mutants with significantly different growth on gradients compared to MGMT wild-type using a Dunnett’s comparison of means test.
A. Growth on 0.02% MNNG Gradient (cm)

B. Growth on 0.05% MNNG Gradient (cm)
Discussion

*ada mutant strains*

The main lesions induced by exposure to MNNG are O\(^6\)-meG and N\(^3\)-meA, which are repaired by Ada and AlkA, respectively (41,42). We wondered whether the lack of AlkA in an *ada* deleted strain would lead to an underestimation of the amount of resistance conferred by MGMT in complementation experiments. To that end we screened four previously reported *ada* strains (PJ1, 3, 5, 6) for MMS and MNNG resistance. As controls we used *alkB*- cells (BJ-1), an *ada* deleted strain (GW7101) and a constitutive *ada* expresser (BS21). Table 2 reports the *ada* sequences for the four PJ strains and for BS21. A stop codon, present in the C-terminus of PJ1, is consistent with decreased levels of transactivation activity originally reported by Lemotte and Walker (30). Similarly, point mutations in the alkyltransferase domain of *ada* for PJ3 and PJ5 (A310V and R322W, respectively) are consistent with a specific defect in alkyl transfer, although the MNNG resistance profiles suggest that PJ3 retains some capacity for O\(^6\)-meG alkyl transfer (Fig.1).

The PJ5 strain turned out to be our best candidate for complementation. PJ5 cells retain substantial resistance to MMS (consistent with an ability to transactivate *alkB*) and a profound sensitization to MNNG (consistent with a deficiency in alkyl transfer activity). Our results are in-line with previous reports showing a strong decrease in MNNG resistance and in O\(^6\)-meG demethylation capabilities for the PJ5
strain relative to AB1157 (wild-type *E. coli* strain) (30,38,44). We hypothesized that the reason for the impaired alkyltransferase capability of the PJ5 strain is due to a bulky tryptophan substitution at position R323, which is located in close proximity to the nucleophilic cysteine residue at position C321. We confirmed this hypothesis by showing that PJ5 cells expressing an R323W mutant of MGMT exhibit levels of MNNG resistance that are comparable to those of the empty vector control (Fig. 2b).

*Role of O\(^6\)-meG repair versus induction of the adaptive response for MNNG protection*

PJ5 cells are moderately more resistant to MNNG than GW7101 cells (Figs. 1 and 2). This observation is consistent with the role of the adaptive response in protecting *E. coli* from the deleterious effects on alkylation, possibly through induction of AlkA and its ability to repair N\(^3\)-meA lesions. However, our results also indicate that MGMT complementation, which almost exclusively repairs O\(^6\)-meG lesions, substantially alleviates MNNG toxicity to *E. coli* cells. Further, our results are remarkably consistent between the two complementation models we utilized (Fig. 2). This concordance between complementation systems and the extent to which MGMT complementation protects from MNNG toxicity highlights the prominent role of O\(^6\)-meG lesions in mediating MNNG cytotoxicity in *E. coli*. These observations suggest that O\(^6\)-meG lesions, in addition to mutagenic, should be considered primarily
cytotoxic in prokaryotic cells. The mechanism of O\textsuperscript{6}-meG cytotoxicity, which is unclear in prokaryotes, should be investigated.

\textit{O\textsuperscript{6}-meG repair profiling of MGMT SNPs and variants}

In this study we tested several previously tested non-synonymous polymorphisms and variants of MGMT for their ability to repair O\textsuperscript{6}-meG lesions. The levels of protection against MNNG toxicity conferred by these MGMT mutants are consistent with previous bacterial, biochemical, and mammalian studies. The common non-synonymous polymorphisms of MGMT (L84F, I143V K178R), present in about 20-25\% of the population, show no to minimal deficiency in O\textsuperscript{6}-meG repair capability (21,32). The more rare variants of MGMT, present in <1\% of the population behave as previously reported as well, with G160R exhibiting robust O\textsuperscript{6}-meG repair despite its close proximity to the active site of MGMT, and W65C exhibiting poor O\textsuperscript{6}-meG repair, most likely due to poor folding stability (12,21,33).

In this study we also tested four previously uncharacterized non-synonymous variants of MGMT. P58S and E166D both confer near wild-type levels of protection against MNNG toxicity, suggesting that these variants probably exhibit no deficiency in O\textsuperscript{6}-meG repair. Both of these mutations occur within \(\alpha\)-helices located on the backside of the protein in areas that do not bind DNA or coordinate zinc residues (8). Although located in the middle of a \(\alpha\)-helix, the E166D variant represents a very conservative change and therefore is unlikely to induce substantial structural changes.
The P58S mutation is not a conservative amino acid substitution but its position (at the top of a α-helix) suggests it may delay the onset of the first turn of the helix without disturbing the helix itself.

The L84F polymorphism and E30K variant of MGMT consistently show a slight decrease in MNNG protection in our assay. Both positions are located adjacent to zinc-coordinating residues, H29 in the case of E30, and H85 in the case of L84. Although the structural role of the zinc ion is unclear (7,46) past studies have shown that the zinc ion increases the reaction rate of MGMT by ~60 fold and lowers the pKa of the active-site cysteine residue (39,46). It is possible that substitutions adjacent to zinc coordinating residues affects zinc binding just enough to slightly lower the reaction rate of these MGMT mutants. This hypothesis would need to be confirmed biochemically.

Perhaps our most striking finding is that the R128Q variant of MGMT confers near wild-type levels of protection in our MNNG assay. The R128 residue of MGMT has been identified as the arginine finger residue that is responsible for stabilizing the Watson-crick base-pairing of the DNA helix once the damaged guanine residue has been flipped into the active site (7,8). Different amino acid substitutions at the R128 position have been reported to either abolish or seriously inhibit the ability of MGMT to repair methylated DNA (7,26), although the effect of a glutamine substitution has not been previously reported. While this result was unexpected, base-flipping enzymes that use a glutamine instead of an arginine finger to stabilize base stacking are not unheard of. Examples include the glycosylase MutY and the endonuclease
EndoIII (13). We conclude that MGMT is able to tolerate both arginine and glutamine residues at this position.

Conclusion

O\textsubscript{6}-meG lesions are of clinical importance, representing one of the main lesions mediating the cytotoxicity of chemotherapeutic drugs such as temozolomide and BCNU (16). We have shown here that natural, non-synonymous variants of MGMT can affect the repair capability of O\textsubscript{6}-meG lesions, although most variants seem to retain near wild-type levels of O\textsubscript{6}-meG repair. However, the activity of the MGMT protein represents only one of many factors contributing to the overall sensitivity of cells to O\textsubscript{6}-meG-inducing agents (12,34). The promoter region of MGMT can be silenced in tumors via epigenetic mechanisms and it is known that a few MGMT mutants, specifically the I143V K178 polymorphism and the G160R variant of MGMT, are less resistant to inhibitors than wild-type MGMT (12,34). Therefore these other factors need to be considered when weighing the options of treating tumors with alkylating agents. As we move closer towards genomics-based personalized medicine and as the mechanisms linking MGMT function with cancer risk become clearer the need to further explore the functional impact of MGMT variants and polymorphisms will increase.
Acknowledgements

This work was supported by the Ann Irwin grant to C.J.T and from the K08 award CA116429 from the NCI to M.C. We thank Dr. Geoff Margison for generously providing us with the PJ1,3,5,6 and BS21 strains and for advice in the early stages of this project.
References


Chapter 3:

Experimental demonstration of a tradeoff between toxicity and repair capacity that likely drove the evolution of 3MeA glycosylase inducible expression.
Abstract

Base excision repair (BER) is a universal mechanism of DNA repair that maintains genomic integrity by replacing damaged nucleotides. The first step in BER involves breaking the glycosidic bond of damaged bases by DNA glycosylases. 3MeA DNA glycosylases specialize in recognition and removal of alkylated nucleotides and can be constitutively-expressed or inducible. Compared to inducible 3MeA glycosylases, constitutively-expressed glycosylases repair a narrower subset of lesions and have lower affinity for their substrates. Here we compare the effect of expressing two close yeast 3MeA glycosylase homologues (S. pombe MAG1 and S. cerevisiae MAG) in a 3MeA glycosylase-deficient strain of E. coli. We find that overexpression of the normally inducible 3MeA glycosylase, scMAG, has a profound deleterious effect on cellular fitness, while overexpression of the normally constitutively expressed glycosylase, spMAG1, has a more moderate effect. We link the deleterious effect of heterologous 3MeA glycosylase expression to levels of enzymatic activity by showing that spMAG1 mutants selected for improved fitness exhibit decreased 3MeA repair and that the fitness of the host E. coli strain inversely correlates to levels of wild-type spMAG1 expression. Our work indicates that constitutive 3MeA glycosylase activity is minimized so as not to have a negative effect on the fitness of the organism. These findings suggest that inducible 3MeA glycosylases arose as a way to bypass the fitness constraints imposed on constitutive glycosylases. Finally, our work points to 3MeA as the critical endogenous target for 3MeA glycosylases,
explaining the narrower substrate specificity of constitutive 3MeA glycosylases.
Introduction

DNA glycosylases cleave glycosylic bonds of damaged nucleobases, removing them from the phosphate backbone of DNA. The resulting abasic site is processed by enzymes that cleave the sugar phosphate backbone (endonucleases), process the resulting free ends (lyases, phosphatases), fill in the resulting gap (DNA polymerases), and finally ligate the DNA sugar phosphate backbone (DNA ligases). Collectively, this form of DNA repair is known as base excision repair (BER).

N3-methyladenine glycosylases (3MeA glycosylases) specialize in repair of alkylation damage. *E. coli* contains two 3MeA glycosylases in its genome: a constitutive-expressed 3MeA glycosylase (tagA) and an inducible 3MeA glycosylase (alkA) (8,16). The TAG glycosylase acts almost exclusively on 3MeA lesions (5,23) while the AlkA glycosylase is known to have a surprisingly broad substrate range and can act on a multitude of endogenously and exogenously produced alkyl lesions in addition to 3MeA (4,27,28). *alkA* mutants are very sensitive to alkylating agents, whereas *tagA* mutants have only moderately increased sensitivity (3,9,19).

The budding yeast *Saccharomyces pombe* contains a constitutive 3MeA glycosylase known as spMAG1 (21), which plays a relatively minor role in the alkylation resistance profile of *S. pombe* (2,15,22). By contrast, its *S. cerevisiae* homolog (scMAG) is inducible and repairs a broader set of alkylation lesions than spMAG1, much akin to *E. coli*’s alkA (1). Indeed, deletion of scMAG1 has a profound effect on the alkylation resistance of *S. cerevisiae* (1,7). Thus,
constitutively-expressed 3MeA glycosylases tend to have narrower substrate specificity and less overall impact on alkylation resistance than inducible ones.

BER is a delicate DNA repair pathway that requires the coordinated action of several enzymes working in succession. Some of the intermediates in the BER repair pathway (notably blocked nicks) are highly cytotoxic. Minimizing the formation of these intermediates requires a high level of control at the expression and the protein interaction levels. Buildup of repair intermediates as a result of imbalanced BER gene expression can become more toxic than the initial lesion (10). Thus, alterations in 3MeA glycosylase expression or activity can be deleterious, particularly in the absence of alkylation challenge. scMAG overexpression for example results in a strong mutator phenotype and compromised viability in both *S. cerevisiae* and *E. coli* strains (11,25). The observed increase in spontaneous mutagenesis is most likely attributed to an abundant production of abasic sites by the glycosylase, as well as by the generation of blocked 5’- and 3’- ends further downstream. Abasic sites are highly mutagenic through the A- rule (14), and blocked ends cause replication fork collapse, which is mutagenic through the intervention of bypass polymerases (13,31). The observed decrease in viability on the other hand is most likely attributed to DNA nicks that are generated by endonuclease processing of abasic sites, as these can easily become DNA strand breaks, which are highly cytotoxic (12).

Here we compare the effect of expressing two close yeast 3MeA glycosylase homologues in an *ada tagA* strain of *E. coli*. We find that overexpression of the inducible 3MeA glycosylase, scMAG, has a profound deleterious effect on cellular
viability while overexpression of the constitutively expressed glycosylase, spMAG1, has a more moderate effect. We link the deleterious effect of heterologous 3MeA glycosylase expression to levels of enzymatic activity by showing that (1) spMAG1 mutants selected for improved fitness also exhibit decreased 3MeA repair and (2) the fitness of the host *E. coli* strain inversely correlates to levels of WT (wild-type) spMAG1 expression. Our work indicates that constitutive 3MeA glycosylase activity needs to be minimized so as not to have a negative effect on the fitness of the organism. These findings suggest that inducible 3MeA glycosylases likely arose as a way to bypass the negative effects of constitutive glycosylase activity on organismal fitness. Our work also points to 3MeA as the critical endogenous target for 3MeA glycosylase repair, explaining the narrower substrate specificity of constitutive 3MeA glycosylases.

**Methods**

*Plasmids and bacterial strains*

All constructs were cloned into the multi-cloning site of the pLitmus 28i vector (N.E.B.) using the XhoI and HindIII restriction enzymes. scMAG1 (RefSeq NM 001179032.1) was cloned into pLitmus 28i using the primers F: AAAAAAACTCGAGATGGGTTCTTCTCAC R: AAAAAAAAGCTTTTAGGATTTTACGAA and spMAG1 (RefSeq NM
001019417.2) was cloned into pLitmus 28i using the primers F:

AAAAAACTCGAGATGGGTTCTTCTCAC R:

AAAAAAAAGCTTTTCAGTTTCTTCTCGG. cDNA of both constructs were generously provided to us by the Eichman laboratory, Vanderbilt University.

All constructs utilized in this paper were transformed into the MV1932 E. coli strain, an alkylation sensitive alkA tag derivative of AB1157 (Fthr-1leu B6 hisG4 proA2 thi-1 argB3 lacY1 galK ara-14 xyl-5 mly-1tsx-33 rpsl31 sup-37) (35). The AB1157 strain itself was used as an alkylation resistant control.

MNNG gradients

pLitmus 28i vectors containing the MAG glycosylases and mutants were transformed into competent MV1932 cells and grown up in the presence of 100µg/ml carbenicillin. Monocultures of these constructs were grown up overnight to saturation and then the following morning diluted down and allowed to grow up to exponential phase. Once reaching exponential phase 50µl of culture mixed with 2ml of top agar were stamped onto a MNNG (N-methyl-N'-nitro-N-nitrosoguanidine) containing gradient using the thin side of a sterile glass microscope slide. For a reference on how gradients are poured see (34) but in brief square petri dishes (Fisher Scientific) are placed on an slope of 10° while a mixture of MNNG plus 25ml LB agar is poured onto the plate. Once dried the petri dish is set flat and a top layer of 25ml LB agar without MNNG is poured on top. 1M MNNG stock solution is prepared by suspending MNNG powder
into 100% DMSO. For reference a 0.02% MNNG gradient is made by mixing 5µl of 1M MNNG with 25ml LB agar. Given the labiality of MNNG, plates need to be made fresh before the experiment and needed to be allowed to cool off individually (no stacking). Gradients were gown at 37°C for 30 hours. Gradient growth is measured in centimeters and the highest point of growth is marked at the end of continuous culture growth and the beginning of growth is marked where the culture starts growing. Length of growth on the gradient is proportional of protection against MNNG toxicity. To confirm all cultures were at similar levels of viability, cultures were plated at a 10⁻⁶ dilution on petri dishes containing carbenicillin and colonies were counted. All experiments were done in triplicate and the error bars represent the standard deviation. MMS (Methyl methanesulfonate) gradients were run in much the same way as MNNG gradients except that the stock solution came as 99% pure liquid (Aldrich).

**Viability assay**

Cultures of MV1932 cells containing the appropriate constructs were grown up to saturation overnight. The next morning cultures were diluted down to an optical density (OD) of 0.15. Cultures were then allowed to grow up for an additional 18 hours, to saturation. Growing cultures to saturation twice negated the big differences observed in the initial inoculation. Cultures were then serial diluted to an appropriate cellular density and then platted on petri dishes containing 100µg/ml carbenicillin.
All experiments were performed in at least triplicate and the error bars represent the standard deviation.

*Genetic library generation and selection*

cDNA of spMAG1 was subjected to random mutagenesis using the GeneMorph II Random Mutagenesis Kit (Agilent Technologies). In brief, we used 100ng of target DNA and subjecting the reaction to 25 cycles of PCR using the suggested PCR program outlined in the GeneMorph II manual with a $T_m$ of 55°C. We used exact 18mer primers of the 5’ and 3’ end of the spMAG1 cDNA to perform the mutagenic PCR as they facilitate easy cloning by megaprime PCR. We then gel purified the PCR product using the Nucleospin extract II kit (E&K Scientific) and cloned the library into pLitmus 28i using megapriming PCR (20). In brief, 250ng of the PCR product was used as the megaprimer and replicated with 50ng of the template plasmid (pLitmus 28i containing wild-type spMAG1) using the PCR protocol outlines in the GeneMorph II EZClone Domain Mutagenesis Kit (Agilent). Afterwards the PCR product was digested with 10 units of DpnI for 1 hour and then transformed into competent Top10 cells (due to their high transformation efficiency). Cells were then plated on pre-warmed petri dishes containing carbenicillin at a high colony density. Petri dishes were left in the 37°C to grow over night. In the morning the petri dishes were washed with 2ml of LB broth and then subsequently mini-prepped. Overall our library contained plasmid DNA from >25,000 individual colonies. After construction,
the library was transformed into Top 10 cells and platted at a low colony density (~100 colonies/ petri dish). 143 random colonies from a few transformations were sent to Sequetech Corporation (Mountain View, Ca) for sequencing to determine the spectrum of the random library. 51% of the sequences analyzed contained non-synonymous mutation, 19% contained synonymous mutations, 20% contained indels, and 11% returned wild-type sequences. Overall we obtained 1.7 mutations per mag1 gene.

To evolve spMAG1 under selection, the MAG1 library was transformed into MV1932 cells and platted at a high density of transformant colonies (>100,000). The plate was then washed with 2ml of LB broth. 10µl aliquots of the wash were then inoculated into 5 separate test tubes containing LB broth and 100µg/ml carbenicillin. Cultures were then grown up to saturation phase over a 24hr period. The next day 10µl of the saturated culture were inoculated into a test tube containing fresh LB and carbenicillin. This process was repeated for 5 days and on the final day cultures were diluted 10^6 and platted on petri dishes containing 100µg/ml carbenicillin. Random colonies from each of the five independent experiments were sent to Sequetech Corporation (Mountain View, Ca) for sequencing.

*7MeG, εA and 3MeA repair assays in vitro*

Constructs were cloned into the pBG100 vector, which contains a 6x histidine tag and a T7 promoter. Protein purifications were performed using standard protocols
described previously (1,26) but in brief a liter of culture in exponential phase was induced with 500uM IPTG and grown for 18 hours at 25°C. Afterwards cells were lysed and homogenized and the protein was purified using a his60 Ni Superflow Resin column (Clonetech), followed by cleavage of the 6xhis tag and a 5ml HiTrap SP HP cation exchange column (GE Technologies).

7MeG and εA assays have been described previously (1) but in brief 6uM of purified protein was incubated with either a fam 5’ labeled double stranded 12mer containing a modified guanine or a P32 labeled double stranded 12mer containing a modified adenine in 1x activity buffer (Hepes pH 7.4 250mM, KCl 500mM DTT 50mM EDTA 10mM). Reactions were run for 240 minutes with aliquots taken at regular intervals. To stop the reaction aliquots of the reaction were mixed with a final concentration of 200mM NaOH. Addition of NaOH converts abasic sites (lesion removal) to single strand DNA breaks. Afterwards the samples were dehydrated and resuspended in Formamide loading dye and run on an acrylamide gel. The gel was imaged and quantified using a Typhoon scanner (GE technologies). Results represent the percent repair of a given lesion at the maximum time point of 240 minutes where no intact DNA left in the labeled DNA would be equal to 100% repair.

3MeA is a labile lesion that cannot be stably placed in an oligo so therefore 3MeA lesion repair was assayed for biochemically using mass spectrometry. In Brief 5uM of protein was incubated for 1hr with calf thymus DNA that had been previously exposed to the methylating agent N-methyl-N-nitrosourea (MNU). After the reaction was stopped with 1M NaCl and the DNA was precipitated, the DNA was then loaded
onto the mass spectrometer. Charge to mass ratio of 3MeA lesions was determined using a 3MeA standard. Removal of 3MeA lesions by the glycosylases was compared to a positive control of 5N HCl. Percent repair of 3MeA lesions is the amount of 3MeA removed by the glycosylases compared to the amount removed by the 5N HCl, which is set at 100%.

*Arabinose induction assay*

The pBAD expression system allows for tightly controlled, titratable protein expression through regulation of arabinose (18). spMAG1 was cloned using SmaI and HindIII into the multi-cloning site of pBAD using the primers F:

```
AAAAAAACCCGGGATGGGTTCTTCTCAC
```

and R:

```
AAAAAAAAAGCTTTTCAGTGGTTTCTTCGG
```

MV1932 cells were transformed with pBAD constructs containing spMAG1, or the K126N E130G N175D triple mutant of spMAG1 and platted on petri dishes containing 100µg/ml kanamycin. Individual colonies were picked and grown up in liquid culture overnight. In the morning cultures were passaged and diluted down to an OD of 0.15 in LB plus an appropriate volume of arabinose and allowed to grow up to saturation for 18 hours. The next day the cultures were serial diluted to 1x10^-3 and platted on petri dishes containing arabinose and 100µg/ml kanamycin. Colonies were counted to determine viability. As a control for arabinose induction, a construct bearing GFP under ara control was used and transformed in parallel with the other constructs. GFP fluorescence was
determined by measuring fluorescence at 570nm excited at 365nm on a UVP imaging system (BioSpectrum).

Statistics

Data are expressed as mean±/ standard deviation (SD). Data were analyzed using Dunnett’s test to test specific hypotheses comparing the performance of various mutants versus an appropriate control, using JMP software (Version 10.0, 2012, SAS Institute). A P-value of <0.05 was considered statistically significant.

Results

MNNG protection in tag ada E. coli cells is a readout for 3MeA repair

MV1932 (tagA ada) E. coli cells are used as a complementation system for 3MeA glycosylases, since these cells lack the constitutive 3MeA glycosylase (tagA) and are unable to express the inducible glycosylase (alkA, whose induction is ada-dependent) (6). To establish the specificity of our complementation system as a readout for 3MeA lesion repair, we compared the effect of expressing a 3MeA glycosylase that almost exclusively repairs 3MeA lesions (TAG) with a 3MeA glycosylase that repair a broad range of substrates (scMAG) and a 3MeA glycosylase that repairs 3MeA but with less affinity than the other two (spMAG1). We found comparable levels of
protection against the $S_N1$ alkylating agent MNNG for both TAG and scMAG (Fig. 1). Our results suggest that repair of lesions other than 3MeA plays little role in this system. Similar results were obtained with the $S_N2$ alkylating agent MMS, consistent with the fact that 3MeA is made by both $S_N1$ and $S_N2$ agents with similar frequency (30). We confirmed that protection was indicative of glycosylase activity, as in all cases mutations that inactivate or decrease glycosylase activity abolished protection (Fig. 1).
Fig. 1. Protection profile of heterologous glycosylase expression against DNA alkylating agents. Measure of continuous growth of *E. coli* MV1932 cells complimented with the TAG, spMAG1, and scMAG glycosylase, respectively against a gradient of the alkylating agents (A) MNNG and (B) MMS. Error bars represent standard deviation of three independent experiments. Samples without error bars were only tested once. * indicates mutants with significantly different growth on gradients compared to their wild type counterpart using a Dunnett’s comparison of means test.
A. Growth on 0.03% MNNG Gradient (cm)

B. Growth on 0.05% MMS Gradient (cm)
Expression of yeast 3MeA glycosylases in E. coli is toxic

When not under the challenge of an alkylating agent, the ectopic expression of our yeast 3MeA glycosylases in MV1932 cells caused a deleterious effect on cellular viability, most noticeably in saturated cultures, which are known to increase the levels of metabolic stress and endogenously produced DNA lesions (29,32) (Fig. 2a). Our results, at least for scMAG, are consistent with previous reports showing that overexpression of scMAG is toxic in an E. coli complementation system (25). We confirmed that the observed cytotoxicity associated with yeast 3MeA glycosylase expression was due to excessive glycosylase activity, because scMAG had a stronger deleterious effect than spMAG1, and because the toxicity of spMAG1 decreases with addition of the D170N mutation, which decreases spMAG1 enzymatic activity by 80% (1) (Fig. 2a). Using our in vivo 3MeA repair assay, we confirmed that scMAG exhibited increased 3MeA repair relative to spMAG1 and that the D170N mutant of spMAG1 has very low levels of 3MeA repair (Fig. 2b).
Fig. 2. Viability and protection profile of yeast MAG glycosylases. (A) Viability of MAG constructs plated at a $10^{-3}$ dilution from a culture left in stationary phase. * indicates strains of *E. coli* that grew significantly different than delta MV using a Dunnett’s comparison of means test. (B) Growth of MAG constructs in centimeters on a gradient of 0.03% MNNG complemented into the MV1932 strain of *E. coli*. Delta MV is a pLitmus 28i vector only. WT *E. coli* is the AB1157 strain of *E. coli* containing delta pLitmus 28i. Error bars represent standard deviation of triplicate experiments. * indicates constructs that grew significantly different than spMAG on MNNG gradients using a Dunnett’s comparison of means test.
Fitness inversely correlates to level of WT spMAG1 expression.

We confirmed that excessive glycosylase activity is deleterious in our complementation system by placing the spMAG1 gene under the control of the ara promoter in a low-copy plasmid vector. We increased spMAG1 expression by inducing the promoter with increasing concentrations of arabinose in the medium and also by expressing spMAG1 from a multicopy plasmid vector. Levels of expression were monitored using a GFP fusion reporter (dotted line, Fig. 3). The results shown in Fig. 3 confirm that residual amounts of 3MeA glycosylase activity are sufficient to confer good viability to E. coli MV1932 cells (see 0% arabinose point), but that past a certain point, expression clearly becomes toxic. Included in the figure is a hypomorphic mutant of spMAG1 (characterized below) in order to show that the observed toxicity is indeed linked to glycosylase activity rather than to some non-specific effect of spMAG1 overexpression.
Fig. 3. MAG viability under an inducible promoter. (A) Viability of MAG wild type or mutant at a $10^{-3}$ dilution from a culture left in stationary phase cloned into the pBAD or pLitmus vector and induced with either 0%, 0.1%, or 1% arabinose. Error bars represent standard deviation of triplicate experiments. * indicates constructs that grew significantly different than spMAG1 in pLitmus 28i using a Dunnett’s comparison of means test. (B) GFP fluorescence as a measure of protein expression for each construct and condition is shown as a measure of GFP fluorescence units.
Directed evolution for reduced spMAG1 toxicity results in the selection of hypomorphic mutants

Expression of both spMAG1 and scMAG on the high copy pLitmus vector exhibits decreased cell viability. Therefore we set out to evolve spMAG1 mutants with improved viability. To do this we generated a random mutant library of spMAG1 using error-prone PCR and selected for mutants with increased viability by passing transformant cultures to saturation several times. Our selection was performed in over five independent experiments run in parallel.

The results of our selection are consistent with our complementation experiments indicating that in MV1932 E. coli cells, 3MeA glycosylase activity contributes positively to survival, as we detected a strong selection against stop codons and indels (we went from 20% in the unselected library to only 5% in the selected one). In addition, we frequently obtained the same mutants in at least two of the five independent selections indicating a surprising amount of convergent evolution events (Table1). One of the mutants (K126N E130G N175D) was obtained independently in all five selections.

This strong convergent evolution is suggestive of some gain-of-function. We confirmed that all the mutants selected in more than one independent experiment (a total of six) indeed had a positive effect on viability relative to WT spMAG. This effect was substantial, with increases in viability of at least 10-fold (Fig. 4a). At the same time, all of these six mutants exhibited a substantial decrease in 3MeA repair in
in vivo (Fig. 4b), suggesting that minimizing toxicity in an *E. coli* host is linked to reducing 3MeA activity of spMAG1.

We decided to characterize the biochemical activity of our most frequently selected mutant, K126N E130G N175D. We performed oligo excision assays using purified spMAG1 protein and oligos containing the lesion in question to quantify the activity of our mutant enzyme on two of its known substrates, 7MeG and etheno-adenine (εA). Since 3MeA is a labile lesion that cannot be stably placed in an oligo, 3MeA lesion repair was assayed by measuring the release of 3MeA lesions from spMAG1 protein incubated with calf thymus DNA that had been previously exposed to the methylating agent MNU (N-Methyl-N-nitrosourea) using mass spectrometry. **Fig. 5** shows that although the triple mutant of spMAG1 still retains activity towards all of its canonical substrates, its level of activity is decreased for all lesions tested. Interestingly, the activity against 7MeG and εA is much more affected than activity against 3MeA, suggesting a selective pressure for maintenance of activity against 3MeA lesions.

Overall our genetic and biochemical analyses suggest that spMAG1 activity in a *tag ada* strain of *E. coli* needs to be reduced to avoid toxicity, although some level of activity is necessary for optimal fitness.
Table 1. spMAG1 selection results. List of spMAG1 non-synonymous mutants pulled from each individual selection experiment.
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
<th>Experiment 4</th>
<th>Experiment 5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>NA</td>
</tr>
<tr>
<td>L122M</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>K126N E130G N175D</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>N42Y V165A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>W208stop</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>I147V T151A A206S</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>R149K</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Others</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>NA</td>
</tr>
</tbody>
</table>
Fig. 4. Viability and protection profile of fitness selected MAG1 mutants. (A) Viability of the six MAG mutant constructs selected for increased cellular fitness plated at a $10^{-3}$ dilution from a culture left in stationary phase. * indicates spMAG1 grew significantly less than all other constructs. (B) Growth of the six MAG mutant constructs selected for increased cellular fitness on a gradient of 0.03% MNNG complemented into the MV1932 strain of *E. coli*. Error bars represent standard deviation of triplicate experiments. * indicates constructs that grew significantly less than spMAG1 on gradients of MNNG.
A. Viability of Saturated Culture

B. Growth on 0.03% MNNG Gradient (cm)
Fig. 5. Repair of individual lesions by MAG1. Percent repair of either 7MeG, εA, or 3MeA lesions by the K126N E130G N175D triple mutant of spMAG1 relative to wild-type control. 7MeG and εA repair was measured by *in vitro* oligo excision and 3MeA repair by mass spectrometry. The results shown are one representative experiment.
Under MNNG challenge, decreasing the deleterious effects of 3MeA activity on fitness prevails over increasing 3MeA activity.

We next forced evolution to maintain 3MeA activity by performing our selection in the presence of the methylating agent MNNG. After transformation of the library into MV1932 cells, we ran our cultures on MNNG gradients and picked colonies growing at the upper threshold of resistance. Out of 28 colonies sequenced, 6 carried a S60P mutation, and 14 were wild-type. We observed a clear enrichment of clones bearing the wild-type sequence or synonymous mutations (from 19% in our unselected library to 36% in our MNNG selected results), a clear indication of positive selection for wild-type sequence. We then characterized the effect on viability of the S60P mutation, along with that of a mutant that appeared only once in our selection (M45V). Neither mutant showed increased levels of 3MeA repair, in fact S60P clearly showed decreased repair levels (Fig. 6a) but both mutants increased viability in saturated cultures (Fig. 6b). Overall, these results suggest it may not be possible to enhance constitutive 3MeA repair by spMAG1 in an *E. coli* complementation system due to a tradeoff between 3MeA glycosylase activity and cellular fitness.
Fig. 6. Protection profile and viability of 3MeA repair selected MAG1 mutants. (A) Growth of the MAG mutant constructs selected for 3MeA repair capability on a gradient of 0.03% MNNG complemented into the MV1932 strain of E. coli. (A) Viability of the MAG mutant constructs selected for 3MeA repair capability plated at a 10^-6 dilution from a culture left in stationary phase. Error bars represent standard deviation of triplicate experiments.
Discussion

*Optimal fitness depends on a narrow window of spMAG1 activity*

The present work indicates that fitness is dependent on a narrow window of 3MeA glycosylase activity, at least in *E. coli*. Complete absence of 3MeA glycosylase activity is deleterious. Delta *tag ada* cells consistently show decreased viability relative to WT *E. coli* ([Fig. 2a](#)) and our selections consistently select against indels and stop codons.

On the other hand, an excess of glycosylase activity can be equally or even more deleterious. This is based on the following lines of evidence:

1) Expression of spMAG1 and of scMAG decreases viability ([Fig. 2a](#), [Fig 6b](#)) of *tag ada* *E. coli* cells

2) scMAG exhibits a stronger cytotoxic effect than spMAG1, correlating with their respective levels of 3MeA glycosylase activity ([Fig. 1, 2](#)).

3) Levels of active spMAG1 expression correlate with cytotoxicity ([Fig. 3](#))

4) Evolution of spMAG1 mutants with decreased toxicity consistently exhibit decreased 3MeA repair ([Fig. 4](#))

5) Wild-type or hypomorphic mutants were selected when forcing the protein to maintain high levels of 3MeA activity ([Fig. 6](#)).

These results strongly suggest that inducible glycosylase genes likely arose as a way to bypass the negative effects of constitutive glycosylase activity on organismal
fitness and explains the overall lower activity of constitutive forms of glycosylases.

*spMAG1 selected mutants: structure, function, & epistatic interactions*

The repeated isolation of the same mutants in independent experiments is remarkable and highlights the narrow window of optimal glycosylase activity. We found six different mutants in more than one independent experiment (listed in Table 1). We found that all six mutants showed decreased 3MeA repair activity. As a quick way to investigate the role of each individual mutation in the observed phenotypic outcome we ran every mutation listed in Table 1 through a SIFT (Sorting Intolerance from Tolerance) algorithm. SIFT analysis estimates how detrimental a substitution at a given position will be to the queried protein based on a combination of phylogenetic and biophysical considerations (17,24). A SIFT substitution probability score functions much the same way as a statistical p value where any substitution score above 0.05 is predicted to not detrimentally affect protein function. The results (Table 2) show that with the exception of 151A, 165A, and 206S all other mutations of spMAG1 are predicted to retain enzymatic function.

Only one of the reoccurring mutants isolated in our selection experiments contained a single non-synonymous substitution: L122M. Based on the 3D structure of spMAG1 (1), substitutions at position 122 likely have an effect on the local packing of the protein. All other mutants have more than one mutation. Based on our SIFT analysis, we found two cases of positive epistasis, *i.e.* two cases in which
deleterious mutations are partially compensated by additional mutations. The first case is the N42Y V165A double mutant. The V165A mutation has a highly significant (<0.05) SIFT score of 0 because it is a residue that helps pack and partially buries the active site of the protein; positions N42 and V165 are in loops across domains, facing each other consistent with the idea that the N42Y mutation may compensate for the deleterious effects of V165A. Similarly, in the I147V T151A A206S mutant, both T151A and A206S mutations have highly significant SIFT scores. Position 151 is part of the helix-loop-helix domain, at the end of helix J and next to the DNA (at a distance of 2.6Å), forming a H-bond with the DNA backbone. Position A206 is located in an internal of cluster of helices, buried, packing the whole little cluster. This position appears to be critical in a variety of glycosylases, as *E. coli* TAG holds a Zn atom at the homologous position and *H. pilori* Mag III holds a cabomoilated Leucine at the homologous position (23). The presence of these two highly deleterious mutations in a mutant that shows decreased but still detectable 3MeA repair activity (Fig. 4) suggests a positive epistatic interaction.

*spMAG1 K126N E130G N175D*

The K126N E130G N175D mutant stood out from the rest, as it was found in all five independent experiments. We found that although this triple mutant of spMAG1 still retains activity towards all of its canonical substrates, its level of activity is less than that of wild type spMAG1 for all lesions tested. Interestingly, in this triple mutant,
activity against 7MeG and εA is much more affected than activity against 3MeA. This observation suggests a selective pressure for maintenance of 3MeA repair, which would explain the fact that although constitutively-expressed 3MeA glycosylases contain a narrowed substrate specificity they typically still retain the ability to repair 3MeA lesions in particular (1,23,33). This implies that 3MeA, represents a more significant endogenous cytotoxic lesion than 7MeG and εA, at least in E. coli.

Compared to the two putative examples of positive epistasis described above, this mutant likely represents the opposite example, i.e. it is likely an example of negative epistasis between mutations that individually have little effect on activity. This is suggested by the non-significant SIFT scores we found for each of the non-synonymous mutations present in this spMAG1 mutant (>0.2 in all three cases, Table 2). Consistent with this hypothesis, all three mutations are required for optimal viability, with K175D having a smaller contribution than the other two mutations (Fig. 7). E130 is in a disordered section of the protein, in the same face as K126. The E130G and K126N mutations may enable the formation of a hydrogen bond between these two residues. The K175D mutation on the other hand is in the active site, close to the DNA backbone and possibly interfering with electrostatic binding (Fig. 8). We confirmed that glycosylase activity was necessary for the optimized survival provided by the K126N E130G N175D triple mutant by showing that the introduction of the D170N mutation decreases survival (Fig. 7).
Table 2. SIFT substitution for mutants. Probably of non-synonymous mutations being functional. 95 homologous sequences compared.

<table>
<thead>
<tr>
<th>Position</th>
<th>Mutant</th>
<th>Substitution observed</th>
<th>SIFT substitution probability</th>
<th>SIFT probability for no mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>N42</td>
<td>N42Y V165A</td>
<td>Y</td>
<td>0.06</td>
<td>N = 0.16</td>
</tr>
<tr>
<td>L122</td>
<td>L122M</td>
<td>M</td>
<td>0.13</td>
<td>M = 0.56</td>
</tr>
<tr>
<td>K126</td>
<td>K126N E130G N175D</td>
<td>N</td>
<td>0.58</td>
<td>K = 0.83</td>
</tr>
<tr>
<td>E130</td>
<td>K126N E130G N175D</td>
<td>G</td>
<td>0.37</td>
<td>E = 1.00</td>
</tr>
<tr>
<td>I147</td>
<td>I147V T151A A206S</td>
<td>V</td>
<td>0.29</td>
<td>I = 1.00</td>
</tr>
<tr>
<td>R149</td>
<td>R149K</td>
<td>K</td>
<td>0.35</td>
<td>R = 1.00</td>
</tr>
<tr>
<td>T151</td>
<td>I147V T151A A206S</td>
<td>A</td>
<td>0</td>
<td>T = 1.00</td>
</tr>
<tr>
<td>V165</td>
<td>N42Y V165A</td>
<td>A</td>
<td>0</td>
<td>V = 1.00</td>
</tr>
<tr>
<td>N175</td>
<td>K126N E130G N175D</td>
<td>D</td>
<td>0.2</td>
<td>N = 0.24</td>
</tr>
<tr>
<td>A206</td>
<td>I147V T151A A206S</td>
<td>S</td>
<td>0.01</td>
<td>A = 1.00</td>
</tr>
</tbody>
</table>
Fig. 7. Viability of the deconstructed K126N E130G N175D mutant of spMAG1. Viability of various MAG constructs plated at a $10^{-3}$ dilution from a culture left in stationary phase. Delta MV is a pLitmus 28i vector only. Error bars represent standard deviation of triplicate experiments.
Fig. 8. Crystal structure of spMAG1. Yellow amino acids represent the HhH DNA binding domain, blue amino acids are additional amino acids that bind the DNA, The red amino acid denotes the position of the catalytic aspartic acid residue necessary for nucleotide cleavage, and light blue denotes the location of the mutations from the K126N E130 N175D triple mutant of spMAG1. Figure is from (1).
Conclusion

3MeA glycosylases can be divided into two categories: those that are constitutively expressed and those that are inducible. These two categories differ in their catalytic efficiency (low for constitutive, high for inducible) and in their substrate specificity (narrow for constitutive and broad for inducible). This divide across categories is particularly striking given that there is no apparent genetic barrier for constitutive glycosylases to evolve broader substrate specificity. Our complementation and directed evolution experiments demonstrate a tradeoff between constitutive 3MeA glycosylase activity and protection against 3MeA cytotoxicity. Our work indicates that constitutive 3MeA glycosylase activity needs to be minimized, leaving cells vulnerable to increased exposure to methylating agents. These findings suggest that inducible 3MeA glycosylases likely arose as a way to bypass the negative effects of constitutive glycosylase activity on organismal fitness. Our work also points to 3MeA as the critical endogenous target for 3MeA glycosylase repair, explaining the narrower substrate specificity of constitutive 3MeA glycosylases.
Acknowledgements

The authors would like to thank Dr. Brandt Eichman and the Eichman laboratory for providing us with the MAG constructs, the pBG100 vector, as well providing a workspace in which to perform the biochemical experiments. The authors would like to thank Dr. Suraj Adhikary for his insight and help with the 7MeG and εA excision assays and Dr. Emily Rubinson for her work on the mass spectrometry data. Funding was provided by the National Cancer Institute [CA116429-04] to MC and by the National Institute of Environmental Health Sciences [ES019625-01A1] to MC.
References:


7. Chen, J., Derfler, B. and Samson, L. (1990) Saccharomyces cerevisiae 3-methyladenine DNA glycosylase has homology to the AlkA glycosylase of E. coli and is induced in response to DNA alkylation damage. *EMBO J.,* 9, 4569-4575.


Chapter 4:

Mutagenic footprint of low-fidelity Pol I replication in *E. coli* reveals an extensive interplay between Pol I and Pol III during ColE1 plasmid replication.
Abstract

ColE1 plasmid replication requires the coordinated action of two DNA polymerases: DNA polymerase I (Pol I) and DNA polymerase III (Pol III). Plasmid replication starts with the extension of an origin of replication (ori) transcript by Pol I. At a point known as “polymerase switch”, the Pol III holoenzyme assembles and completes plasmid replication. Using mutations generated during ColE1 plasmid replication in vivo by a low fidelity-mutant of Pol I (LF-Pol I), we found an area of leading strand that appears to be replicated exclusively by Pol I. The mutation spectrum for LF-Pol I in this section of sequence is highly biased between complementary pairs, allowing us to use individual mutations as indicators of strandedness. We validated this mutation footprinting approach by showing that we derive a consistent mutation spectrum regardless of which strand or portion of the sequence the spectrum is derived from. Our mutation footprint provides the following mechanistic insights on the role of Pol I during ColE1 plasmid replication: 1) locates the polymerase switch at ~200 bp downstream of replication initiation; 2) shows that downstream of the polymerase switch, Pol I replicates both strands with comparable frequency, suggesting Pol I may be capable of coordinated double-strand replication; 3) provides evidence for a gap-filling role of Pol I associated with termination of replication. Finally, we detect differences in replication fidelity produced by changes in culture conditions, illustrating how our strand-specific footprinting approach can be used to dissect genomic and physiologic variables modulating Pol I fidelity in vivo.
Introduction

ColE1 plasmids are the basis for most recombinant expression vectors in *E. coli* and represent convenient models for mechanistic studies of DNA repair and replication. ColE1 plasmid replication is initiated by the transcription of a ~600 nt sequence known as the plasmid origin of replication (*ori*), generating an RNA primer known as RNA II. RNA II extension by Pol I initiates leading-strand synthesis (12), which facilitates replisome assembly by exposing a primosome assembly signal on the leading strand (18).

A critical component of the replisome is the Pol III holoenzyme (Pol III HE). This complex contains two core subassemblies. Each core subassembly consists of three tightly-bound subunits: α (dnaE, the DNA polymerase), ε (dnaQ, the 3’→5’ proofreading exonuclease), and θ (holE, stabilizer for the ε subunit). The two cores are connected by a ring-shaped sliding-clamp processivity factor, creating a dimeric polymerase unit that replicates both strands (leading and lagging) coordinately at high speed (for reviews see (13,19)). While leading-strand synthesis by the Pol III HE is continuous, by 5’ to 3’ elongation of the Pol I-synthesized leading-strand, the lagging strand is synthesized 3’ to 5’ in fragments that are assembled together. DnaG, a DNA primase, synthesizes short RNA primers that are elongated by Pol III; when Pol III reaches the end of the Okazaki fragment, it is replaced by Pol I, which removes the RNA primer through its 5’→3’ exonuclease activity and fills in the remaining gap.

Fluorescence localization of tagged core Pol III subunits using single-molecule
microscopy during replication in living *E. coli* cells indicates that a new Pol III molecule is used for each Okazaki fragment (15).

We previously created a low fidelity variant of Pol I (LF-pol I) by mutating three key determinants of fidelity: I1709N in motif A, A759R in motif B, and A424D in the proofreading domain (5). Expression of this variant in JS200 cells, a *polA12* (Pol I temperature-sensitive) strain of *E. coli*, under restrictive conditions leads to the introduction of random mutations during ColE1 plasmid replication (5). These mutations can be considered a footprint of Pol I replication. In a previous article (1), a thorough analysis of the footprint of error-prone Pol I replication in ColE1 plasmids allowed us to establish the following: 1) Pol I replication extends well beyond the ~100 nt of RNA II primer extension reported *in vitro*; 2) the likely location of Okazaki processing sites on the plasmid; 3) the extent of processing by Pol I at Okazaki processing sites, which we found is limited to ~20 nucleotides.

The main limitation of our previous article was our inability to ascribe strandedness to individual mutations. All we could say was that mutations were more likely to correspond to leading or lagging strand, based on an asymmetric representation of complementary pairs. Here we refine our original approach by estimating the mutation spectrum of LF-Pol I on the leading strand and generating a strand-specific footprint of LF-Pol I replication of the plasmid. We then validated our approach by deriving comparable mutation spectra from either strand and before and after the polymerase switch, reasoning that the mutation spectrum is an intrinsic property of
the polymerase and therefore (barring local sequence context effects) should be the same regardless of strand or section of plasmid sequence considered.

Our current analysis identifies the location of the polymerase switch 170 to 250 nt downstream of DNA replication initiation. Downstream of the polymerase switch, Pol I replication shows little strand preference, suggesting Pol I may be capable of coordinated double-strand replication, consistent with previous reports of a functional redundancy between Pol I and the α subunit of Pol III. In addition, at the 3’ end of plasmid replication we discovered a clear bias for lagging-strand synthesis, pointing to a role for Pol I in termination of lagging-strand replication.

Methods

Bacterial strains

JS200 (SC-18 recA718 polA12ts uvrA355 trpE65 lon-11 sulA1) cells were used as our host strain. The polA12 allele encodes a point mutation in Pol I (G544D) that interferes with the coordination between the polymerase and the 5’→3’ exonuclease activities (4). This Pol I mutant exhibits reduced temperature stability and activity at 42 °C (25). RecA718 is a sensitized allele of RecA, resulting in SOS induction under conditions that are restrictive for polA12 (7).

Plasmid constructs

Our mutagenic plasmid expressing LF-Pol I (muta-plasmid) was generated by cloning of the mutant Pol I sequence into a pHSG576 vector between the HindIII/EcoRI
restriction sites, and bears chloramphenicol resistance (22). pGFPuv (with
carbenicillin resistance) was obtained from Clontech (Mountain View, CA). The
pLitmus ALKBH1 plasmid was generated by cloning the ALKBH1 cDNA sequence
(GenBank: BC025787.1) into the multi-cloning site of the pLitmus 28i vector
between the XhoI/HindIII restriction sites.

Media and Supplies

Growth media LB Agar and LB broth were purchased from Fisher Scientific and
prepared according to vendor specifications. Some mutagenesis experiments were
carried out in 2XYT rich media containing 0.016g/ml Bacto Tryptone, 0.01g/ml
Bacto Yeast Extract and 0.005g/ml NaCl suspended in deionionized water. The
antibiotic concentrations used for marker selection are: 30µg/ml (chloramphenicol),
and 50µg/ml (hTK) or 100µg/ml (all other libraries) (carbenicillin). All DNA
isolation procedures were performed using Machery Nagel’s Nucleospin Plasmid
miniprep. Sequencing was carried out by Sequetech (Mountain View, CA) using the
following sequencing primers; attP2 (CAGGAAACAGCTATGAC) and Blac5
(TTACGGTTCTGGCCTTTTGC) for pGFPuv and MC360
(CTTGCCACTTGCTGACGG) for ALKBH1 libraries, respectively.

Error-prone pol I Mutagenesis

The target plasmid, a ColE1 plasmid bearing the gene of interest, was transformed
into JS200 cells carrying muta-plasmid, the pSC101 (Pol I-independent) plasmid
bearing our low-fidelity Pol I. When these transformants are grown under restrictive
conditions, low-fidelity Pol I is the functional polymerase present in the cell, introducing random errors during replication of the ColE1 target plasmid.

Liquid mutagenesis: mutagenesis in liquid culture was performed by switching a culture grown under permissive conditions (LB, 30°C, exponential) to restrictive conditions (2XYT, 37°C, saturation) as described in (5). Briefly, ~100ng of the target plasmids (pGFPuv or pLitmusALKBH1) were transformed into electrocompetent JS200 muta-plasmid cells (for preparation of competent cells, see (24)). The transformants were resuspended in 1ml LB Broth, recovered for 1h at 30°C, and plated at 30°C on LB Agar plates containing 100 µg/ml carb. A single colony was picked from each plate, inoculated into 4ml LB broth and grown at low density at permissive temperature (30°C). For mutagenesis, an aliquot of the overnight culture (dilution factor 1:10³ to 1:10⁵) was transferred into 4ml of 2XYT media (pre-warmed at 37°C), and grown shaking at 37°C for 1 or 3 days to reach complete saturation or hypersaturation (24). Following mutagenesis, plasmid DNA was isolated using Machery Nagel’s Nucleospin Plasmid miniprep kit.

Solid plate mutagenesis:

Electrocompetent JS200 cells carrying muta-plasmid, the pSC101 (Pol I-independent) plasmid bearing our low-fidelity Pol I were transformed with ~100ng of the target plasmids (pGFPuv or pLitmusALKBH1). Cells were allowed to recover at 37°C in LB broth for one hour. Cells were then plated on pre-warmed Petri dishes containing chloramphenicol and carbenicillin (to select for both the pSC101 and the target
plasmid) at a high colony density (>100,000 colonies). Petri dishes were left in the 37°C to grow over night. In the morning the Petri dishes were washed with 2ml of LB broth and then subsequently mini-prepped. This constituted one round of solid plate mutagenesis.

Iteration of mutagenesis and sequencing

The mutagenesis procedures described above were repeated to increase the mutation frequency as described in detail in (24). Briefly, the plasmid library recovered from the initial round of mutagenesis was retransformed into fresh JS200 muta-plasmid cells at 30°C, and transformant colonies were washed, inoculated into 4ml of 2XYT media and grown to saturation at 37°C (liquid protocol) or retransformed JS200 muta-plasmid cells were directly plated at 37°C (solid plate mutagenesis). These procedures were repeated until the desired mutation frequency was reached. Individual plasmids were identified through transformation of a small amount of plasmid DNA (50-100ng) into BL21 cells. From this transformation, individual colonies were sequenced. In Supplementary table 1 for each clone present in our libraries we list number of mutagenesis cycles, sequence coverage, and mutations found. This information is summarized in Table 1 of the main text.

Hotspot identification

We investigated the spatial distribution of mutations by calculating the distribution of distances between mutant positions for all the mutations included in this study (501 mutations at 335 positions for GFP and 260 mutations at 179 positions for ALKBH1).
A “hotspot index” was calculated, defined as the number of mutations that can be grouped as being in the same or adjacent positions. Moving along the sequence we start counting at 1 until we have exhausted the number of mutations at that position. We then check the next nt position and continue to raise the count if it has one or more mutations. If not, the count is logged and the count starts over, continuing forward on the strand. Individual hotspot indexes are plotted in Suppl. Fig. 1 and the cluster size distribution for these indices is shown in Suppl. Fig. 2. The vast majority of clusters fall in the 1-2 mutation category (n=303). However, 16 of these clusters deviate from the overall distribution, with ≥ 6 mutations. The sequence of these clusters is shown in Suppl. Table 2. Each cluster comprises a moderate number of mutations, with an average of 9 mutations per cluster and a standard deviation of 2.5. These hotspots tend to be GC-rich, with fifteen of these sites occurring in runs of 3 or more C or G residues. The only recognizable motif was “5’-CCA/TA/T-3’”, found in seven of the hotspots (highlighted with a grey box in Table 2). Mutations in these clusters (133 mutations in total) were considered “hotspots” and excluded from the footprint as unrepresentative of randomly distributed mutations made by the polymerase.

*Strandedness markers*

Leading-strand marker mutations are defined as the most frequent of the complementary pairs: A→G, C→T, A→T, and G→T. Lagging-strand marker mutations are defined as the least frequent of the complementary pairs: T→C, G→A,
T→A, and C→A. Due in part to low representation in our database the strandedness of C→G vs. G→C was inconclusive. The frequency of T→G and A→C mutations in the area of single-stranded Pol I replication was also too low to ascribe strandedness unambiguously but we designated T→G as a marker for leading-strand synthesis based on the fact that we did see a moderate enrichment for T→G mutations in that area and that we saw the converse in putative Okazaki processing sites: an enrichment for A→C mutations.

Identification of Okazaki processing sites

Putative Okazaki processing sites within the sequence coverage area were identified using the approach previously described in (1). Briefly, clusters of consecutive marker lagging-strand mutations located at a short distance from each other (d≤5; the average distance being 7.4) were plotted. Based on the cluster size distribution, clusters with 6 or more mutations were considered significant. This identified the following positions as likely Okazaki processing sites: positions 179-191; 559-578; 910-926; and 1035-1045 (sites I-IV in Suppl. Fig. 3a). Of these, site I (comprising positions 179-191) is likely artifactual because it falls squarely in a mutation hotspot as defined in Suppl. Fig.1a and it does not show the same level of enrichment for lagging-strand mutations seen in the other three sites (Suppl. Fig. 3b). For the remainder of the analysis we considered only sites II, III and IV as legitimate Okazaki processing sites.
Statistical methods

To aid in the identification of the hotspots, we simulated the null distribution of hotspots under the assumptions that for each library, coverage was the same across clones and that mutations occurred independently with probability $p$ at each nt position in each clone. We calculated 1,000,000 Monte Carlo replicates for each library, then estimated the right tail probabilities using the empirical distribution.

To test for a statistically significant difference in mutation frequency between leading and lagging strands in the area beyond the putative switch, we binned the mutation counts into 50 nt intervals, then performed a Wilcoxon Signed Rank test using these counts ($p=0.26$).

In several instances, we performed a Fisher’s exact test on observed mutational spectra from different locations or libraries in order to ascertain whether or not they were different at a statistically significant level. The reported p-values were estimated by Monte Carlo with 100,000 simulations.

All simulations and statistical tests were performed using standard R packages.
<table>
<thead>
<tr>
<th>Library</th>
<th>Form</th>
<th># cycles</th>
<th># clones</th>
<th>Coverage (kbp)</th>
<th># mut</th>
<th>Mut frequ $\times 10^3$</th>
<th>Mut freq/cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFP-A</td>
<td>Suspension</td>
<td>1</td>
<td>154</td>
<td>305</td>
<td>184</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>GFP-B</td>
<td>Suspension</td>
<td>2</td>
<td>185</td>
<td>55.9</td>
<td>55</td>
<td>0.98</td>
<td>0.49</td>
</tr>
<tr>
<td>GFP-C</td>
<td>Colony</td>
<td>4</td>
<td>105</td>
<td>116.7</td>
<td>262</td>
<td>2.24</td>
<td>0.56</td>
</tr>
<tr>
<td>ALKBH1</td>
<td>Colony</td>
<td>4</td>
<td>120</td>
<td>80.2</td>
<td>272</td>
<td>3.39</td>
<td>0.85</td>
</tr>
<tr>
<td>Total plankt.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total colony</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 Metrics for the libraries included in this study.
Fig. 1. Identification of mutation hotspots through clustering analysis. Hotspot mutation index, defined as the number of consecutive mutations that can be found at a distance of $\leq 1$ nucleotides from each other is shown on the y-axis relative to the distance (in nucleotides) of mutant positions from the RNA/DNA switch (x-axis). **a** GFP libraries. All 501 mutations from our 3 GFP libraries were included. The average distance between the 337 mutant positions was 2.52 nucleotides. **b** ALKBH1 libraries. All 257 mutations were included. The average distance between the 170 positions was 2.75 nucleotides.
**Fig. 2. Distribution of hotspot indices** shown in Fig. 1 for the GFP libraries (a) and for the ALKBH1 library (b). These indices (y-axis) quantify the number of mutations found in the same or adjacent position for given nucleotide positions.
Fig. 3. Identification of putative Okazaki primer processing sites.  

**a. Clustering analysis.** Number of consecutive positions with signature lagging-strand mutations that can be found at a distance of ≤5 nucleotides from each other (y-axis) relative to the distance (in nucleotides) of mutant positions from the RNA/DNA switch (x-axis). The average distance between lagging-strand mutant positions was 7.39 nucleotides. Clusters considered significant (n>5 mutations) are labeled in roman numerals. The nucleotide positions for these sites are: 179-191; 559-578; 910-926; and 1035-1045.  

**b. Lagging-strand mutation enrichment.** Mutations within the sites defined by clustering analysis (roman numerals on x-axis) were classified as signature leading- or lagging-strand as described in the methods and their ratio calculated. Columns represent ratio of lagging vs. leading at each of these sites.
Results

Replication of a neutral sequence by LF-Pol I generates random mutations whose distribution identifies Pol I templates with high resolution (1). The main limitation of this polymerase template-mapping approach is our inability to identify the strand where mutations originally occurred. In a previous article, we addressed this problem using biases in the distribution of complementary mutations to define probabilistic markers, *i.e.* markers that indicated a higher probability of originating in one strand vs. the other (1). For each complementary pair, the most frequent mutation (A→G; C→T; A→T; G→T) was designated as a marker for leading-strand synthesis and the least frequent, complementary mutations (T→C; G→A; T→A; C→A), as markers for lagging-strand synthesis. Combined with a clustering analysis, this approach identified sites of Okazaki primer processing by Pol I. However, without knowing the mutation spectrum of LF-Pol I on a single strand, we were unable to adequately interpret the mutation footprint outside areas of Okazaki primer processing.

Here our goal is to use LF-pol I mutations to distinguish the role of Pol I from that of the Pol III HE during ColE1 plasmid replication. The present analysis includes two new libraries: one targeting GFP and another one targeting human ALKBH1. These libraries are of high quality because they underwent multiple rounds of mutagenesis (n=4), which increases the mutation density and facilitates the identification of Okazaki processing sites (1), and because they involve sequences that are neutral, *i.e.* sequences that provide no significant fitness advantage or disadvantage to the host.

The sequence coverage for these libraries on their respective plasmids is shown in Fig.
l and their metrics are summarized in Table 1. The new LF-Pol I mutagenesis data can be found in Suppl. Table 1, broken down by individual clones. In addition, our previously-reported GFP library (1) was also included in the present analysis (Table 1).

We found little evidence of clonal selection in our GFP or ALKBH1 libraries, confirming that these genes represent largely neutral targets in our system. In addition, generating a footprint for Pol I replication requires a random distribution of mutations along the sequence. We investigated the spatial distribution of mutations by calculating an index (which we named hotspot index) that measures local clustering of mutations (Methods; Suppl. Fig. 1). In both libraries we found strong evidence of mutation clustering (Suppl. Fig. 2). Suppl. Table 2 shows the local sequence in these areas of decreased Pol I fidelity. These hotspots tend to be GC-rich, with fifteen of these sites occurring in runs of 3 or more C or G residues. We also found one recognizable motif: 5’-CCA/TA/T-3’, which is found in seven of the hotspots. Given the non-random spatial distribution of these mutations (with multiple hits in the same or adjacent positions) and their apparent sequence-context dependence, mutations in these clusters were analyzed separately.

Our clustering analysis aimed at identifying potential Okazaki Processing Sites (OPS)(1) identified the following candidate OPSs in our new GFP library: positions 179-191; 559-578; 910-926; and 1035-1045 (sites I-IV in Suppl. Fig. 3a). Of these, site I (comprising positions 179-191) is likely artifactual because it falls squarely in a
mutation hotspot as defined in Suppl. Fig. 1a and it does not show the same level of enrichment for lagging-strand mutations seen in the other three sites (Suppl. Fig. 3b). For the remainder of the analysis we considered only sites II, III and IV as legitimate Okazaki processing sites. These sites were removed from our replication footprint, as they are associated with a separate role of Pol I in plasmid replication (processing of Okazaki primers).

In our GFP libraries we noticed an area (between positions 70 and 170) where the ratio between marker leading- and marker lagging-strand mutations, is extraordinarily high, in the order of ten-to-one. This trend encompassed all complementary pairs, suggesting that this unique distribution of mutations is likely attributable to strand preference. Indeed, we found a similar area of very high (>10-fold) bias in our new ALKBH1 library (which has a different sequence than the GFP library at these positions), confirming this bias is independent of sequence context. Fig. 2 shows the ratio of leading-strand vs. lagging-strand marker mutations at 50nt intervals for the whole area of coverage for the GFP and ALKBH1 libraries (Fig. 2a and Fig. 2b, respectively). A substantial bias for leading-strand replication is indicated with dark grey arrows. We interpret the end of the high bias for leading-strand synthesis as indicative of the polymerase switch (inverted, white triangle).

Beyond the putative switch, the ratio of marker leading vs. marker lagging strand mutations for the GFP library stayed around 1. The average between leading vs. lagging strand ratio over 18 intervals was 1.3, with a standard error of 0.33 (Fig. 2a).
The absolute frequency of mutation for each strand in our new GFP library is shown in Fig. 2c at 50 nucleotide intervals, with leading strand mutations in grey circles and lagging-strand mutations in white circles. Again, we detected no statistical difference in the absolute frequency of Pol I mutagenesis between leading and lagging-strand mutations in the area beyond the putative switch (p value for Wilcoxon test=0.2). These results indicate that Pol I does replicate plasmid sequence beyond the switch, and strongly suggest that Pol I shows little strand preference in this area of sequence. An average above 1 (1.3) may indicate a slight preference for leading-strand synthesis but this observation doesn’t detract from the conclusion that both strands are being replicated in this area. Finally, we found another sharp change in the leading vs. lagging-strand mutation ratios at the 3’ end of plasmid replication, going from close to 1 (no strand preference) to negative (predominantly lagging; Fig. 2b, white arrows). This new switch in strand specificity is likely associated with replication termination and points to a special role of Pol I in lagging-strand synthesis in this area (see Discussion).
Fig. 4. Sequence coverage. The sequence coverage for the two libraries presented here is shown on circular representation of the plasmid showing ORFs, Col E1 plasmid origin of replication, and point of replication initiation a. GFP library. b. ALKBH1 library.
Fig. 5. Comparison of leading-strand vs. lagging-strand marker mutations. Ratios of leading- versus lagging-strand mutations (as defined in Methods) are shown for 50 nucleotide intervals at increasing distance from replication initiation. Only intervals with at least 10 mutations are shown and hotspots, defined as areas with 5 or more mutations in the same or contiguous positions (Methods) were removed. On the X-axis the number means the end of the interval, so “50” means 0-50 sequence interval, and “-200” means -250 to -200 sequence interval. The location of DNA replication initiation and switch are indicated with a black and a white inverted triangle, respectively. Areas of high bias for leading-strand replication are highlighted with dark grey arrows, and high bias for lagging-strand replication with white arrows a. GFP libraries. Both liquid and solid plate libraries are included and mutation hotspots were removed prior to the analysis (methods). The following 3 (out of 21) intervals comprised fewer than 10 mutations and were excluded from the analysis as unrepresentative: 470-520, 1170-1220 and 1220-1270. b ALKBH1 library. c. Mutation frequency, expressed as number of point mutations per $10^3$ nucleotides (in logarithmic scale), is shown for the solid plate GFP library. Leading-strand mutations are shown as grey circles, and lagging-strand mutations as white circles. The proposed location of the switch is indicated with an inverted white triangle.
a

![Graph a](image)

Switch

![Graph b](image)

Replication initiation

Switch

Distance from DNA replication initiation (nucleotides)
Mutation frequency (mutations per 10^3 nucleotides)

Distance from DNA replication initiation (nucleotides)

Switch
Given a representative sample size, if both strands are replicated with comparable frequency, each mutation of a complementary pair should be equally represented (Fig. 4). The strong bias for marker Pol I leading-strand mutations upstream of the polymerase switch suggests that in this particular section of sequence Pol I plays little or no role in lagging-strand replication. Therefore mutations found in this section of sequence likely represent the mutation spectrum of the LF-Pol I polymerase on a single-stranded template. This inferred spectrum of LF-Pol I *in vivo* is shown in Fig. 3a, with individual mutations listed grouped by complementary pairs in order to facilitate seeing differences indicative of possible strand bias. For comparison, Fig. 3b shows the mutation spectrum reported for proofreading-deficient (Pol I exo-) mutant *in vitro* (2). The overall profile is consistent between the two error-prone Pol I mutants. The A→G/ T→C pair is the only clear exception, but this pair is a known preferred substrate for mismatch repair and therefore expected to be corrected *in vivo* in MMR-competent cells (14,21).

Note the clear differences in frequency found between the complementary pairs. In the case of LF-Pol I (Fig. 3a), the differential representation between the three most abundant mutation pairs (A→G/ T→C; C→T/ G→A; and A→T/ T→A), is dramatic: between 10- and 20-fold. This strong bias between complementary pairs means that certain mutations can now be approximated to be physical indicators of strandedness, with the most frequent mutation of the pair representing leading-strand synthesis and the other partner representing lagging-strand synthesis. This concept is illustrated in
Fig. 4 for the C→T/ G→A complementary pair. If the frequency of C→T mutations introduced by the polymerase is much higher than that of G→A mutations, and the sequenced strand is the leading strand, C→T mutations can be approximated to indicate leading-strand synthesis and G→A mutations to indicate lagging-strand synthesis. The same applies to other complementary pairs with strong differences in frequency. We reasoned that since the mutation spectrum is largely an intrinsic property of the polymerase, we should derive a consistent mutation profile regardless of strand or section of the plasmid considered. Specifically, we made the following three predictions:

1) The leading-strand mutation spectrum should be comparable regardless of whether it is derived from mutations occurring before or after the polymerase switch.

2) The mutation spectrum derived from lagging-strand sequence should be consistent with the spectrum derived from leading-strand sequence.

3) Perturbing the mutation spectrum of the polymerase should produce consistent changes across the three sequence compartments: 1) leading strand before the polymerase switch, 2) leading-strand downstream of the switch, and 3) lagging-strand downstream of the switch.
Fig. 6. Mutation spectrum for error-prone Pol I. a Area of high leading-strand replication bias. Each complementary pair is shown on the x-axis, with the frequency of occurrence (% in logarithmic scale) on the y-axis. The mutations include GFP and hTK libraries generated in suspension culture described in (1) (n=33 and 30 mutations, respectively), and GFP and ALKBH1 libraries generated on solid plates (n=20 and 40 respectively). Hotspots, defined as areas with 5 or more mutations in the same or contiguous positions (methods) were excluded form this analysis. b. In vitro fidelity of 3’→5’ exo domain-knockout is shown for each complementary pair, as reported in (2). The frequency of mutation is expressed as percentage, in logarithmic scale to facilitate visual comparison with panel a.
Fig. 7. Rationale for determining strand specificity based on frequency of complementary mutations. C→T is shown as an example. If the error rate of the polymerase for C→T >>> G→A, leading-strand mutations appear as C→T and lagging-strand mutations as G→A. The frequency of complementary strand mutations correlates directly with the strand preference of the polymerase: C→T with the frequency of replication of leading-strand, and G→A with the frequency of replication of lagging-strand (marker for strand preference). This is in contrast to the scenario where C→T > G→A, in which case C→T is only more likely to be leading-strand but not an unambiguous indicator of strandedness, since it can also correspond to a G to A in the lagging strand. In either case the ratio of C→T vs. G→A can be used then to establish the strand preference of the polymerase. Three scenarios shown: 1) no strand preference; 2) preference for leading-strand; 3) preference for lagging-strand. Both strands are shown, and the light grey box highlights the leading (sequenced) strand. The thickness of the arrows is proportional to the mutation frequency. Dashed lines represent decreased replication preference. Mutations introduced during replication of this (non-preferred) strand are denoted in small letters.
Fig. 5a compares the mutation spectra of GFP derived from three different areas of sequence: leading before polymerase switch (white columns), leading-strand after the switch (light grey columns), and lagging-strand after the switch (dark grey columns). The results are strikingly similar in all three areas, with C→T representing around 60% of the total, A→G ~20%, A→T ~10%, and the remainder constituting <5%.

Fisher’s exact test confirms that there is no statistical difference between the spectra derived from leading before switch, leading after switch, and lagging after switch sequence (p values=0.94), thus confirming predictions #1, and #2. In the hotspots we find a different mutation profile, with a stronger predominance of C→T mutations (80%) at the expense of A→G mutations, which represent less than 10%. However, again we derive very similar numbers from either strand, further supporting the accuracy of our strand mapping (Fig. 5b).

To test prediction #3 we reasoned that growth conditions might affect the fidelity of LF-Pol I in vivo. Specifically, cells grown in structured environments such as biofilms are known to experience different forms of genotoxic stress compared to cells grown in planktonic form (3,6). We compared our new GFP library (generated using a solid plate protocol), to our original random mutant libraries generated in cultures grown in suspension in liquid media. While the overall mutation frequency was comparable in both libraries (Table 1), we did detect significant differences in the mutation spectrum: in our liquid culture libraries, C→T mutations represented only ~40% of the total (compared to 60% for cells grown as colonies), while the representation of
A→G and A→T mutations increased substantially, from 20 to 30% and from 10 to 20%, respectively. Differences in spectrum between the two libraries were confirmed statistically using the Fisher’s exact test (p=0.046). Notably, comparing liquid and solid libraries we observed consistent differences in the mutation spectrum of LF Pol I regardless of whether we looked at mutations on the leading-strand before the polymerase switch or at mutations on either strand downstream of the polymerase switch (Fisher’s exact test p value 0.74; Fig. 5b). A shift in mutation spectrum upon perturbation of polymerase fidelity in vivo that is consistent across strands and areas of sequence neatly confirms prediction #3.

Our ability to derive a consistent mutation spectrum regardless of strand or section of sequence (predictions #1,2), and observing a consistent shift in the mutation spectrum upon perturbation of replication fidelity (prediction #3) support the accuracy of our strand mapping approach. Finally, our ability to detect subtle differences in replication fidelity produced by changes in culture conditions illustrates how our mutation footprinting approach can be used as a tool to dissect genomic and physiologic variables modulating Pol I fidelity in vivo.
Fig. 8. Concordance between mutation spectra. The relative proportion of marker leading-strand mutations is shown for leading-strand sequence before functional replisome assembly at position 170 (white columns), and after assembly, with light grey columns for leading strand and dark grey columns for lagging strand. a. Plate mutagenesis protocol, excluding hotspots and OPS; n=17 (<170), n=155 (>170 leading), n=62 (>170 lagging). b Plate mutagenesis protocol, hotspots only, excluding OPS; n=69 (>170 leading), 25 (>170 lagging). c. Suspension culture mutagenesis protocol, excluding hotspots and OPS; n=26 (<170), 139 (>170 leading), 55 (>170 lagging).
Discussion

The present work is based on random mutant library data generated by LF-Pol I replication of a ColE1 plasmid. Given the high frequency of LF Pol I mutagenesis (3 orders of magnitude above spontaneous mutation levels), the ColE1 plasmid mutations we present here can be assumed to be essentially free from Pol III replication errors, from DNA damage-induced mutagenesis, or from mutations introduced by error-prone accessory factors such as Pol IV and Pol V.

To use LF-Pol I mutations as a footprint for Pol I replication, areas of unusual mutation density needed to be avoided as unlikely to be representative of random errors made by the polymerase. We identified these areas by defining a hotspot index that measures local mutation density (see Methods), shown in Suppl. Fig. 1 and finding clusters deviating from the expected distribution (Suppl. Fig. 2). Using this approach, we identified 15 mutation clusters in the GFP library and 5 in our ALKBH1 library, listed in Suppl. Table 2. Each cluster comprises a moderate number of mutations, with an average of 9 mutations per cluster and a standard deviation of 2.5. We also found an overrepresentation of GC runs and of the CCA/TA/T motif. This profile suggests that our hotspots correspond to areas of moderate sequence instability that is at least partially dependent on local sequence context. Compared to the rest of the sequence, hotspots tend to have more C→T mutations (80%, compared to 60% in non-hotspots) at the expense of A→G mutations (<10%, compared to 20%; Fig. 5a,b). The mutation profile of hotspots argues against an involvement of local
MMR deficiency, as this deficiency would be expected to lead to a relative increase in A:T→G:C mutations instead (14,21).

We also mapped the presence of Okazaki processing sites in our GFP library to distinguish mutations originating from Okazaki primer processing from mutations generated by Pol I extension. Our new analysis confirmed two of the four Okazaki processing sites in pGFPuv plasmid sequence tentatively identified in our previous study: II (at positions 559-578) and IV (at positions 1035-1045) (1). Our analysis also identified a new candidate Okazaki processing site (site III), at positions 910-926. The new site is less than 100 nt away from site IV, and may represent an alternate location for that site.

A key finding of the present work is a ~100 nt area of leading-strand that appears to be replicated exclusively by Pol I (Fig. 2, dark grey arrows). The mutation spectrum for LF-Pol I in this section of sequence is highly biased between complementary pairs, allowing us to approximate mutations to strandedness indicators. The resulting strand-specific mutation footprint of LF-Pol I is diagrammed in Fig. 6. We found three distinct areas based on the ratio of leading vs. lagging-strand mutation: 1) an area close to plasmid replication initiation (up to position ~200) with a clear predominance of leading-strand mutations; 2) most of the remaining sequence, which shows little strand bias; and 3) an area associated with replication termination, where we see a clear predominance of lagging-strand mutations.
The practical absence of Pol I lagging-strand mutations in an ~100 nt area of sequence close to replication initiation suggests that Pol III may be taking over lagging-strand replication of this section sequence. Since the Pol III HE generally replicates both strands, it is unclear whether Pol III replication here involves only the subassembly responsible for leading-strand replication, suggesting it may assemble earlier than the other core subassembly or whether the polymerase subunit (α) of Pol III is responsible for the replication of this stretch of sequence. Given this ambiguity in Fig. 6, this area has been designated as “Pol III” to distinguish it from “Pol III HE” replication.

We interpreted the transition to a more balanced template preference as corresponding to the previously-described polymerase switch where the Pol III holoenzyme takes over replication of both leading and lagging-strands. Another area of strand bias for LF-Pol I that we identified is the 3’ end of replicated sequence. We interpret this bias as indicative of a special role of Pol I during replication termination. Pol I may be involved in resolving gaps produced by asynchronous termination between the two strands. If leading-strand replication was completed first, disassembly of the Pol III HE would leave a gap in the lagging strand. Given that Pol I is a polymerase specialized in short-gap repair (16,20), a gap-filling role during termination of ColE1 plasmid replication is a very plausible mechanism to account for the observed lagging-strand synthesis preference at the 3’ end.
Downstream of the putative polymerase switch, we found that LF-Pol I replication produces mutations in both strands with comparable frequency (Fig. 2), indicating that in this area Pol I shows little strand preference. Extensive double-stranded plasmid replication by LF-Pol I likely accounts for the remarkably balanced base pair substitution profile reported for LF Pol I-generated libraries (26) despite the unbalanced spectrum reported in vitro for Pol I exo- (2) and inferred in vivo in this work. For example, on a given strand LF-polymerase makes almost exclusively C→T mutations, but since at the same time the polymerase makes a comparable number of G→A mutations when it replicates the opposite strand, a balanced representation for this complementary pair is achieved (Fig. 4). Thus, for polymerases that replicate both strands of DNA, achieving a balanced mutation spectrum only necessitates a high frequency of mutation of one nucleotide substitution for each complementary pair. It would be surprising if this strategy for achieving a balanced generation of genetic diversity hasn’t been exploited more often in nature.

The most parsimonious explanation for double-stranded replication by Pol I is the functional incorporation of Pol I into the Pol III holoenzyme. This proposition is supported by three types of arguments: (1) It is hard to envision a mechanism that would produce balanced double-strand synthesis without any coordination between synthesis of the two strands; the Pol III HE has the molecular machinery already in place for the coordinated replication of the two strands. (2) Topologically, the change in Pol I template preference coincides with the polymerase switch, suggesting that double-strand replication by Pol I and Pol III HE assembly are mechanistically linked
There are precedents for polymerase exchanges involving the replication fork during DNA replication: examples of this “polymerase tool belt” scenario include Pol II, Pol IV and Pol V (reviewed in (8,23)). While these examples involve highly localized transactions in the context of tolerization to DNA damage, the earlier observation that Pol I is essential for survival in the absence of a functional Pol III α subunit (17), is consistent with the idea that Pol I can functionally replace the polymerase subunit of Pol III during normal replication.

Based on mutation frequency, double-stranded replication by Pol I appears to occur only in a fraction of the plasmids. In the case of our GFP libraries, the mutation frequency in the 100 nucleotides immediately upstream of the switch is 2-3 fold higher than the average frequency at areas further downstream. Since this replication only involves one strand it would be expected to have only half the mutation load. This suggests that Pol I may be responsible for the double-stranded replication of 15-25% of plasmids. We ignore the functional significance of Pol I double-stranded replication since it appears to be redundant with Pol III HE replication. Since our libraries were generated in saturated cultures, the apparent partial functional replacement of Pol III by Pol I may be a component of a more general stress response of cells under conditions of starvation.

A critical test for our approach for generating a strand-specific mutation footprint was showing that we derive the same spectrum (which is likely an intrinsic property of the polymerase) from either strand. We demonstrated that this is the case both before and
after the polymerase switch and that perturbations in polymerase fidelity produce consistent changes in both strands (Fig. 5a,c). These striking results strongly support our strand-specific footprinting approach.

We found that changing culture conditions from growth in the structured environment of a colony to growth in planktonic form produces a shift in the mutation spectrum of LF-Pol I without substantially changing the mutation frequency. Mismatch repair (MMR) depletion has been reported under conditions of starvation (reviewed in (9,10)). The two culture conditions used to generate our libraries involve starvation, suggesting some level of MMR depletion is likely in both cases. However, substantial mismatch repair deficiency would lead to a dramatic increase in overall mutation frequency and to telltale changes in mutation spectrum, increasing transitions with a preference for A:T→G:C over C:G→T:A mutations (14,21). Thus, the lack of significant change in overall LF-Pol I mutation frequency comparing liquid and solid cultures and the nature of the changes in mutation spectrum argues against differences in mismatch repair functionality as significant contributors to the observed shift in mutation spectrum. This shift in mutation spectrum depending on culture conditions described here may reflect other differences in the physiological state of cells such as differences in dNTP pools, whose effect on replication fidelity in vivo has been elegantly established through ribonucleotide reductase overexpression (11).

In sum, we found a way to determine a strand-specific mutation footprint for LF-Pol I in vivo. This approach provided new insights into the interplay between Pol I and Pol
III polymerases during initiation and termination of plasmid replication. Finally, by detecting differences in the mutation spectrum of LF-Pol I comparing libraries generated in cells grown in liquid culture versus libraries generated in colonies, we illustrate how our approach can be used more generally to investigate the impact of genomic and physiologic variables on the polymerase fidelity of Pol I in vivo.
Fig. 9. **Footprint of LF-Pol I plasmid replication.** Strand-specific mutation patterns are mapped on a generic ColE1 plasmid. Both strands are shown, and the points of replication initiation and Pol III holoenzyme assembly are indicated with inverted triangles. Replication is shown as solid lines, with arrows pointing the direction of replication. Dark grey lines represent areas of no significant strand bias are attributed to replisome synthesis, and assume Pol I is part of the replisome. Black lines represent Pol I replication in areas showing significant Pol I strand bias. During replication initiation, lagging-strand synthesis is attributed to Pol III, possibly before assembly of the Pol III replisome (light grey line), beginning at the switch but not extending all the way to the point of DNA replication initiation. During termination of DNA synthesis, a bias for lagging-strand synthesis is attributed to filling of a gap left by premature disassembly of the holoenzyme.
Acknowledgements

I thank Jordan Yoder, David Alexander, Jaime Hernandez, and Yueling Loh for their contributions to this chapter. I would like to thank Rachel Karchin for recruiting JY and YL to this project through her Computational Biology class, Musgrove for careful proofreading of an early version of the manuscript and Thomas Kunkel and Roel Schaaper for critical input on the manuscript ahead of publication. This project was partially supported through a K08 award of the NIH (CA116429-04) to MC.
References:


Chapter 5:

Conclusions and future research
The projects described in this dissertation provide insight into and tools for studying DNA replication and repair.

**Project 1: Characterization of O\textsuperscript{6}-meG repair by naturally occurring MGMT mutants provides insight into MGMT structure and function**

The aim of this project was to determine if any of the naturally occurring non-synonymous polymorphisms or variants of human methyl guanine methyl transferase (MGMT) have deficiencies in the repair of O\textsuperscript{6}-meG lesions as well as to assess the relative toxicity of O\textsuperscript{6}-meG lesions to *E. coli*. First I developed an *E. coli* complementation system specific for O\textsuperscript{6}-meG lesion repair by characterizing several strains of *E. coli* with respect to their ability to resist toxicity by two alkylating agents, each with a different spectrum of lesions they induce. I found two strains of *E. coli* that I could use to assay proteins for O\textsuperscript{6}-meG lesion repair: one with a completely deleted ada gene and one that contained a point mutation in the transferase domain of ada. Then I used both systems in order to determine how well the MGMT mutants repaired O\textsuperscript{6}-meG lesions. I found that most naturally occurring MGMT mutants repair O\textsuperscript{6}-meG lesions with near wild type levels of activity. I also discovered that the arginine plug residue of MGMT, important for base stacking properties while the lesion is flipped into the active site of MGMT, could be substituted for a glutamine. While it is not unheard of for DNA modifying enzymes to contain a glutamine plug residue, this is the first example, to our knowledge, showing that one plug residue can
be substituted for another. Finally I showed that O\textsuperscript{6}-meG lesions are indeed cytotoxic to bacteria, rebutting previous data showing that O\textsuperscript{6}-meG lesions are primarily mutagenic (5).

Future research should focus on the ability of naturally occurring mutants of MGMT to resist MGMT inhibitors (such as O\textsuperscript{6}-benzylguanine) and to influence cancer risks. I believe in vitro studies of O\textsuperscript{6}-meG lesions repair and O\textsuperscript{6}-benzylguanine resistance would complement and add depth to my heterologous expression experiments. In addition I believe in vivo mammalian cell studies could help provide insight into clinical analyses of cancer risks. While performing the research for chapter 2 of my dissertation I also transfected several MGMT mutants into K562 cells, a MGMT- myeloid progenitor cell line (3). Preliminary results show that the I143V K178R polymorphism of MGMT, present in about 20% of the population grows significantly slower than wild type MGMT. Currently research in the Camps laboratory is ongoing to validate this find and to determine if the I143V K178R polymorphism does indeed have tumor suppressor properties.

**Project 2: Experimental demonstration of a tradeoff between toxicity and repair capacity that likely drove the evolution of 3MeA glycosylase inducible expression**

Initially the aim of this project was to determine residues and structures important for the recognition and repair of 3MeA lesions. To that end I tried to exploit the structurally similarities and repair differences between scMAG and
spMAG1 and to select mutants of spMAG1 that had increased 3MeA repair capabilities. Unfortunately the project could not provide insight into 3MeA repair because we were unable to evolve spMAG1 to have increased 3MeA repair. Instead we observed a clear enrichment for wild-type spMAG and hypomorphic mutants. This observation led to the third chapter of my thesis. The aim of this study changed to determine how activity levels of glycosylases affect organismal fitness. The results of this study show that glycosylase induced cellular toxicity is linked to catalytic activity and that decreasing the enzymatic activity of spMAG1 alleviates the toxic effect this protein has on cellular viability. I was able to prove this point through two different experimental systems. First I used the random mutant library I originally designed to select for increases in 3MeA repair to select more viable versions of spMAG1. It turned out that all spMAG1 mutants selected for increased viability had decreased enzymatic activity. Secondly I showed a rescue of the cytotoxic phenotype by limiting the expression of spMAG1 under an inducible promoter system. Our results suggest that inducible 3MeA glycosylases likely arose as a way to bypass the negative effects of constitutive glycosylase activity on organismal fitness.

I believe bioinformatic and phylogenic approaches detailing the evolution, overall fold structure, and substrate specificity of constitutive versus inducible glycosylases would greatly complement my experimental findings and strengthen the hypotheses put forth in chapter 3 of my dissertation. In addition I believe the idea put forward in chapter 3 that inducible genes evolved as a way to bypass the activity constraints placed on constitutively active genes should be evaluated in other
constitutive/inducible homologous gene pairs. Finally, placing spMAG1 under an alkylation specific inducible promoter may alleviate its fitness constraints and allow the gene to evolve increased activity towards 3MeA lesions.

Project 3: Roles of DNA polymerase I in leading and lagging-strand replication defined by a high-resolution mutation footprint of ColE1 plasmid replication

The aim of this project was to use the mutagenic footprint left by a low fidelity DNA polymerase I to determine, with high resolution, how Pol I is involved with ColE1 plasmid replication. This study set out to answer three specific questions: 1. To determine how long the initial transcript synthesized by Pol I is during leading strand initiation 2. To determine how sharp the transition between Pol I and Pol III is after initial leading strand extension by Pol I and 3. To determine the size of Okazaki processing sites by Pol I. In order to do this first we determined the mutational spectrum left by our LF Pol I. Then we assigned certain mutations as signature leading strand and signature lagging strand mutations based on their proximity to the RNA/DNA switch site on the ColE1 plasmid and their relative abundance. Once we had determined signature mutations we were able to sequence a library of mutants generated by LF Pol I in order to answer our questions. We determined that the transition between Pol I and the Pol III replisome is relatively sharp and that Pol I only synthesizes about the first 200 nucleotides on the leading strand directly following the RNA/DNA switch site. This observation is much shorter than
previously published results showing that Pol I synthesizes up to the first kilobase on the leading strand but is in line with an initial biochemical report by Tomizawa, indicating a transcript size of ~100 nts (1,2). Our most surprising find and possibly the most novel was that we discovered a small fraction of plasmids are fully synthesized by Pol I without the help of Pol III. This important point could have new implications for the role of Pol I during chromosomal replication. Finally, by interrogating areas that contained a high density of signature lagging strand mutations we were able to discover sites of Okazaki processing within the first kilobase of nucleotides synthesized.

Future research on Pol I replication should be extended to the chromosome. Now that we have established and verified strandness for the mutations caused by our LF Pol I variant the Camps laboratory should take advantage of the recent advances in technology and decreases in the costs of whole genome sequencing and use deep sequencing coupled with our LF Pol I JS200 complementation system to determine, with high resolution, what sections of the chromosome are replicated by Pol I. Also by perturbing the growth conditions of *E. coli* (planktonic versus biofilm associated growth) much could be learned about how replication is altered during times of stress response. This work would have importance relevance to evolutionary adaptation. Future research should also focus on creating other low fidelity polymerases so that their role in *in vivo* replication can be studied more in depth. Research using these techniques is already occurring at NIEHS in the laboratory of Dr. Thomas Kunkel (4). Finally, the JS200 complementation system utilized in chapter 4 with great success.
should be used to characterize other polymerases that can complement the strain. Currently a graduate student in the Camps laboratory is complementing the JS200 with HIV RT in order to determine the fidelity of RT and some of its commonly found antiretroviral drug-selected variants. This experimental model could be used to characterize other viral polymerases of clinical relevance.

The future research outlines for each project would advance and expand the projects and results presented in this dissertation.
References


