

https://dspace.org/

Mutation rates in mycobacterial hosts with altered Dna metabolic activity

Barichievy, Samantha

2006-02-08
http://hdl.handle.net/10539/163

Downloaded from a DSpace repository. https://dspace.org/

Chapter 1

Introduction

1.1 The worldwide burden of tuberculosis

Mycobacterium tuberculosis is an obligate human pathogen that causes the disease tuberculosis (TB). This pathogen has evolved with its human host, indiscriminate of nation or ethnicity and infected early man, with records of active TB disease as early as 3000 BC (van Helden, 2003). Following decades of co-existence, factors such as human migration, urbanisation and changes in hygiene levels have served as mechanisms to accelerate the spread of the bacterium. Recent assessments of the burden of TB disease indicate that 1.8 billion people are infected with *M. tuberculosis* and of these, 16.2 million people are currently suffering from active disease (Navin *et al.*, 2002). Each new year, TB cases lead to approximately 2 million deaths or 23% of global fatalities attributable to infectious diseases for that year (Dye *et al.*, 1999; Caws and Drobniewski, 2001). Despite the fact that the disease is preventable and treatable, TB continues to expand into a global crisis prompting the World Health Organisation (WHO) to declare it a global health emergency in 1993. The WHO insists that there is a slow but steady decline in the number of *M. tuberculosis* infections in the industrialised world, but the developing countries remain part of a far more alarming picture. Understanding the biology of the pathogen and the interactions it has with the host immune system are key factors in vaccine and drug development. Fortunately after the slump in TB related science following World War II, much progress has been made over the past 15 years in gathering information about the TB pathogen.

People infected with *M. tuberculosis* have a lifetime risk of developing the disease that ranges from 10% to 20%. In people co-infected with Human Immunodeficiency Virus (HIV) and *M. tuberculosis*, the risk of developing active TB increases to 10% per annum (Corbett *et al.*, 2003). Countries suffering from a HIV epidemic are thus the worst affected by the TB burden. Currently Asia, Eastern Europe and sub-Saharan Africa are the worst affected geographic regions, with TB as the most common cause of morbidity in HIV positive (HIV+) adults. In Africa 31% of all new TB cases in adults are attributable to HIV infection. In South Africa alone there are an estimated 2 million TB and HIV co-infected adults (Corbett *et al.*, 2003). Many cases of TB are due to primary infections but a critical number of people develop reactivation disease (Young and Duncan, 1995a). These latter patients suffer a primary *M. tuberculosis* infection that does not develop into a symptomatic TB disease. Instead, the mycobacteria enter a latent stage of infection and the patients remain asyptomatic. Reactivation disease develops when 'dormant' *M. tuberculosis* bacilli respond to an environmental signal, such as immune suppression linked to HIV infection, and re-emerge causing active disease. HIV+ individuals are thus more susceptible to TB infection and, in addition to this, place HIV negative (HIV-) people at a greater risk of contracting TB disease due to the increased transmission rates (Corbett *et al.*, 2003).

As with most infectious diseases, the poor are worst affected by TB as a result of poor living conditions (overcrowding), malnutrition and a lack of access to drug treatment programs. It has been

stated that "The poor will remain a breeding ground for the 'white plague' until we realise that it is not only the microbe *Mycobacterium tuberculosis* that is causing the disease" (van Helden, 2003). TB as a disease was not a major threat to mankind until humans began living in close communities and changing their diets. Industrialised countries have better access to disease prevention and treatment programs. However, the management of TB cannot focus solely on the pool of infected individuals in industrialised areas as, regardless of socio-economic location, those infected with the disease place those who are not at an increased risk of infection. Even though industrialised countries share less of the disease burden, TB remains a global responsibility. Treatment of the disease in an immunocompetent patient involves a minimum of 6 months with multiple antibiotics (Glickman and Jacobs Jr, 2001). This curative regimen is less than optimal as poor patient compliance is common. Eradication of TB requires a thorough understanding of the biology of the infectious agent, development of more amenable treatment regimens and a review of the currently available vaccine.

1.2 The BCG vaccine

The outcome of infection with *M. tuberculosis* depends strongly on the response of the host immune system (Cosma *et al.*, 2004; Pozos and Ramakrishnan, 2004). In most cases, active disease will not occur because infection with *M. tuberculosis* induces an immune response sufficient for protection against the establishment of a primary disease state (Young and Duncan, 1995b; Young and Duncan, 1995a; Cosma *et al.*, 2004). Under optimal conditions, a good vaccine will induce an immune response as close to that of a natural infection, thus reducing the occurrence of primary disease should a virulent infectious agent subsequently challenge the host. Calmette and Guérin developed the BCG (bacille Calmette-Guérin) vaccine by *in vitro* passaging of a virulent strain of the very closely related organism, *Mycobacterium bovis*, until the pathogen was rendered avirulent (Mahairas *et al.*, 1996). BCG has been administered globally since 1921 resulting in over 3 billion people having received this vaccine. The administration of BCG by multiple puncture together with the production of freeze dried BCG advanced large-scale vaccinations. Initially, oral administration of the live attenuated vaccine was used but more recent methods include subcutaneous injection of a live recombinant multivalent vaccine, effective against a number of mycobacterial pathogens.

Repeatedly, the BCG vaccine has been found to protect against primary infection in children but has been only partly successful against pulmonary disease in adults (Fine, 1990; Colditz *et al.*, 1994). One favored explanation for its poor efficacy is the fact that the BCG vaccine was generated from a non-human mycobacterial pathogen. Although *M. tuberculosis* and *M. bovis* share a 99% similarity in their DNA sequence (Gordon *et al.*, 2001; van Helden, 2003), in retrospect, an attenuated *M. tuberculosis* strain may have produced a vaccine with better efficacy in humans. Since the early 1940's many countries the world over have cultivated the BCG vaccine for their own immunisation programs. Consequently, a great number of BCG strains have been produced and molecular analysis has revealed chromosomal variations between them (Jungblut *et al.*, 1999). Thus worldwide

immunisations against TB have not been used with the same BCG vaccine, which immediately compromises the ability to assess the efficacy of the BCG vaccine in any given trial. In addition, the efficacy of the BCG vaccine may be lowered in areas where environmental mycobacteria are present at high concentrations in local soil and water supplies (Snewin *et al.*, 2001). Studies have shown that people exposed to mycobacteria in their surroundings may inadvertently be priming their immune systems against successive challenge with either pathogenic *M. tuberculosis* or the BCG vaccine strain. However, the results are unclear as background immunisation from the environmental bacteria also elicits an immunity in the negative control patients, thereby lowering the apparent efficacy of the BCG immunisation (Snewin *et al.*, 2001; Brandt *et al.*, 2002).

To complicate matters even further, there is significant geographical variability in the efficacy of BCG, the reasons for which remain largely unknown. Recently, a study that used epidemiological data in conjunction with genetic analysis, demonstrated a stable association between the human host and TB pathogen populations (Hirsh *et al.*, 2004). A human host's geographical location of origin is predictive of the specific strain of *M. tuberculosis* carried and this association remains strong even if transmission takes place somewhere outside of the host's birth place. Compounding this, genetic variability in pathogen populations, such as *Plasmodium* or *Streptococcus pneumoniae*, has been shown to render vaccines raised against them ineffective. This may also be true of the relationship between the various *M. tuberculosis* strains and the BCG vaccine, but currently, no efficacy studies have been done using distinct *M. tuberculosis* strains, BCG vaccine strains and their human host populations. Auxotrophic vaccines have been shown to protect against *M. tuberculosis* challenge in immuno-compromised mice and may represent a potentially safe and useful vaccine candidate against TB in HIV+ populations (Guleria *et al.*, 1996). The poor efficacy of BCG against the most common form of TB and the paucity of alternative vaccines remains a hurdle in the fight to eradicate the disease. Vaccines represent the host immunity on the scale of a whole organism. To understand disease prevention at a cellular level, it is important to take a closer look at the host immune response.

1.3 Macrophages, granulomas and the host response

Macrophages are one of the key players in the host immune response towards long-term bacterial intracellular pathogens. Using a variety of cell surface associated receptors, macrophages interact with and internalise foreign bacteria, undergo maturation and then assault the pathogen with acid, lytic enzymes, oxygenated lipids and potent reactive oxygen and nitrogen intermediates (Nathan and Shiloh, 2000). Using transcriptome analysis studies, it has been shown that naïve phagosomes containing *M. tuberculosis* provide an environment that damages DNA, impairs cell envelopes, deprives the bacteria of iron and carbohydrates, and supplies fatty acids (Nathan and Shiloh, 2000). Once exposed to Interferon gamma (IFN-γ), macrophages become activated, generate Nitric Oxide (NO) and further expose the internalised *M. tuberculosis* to nitrosative and oxidative intermediates as

well as cause inhibition of aerobic respiration pathways (Schnappinger *et al.*, 2003). Despite this hostile intracellular environment, *M. tuberculosis* continues replicating in an uncontrolled manner until the rupture of the infected macrophages. This allows the infection to spread to other cells and for extracellular replication of the pathogen to occur. As *M. tuberculosis* numbers increase, the host lymph system supplies a ready vehicle for transport to many other organs, resulting in disseminated disease. Fortunately for most individuals, the host immune response is able to contain the infection within three weeks and a state of latent disease develops (Toossi and Ellner, 1998).

Humans are the only hosts capable of being infected by *M. tuberculosis* and transmission occurs almost exclusively by aerosolised droplets containing infectious organisms (Glickman and Jacobs Jr, 2001). When an individual suffering from pulmonary TB coughs, droplets containing the infectious agent are spread. If these are inhaled by an uninfected individual, a new disease cycle begins. Once inhaled, the bacilli lodge in the lung alveoli where the host's most proficient phagocytes, the alveolar macrophages, become the specific cell within which the bacteria reside (Toossi and Ellner, 1998; Glickman and Jacobs Jr, 2001). In order to gain access to these particular cells, *M. tuberculosis* has developed a number of manipulative strategies. Complement receptors (Schlesinger *et al.*, 1990), fibronectin receptors (Ratliff *et al.*, 1988) and mannose receptors (Schlesinger, 1993) are all host cell surface associated molecules exploited by the bacteria to obtain entry into the intracellular space. Once inside the macrophage, *M. tuberculosis* can replicate at this initial pulmonary infection site and it does so within the phagosome. The pathogenic organism further enhances its own chances of survival by modulating the maturation of the phagosomal compartment. The strategies employed include alkalinisation of the vacuole through exclusion of the proton ATPase (Sturgill-Kosycki *et al.*, 1994), prevention of phagosome–lysosome fusion (Armstrong and Hart, 1971) and inhibition of nitric oxide synthase recruitment (Miller *et al.*, 2004).

M. tuberculosis is probably most renowned for its ability to enter a persistent state within the host that can last for decades or until a deficiency in host immunity allows for reactivation of the pathogen. Within the intracellular space of a macrophage, it is postulated that oxygen and nutrients become growth limiting factors. This seems especially plausible when examining the morphology of a granuloma, the classical hallmark of pulmonary TB. Granulomas are an organised collection of cells, local host tissue and cellular debris. Guinea pigs are assumed to have granuloma formation closest to that of humans and in this model, the morphology is characterised by a caseous necrotic core surrounded by a layer of mature *M. tuberculosis* infected macrophages, and a subsequent layer of lymphocytes with fibrosis (Turner *et al.*, 2003). The host immune system bombards the structures with additional leukocytes and fibroblasts to try and prevent bacterial dissemination. There is also an extravagant interaction of cytokines to help facilitate the coordinated migration and integration of cells within the granuloma. Upregulation of chemokines, receptors, adhesion and integrin molecules all ensure a correct temporal and spatial influx of cells (Saunders and Cooper, 2000) that result in an

'igloo' inside which *M. tuberculosis* persists. The synchronisation of the process also ensures retention of cells in the granuloma over time (Turner *et al.*, 2003).

As part of the human immune response to a TB infection, CD4+ cells secrete IFN-γ that activates macrophages. Once activated, these latter cells produce Nitric Oxide Synthase (NOS2) that in turn generates nitric oxide (NO) and restricts *M. tuberculosis* replication within the macrophage. NOS2 has been shown to localise to phagosomes only within activated macrophages using the actin cytoskeleton of the host cell for recruitment (Miller *et al.*, 2004). Mycobacteria are able to block this recruitment and so prevent the formation of NO. This specific pathogenic defence mechanism relies on the action of the host IFN-γ induced NOS2 pathway and when successful, allows *M. tuberculosis* to replicate within an otherwise lethal environment. The host cells are not without alternatives though. An IFN-γ dependant, NOS2 independent pathway exists, whereby TB infection can be controlled without the reliance on NO. Host guanosine triphosphatases (GTPases) such as LRG-47 serve as trafficking molecules that are functionally distinct from NOS2 and help in the acidification of macrophages (MacMicking *et al.*, 2003). LRG-47 is hypothesized to affect intracellular pH levels by influencing the number of adenosine triphosphate (ATP) dependant vacuolar proton pumps on the macrophage surface, thus dropping the pH to levels that even mycobacteria cannot withstand. Counteractively, *M. tuberculosis* can alkalinise its surroundings simply by halting NO synthesis or limiting the influx of hydrogen atoms into the immediate environment. Thus the TB pathogen tips the balance of sterilisation versus replication in its favour.

1.4 Chemotherapy

Host immunity clearly is not always sufficient to contain a TB infection. As a result antibiotics are used as an exogenous agent to fight this disease. Following 1993, the WHO implemented a DOTS regimen (Directly Observed Treatment, Short course) to accelerate the control of TB worldwide. This program consists of five components: (1) government commitment, (2) diagnosis of TB disease by sputum smears, (3) short-course chemotherapy with directly observed supervision of the drug taking, (4) regular drug supplies to clinics, hospitals and doctors and (5) competent record-keeping to ensure that progress is made (Mitchison, 2004). Despite a number of countries across the world adopting the DOTS strategy, there has been limited success especially in the developing world. Political indifference, the complication of HIV and TB co-infection, and the increasing prevalence of multi-drug resistant *M. tuberculosis* (MDR-TB) have proven to be major obstacles towards the containment of this disease in most developing countries. However, one notable exception is Uganda (Gandy and Zumla, 2003). Through a vigorous and open public awareness campaign, the Ugandan government was able to explain the origin of HIV infection and how preventative methods could be used, not only to prevent HIV infection but also to curb the spread of TB. The nationwide education campaign led to a decline in the number of HIV related TB infections and an overall reduction in TB disease in that country (Mitchison, 2004). Through education, the rate of new TB infections can be reduced but the

tedious chemotherapeutic regimen leads to poor patient compliance and exacerbates the development of MDR-TB.

The existing front line drug regimen used in TB treatment includes two months of Isoniazid (INH), Pyrazinamide (PZA), Rifampicin (Rif) and either Ethambutol (EMB) or Streptomycin (Str) followed by four months of INH and Rif. Generally, their modes of action are well understood and their relevance indicates that their bacterial targets are essential. To be effective, the antibiotic must penetrate the mammalian cell membrane (the macrophage), remain stable within the highly oxidative environment of the macrophage and still attain concentrations at which the drug is able to affect its target (Arain *et al.*, 1996). The initial uncontrolled replication of *M. tuberculosis* within alveolar macrophages allows pathogen numbers to increase dramatically. Coupled with the formation of a fairly impenetrable granuloma, antimycobacterial drugs are only partly successful at maintaining inhibitory concentrations at the sites of action. Even though the granuloma is a dynamic structure, the cellular morphology is well organised over time (Cosma *et al.*, 2004) and it remains a formidable three dimensional barrier to antibiotics. Drugs do manage to reach their targets and bacterial cell numbers may decline but the antibiotic concentration barrier created across the granuloma ensures that the bacteria in the necrotic core receive far less exposure to the drugs than those replicating extracellularly. Under these circumstances, antimycobacterial resistance inevitably becomes a problem, especially when treatment is intermittent or inadequate.

M. tuberculosis has evolved many mechanisms of drug resistance, all of which involve genetic alterations in a gene(s) associated with the action of a drug. In addition to chromosomally acquired resistance mechanisms, *M. tuberculosis* has an intrinsic resistance mechanism, the mycobacterial cell envelope. This structure has 2 features, a thick hydrophobic cell wall and a capsule-like polysaccharide coat (Figure 1). Together, they form a highly effective permeability barrier to most compounds (Brennan and Nikaido, 1995; Draper, 1998). The mycobacterial cell wall contains lipids, called mycolic acids, which are bound to arabinogalactans. These in turn are covalently linked to peptidoglycans (Trias and Benz, 1994). Mycolic acids within closely related bacteria such as *Corynebacterium* and *Nocardia* species have forty to sixty carbon atoms. However, mycolic acids within mycobacteria contain seventy to ninety carbon atoms (Brennan and Nikaido, 1995; Draper, 1998). The unusual length of these long chain fatty acids and the presence of arabinogalactan contribute to the poor permeability of the cell wall. In addition to this formidable barrier, mycobacteria have other mechanisms to resist drugs. They use efflux systems to pump unwanted chemicals, including drugs, out of their compartments and produce hydrolytic enzymes, capable of modifying drugs. These factors all contribute to the general 'natural resistance' that mycobacteria have against many chemotherapeutic compounds (reviewed in Zhang and Telenti, 2000).

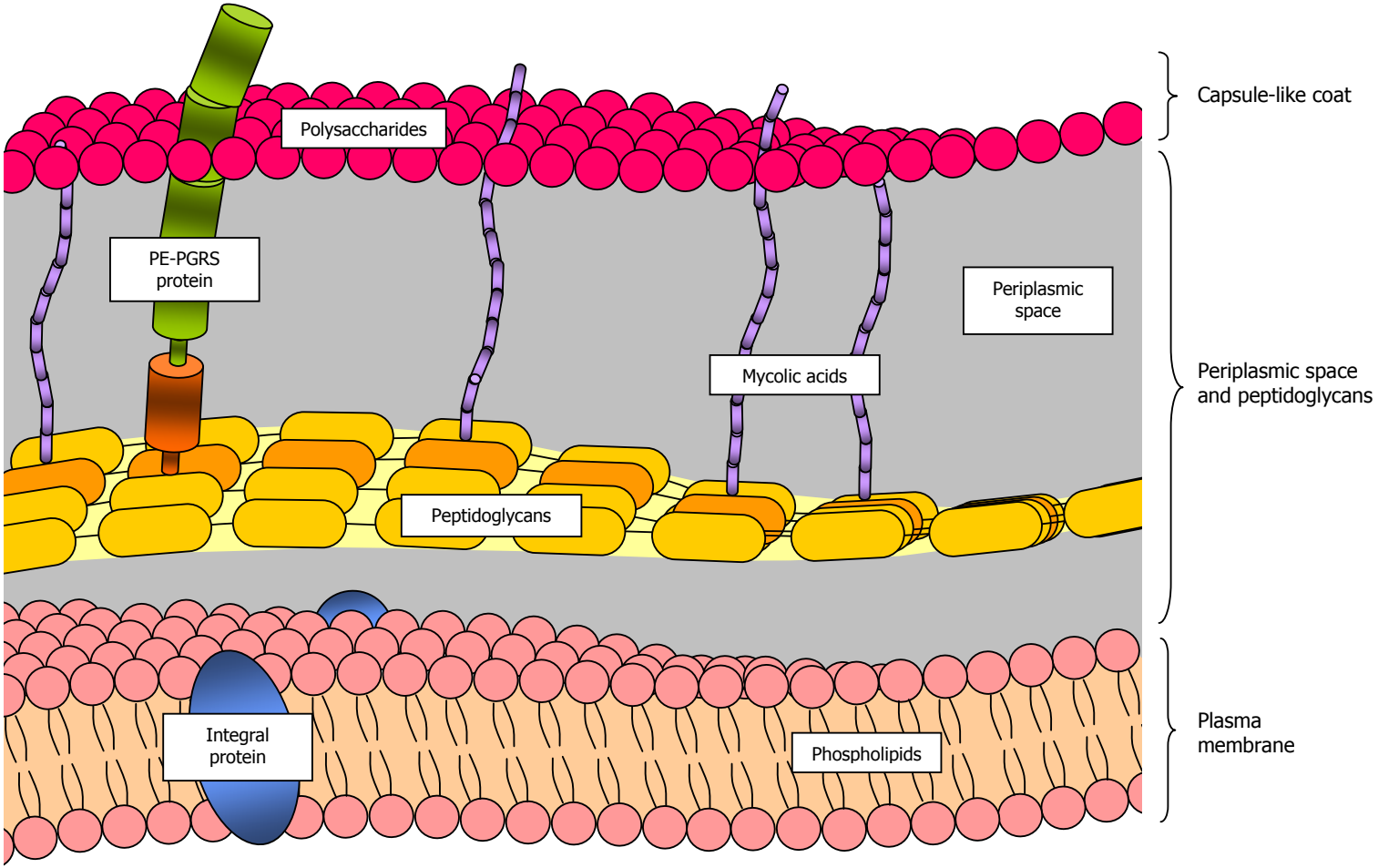


Figure 1: The mycobacteral cell envelope. The plasma membrane is composed of a phospholipid bilayer with integral proteins. The periplasmic space contains the peptidoglycan (indicated by yellow chain) and arabinogalactan (indicated by orange chain) with long chain mycolic acids covalently bound to the latter polymers. The outer capsule-like coat is composed mostly of polysaccharides. PE-PGRS proteins that are localised at the cell surface are possibly bound to the peptidoglycan layer allowing them to interact with the extracellular milieu.

1.5 MDR-TB

Presently MDR-TB, defined as strains of *M. tuberculosis* that are resistant to at least INH and Rif (Ramaswamy and Musser, 1998), may be growing as rapidly as 400 000 cases each year and presents a major problem in areas such as Estonia, Latvia, China's Henan Province and the Russian Federation. A recent study has shown that areas where drug sensitive *M. tuberculosis* strains have been successfully limited through anti-TB drug use, frequently evolve into areas with a high prevalence of MDR-TB (Blower and Chou, 2004). Another study has revealed that the future burden of MDR-TB depends strongly on the relative fitness of these strains to compete against drug sensitive *M. tuberculosis* and other less fit MDR-TB strains (Cohen and Murray, 2004). Together, these studies imply that the current epidemiological measures are insufficient at limiting MDR-TB as no efforts are aimed specifically at these strains. Front line drugs are of no use in the treatment of MDR-TB and so second line drugs, such as Capreomycin, Ethionamide and Para-aminosalicylic Acid are used in these cases. Unfortunately, the efficacies of both front and second line TB drugs are lessened through poor patient compliance, an increasing pattern of antibiotic misuse, and the effect of co-infection with HIV. It seems plausible that the problems associated with enduring TB treatment and granuloma morphology are augmenting the increasing incidence of MDR-TB.

Theoretically, sequential accumulation of mutations conferring resistance to single antimycobacterial drugs may lead to a MDR-TB strain. If the acquisition of resistance is random and in a given mycobacterial population 1 in 10^6 bacteria develop INH resistance and 1 in 10^8 cells mutate to develop Rif resistance, then the likelihood a single bacterium has acquired MDR status is 1 in 10^{-14} (Harkin and Harris, 1995). Alternately, MDR-TB strains may arise through a single-step process, for example if a solitary 'MDR element' existed and was passed on to a drug sensitive species. Mutations which alter the cell wall so dramatically that no drug could enter the cytoplasmic space may also result in a MDR phenotype. Mechanisms mediating multi-drug resistance in other pathogenic bacteria, such as the transfer of a single 'MDR element' by plasmid conjugation or transposable elements, do not seem to occur in *M. tuberculosis* (Ramaswamy and Musser, 1998; Zhang and Telenti, 2000). This lends weight to the idea of sequential mutations leading to MDR. The ability to accurately genotype clinical isolates of *M. tuberculosis* has shown that globally, TB has been propagated by thousands of different genotypes of this mycobacteria (Victor *et al.*, 2004). Furthermore, the strains differ with respect to their frequencies relative to different areas. This is also true of MDR-TB. In a recent study from the Western Cape Province, South Africa, 40% of TB isolates collected were multi-drug resistant (Streicher *et al.*, 2004). In comparison, 33% of TB isolates in Estonia and 43% of TB isolates in Archangel, Russia had a MDR phenotype (Werngren and Hoffner, 2003).

In a study from 1995, IS*6110* Restriction Fragment Length Polymorphism (RFLP) and spoligotyping patterns showed that the largest proportion of *M. tuberculosis* strains from Beijing in China, subsequently named the Beijing family, shared a high degree of similarity (Kremer *et al.*, 1999). These

strains have drawn attention because they are often associated with MDR-TB and have been identified as being responsible for several global TB outbreaks, including ones in several states in the USA, in South Africa, Russia and Spain (Caminero *et al.*, 2001). The strong association between *M. tuberculosis* Beijing strains and multi-drug resistance, coupled with their worldwide dissemination has raised the question of whether these strains have an enhanced capacity to develop drug resistance (Rad *et al.*, 2003; Werngren and Hoffner, 2003). *M. tuberculosis* does not have plasmids and cannot transfer DNA horizontally between strains, thus all adaptations to antibiotics occur via chromosomal mutations only. A study showed that *M. tuberculosis* strains from the Beijing family had acquired missense mutations in three putative *mut* genes, including two of the *mutT* type (Rv3908 and *mutT2*) and *ogt* (Rad *et al.*, 2003). The *mutT* genes are involved in mismatch repair via recycling of the nucleotide pool, while *ogt* is involved in the reversal of promutagenic O-alkylated DNA damage (Durbach *et al.*, 2003). The observation that these point mutations were unique to Beijing strains and that they occurred in genes involved in DNA damage repair, led the authors to suggest that these defects may render these strains hypermutable.

Molecular analysis of the Beijing strains in the above study showed that their multi-drug resistant status evolved sequentially suggesting a possible pathway for the clonal expansion of the family (Rad *et al.*, 2003). The suggested hypermutable state of these strains was also proposed as a partial explanation for the successful adaptation of Beijing strains to their host environment. However, a subsequent study showed that the association of the Beijing strains with multi-drug resistance was not due to an altered ability to develop resistance (Werngren and Hoffner, 2003). In this latter study, it was hypothesized that an elevated mutation rate for a *M. tuberculosis* Beijing isolate is associated with an increased acquisition of drug resistance, resulting in a more rapid selection of the drug resistant bacteria. The Beijing strains exhibited an acquisition of resistance comparable to non-Beijing strains as shown by mutation rates that were not significantly different. The observation that Beijing strains are no more prone to generating resistance than are non-Beijing strains suggests that a hypermutable state is not responsible for the association between the Beijing genotype and multi-drug resistant status. Importantly, the development of drug resistance in Beijing strains was not tested under conditions of suboptimal therapy. Furthermore, virulence, fitness and transmissibility have not been yet tested in these strains (Werngren and Hoffner, 2003) and so the reasons for the widespread dissemination and prevalence of these strains remain unclear.

It must be noted that not all MDR-TB strains belong to the Beijing family. Studies on clinical isolates from rural districts in the Western Cape Province of South Africa have revealed at least three different MDR-TB strains are widespread and recently transmitted in that area (Streicher *et al.*, 2004; Victor *et al.*, 2004). Strains belonging to the *M. tuberculosis* Beijing family, strain family 11 (F11) and strain family 28 (F28) were identified. Furthermore, a large pool of INH-monoresistant isolates has been identified and is of great concern as they may serve as a source for the development of future MDR-TB. Interestingly, strain F11 is more prevalent (21.4%) than Beijing strains (16.5%) and are as

successful at contributing to the MDR-TB problem in Western Cape communities in South Africa (Victor *et al.*, 2004). Isolates of F11 not only contribute to the TB epidemic in South Africa but have also been located in twenty five countries around the world and on four different continents. Unfortunately, as in the case of the Beijing strains, the factor(s) underlying the epidemiology of the F11 strains have not yet been elucidated. Mutations certainly produce genetically different individuals that can compete for resources, the result of which is the selection of the best adapted genotype(s). The rate and spectrum of mutations is largely under genetic control and can involve different mutation mechanisms, also known as mutators.

1.6 Environment-dependent and heritable mutators

1.6.1 Environment-dependent mutators

Alterations in the mutation rate of an organism can occur in an environment-dependent or heritable manner (Metzgar and Wills, 2000). Environment-dependent changes are the result of the induction or suppression of mutator mechanisms (mutators) that always affect the mutation rate of an organism globally. A good example of an environment-dependent mutator is a group of enzymes called DNA mutator polymerases or error-prone (EP) polymerases (Radman, 1999; Friedberg *et al.*, 2002; Friedberg, 2003). These specialised polymerases are characterised by their ability to copy cognate lesions that may be non-instructive for replicative polymerases. When acting on their preferred template (containing a particular type of lesion), EP polymerases operate with high fidelity (Friedberg *et al.*, 2002). However, when operating on non-cognate lesions, they exhibit reduced genetic fidelity, a factor that results in the generation of mutations (Rattray and Strathern, 2003). One family of EP polymerases is the DinB-Rad30-Rev1 or Y-family of polymerases (Yang, 2003). The founding members of this family include the DinB and UmuC proteins from *E. coli*, now known as PolIV and PolV respectively. PolIV is encoded by *dinB* and may function in the re-initiation of stalled replication forks (Goodman, 2002) or in bypassing bulky lesions (Shen *et al.*, 2002). Mutations in *umuC* and *umuD* in *E. coli* revealed that these genes encode PolV, proficient at by-passing UV-induced pyrimidine dimers (Reuven *et al.*, 1999; Tang *et al.*, 2000).

M. tuberculosis has two genes that show homology to the UmuC branch of the Y-family, namely *dinX* (Rv1537) and *dinP* (Rv3056) (Cole *et al.*, 1998). *Mycobacterium smegmatis* also has conserved regions for *dinX* (MSMEG3178) and *dinP* (MSMEG1002) as well as a third paralogue referred to as *dinP3* (MSMEG6405) (Figure 2). Outside of the Y-family of DNA mutator polymerases is another low fidelity DNA polymerase, the recently characterised DnaE2 (Boshoff *et al.*, 2003). The *dnaE2* gene of *M. tuberculosis* (Rv3370c) has been shown to be up-regulated *in vitro* by various DNA damaging agents as well as during infection in a murine model. Exposure of *M. tuberculosis* to UV irradiation, mitomycin C and hydrogen peroxide results in an up-regulation of *dnaE2* that in turn leads to induced mutagenesis and a direct emergence of drug resistance *in vivo*. It has been suggested that the gene

product, DnaE2, is a polymerase with relaxed fidelity that mediates survival of *M. tuberculosis* by inducing mutagenesis through EP, translesion synthesis across sites of DNA damage. This induced mutagenesis affects bacterial resistance to drugs such as Rif in a *dnaE2*-dependant manner. Resistance to Rif occurs due to mutations in the *rpoB* gene encoding the RNA polymerase β subunit and exposure to UV irradiation causes a *dnaE2* dependant increase in bacterial mutation rate (Boshoff *et al.*, 2003; Friedberg and Fischhaber, 2003). Due to the fact that DnaE2 causes a genome-wide increase in mutation rate in response to an exogenous factor, this polymerase is also classed as an environment-dependent mutator.

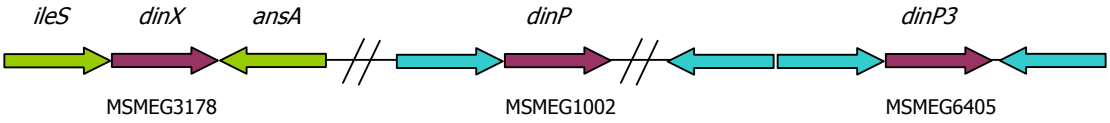


Figure 2: Chromosomal organisation of the conserved loci of *dinX* and *dinP* in *M. smegmatis* as well as a third homologous *umuC*-like gene, *dinP3*. The genes are indicated by dark red arrows with the gene numbers as annotated by TIGR (http://www.tigr.org) indicated at the bottom of the figure. These genes encode putative DNA polymerases belonging to the UmuC-DinB-Rad30-Rev1 or Y family of polymerases. The green arrows indicate other annotated genes in the *M. smegmatis* genome. The blue arrows indicate genes in *M. smegmatis* with homology to annotated genes from *M. tuberculosis* (Cole *et al.*, 1998).

1.6.2 Heritable global mutators

Heritable mutators are mechanisms that are independent of the environment. They may act resulting in a global increase in mutation rate or alternatively they may be local (Metzgar and Wills, 2000). Heritable global mutator phenotypes have been linked to mutations within genes involved in DNA repair, especially the methyl-directed mismatch repair (MMR) system. Indeed, the mutations within the *mutT*, *mutT2* and *ogt* genes of *M. tuberculosis* Beijing strains is what led to the suggestion that these may be mutator strains (Rad *et al.*, 2003). However, the observation that the global mutation rates of these strains were no different to that of non-Beijing strains dismissed this suggestion (Werngren and Hoffner, 2003). In order for a heritable global mutator to become fixed within a

population, second-order selection has to occur, i.e. some kind of advantageous mutation that is linked to the mutator has to be generated (Weber, 1996). Asexual reproduction then ensures the linkage is maintained and the mutator is inherited. Importantly, a globally increased mutation rate is only favoured if the normal rate of mutation is the limiting factor in the adaptation to a selective pressure (de Visser *et al.*, 1999). Heritable global mutator phenotypes give rise to progeny with elevated mutation rates as well. Contrastingly, although environment-dependent mutators cause a globally increased mutation rate, this mutator phenotype is not inherited and the progeny of these cells have normal mutation rates (Metzgar and Wills, 2000).

1.6.3 Heritable local mutators

Local mutators are the result of a unique sequence that due to its characteristics predisposes regions of the genome to certain mutations (Metzgar and Wills, 2000). The contingency loci of pathogenic bacteria provide a good example of this kind of mutator. Pathogenic prokaryotes with these sequences are able to affect mutation rates locally, i.e. just within particular genes. Generally contingency loci code for antigens that are expressed on the cell surface. Furthermore, the expression of these antigens can be alternated between states of on and off at high rates. This allows the pathogens to rapidly alter their antigen display without incurring the deleterious mutations linked to a global increase in mutation rate. The switching mechanism of contingency loci results from the hypermutable properties of tandemly arranged repeat sequences that are located within the relevant gene(s). During DNA replication, slippage of the repeats resulting in insertion or deletion mutations occurs (Robertson and Meyer, 1992). Clearly, pathogens that are able to readily adapt to their changing host environments are at an advantage. Pathogenic mycobacteria have been shown to contain abundant unique repeat sequences which may code for proteins that interact with host cells (Cole *et al.*, 1998). An example of these repeats are the PE and PPE gene families, which due to their genetic make-up provide good substrates for strand slippage events mediated by normal replicative polymerases. Are these genes local heritable mutators? Currently, that question remains unanswered.

1.7 Identification of the PE and PPE multigene families in the genome of *M. tuberculosis*

A landmark in the field of TB research was the completion of the genome sequence of *M. tuberculosis* strain H37Rv in 1998 (Cole et al., 1998). The genome of this organism comprises 4 411 529 bp with a 65.6% guanine (G) + cytosine (C) content. The G+C content of the genome was plotted (Brosch *et al.*, 2000) in order to identify regions with atypical nucleotide organisation, which may be linked to pathogenicity islands (Hacker *et al.*, 1997). The distribution was found to be uniform; however, several genomic areas were flagged with a G+C content > 80%. Further analysis showed that these areas are characterised by repeat sequences $^{5'}$CGGCGGCAA$^{3'}$ or $^{5'}$GCCGGTGTTG$^{3'}$ that code for two unrelated families of acidic, glycine rich proteins called the PE and PPE families (Cole *et al.*, 1998). An astonishing 10% of the coding capacity of the *M. tuberculosis* H37Rv genome is devoted to these two,

large multigene families whose names derive from a conserved proline-glutamate (PE) or proline-proline-glutamate (PPE) residue group at the N-terminal of the encoded protein. In the PE family, the N-terminal domain has ~110 amino acids while the corresponding PPE counterpart has ~180 amino acids (Figure 3). The C-terminal domains of both families are divergent and together with the polymorphisms in their respective N-terminal domains, extensive redundancy in these two gene families has been created (Glickman and Jacobs Jr, 2001).

Phylogenetic analysis has separated the 99 members of the PE protein families into various groups, the largest of which is the polymorphic GC rich sequence (PGRS) gene subfamily. These latter genes are the most abundant repetitive sequences in the TB complex occurring at 26 to 30 distinct loci (Kanduma *et al.*, 2003). PGRS elements are derived from a 9 base pair (bp) repeat $^{5'}$CGGCGGCAA$^{3'}$ arranged in tandem copies that translate to Gly-Gly-Ala, Gly-Gly-Asn motifs or variations of these (Poulet and Cole, 1995b).There are 68 members in the PPE family, divided into 3 groups, one of which comprises proteins with major polymorphic tandem repeats (MPTRs). This class is characterised by the existence of multiple tandem copies of an Asn-X-Gly-X-Gly-Asn-X-Gly motif that cause the C-terminus of these proteins to vary greatly in length (Cole *et al.*, 1998). The second subgroup in the PPE family is characterised by a conserved motif at amino acid position 380 while the third subgroup is simply related by the common N-terminal domain.

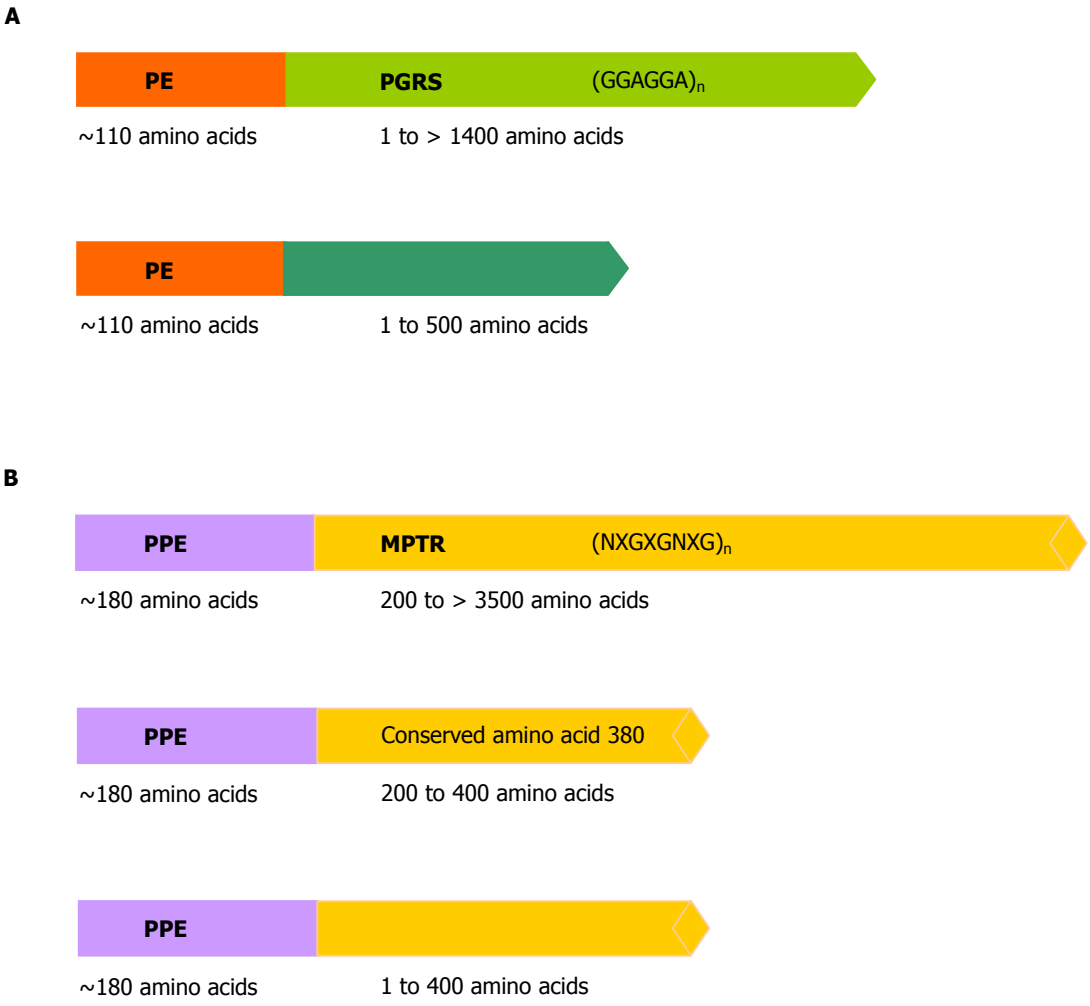
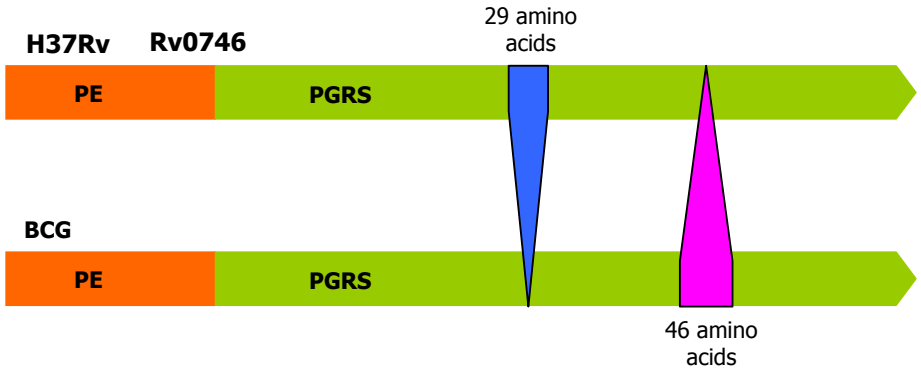


Figure 3: Members of the PE and PPE protein families. (A) The two main groups within the PE family of which the PE-PGRS proteins are the most abundant. Both groups share a common N-terminal domain sequence comprised of ~110 amino acids (indicated by orange blocks). The C-terminal domains are divergent and range in length from 1 to > 1400 amino acids (indicated by green blocks). The PE-PGRS proteins have a common amino acid sequence, GGAGG, repeated in tandem. (B) The three main groups within the PPE family. All three groups share a common N-terminal domain composed of ~180 amino acids (indicated by purple blocks). The C-terminal domains are divergent and range in length from 1 to > 3500 amino acids (indicated by yellow blocks). The PPE-MPTR proteins have a repeat sequence NXGXGNXG where X can be any amino acid (Cole *et al.*, 1998).

1.8 Polymorphisms in the PE and PPE genes

Although initial research identified the PE and PPE genes within the *M. tuberculosis* H37Rv laboratory strain, *M. marinum*, *M. gordonae*, *M. kansasii*, *M. bovis*, and *M. smegmatis* also contain these genes within their genomes (Ross *et al.*, 1992; Richardson *et al.*, 2004). Importantly, 60 PE-PGRS genes have also been identified in the clinical strain *M. tuberculosis* CDC1551 and these genes are most variable when compared to the laboratory strain H37Rv (Fleischmann *et al.*, 2002). Comparative sequence analysis of *M. tuberculosis* H37Rv and *M. bovis* BCG Pasteur indicated that the latter, non-pathogenic strain either lacks certain PE and PPE genes or contains corresponding genes with deletions or insertions. Of relevance to this study is the polymorphism between the *M. bovis* BCG Pasteur PE-PGRS gene that corresponds with the *M. tuberculosis* H37Rv gene Rv0746 (*PE_PGRS9*) (Figure 4). The BCG vaccine strain contains a 46-codon insertion and a 29-codon deletion in the PGRS portion of the Rv0746 gene counterpart (Cole, 1998). *M. leprae*, the organism considered to have the minimal mycobacterial gene set, has no intact PE-PGRS open reading frames (ORFs) while *M. smegmatis*, a non pathogenic saprophyte, has a larger genome than *M. tuberculosis* but fewer PE and PPE genes. Clearly, there is considerable inter-strain variation in PE and PPE genes and significant differences in the PE and PPE gene complements between mycobacterial species. The fact that *M. tuberculosis* has retained so many of these genes suggests they may be playing an important role in pathogenesis but the significance of the variation is presently unclear.

A



B

```
H37Rv       ---GSGAPGGAGGAAGLWGTGGAGGAGGSSAGGGGAGGAGGAGGWLLGDGGAGGIGGAST
BCG         ---GSGAPGGAGGAAGLWGT----------------------------GGAGGIGGAST

H37Rv       VLGGTGGGGGVGGLWGAGGAGGAGGTGLVGGDGGAGGAGGTGGLLAGLIGAGGGHGGTGG
BCG         VLGGTGGGGGVGGLWGAGGAGGAGGTGLVGGDGGAGGAGGTGGLLAGLIGAGGGHGGTGG

H73Rv       LSTNGDGGVGGAGGNAGMLAGPGGAGGAGGDGENLDTGGDGGAGGSAGLLFGHGGAGGAG
BCG         LSTNGDGGVGGAGGNAGMLAGPGGAGGAGGDGENLDTGGDGGAGGSAGLLFGHGGAGGAG

H37Rv       GFGFLGGDGGAGGNAGLLLSSGGAGGFGGFGTAGGVGGAGGNAGWLGF------------
BCG         GFGFLGGDGGAGGNAGLLLSSGGAGGFGGFGTAGGVGGAGGNAGWLGFGGAGGIGGIGGN

H37Rv       ---------------------------------GGAGGVGGSAGLIGTGGNGGNGGTGA
BCG         ANGGAGGNGGTGGQLWGSGGAGGEGGAALSVGDTGGAGGVGGSAGLIGTGGNGGNGGTGA

H37Rv       NAGSPGTGGAGGLLLGQNGLNGLP-----------------------------------
BCG         NAGSPGTGGAGGLLLGQNGLNGLP-----------------------------------
```

Figure 4: Sequence variation that exists between *M. tuberculosis* H37Rv and *M. bovis* BCG in the PE-PGRS gene Rv0746 (Cole *et al.*, 1998). (A) The BCG protein has a deletion of 29 amino acids (indicated by blue arrow) and an insertion of 46 amino acids (indicated by pink arrow) compared to the sequence in the *M. tuberculosis* H37Rv counterpart. (B) Specific amino acid sequence variation between H37Rv (upper line of sequence) and BCG (lower line of sequence) in gene Rv0746. The 29 amino acid deletion in BCG is indicated by blue letters. The 46 amino acid insertion in BCG is indicated by pink letters. The Morf2 substrate used in this study was based on the 29 amino acid deletion between the 2 sequences (refer to Section 2.10.1).

1.9 Localisation of proteins encoded by PE and PPE genes

Technically, research on PE and PPE genes and their encoded proteins is very challenging due to the large sizes of these gene families, the high degree of similarity between genes within a given family, their repetitive structure and their unusually high G+C-richness. Studies have shown that some PE-PGRS proteins bind to fibronectin, suggesting a potential mechanism of interaction between the pathogen and its human host cells (Espitia *et al.*, 1999). Indeed, research has shown that DNA vaccines that express PE-PGRS proteins elicit an antibody response. These antibodies then recognise PE-PGRS antigens on mycobacterial surfaces (Brennan *et al.*, 2001; Banu *et al.*, 2002). Immunofluorescence studies of PE and PPE proteins within mycobacterial cells, have revealed that many of them are localised to the cell wall (Banu *et al.*, 2002). Green fluorescent protein (GFP) has been tagged onto a PE-PGRS gene Rv1818c of *M. tuberculosis* and fluorescent microscopy showed that the fusion protein is localised to the cellular poles in the mycobacterial cell wall and exposed to extracellular constituents. Analysis revealed that the PE portion of the protein is specifically responsible for targeting the entire protein to the cell wall compartment (Delogu and Brennan, 2001). The host immune response to PE-PGRS proteins, together with the presence of certain PE-PGRS proteins at the mycobacterial cell surface implies that these proteins may be interacting with the host immune system.

1.10 Potential biological functions of proteins encoded by PE and PPE genes

To date, the biological functions of PE and PPE proteins have not been elucidated. The presence of PE-PGRS proteins at the cell surface alters colony morphology and has been linked to differences in virulence between mycobacterial strains. This was highlighted in a recent example using a *M. bovis* BCG strain that had been complemented with a PPE protein. The complemented strain was more virulent than the parent and showed altered colony morphology (Pym *et al.*, 2002). Other studies have also shown that the expression of a PE-PGRS gene (*mag* 24-1) by *M. marinum* is associated with replication in macrophages and persistence in granulomas in the frog model of infection (Ramakrishnan *et al.*, 2000). *M. marinum* is closely related to *M. tuberculosis* genomically, thus it is likely that these two organisms employ related strategies of replication and persistence in their host cells. Furthermore, the frog model of infection best emulates the granulomas in the human lung and so these host / pathogen interactions are particularly interesting. Serology experiments demonstrated that antibodies in human sera of TB infected patients react with surface antigens on mycobacteria namely, PE-PGRS protein Rv1759c (Espitia *et al.*, 1999) or the PGRS domain of Rv3367 (Singh *et al.*, 2001). Complementary results from another study showed that antibodies raised against a number of different PE-PGRS proteins can cross react, suggesting that proteins in this family share common epitopes (Banu *et al.*, 2002). Together, these data show that some PE-PGRS proteins are present in the host during a TB infection.

The localisation of certain PE-PGRS gene encoded proteins at the cell surface infers other potential functions such as lending to the structural integrity of the cell wall. A similarity can been drawn between the elastic properties of plant cell walls, fibrous proteins from silk and the PGRS domains of PE-PGRS proteins. All three groups of proteins are glycine rich and because the former two proteins have structural functions (Ye *et al.*, 1991; Hayashi and Lewis, 2000), a similar role has been proposed for PE-PGRS proteins. With this in mind, it is then intriguing to visualise the mycobacterial cell wall and the specific localisation of PE-PGRS proteins at the cell surface (refer to Figure 1 Section 1.4). As previously mentioned the PE portion of Rv1818c is responsible for targeting the entire protein to the cell wall compartment (Delogu and Brennan, 2001). Certainly, if localisation of PE and PPE proteins at the cell surface does occur, then they could be lending to the structural integrity of the wall. Other roles that have been proposed for PE and PPE proteins include functioning as storage proteins for asparagines (Cole, 1999) and esterase or lipase activity (Cole, 1998). Asparagine is a rare amino acid that is a nitrogen source for *M. tuberculosis* and so PPE proteins in particular have been postulated to serve as a store for these building blocks. Alternative studies have revealed that PE gene Rv3097c has a C-terminus sequence that codes for a peptide with predicted esterase or lipase activity but this function remains to be validated.

The inter-strain polymorphisms between members of the PE and PPE families suggest an inherent ability of *M. tuberculosis* to introduce strain variability using repeat sequences (Cole *et al.*, 1998). Further speculation surrounds these polymorphisms as they have intriguing parallels with hyper-variable, contingency loci that play an important part in generating antigenic diversity in other bacterial pathogens (Robertson and Meyer, 1992). Pathogens such as *Neisseria gonorrhoeae*, *Helicobacter pylori* and the Epstein-Barr Virus (EBV) all contain repeat sequences with analogy to the PE-PGRS genes of *M. tuberculosis*. In the former bacteria, the repeat sequences are rearranged, often by strand slippage during chromosomal replication, to create genetic variation in genes involved in virulence and pathogenicity. Some of these pathogens use their repeat sequences to manipulate host cell processes thereby improving their chances of intracellular survival. A similar observation has been made in *M. tuberculosis* where PE-PGRS proteins fused to GFP were shown to be less susceptible to intracellular proteosomal degradation compared to other mycobacterial proteins (Brennan and Delogu, 2002). This resistance to degradation suggested a mechanism whereby *M. tuberculosis* could inhibit processing of PE-PGRS proteins, within antigen presenting cells, which would otherwise have elicited an immune response.

The numerous similarities between repeat sequences in *M. tuberculosis* and other pathogens makes it tempting to assign the PE and PPE gene families a related function in immunopathogenicity. This is especially evident when the vast repertoire of variation linked to changes in repeat sequences is studied. Antigenic diversity, persistence and colonisation factors as well as virulence determinants all result from alterations in nucleotide repeat sequences. Often the mechanism to create genetic diversity involves slipped-strand mispairing and it has been suggested that this may also frequently

occur in the G+C rich domains of PE-PGRS genes. Furthermore, GC mispairing is fairly stable, thereby allowing genetic diversity created by mispaired bases to be inherited (Poulet and Cole, 1995a). Coupled with the pool of numerous PE-PGRS genes, there seems to be a likely and simple method available to create a wide range of mechanisms that would aid intracellular survival of pathogenic mycobacteria. Furthermore, a recent study has shown that diversity of the and PPE genes could be achieved through transcriptional regulation (Voskuil *et al.*, 2004). Global expression profiles demonstrated that different PE and PPE genes are expressed as *M. tuberculosis* encounters different environments. Therefore, aside from mutational mechanisms, PE and PPE genes may use gene regulation to generate diversity. For a better appreciation of the implications that variation in the PE and PPE gene families can cause, it is important to understand how other pathogens utilise their genetic repeat elements to their advantage.

1.11 Genetic diversity of pathogens involving repeat sequences: antigenic and phase variation

In pathogens, genetic variation can lead to clinically relevant phenotypic alterations, including phase and antigenic variation. Phase variation is the reversible loss or gain of a surface molecule whereas antigenic variation concerns the composition of that surface structure (Robertson and Meyer, 1992). The phenotypic variation of surface components is important as these structures are usually involved in host colonisation mechanisms or enhancing pathogen survival within the host. The ability of pathogens to adapt to changing environments thus rests on their capacity to generate novel phenotypes. Genetic variation has been well researched using a number of pathogens including *Bordetella pertussis* (Willems *et al.*, 1990), *N. gonorrhoeae* (Stern *et al.*, 1986; Murphy *et al.*, 1989; Makino *et al.*, 1991), *Mycoplasma hominis* (Ladefoged *et al.*, 1995), *H. pylori* (Josenhans *et al.*, 2000) and the EBV (Levitskaya *et al.*, 1995; Levitskaya *et al.*, 1997). The dominant mechanisms of variation involved in all of these examples are recombinational events and slipped-strand mispairing during DNA replication, both of which are associated with genomic repeat sequences (Banu *et al.*, 2002). As previously mentioned, slipped-strand mispairing to create variations in repeat sequences, has also been postulated to occur in the G+C rich domains of PE-PGRS genes and at a higher rate than variations in A+T rich domains (Poulet and Cole, 1995a).

1.11.1 Antigenic variation

Functionally, genetic variation can provide the means for a pathogen to evade host immune surveillance using different antigens. *Aniplasma marginale*, a tick-borne rickettsia responsible for causing haemoparasitic disease in cattle, produces variable major surface proteins that coat the organism. The peptides are encoded by a major surface protein 1 (MSP-1) domain containing a number of tandem repeat sequences. Alterations in the polypeptides are generated by recombination and slipped-strand pairing between the tandem repeats and size differences in the proteins correlate

directly with the number of repeats (Allred *et al.*, 1990). Similarly group A streptococcal bacteria, aetiological agents of many human blood infections, produce major surface molecules, called M proteins, to evade host immunity. These peptides are dimeric coiled-coil molecules varying in size between 41 and 80 kDa. The range of size is dependent on homologous recombination events between large repeat regions of the genes encoding the M protein (Hollingshead *et al.*, 1987). Illustrating another example, *Borrelia hermsii*, a spirochete responsible for causing Relapsing Fever, maintains a surface coat composed of variable lipoproteins. The differing primary structures of these proteins determines the serotype of each strain and the antigenic variation is associated with DNA rearrangements of all, or part of, an antigen specifying gene (Meier *et al.*, 1985).

The Gram negative coccus, *N. gonorrhoeae*, induces sexually transmitted disease in humans and expresses several types of variable surface molecules. The most thoroughly studied among these are the pilus and opacity (Opa) proteins, both essential for bacterial colonisation in human mucosa. Numerous studies have revealed mechanisms of phase and antigenic variation in *N. gonorrhoeae* Opa proteins (Meyer and van Putten, 1989; Murphy *et al.*, 1989; Makino *et al.*, 1991; van Putten, 1993). Linked to antigenic variation, the 5' region of various Opa genes has identical pentameric pyrimidine units (CTCTT) that encode the hydrophobic core of the signal peptide. Homologous recombination during DNA replication may occur between these repeats, leading to antigenic variation, however a recombination independent mechanism has also been hypothesized. The CTCTT repetitive sequence, comprised only of pyrimidine nucleotides, may cause the DNA polymerase to slip during replication leading to silent mutations. When translated, these mutations may be enough to create an antigenic variant (Stern *et al.*, 1986). The signal peptide sequences with their encoded repeats are only present in a few copies within each cell, thus silent mutations or recombination events between different Opa genes may be essential for increasing the number of variant epitopes produced by a single cell.

1.11.2 Phase variation

Each Opa protein of *N. gonorrhoeae* is capable of independent phase variation between on and off expression states (Mayer, 1982; Black *et al.*, 1984). All *opa* genes are transcribed off their own functional promoter but not all are translated into proteins. This is due to the expression control system that is dependent upon repetitive sequences in the hydrophobic core region of the protein. The number of repeat units is variable within each gene and determines the reading frame of the *opa* gene (Figure 5). Between 7 and 28 repeats can be present and depending on the number of units, the expression of individual *opa* genes can be switched on or off (Stern *et al.*, 1986; Meyer and van Putten, 1989). Genes that are turned on and thus translated into functional Opa proteins, have in-frame rearrangements of the CTCTT repeats. Those genes that are switched off show out-of-frame rearrangements (Murphy *et al.*, 1989). Expression control events are RecA independent showing that phase variation of Opa antigens relies only on the special characteristics of the repeat sequences. Mycoplasmas, the smallest free living prokaryotes, use a similar mechanism to generate diverse

surface antigens. *M. hominis* has variable adherence associated (Vaa) proteins that serve as adhesion molecules. The mature Vaa lipoprotein product contains 1 to 4 almost identical tandem repeat units totaling 121 amino acids in the central portion of the protein. Acquisition or depletion of these repeats leads to size variation of the subsequent lipid linked peptide, a result of insertions or deletions of short in-frame oligomeric sequences within the repeat regions of the genes (Zhang and Wise, 1996).

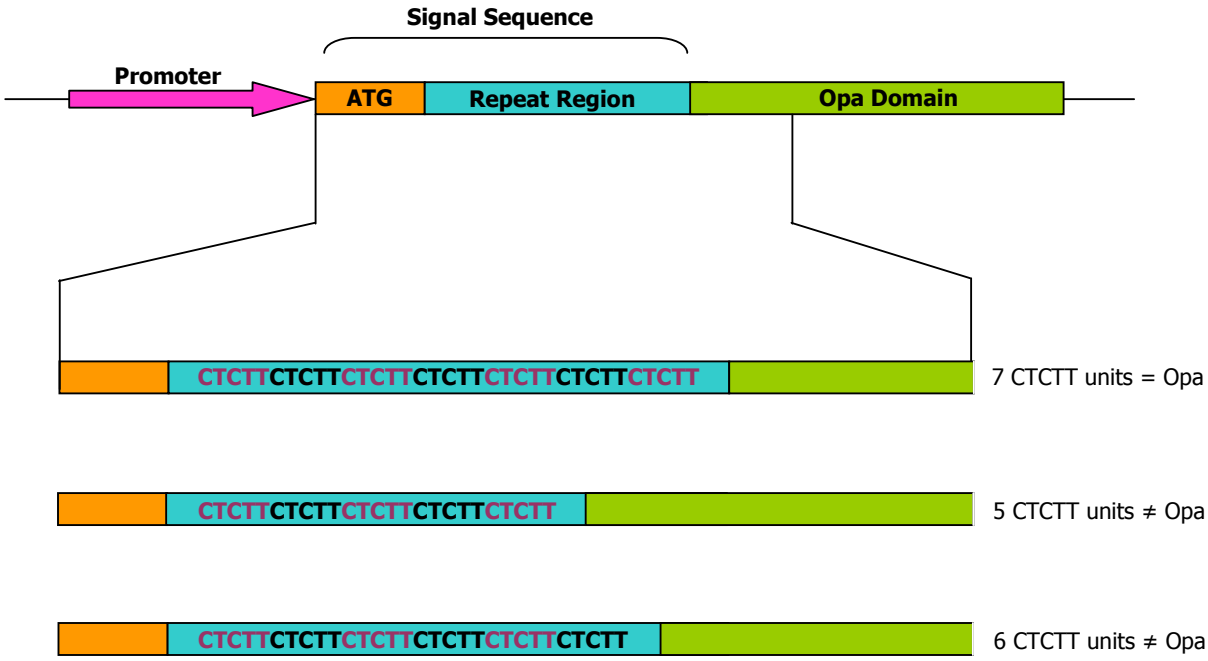


Figure 5: Organisation of an *opa* gene showing the control mechanism of gene expression. The pink arrow indicates the functional promoter, part of each unique *opa* gene. The orange block indicates the ATG start codon of the *opa* gene. The blue block indicates the repeat region of the gene containing various numbers of the CTCTT repeat unit (coloured pink or black). Together with the ATG start codon, the repeat region forms the hydrophobic signal sequence. The green block indicates the downstream portion of the *opa* gene. The number of CTCTT repeat units within the repeat region affects the open reading frame of the *opa* gene: if the reading frame is incorrect (for example if 5 or 6 CTCTT units are present), no Opa protein is produced. If the repeat units are in frame (for example if 7 CTCTT units) a functional Opa protein will be translated. Alterations in the reading frame between the CTCTT repeats are RecA independent and hypothesized to be due to replication error (Meyer and van Putten, 1989).

Phase variation in *H. pylori* is linked to the organism's state of motility in the viscous environment of the human stomach. As the major causative agent of gastritis and stomach ulcers in humans, *H. pylori* relies on its unipolar sheathed flagellae to colonise and persist in the host organ (Kostrzynska *et al.*, 1991). The regulation of flagellar synthesis and related bacterial motility in *H. pylori* is unique to this organism and once again involves simple repeat sequences in the genome (Alm *et al.*, 1999). The *fliP* gene of *H. pylori* contains adjacent cytidine residues and codes for peptides that make up the flagellar export apparatus. Quite simply, *H. pylori* uses single base pair, slipped-strand mediated frameshifts in the cytidine run of the *fliP* gene to shut down or start up flagellar synthesis and hence switch motility off and on (Josenhans *et al.*, 2000). In contrast, EBV uses repeat sequence rearrangements to evade host immunity. When phagocytosed, the proteins from most foreign organisms are degraded via the ubiquitin / proteosome pathway and the antigenic products are presented on the cell surface using major histocompatibility complex (MHC) class I molecules. Roaming cytotoxic T lymphocytes (CTLs) interact with MHC class I / antigen complexes and subsequently the infection is contained. EBV expresses Epstein-Barr virus nuclear antigen (EBNA) 1 that contains a repeat domain comprised of G and A residues. These repeats may prevent antigen processing by blocking the proteolytic machinery of the ubiquitin pathway needed to produce antigenic units (Levitskaya *et al.*, 1995). Without antigens, the MHC class I molecules avoid immune surveillance by CTLs and so uncontrolled proliferation of the EBV continues (Levitskaya *et al.*, 1997).

The examples illustrated above demonstrate how simple nucleotide repeat sequences can mutate or rearrange to create genetic variation. All these events involve some kind of misalignment between repetitive sequences during the course of chromosomal replication. One common feature is the proximity of the repeats – they need to be in a close vicinity, probably because their misalignment occurs within a single replication fork (Lovett, 2004). Misalignment events can occur in *trans* i.e. between a nascent strand and a repeat site on the template strand, or in *cis* i.e. between two repeat sites on the same strand. Furthermore, misalignment between repeats can cause deletions or duplications involving the repeat regions and any intervening bases (Figure 6) (Schaaper *et al.*, 1986; Schaaper, 1988). A second common feature of misaligned rearrangements is the lack of involvement of RecA proteins (Meier *et al.*, 1985; Hollingshead *et al.*, 1986; Murphy *et al.*, 1989; Allred *et al.*, 1990; Jonsson *et al.*, 1992; Levitskaya *et al.*, 1997; Josenhans *et al.*, 2000). The mechanisms responsible for repeat sequence manipulation remain unclear and many questions remain. Are slippage events random? Are low fidelity polymerases involved in these kinds of mutations? Whatever the answers may be, it is clear that repeat regions create genetic diversity and pathogens, including *M. tuberculosis*, seem to have capitalised on this.

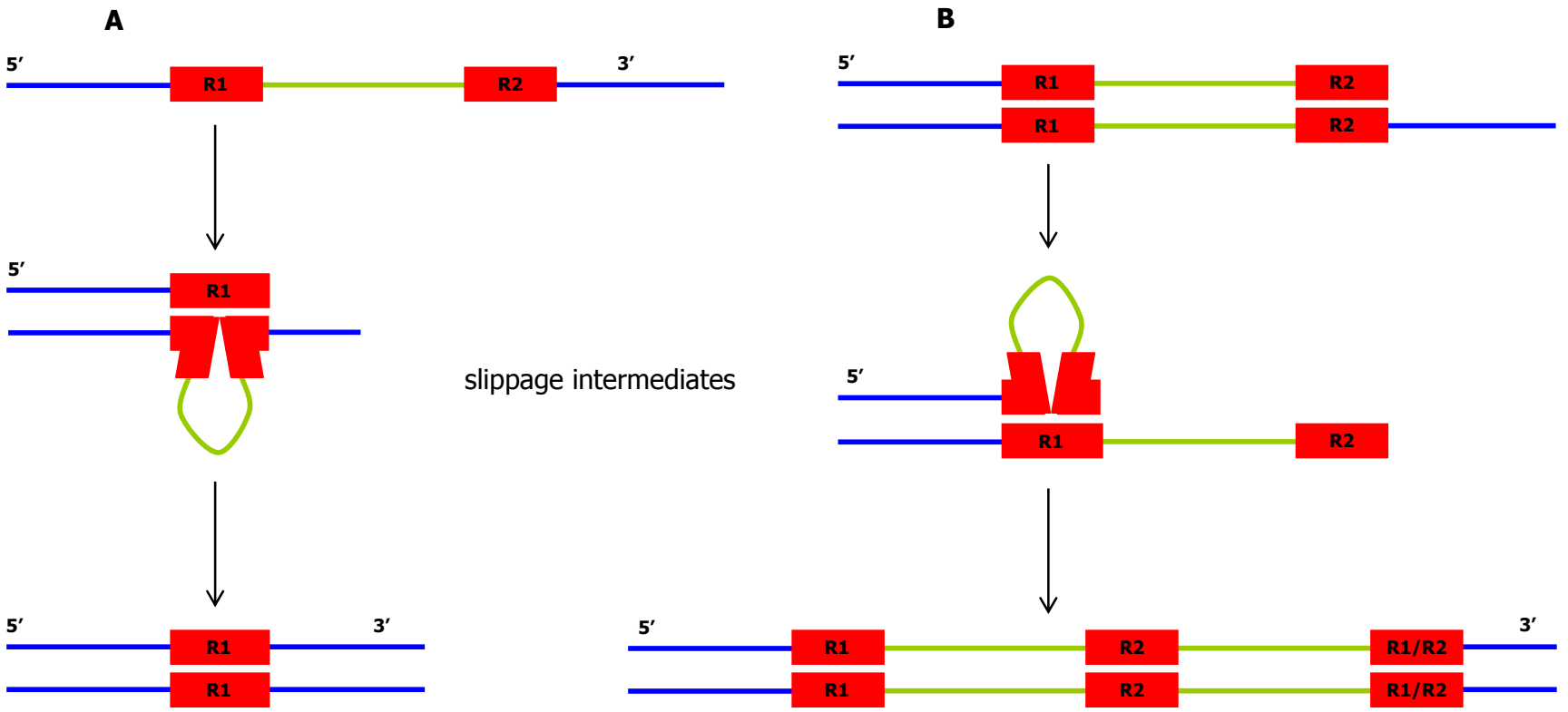


Figure 6: RecA independent misalignment events involving repeat sequences. (A) Deletion misalignment events are formed when there is backward slippage between repeat sequences (indicated by red blocks). After replication of the slippage intermediate, one repeat and the intervening DNA sequence (indicated by green line) are deleted. (B) Duplication misalignment events are formed when there is forward slippage between the repeat sequences. Replication of this slippage intermediate produces an expanded number of repeats including a duplication of the intervening sequence (Lovett, 2004)).

1.12 Mutation rate analysis

1.12.1 The original mutation rate experiments of Luria and Delbrück

Historically there has been much debate over whether mutations in the genome of an organism are spontaneous or acquired. A landmark study in 1943 by Salvador Luria and Max Delbrück showed that the ability of bacteria to rapidly respond to changes in their environment are a normal consequence of random gene mutation followed by selection in the growing population (Luria and Delbrück, 1943). Luria and Delbrück used an *Eschericia coli* and T1 bacteriophage system to determine the mutation rates of phage sensitive *E. coli* to phage resistance. They began with two hypotheses: 1) the Mutation Theory argued that phage resistance is a random genetic event that occurs prior to the onset of the selective pressure (i.e. the application of the phage cells to the bacterial population); and 2) the Acquired Immunity Theory stated that because every cell within a population has a small but constant and independent probability of undergoing a mutation to phage resistance within its lifetime, those cells that escape the selective pressure do so at random. During exponential growth, every cell within a bacterial population has a low probability of sustaining a mutation. After sustaining a mutation, the cell produces a clone of like-mutants and the size of this clone depends on the time point during the population's growth at which the mutation occurred. Thus the final number of mutants from a culture depends on a stochastic process. The fluctuating distribution of the number of mutants in a set of parallel cultures is the key factor used by Luria and Delbrück to determine mutation rates for bacterial populations (Luria and Delbrück, 1943).

If the first of their theories is followed and resistance mutations are assumed to be present at random prior to encountering a selective pressure, then the proportion of resistant cells within a population would increase over time. This is because existing mutants continue to reproduce and increase the pool of resistant cells while new mutants are also added to the pool (Figure 7A and 7B). Furthermore, the culture would contain clones of resistant bacteria of various sizes depending upon when during the growth phase the mutation to resistance occurred. On the other hand, if resistance is assumed to be acquired only upon exposure to the selective pressure, the proportion of resistant cells within a population over time would remain constant because every acquisition of resistance is considered to be an independent event (Figure 7C and 7D). Further, because it is assumed that every acquisition of resistance is an independent event with no genetic component, resistant cells and their descendants would arise as separate and scattered individuals. To distinguish between the two hypotheses, cultures could be examined to see if the proportion of resistant bacteria increases over time, or groups of bacteria could be examined to see if genetically related cells display the same resistance mutations (Robbins, 2001). Due to technical difficulties present at the time, Luria and Delbrück obtained their conclusions by analysing resistance in familial groups of bacteria.

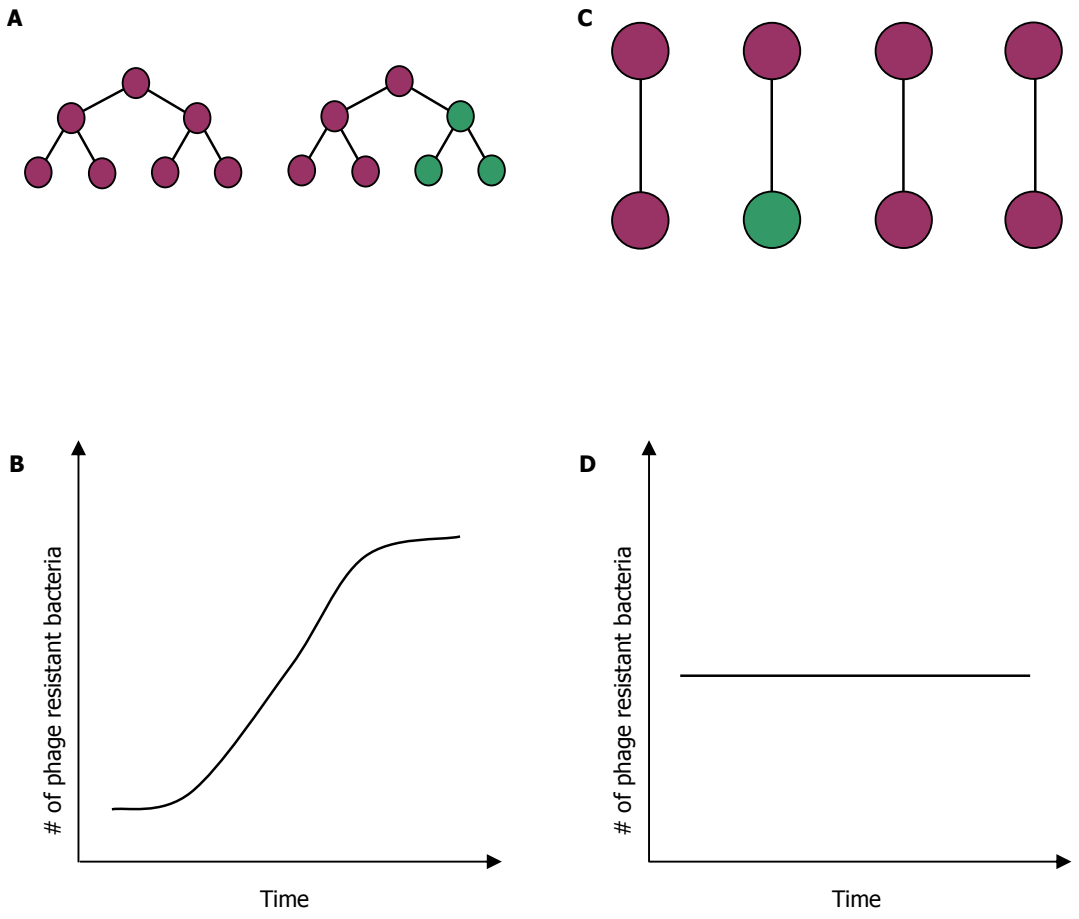


Figure 7: Schematic diagrams indicating the differences between the Mutation Hypothesis (A and B) and the Acquired Immunity Hypothesis (C and D) tested by Luria and Delbrück (Luria and Delbrück, 1943). (A) According to the Mutation Hypothesis, resistant bacteria will occur as a group of closely related individuals as each mutant gives rise to clonal progeny. Furthermore if resistance is due to random genetic events occurring prior to the onset of a selective pressure, the number of resistant cells within a bacterial population will increase over time (B) because the existing mutants grow and divide while new mutants are also added to the resistance pool. Alternately, the Acquired Immunity Hypothesis states that cells which mutate to survive the selective pressure do so at random under the selective conditions. Their descendants may not do so and their subsequent resistance distributions will also be random (C). In addition, if resistance mutations are acquired only once a selective pressure is present, the number of resistant cells within a population will remain constant over time (D). This is because each cell has a small, constant and independent probability of sustaining a resistance mutation prior to selection within its lifetime. For simplicity, each cell in Figure 7C only gives rise to a single progeny. Resistant bacteria are indicated by green circles while non-resistant bacteria are indicated by purple circles.

In order to test their hypotheses, Luria and Delbrück argued that if the distribution of mutants within a culture was random, the Poisson distribution would be the appropriate model to apply to such randomly distributed populations. According to this distribution, the variance would be significantly different from the mean of the population, highlighting the fluctuating dispersal of mutants within the population. Their experiment used a single liquid control culture from which a number of aliquots were taken. In parallel, ten independent liquid cultures were set up, from which phage resistant mutants were also recovered. As shown in Table 1, the experimental differences were evened-out in the single culture because the samples reflected the average of all mutation events. Therefore, the number of mutants in each sample was nearly identical and so little fluctuation is observed. In direct contrast, the number of mutants recovered from each of the ten independent parallel cultures shows considerable variation. The amount of fluctuation in the data supported only one of the two hypotheses. When the mean and variance values for the two data sets are compared, it is obvious that the Poisson distribution fitted the second group. Clearly the high degree of fluctuation in the ten parallel cultures supported the notion of spontaneous mutation prior to selection and this indeed was what Luria and Delbrück concluded. Variance was used to measure the extent of fluctuation and for this reason the experiment was called the Fluctuation Test.

Table 1: Luria and Delbrück's Fluctuation Test confirming the Mutation Theory of spontaneous mutation. The mean is the average number of phage resistant bacteria within a set of parallel cultures. The variance is a measure of the fluctuating spread of variables about the mean.

	Number of phage resistant bacteria	
Sample number	**Single control culture**	**Independent parallel cultures**
1	14	6
2	15	5
3	13	10
4	21	8
5	15	24
6	14	13
7	26	165
8	16	15
9	20	6
10	13	10
Mean	**16.7**	**26.2**
Variance	**15.0**	**2178.0**

1.12.2 The Fluctuation Test

The Fluctuation Test determines the distribution of mutant numbers in a parallel set of cultures and the mutation rate is then obtained by analysing that distribution. Various calculations exist to determine the mutation rate, all of which solve for a common parameter *m*. The definition of *m* is the probable number of mutations per culture that gave rise to the observed distribution of mutants. Importantly, *m* is the number of mutational events, and *not* the number of mutant cells that carry a given mutation. A mutation rate (μ) is defined as the probability that a cell will endure a mutation during its lifetime, thus *m* can be converted to μ by dividing it by a function of the total number of cells at risk (Rosche and Foster, 2000). Luria and Delbrück calculated *m* using mean data values and called their algorithm the Luria-Delbrück Method of Means. They devised a second method to estimate *m* that specifically deals with cultures in which no mutants arise or if mutants are selected in liquid medium. This is called the P_o Method as it relates to the Poisson distribution associated with data sets where some tubes yield no mutants (Luria and Delbrück, 1943). Subsequently, Lea and Coulsen calculated *m* using a data set median as their base parameter and called theirs the Lea-Coulsen Method of the Median (Lea and Coulsen, 1949). The accepted gold standard algorithm to estimate *m* is the MSS Maximum Likelihood Method devised by and named after the authors Ma-Sandri-Sarkar (Rosche and Foster, 2000). Once *m* has been calculated using the appropriate method, μ can then be determined.

As with all mathematical models relating to biological processes, there are a number of underlying assumptions inherent in the Fluctuation Test: 1) the probability of a cell sustaining a mutation is constant per cell per lifetime, 2) the probability of a mutation per cell lifetime does not vary during the growth of the culture i.e. the probability of mutation is not affected by different growth phases, 3) the proportion of mutants within a population is always small, 4) the starting number of cells (N_o) is negligible compared to the final number of cells (N_t) in a culture, 5) the growth rates of both mutants and non-mutants are the same, 6) death of any cells in the population is negligible, 7) reverse mutations from mutant back to wild type are negligible, 8) all mutants are detected and 9) no mutants arise after the selection is imposed (Rosche and Foster, 2000). Obviously certain assumptions are not met under all biological conditions, for example: Luria and Delbrück's first assumption only allows for spontaneous mutagenic events. Chemical, biological or physical agents added to the system may cause variable changes in the probability of a cell undergoing a mutation within its lifetime, thereby making their method inapplicable (Kendal and Frost, 1988). However, as long as sound experimental protocol is followed and all methodology is documented and reported, departures from the assumptions can be readily explained. Furthermore, the method used to obtain *m* and μ values should be recorded to allow for standardisation and comparison.

The specifics of a Fluctuation Test can be adapted according to the system required but the general experiment begins with a set of parallel cultures, each inoculated with a small number of identical

non-mutant cells (N_o), that are then grown under optimal conditions. It is important to note that there must be no pre-existing mutants in the culture, thus all tubes in a parallel set should contain the minimum number of cells required to produce a viable culture without containing any initial mutants (Rosche and Foster, 2000). Following the required growth period, the total number of cells in each culture vessel (N_t) is determined by plating dilutions onto non-selective agar. The complete number of mutants per culture tube is ascertained by plating the entire remaining culture onto selective medium. The distribution of mutants can then be used to calculate *m* and subsequently μ (Luria and Delbrück, 1943; Jones *et al.*, 1994; Rosche and Foster, 2000). The precision of the Fluctuation Test lies in the reproducibility of the entire experiment while the accuracy is determined by how close the *m* value is to the actual underlying number of biological mutations that occur in the culture. Both of these parameters depend on the specific method used to generate *m* and μ. Fluctuation Tests therefore provide a tool for the accurate and biologically relevant assessment of underlying mutation mechanisms within bacterial populations.

1.13 Aims of this study

Studies have shown that certain PE-PGRS proteins play a direct role in virulence in *M. marinum* (Ramakrishnan *et al.*, 2000), that PE-PGRS proteins are cell surface associated (Brennan *et al.*, 2001), and that some PE-PGRS proteins are variable surface antigens (Banu *et al.*, 2002). Genome comparisons of clinical isolates of *M. tuberculosis* have also confirmed the polymorphic character of some of these genes, which suggest that these genes may function as heritable local mutators (Metzgar & Wills, 2000) analogous to the contingency loci found in other pathogenic bacteria. These findings underscore the importance of measuring mutation rates at PE-PGRS loci and of investigating the underlying mutational mechanisms. Against this background, the specific aims of this study were as follows:

1. To validate a reporter assay designed to investigate mutations in a PE-PGRS repeat-containing sequence in a *M. smegmatis* host system;
2. To use this assay system to measure the rates of mutations within the PE-PGRS repeat-containing sequence and to investigate the mutational spectra by genotypically characterising the products of mutational events;
3. To investigate the underlying mechanisms of mutagenesis at such loci by determining the rates and spectra of mutations in host strains with specific defects in DNA metabolism.

Chapter 2

Materials and Methods

2.1 Media and general solutions

The details of the growth media as well as general solutions employed in this study are outlined in Appendix A.

2.2 Storage of bacterial strains

All *E. coli* and *M. smegmatis* strains used in this study were stored at -75°C in 33% glycerol (Merck, Germany).

2.3 Culture conditions

The *E. coli* host strain used in this study was DH5α (Hanahan, 1983). *E. coli* cultures were grown with shaking (350 rpm) in Ehrlenmeyer flasks, at 37°C, in a New Brunswick Series 25 orbital incubator (New Brunswick Scientific Co; Inc, USA). *E. coli* strains were grown in Luria Bertani broth (LB) or on Luria Bertani Agar (LA) plates (refer to Appendix A), with the appropriate selective antibiotic (refer to Section 2.4).

Unless otherwise indicated all *M. smegmatis* strains were grown in MADC-Tw [Middlebrook 7H9 broth (Difco, USA - refer to Appendix A) supplemented with ADC (Merck, Germany - 50 g / L albumin, 20 g / L dextrose and 0.04 g / L catalase) and 0.05% Tween 80 (Sigma, Germany); Jacobs et al., 1991] in Ehrlenmeyer flasks or in 15 mL or 50 mL disposable tubes (Sterilin, UK) in a New Brunswick Series 25 orbital incubator. Unless otherwise specified, the solid medium used for culturing *M. smegmatis* was Middlebrook 7H10 agar (Difco, USA – refer to Appendix A) supplemented with OADC (Merck, Germany – 0.5 g / L oleic acid, 50 g / L albumin, 20 g / L dextrose and 0.04 g / L catalase) containing the appropriate selective antibiotic. The *M. smegmatis* strains used in this study are shown in Table 2.

Table 2: *M. smegmatis* strains used in this study

***M. smegmatis* strains**	**Characteristics**	**Source or reference**
mc^2 155	High frequency transformation mutant of *M. smegmatis* ATCC 607	(Snapper, S.B., Melton, R.E., *et al*, 1990)
mc^2 155 ΔrecA	*recA* deletion mutant of mc^2 155; Δ*recA*	Dr. E.E. Machowski
mc^2 155 ΔY	Derivative of mc^2 155 lacking three Y-family EP polymerase-encoding genes; Δ*dinX*, Δ*dinP*, Δ*dinP*::*hyg*, Δ*dinP3*; Hyg^R	Mr. D.F. Warner
mc^2 155 ΔL	Derivative of m^2 155 ΔY carrying a *dnaE2* deletion mutation; Δ*dnaE2*; Hyg^R	Dr. E.E. Machowski
mc^2 155 *his*G380E	His auxotrophic mutant of mc^2 155 carrying a G380E mutation in HisD	Dr. E.E. Machowski

2.4 Medium supplements

Antibiotics were used at the following concentrations: (1) *E. coli*: ampicillin (Amp), 200 µg / mL; gentamicin (Gm), 2 µg / mL in liquid media and 5 µg / mL in agar plates; (2) *M. smegmatis*: Gm, 2 µg / mL in liquid media and 5 µg / mL in agar plates; hygromycin B (Hyg), 50 µg / mL; kanamycin (Km), 25 µg / mL; rifampicin (Rif) 150 µg / mL. All antibiotics were added to cooled media (55°C) prior to pouring plates or were added to liquid media at room temperature. For growth of histidine auxotrophic strains (His-), 7H9 broth and 7H10 agar plates were supplemented with L-histidine (Roche Biochemicals, Germany) at a concentration of 100 µg / mL.

2.5 Plasmid DNA

All plasmids used in this study are listed in Table 3. Plasmid DNA was isolated as detailed in Section 2.7.1. Following isolation, plasmid DNA was stored in deionised sterile water at -20°C.

Table 3: Plasmids used in this study (refer to Appendix B for plasmid maps)

Plasmids	Characteristics	Source or reference
p9/19	Derivative of p164.1 carrying the Morf2 reporter and 'aph (cryptic aminoglycoside phosphotransferase cassette) as a *Xba*I-*Asp*718 fragment; Amp^R	S. Durbach, PhD thesis, University of the Witwatersrand, 2002
pML10	pSUP104 carrying the 2.8 kb Gm^R cassette as a *Bam*HI-*Hin*dIII fragment from pH1IJ; Gm^R	(Labes *et al.*, 1990)
pHINT	Integrative shuttle vector that integrates at the mycobacterium *attB* locus; Hyg^R	(O'Gaora *et al.*, 1997)
pGINTO	Derivative of pHINT carrying the 2.8 kb Gm^R cassette from pML10 as a *Bam*HI-*Hin*dIII fragment; integrative shuttle vector that integrates at the mycobacterium *attB* locus, Gm^R	This work
pGRAK	Derivative of pGINTO; integrating vector that integrates at the mycobacterium *attB* locus, contains the Morf2 reporter fragment and 'aph, Gm^R	This work
pOLYGaa	Derivative of pAGAN11 (obtained from Dr. T. Parish) carrying the inducible acetamidase promoter as a *Bam*HI insert; multicopy shuttle vector; Hyg^R	Mr. D.F. Warner
pOLYGaadinP	Derivative of pOLGYaa carrying the *M. tuberculosis dinP* gene under control of the inducible acetamidase promoter; Hyg^R	Mr. D.F. Warner
pOLYGaadinX	Derivative of pOLGYaa carrying the *M. tuberculosis dinX* gene under control of the inducible acetamidase promoter; Hyg^R	Mr. D.F. Warner

2.6 Transformation of cells

Unless otherwise stated, all microcentrifugations were performed at 10 000 x *g* at room temperature in an Eppendorf 5415D microfuge (Eppendorf AG, Germany).

2.6.1 *E. coli* DH5α

The method used to prepare competent *E. coli* DH5α cells was adapted from Sambrook *et al*. (1989) as follows. *E. coli* cells were grown up in 100 mL of LB until log phase as measured by OD_{600} of 0.4. Cells were then chilled for 20 min on ice. Centrifugation followed for 5 min at 1500 x *g* and 4°C. The pellet was then gently resuspended in 10 mL of ice cold 0.1 M $CaCl_2$ (Merck, Germany) and incubated for 120 min on ice. Cells were then harvested by centrifugation and resuspended in 1 mL of 0.1 M $CaCl_2$ solution. Cells prepared in this way were transformation competent. Transformations were carried out as follows. DNA (50 – 100 ng) was added to a 100 µL aliquot of competent cells and the mixture was incubated on ice for 15 min. The cells were then heat shocked for 90 s at 42°C. After the addition of 800 µL of rich 2TY broth (refer to Appendix A), the cells were incubated for 60 min at 37°C without shaking, to allow for phenotypic expression of antibiotic resistance genes, prior to plating dilutions onto solid media supplemented with the appropriate antibiotic.

2.6.2 Electroporation of *M. smegmatis*

M. smegmatis strains were transformed by electroporation following a protocol adapted from Jacobs Jr *et al*. (1991). A single colony of *M. smegmatis* was inoculated into 100 mL MADC-Tw and grown to log phase (OD_{600} ~ 0.9). The cells were chilled for 20 min on ice and harvested by centrifugation at 4°C for 5 min at 1500 x *g* in a pre-chilled Beckman JA20 rotor (Beckman Coulter, USA). The pellet was resuspended in 40 mL of filter sterilised 10% glycerol and harvested as above. The cells were washed a total of four times in the same manner using 40 mL, 25 mL and 10 mL (twice) of 10% glycerol. After the final wash, the pellet was resuspended in 2mL of 10% glycerol and these electro-competent cells were used immediately. Electroporations were carried out in a BioRad Gene Pulser XCell™ electroporator (BioRad, USA), using 0.2 cm electrode gap cuvettes (BioRad, USA). The following settings were used: 1000 Ω, 25 µF and 2.5kV. Generally, 10 - 100 ng of plasmid DNA was added to a 200 µL aliquot of electro-competent cells, in a pre-chilled electroporation cuvette. Following electroporation, 800 µL of 2TY broth was added. The cells were incubated for 180 min at 37°C without shaking, to allow for phenotypic expression of antibiotic resistance genes, prior to plating appropriate dilutions onto selective solid media.

2.7 DNA Extraction Protocols

2.7.1 Isolation of plasmid DNA

Plasmid DNA was isolated from *E. coli*, as per the methodology outlined in Sambrook *et al*. (1989). Plasmid-containing colonies grown on LA/ Amp agar plates were used to inoculate 5 mL of LB and the culture was grown to stationary phase (± 9 hours). The culture was dispensed into 1.5 mL microfuge tubes (Whitehead Scientific, South Africa) and harvested by centrifugation for 90 s. The supernatants were discarded and the pellets were resuspended in 100 µL of cold (4°C) Solution I (refer to Appendix A). Once resuspended, 200 µL of Solution II (refer to Appendix A) was added and the cells were lysed for 5 min on ice. Thereafter, 150 µL of Solution III (refer to Appendix A) was added to precipitate protein and genomic DNA by incubation for 5 min on ice. The mixture was centrifuged for 15 min and the supernatant transferred to a fresh 1.5 mL microfuge tube. RNase A (Roche Biochemicals, Germany) was added to a final concentration of 100 µg/mL and the mixture was incubated for 15 min at 42°C. An equal volume (400 µL) of isopropanol (Merck, Germany) was added followed by centrifugation for 15 min. The DNA pellet was then resuspended in 450 µL of TE buffer (refer to Appendix A) before being re-precipitated as described in Section 2.7.3.

Following the final precipitation of DNA and resuspension in deionised sterile water, plasmid quality and quantity were assessed spectrophotometrically. The absorbance of the plasmid sample was measured at 260 and 280 nm, with a diluent used to re-zero the spectrophotometer at each wavelength. The quality of plasmid DNA was calculated as a ratio of the absorbance at 260 nm (OD_{260}) to that at 280 nm (OD_{280}). The ratio for pure DNA was 1.8 and any contamination by protein or phenol caused the ratio to be significantly lower than this value. The quantity of plasmid DNA (ng / µL) was calculated by multiplying the OD_{260} value by the product of the dilution factor and 40.

2.7.2 Cetyltrimethylammoniumbromide (CTAB) method for isolation of chromosomal DNA from *M. smegmatis*

M. smegmatis was grown in liquid medium overnight to early stationary phase (± 16 hours). Cells were heat-killed for 30 min at 65°C. The cells were harvested by centrifugation for 90 s and the pellet was resuspended in 450 µL of TE buffer containing 50 µL of lysozyme (10 mg / mL - Roche Biochemicals, Germany). The mixture was then incubated for 60 min at 37°C. Following incubation, 70 µL of 10% SDS (Fluka, Switzerland) and 6 µL of proteinase K (10 mg / mL - Roche Biochemicals, Germany) were added and the mixture was incubated for 120 min at 65°C. CTAB / NaCl solution (Separations, South Africa / Merck, Germany - refer to Appendix A) was pre-warmed while 100 µL of 5M NaCl was added to the mixture. Once warmed, 80 µL of CTAB / NaCl solution was added followed by further incubation for 10 min at 65°C. An equal volume (700 µL) of chloroform:isoamyl alcohol solution (24:1 - SMM Instruments, South Africa / Merck, Germany) was then added and the mixture

was centrifuged for 5 min. The upper aqueous phase was transferred to a fresh 1.5 mL microfuge tube and an equal volume (~500 µL) of isopropanol was added. The DNA was recovered by centrifugation for 20 min. The pellet was resuspended in 450 µL of TE buffer and then precipitated again as described in Section 2.7.3.

2.7.3 Precipitation of DNA

Following plasmid or chromosomal DNA isolation, pellets were resuspended in 450 µL of TE buffer. The pellet was then precipitated again by the addition of one tenth the volume of 3M sodium acetate (pH 5.2 – Merck, Germany) and 2.5 volumes of ice cold (-20^{o}C) absolute ethanol (BDH Laboratories, UK or Merck, Germany). Centrifugation followed for 20 min at 4^{o}C. The pellet was then washed in 1mL of 70% ethanol, vacuum dried in an Eppendorf Concentrator 5301 vacutainer (Eppendorf AG, Germany) and resuspended in an appropriate volume of deionised sterile water.

2.8 Agarose gel electrophoresis

2.8.1 Preparation of gels

1% agarose gels (Separations, South Africa) and 4% Nusieve® GTG® gels (Cambrex, USA) were prepared as per the methodology outlined in (Sambrook *et al.*, 1989). Briefly, agarose powder was mixed with an appropriate volume of cold (4^{o}C) TE buffer. The mixture was heated until all the gel powder had dissolved. Prior to pouring, 0.5 µL per 30 mL gel of ethidium bromide (EtBr – 10 mg / mL – Sigma, Germany) was added. Nusieve® GTG® gels were allowed to set overnight prior to use.

2.8.2 Electrophoresis

Both 1% agarose and 4% Nusieve® GTG® gels were run in Hoefer gel tanks (Hoefer, UK). The walls of the gel tanks were filled with polyethylene glycol (PEG – Merck, Germany) to ensure even heat conduction. All DNA samples, controls and appropriate molecular weight markers (Molecular Weight Marker λIII and λIV – Roche Biochemicals, Germany) were mixed with loading dye (refer to Appendix A) and loaded into the wells of the gel. Gel tanks were then sealed and connected to a power pack (BioRad, USA). For the separation of chromosomal DNA and pGRAK reporter-based PCR products, gels were run for 2.5 hours at 80V. For the separation of *his*G380E reporter-based PCR products and restriction endonuclease digests, gels were run for 1 hour at 100V. Electrophoresis products were viewed using a GelDoc system (BioRad, USA).

2.8.3 Purification of DNA from agarose gels

DNA fragments were separated by agarose gel electrophoresis and were purified either by agarase digestion (Roche Biochemicals, Germany) or using a Geneclean Kit (Q-Biogene, USA), as follows:

(A) Agarase protocol. DNA fragments were separated using Seaplaque® GTG low melting agarose (FMC Bioproducts, USA). The DNA fragment was excised from the gel and the slice weighed. The gel slice was incubated for 10 min at 65°C until melted when 4 µL of agarase buffer per 0.1 g of gel was added. The mixture was then moved to 42°C and 1 µL of agarase enzyme was added. The mixture was incubated for 30 min with agitation at 42°C. A further 0.5 µL of agarase was added and the mixture incubated for a further 30 min with agitation at 42°C. The mixture was placed on ice for 15 min to allow any residual gel to set. The mixture was then centrifuged for 15 min at 4°C, the supernatant was transferred to a fresh microfuge tube and the DNA was precipitated.

(B) Geneclean protocol. The DNA band was excised from a 1% agarose gel and the gel slice weighed to determine the volume of 6M sodium iodide (NaI) to be added (3 X volume of NaI to DNA solution). The gel slice with NaI was incubated for 10 min at 65°C until melted whereupon 5 µL of glassmilk was added. The mixture was incubated for 5 min at room temperature followed by centrifugation for 10 s. The supernatant was removed and the pellet was resuspended in 700 µL of New Wash® followed by centrifugation for 30 s. The New Wash® steps were repeated for a total of four washes. After the final centrifugation, the supernatant was removed and the pellet vacuum dried in an Eppendorf Concentrator 5301 vacutainer prior to being resuspended in an appropriate volume of deionised sterile water.

2.9 DNA manipulations

2.9.1 Restriction endonuclease digests

Plasmid DNA was digested for 60 min with the appropriate restriction endonucleases (AEC Amersham, UK or Roche Biochemicals, Germany) in supplied buffers at 37°C, unless otherwise specified by the manufacturer. Where necessary, appropriate volumes of bovine serum albumin (BSA) and phosphate buffered saline (PBS) were also included (Roche Biochemicals, Germany). *M. smegmatis* chromosomal DNA was digested overnight with the appropriate restriction endonucleases in supplied buffers at 37°C.

2.9.2 Blunt-ending of 3' or 5' overhangs

For blunt-ending of 5' overhangs, 0.5 U of T7 DNA polymerase (Promega, USA) and an appropriate volume of supplied buffer were added to linearised DNA, followed by incubation for 10 min at 37°C.

For blunt-ending of 3′ overhangs, one unit of Klenow enzyme (Roche Biochemicals, Germany) and an appropriate volume of supplied buffer which included dNTPs (Sigma, Germany) were added to linearised DNA, and the mixture was incubated for 20 min at 30°C.

2.9.3 Removal of 5′ phosphate from DNA

Either shrimp alkaline phosphatase (SAP – Roche Biochemicals, Germany) or calf intestinal phosphatase (CIP – Roche Biochemicals, Germany) was used to 5′-dephosphorylate DNA. Briefly, a maximum of 4 µL of buffer and 2 µL of enzyme was added to 36 µL of restricted DNA. The mixture was incubated for 30 min at 37°C. If SAP was used, the mixture was then placed at 60°C for 10 min to inactivate the enzyme. If CIP was used, following incubation at 37°C, water was added to a final volume of 100 µL. The enzyme was removed by phenol / chloroform (Merck, Germany / SMM Instruments, South Africa) extraction as described in Section 2.7.3.

2.9.4 Ligations

Ligations were carried out using the Fastlink™ DNA Ligation Kit (Epicentre Technologies, USA) according to the manufacturer's instructions. Briefly, linearised vector DNA was blunt ended using T7 DNA polymerase or Klenow enzyme where appropriate. The vector was then 5′ -dephosphorylated using CIP or SAP. Both the linearised insert and vector DNA were run on 1% agarose gels to determine their concentrations relative to a standardised molecular weight marker. Molar ratios of DNA were calculated using the following equation:

[(ng vector x kb size of insert) / kb size of vector] x ratio wanted = ng insert

For sticky ended ligations, 1:1 to 1:3 ratios of DNA were used with the appropriate volume of supplied buffer and the total required amount of T4 ligase. These ligation reactions occurred for 30 min at room temperature. For blunt-ended ligations, 1:5 to 1:10 ratios of DNA were used with the appropriate volume of supplied buffer and half the required amount of T4 ligase. These ligation reactions occurred for 90 min at room temperature.

2. 10 The Morf2 reporter assay

2.10.1 Design of the Morf2 substrate

As described in Figure 4 (refer to Section 1.8), the *M. bovis* BCG PE-PGRS gene corresponding to the *M. tuberculosis* gene Rv0746 has a 29 amino acid deletion in the PGRS portion of the protein. In preliminary work, Dr. S. Durbach designed a reporter assay construct for investigating mutational mechanisms at PE-PGRS loci in *M. smegmatis* (S. Durbach, PhD thesis, University of the

Witwatersrand, 2002). The reporter was modeled on the deletion in Rv0746, and included an engineered stop codon within the coding sequence. The reporter was cloned into an integrating vector that uses the phage attachment site (*attP*) on the plasmid and the bacterial attachment site (*attB*) of *M. smegmatis* mc^2 155 to integrate into the bacterial chromosome. The reporter substrate is GC rich and comprised of five regions, organised as shown in Figure 8. The 5'-end of the sequence has a *M. tuberculosis* antigen 85A (Ag85A) promoter region (P_{Ag85A}), ribosome binding site (RBS) and ATG start codon. The third base of the ATG start codon (G) forms the first base of a *Bam*HI site (G↓GATCC) that in turn marks the start of the GC substrate's fourth region, designated as Morf2. The Morf2 sequence contains three PGRS repeat elements ($^{5'}$GGCCGGGGCCGG$^{3'}$) that each encode a Gly-Gly-Ala-Gly-Gly repeat and are designated as R1, R3 and R2, respectively. Repeats R1 and R3 are separated by a 15 bp intervening sequence (IV) while R3 and R2 are separated by 7 bp. The 15 bp IV between R1 and R3 was modified from the native *M. tuberculosis* Rv0746 sequence to contain a point mutation that converts a serine to a stop codon (T**C**G → T**A**G). The 3' end of the Morf2 region is marked by a *Nhe*I restriction site and forms a translational fusion with a cryptic aminoglycoside phosphotransferase gene ('*aph*) that encodes Km resistance (Km^R). The '*aph* gene is cryptic as it does not contain a start codon.

In the assay system, any expression of functional reporter activity (observed as Km^R) under the control of the Ag85A promoter, requires a mutational event that bypasses the TAG stop codon between R1 and R3 to form a translational fusion to the Aph reporter.

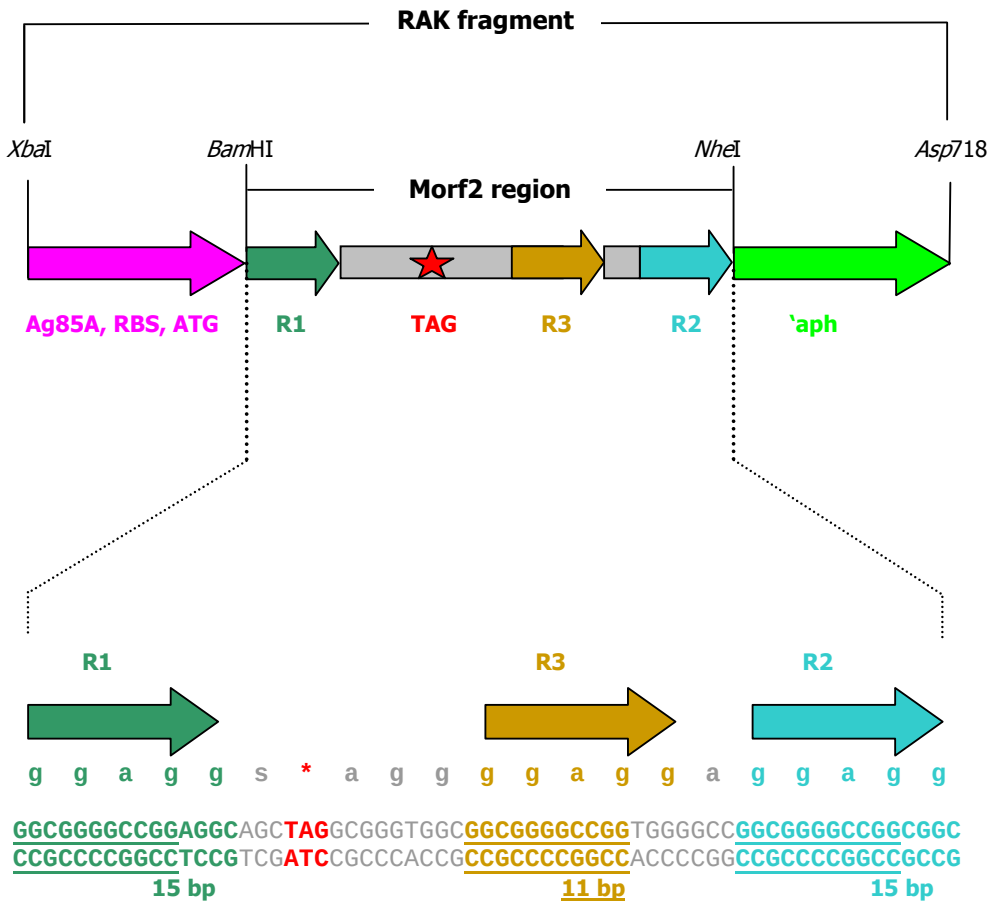


Figure 8: Design of Km^R based reporter, pGRAK that integrates at the *attB* locus of *M. smegmatis* mc^2 155 strains. The RAK fragment is defined by a *Xba*I site at the 5′ end and an *Asp*718 site at the 3′ end. The pink arrow indicates the Antigen 85A promoter (Ag85A) with ribosome binding site (RBS) and ATG start codon. The Morf2 region is defined by a *Bam*HI site at the 5′ end and a *Nhe*I site at the 3′ end. This region is comprised of three PE-PGRS repeat sequences, called R1 (green block), R3 (yellow block) and R2 (blue block) respectively, and an engineered TAG stop codon (red star) in the intervening sequence (grey block) between R1 and R3. The 3′ end of the Morf2 region is fused in frame to a cryptic Km^R cassette ('aph) indicated by the bright green arrow. A closer look at the DNA sequence making up the Morf2 region indicates that the PE-PGRS repeats, R1, R2 and R3 share an 11 base pair (bp) core repeat sequence (indicated by underlining of bases). R1 and R2 share a 15 bp repeat sequence, except for one base difference. The red bases and stars indicate the TAG stop codon. The green bases and arrows indicate repeat 1 (R1). The yellow bases and arrows indicate repeat 3 (R3). The blue bases and arrows indicate repeat 2 (R2).

2.10.2 Cloning of the pGRAK reporter

Prior work had indicated that although the Morf2 substrate was functional, the Gm resistance cassette contained in the integrative vector carrying Morf2 was not fully functional and did not allow selection for the vector in *M. smegmatis* grown in liquid media supplemented with Gm. To address this problem, the pGRAK vector, which contains a Gm resistance cassette from an alternative source was constructed as follows. The vector, pHINT, was digested with *Eco*RI (G↓AATTC) to generate two fragments, one of which lacked the Hyg resistance gene (*hyg*). This fragment, designated as INTO (4.2 kb), was isolated, and the ends were blunted using Klenow enzyme. A Gm resistance cassette was excised from pML10 as a 3 kb *Pst*I (CTGCA↓G) fragment and the ends were blunted using T7 polymerase. The resulting fragment, Gm3, and was ligated to INTO to create the plasmids, pGINTOa and pGINTOb, which differ in the orientation of the Gm resistance gene. Since pGINTOa (but not pGINTOb) conferred resistance of *E coli* and *M. smegmatis* to Gm, this vector was selected for further use.

The RAK fragment was excised from p9/19 as a 1.1 kb *Xba*I/*Asp*718 (T↓CTAGA / G↓GTACC) cassette and blunt-ended using Klenow enzyme. pGINTOa was digested using *Hin*dIII (A↓AGCTT) followed by CIP or SAP removal of the 5' phosphate. The vector and blunt-ended RAK insert were ligated to create the plasmid, pGRAK. To confirm the integrity of the Morf2 region of pGRAK, PCR primers (ANAF / ANAR) were designed to flank the *Bam*HI and *Nhe*I sites that marked the 5' and 3' ends of the region, respectively. The region was amplified by PCR and the product was sequenced using the sequencing primer, MFSF. This analysis confirmed that the Morf2 sequence in pGRAK was correct. This, together with the fact that pGRAK conferred Gm resistance in *M. smegmatis* during growth in liquid media, confirmed the suitability of this reporter plasmid for use in mutational analysis in a Fluctuation Test format.

2.11 Fluctuation Tests

The optimisation of the methodology for measuring mutation rates in *M. smegmatis* by fluctuation analysis formed a large part of this study. Thus a general protocol is discussed below but specific details related to each bacterial strain used are contained in Section 3.2 as well as Appendix C.

2.11.1 Experimental protocol for mutation rate assessment

A single colony of *M. smegmatis* was picked from a fresh electroporation or streaked plate and inoculated into 5 mL of MADC-Tw in a 50 mL Ehrlenmeyer flask. The culture was grown to an OD_{600} of 1.2 to 1.5. Within this OD_{600} range, the density of the culture was expected to be ~ $Log_{10}8$ CFU / mL (refer to Section 3.1.1). The actual cell density in the inoculum was confirmed by plating duplicate

serial dilutions onto Middlebrook 7H10 plates. Appropriate dilutions of the inoculum were also plated onto Middlebrook 7H10 plates supplemented with relevant antibiotics used for selective growth to determine the number of pre-existing mutants (r) within the inoculum culture. Incubation of plates followed at 30^{o}C and subsequent colony forming units (CFUs) were scored. The r value of the inoculum determined the dilution required to ensure that the N_o (initial) culture did not contain any pre-existing mutants.

To circumvent the technical problem of clumping, common to aged mycobacterial cultures, an elaborate method was used to dilute the inoculum from ~Log_{10}8 CFU / mL to ~Log_{10}2 CFU / mL as follows. 1 mL of the inoculum (~Log_{10}8 CFU / mL) was transferred to a 250 mL Ehrlenmeyer flask containing a magnetic stirrer bar and 99 mL of fresh media containing appropriate antibiotics where necessary. This created a culture with ~Log_{10}6 CFU / mL. The culture was continually stirred using a magnetic stirrer pad (Boeco, Germany). Whilst stirring, 1 mL of the Log_{10}6 CFU / mL culture was transferred to another flask as before to create a culture containing ~Log_{10}4 CFU / mL. Whilst stirring, 10 mL of the Log_{10}4 CFU / mL culture was transferred to a 250 mL Ehrlenmeyer flask containing a magnetic stirrer bar and 90 mL of fresh media containing appropriate antibiotic where necessary. This created a culture with ~Log_{10}3 CFU / mL. The same procedure continued until the final culture contained ~Log_{10}2 CFU / mL. These latter two cultures, containing Log_{10}3 CFU / mL or Log_{10}2 CFU / mL were designated as the N_o cultures. The actual number of cells per mL in the N_o cultures were determined by plating neat aliquots onto duplicate Middlebrook 7H10 plates.

To ascertain N_t values (final population numbers), 1.5 to 2.8 mL of each of the N_o cultures were aliquoted into each of 7 to 30 parallel 15 mL disposable culture tubes. The number of tubes used depends on the *m* value. Generally, the higher the *m* value, the fewer number of tubes that are needed to calculate μ (Figure 9). The initial Fluctuation Tests in this study were completed with 30 parallel tubes but once the parameters of the Fluctuation Test had been refined, sets of 7 parallel tubes were used to generate reliable μ values. The culture tubes were placed in a beaker of water, secured with paper towel and grown to log (2 days or ~ 48 hours growth) and saturation (6 days or ~ 114 hours growth) phase. At log phase, appropriate serial dilutions from each tube were plated onto Middlebrook 7H10 plates. These plates were incubated at 30^{o}C and subsequent CFUs scored as N_t values. The entire remaining culture within each tube (~1 to 2 mL) was grown as before to saturation phase. At saturation phase, appropriate serial dilutions from each tube were plated onto Middlebrook 7H10 plates. These plates were incubated at 30^{o}C and subsequent CFUs scored as N_t values. The entire remaining culture within each tube was plated neat onto Middlebrook 7H10 plates supplemented with relevant antibiotics used for Km^R mutant selective growth, incubated at 30^{o}C and subsequent CFUs scored as mutants. Prior to creating serial dilutions, all 15 mL culture tubes containing culture were vortexed for a minimum of 30 s. All serial dilutions were done using filter sterilised 0.5% Tween 80 and when creating a serial dilution series, all 1.5 mL microfuge tubes containing solution were vortexed for a minimum of 5 s per tube.

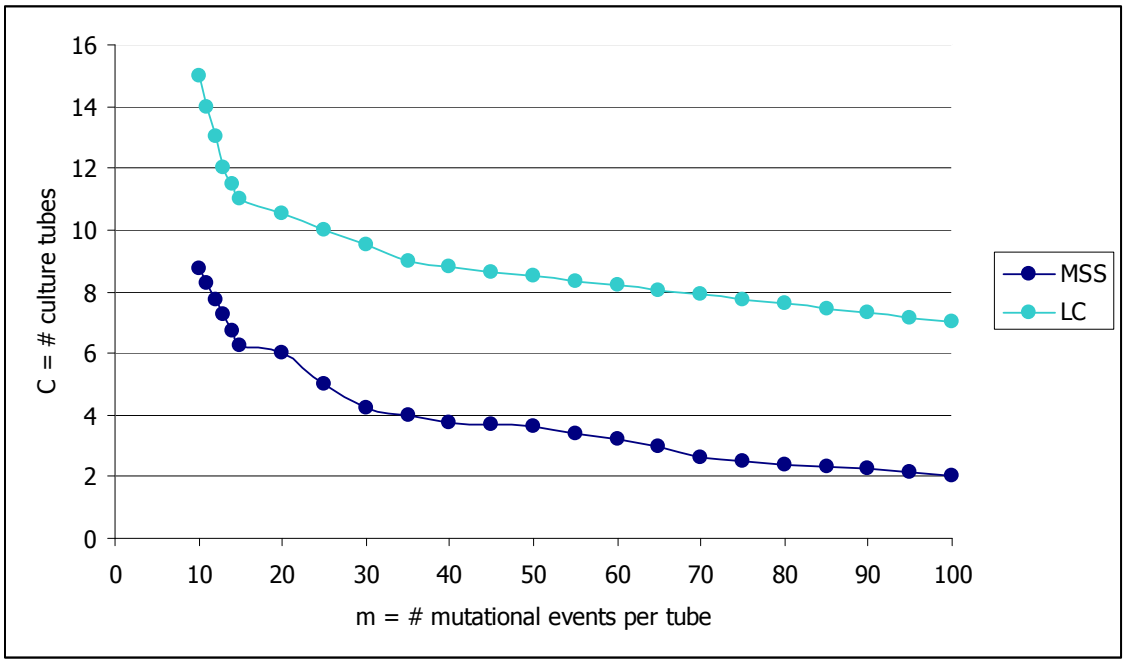


Figure 9: Number of culture tubes (*C*) versus *m*, the number of mutational events per culture tube. For any *m* value between 10 and 100, the number of cultures tubes required to achieve a theoretical precision of 20% using the Lea-Coulsen Method (LC – shown in light blue) or MSS Maximum Likelihood Method (MSS – shown in dark blue) can be determined from the graph. Using either method to generate *m*, the number of cultures tubes required decreases as the *m* value increases. However, due to the iterative nature of the MSS Method, the number of culture tubes required by this method is lower overall.

2.11.2 Adaptation of experimental protocol for mutation rate assessment of mc^2 155 *his*G380E strain

Fluctuation Tests on the *M. smegmatis* mc^2 155 *his*G380E strain were performed using the same general protocol as described in section 2.11.1, but with the following modifications. Prior to assessing mutant numbers in the N_o and N_t populations, cultures were washed three times in 0.5% Tween 80 to remove residual L-histidine supplement in the media. The washed aliquots were plated onto plain Middlebrook 7H10 plates, grown at 30°C and subsequent CFUs scored as *his* revertants, i.e. colonies that had reverted from His auxotrophy (His-) to His prototrophy (His+).

2.11.3 Calculation of mutation rates

Many methods exist to calculate μ, each of which relies on the generation of the variable *m*. For this study, the Lea-Coulsen Method of the Median was used to generate *m* values for any set of data where the number of mutants per tube (r) was greater than zero (r > 0). The P_o Method was used to generate *m* for any data set where at least one tube contained no mutants (r = 0). Once the initial *m* value had been estimated using either of these methods, the *m* values were refined by iteration using the MSS Maximum Likelihood Method. From these iterative values, an *m* value (that maximised Equation 17) was determined which in turn was used to calculate a μ value. A set of Microsoft Office Excel spreadsheets was generated using algorithms adapted from Rosche and Foster (2000) to calculate the initial *m* value using either the P_o or the Lea-Coulsen methods as follows:

- P_o Method:

P_o = (number of tubes yielding r = 0) / total number of tubes……Equation 1
Example: P_o = (2 / 15)

From Equation 1:

$m = - \text{Ln} (P_o)$……Equation 2
Example: $m = - \text{Ln} (2 / 15) = 2.02$

- Lea-Coulsen Method of the Median:

$f / m - \text{Ln} (m) = 1.24$……Equation 3

Iterative values of *m* were generated by the MSS Maximum Likelihood Method according to:

$P_o = e^{-m}$……Equation 4
$Pr = m / r [Pi / (r - 1 + 1)]$……Equation 5
where $1 < r < 150$

$$f(r|m) = \prod_{i=1}^{C} f(r_i|m)$$……Equation 6

where $f(r|m) = Pr$

Example from Equations 2 and 5:

r = 1: $P1 = 2 / 1 [P0 / (1 - 0 + 1)]$

r = 2: $P2 = 2 / 2 [P0 / (2 - 0 + 1) + P1 / (2 - 1 + 1)]$

r = 3: $P3 = 2 / 3 [P0 / (3 - 0 + 1) + P1 / (3 - 1 + 1) + P2 / (3 - 2 + 1)]$ etc.

r remained the same for each series and i increased by 1 for each term within the series of Equation 5. For any value of r > 150, the proportion added to the total figure was so small that all r > 150 values were fitted into the equation where r = 150.

The Pi values generated from Equation 5 were then multiplied by the power of the rank of the data (Equation 17 in Rosche and Foster, 2000) to generate iterative *m* values. The values from Equation 17 were plotted against the iterative *m* values to generate a graph (Figure 10). The *m* value that correlated with the maximum Equation 17 value was read off the graph and used to calculate μ.

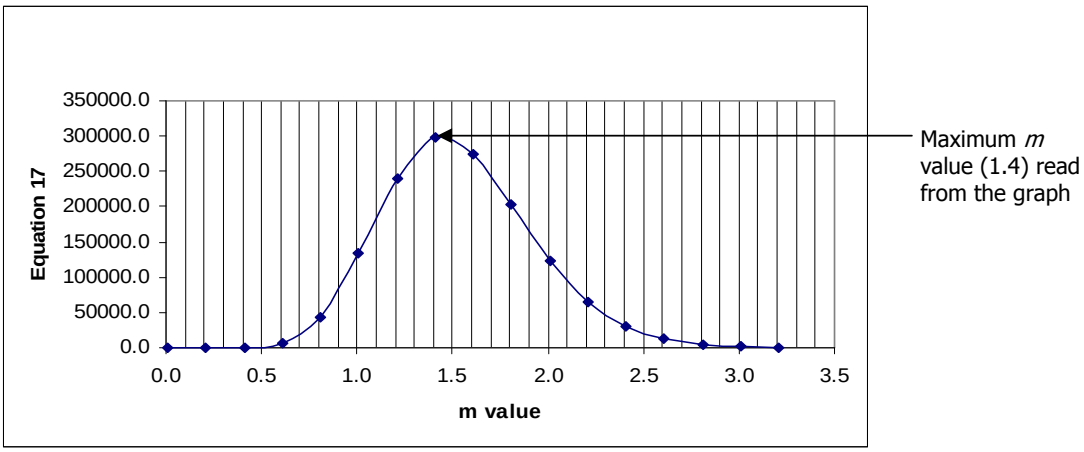


Figure 10: The power of the rank of a data set (Equation 17) is plotted against iterative *m* values to generate a 2 tailed curve. The *m* value that correlates with the maximum Equation 17 value is read from the graph and this value is used to generate μ. In the example illustrated, the initial *m* value is 1.6 while the *m* value that maximises Equation 17 from the graph is 1.4.

- Calculation of μ

The *m* value that correlated with the maximum Equation 17 value generated from the MSS Maximum Likelihood Method was used to calculate μ according to:

$\mu = m / (1.44 \times \text{average } N_t)$.....Equation 7

Example: $\mu = 1.40 / (1.44 \times 6.60 E + 09)$

$\mu = 1.47 E - 10$

The classical representation of Luria and Delbrück data is illustrated as Log (Pr) versus Log (r) where:

$Pr = 2m / r$.....Equation 8

The data indicates the relationship between the number of mutational events within a tube and the number of CFU's that were scored on a plate (Figure 11).

Example: If *m* = 1 and there were 50 Km^R CFUs scored on a plate, the relationship between the variables would not be tight and the conclusion from the data would be that the Luria Delbrück model is poor as it is not a good reflection of what happened in the tubes to give rise to the observed distribution of mutants. However, as the relationship between r and Pr tightens, the R^2 value approaches 1. This suggests that the Luria Delbrück model is a good indication that the observed values are close to the expected values. All data sets for all pGRAK based host strains showed R^2 values = 1 (data not shown). The slope of the graph in Figure 11 correlates with the time point at which the *m* and μ values were assessed. At log phase (Figure 11A) the slope is steep because fewer mutational events (*m*) led to the observed mutants (r). At saturation phase (Figure 11B), the results are analogous: the slope is less steep as a greater *m* value led to the corresponding r values.

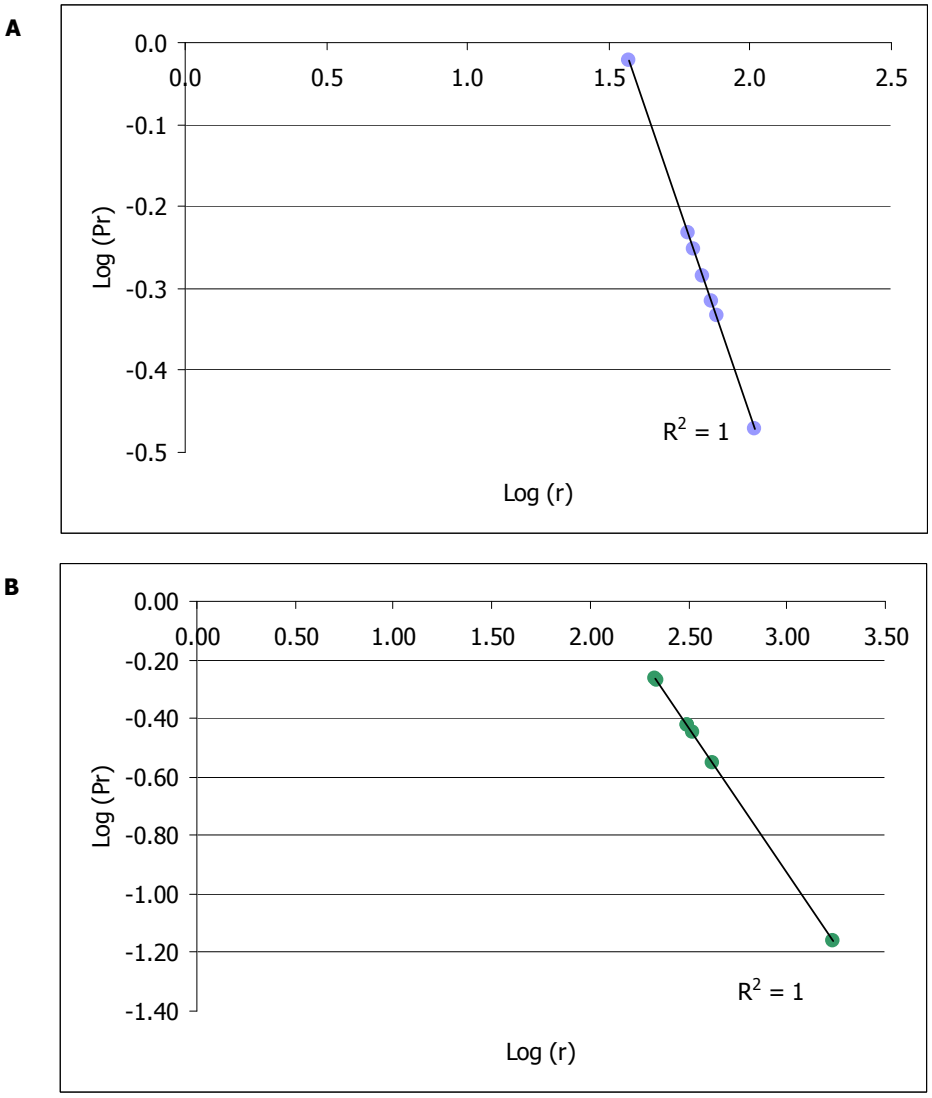


Figure 11: Classical representation of the Luria Delbrück model. The data indicates the relationship between the number of mutational events (*m*) within a tube and the number of Km^R CFU's that were scored (r). The relationship between r and Pr is tight, i.e. the R^2 value is close to 1 if the Luria Delbrück model is a good representation of the underlying biological events. The slope of the graph indicates the time point at which the data was collected. (A) During log phase growth, fewer mutational events lead to the observed number of Km^R mutants and so the slope of the graph is steep. (B) At saturation phase growth, the slope is less steep as more mutational events occurred to give rise to the observed Km^R mutants.

2.12 PCR and sequencing

The oligonucleotide sequences used in this study are listed in Table 4.

Table 4: Oligonucleotide sequences used for PCR and sequencing

Name	Sequence (5' to 3')
ANAF	GGTTGACTACACGAGCACTG
ANAR	CTCGCGATAATGTCGGGCAA
attBS1	ACGTGGCGGTCCCTACCG
attBS2	ACAGGATTTGAACCTGCGGC
attL2	CTTGGATCCTCCCGCTGCGC
attL4	AATTCTTGCAGACCCCTGGA
DF2	CCACCGACCGCGAACTGGCC
HisCR	CGCCCCTGCAGTACGGCGACTTGCCCCGC
MFSF	GCAGTCTGACCTAATTCAGG

2.12.1 PCR of pGRAK based strains

To assess the spectrum of mutations within the pGRAK Morf2 region, Km^R colonies within each pGRAK based Fluctuation Test were picked at random from Middlebrook 7H10 plates supplemented with Km^{25}. Chromosomal DNA was extracted by boiling each colony for 20 min in a solution of 40 µL chloroform and 20 µL of deionised sterile water at 100^{o}C, followed by centrifugation for 5 min. An aliquot from the aqueous phase was used immediately as PCR template or alternatively, stored at -20^{o}C. A forward (ANAF) and a reverse (ANAR) PCR primer set was used to amplify a 282 bp sized fragment spanning the Morf2 region within the pGRAK reporter construct. A forward (ANAF) and reverse (RK aph) primer set were used to amplify a 1002 bp region spanning the Morf2 region and the cryptic *aph* cassette (3' end of cassette marked by *Kpn*I site) within the pGRAK reporter (Figure 12).

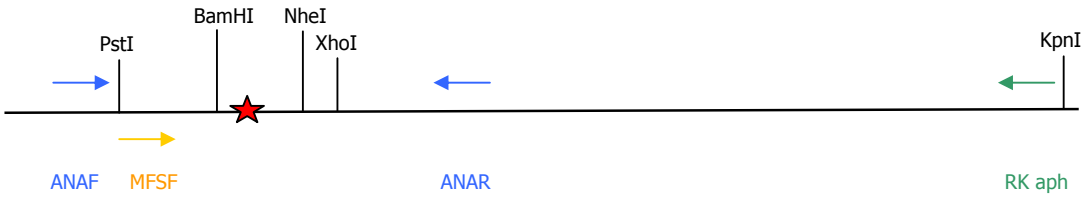


Figure 12: Region of the pGRAK reporter plasmid showing the Morf2 region (between *Bam*HI and *Nhe*I sites) and downstream cryptic 'aph cassette (*Nhe*I to *Kpn*I) with relevant PCR and sequencing primers. ANAF and ANAR are forward and reverse PCR primers respectively that amplify a 282 bp product. MFSF is a forward sequencing primer. RK aph is a second reverse PCR primer used in conjunction with ANAF to amplify the Morf2 region and the cryptic 'aph cassette. The red star indicates the TAG stop codon within the Morf2 region.

A master reaction mix was created using ingredients from a Faststart™ Kit (Roche Biochemicals, Germany) for each PCR as follows: 6.6 µL of deionised sterile water, 1.5 µL of buffer supplemented with $MgCl_2$ and 3 µL of 5 X GC rich buffer were aliquoted into a microfuge tube cooled on ice. To this mixture, 2.4 µL of dNTPs (1.25 uM each) and 0.75 µL of each of the forward and reverse primers (10uM) were added. Finally, 0.1 µL of Faststart™ TAQ polymerase was added to the mixture. All ingredients were mixed by gentle flicking of the microfuge tube. A total volume of 15 µL of master mix was aliquoted per 0.2 mL PCR tube (Whitehead Scientific, South Africa). To this was added 2 µL of PCR template. Despite the fact that Faststart™ TAQ polymerase is stable at room temperature (24°C), all PCR reactions were set up as quickly as possible and all ingredients and tubes were maintained on ice.

All PCR reactions were carried out using either a Hybaid PCR Express (Hybaid, UK) or an Eppendorf Mastercycler Gradient (Eppendorf AG, USA) thermocycler. Initial denaturation for 4 min at 95°C ensured that all PCR templates present had been denatured. A set of 40 cycles then followed with each cycle comprising a denaturation step at 94°C for 25 s, an annealing step at 60°C for 25 s and an elongation step at 72°C for 30 s. Once the final cycle had been completed, a final elongation step for 5 min at 72°C was performed to ensure that all PCR templates had been completely replicated. All PCR products were then maintained at a holding temperature of 15°C until they were either analysed by gel electrophoresis or transferred to a freezer for storage at -20°C.

2.12.2 PCR of *M. smegmatis* mc^2 155 *his*G380E strain

The auxotrophic phenotype of the mc^2 155 *his*G30E mutant of *M. smegmatis* is due to a GGG (Gly) → GAG (Glu) mutation at amino acid 380 of the *hisD* gene involved in His biosynthesis. This mutation corresponds exactly to the *his5* mutation described by Hinshelwood and Stoker (1992). Since the Gly380 residue is essential for HisD function, this auxotrophic mutant provides a system for measuring the rate of one particular base substitution mutation, namely A → G, which results in reversion to His prototrophy. To confirm the genotype of revertants, colonies from each mc^2 155 *his*G380E based Fluctuation Test were picked at random from Middlebrook 7H10 plates, chromosomal DNA was extracted as described in Section 2.12.1 and an aliquot of the supernatant was used immediately as PCR template or stored at -20°C for later use. A forward (DF2) and reverse (HisCR) PCR primer set was used to amplify a 536 bp sized fragment spanning codon 380 within the *hisD* gene of *M. smegmatis*. A master reaction mix was created for each PCR as described in Section 2.12.1. All PCR reactions were carried out as described in Section 2.12.1, but used 30 cycles only of the denaturation, annealing and elongation steps. PCR products were separated by gel electrophoresis, purified and sequenced.

2.12.3 PCR of *attL* and *attR* regions

Integrating vectors, such as pGRAK, use the L5 phage derived attachment site (*attP*) located on the plasmid and the bacterial attachment site (*attB*) located on the chromosome of the host bacteria to integrate (Pena *et al.*, 1997). During integration, the *attP* and *attB* sites are rearranged by the phage encoded integrase to create the left (*attL*) and right (*attR*) sites. To ensure correct integration of the pGRAK reporter into the various *M. smegmatis* host strains, PCR primers were designed to flank the *attL* and *attR* sites. A forward (attBS2) and reverse (attL4) PCR primer set was used to amplify a 320 bp sized fragment spanning the *attL* region while a forward (attL2) and reverse (attBS1) PCR primer set was used to amplify a 282 bp sized fragment spanning the *attR* region (Figure 13). Analysis of PCR products was based only on size according to gel electrophoresis.

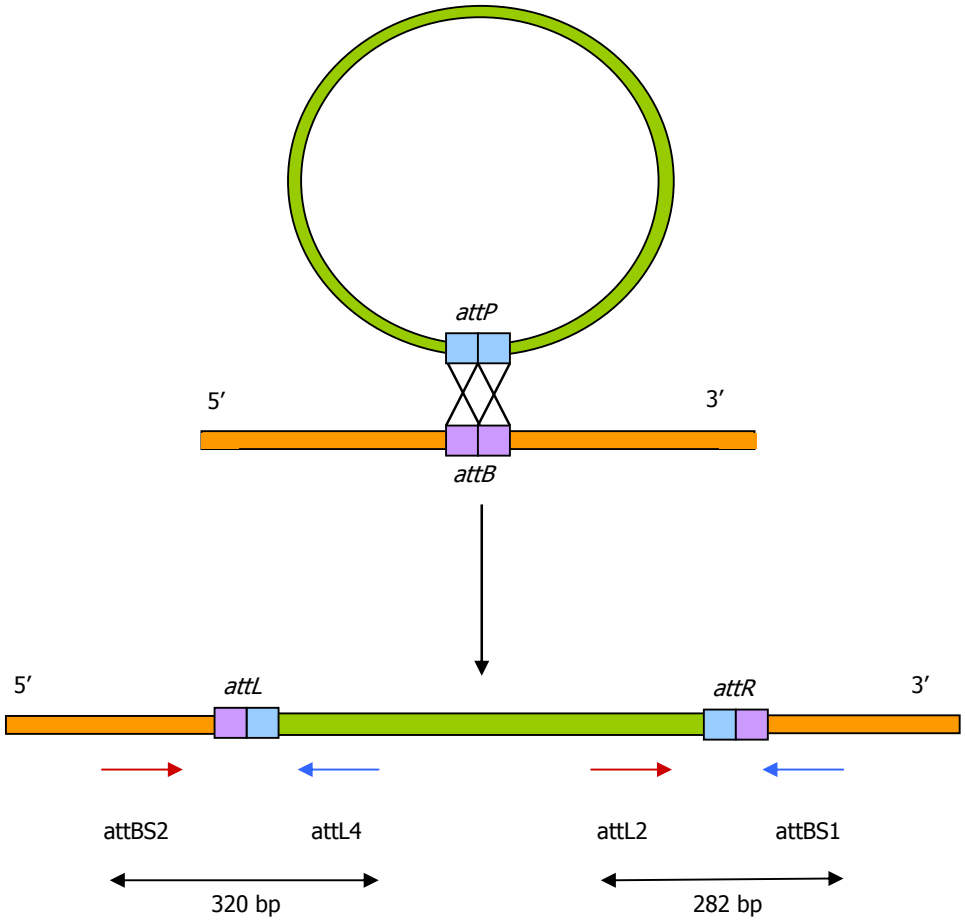


Figure 13: Mechanism of integration of L5 derived vectors into the *M. smegmatis* mc^2 155 chromosome. The *attP* site on the plasmid (indicated by blue blocks on the closed green circle) lines up with the corresponding *attB* site on the bacterial chromosome (indicated by the purple blocks on the orange line). Using the phage integrase enzyme (encoded on the plasmid) the *attP* and *attB* sites are recombined, generating the *attL* and *attR* sites. All genetic material contained within the plasmid is linearised and becomes part of the bacterial chromosome. PCR primers (indicated by blue and red arrows) were designed to flank the *attL* and *attR* regions to ensure correct integration of the pGRAK vector into the *M. smegmatis* based strains used in this study.

2.12.4 DNA sequencing

All sequencing was carried out by Inqaba Biotech (Pretoria, South Africa) on a Spectrumedix 2410 Capillary Electrophoresis automated DNA sequencer using Big Dye Terminator V3.1 software from ABI for data analysis. A MFSF primer was used to sequence Morf2 PCR products. A HisCR primer was used to sequence *his*G380E PCR products.

Chapter 3

Results

3.1 Optimisation of experimental parameters

Prior to analysis of mutation rates by Fluctuation Tests, it was necessary to establish the growth parameters of *M. smegmatis* mc^2 155 in liquid media and to optimise the culture conditions in order to obtain reproducible N_t values.

3.1.1 Growth curve of *M. smegmatis* mc^2 155

To test the growth parameters of *M. smegmatis* mc^2 155, a 24-h growth curve was completed. At 3 hourly intervals, samples were removed from a batch culture for CFU and OD_{600} determination. The log phase growth of *M. smegmatis* mc^2 155 is defined by CFU / mL counts of $Log_{10}7$ to $Log_{10}8$ (Kana and Mizrahi, 2004). Over the 24-h growth period, the CFU / mL counts increased from $Log_{10}6$ to $Log_{10}8$ (Figure 14). The exponential phase of growth, as defined above, occurred between 12 and 20 hours post inoculation. Thereafter, the culture entered saturation phase. All log phase cultures used in Fluctuation Tests in this study were defined as such by their number of hours (16 to 20 hours) post inoculation and verified by CFU / mL counts.

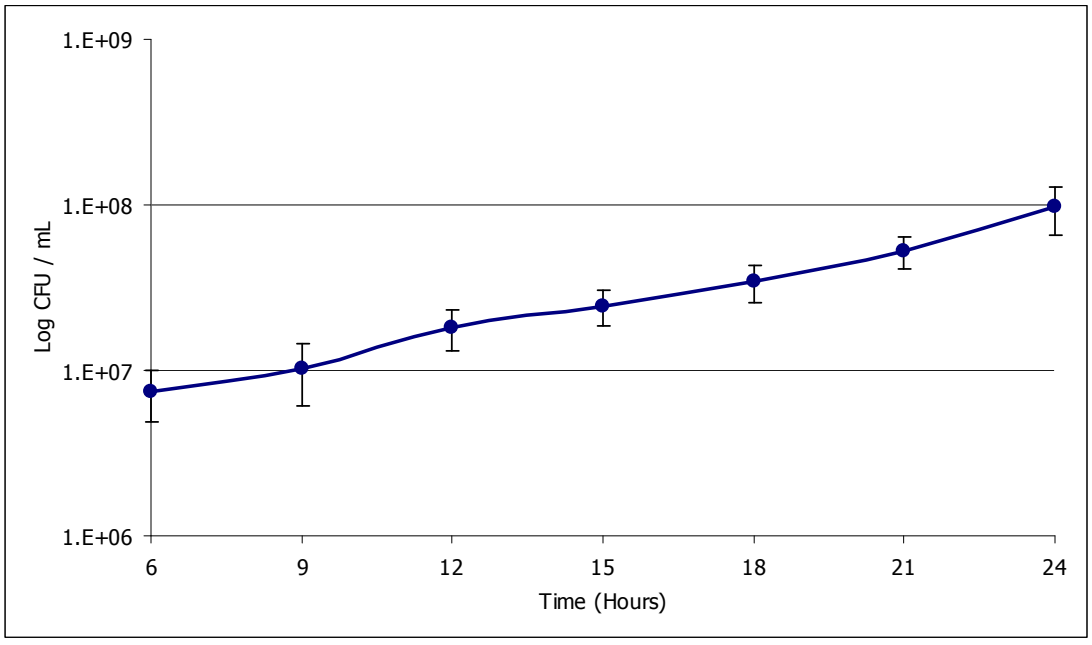


Figure 14: Growth parameters of *M. smegmatis* mc^2 155 grown in MADC-Tw. Log CFU / mL versus Time (Hours). Data points are averages of two biological replicates assessed in duplicate. Log phase growth, as defined by CFU / mL values of $Log_{10}7$ to $Log_{10}8$, was observed between 12 and 20 hours post inoculation. Thereafter, stationary phase growth occurred.

3.1.2 Confirmation of the integrity of the Morf2 region within the pGRAK reporter

Prior to use in Fluctuation Tests, the Gm resistance cassette and Morf2 region within the pGRAK reporter were assessed (refer to Section 2.10.2) to verify that the cloning strategy had resulted in formation of the correct plasmid. Restriction digestion of the Gm resistance cassette revealed two previously non-annotated restriction endonuclease sites, namely *Eco*RV and *Bgl*II (Figure 15A). Sequence analysis of the Morf2 region revealed the construct to be correct (Figure 15B).

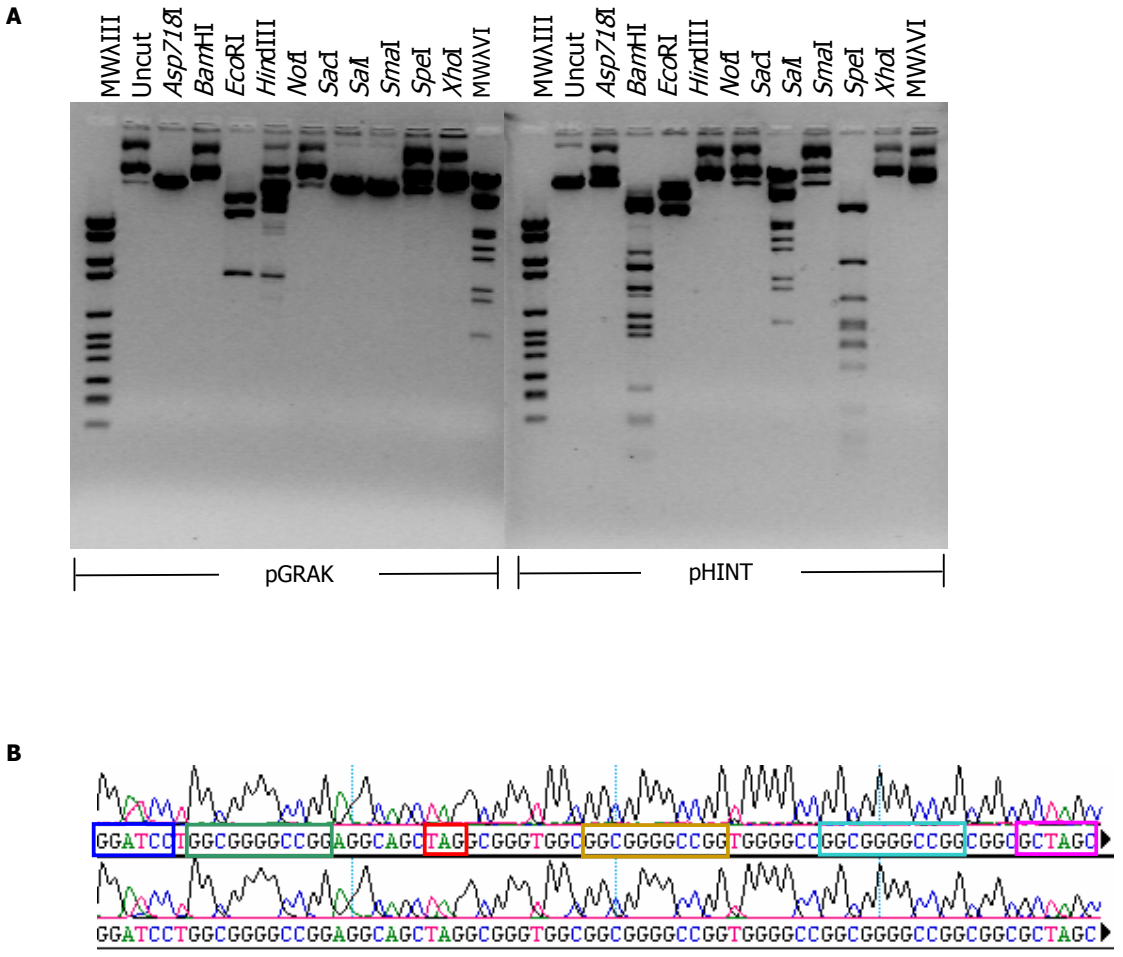


Figure 15: Confirmation of the integrity of the Morf2 region. (A) Restriction digestion of the Gm^R cassette in the pGRAK reporter. Thirty different restriction endonucleases were used to screen the Gm^R cassette from pGRAK. pHINT was digested with the same endonucleases as a control. All restriction digests were separated on agarose gels and selected examples are illustrated. From these and other restriction digests (data not shown), it was deduced that the Gm^R cassette contains *Eco*RV, *Bgl*II, *Sph*I, *Bst*BI and *Not*I restriction sites, the first two sites being previously non-annotated. MW λIII: Molecular Weight Marker λIII, MW λVI: Molecular Weight Marker λVI, Uncut: uncut control plasmids pGRAK or pHINT. (B) Chromatogram data of the Morf2 region within the pGRAK reporter. The top panel indicates a control sample of the Morf2 region while the lower panel indicates a sample sequence selected for comparison. The 5′ end of the Morf2 region is defined by a *Bam*HI site (G↓GATCC highlighted by blue block) and the 3′ end is defined by a *Nhe*I site (G↓CTAGC highlighted by pink block). The stop codon (TAG highlighted by red block) is present in both sequences as are the direct repeat sequences R1 (highlighted by green block), R2 (highlighted by yellow block) and R3 (highlighted by blue block) respectively ($^{5'}$GGCCGGGGCCGG$^{3'}$).

3.1.3 Confirmation of site-specificity of pGRAK reporter

Having determined the basic growth characteristics of the host bacterium and confirmed that the reporter was correct, it was then necessary to transform *M. smegmatis* mc^2 155 with pGRAK and determine a range of Gm concentrations over which the plasmid would be maintained, thereby allowing for survival and growth of the transformed strain in the presence of antibiotic. To confirm the site-specificity of integration of the reporter, the *attL* and *attR* regions of all pGRAK based strains used in this study were amplified by PCR (refer to Section 2.12.3). The results of selected *attL* (PCR product size = 320 bp) and *attR* (PCR product size = 282 bp) PCR products are shown in Figure 16.

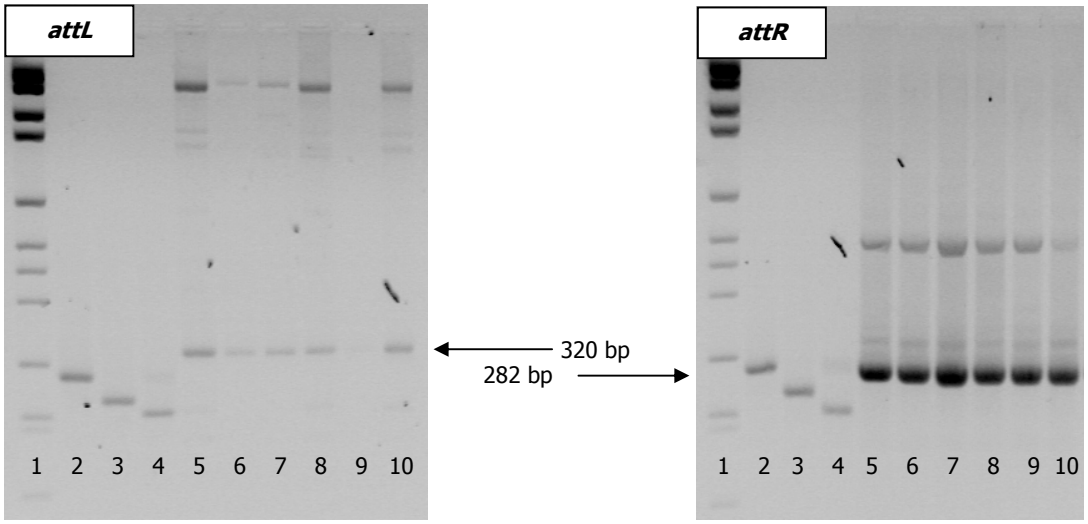


Figure 16: PCR products of *attL* and *attR* regions of selected pGRAK based strains used in this study. Only correct integration events would have yielded the expected PCR product. The left hand gel indicates *attL* PCR products (320 bp) and the right hand gel indicates *attR* PCR products (282 bp). Lanes 1: Molecular Weight Marker λVI, Lanes 2: control PCR product of 282 bp in length, Lanes 3: control PCR product of 263 bp in length and Lanes 4: control PCR product of 245 bp in length. Lanes 5 to 10: *attL* and *attR* PCR products from separate colonies that were picked from Gm^5 plates post electroporation.

3.1.4 Gm concentration range

Following confirmation of the correct integration of the pGRAK reporter, a range of Gm concentrations over which plasmid growth could be maintained was determined. This was important because integrative vectors have been found to be prone to loss in the absence of drug selection (Lee *et al.*, 1991; Pena *et al.*, 1997). The Gm cassette within the pGRAK reporter allowed for selection of clones that maintained the vector in the presence of antibiotic. Wild type *M. smegmatis* mc^2 155, mc^2 155 (pGRAK) and mc^2 155 (pGINTOa) strains were grown separately both in liquid (MADC-Tw) and on solid (Middlebrook 7H10 agar plates) media that contained Gm at each of the following concentrations (µg / mL): 20, 10, 5, 2, 1, 0.5, 0.1, 0.05 and 0.01.

In MADC-Tw, all three strains grew at Gm concentrations at or below 1 µg / mL (Figure 17A). At an antibiotic concentration of 2 µg / mL in liquid medium, mc^2 155 did not exhibit any growth, mc^2 155 (pGINTOa) exhibited poor growth (defined as an increase in OD_{600} values but few countable CFUs when plated on agar) and mc^2 155 (pGRAK) grew well (defined as CFUs of $Log_{10}7$ when plated on agar). None of the cultures grew well in liquid medium that contained Gm at concentrations higher than 2 µg / mL.

On Middlebrook 7H10 agar plates, mc^2 155 did not grow when the antibiotic concentration was higher than 0.1 µg / mL while both mc^2 155 (pGRAK) and mc^2 155 (pGINTOa) grew well at this and lower Gm concentrations (Figure 17B). At Gm concentrations at or above 0.5 µg / mL, mc^2 155 (pGINTOa) grew poorly while mc^2 155 (pGRAK) grew well. Neither of these cultures grew well on solid agar that contained an antibiotic concentration higher than 10 µg / mL.

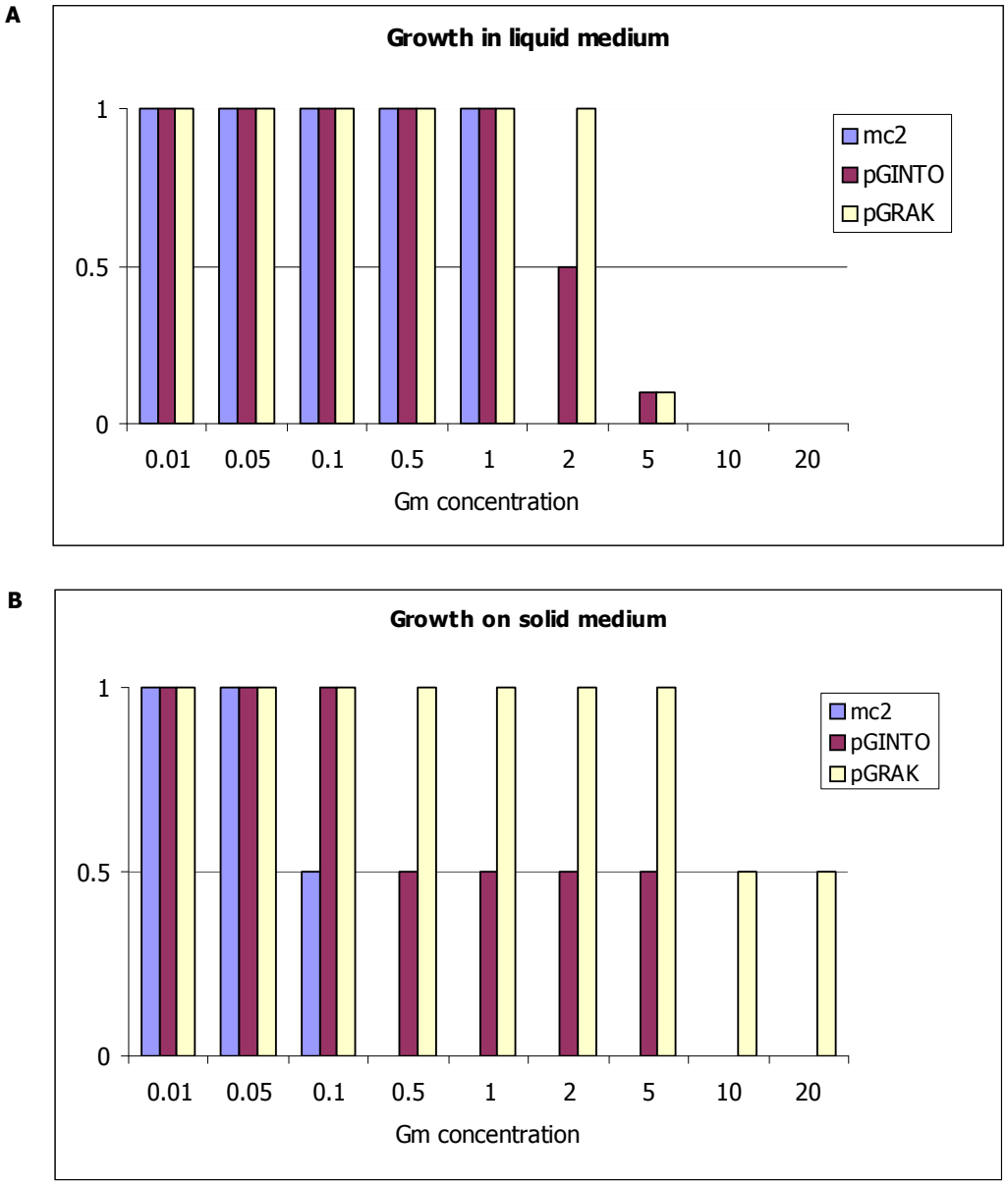


Figure 17: Growth of mycobacterial cultures with and without Gm based plasmids over a range of antibiotic concentrations. (A) Growth of cultures in liquid medium. Wild type *M. smegmatis* (indicated as mc2 in key) did not grow at Gm concentrations higher than 1 µg / mL. pGINTO based cultures grew poorly while pGRAK based cultures grew well at Gm concentrations of 2 µg / mL. Neither Gm based plasmid facilitated growth in liquid medium at Gm concentrations higher than 2 µg / mL. (B) Growth of cultures on solid agar medium. Wild type *M. smegmatis* did not grow at Gm concentrations higher than 0.1 µg / mL. pGINTO based cultures grew poorly at any Gm concentration higher than 0.5 µg / mL. pGRAK based cultures grew well at Gm concentrations up to and including 5 µg / mL. At higher concentrations, pGRAK did not facilitate growth on solid agar. Poor growth was defined as an increase in OD_{600} values but few countable CFUs when plated on agar (indicated by arbitrary value of 0.5 graphs in A and B). Good growth was defined as CFUs of $Log_{10}7$ when plated on agar (indicated by arbitrary value of 1 on graphs in A and B).

Based on the results shown in Figure 17, it was decided to use Gm at a concentration of 2 µg / mL in liquid medium and 5 µg / mL in solid agar for all Fluctuation Tests carried out in this study, as these concentrations of Gm would ensure maintenance of the integrating pGRAK vector. Further, all Fluctuation Tests in this study used Gm to select for cells that contained the pGRAK reporter in its native form (i.e. no mutational event had occurred to abolish the TAG codon within the Morf2 region) while Km was used to select for mutants that had acquired Km^R as a result of a mutational event within the Morf2 region of the reporter.

3.1.5 Antibiotic cross-resistance

Both Gm and Km fall into the class of aminoglycoside antibiotics that act on the bacterial ribosome to inhibit translation of proteins. Resistance to these antibiotics can occur due to mutations in the ribosomal target. Gm acts on the L6 portion of the 50S subunit while Km acts on the 30S subunit and cross-resistance between the two antibiotics, whereby bacteria carrying a resistance mutation to Km are also resistant to Gm, has been observed (Traub and Fukushima, 1979; Adwan *et al.*, 1998). This phenomenon was evident during initial growth experiments involving the pGRAK reporter construct in *M. smegmatis* mc^2 155. Cultures were grown in liquid media containing Gm at a concentration of 2 µg / mL and when aliquots were plated onto Middlebrook 7H10 plates containing Km at a concentration of 10 µg / mL (Km^{10}), 'background' growth was observed, as evidenced by the appearance of small colonies of varying sizes. Genotypic analysis by PCR of the background CFUs revealed that the clones had not acquired a Km^R mutation.

To test whether the Gm^R mutation in these cultures was resulting in cross-resistance to Km^{10}, the Km concentration was increased to Km^{25} on solid media. At the higher Km concentration, the 'background' growth was no longer evident and only discrete Km^R colonies were observed (Figure 18). In this case, genotypic analysis of randomly selected Km^R clones by PCR analysis revealed various mutations within the Morf2 region of the pGRAK reporter to by-pass the TAG stop codon and allow for expression of a functional Aph fusion protein (refer to Section 3.4). The mutational events revealed by the genotypic analysis were consistent with the formation of discrete Km^R clones on the Km^{25} agar plates.

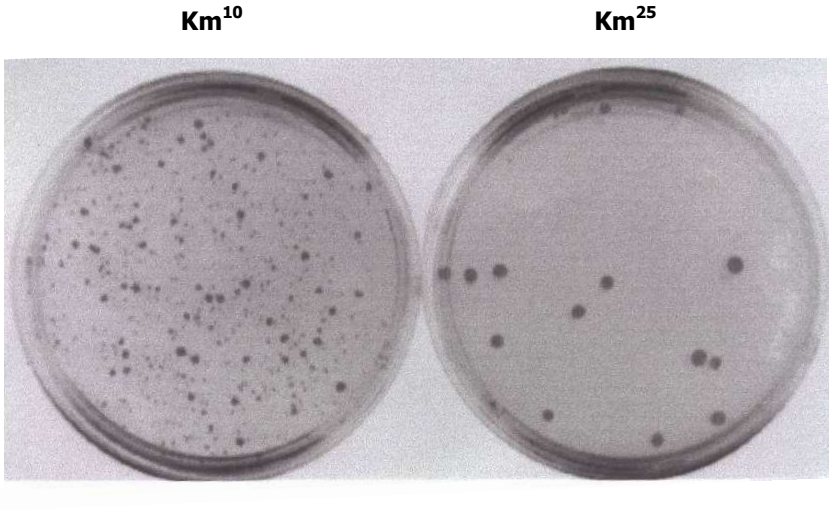


Figure 18: Cross-resistance between two aminoglycoside antibiotics, Gm and Km. Cultures grown in liquid media containing Gm at a concentration of 2 µg / mL developed cross resistance to low Km concentrations (indicated by 'background' growth on agar plate containing Km^{10} on left hand side of figure). When the same Gm^R cultures were plated on agar plates containing a higher Km concentration, the 'background' growth no longer appeared and only discrete Km^R CFUs grew (right hand side of figure). Sequence analysis of these Km^R CFUs revealed various mutations within the Morf2 region of the pGRAK reporter to by-pass the TAG stop codon and allow for expression of a functional Aph fusion protein.

3.1.6 Optimisation of methodology to minimise variability in N_t values

The single most important factor used in the calculation of a mutation rate (µ) by fluctuation analysis is the final population number (N_t) within each tube. Therefore, it is vital that the N_t values fluctuate as little as possible within a set of parallel cultures. The two factors that influence this parameter most significantly are: (1) the growth media used; and (2) the growth phase at which the N_t values are assessed. Linked to the second factor is the complication caused by clumping of bacteria within the culture, which is a problem of particular relevance to mycobacteria.

(1) *Growth media*. As the N_t value of a culture is a reflection of its growth, this value is dependent on the content of the growth medium. Three different liquid media, namely LB, 2TY and MADC-Tw were used to evaluate how the respective media constituents affected cell growth and consequently N_t values in a Fluctuation Test. LB is a general purpose medium that supports the growth of many microorganisms, 2TY is an enriched medium fortified with yeast extract, whereas MADC-Tw is a medium specifically formulated for the growth of mycobacteria (Jacobs Jr *et al.*, 1991). The N_t values fluctuated significantly when parallel cultures were grown in LB or 2TY but were reproducible when cultures were grown in MADC-Tw (Figure 19). These observations suggested that MADC-Tw was best suited for use in fluctuation analysis, and all Fluctuation Tests were therefore performed using this culture medium.

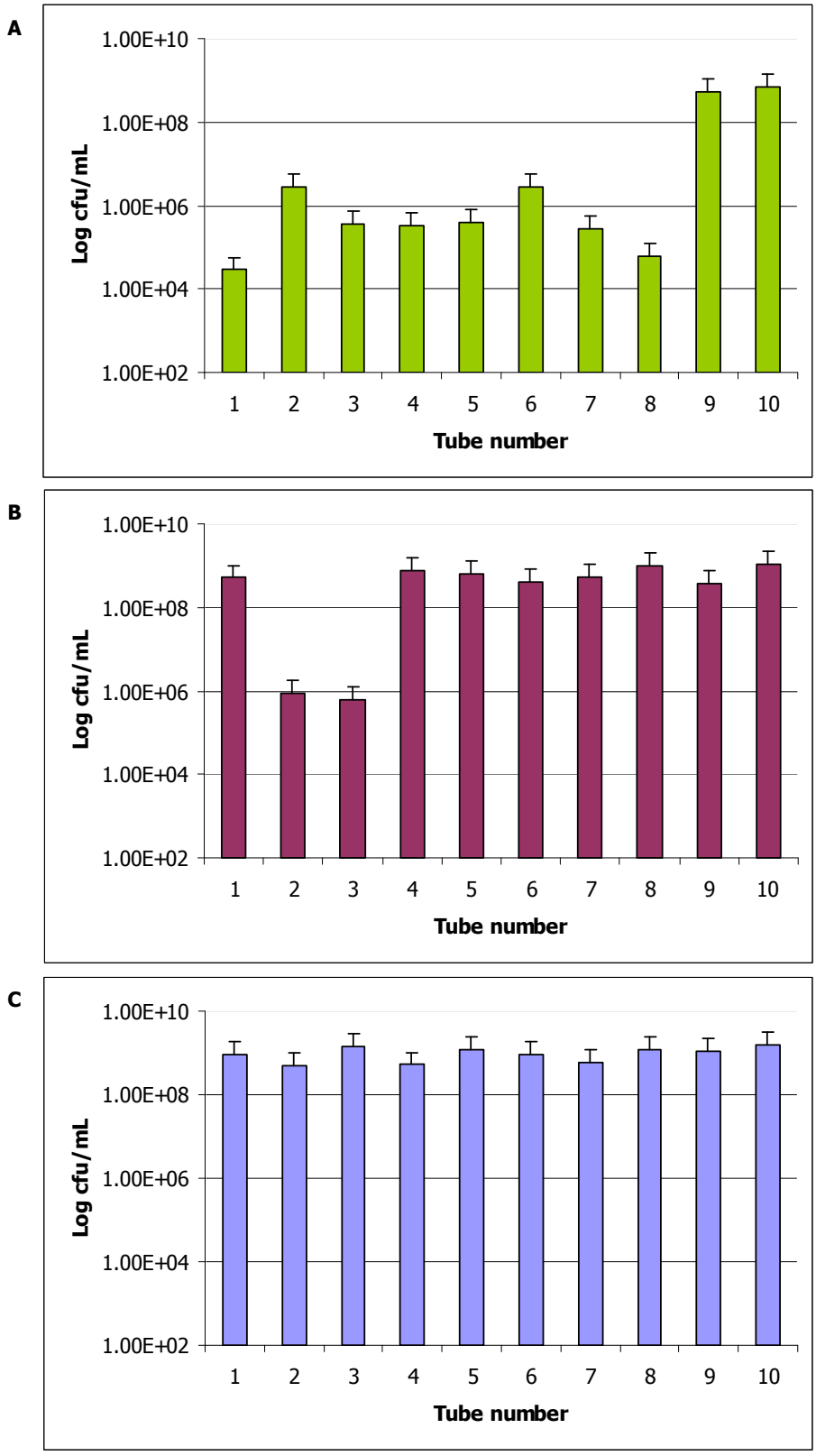


Figure 19: Fluctuation in N_t values was dependant on the growth medium used. Ten parallel tubes that contained the same culture were grown in three different media that each contained Gm at a concentration of 2 µg / mL to maintain the plasmid within the bacteria. (A) N_t values of mc^2155 (pGRAK) cultures grown in general LB media. (B) N_t values of mc^2155 (pGRAK) cultures grown in enriched 2TY media. (C) N_t values of mc^2155 (pGRAK) cultures grown in mycobacterium-specific MADC-Tw.

(2) *Growth phase*. Mycobacteria have thick hydrophobic cell walls and as a result, these organisms tend to clump together when cultures are grown beyond log phase. Clumping of cultures is a serious confounding factor as it results in a significant variation of N_t values between parallel cultures and it affects the ability to score mutants because cell pellets are precluded from being completely resuspended prior to plating. Consequently, the calculation of μ is unreliable. It was found that the extent of clumping could be reduced by placing the culture tubes in a beaker of water and securing the tubes with paper towels, thereby creating a 'water jacket' in which cultures were incubated (refer to Appendix D). Cultures that had clumped were resuspended by aspiration with a 1 mL syringe needle prior to assessing the N_t values. Furthermore, when assessing N_t values, all serial dilutions were made in 0.5% Tween 80. In conjunction, these three techniques proved effective in minimising the effects of clumping of log or saturation phase cultures on N_t value determination, as assessed by the high degree of concordance in N_t values in parallel culture tubes (Figure 20).

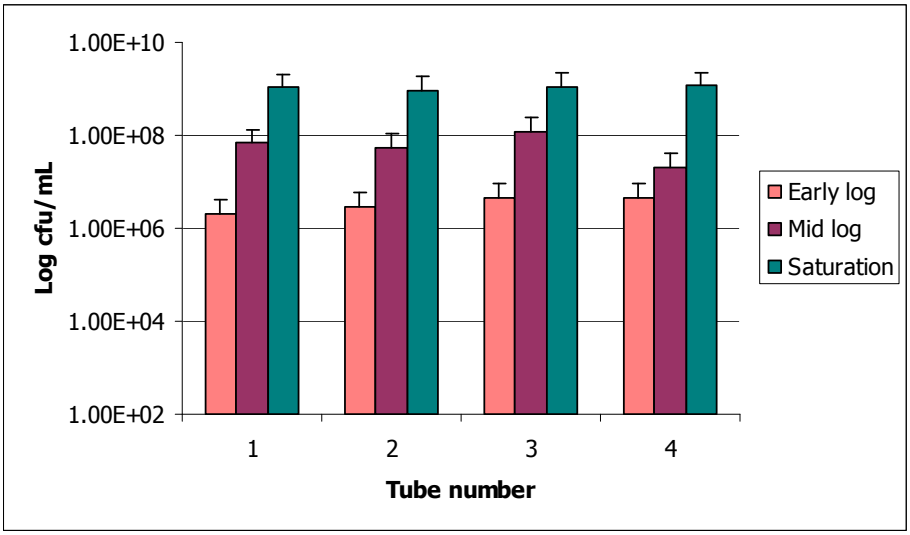


Figure 20: N_t values assessed at different time points during growth of mc^2 155 (pGRAK) in MADC-Tw. Early log cultures were assessed after 24 hours, mid log cultures after 48 hours and saturation cultures after 114 hours of growth. In each case, N_t values were assessed from the same tube that was then cultured further. The N_t values (Log CFU / mL) increased over time from $Log_{10}6$ CFU / mL at early log phase to $Log_{10}9$ CFU / mL at saturation phase, however at each time point, the N_t values were comparable between tubes.

3.1.7 Dependence of the μ value on the growth phase of the culture

Having established a methodology to ensure reproducibility of N_t values between parallel cultures scored at various stages of growth, it was necessary to determine the dependence, if any, of the μ value on the growth phase of the culture. In preliminary experiments, numbers of Km^R mutant colonies and N_t values were assessed at early log, mid log and saturation phase time points using the pGRAK reporter. However, at early log, no mutant colonies grew on Km selective plates, presumably as a result of the low N_t values of the cultures. Mutation rates could be assessed at later stages of growth: at mid log, μ was found to be $(1.56 \pm 0.02) \times 10^{-8}$, and at saturation phase, the μ value was $(2.54 \pm 0.70) \times 10^{-8}$. These values were not significantly different from each other ($P = 0.65$) suggesting that the mutation rate assessed using the pGRAK reporter did not change significantly as a function of the growth phase of the culture. Nonetheless, all subsequent μ values were routinely assessed at both the mid-log and saturation phases of growth.

3.2 Mutation rate assessment: Fluctuation Test data

3.2.1 Overall *m* and N_t values at log and saturation phase

The pGRAK reporter provided a tool for the assessment of mutant numbers that were subsequently used to generate a mutation rate, μ (refer to Section 2.11.3). As previously mentioned, the calculation of μ involved two key factors namely the *m* value and the average N_t value for a set of parallel tubes in a Fluctuation Test. The data in Table 5 summarise selected examples of these two factors. The initial *m* values as calculated by either the P_o Method or the Lea-Coulsen Method of the Median were not significantly different from the final *m* value as calculated by the MSS Maximum Likelihood Method. However, it was important to use the more accurate latter *m* value when calculating μ as there was a higher degree of confidence associated with the iterative nature of the MSS Maximum Likelihood calculation.

The average N_t values summarised in Table 5 differed according to the time point at which the N_t values were assessed. At log phase, N_t values were approximately $Log_{10}8$ CFU / mL and at saturation phase they were approximately $Log_{10}9$ CFU / mL. Consistent with these data are the *m* values at the same time points. Generally, log phase *m* values were low (*e.g.* 3.9) while the saturation phase counterparts were higher (*e.g.* 6.4). In cases where a jackpot had occurred, the median *m* values were as high as 100. A jackpot event is defined as a very early mutational event that allows for clonal expansion and a very high representation of that particular type of mutation within the culture. There were no instances where *m* value was 1, i.e. where one single mutational event led to the observed distribution of mutants. From these data, it was clear that no single host strain should have given rise to mutants with the same genotype, especially in the case of a jackpot event.

The calculation of μ relies heavily on the accuracy of the N_t value. However, higher N_t values (such as those calculated at saturation phase) are offset by higher *m* values at the same time point. Lower N_t and *m* values, such as those assessed at log phase, lead to the same result. Therefore, the time point at which N_t values were assessed did not affect the calculation of μ. Using the N_t and *m* values from Table 5, μ values were calculated for various host strains. The μ data summarised Table 5 are average values for two biological replicates of each strain.

Table 5: *m* values, average N_t values and μ values assessed at different time points in various mycobacterial strains

Background strain	**Time point assessed**	**Initial *m* value**	**MSS calculated *m* value**	**Average N_t value (CFU / mL)**	**Mutation rate, μ ($\mu = [Ln(2).m]/N_t$)**
mc^2 155 (pGRAK)	Log	2.7	5.3	$(5.94 \pm 0.57) \times 10^{+8}$	$(2.45 \pm 0.71) \times 10^{-8}$
mc^2 155 ΔrecA (pGRAK)	Saturation	76	76	$(2.96 \pm 0.20) \times 10^{+9}$	$(2.30 \pm 1.40) \times 10^{-8}$
mc^2 155 (pGRAK / pOLYGaa)	Log	6.5	6.4	$(5.31 \pm 0.60) \times 10^{+8}$	$(3.30 \pm 0.27) \times 10^{-8}$
mc^2 155 (pGRAK / pOLYGaadinP)	Saturation (jackpot)	100	100	$(1.58 \pm 0.17) \times 10^{+9}$	$(3.32 \pm 1.12) \times 10^{-8}$
mc^2 155 (pGRAK / pOLYGaadinX)	Log	3.9	4.1	$(3.94 \pm 0.43) \times 10^{+8}$	$(1.78 \pm 1.12) \times 10^{-8}$
mc^2 155 ΔY (pGRAK)	Log	16.8	17.8	$(2.48 \pm 0.23) \times 10^{+8}$	$(2.64 \pm 0.67) \times 10^{-8}$
mc^2 155 ΔL (pGRAK)	Saturation	50	50	$(1.31 \pm 0.11) \times 10^{+9}$	$(2.19 \pm 0.45) \times 10^{-8}$
mc^2 155 (Rif resistance)	Saturation	1.3	1.0	$(3.45 \pm 0.32) \times 10^{+9}$	$(4.02 \pm 2.76) \times 10^{-10}$
mc^2 155 *his*G380E	Saturation (jackpot)	96	88	$(4.43 \pm 0.15) \times 10^{+8}$	$(4.86 \pm 4.32) \times 10^{-7}$

3.2.2 μ values in pGRAK reporter-based strains

The rate of mutation to Km^R using the pGRAK reporter was initially determined in the control strain, namely *M. smegmatis* mc^2 155 carrying the pGRAK reporter, and was found to be $(2.45 \pm 0.71) \times 10^{-8}$. This value provided a benchmark rate against which the rate of mutation in other genetic backgrounds could be compared. Application of the pGRAK reporter in the *recA*-deficient mutant, $mc^2$155 ΔrecA gave a mutation rate of $(2.30 \pm 1.40) \times 10^{-8}$. This μ value was not significantly different from that observed for mc^2 155 (pGRAK) (P = 0.82).

Mutation rates were then determined in *M. smegmatis* mc^2 155 strains that over-expressed Y-family EP polymerases to test the contribution of these alternate DNA polymerases to the generation of mutations at PE-PGRS loci. These cultures were grown in the presence of Gm and Hyg antibiotics to ensure that both the integrating pGRAK and the replicative pOLYGaa-based vectors were maintained within the cultures. The rate of mutation in the control strain, which corresponds to mc^2 155 (pGRAK) carrying the empty expression vector, pOLYGaa, was $(3.30 \pm 0.27) \times 10^{-8}$. This value was not significantly different from that observed for mc^2 155 (pGRAK) itself (P = 0.23), confirming that the presence of the replicating expression vector had no effect on the mutation rate. Mutation rates were then assessed in the mc^2 155 (pGRAK / pOLYGaadinP) and mc^2 155 (pGRAK / pOLYGaadinX) strains, which over-express *M. tuberculosis dinP* and *dinX* respectively. The mutation rates were $(3.32 \pm 1.12) \times 10^{-8}$ and $(1.78 \pm 1.12) \times 10^{-8}$ which are not significantly different from each other (P = 0.20) or from the rates in the control strains, mc^2 155 (pGRAK) and mc^2 155 (pGRAK / pOLYGaa); P= 0.20 and 0.48 respectively. The mutation rate was then assessed in a strain lacking all three recognisable Y-family polymerase encoding genes in *M. smegmatis*, namely *dinP*, *dinX* and *dinP3*. This strain was designated mc^2 155 ΔY (pGRAK) and the observed mutation rate was $(2.64 \pm 0.67) \times 10^{-8}$. This value was not significantly different from that observed for the control strains (P = 0.21 and 0.94 respectively) or from the rates in the strains that over-express DinP and DinX (P = 1.0).

Since the mutation rate to Km^R measured using the pGRAK reporter system was insensitive to alterations in the levels of expression of DinP and DinX, the effect of the C-family EP polymerase, DnaE2, was then investigated. DnaE2 was recently shown to be significantly damage-inducible and to perform EP translesion synthesis under DNA damage conditions such as those induced by UV light (Boshoff *et al.*, 2003). To assess the specific contribution of DnaE2 to the mutational pathway reported by pGRAK, the mutation rate was assessed in a strain that is deficient for all the aforementioned Y-family polymerase genes as well as *dnaE2* (mc^2 155 ΔL). The mutation rate to Km^R in mc^2 155 ΔL (pGRAK) was $(2.19 \pm 0.45) \times 10^{-8}$, a value that was not significantly different from that observed for mc^2 155 ΔY (pGRAK) or the control strains (P = 0.08 and 0.52 respectively).

3.2.3 Rate of mutation to Rif^R

In order to have a benchmark mutation rate against which the rates determined using the pGRAK reporter could be compared, the rate of mutation rate to Rif^R was determined in wild type *M. smegmatis* mc^2 155. Rif^R arises primarily via base substitution mutations in the β subunit of RNA polymerase (Billington *et al.*, 1999), in a region known as the 'Rif-resistance determining region', which is located between codons 507 and 533 of the *rpoB* gene of *M. tuberculosis*. More than 35 resistance alleles have been identified in this region of the *rpoB* gene of *M. tuberculosis*. Rif^R alleles in clinical isolates of *M. tuberculosis* show varying frequencies with His_{526} → Tyr and Ser_{531} → Leu together accounting for 64% of all mutations (Billington *et al.*, 1999; Morlock *et al.*, 2000; Mariam *et al.*, 2004). The rate of mutation of *M. smegmatis* mc^2 155 to Rif^R was calculated to be (4.02 ± 2.76) x 10^{-10}, which is markedly lower than the rates of mutation of the various *M. smegmatis* strains to Km^R, as assessed using the pGRAK reporter (P = 0.02).

3.2.4 Rate of A→G base substitution mutagenesis using the *his*G380E reporter

The rate of a specific base substitution mutation was then assessed using the *his*G380E reporter in an assay designed to monitor the rate of reversion of the hisG380E auxotrophic mutant of mc ^{2}155 to His prototrophy (refer to Section 2.12.2). The His reversion assay differs from the pGRAK-based assay that the former applies a non-lethal selection and uses a nutritional requirement, rather than antibiotic resistance, as a selection method. The rate of reversion of the *his*G380E mutation in a wild type background was found to be (4.86 ± 4.32) x 10^{-7}. This rate of reversion by base substitution mutagenesis was not markedly higher than the rate of mutation to Km^R assessed for any of the pGRAK based strains (P = 0.18). Sequence analysis of randomly selected clones from un-supplemented plates confirmed that revertants contained the expected A → G transition mutation (E380G) which restored the Gly residue at position 380 in HisD (Figure 21).

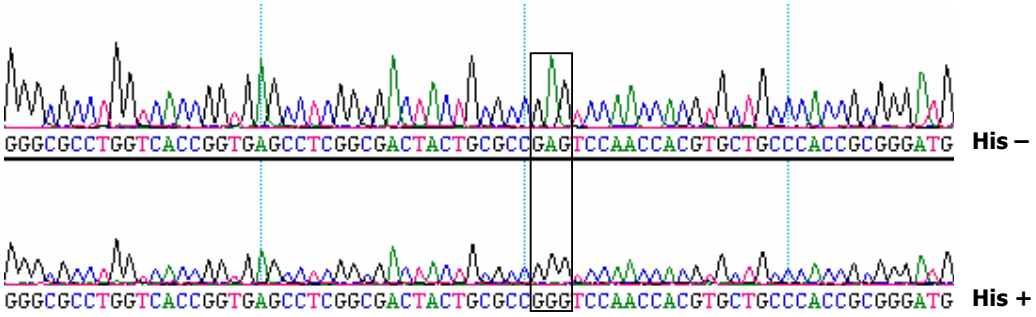


Figure 21: Chromatogram data showing the wild type (GAG) Histidine auxotroph (His-) sequence (upper panel) and the revertant (GGG) Histidine prototroph (His+) sequence (lower panel) in selected mc^2 155 *his*G380E clones. The reversion mutation from GAG to GGG is highlighted by the black box. The base substitution mutation results in a codon change from glutamate (E) to glycine (G).

3.2.5 Statistical analysis of results

Statistical analyses of the results in this study were difficult to determine due to the complications associated with analysis of non-parametric data. It was difficult to determine Confidence Limits (CL) for *m* and μ as the distributions of these values were all shown to be Poisson. This was because not all the cells within a culture would respond in the same way to an applied stress (such as antibiotic presence in the medium). The CL could have been calculated according to Ln (*m*) ± (1.96 . σ) where $\sigma = (1.225m^{-0.315})/\sqrt{C}$, but the results of these equations were based on the weak assumptions associated with non-parametric data sets. Furthermore, μ values from different Fluctuation Tests were compared using only a Student's T-test. No other methods have yet been devised to compare different μ values and while the T-Test is accurate, it relies on data sets that contain numerous samples. This was not the case in this study where two μ values were compared to each other. The lack of robust statistical analyses has been noted as a limitation to Fluctuation Tests (Kendal and Frost, 1988), while the mathematical complexity required to amalgamate the data and arrive at a μ value can be daunting. Despite these factors, Fluctuation Tests remain a reliable, accurate and reproducible means of assessing μ that indicate the underlying mutagenic mechanisms in a bacterial population.

3.3 Mutation frequencies

Due to the complex technical nature of mutation rate analysis, mutation frequencies have been favoured when determining the fraction of mutant bacteria within a set of replicate cultures. As previously described, every cell within a culture has an equal probability of undergoing a mutation at any point within its lifetime. This is a stochastic process and thus if a mutation frequency is used, early mutational events are lost when they are averaged out, giving a distorted reflection of the distribution of mutations within the few replicate cultures (Martinez and Baquero, 2000).

To validate the use of mutation rates versus mutation frequencies, both values were calculated for a set of culture tubes containing one of the control strains mc^2 155 (pGRAK). Mutants were assessed as those that had developed Km^R and mutation frequency (f) was calculated as the ratio of the average number of mutants to the total number of cells. The average N_t value used in all calculations was 9.56 x 10^{+8}. The results are indicated in Table 6 and include the r values (number of Km^R mutants per tube), the frequency of mutants per tube as well as the overall frequency of mutants across all tubes in the set. The comparative μ value is also given.

Table 6: Comparison of mutation frequencies and mutation rates for Km^R in mc^2 155 (pGRAK)

Tube #	**1**	**2**	**3**	**4**	**5**	**6**	**7**
r	296	76	221	3364	487	532	1820
f per tube	3.1×10^{-7}	8.0×10^{-8}	2.3×10^{-7}	3.5×10^{-6}	5.1×10^{-7}	5.6×10^{-7}	1.9×10^{-6}
f overall	**4.1×10^{-6}**						
μ overall	**2.5×10^{-8}**						

Based on the overall f and μ values in Table 6, there was a significant difference between mutation rate and mutation frequency values for the same data set (P = 0.05). A recent study of mutation rates and frequencies in the *M. tuberculosis* W-Beijing genotype reported similar discrepancies between f and μ for the same data set (Werngren and Hoffner, 2003). These data serve to emphasize that the variation of r values in a set of tubes is lost when a frequency calculation is used. The variation is a reflection of the underlying stochastic process of random mutation within those tubes. A μ calculation takes this variation into account and thus it is important to use μ and not f when analysing mutant distributions.

3.4 Spectrum of mutations in pGRAK based strains

3.4.1 From Fluctuation Test to mutant genotyping

To clarify the procedures followed, Figure 22 illustrates the methods employed, from μ generation via a Fluctuation Test to sequence analysis, in order to assess the spectrum of Km^R genotypes generated using the pGRAK reporter.

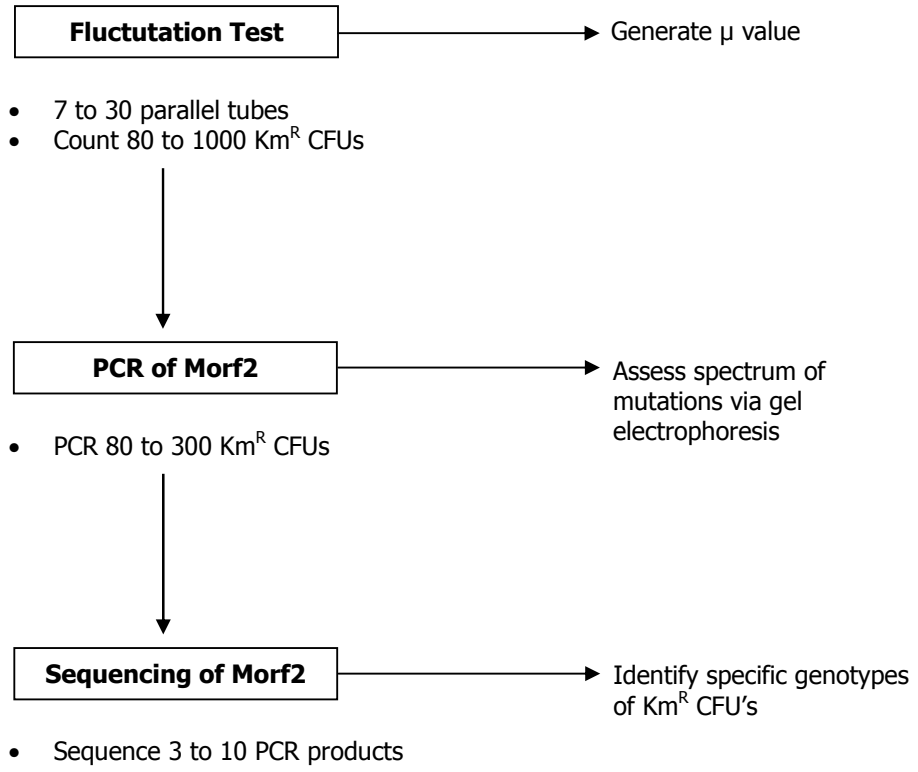


Figure 22: Flow diagram outlining the experimental procedures followed to generate μ values and to assess the spectrum of mutations per mycobacterial host background. Fluctuation Tests generated a μ value and Km^R CFU's. A random selection of these mutants were subsequently subjected to PCR to determine the type of mutation that had occurred in the Morf2 region of the pGRAK reporter that led to Km^R. Between 80 and 300 Km^R CFUs per Fluctuation Test were assessed by PCR, depending on the number of jackpots that had occurred. Once PCR was complete, a select group of PCR products were sequenced to verify the specific genotypic mutation that had occurred.

3.4.2 Multiple genotypes recovered from a single tube

The number of Km^R mutants observed on a single agar plate within a Fluctuation Test ranged from 8 to > 1000 CFUs. Clearly, only a sample of the mutants could be genotypically assessed by PCR. Between 80 and 300 Km^R mutants per Fluctuation Test were analysed, depending on the number of jackpot events that had occurred. Figure 23 illustrates a gel electrophoresis analysis of the PCR products generated from Km^R clones present within a single tube in a Fluctuation Test. Clearly, there was more than one type of mutation event that gave rise to the observed distribution of mutants in this tube. This observation was also evident in tubes that had undergone a jackpot event. To emphasise, even tubes that had undergone a very early mutation event to Km^R, showed a wide spectrum of Km^R mutations as opposed to a single genotype, as may have been expected from a culture in which an early mutation event had occurred.

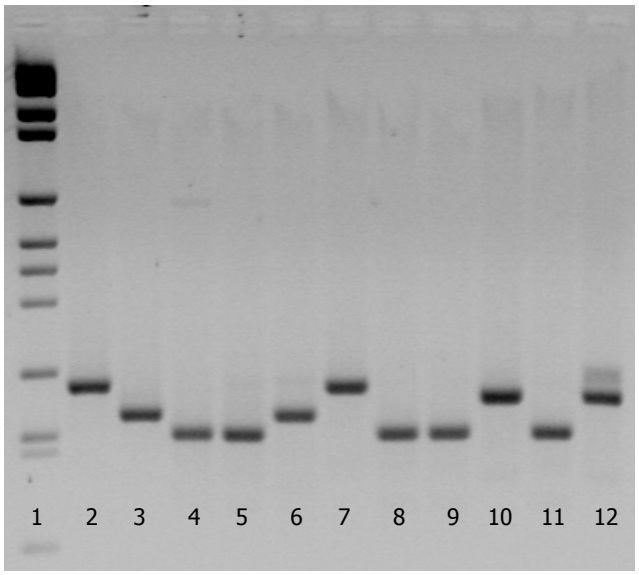


Figure 23: Spectrum of Km^R mutations observed in mc^2 155 (pGRAK) CFUs as shown by gel electrophoresis. Lane 1: Molecular Weight Marker λVI. Lanes 2 to 12: PCR products from separate Km^R colonies that were picked from a single tube. The PCR products vary in size, indicating a number of underlying mutational events.

Sequence analysis of the Km^R mutants revealed a wide spectrum of mutations in the Morf2 region. The structure of the Morf2 substrate was such that it would allow point mutations as well as slippage deletion and / or insertion-deletion events to be observed. Of the sequenced products obtained from these experiments, a total of 11 different genotypes resulting in a Km^R phenotype were identified (Table 7). The mutations fell into three major groups:

A. *Point mutations*. These included: 1) mutations within the stop codon that converted it from a stop codon to one that encoded an amino acid; and, 2) point mutations downstream of the TAG stop codon that formed a new start codon (GCG → GTG).
B. *Deletion mutations*. Five distinct mutations were identified in this class, which included pure deletion or partial deletion / partial duplication mutations. Pure deletion mutations were: 1) a deletion mutation that eliminated the sequence from the start of R1 to the start of R3, including the stop codon; and 2) a deletion mutation that eliminated the sequence between the start of R1 and the start of R2, including the stop codon, R3 and the intervening sequence. Partial deletion / partial duplication events included: 1) mutations resulting from rearrangements between R1 and R3 or R1 and R2 with an in-frame overlap of two complementary bases; and 2) duplications of the IV combined with partial deletion / rearrangement events between the repeats.
C. *Downstream mutations*. This class included two clones that had grown on Km^{25} agar plates but were shown to contain the wild type Km sensitive (Km^S) Morf2 sequence. Subsequent PCR and sequence analysis of the sequence downstream of the Morf2 region revealed new ATG start codons. One clone had a duplication of part of the '*aph* cassette that resulted in the formation of a new, in-frame start codon, while the second clone had a single point mutation within the first nineteen amino acids of the '*aph* cassette that had generated a new, in-frame start codon.

Table 7: Spectrum of mutational events in the Morf2 region of the pGRAK reporter as well as mutations within the 'aph cassette that resulted in Km^{R} [a]

Sequence	
Key[b]	*Bam*HI R1 Stop R3 IV R2 *Nhe*I
Km^{S} substrate	GGATCCTGGCGGGGCCGGAGGCAGCTAGGCGGGTGGCGGCGGGGCCGGTGGGGCCGGCGGGGCCGGCGGCGCTAGC
Point mutations	
TAG → TTG (Leucine) TAG → TGG (Tryptophan) TAG → CAG (Glutamine) New GTG start codon	GGATCCGGGCGGGGCCGGAGGCAGCTTGGCGGGGGGCGGCGGGGCCGGTGGGGCCGGCGGGGCCGGCGGCGCTAGC GGATCCTGGCGGGGCCGGAGGCAGCTGGGCGGGTGGCGGCGGGGCCGGTGGGGCCGGCGGGGCCGGCGGCGCTAGC GGATCCTGGCGGGGCCGGAGGCAGCCAGGCGGGTGGCGGCGGGGCCGGTGGGGCCGGCGGGGCCGGCGGCGCTAGC GGATCCTGGCGGGGCCGGAGGCAGCTAGGTGGGTGGCGGCGGGGCCGGTGGGGCCGGCGGGGCCGGCGGCGCTAGC
Deletion mutations	
R1 to R3 R1 to R2 R1 overlaps R3 R1 overlaps R2 R1 to R3 and back plus IV	GGATCCT----------GGCGGGGCCGGTGGGGCCGGCGGGGCCGGCGGCGCTAGC GGATCCT----------GGCGGGGCCGGCGGCGCTAGC GGATCCT----------GGCGGGGCCGGCGGGGCCGGTGGGGCCGGCGGGGCCGGCGGCGCTAGC GGATCCT----------GGCGGGGCCGGCGGGGCCGGCGGCGCTAGC GGATCCT----------GGCGGGGCCGGTGGGGCCGGCGGGGCCGGTGGGGCCGGCGGGGCCGGCGGCGCTAGC
Downstream mutations	
Duplication and new ATG	GCTAGCCATATTCAACGGGAAACGTCTTGCTCGAGGCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGT ATAAAGCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAAA
New ATG start codon	GCTAGCCATATTCAACGGGAAACGTCTTGCTCGAGGCCGCGATTAAATTCCAACATGGATGCTGATTTATATGGGT ATAAATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAA

a. The corresponding DNA sequencing chromatograms of the mutations described are given in Appendix E.

b. The *Bam*HI site denotes the 5'-end of the Mor2 region, whereas the *Nhe*I site denotes the 3'-end of this region and the start of the '*aph* cassette

The results in Table 8 show the relative percentages of the three major groups of mutations. Regardless of the genetic background of the host strain, deletion mutations were the most common type of mutation, with the first two types of deletion mutations shown in Table 7 predominating. Furthermore, the downstream mutations were the least common type of mutation and were not represented in all background strains. Any sampling errors that may have occurred in the selection of Km^R mutants for genotyping by PCR analysis may have biased the observed spectrum of mutations. However, without knowledge of the spectrum of Km^R mutations from every colony in an entire Fluctuation Test, a sampling error could not be evaluated. An interesting observation is that point and downstream mutations were evident within a set of parallel culture tubes, even though these types of mutations were in the minority.

The possibility that some of the mutant genotypes may have arisen via sequential mutational events was considered. For example, a point mutation to abolish the TAG codon may have undergone a second mutational event to form a R1 to R3 deletion mutant that finally mutated to form a R1 to R2 deletion mutant. However, two lines of evidence showed that the mutations were not sequential. Firstly, the mutations were shown to be stable over time: clones with specific Km^R mutations that had been sequenced and classified into one of the 3 categories, were serially passaged four times. Sequence analysis indicated that regardless of the type of mutation, the genotypic change was heritable and remained within the population through a number of successive generations (data not shown). Secondly, the most common types of mutations were assessed at log and saturation phase in a single host background, mc^2 155 (pGRAK). At log phase, the most common mutation was an R1 to R2 deletion, but at the later time point, the most common mutation was an R1 to R3 deletion. Clearly, the latter mutant could not have arisen from the former. Finally, it did not seem plausible that a mutant that had acquired Km^R would undergo a second or third mutation event that would not alter the overall phenotype and be represented sufficiently in a Km^R population to be picked from a plate when sampling.

Table 8: Relative percentages of three different groups of mutations per pGRAK-based host strain used in this study

	Mutation group			
Background strain	**Point**	**Deletion**	**Downstream**	**Most common mutation**
mc^2 155 (pGRAK)	9	90	1	Deletion
mc^2 155 ΔrecA (pGRAK)	8	92	0	Deletion
mc^2 155 (pGRAK / pOLYGaa)	10	88	2	Deletion
mc^2 155 (pGRAK / pOLYGaadinP)	8	90	2	Deletion
mc^2 155 (pGRAK / pOLYGaadinX)	20	80	0	Deletion
mc^2 155 ΔY (pGRAK)	9	91	0	Deletion
mc^2 155 ΔL (pGRAK)	13	87	0	Deletion

Chapter 4

Discussion and Conclusions

4.1 Optimisation of experimental parameters

4.1.1 Confirmation of the integrity of the integrating and replicative vectors

The pGRAK reporter system used a vector that integrated into the chromosome of the host bacterium. Within the experimentally determined Gm concentration range, the Gm^R cassette within the reporter allowed for selection of cells carrying the reporter while excluding cells that had either lost the construct or retained a wild type genotype (refer to Section 3.1.3). In mycobacterial host strains that had both the integrating pGRAK and the replicative pOLYGaa (or pOLYGaa-based) vectors, both Gm and Hyg antibiotics were included in the media to ensure that both constructs were maintained within the culture. The Gm^R cassette was carried on the former plasmid while Hyg^R was conferred by the latter. Thus, only cells expressing both antiobiotic resistance genes could have grown in the presence of these drugs. The pGRAK integrating vector was always electroporated into a host background that already contained the appropriate replicative vector. This measure was taken to minimise the amount of culture time available for KmR mutations to accumulate. This would have led to a high number of pre-existing mutants in the culture which would have affected the accuracy of the Fluctuation Test.

4.1.2 Optimisation of Fluctuation Test methodology for use in mycobacteria

The mycobacterial cell wall contains arabinogalactans and long chain (70 to 90 carbon atoms) fatty acids, mycolic acids, both unique to these bacteria (Draper, 1998). As a consequence, mycobacterial cell walls are lipid rich and when grown in liquid medium, cultures tend to clump. This problem created specific challenges which had to be experimentally addressed in order for fluctuation analysis to be used as a means of assessing mutation rates in these organisms. The growth medium employed had a significant effect on the variability of the N_t values across a set of parallel tubes (refer to Section 3.1.4). N_t values were assessed in LB, 2TY and MADC-Tw, all supplemented with the detergent Tween 80 (0.1%), which was added to minimise mycobacterial clumping. The fact that LB and 2TY cultures still clumped, even in the presence of Tween 80, suggests that these media are not well suited for the growth of mycobacteria. MADC-Tw has been specifically formulated for use in mycobacteria and the fact that N_t values were stable in this medium reinforced the need for correct culture conditions to be used in order to maximise the reproducibility of results, especially in the case of N_t values. Although the growth phase did not affect the calculation of μ (refer to Section 3.1.4), the preparation of serial dilutions to assess N_t values was technically more challenging in older cultures. The use of a 'water jacket' proved very effective at reducing clumping and was probably successful because it created a local environment of constant temperature in which the parallel tubes could incubate. Any tubes that did contain clumped cultures were aspirated with a 1 mL syringe needle and this too proved effective at breaking up residual clumps. There was a concern that cells that had adhered to the needle or walls of the tube may affect the serial dilutions and subsequent N_t values, but thorough vortexing of

all tubes and the addition of extra Tween 80 when preparing serial dilutions ensured that inter-tube variation was minimised.

4.1.3 Deviations from Luria and Delbrück assumptions

In 1943 when Luria and Delbrück designed the Fluctuation Test, they did so with only basic knowledge of growth kinetics in bacterial cultures. Since then, the understanding of the biology and growth cycles of bacteria has been widely expanded and their assumptions (refer to Section 1.12.3) need to be adapted accordingly. For instance, it was recommended that each tube within a parallel series be inoculated with a single colony (Rosche and Foster, 2000). However, this would have caused a variation in the initial population numbers (N_o) and potential variation in the subsequent final population numbers (N_t) per tube. Accordingly, the experimental protocol used in this study was adapted according to the method of Rosche and Foster (2000) to ensure that each tube within a parallel series was inoculated from a single culture with a known number of cells per mL (i.e. the two N_o cultures with either $Log_{10}2$ CFU / mL or $Log_{10}3$ CFU / mL).

As previously emphasized, the calculation of μ depends heavily upon the accuracy of the average N_t value from a single Fluctuation Test (refer to Section 2.11.1). As such, it is extremely important that the N_t values across a set of parallel tubes be as comparable as possible. One of Luria and Delbrück's assumptions is that in a culture, all cells are at the same point in their growth phase and are dividing at the same rate (Luria and Delbrück, 1943). In this study, N_t values were assessed at log and saturation phase. While both time points provided reproducible N_t values from parallel tubes, generally saturation phase cultures showed the least N_t variation between tubes. During this growth period specifically, the individual cells within the cultures may not have all been at the same point within their growth phase thereby negating Luria and Delbrück's assumption. Although theirs was a reasonable first approximation, it now has to be tempered by experimental evidence that growth rates differ periodically through the cell cycle (Kendal and Frost, 1988).

Luria and Delbrück only made provision for spontaneous mutations that occurred prior to plating on selective media. In the case of non-lethal selection, such as reversion to His+, post-selection mutation was a potential problem (McKenzie and Rosenberg, 2001; Rosenberg, 2001). To emphasize, any mutation events that occurred <u>after</u> plating on selective media may have caused the estimated number of mutational events (*m* value) to be over inflated thereby affecting the calculation of μ. For example, if the *hisD* allele in the *his*G380E reporter had been leaky, His- cells could have grown enough on the selective agar to produce a mutation following plating, or reversion mutations could have happened in the absence of growth (a so called 'adaptive mutation'). To avoid these post-selection mutation problems, cells were washed prior to plating and CFU's were counted as soon as they appeared.

Standardisation of experimental parameters is obviously important if comparisons between studies are to be made. One suggestion is that a single continuous culture be used for all Fluctuation Tests but this may introduce periodic selection (Rosche and Foster, 2000). This term simply explains the phenomenon whereby successive mutants within a single continuous culture each display better adaptation to their environment and are not comparable. Evidently, the Km^R mutations acquired by pGRAK reporter carrying strains were not sequential (refer to Section 3.4.2) suggesting that periodic selection would have been unlikely to occur in this system. Despite this observation, each Fluctuation Test was started from a single cell cultured from a fresh electroporation of the pGRAK reporter into the relevant host strain. This ensured that no siblings or adaptive mutants were present in the starting inoculum.

4.2 Mutation rate assessment

The μ value for each strain used in this study was calculated from the pool of mutation profiles, i.e. regardless of the type of Km^R mutation, each and every Km^R mutant scored per Fluctuation Test was included in the calculation of μ for that strain (refer to Table 5 in Section 3.2.1). Therefore, the calculated μ values pertain to the collective pool of Km^R mutants per strain. Due to the potential sampling bias with regard to the PCR of Km^R mutants, individual μ values for each type of mutation could not be accurately calculated. This limitation applies equally to other mutational targets (e.g. *rpoB*) in which distinct mutations, each of which occurs at a particular mutation rate, can give rise to the scored phenotype (in this case, resistance to Rif).

4.2.1 μ value in the *recA*-deficient pGRAK reporter-based strain

The rate of mutation to Km^R using the pGRAK reporter was initially determined in the control strain, namely mc^2 155 (pGRAK) and this value provided a benchmark against which the rates of mutation in other genetic backgrounds could be compared. Application of the pGRAK reporter in the *recA*-deficient mutant, mc^2155 ΔrecA gave a mutation rate that was not significantly different from the control strain. This suggested that pGRAK reporter-based mutations to Km^R were independent of RecA function. RecA is a protein involved in recombination events. One of its roles is to recognise regions of homology between 2 strands of DNA, bind to the relevant region and promote branch migration resulting in the exchange of homologous genetic material (Echols and Goodman, 1991; Kuzminov, 1999). Other evidence has shown that RecA is also involved in translesion synthesis of damaged DNA as well as stimulation of autocatalytic cleavage of LexA thus stimulating the SOS response (Duigou *et al.*, 2004). Importantly, RecA usually promotes recombination between long DNA sequences as short DNA templates are too small to support RecA filament formation (Tippin *et al.*, 2004).

In *E. coli*, RecA catalyses homologous recombination and strand exchange between DNA molecules (Echols and Goodman, 1991). However, recombination also occurs via RecA-independent mechanisms. DNA repeat sequences in *Bacillus subtilis* have been shown to recombine by slipped-strand mispairing, a mechanism in which the nascent DNA that has separated during DNA replication, reanneals with a second repeat thereby generating a deletion in the newly synthesized strand (Bruand *et al.*, 2001). This process occurs without the input of RecA suggesting that the stimulation of slipped-strand mispairing is also RecA-independent. The fact that the μ value in the *recA*-deficient strain was no different from the pGRAK-based control strain suggests that the mutation mechanisms induced by the pGRAK reporter may be occurring via RecA-independent slipped-strand mispairing. The close proximity in the Morf2 region of the repeat sequences to one another supports the idea of slipped-strand mispairing.

4.2.2 μ values in pGRAK reporter-based strains with altered levels of expression of Y-family and C-family polymerases

High fidelity replicative polymerases generate spontaneous mutations at a rate of 10^{-4} to 10^{-5} per base pair. Action of the 3′ to 5′ exonucleolytic proofreading function associated with the replicative polymerase increases the fidelity of replication to between 10^{-5} and 10^{-7} per base pair (Tippin *et al.*, 2004). The high fidelity PolIII enzyme is involved during chromosomal replication until the replication fork becomes stalled due to damaged DNA bases. Studies have shown that, in response to various forms of DNA damage, EP DNA polymerases, such as those from the Y-family, take over the role of PolIII, repair the damage and then relinquish the role of replication back to the replicative polymerase (Tippin *et al.*, 2004). Occasionally, the EP polymerases copy undamaged DNA thereby offering flexibility in the genetic code – a factor that enables pathogens to readily adapt to stressful environments. Typically, EP polymerases display mutation rates of between 10^{-1} and 10^{-3} per base pair of undamaged DNA (Tippin *et al.*, 2004).

To test the contribution of these alternate DNA polymerases to the generation of mutations at PE-PGRS loci, the rates of mutation to Km^R were determined using the pGRAK system in *M. smegmatis* mc^2 155 strains that over-expressed Y-family EP polymerases. The rate of mutation in the control strain carrying the empty expression vector, pOLYGaa was not significantly different from that observed for mc^2 155 (pGRAK) itself, confirming that the presence of the replicating expression vector had no effect on the mutation rate. Mutation rates assessed in *M. smegmatis* strains which over-express *M. tuberculosis dinP* and *dinX*, were not significantly different from each other or from the rates in the two aforementioned control strains. When the mutation rate was assessed in the strain that lacked all three recognisable Y-family polymerase encoding genes (namely *dinP*, *dinX* and *dinP3*), once again no significant difference occurred. The contribution of the C-family EP polymerase, DnaE2, to the mutational pathway reported by pGRAK, was then assessed by mutation rate analysis in a multiple deletion strain lacking all of the aforementioned Y-family polymerase genes as well as *dnaE2*.

Again, the rate of mutation to Km^R in this background showed no significant difference to either of the control strains, or the strains that either over- or under-express one or more of the Y-family polymerases.

In conjunction, these results argue against a role for any of the EP polymerases in the mutation pathways observed using the pGRAK reporter under the conditions tested, and reinforces the notion that these mutations are formed by slipped-strand mispairing between adjacent repeats during chromosomal replication mediated by the replicative PolIII holoenzyme. Notably, DnaE2 has been shown to function in induced base substitution mutagenesis only under conditions in which chromosomal DNA damage has been deliberately introduced by exposure to UV or treatment with mitomycin C (Boshoff *et al.*, 2003). However, mutation rates were not tested under these conditions in the present study as the DNA damaging treatments would have caused changes in the probability of the cells undergoing a mutation within their growth cycle, thereby negating the Luria / Delbrück assumption that a cell has a constant probability of undergoing a mutation within its lifetime (refer to Section 1.12.3). As a result, the potential contribution of DnaE2 to mutagenesis at PE-PGRS loci under DNA damaging conditions remains unknown.

4.2.3 Rates of base substitution mutagenesis in *M. smegmatis*

The rate of spontaneous base substitution mutagenesis in wild type *M. smegmatis* mc^2 155 was measured by determining the rate of mutation to Rif^R. The mutation rate determined in this study [$(4.02 \pm 2.76) \times 10^{-10}$] is comparable to the μ value of 2.25×10^{-10} per bacterium per generation for spontaneous mutagenesis of *M. tuberculosis* H37Rv to Rif^R (David, 1970). In contrast, the mutation rate to Rif^R in this study was lower than those previously reported for *M. tuberculosis* Beijing (1.1×10^{-8}) and non-Beijing (1.3×10^{-8}) strains (Werngren and Hoffner, 2003). However, it was noted that the rates of mutation to Rif^R observed in both the Beijing and non-Beijing strains were calculated from small samples and so there were too few mutants ($1.6 < r < 7.7$) for an exact calculation of μ.

The rate of base substitution mutagenesis in *M. smegmatis* was also assessed using a different reporter system, which monitored the rate of reversion of the *his*G380E allele by A→G transition mutagenesis (refer to Section 2.12.2). The μ value for the mc^2 155 *his*G380E strain was higher [$(4.86 \pm 4.32) \times 10^{-7}$] than the μ value observed for any of the pGRAK based strains, although statistical analysis revealed that the difference was not significant ($P = 0.18$). However, the average N_t value for parallel cultures of $mc^2$155 *his*G380E grown to saturation phase (6 days) was a log lower ($\sim Log_{10}8$) than those recorded for the same time period in the pGRAK based assays ($\sim Log_{10}9$). Comparably, the average N_t values from the Rif^R based assays were the same as those for the pGRAK based strains. As noted previously, the calculation of μ relies heavily on the accuracy of the N_t values across a set of parallel tubes. The observation that the average N_t values for the Rif^R and pGRAK based assays are

similar means their corresponding μ values can be compared. The lower average N_t value for the His reporter-based assay means the corresponding μ value cannot be compared to those obtained for the pGRAK based strains. Furthermore, the discrepancy between the pGRAK and His reporter-based μ values reflects the affect of low N_t values on the calculation of μ.

4.2.4 Limitations to reporter-based assays

Overall, the lack of differences in μ values between the pGRAK reporter-based strains suggests that the DNA metabolic enzymes tested, namely EP DNA polymerases and RecA, are not involved in the mutational events detectable using the pGRAK reporter. The lack of a role for RecA in producing the spectrum of pGRAK-derived mutations shown in Table 7 is consistent with the lack of involvement of this protein in strand slippage at repeat elements in other bacteria (Allred *et al.*, 1990; Jonsson *et al.*, 1992; Levitskaya *et al.*, 1997; Josenhans *et al.*, 2000). However, like any mutational target, the pGRAK reporter is inherently limited in terms of the mutational events that can be scored, which in this case, are only those that give rise to viable mutants that can form colonies on selective media. In the case of the pGRAK reporter system, these would be mutational events that lead to the expression of a functional Aph fusion protein that can confer Km^R. Further, conclusions can only be drawn from those experimental conditions that were tested, which might not allow potentially important contributions to be revealed. For example, determining the potential contribution of DnaE2 may require the assay to be performed under conditions in which this protein is induced, such as UV irradiation or exposure to mitomycin C.

4.3 Spectrum of mutations associated with PE-PGRS genes

It was interesting to observe all the different types of mutational events within the Morf2 region that abolished the TAG codon resulting in the production of a functional *aph* transcript. These included simple point mutations as well as more complex mutational events involving deletions, deletion-insertions and other rearrangements between proximal and / or distal repeat sequences. Certainly the lack of RecA involvement suggests a 'stuttering' by the mycobacterial replication complex (PolIII) when it encounters sequences that contain a number of PGRS repeats in close proximity to one another. It seems highly plausible that no specialised polymerases are recruited to the replication fork when it encounters such sequences and that the mutational events are simply a consequence of slipped-strand mispairing during replication by the normal replicative polymerase.

When PE-PGRS sequence polymorphisms were first identified at PE-PGRS and other G+C-rich loci in *M. tuberculosis*, it was proposed that the most likely mechanism underlying the production of repeat polymorphisms would be slipped-strand mispairing (Poulet and Cole, 1995). The results presented in this study provide the most compelling evidence to date in support of such a mechanism. The fact that the same spectrum of mutations occurred at log and saturation phase using the pGRAK reporter,

reinforces the idea that these events are replication dependant and not a result of an aging bacterial population. Combined with the lack of change in μ in the EP polymerase related strains in this study, this further reinforces the idea that the mutations in the PE-PGRS region of pGRAK investigated in this study were a consequence of replication dependent strand slippage. It was also interesting to note that despite the predominance of slipped-strand mispaired mutations, point mutations were still detected. Therefore, although mutations caused by slipped-strand mispairing are common in PE-PGRS-containing sequences, they do not occur at a rate that is sufficiently high to mask base substitution mutations within the same locus.

The wide spectrum of mutations gleaned in this study is in line with the sequence variation observed in clinical isolates of *M. tuberculosis* (Brennan *et al.*, 2001; Banu *et al.*, 2002; Brennan and Delogu, 2002; Fleischmann *et al.*, 2002). Therefore, although the results here imply that PE-PGRS loci are hypermutable, they are only moderately so. Furthermore, while the mutation rates exhibited in the pGRAK reporter-based assays were higher than those observed for Rif^R in wild type cells, they were in line with the spectrum of mutations that occurred. Notably, a recent study using TraSH methodology showed that three genes (Rv1807, Rv3872 and Rv3873) from the PE and PPE families were essential for growth in mice (Sassetti *et al.*, 2003). Another TraSH-based study by the same group showed that fifty-two PE and thirty-one PPE genes were non-essential for the growth of *M. tuberculosis* H37Rv *in vitro* (Sassetti *et al.*, 2003). These data suggest that either most of the PE and PPE genes are able to complement each other or that they are required under conditions that were not tested (Sassetti and Rubin, 2003).

Regardless, the fact that 10% of the genome or 4% of the proteome of *M. tuberculosis* H37Rv is encoded by the PE and PPE gene families implies that they are playing an important biological role in ensuring mycobacteria can survive and interact with the constituents in their immediate environments. This has implications for TB vaccine design because if PE-PGRS proteins are antigens that can illicit an immune response, they should be considered as possible epitope-based vaccines. PPE antigen Mtb39A (Rv1196) has already been tested as a subunit vaccine against TB infection (Reed *et al.*, 2003). This type of vaccine has been shown to provide advantages over whole-protein-based vaccines (Fonesca *et al.*, 2001) and if the vast choice of potential PE and PPE proteins is considered, there are a large number of potential alternatives to the current BCG vaccine.

4.4 Conclusions

A number of important conclusions can be drawn from this study. Firstly, a reliable and reproducible methodology was developed for measuring mutation rates in *M. smegmatis* by fluctuation analysis. The experimental value of the pGRAK reporter was evident by its ability to reveal a spectrum of mutations that was consistent with the types of mutation observed at PE-PGRS loci in clinical isolates of *M. tuberculosis*. The evidence obtained using this reporter system suggests that slipped-strand mispairing between proximal and distal PGRS sequences located *in cis* is the predominant type of mutational event at such loci. Moreover, slipped-strand mispairing at such loci occurs at a moderately higher rate than base substitution mutagenesis and is mediated by the normal replicative polymerase.