Molecular Microbiology (1998) **29**(6), 1331–1339

MicroGenomics

DNA repair in *Mycobacterium tuberculosis*. What have we learnt from the genome sequence?

Valerie Mizrahi[1,2]* and Susan J. Andersen[1]
[1]*Molecular Biology Unit, South African Institute for Medical Research, PO Box 1038, Johannesburg 2000, South Africa.*
[2]*Department of Haematology, University of the Witwatersrand, Johannesburg, South Africa.*

Summary

The genome sequence of *Mycobacterium tuberculosis* was analysed by searching for homologues of genes known to be involved in the reversal or repair of DNA damage in *Escherichia coli* and related organisms. Genes necessary to perform nucleotide excision repair (NER), base excision repair (BER), recombination, and SOS repair and mutagenesis were identified. In particular, all of the genes known to be directly involved in the repair of oxidative and alkylative damage are present in *M. tuberculosis*. In contrast, we failed to identify homologues of genes involved in mismatch repair. This finding has potentially significant implications with respect to genome stability, strain variability at repeat loci and the emergence of chromosomally encoded drug resistance mutations.

Introduction

Efficient systems have evolved to repair the damage sustained by cellular genomes as a result of hydrolysis and exposure to agents that alkylate and oxidize DNA (Sancar, 1994; Seeberg *et al*., 1995). In the case of an intracellular pathogen such as *Mycobacterium tuberculosis*, the ability to repair the DNA damage caused by exposure to reactive oxygen intermediates (ROIs) produced by the phagocytic host cell is likely to play a particularly important role in ensuring the survival of the organism in this environment (Demple and Harrison, 1994). DNA repair systems are also bound to be essential for repairing lesions sustained during the dormant phase of *M. tuberculosis*. Consistent with this notion is that fact that genes such as *xthA*, *dps*, *aidB*, *ada* and *recF*, which are involved in the prevention and repair of DNA damage, are induced in stationary phase (Loewen and Hengge-Aronis, 1994; Taverna and Sedgwick, 1996; Martinez and Kolter, 1997; Villarroya *et al*., 1998). In addition to the damage caused by spontaneous hydrolysis, endogenous alkylating agents known to cause DNA damage in starving bacteria (Taverna and Sedgwick, 1996; Sedgwick, 1997) might also wreak havoc on the genome during this phase, suggesting that the repair of DNA damage both during and in the early stages of exit from dormancy would be necessary for the reactivation of viable organisms. Although some progress has been made towards defining the biochemical properties of specific repair enzymes such as RecA (Davis *et al*., 1991, 1992; Colston and Davis, 1994; Kumar *et al*., 1996) and DNA polymerase I (Huberts and Mizrahi, 1995; Gordhan *et al*., 1996; Mizrahi and Huberts, 1996), and towards understanding the nature of the SOS response in mycobacteria (Durbach *et al*., 1997; Movahedzadeh *et al*., 1997; Papavinasasundaram *et al*., 1997), our knowledge of mycobacterial DNA repair systems remains in its infancy. However, as in all other areas of *M. tuberculosis* research, the recent completion of the genome sequences of a laboratory strain (H37Rv; Cole *et al*., 1998) and a clinical isolate (CSU#93) of *M. tuberculosis* by researchers at the Sanger Centre and at The Institute for Genome Research (TIGR), respectively, has created the opportunity of investigating the genetics of DNA repair in this organism at a previously unattainable level (Young, 1997). We have searched the *M. tuberculosis* genome sequence databases for homologues of genes known to be involved in DNA repair in *Escherichia coli* and other organisms (Friedberg *et al*., 1995), and in this paper we report the results of these searches and discuss the implications of our findings.

Results

The genes have been classified into three functional categories – (i) excision repair and DNA damage reversal, (ii) recombinational repair, (iii) SOS repair and mutagenesis – and the results of the database searches are shown in Tables 1–3.

Received 23 April, 1998; revised 24 June, 1998; accepted 26 June, 1998; *For correspondence. E-mail 075val@chiron.wits.ac.za; Tel. 11 489 9370; Fax 11 489 9001.

Table 1. Genes involved in DNA damage reversal and excision repair.

Gene	Function in damage reversal or excision repair	Function	*M. tuberculosis* (+/−) database location
phrB	Deoxyribopyrimidine photolyase	DR	−
ada	O^6 methylguanine DNA methyltransferase	DR and BER	+ MTCY130.02c; *ada* part of *adaA::alkA*
ogt	O^6 alkylguanine DNA alkyltransferase	DR	+ MTCY130.01c; either *ogt* or *adaB*
mutT	8-Oxo-dGTPase	Removal of damaged nt	+ MTCI65.27
dut	dUTPase		+ MTCY05A6.18c
tag1	3-Methyladenine DNA glycosylase I	BER	+ MTCI364.22
alkA	3-Methyladenine DNA glycosylase II; has BER hypoxanthine DNA glycosylase activity	+	MTCY130.02c; *alkA* part of *adaA::alkA*
mpg[a]	**3-Methylpurine DNA glycosylase**	**BER**	+ **MTCI125.10**
ung	Uracil DNA glycosylase	BER	+ MTCY349.11
mug	G:T/U mismatch-specific DNA glycosylase	BER	−
mutY	A/G-specific adenine glycosylase; has AP lyase activity; removes adenine mispaired with 8-oxoG	BER	+ MTV024.07
nth	Endonuclease III (TG DNA glycosylase); has AP lyase activity; excises oxidized pyrimidines (ring saturation and ring fragmentation derivatives of thymine)	BER	+ MTV025.022c
fpg	Formamidopyrimidine DNA glycosylase – excises 8-oxoG and FaPy lesions; has AP lyase activity and a dRPase activity	BER	+++ MTCY338.13c; also cross-matches with MTV008.20c, MTCY71.37 (probably Nei), and MTCY10D7.30c
nei	Endonuclease VIII – functionally similar to Nth, but with structurally similar to Fpg; has AP lyase activity; excises oxidized pyrimidines	BER	+ MTCY71.37; also cross-matches MTCY338.13c (probably Fpg), MTV008.20c and MTCY10D7.30c
nfi	Endonuclease V; deoxyinosine 3′-endonuclease; also cleaves DNA containing urea residues, AP sites and mismatches	BER	−
xthA	Exonuclease III (AP endonuclease)	BER	+ MTCY22G10.24c
nfo	Endonuclease IV (AP endonuclease)	BER	+ MTI376.04c
uvrA	Excinuclease; DNA-binding ATPase	NER	+ MTCY06H11.02
uvrB	Excinuclease subunit; helicase	NER	+ MTCY01B2.25
uvrC	Excinuclease subunit; nuclease	NER	+ MTCY21B4.38. A second *uvrC*-like ORF, MTCY190.02, was also identified. This ORF consists of a fusion of the ε-subunit of DNA PolIII (*dnaQ*) to the N-terminal domain of UvrC.
uvrD	Helicase II	NER/other	+ MTCY10D7.25c
recA	Regulation of NER	NER	+ MTRECA; MTV002.02c
lexA	Regulation of NER	NER	+ MTLEXA; MTCY05A6.41
mfd	Coupling of transcription and NER (TCRF)	NER	+ MTCY10G2.29c
ercc3[a]	**XPB/ERCC3 helicase (eukaryotic)**	**NER/other**	+ **MTV043.54c**
dam	A-specific methylase	MR	−
dcm	C-specific methylase	MR	−
mutH	Long-patch GATC-specific endonuclease	MR	−
mutL	Binds to heteroduplex-MutS complex	MR	−
mutS	Mismatch recognition and binding	MR	−
vsr	Short-patch repair endonuclease	MR	−
recJ	DNA deoxyribophosphodiesterase (dRPase)	−	
polA	DNA polymerase I (5′–3′ exonuclease and DNA polymerase)	Post-incision events	+ MTCY01B2.21; contains both 5′–3′ exonuclease and DNA polymerase domains. Cross-matches to MTCY49.30, which encodes a putative 5′–3′ exonuclease. Genes encoding the other major 5′ exonucleases, namely RecBCD (Table 2) and exonuclease VII (*xseA* and *xseB*), were also found in *M. tuberculosis*
polB	DNA polymerase II		−
polC	DNA polymerase III holoenzyme		++ MTCY48.18c; MTV004.28c; only the *dnaE* genes encoding the replicative subunit of PolIII are specified herein
yshC	DNA polymerase β		+ MTCY1A6.12; putative identification
ssb	SSB; single-stranded binding protein		++ MTCY21D4.17; MTV008.34c
lig	DNA ligase (NAD dependent)		+ MTV012.28c
lig[a]	**DNA ligase (ATP dependent)**		++[b] **MTCY22D7.19c, MTV025.079**

DR, damage reversal; BER, base excision repair; NER, nucleotide excision repair; MR, mismatch repair.
A plus sign denotes the presence of a putative homologue in the Sanger genome sequence database. The homologues were identified as described under 'Experimental procedures'. All genes have been located on annotated Sanger cosmid sequences in GenBank, and therefore only these descriptions are given. The same genes were found in the TIGR database (not shown). However, owing to complications arising from errors in the sequence data, which were preliminary at the time of the analysis, the TIGR ORFs were not analysed further.
a. The genes shown in bold have no homologue in *E. coli*.
b. The related, hypothetical proteins MTCY08D9.01c, MTCY06A4.13c and MTV025.078c also show homology to ATP-dependent ligases.

Table 2. Genes involved in recombination.[a]

E. coli gene	Function in genetic recombination	*M. tuberculosis* (+/−) database location
recB	Exonuclease V; ATPase, ds DNA exonuclease; chi-specific endonuclease	+ MTCY20H10.11c
recC		+ MTCY20H10.12c
recD		+ MTCY20H10.10c
recE	5′–3′ ds DNA exonuclease (exonuclease VIII)	−
recT	DNA renaturation	−
recJ	5′–3′ ss DNA exonuclease	−
recQ	Helicase	[b] Closest match is RNA helicase (MTCY50.29c)
recN	Unknown; ATP-binding consensus sequence	+ MTCI125.18
recA	ATPase and strand exchange	+ MTRECA; MTV002.02c
radA	Recombination	+ MTV024.03
recF	ss DNA binding protein; modulates assembly and disassembly of RecA filaments (with RecO and RecR)	+ MTCY10H4.01
recO	Modulates assembly and disassembly of RecA filaments (with RecF and RecR)	[c] Closest match is MTCY27.18; 28% identity over 156 aa overlap
recR	Modulates assembly and disassembly of RecA filaments (with RecF and RecO)	+ MTV025.063c
ssb	Single-stranded binding protein	++ MTCY21D4.17; MTV008.34c
ruvA	Binds to Holliday junctions	+ MTCY227.08
ruvB	Branch migration; ATPase	+ MTCY227.09
ruvC	Endonuclease; resolves Holliday junctions	+ MTCY227.07
recG	ATPase; promotes branch migration	+ MTCY349.14
sbcB	Exonuclease I	−
sbcC	ATP-dependent ds DNA exonuclease	−
sbcD		+ Closest match is MTCY50.05c
topA	Topoisomerase I	+ MTCY15C10.06
gyrA	DNA gyrase	+ MSGGYRAB; MTCY10H4.04
gyrB	DNA gyrase	+ MSGGYRAB; MTCY10H4.03
polA	DNA polymerase I	+ MSGPOLA; MTCY01B2.21
ligA	DNA ligase	+ MTV012.28c
uvrD	Helicase II	+ MTCY10D7.25c
helD	Helicase IV	[b] Closest match is UvrD (MTCY10D7.25c)

a. Taken from Kowalczykowski *et al.* (1994).
b. Although convincing homologues of *recQ* and *helD* were not identified, three co-directional, putative ATP-dependent DNA helicases were found on an 11.5 kb region of the genome (MTCY07D11.24, MTV014.45c and MTV014.42c).
c. WU-BLAST produced a *P*-value of 0.054 for the search using *E. coli* RecO as the query sequence. The advanced gapped BLASTP program, in searching MTCY27.18 against the NR database, produced a *P*-value of 0.0012 for *S. typhimurium* RecO, which is indicative of homology.

Excision repair and DNA damage reversal

Mismatch repair. Prokaryotic and eukaryotic cells share the ability to repair mismatched basepairs, which may arise in several ways, such as by the introduction of errors during DNA replication and by the formation of a heteroduplex between homeologous DNA strands during a recombinational event (Friedberg *et al.*, 1995; Modrich and Lahue, 1996). This pathway plays a critically important role in mutation avoidance and genome stabilization.

Long-patch mismatch repair. Detailed studies of methyl-directed mismatch repair in *E. coli* and *Streptococcus pneumoniae* have provided mechanistic insights into this highly conserved system that targets repair to the daughter strand of DNA (Modrich and Lahue, 1996). At the core of this system are MutS and MutL, which initiate repair by sequential binding to the mismatch (Grilley *et al.*, 1989). Interestingly, we failed to identify homologues of the *mutS*, *mutL*, *mutH* and *dam* genes in *M. tuberculosis*. The absence of *mutS* and *mutL* is particularly noteworthy as these genes are highly conserved in bacteria, yeast, mice and human cells (Modrich and Lahue, 1996) and are widely distributed in prokaryotes (Koonin and Galperin, 1997). However, the absence of these genes from the 'minimal' bacterial gene complement deduced from *Mycoplasma* (Fraser *et al.*, 1995) suggests that this mismatch repair system is nonetheless dispensable.

Other possible mechanisms for mismatch repair. *M. tuberculosis* also lacks the *dcm* and *vsr* genes, which participate in short-patch mismatch repair in *E. coli*. However, a *mutY* homologue was identified, suggesting that the encoded glycosylase may repair A•G and A•C mismatches by adenine excision. The UvrABC excinuclease (see below) might also excise mismatches from one or other strand (Huang *et al.*, 1994), but it is important emphasize that mismatch repair via these mechanisms could lead to *mutation fixation* rather than *mutation avoidance* by virtue of the lack of parental- versus daughter strand discrimination.

Table 3. Genes involved in SOS repair and mutagenesis.[a]

E. coli gene	Induced physiological responses and/or function	*M. tuberculosis* (+/−) database location
lexA	Regulator of SOS response	+ MTLEXA; MTCY05A6.41
recA	Regulator of SOS response and recombination	+ MTRECA; MTV002.02c
recN	Double-stranded break repair	+ MTCI125.18
ruvA	Daughter strand gap repair	+ MTCY227.08
ruvB		+ MTCY227.09
uvrA	Excision repair and long-patch repair	+ MTCY06H11.02
uvrB	Excision repair and long-patch repair	+ MTCY01B2.25
uvrD	Excision repair; helicase II	+ MTCY10D7.25c
dinP	Mutagenesis	+ MTCY22D7.25c; related to *B. subtilis uvrX*, *yqjH* and *yqjW* and cross-matches with MTCY48.28c (probably *umuC*)
umuC	Translesion by-pass (mutagenesis)	+ MTCY48.28c; related to *B. subtilis uvrX*, *yqjH* and *yqjW* and cross-matches with MTC22D7.25c (probably *dinP*)
umuD	Translesion by-pass (mutagenesis)	−
polB	DNA polymerase II	−
dinG	Helicase	+ MTCY130.14c

a. Taken from Friedberg *et al.* (1995). The only genes included in this analysis are those known to be *directly* involved in SOS repair and mutagenesis. Other SOS-inducible genes such as *sulA*, *dinD*, *dinH*, *dinF* and *dinI* have been excluded, as they are beyond the scope of this work.

Does *M. tuberculosis* possess a non-orthologous mismatch repair system? Although mutS has been identified in both *Methanobacterium methanoautotrophicum* (Smith *et al.*, 1997) and *Pyrococcus horikoshii* (URL:http://www.bio.nite.go.jp/BLAST/blast.html), other archaeons such as *Methanococcus jannaschii* (Koonin and Galperin, 1997) and *Archaeoglobus fulgidus* (Klenk *et al.*, 1997) appear to lack the *mutS* and *mutL* genes. The fact that *Sulfolobus acidocaldarius* has a normal spontaneous mutation rate (Jacobs and Grogan, 1997) implies that non-orthologous displacement of the MutHLS mismatch repair system might have occurred in at least some archaeal species (Koonin *et al.*, 1996). By analogy, it is therefore conceivable that *M. tuberculosis* possesses an alternative, unidentified mismatch repair system. The observation that the mutation rate for *M. tuberculosis* growing *in vitro* (David and Newman, 1971) is normal would appear to support this idea on the basis of the mutator phenotype of mismatch repair-defective mutants of *E. coli* and related organisms (Cox, 1976; LeClerc *et al.*, 1996). Indeed, the normal mutation rate has been used to explain the striking lack of synonomous substitutions in structural genes of *M. tuberculosis* by arguing that this organism is evolutionarily young (Kapur *et al.*, 1994; Sreevatsan *et al.*, 1997). However, if the fidelity of DNA replication in a slow-growing organism such as *M. tuberculosis* is significantly higher than that observed in *E. coli*, the need for post-replicative mismatch repair and hence its relative contribution to the overall fidelity of chromosomal replication would be reduced. Therefore, in the absence of experimental data on the fidelity of replication, the mutation rate, *per se*, does not necessarily constitute proof of a non-orthologous mismatch repair system in *M. tuberculosis*.

Is *M. tuberculosis* heritably mismatch repair defective? In our view, this alternative hypothesis should be considered on the basis of the following observations. (i) An unexpectedly high frequency of mismatch repair-defective mutators was found in pathogenic strains of *E. coli* and *Salmonella* (LeClerc *et al.*, 1996), implying that such a phenotype allows the pathogen to readily adapt to its environment (Taddei *et al.*, 1997). (ii) Studies of adaptive mutagenesis in *E. coli* (reviewed by Rosenberg, 1997) have reinforced the view that mismatch repair-defective mutator activity might be advantageous under conditions of stress, such as those encountered during stationary phase, nutrient starvation, immune surveillance and therapeutic intervention, all of which apply to *M. tuberculosis*. Therefore, adaptive mutagenesis, which occurs in non-dividing or slowly growing cells under selection for a particular phenotype, may play an important role in the adaptation of *M. tuberculosis* to its environment, particularly if the organism is a defective mismatch repair. First, the genes required for recombination-dependent adaptive mutagenesis are present in *M. tuberculosis* (see below). Second, DNA lesions such as double-stranded breaks and oxidative damage required to initiate this and other stationary-phase mutational pathways (Bridges, 1997; Bridges and Timms, 1997; Rosenberg, 1997) are probably formed in this organism (McFadden, 1996). Thirdly, all known drug resistance mutations in *M. tuberculosis* are chromosomally encoded (Cole and Telenti, 1995; Parsons *et al.*, 1997), suggesting that mutations conferring resistance are readily fixed in the genome of this organism by direct drug selection. As stationary-phase mutations have been shown to include base substitutions, frameshifts and small deletions, it is possible that at least some of the drug-resistant mutations of this type observed in the *katG* (Zhang and Young, 1994) and *pncA* (Scorpio *et al.*, 1997) genes of *M. tuberculosis* might have arisen via adaptive mutagenesis. We also note that the potential clinical relevance of adaptive mutagenesis was recently confirmed for ciprofloxacin resistance in

non-dividing *E. coli* cells (Riesenfeld *et al*., 1997). (iii) In addition to its role in post-replicative error correction, the mismatch repair system also provides a barrier to recombination between partially diverged ('homeologous') sequences (Matic *et al*., 1996). The lack of a mismatch repair system might therefore facilitate intrachromosomal exchanges between tandem-repeated, interspersed and homeologous sequences. The ubiquitous repeats found in the genome of *M. tuberculosis* include the direct repeat (DR) sequence, the GC-rich repetitive sequence found in genes encoding the PGRS group of Gly Ala-rich, PE family proteins, the MPTR repeats occurring in genes encoding PPE family, Gly Asn-rich proteins, MIRUs (Supply *et al*., 1997; Magdalena *et al*., 1998), as well as numerous exact tandem repeats (ETR) (Groenen *et al*., 1993; van Soolingen *et al*., 1993; Zhang and Young, 1994; Poulet and Cole, 1995a,b; Frothingham and Meeker-O'Connell, 1998). In light of the high frequency at which polymorphic and exact repeats occur in the genome of *M. tuberculosis*, the relaxation in recombinational control between homeologous sequences and the inability to correct polymerase slippage errors on templates containing simple repeats resulting from a lack of mismatch repair may provide a possible mechanism for generating strain variation that is distinct from that attributable to the movement of mobile elements. In this respect, a genome-wide comparison between H37Rv and CSU#93 focused on the similarities and differences in repeat sequences should provide an important adjunct to the information on strain variability at repeat loci acquired from molecular epidemiological studies (Frothingham and Meeker-O'Connell, 1998; Magdalena *et al*., 1998).

Base excision repair. The BER pathway is primarily responsible for correcting damage caused by hydrolysis and exposure to ROIs and other metabolites that oxidize and alkylate DNA (Lindahl, 1993; Sancar, 1994; Seeberg *et al*., 1995). Endogenous agents such as S-adenosylmethionine are capable of non-enzymatically alkylating the ring nitrogens of purine residues to form 7-MeG and 3-MeA, of which the latter is a particularly important cytotoxic lesion that blocks replication and hence is a threat to viability. Three distinct 3-MeA-glycosylase-encoding genes were found, of which *alkA* may be alkylation damage inducible. The genome sequence also revealed a number of genes involved in the repair of oxidized pyrimidines (*nei* and *nth*) and purines (*fpg*). The identification of *mutY*, *mutT* and *fpg* genes suggest that the 'GO system' (Michaels and Miller, 1992) for protection against the starvation-associated mutagenic effects of 8-oxo-G (Bridges *et al*., 1996) is intact in *M. tuberculosis*. Finally, two AP-endonuclease-encoding genes, both of which belong to oxidative defence regulons in *E. coli* (Demple, 1991), were found in *M. tuberculosis*, as were genes encoding PolI, PolIII and DNA ligases.

Nucleotide excision repair. Genes encoding the UvrABC excinuclease (Sancar, 1994) were identified in *M. tuberculosis*, as were *uvrD* and *polA* (Huberts and Mizrahi, 1995; Mizrahi and Huberts, 1996). This organism also contains the XPB/ERCC3-like helicase identified in *M. leprae* (Poterszman *et al*., 1997) and although the *mfd* gene was identified, the lack of *mutS* and *mutL* might adversely affect the efficiency of transcription-coupled NER (Mellon and Champe, 1996).

Reversal of DNA damage. The *phrB (phr)* gene was not found in *M. tuberculosis*. Moreover, a homologue of the *B. subtilis spl* gene (Fajardo-Cavazos *et al*., 1993) was not identified, suggesting that *M. tuberculosis* relies on NER for the repair of UV-induced damage. However, an operon encoding proteins involved in the reversal of alkylation damage was identified. As discussed above, nitrosated compounds represent a potentially important endogenous source of O^6MeG or O^4MeT in dormant tubercle bacilli. The notion that such compounds might be formed in *M. tuberculosis* is consistent with the fact that genes encoding the nitrate reductase A and molybdopterin cofactor biosynthetic proteins required for their formation (Taverna and Sedgwick, 1996) are present in its genome.

Recombinational repair

The *M. tuberculosis* genes involved in these processes are shown in Tables 2 and 3. The unusual structure (and potential implications) of the *M. tuberculosis recA* gene, which plays a central role in recombination, damage tolerance and induction of the SOS response, has been described in detail elsewhere (Davis *et al*., 1991, 1992; Colston and Davis, 1994; McFadden, 1996) and is therefore not discussed herein. The RecBCD pathway appears to be intact, suggesting that *M. tuberculosis* possesses the enzymatic machinery required for double-stranded break repair and conjugal transfer of DNA, although the latter process has yet to be demonstrated in *M. tuberculosis* (Parsons *et al*., 1998). In contrast, the absence of *recJ* suggests that the RecF pathway is defective (Kowalsczykowski *et al*., 1994; Tseng *et al*., 1994). We note that these conclusions on recombinational pathways are in close agreement with those reached by McFadden (1996).

SOS repair and mutagenesis

M. tuberculosis has been shown to possess the regulatory elements of a functional SOS system, and with the exception of *polB* and *umuD*, contains all of the remaining genes associated with SOS repair and mutagenesis in *E. coli*. Although considerable progress has been made towards understanding the sequence requirements for DNA binding by *M. tuberculosis* LexA (Durbach *et al*., 1997; Movahedzadeh *et al*., 1997; Dullaghan and Davis, 1998), a

consensus sequence has yet to be established. For this reason, a systematic scan of the genome sequence for potential LexA binding sites was not included in this study. However, we note that at least one gene that is not SOS inducible in *E. coli* (*ruvC*) appears to be LexA-regulated in *M. tuberculosis* (Colston *et al*., 1998).

Discussion

This study has provided a useful framework upon which future experimental studies on DNA repair in *M. tuberculosis* could be based. The most significant finding was the lack of a MutS-based mismatch repair system, which has potentially important implications with respect to genome stability, strain variability and the occurrence of chromosomally encoded drug resistance mutations. In this respect, the most important questions that warrant experimental investigation are as follows. (i) Is the requirement for mismatch repair of post-replicative errors obviated by an unusually high fidelity of DNA replication in *M. tuberculosis*? (ii) Does the lack of a mismatch repair-mediated correction of polymerase slippage errors provide a mechanism for generating sequence variation at repeat sequences? (iii) Does adaptive mutagenesis play a role in the emergence of drug resistance mutations? This study has also underscored the need to study the role of DNA repair in the dormant phase and during the exit from dormancy. In one possible scenario, the capacity for completing excision repair and recombinational repair by polymerization (gap-filling) may be severely limited by low dNTP pools in dormant (non-replicating) bacteria, in which case AP sites, single-stranded gaps and double-stranded breaks would accumulate. For dormant *M. tuberculosis* to reactivate in a viable form, the repair of such lesions would constitute an obligate step in the early stages of exit from dormancy before, or in synchrony with, the resumption of chromosomal DNA replication. Recombinational repair would obviously be facilitated at this stage if persistent bacilli indeed exist in a diploid state, as was recently suggested by Parrish *et al.* (1998). Experimental testing of this and other models may yield important insights relevant to the development of novel drugs for the treatment of dormant tubercle bacilli.

Experimental procedures

M. tuberculosis *nucleotide sequence databases*

The nucleotide sequence of the H37Rv laboratory strain of *M. tuberculosis* was from the Sanger Centre (http://www.sanger.ac.uk/Projects/M_tuberculosis/blast_server.shtml; contiguous sequence database (TB.seq) and the database containing all annotated cosmids and other sections that have been submitted to EMBL). These sequence data were produced by the *M. tuberculosis* Sequencing Group at the Sanger Centre and can be obtained from ftp://ftp.sanger.ac.uk/pub/tb/sequences (Cole *et al*., 1998). The unfinished nucleotide sequence of the CSU#93 clinical isolate of *M. tuberculosis* was accessed via the TIGR BLAST server located at the National Center for Biotechnology Information (NCBI; Bethesda, MD, USA; http://www.ncbi.nlm.gov/BLAST/tigrbl.html; 370 contigs). These preliminary sequence data were produced by the *Mycobacterium* Sequencing Group at TIGR.

Database searches and protein sequence comparisons

Searches against the six reading frame translation of the *M. tuberculosis* nucleotide sequence databases located at the Sanger Centre were performed at the *M. tuberculosis* BLAST server using the TBLASTN executable of the WU-BLAST program, which is based on the BLAST2 algorithm (Altschul and Gish, 1996). Searches against the six reading frame translation of the TIGR *M. tuberculosis* nucleotide sequence database located at the NCBI were performed using the TBLASTN executable of the gapped BLAST program (Altschul *et al*., 1997). The query used for the BLAST search was the sequence of the corresponding protein from *E. coli* and/or *B. subtilis* (obtained from GenBank). All query sequences were filtered for regions of low compositional complexity before the database search using the SEG program (Wootton and Federhen, 1996). Under these conditions, a P-value ≤ 0.001 produced by WU-BLAST was considered to be indicative of homology (Koonin *et al*., 1997). To confirm the assignment of the *M. tuberculosis* open reading frames (ORFs) thus identified, those displaying the greatest similarity to the query sequence were further analysed by performing advanced BLASTP searches against the protein version of the non-redundant (NR) database maintained at the NCBI, as well as BLASTP + BEAUTY searches against the NR database resident on the Baylor College of Medicine BLAST + BEAUTY server, using the gapped BLAST (Altschul *et al*., 1997) and BEAUTY programs (Worley *et al*., 1995; Smith *et al*., 1996) (http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html). For cases in which searches of the *M. tuberculosis* databases with the full-length query sequence(s) indicated potential absence of the encoding gene (P-value >0.001 from WU-BLAST search), additional advanced BLASTP and TBLASTN searches with performed, wherever possible, using conserved motif query sequences that were characteristic of the protein in question, and which were obtained from the PROSITE (Bairoch *et al*., 1997) and BLOCKS databases (http://www.seqnet.dl.ac.uk). If these searches yielded no consistent indication of homology, the gene was scored as absent. Multiple sequence alignments were carried out using the MEGALIGN program (DNASTAR).

Acknowledgements

The authors gratefully acknowledge the financial support of the Glaxo Wellcome *Action TB* Initiative, the South African Medical Research Council, and the South African Institute for Medical Research Foundation. We are deeply indebted to Stewart Cole and to Bart Barrell, Rob Fleischmann and their colleagues at the Sanger Centre and TIGR for determining, and making publicly available, the genome sequences upon which this work was based. We are most grateful to Steven Durbach, Selwyn Quan, Stephanie Dawes, Helena Boshoff and other members of the Mizrahi laboratory for

assistance with the database searches and for stimulating discussions. We thank Erik Böttger, Susan Rosenberg, Philip Supply, Harvey Rubin, Paul van Helden, Lafras Steyn, Ruth McAdam, Ken Duncan and Keith Klugman for constructive comments and one of us (V.M.) is particularly grateful to Julian Davies for providing the provocative ideas which prompted this study.

References

Altschul, S.F., and Gish, W. (1996) Local alignment statistics. *Methods Enzymol* **266:** 460–480.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Bairoch, A., Bucher, P., and Hofmann, K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res* **25:** 217–221.

Bridges, B.A. (1997) Hypermutation under stress. *Nature* **387:** 557–558.

Bridges, B.A., Sekiguchi, M., and Tajiri, T. (1996) Effect of *mutY* and *mutM*/*fpg-1* mutations on starvation-associated mutation in *Escherichia coli*: implications for the role of 7,8-dihydro-8-oxoguanine. *Mol Gen Genet* **251:** 352–357.

Bridges, B.A., and Timms, A.R. (1997) Mutation in *Escherichia coli* under starvation conditions: a new pathway leading to small deletions in strains defective in mismatch correction. *EMBO J* **16:** 3349–3356.

Cole, S.T., and Telenti, A. (1995) Drug resistance in *Mycobacterium tuberculosis*. *Eur Respir J* **20** (Suppl.): 701S–713S.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., *et al*. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393:** 537–544.

Colston, M.J., and Davis, E.O. (1994) Homologous recombination, DNA repair, and mycobacterial *recA* genes. In: *Tuberculosis. Pathogenesis*, *Protection* and *Control*. Bloom, B.R. (ed.). Washington, DC: American Society for Microbiology Press. pp. 217–226.

Colston, M.J., Davis, E.O., Movahedzadeh, F., Papavinasasundaram, K.G., Thomas, N., Jenner, P., and Bachrach, G. (1998) The mycobacterial response to DNA damage and intracellular stress. Keystone Symposium on TB: Molecular Mechanisms and Immunologic Aspects. Abstract Y1: 007.

Cox, E.C. (1976) Bacterial mutator genes and the control of spontaneous mutation. *Annu Rev Genet* **10:** 135–156.

David, H.L., and Newman, C.M. (1971) Some observations on the genetics of isoniazid resistance in the tubercle bacilli. *Am Rev Respir Dis* **104:** 508–515.

Davis, E.O., Sedgwick., S.G., and Colston, M.J. (1991) Novel structure of the *recA* locus of *Mycobacterium tuberculosis* implies processing of the gene product. *J Bacteriol* **173:** 5653–5662.

Davis, E.O., Jenner, P.J., Brooks, P.C., Colston, M.J., and Sedgwick, S.G. (1992) Protein splicing in the maturation of *M. tuberculosis* RecA protein: a mechanism for tolerating a novel class of intervening sequence. *Cell* **71:** 201–210.

Demple, B. (1991) Regulation of bacterial oxidative stress genes. *Annu Rev Genet* **25:** 315–337.

Demple, B., and Harrison, L. (1994) Repair of oxidative damage to DNA: enzymology and biology. *Annu Rev Biochem* **63:** 915–948.

Dullaghan, E.M., and Davis, E.O. (1998) Sequence requirements for *Mycobacterium tuberculosis* LexA binding. Keystone Symposium on TB: Molecular Mechanisms and Immunologic Aspects. Abstract Y1: 412.

Durbach, S.I., Andersen, S.J., and Mizrahi, V. (1997) SOS induction in mycobacteria: analysis of the DNA-binding activity of a LexA-like repressor and its role in DNA damage induction of the *recA* gene from *Mycobacterium smegmatis*. *Mol Microbiol* **26:** 643–653.

Fajardo-Cavazos, P., Salazar, C., and Nicholson, W.L. (1993) Molecular cloning and characterization of the *Bacillus subtilis* spore photoproduct photolyase (*spl*) gene, which is involved in repair of UV-irradiation-induced DNA damage during spore germination. *J Bacteriol* **175:** 1735–1744.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., *et al*. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397–403.

Friedberg, E.C., Walker, G.C., and Siede, W. (1995) *DNA Repair and Mutagenesis*. Washington, DC: American Society for Microbiology Press.

Frothingham, R., and Meeker-O'Connell, W.A. (1998) Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. *Microbiology* **144:** 1189–1196.

Gordhan, B.G., De Meyer, A.R., Andersen, S.J., and Mizrahi, V. (1996) Construction by homologous recombination and phenotypic characterisation of a DNA polymerase domain *polA* mutant of *Mycobacterium smegmatis*. *Gene* **178:** 125–130.

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

Groenen, P.M., Bunschoten, A.E., van Soolingen, D., and van Embden, J.D. (1993) Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol Microbiol* **10:** 1057–1065.

Huang, J.C., Hsu, D.S., Kazantsev, A., and Sancar, A. (1994) Substrate spectrum of human excinuclease: repair of abasic sites, methylated bases, mismatches, and bulky adducts. *Proc Natl Acad Sci USA* **91:** 12213–12217.

Huberts, P., and Mizrahi, V. (1995) Cloning and sequence analysis of the gene encoding the DNA polymerase I from *Mycobacterium tuberculosis*. *Gene* **164:** 133–136.

Jacobs, K.L., and Grogan, D.W. (1997) Rates of spontaneous mutation in an archaeon from geothermal environments. *J Bacteriol* **179:** 3298–3303.

Kapur, V., Whittam, T.S., and Musser, J.M. (1994) Is *Mycobacterium tuberculosis* 15,000 years old? *J Infect Dis* **170:** 1348–1349.

Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., *et al*. (1997) The complete genome sequence of the hyperthermophilic, sulfate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390:** 364–370.

Koonin, E.V., and Galperin, M.Y. (1997) Prokaryotic genomes:

the emerging paradigm of genome-based microbiology. *Curr Opin Genet Dev* **7:** 757–763.

Koonin, E.V., Mushegian, A.R., and Bork, P. (1996) Non-orthologous gene displacement. *Trends Genet* **12:** 334–336.

Koonin, E.V., Mushegian, A.R., Galperin, M.Y., and Walker, D.R. (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol Microbiol* **25:** 619–637.

Kowalczykowski, S.C., Dixon, D.A., Eggleston, A.K., Lauder, S.D., and Rehrauer, W.M. (1994) Biochemistry of homologous recombination in *Escherichia coli*. *Microbiol Rev* **58:** 401–465.

Kumar, R.A., Vaze, M.B., Chandra, N.R., Vijayan, M., and Muniyappa, K. (1996) Functional characterization of the precursor and spliced forms of RecA protein of *Mycobacterium tuberculosis*. *Biochemistry* **35:** 1793–1802.

LeClerc, J.E., Li, B., Payne, W.L., and Cebula, T.A. (1996) High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science* **274:** 1208–1211.

Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature* **362:** 709–715.

Loewen, P.C., and Hengge-Aronis, R. (1994) The role of the sigman factor σ^S (KatF) in bacterial global regulation. *Annu Rev Microbiol* **48:** 53–80.

McFadden, J. (1996) Recombination in mycobacteria. *Mol Microbiol* **21:** 205–211.

Magdalena, J., Vachee, A., Supply, P., and Locht, C. (1998) Identification of a new DNA region specific for members of *Mycobacterium tuberculosis* complex. *J Clin Microbiol* **36:** 937–943.

Martinez, A., and Kolter, R. (1997) Protection of DNA during oxidative stress by the non-specific DNA-binding protein Dps. *J Bacteriol* **179:** 5188–5194.

Matic, I., Taddei, F., and Radman, M. (1996) Genetic barriers among bacteria. *Trends Microbiol* **4:** 69–72.

Mellon, I., and Champe, G.N. (1996) Products of DNA mismatch repair genes *mutS* and *mutL* are required for transcription-coupled nucleotide-excision repair of the lactose operon in *Escherichia coli*. *Proc Natl Acad Sci USA* **93:** 1292–1297.

Michaels, M.L., and Miller, J.H. (1992) The GO system protects organisms from the mutagenic effects of the spontaneous lesion, 8-hydroxyguanine (7,8-dihydro-8-oxoguanine). *J Bacteriol* **174:** 6321–6325.

Mizrahi, V., and Huberts, P. (1996) Deoxy- and dideoxynucleotide discrimination and identification of critical 5′ nuclease domain residues of the DNA polymerase I from *Mycobacterium tuberculosis*. *Nucleic Acids Res* **24:** 4845–4852.

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

Movahedzadeh, F., Colston, M.J., and Davis, E.O. (1997) Characterization of *Mycobacterium tuberculosis* LexA: recognition of a Cheo (*Bacillus*-type SOS) box. *Microbiology* **143:** 929–936.

Papavinasasundaram, K.G., Movahedzadeh, F., Keer, J.T., Stoker, N.G., Colston, M.J., and Davis, E.O. (1997) Mycobacterial *recA* is cotranscribed with a potential regulatory gene called *recX*. *Mol Microbiol* **24:** 141–153.

Parrish, N.M., Dick, J.D., and Bishai, W.R. (1998) Mechanisms of latency in *Mycobacterium tuberculosis*. *Trends Microbiol* **6:** 107–112.

Parsons, L.M., Driscoll, J.R., Taber, H.W., and Salfinger, M. (1997) Drug resistance in tuberculosis. *Infect Dis Clin N Am* **11:** 905–928.

Parsons, L.M., Jankowski, C.S., and Derbyshire, K.M. (1998) Conjugal transfer of chromosomal DNA in *Mycobacterium smegmatis*. *Mol Microbiol* **28:** 571–582.

Poterszman, A., Lamour, V., Egly, J.M., Moras, D., Thierry, J.C., and Poch, O. (1997) A eukaryotic XPB/ERCC3-like helicase in *Mycobacterium leprae*? *Trends Biochem Sci* **2:** 418–419.

Poulet, S., and Cole, S.T. (1995a) Repeated DNA sequences in mycobacteria. *Arch Microbiol* **163:** 79–86.

Poulet, S., and Cole, S.T. (1995b) Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*. *Arch Microbiol* **163:** 87–95.

Riesenfeld, C., Everett, M., Piddock, L.J., and Hall, B.G. (1997) Adaptive mutations produce resistance to ciprofloxacin. *Antimicrob Agents Chemother* **41:** 2059–2060.

Rosenberg, S.M. (1997) Mutation for survival. *Curr Opin Genet Dev* **7:** 829–834.

Sancar, A. (1994) Mechanisms of DNA excision repair. *Science* **266**: 1954–1956.

Sedgwick, B. (1997) Nitrosated peptides and polyamines as endogenous mutagens in O^6-alkylguanine-DNA alkyltranferase deficient cells. *Carcinogenesis* **18:** 1561–1567.

Scorpio, A., Lindholm-Levy, P., Heifets, L., Gilman, R., Siddiqi, S., Cynamon, M., and Zhang, Y. (1997) Characterization of *pncA* mutations in pyrazinamide-resistant *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* **41:** 540–543.

Seeberg, E., Eide, L., and Bjoras, M. (1995) The base excision repair pathway. *Trends Biochem Sci* **20:** 391–397.

Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J.J., Aldredge, T., *et al.* (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J Bacteriol* **179:** 7135–7155.

Smith, R.F., Wiese, B.A., Wojzynski., M.K., Davison, D.B., and Worley, K.C. (1996) BCM Search Launcher: an integrated interface to molecular biology database search and analysis services available on the World Wide Web. *Genome Res* **6:** 454–462.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., and Musser, J.M. (1997) Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA* **94:** 9869–9874.

Supply, P., Magdalena, J., Himpens, S., and Locht, C. (1997) Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol Microbiol* **26:** 991–1003.

Taddei, F., Matic, I., Godelle, B., and Radman, M. (1997) To be a mutator, or how pathogenic and commensal bacteria can evolve rapidly. *Trends Microbiol* **5:** 427–428.

Taverna, P., and Sedgwick, B. (1996) Generation of an endogenous DNA-methylating agent by nitrosation in *Escherichia coli*. *J Bacteriol* **178:** 5105–5111.

Tseng, Y.C., Hung, J.L., and Wang, T.C. (1994) Involvement of RecF pathway recombination genes in postreplication repair in UV-irradiated *Escherichia coli* cells. *Mutat Res* **315:** 1–9.

van Soolingen, D., de Haas, P.E., Hermans, P.W., Groenen, P.M., and van Embden, J.D. (1993) Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of *Mycobacterium tuberculosis*. *J Clin Microbiol* **31:** 1987–95.

Villarroya, M., Pérez-Roger, I., Macián, F., and Armengod, M.E. (1998) Stationary phase induction of *dnaN* and *recF*, two genes of *Escherichia coli* involved in DNA replication and repair. *EMBO J* **17:** 1829–1837.

Wootton, J.C., and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266:** 554–571.

Worley, K.C., Wiese, B.A., and Smith, R.F. (1995) BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res* **5:** 173–184.

Young, D.B. (1997) TB research; entering the post-genomic era. *Mol Med Today* **3:** 6–7.

Zhang, Y., and Young, D. (1994) Strain variation in the *katG* region of *Mycobacterium tuberculosis*. *Mol Microbiol* **14:** 301–308.