THE MEASUREMENT OF GENETIC DIVERSITY IN *MYCOBACTERIUM TUBERCULOSIS* USING RANDOM AMPLIFIED POLYMORPHIC DNA PROFILING

THESIS

Submitted in fulfilment of the requirements for the Degree of

MASTER OF SCIENCE

of

RHODES UNIVERSITY

by

Sharon M Richner

January 2000

TABLE OF CONTENTS

ABSTRACT

Mycobacterium tuberculosis has caused a resurgence in pulmonary disease in both developed and developing countries in recent times, particularly amongst people infected with the human immunodeficiency virus. The disease has assumed epidemic proportions in South Africa and in the Eastern Cape Province in particular. Of further concern is the isolation of increasing numbers of multiply drug resistant strains. Knowledge of the genetic capability of this organism is essential for the successful development of novel antibiotics and vaccines in an attempt to bring the global pandemic under control. Measurement of the genetic diversity of the organism may significantly contribute to such knowledge, and is of vital importance in monitoring epidemics and in improving treatment and control of the disease. This will entail answering a number of questions related to the degree of genetic diversity amongst strains, to the difference between urban and rural strains, and between drug resistant and drug sensitive strains, and to the geographical distribution of strains. In order to establish such baseline information, RAPD profiling of a large population of isolates from the western and central regions of the Eastern Cape Province was undertaken. A smaller number of drug resistant strains from a small area of KwaZulu-Natal were also analysed, with a view to establishing the genetic difference between strains from the two provinces. Cluster analysis, analysis of molecular variance and Geographical Information Systems technology were used to analyse the RAPD profiles generated. An unexpectedly high degree of genetic diversity was detected in strains from both provinces. While no correlation was seen between genetic diversity and either urban-rural situation or geographical location, a small degree of population structure could be correlated with drug resistance in the Eastern Cape. Furthermore, a significant degree of population structure was detected between strains from the two provinces, although this was still within the parameters for conspecific populations. Future work is necessary to further characterise strains from rural areas of both provinces, as well as from the eastern region of the Eastern Cape in an attempt to pinpoint the cause of the separation of the provincial populations.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AFLP	amplified fragment length polymorphism
AIDS	acquired immunodeficiency syndrome
AMOVA	analysis of molecular variance
AP-PCR	arbitrarily primed PCR
BCG	Bacille Calmette-Guerin
CCV	cophenetic correlation value
CFU	colony forming units
CTAB	*N*-cetyl-*N,N,N*-trimethyl-ammonium bromide
DAF	DNA amplification fingerprinting
DGGE	denaturing gradient gel electrophoresis
DNA	deoxynucleic acid
dNTP	deoxynucleotide triphosphate
DOTS	Directly Observed Treatment, Shortcourse
DR	direct repeat
DVR-PCR	direct variable repeat PCR
EDTA	ethylene diamine trichloracetic acid
FOSA	Friends of the Sick Association
GC	guanine and cytosine
GIS	Geographical Information System
HIV	human immunodeficiency virus
HSRC	Human Sciences Research Council
INH	isoniazid
IS	insertion sequences
J	Jaccard coefficient
LRF	large restriction fragment
M	migration rate
MAAP	multiple arbitrary amplicon profiling
MDR	multiple drug resistance
MDR-TB	multiple drug resistant tuberculosis
$MgCl_2$	magnesium chloride

MLEE	multilocus enzyme electrophoresis
MPTR	major polymorphic tandem repeat
NaCl	sodium chloride
NAP	*p*-nitro-α-acetyleamino-β-hydroxy-propiophenone
NJ	Neighbour Joining
NL	Nei and Li's coefficient
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
P.E.	Port Elizabeth
PFGE	pulsed-field gel electrophoresis
PIM	Phenetic Ingroup Method
RAMPO	random-amplified microsatellite polymorphism
RAPD	random amplified polymorphic DNA
RFLP	restriction fragment length polymorphism
SAHN	sequential agglomerative hierarchical non-overlapping
SAIMR	South African Institute for Medical Research
SANTA	South African National Tuberculosis Association
SDS	sodium dodecyl sulphate
SMC	simple matching coefficient
SSCP	single-strand conformational polymorphism
SSR	simple sequence repeat
TGGE	thermal gradient gel electrophoresis
TIF	tagged image file
UK	United Kingdom
UPGMA	unweighted pair group method using arithmetic averages
VNTR	variable number of tandem repeats
V	volts
WHO	World Health Organisation

ACKNOWLEDGEMENTS

This study was largely funded by Glaxo-Wellcome (UK), and administered by the Medical Research Council, as part of the South African section of the global *Action TB* initiative. My sincere appreciation to my supervisor, Professor Ralph Kirby, for his assistance and guidance, and for his belief in myself and the project. This study would not have been possible without the help and encouragement of the Staff and students of the Department of Biochemistry and Microbiology, especially Dr Stephanie Burton, who proofread the work in progress; Jacqui Goodwin, Susan Brown and Philippa Norman, fellow students in the Molecular Genetics Laboratory, who shared their insights and technical expertise freely; the support of the Technical Staff. My sincere appreciation to the following, who all made a vital contribution to this study: Jill Meiring and staff, SAIMR P.E., for provision of the cultures; Dr Rob Harris and Mrs Caryll Tyson, Geodatec, for the Eastern Cape GIS maps; Mrs Susan Abraham, Geography Department, Rhodes University, for the Natal map; Mr Rob Cross and staff, Rhodes University's Electron Microscopy Unit, for assistance with scanning of figures; Paul and Luc Vauterin, and Herre Heersma for assistance with GelCompar software; Laurent Excoffier, University of Geneva, for assistance with the Arlequin software; Gaby Pfyffer and Amalio Telenti of the universities of Zürich and Berne, respectively, for sharing their insights and expertise; Mrs Monteith, P.E. Health Department, and SANTA National, for providing reams of demographic information; Dr Fanny Kiepiela, Virology Department, University of Natal, for assistance with the Natal study; Dr Fernanda da Silva-Tatley and staff, Medical Biochemistry, University of Cape Town, for assistance with DNA extraction protocols; HSRC, for provision of a digital map of the Eastern Cape clinics; Dr Nigel Barker, Botany Department, Rhodes University, for guidance in analysis of results; Dr Sunday Oghiakhe, formerly of the Zoology Department, Rhodes University, for providing much needed equipment; Mike Lesar, SANTA Eastern Cape, for ongoing encouragement and interest; Professor Paul van Helden, Dr John Hauman and Dr Rob Warren, MRC Centre for Molecular and Cellular Biology, University of Stellenbosch, for willingness to share their knowledge. It has been a privilege to interact with these and many others during the course of this study. And then, last but not least, my warmest thanks and love go to my husband and best friend, Jürg; my mother, Yvonne Cleland; my South African and Swiss family and friends, without all of whom I would not have been able to do any of this; and to the One whose love has made everything possible.

CHAPTER 1
INTRODUCTION

Mycobacterium tuberculosis has been present in the human population since antiquity, as evidenced by signs of tubercular infection in Neolithic skeletal remains, as well as in spinal column fragments of Egyptian mummies dating from 2400 BC (Festenstein and Grange, 1991). The disease was identified by 460 BC as the most widespread of the times (NJMS National Tuberculosis Center, 1996). Early Hindu writings, and those of Greek and Roman historians, describe the signs and symptoms of the disease that was then called consumption or phthisis. Bacilli have also been detected in the internal organs of pre-Columbian Peruvian mummies in South America, calling into question the belief that Columbus introduced tuberculosis to the Americas.

It is thought that tuberculosis was originally a zoonose which succeeded in making the transition to humans at the time of the domestication of animals by early societies (Taylor *et al.*, 1999). This hypothesis has been substantiated by the detection of microbiological and biochemical characteristics in *M. tuberculosis* that are almost identical to those in *M. bovis*, the organism responsible for causing disease in cattle and other wildlife species, as well as humans (Taylor *et al.*, 1999). Nucleotide sequence analysis has strengthened this hypothesis by seemingly revealing an absence of allelic variation in certain genes amongst modern isolates of *M. tuberculosis,* which seems to suggest an evolutionary origin for the organism within the last 15 000 to 20 000 years (Taylor *et al.*, 1999). Furthermore, the apparent lack of divergence following gene duplication is consistent with the hypothesis that the organism is relatively young in evolutionary terms (Cole *et al.*, 1998). Techniques such as spoligotyping, which are proving useful in the field of microbial paleogenetics, have recently shown the existence of a closer genetic relationship between human isolates of *M. tuberculosis* and caprine, as opposed to bovine, *M. bovis* isolates (Taylor *et al.*, 1999). This suggests transmission from goats, rather than cattle, as the possible origin of the disease in humans, which would correlate well with the prior domestication of goats in Southwest Asia (Taylor *et al.*,1999).

In 1882 the staining technique developed by Robert Koch enabled visualisation of the bacteria responsible for the disease (Collins, 1998). Tuberculosis has always been inextricably linked to

the poverty, malnutrition and overcrowded living conditions of disadvantaged members of society. Young children, adolescents and the elderly may be affected more than other age groups. No other disease has been so prevalent and widespread over such an extensive period of time. The tuberculosis mortality rate in nineteenth century Britain was higher than the combined rate due to smallpox, measles, whooping cough, scarlet fever and typhus, while the disease is estimated to have caused nine million deaths in nineteenth century France (Metcalf, 1991).

Human tuberculosis presents itself in two forms - primary and postprimary. Primary infection results at the time of first exposure to the organism, and may or may not proceed to active infection. Active infection occurs in about 10% of newly infected people due to the inability of the immune system to contain the organism at the original focus of infection (Festenstein and Grange, 1991). Where the immune system does contain the infection, one or more organisms may remain at the original focus in a dormant, though viable, form. Postprimary infection may then occur at a later stage as a result of endogenous reactivation due to debilitation of the immune system by a disease such as acquired immunodeficiency syndrome (AIDS) caused by the human immunodeficiency virus (HIV) (Kochi, 1991). Other factors that predispose individuals to postprimary infection include malnutrition, overcrowding, stress and hormonal imbalance. Postprimary infection occurs most frequently in industrialised countries amongst the elderly (Kochi, 1991). The scenario is different in developing countries, with all age groups being susceptible to primary and postprimary infection.

1.1 Tuberculosis in the twentieth century

The development of artificial bacterial culture media in the early part of this century made possible the *in vitro* cultivation of *Mycobacterium bovis*. Attenuated cultures of low virulence were then used as vaccines, which provided protection against infection with *M. tuberculosis*. The subsequently-named Bacille Calmette-Guerin (BCG) vaccine contributed to the successful eradication of tuberculosis in the United Kingdom. However, the vaccine has not been as effective in South Africa, even though routine BCG vaccination forms an integral part of primary health care, with 94.8% of children having received it by their first birthday (Department of Health, 1998b). The possible reasons for its inefficacy will be addressed at a later stage. BCG vaccination has been discontinued in America in order to facilitate the use of the tuberculin skin test for detection of primary infection (Wayne and Kubica, 1986).

Early antimicrobial agents had no effect on *M. tuberculosis*. However, the development of streptomycin in 1943 made a significant impact on the disease process (Festenstein and Grange, 1991). However, the new antibiotic produced undesirable side effects and led to the early development of resistant strains of the organism. Fortunately, a rapid succession of anti-tuberculosis drugs were developed in the following years and it was demonstrated that combination therapy with two or three of these was able to effect a cure and overcome the problem of resistance. By the middle of this century, it was thought that the scourge of tuberculosis was fast being banished from large areas of the globe, with the disease having been transformed into one that was eminently curable.

1.1.1 Factors responsible for the resurgence of tuberculosis

Industrialised countries showed a steady drop in tuberculosis incidence until approximately the mid-1980s (NJMS National Tuberculosis Center, 1996). Consequently, basic tuberculosis research and drug development ground to a halt in the 1970s, especially in Britain and America (Brower, 1996). However, tuberculosis remained endemic in Africa, Asia and South America due to poor socio-economic conditions. Then, from the mid-1980s, a steady increase in the incidence of tuberculosis in both developed and developing countries was seen, which has been attributed to a number of factors. The past three or four decades have seen an increasing amount of population movement from developing into developed countries. As a result approximately 50% of tuberculosis cases in many industrialised countries occur amongst such immigrants (World Health Organisation (WHO), 1998a). Between 1985 and 1991, the number of cases reported in the United States rose from 22 201 to 26 283, with the incidence in New York City alone doubling between 1983 and 1992 (Small and Moss, 1993). In developing countries, the steady increase in tuberculosis rates over the past three decades can be attributed in part to population mobility due to ongoing military conflict. Epidemiological surveillance and treatment of such populations is extremely difficult, and it is estimated that approximately fifty percent of the world's refugees may be suffering from tuberculosis (WHO, 1998a).

Twenty two countries were identified in 1997 as having the highest tuberculosis burden, with twelve of these having an estimated incidence of over 250 cases per 100 000 population (Table 1.1) (WHO, 1999). Of these,. The South African incidence rate may be affected by the use of a higher total population figure than was subsequently determined by the 1996 Census, which estimated it to be in the region of only 38 million (Central Statistical Services, 1998).

Table 1.1 Estimated incidence of tuberculosis: top twelve countries, 1997 (WHO, 1999)

Country	Population x 1000	Incidence rate per 100 000
Cambodia	10 516	539
Zimbabwe	11 682	538
South Africa	43 336	392
Afghanistan	22 132	333
Uganda	20 791	320
Philippines	70 724	310
Tanzania	31 507	308
Kenya	28 414	297
Indonesia	204 323	285
Peru	24 367	265
Democratic Republic of Congo	48 040	263
Ethiopia	60 148	260

The second factor, the HIV/AIDS pandemic, has probably played the greatest role in the worldwide resurgence of tuberculosis in both developed and developing countries (Festenstein and Grange, 1991). Approximately one third of the increase in the incidence of tuberculosis in the last five years is estimated to be due to infection with HIV (WHO, 1998a). Nearly 40% of the 33 million people living with HIV/AIDS are also infected with *M. tuberculosis* (Narain, 1999). It is estimated that, by the end of the century, HIV infection will have caused approximately 1.5 million cases of tuberculosis annually that otherwise would not have occurred (WHO, 1998a). This is due to the immunosuppressive effect that HIV exerts, which neutralises the immune system's protection against active tuberculosis. Tuberculosis then results from the reactivation of an endogenous tuberculous lesion or from recent exogenous infection (Festenstein and Grange, 1991). It is the most frequently occurring of the secondary infectious diseases in people infected with HIV (WHO, 1998a). HIV infection seems to accelerate the course of tuberculosis, leading to premature death in AIDS patients (Small and Moss, 1993). To date, tuberculosis has accounted for almost one-third of AIDS deaths worldwide (WHO, 1998a).

A third important factor pertains to poor management of tuberculosis treatment programmes (WHO, 1998a). The administration of a combination of effective antibiotics at an effective dosage and for a sufficient length of time, is vital in the treatment of this disease. The unreliable supply of such antibiotics, especially in developing countries, plays a major role in the breakdown of tuberculosis control programmes.

Such poorly supervised and inadequate tuberculosis treatment programmes have also played an important role in the increased development of drug resistant strains over the past decade, which is the fourth factor to which the modern tuberculosis epidemic can be attributed (WHO, 1995). Patient non-compliance in both developed and developing countries has further contributed to the development of multiply drug resistant (MDR) strains (NJMS National Tuberculosis Center, 1996). Such multiple drug resistant tuberculosis (MDR-TB) is extremely difficult and expensive to treat, and may be fatal should the strain develop resistance to all available antibiotics (WHO, 1998a). As many as 50 million people worldwide may be infected with MDR-TB (WHO, 1995).
These four factors have together been responsible for the fact that tuberculosis today kills more young people and adults than any other infectious disease. The global picture has become increasingly ominous, with one third of the world's population having been exposed to the organism, approximately 20 million of which were cases of active infection at the beginning of the 1990s (Kochi, 1991; Bloom and Murray, 1992). Today tuberculosis is the single leading cause of death in developing countries, where it is responsible for 26% of avoidable adult deaths (Eastern Cape Department of Health, 1997).

1.1.2 Tuberculosis and HIV/AIDS in Africa and southern Africa

Tuberculosis was the leading cause of death in sub-Saharan Africa at the end of the 20th century, as well as the biggest contributor to the disease burden (Eastern Cape Department of Health, 1997). In several African countries, tuberculosis cases have doubled or even trebled in the past ten years (WHO, 1998a). The 1996 tuberculosis demographics of the sub-continent of southern Africa, as estimated by the WHO and officially reported by the countries in that area, demonstrate the difficulty in obtaining an accurate idea of the degree of disease existing there (Table 1.2) (WHO,1998c).

Table 1.2 Tuberculosis in southern Africa in 1996 (WHO, 1998c)

Country	Estimated number of tuberculosis cases	Officially reported number of tuberculosis cases
Angola	25 166	15 424
Botswana	5 936	6 630
Lesotho	5 195	4 361
Malawi	17 032	20 630
Mozambique	33 634	18 443
Namibia	6 300	6 773
South Africa	105 983	91 578
Swaziland	1 762	3 893
Tanzania	57 594	44 416
Zambia	28 549	40 417
Zimbabwe	23 678	35 735

South Africa carries the heaviest tuberculosis burden on the sub-continent, with more than twice as many reported cases as Tanzania, the country with the next highest number of cases.

It is difficult at present to determine the incidence of tuberculosis in South Africa. The provincial figures are probably not an accurate reflection of the amount of tuberculosis that actually exists in the country, due to the reorganisation of provincial health departments since 1994 which will undoubtedly have resulted in a significant amount of under-notification (Table 1.3) (Department of Health, 1999). This is thought to be the case in the Eastern Cape Province in particular, where three health departments were amalgamated.

As previously mentioned, Africa finds itself in the grip of a devastating AIDS epidemic at the end of the twentieth century. One in every 40 people in Africa is estimated to be HIV-positive, compared with the worldwide figure of one in every 250. The situation is the most severe in sub-Saharan Africa, which is home to 67% of the people who became infected with HIV in 1998 (Table 1.4) (WHO, 1998d).

Table 1.3 Incidence of tuberculosis in South Africa in 1997 (Department of Health, 1999)

Province	Population	Tuberculosis incidence rate per 100 000
Northern Cape	840 321	523
Western Cape	3 956 875	496
Free State	2 633 504	334
Eastern Cape	6 302 525	316
KwaZulu-KwaZulu-Natal	8 417 021	244
Gauteng	7 348 423	241
Mpumalanga	2 800 711	105
North West Province	3 354 825	146
Northern Province	4 929 368	94
Total	40 583 573	255

Table 1.4 Regional distribution of persons living with HIV/AIDS in 1998 (WHO, 1998d)

Region	Number of cases
Sub-Saharan Africa	22 500 000
South & South-East Asia	6 700 000
South America	1 400 000
North America	890 000
East Asia & Pacific	560 000
Western Europe	500 000
Caribbean	330 000
Eastern Europe & Central Asia	270 000
North Africa & Middle East	210 000
Australia & New Zealand	12 000
Total	33 400 000

The situation in South Africa is very serious, as seen from the results of the seventh national antenatal clinic survey carried out in 1996 (Department of Health, 1997). These surveys form the cornerstone of HIV surveillance in the absence of an official AIDS notification policy. Figure 1.1 shows the annual increase in the level of HIV infection among women attending antenatal clinics

from 1990 to 1996. A sharp rise in HIV positivity occurred in 1994, with a continuing upward trend in the following two years. The exponential growth curve predicted a steep climb in HIV prevalence in the years ahead, which has indeed been the case. The HIV cases and AIDS deaths for the western region of the Eastern Cape for the same period of time reflect the same exponential growth, as can be seen in Figure 1.2 (City Health Department, 1997).

Table 1.5 compares the results of the 1996 antenatal clinic survey with those of the previous two years, according to province. The sharp increase of 35% between 1995 and 1996 in the total number of HIV positive pregnant women can largely be attributed to an unexplained three-fold rise in prevalence in the North West province. The increase in HIV positivity rate in the Eastern Cape between 1995 and 1996 was also in the order of 35%. Mpumalanga was the only province where the prevalence rate fell between 1995 and 1996.

Table 1.5 HIV positivity rate amongst antenatal clinic attendees in South Africa between 1994 and 1996 (Department of Health, 1997)

Region	1994 (%)	1995 (%)	1996 (%)
Western Cape	1.16	1.66	3.09
Eastern Cape	4.52	6.00	8.10
Northern Cape	1.81	5.34	6.47
Free State	9.19	11.03	17.49
KwaZulu-Natal	14.35	18.23	19.90
Mpumalanga	12.16	16.18	15.77
Northern Province	3.04	4.89	7.96
Gauteng	6.44	12.03	15.49
North West	6.71	8.30	25.13
South Africa	**7.57**	**10.44**	**14.17**

1.1.3 Tuberculosis prevention and control

Figure 1.3 shows the potential trends in tuberculosis infection in South Africa for the decade 1995 to 2005 (Medical Research Council, 1997). Were the dual epidemic left to rage uncontrolled, the incidence of tuberculosis would continue to rise dramatically. Effective control of both would result in tuberculosis incidence being approximately the same in 2005 as it was in 1995, with hope for a decline in subsequent years.

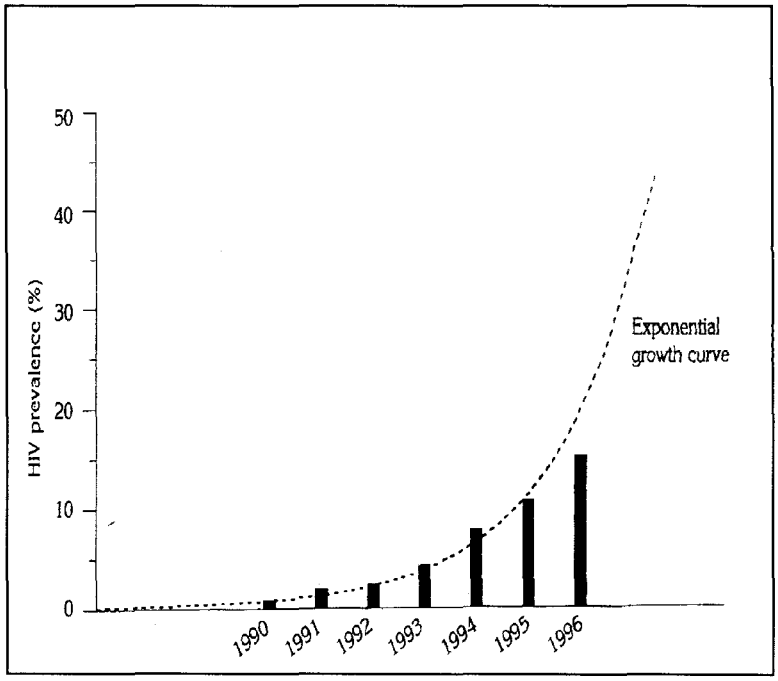


Figure 1.1 Prevalence of HIV positivity in antenatal clinic attendees in South Africa between 1990 and 1996 (Department of Health, 1997)

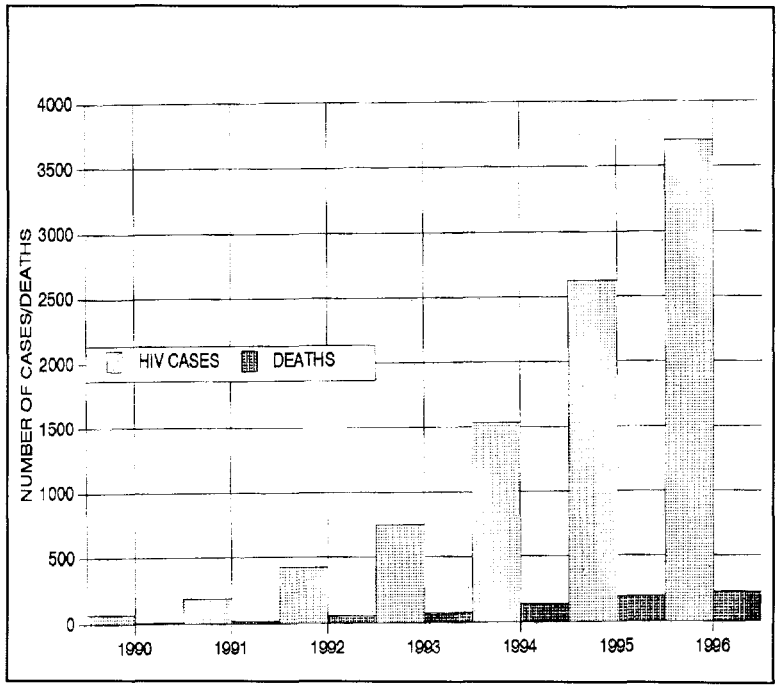


Figure 1.2 HIV positive cases and AIDS deaths in the western region of the Eastern Cape between 1990 and 1996 (City Health Department, 1997)

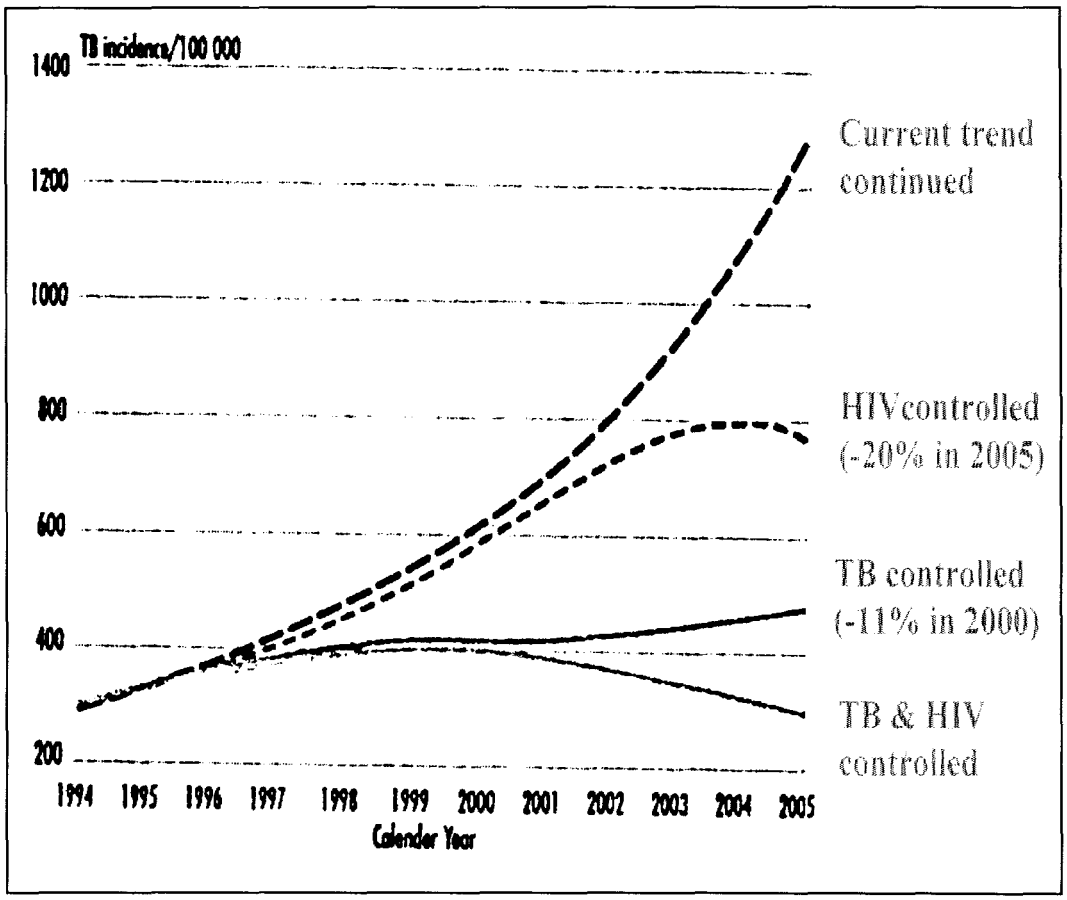


Figure 1.3 Potential trends in tuberculosis and HIV infections in South Africa between 1995 and 2005 (Medical Research Council, 1997)

However, an effective tuberculosis prevention and control strategy involves meeting a number of criteria that are critical for success. These include political commitment, improved case detection, a regular and reliable drug supply and an efficient recording and reporting system (Department of Health, 1998a). Another important aspect involves the use of a standardised short-course antibiotic treatment regimen. This was introduced in South Africa in 1996 and involves the administration to all newly diagnosed patients of four antibiotics (isoniazid, rifampicin, pyrazinamide and ethambutol) for a period of two months, followed by just isoniazid and rifampicin for another four months (Department of Health, 1996). Patients whose treatment has failed, usually as a result of non-completion of the initial course of therapy, need to be retreated for a further eight months. The first two months of the retreatment regimen involve the administration of the four antibiotics mentioned before as well as of streptomycin.

Multiple drug resistance poses a serious threat to any attempts to establish an effective tuberculosis control programme. A strain of *M. tuberculosis* is categorised as MDR on the basis of resistance at least to both isoniazid and rifampicin (Department of Health, 1996). MDR-TB infection may occur in one of two ways. Acquired resistance can be developed by the existing infecting strain during the course of treatment, usually as a result of poor patient compliance or faulty antibiotic prescription. Primary resistance, on the other hand, involves resistance to one or more antibiotics in a patient who has not received prior tuberculosis treatment and has thus been infected by an MDR strain via person to person transmission. Studies have shown that such resistant strains are not easily transmitted due to a concomitant loss of virulence, and that primary resistance occurs more frequently amongst immunocompromised than it does amongst immunocompetent people (Gillespie & McHugh, 1997). MDR-TB outbreaks have been shown to occur in particular among HIV-infected individuals (Telenti, 1997a).

The threat posed by increasing drug resistance makes it imperative to understand the genetic basis of such resistance, particularly with a view to developing new drugs (Telenti, 1997a). It is known that most bacteria use a number of strategies to achieve resistance. These include barrier mechanisms such as decreased cell wall permeability and efflux pumps, the existence of degrading or inactivating enzymes such as β-lactamases and aminoglycoside acetyl transferases, and drug-target modification, where a single mutation can occur in a key gene (Telenti, 1997a; Cole *et al.*, 1998).

The genetic information necessary to achieve such resistance may be acquired from exogenous mobile genetic elements such as plasmids or transposons, or may reside in the chromosomes of the organism due to spontaneous mutation. Mycobacteria are similar to most other bacteria in possessing the ability to use several of these strategies. The first two confer a natural resistance to many of the antibiotics used successfully against other bacteria. However, the development of resistance to specific anti-tuberculosis drugs depends on the third mechanism. Here, mycobacteria differ from many other bacteria in two respects. Development of resistance to multiple antibiotics involves a stepwise accumulation of individual mutations in several independent genes, as opposed to the "block" acquisition that is usually the case with other bacteria (Telenti, 1997a). The second, and most significant, difference is the apparent absence in *M. tuberculosis* of the means to acquire resistance from other bacteria or mycobacteria via mobile exogenous elements such as plasmids (Telenti, 1997a). Transmission of resistance is thus vertical as opposed to horizontal. Thus, a large proportion of mycobacterial resistance develops as a result of the spontaneous occurrence of point mutations, in response to selective pressure from exposure to the antibiotic, at a number of gene sites, and with subsequent vertical transmission of this property to offspring. Eleven genes have been associated with drug resistance: *katG, inhA* and *ahpC* (isoniazid), *rpoB* (rifampicin), *embAB* (ethambutol), *pcnA* (pyrazinamide), *gyrA, gyrB* and *lfrA* (fluoroquinolones) and *psL* and *rrs* (streptomycin) (Telenti, 1997a, b & c).

The organism's unique, lipid-rich cell wall has also been shown to play a vital role in the development of drug resistance. Until recently, isoniazid was most effective as an inhibitor of the synthesis of mycolic acid, which is the unique lipid component of the mycobacterial cell wall. Increasing acquisition of resistance to this staple antibiotic has resulted in a search for other agents able to target cell wall components. The identification of enoyl reductase, an enzyme that is involved in the biosynthesis of mycolic acid, led to the development of so-called kyolic acids that are able to interfere with its action (Brower, 1996). The recent decoding of the genome sequence of H37Rv has provided much information on new drug targets linked to the complex fatty acid pathways involved in the biosynthesis of the mycolic acid layer of the cell wall (Cole *et al.*, 1998; Orme and Belisle, 1999). Other potential drug targets are the mycocerocic acids unique to virulent *M. tuberculosis* isolates, as well as the large amounts of lipoarabinomannan which span the entire envelope (Orme and Belisle, 1999).

Tuberculosis research was given fresh impetus in 1993 when GlaxoWellcome (London) commenced a worldwide five year collaborative research effort, known as *Action TB*. An important aspect of

this programme involved the search for new *M. tuberculosis* targets for antimicrobial agents and developing new antibiotics (Brower, 1996). Progress was made in the area of trehalose biosynthesis, as well as of glycosolation inhibition. It was shown that the two enzymes involved in trehalose synthesis provide potential targets for a drug whose action would serve to interrupt the trehalose pathway. Research in Canada focussed on genes with transporter functions that are involved in antibiotic resistance, on the relationship between signal transduction in normal and tuberculosis-infected macrophages, on the development of a subunit vaccine, and on the role of the complement system and intracellular pathways in mycobacteria-macrophage interaction (Brower, 1996). It remains for pharmaceutical companies to convert these findings into new and effective antibiotics.

Another important aspect of an effective tuberculosis control programme is a strategy based on directly-observed therapy, which involves the use of both health workers and volunteers from the patient's community, who undertake to physically observe the patient taking the daily medication.. It is mooted as the most effective means of ensuring that patients on tuberculosis therapy complete the course, thus minimising the selection of resistant mutants. It has formed the backbone of a global strategy, called Directly Observed Treatment, Shortcourse (DOTS), which was instituted by the World Health Organisation in 1993 (Blinkhoff, 1999). Ninety five countries out of 212 worldwide have adopted the DOTS programme, with 63 of those implementing the strategy countrywide (Blinkhoff, 1999). Cure rates have been reported as ranging from 70% to 90%. South Africa implemented the DOTS strategy in 1997, at which time the tuberculosis treatment success rate was at 69% (WHO, 1999). It has not yet been possible to establish the degree of success attained in this country.

Although the effectiveness of the BCG vaccine has been variable between countries and has not been as successful in South Africa as in many Western countries, vaccination is still viewed as a desirable component of any national tuberculosis control programme and is thought by some to be the only long-term solution to the global pandemic (Orme and Belisle, 1999). BCG continues to provide satisfactory protection for newborns in most countries. It has been highly effective against tuberculosis in England, but less so in South India (Jacobs *et al.*, 1987; Young *et al.*, 1985). A recent study has shown that BCG cultures kept in different laboratories around the world have undergone various gene deletions, resulting in distinct substrains (Orme and Belisle, 1999). While there is as yet no evidence that such deletions are the basis for variable vaccine efficiency, linking

them to changes in immunogenicity and protective ability might explain BCG's variable performance, as well as provide a way to genetically manipulate this vaccine to improve its efficacy. Recent nucleotide sequencing studies of specific genes have shown a seeming lack of interstrain genetic diversity in *M. tuberculosis* which indicates that an effective tuberculosis vaccine should be able to contribute significantly to controlling the disease (Cole *et al.*, 1998). Recent advances in basic mycobacterial research have made the development of a new vaccine a realistic possibility. Much research is needed into immunological markers of protection in order to provide information on parameters such as dosage and vaccination route (WHO, 1998b). In addition, candidate vaccines need to meet a number of important public health criteria. Such vaccines must be highly immunogenic, they should not interfere with the widely used tuberculin skin test and they should be capable of being administered to immunocompromised individuals (Foulds, 1997). Strains of mutant BCG have recently been developed, which have been shown to give equal or greater protection against tuberculosis infection in mice than does the BCG vaccine (Brower, 1996). In addition, these strains have been shown to be safe in immunodeficient animals, which provides hope for the development of a vaccine for AIDS patients. However, there is concern about the role of animal models in preclinical testing of vaccines, as vaccines that protect animals may not necessarily protect humans (Foulds, 1997). Some researchers believe that animals are best used for demonstrating basic immunological responses and safety (Foulds, 1997).

It has been shown that the use of cytokines in conjunction with vaccines may be one way of improving their protective effect (Foulds, 1997). The administration of IL-2 and IL-12, along with *M. tuberculosis* culture filtrate proteins, dramatically improved the protective response in mice. In another study, the genes for five cytokines were inserted into BCG and a stronger immune response was demonstrated in mice than was the case with unmodified BCG (Brower, 1996). Vaccines that consist of the deoxyribonucleic acid (DNA) which encodes antigenic *M. tuberculosis* proteins have been shown to provoke an immune response that is similar to that obtained with a live, attenuated vaccine (Brower, 1996). A vaccine consisting of a plasmid that encodes a 65 kDa heat shock protein was shown to produce both cellular and humoral immunity in mice (Brower, 1996). Another approach has involved the stimulation of the T-cell response to *M. tuberculosis* by means of antigenic peptides. Such a vaccine could be used in conjunction with antibiotics as an indirect immunostimulatory approach (Brower, 1996). More recent approaches have involved the use of transposon mutagenesis to develop autotrophic mutants, the use of existing BCG vaccine as a recombinant vector and the use of subcellular fractionation to produce subunit vaccines (Orme and Belisle, 1999).

Elucidation of the proteomics of the organism has led to the discovery of two extensive glycine-rich protein families (PE and PPE) which, it is hoped, will prove to be immunogenic (Cole *et al.*, 1998).

1.2 The significance of genetic diversity in disease control

From the above, it is clear that knowledge of the genetic capability of this organism is vitally important when aiming to develop novel tuberculosis antibiotics and vaccines in an attempt to bring the global tuberculosis pandemic under control. An important aspect of the genetic capability of *M. tuberculosis* is the degree of genetic diversity that exists within bacterial populations in a community. The ability to measure genetic diversity and to differentiate amongst bacterial strains in such a population is of vital importance in monitoring disease outbreaks, tracking sources of transmission, monitoring the efficacy of antibiotic treatment, and in improving disease control in general (Goyal *et al.*, 1994; van Soolingen *et al.*, 1994).

Determination of such genetic diversity focuses essentially on the evolutionary changes that take place in bacterial populations (Whittam, 1995). Such changes usually occur at a much faster rate in bacteria than in eukaryotic organisms, resulting in the existence of a significant degree of genetic diversity in most bacterial populations (Maynard Smith, 1995). One of the reasons for the faster rate of evolution in bacteria is the ability of short pieces of DNA of between 100 to 1000 bases to be exchanged horizontally between two bacterial cells of the same, or of different, species. It has been suggested that short segments of DNA, known as insertion sequences (IS), play an active role in such horizontal transmission of genetic information among bacterial cells (Werner *et al.*, 1995). Such horizontal gene transfer occurs at a slower rate in certain bacterial species, which has been demonstrated particularly in populations of *Escherichia coli* and certain species of the *Salmonella* genus (Spratt *et al.*, 1995). However, the majority of genetic diversity occurs in a conspecific population as a result of asexual replication, which is due to the occurrence of mutations, such as inversions, duplications, deletions, transpositions and point mutations, that arise spontaneously when replication errors occur during periods of exponential growth in bacterial cells. Interestingly, recent studies have shown that mutations may even occur in non-replicating bacterial cells (Foster, 1995). The population structure of bacteria where genetic diversity is due more to mutation than to recombination, is characterised by clonality, with strong linkage disequilibrium existing between

alleles (Spratt *et al.*, 1995). The evolutionary history of such populations is dominated by periodic selection and stochastic extinction, and the existence of independent, non-recombining lineages that are identical by descent. Clonal populations are further characterised by the extensive geographical distribution of the same strains. Members of a particular clone are not necessarily genetically or phenotypically homogeneous (Whittam, 1995). A clone is thus an extant group of bacteria within a species, having many similarities derived from a common ancestor that are not shared by other strains.

Until the beginning of the 1990s, phage typing had been the only available method for the differentiation of strains of *M. tuberculosis* and for the determination of genetic diversity, but it was time-consuming and cumbersome, and relatively few types could be distinguished (Crawford and Bates, 1984; van Soolingen *et al.*, 1991). Over the past decade, tuberculosis epidemiology has been revolutionised by the development of a variety of techniques that generate molecular markers, which enable closely related genotypes to be distinguished and permit the measurement of genetic diversity (Karp *et al.*, 1997). This has given rise to the new discipline of molecular epidemiology, which is the integration of molecular biological techniques which identify specific strains of organisms by means of genome analysis, with the medical discipline of epidemiology which investigates the distribution of disease in human populations (Small and Moss, 1993). Molecular epidemiology has been used increasingly in the study of tuberculosis, with a view to gaining a better understanding of the modern epidemic.

1.2.1 Molecular marker techniques

A wide variety of molecular marker techniques can be utilised for measuring genetic diversity in bacterial populations. Such techniques can be grouped into a number of categories, depending on whether or not the polymerase chain reaction (PCR) is used, and whether or not specific primers are used. Amongst the non-PCR methods are restriction fragment length polymorphism (RFLP) analysis (Peillon *et al.*, 1994), simple sequence repeats (SSRs) or microsatellite analysis (Peillon *et al.*, 1994), and variable number of tandem repeats (VNTR) or minisatellite fingerprinting (Dallas, 1988; Bridge *et al.*, 1997). Minisatellites such as $(GTG)_5$ have been particularly useful in the typing of strains of *M. tuberculosis* that have few or no insertion sequences, as RFLP typing of such strains is acknowledged to be problematic (Wiid *et al.*, 1994). However, PCR-based techniques have been used more widely as they use less DNA, obviate the necessity of using lengthy probe

hybridization steps and do not require the expertise and safety precautions necessary for the handling of radioactive isotopes (Chevrel-Dellagi *et al.*, 1993). Such methods fall into two categories: targeted PCR, where specific primers are used to generate diversity data from specific sequences of DNA, and multiple arbitrary amplicon profiling (MAAP), where arbitrary primers are used to amplify a range of polymorphic DNA products (Karp *et al.*, 1997).

Targeted PCR techniques are incorporated into a wide range of molecular genetic procedures, including cDNA cloning, mutagenesis, DNA sequencing, gene walking and *in situ* hybridisation (Cobb and Clarkson, 1994). DNA sequencing used in conjunction with PCR amplification of specific genes or portions of genes, provides the most specific way of measuring genetic diversity (Karp *et al.*, 1997). Gel systems such as thermal gradient gel electrophoresis (TGGE), denaturing gradient gel electrophoresis (DGGE) and single-strand conformational polymorphism (SSCP) can detect variations down to a single base pair.

MAAP techniques include the random amplified polymorphic DNA (RAPD), arbitrarily primed PCR (AP-PCR) and DNA amplification fingerprinting (DAF) methods (Williams *et al.*, 1990; Welsh and McClelland, 1990; Caetano-Anollés *et al.*, 1991). The latter two differ from RAPD in primer length, stringency conditions and fragment detection. A more recently developed technique, known as amplified fragment length polymorphism (AFLP), employs a combination of RFLP and arbitrary primer PCR technology (Karp *et al.*, 1997). The random-amplified microsatellite polymorphism (RAMPO) method results in the informativeness of RAPD profiles being increased due to the detection of second-level amplification products using a microsatellite-complementary primer (Ramser *et al.*, 1997).

The complexity and cost of many molecular marker techniques are distinct disadvantages when measuring the genetic diversity of large populations. Most methods are time-consuming, use relatively large quantities of purified DNA, and require prior knowledge of DNA sequence. In addition, they are sensitive to the DNA shearing that inevitably occurs during extraction procedures and which is difficult to avoid when working with *M. tuberculosis*. Furthermore, the genetic information produced by these methods usually only relates to a relatively small portion of the genome. The relative rapidity, cost-effectiveness and technical simplicity of MAAP techniques have thus resulted in their widespread application to the measurement of genetic diversity in both prokaryote and eukaryote DNA (Small and Moss, 1993). They have been refined to the extent of

being able to provide reliable results using small concentrations of DNA which is not necessarily of the highest grade (Small and Moss, 1993; Palittapongarnpim *et al.*, 1993b). The utilisation of low stringency PCR amplification with single primers of arbitrary sequence makes prior knowledge of DNA sequence unnecessary and results in the generation of strain-specific profiles of anonymous, polymorphic DNA fragments (Williams *et al.*, 1990). Numerous primers can be used to generate a battery of molecular markers, resulting in a significant proportion of the genome being surveyed.

Amongst the first of the MAAP techniques to be described was RAPD. In the definitive study, Williams *et al.* (1990) amplified human, soybean, corn and *Neurospora crassa* DNA with arbitrary primers, demonstrating an optimum annealing temperature of 36°C, with amplification failure at temperatures above 40°C. Primer sequences of nine or ten nucleotides were optimal, with a GC composition of between 50% and 80% (Williams *et al.*, 1990). The importance of incorporating negative controls to exclude primer artifacts or exogenous DNA was demonstrated. Their results demonstrated that primers of differing sequence generated different RAPD profiles from the same DNA. Certain bands were common to all individuals, but sufficient unique RAPD markers were detected, using agarose gel electrophoresis, to allow the differentiation of even closely related strains of the same species. The reproducibility of RAPD markers was also demonstrated. Of great significance was the fact that this technique was able to generate DNA markers in genomic regions which would be inaccessible to RFLP analysis due to the presence of repetitive DNA sequences. The usefulness of the method was further demonstrated by the fact that single nucleotide changes in primer sequence resulted in a complete change in the RAPD profile of a particular species, indicating an ability to detect single base changes. Other sources of polymorphisms were also shown to exist, such as deletions of a priming site, insertions that rendered priming sites too distant to support amplification, and insertions that changed the size of a band without preventing its amplification. Nearly all RAPD markers are dominant, which means that it is not possible in diploid organisms to determine whether a particular DNA segment has been amplified from a locus that is heterozygous or homozygous (Williams *et al.*, 1990). However, dominance is not an issue with bacteria due to their haploid nature. The RAPD technique has subsequently found wide application in the fields of genetic mapping, plant and animal breeding, and population genetic studies.

Welsh and McClelland (1990) showed independently, using similar methodology, that primers of arbitrary sequence can be used for detecting genetic diversity. AP-PCR used primers of 20 and 34 bases in length, in conjunction with amplification conditions that consisted of one low stringency cycle with an annealing temperature of 40°C, followed by 30 to 40 high stringency cycles with an annealing temperature of 60°C. Shortly after these two landmark studies, Caetano-Anollés *et al.* (1991) developed the DAF method which demonstrated that arbitrary primers consisting of only five bases in length were able to generate detailed and relatively complex, polymorphic DNA profiles. Differences included the use of higher primer concentrations, lower template DNA concentrations, and visualisation of amplification products using polyacrylamide gel electrophoresis and silver staining. DAF profiles thus consisted of a greater number of markers than did RAPD profiles. Multiplexing of primers, where two primers are used in combination, was also shown to be able to generate meaningful DNA markers (Caetano-Anollés *et al.*, 1991). The use in this technique of low concentrations of DNA template is not ideal, as it may result in inconsistent fingerprints, with weakly amplified products appearing as bands of fainter intensity. Potential applications of DAF include identity testing, population and pedigree analysis, phylogenetic studies, genetic mapping and molecular characterisation of near isogenic lines (Caetano-Anollés *et al.*, 1991). Family studies showed the heritability of DAF markers, with offspring inheriting variant bands from either parent.

Of the three arbitrary amplification techniques surveyed above, the RAPD method has gained widest acceptance. PCR cycling using a single low stringency annealing temperature is more advantageous than the two-step annealing protocol of the AP-PCR method. Furthermore, the use of agarose gel electrophoresis is simpler than the PAGE and silver stain protocols of DAF, and is able to generate sufficient genetic information. Reliable and reproducible results, which have contributed greatly to knowledge of the genetic diversity of a wide variety of organisms, have been obtained using this technique wherever care has been taken to standardise the various elements of the RAPD-PCR reaction (Weeden *et al.*, 1992). The quality and concentration of DNA template, the composition of the reaction cocktail and the cycle conditions all need to be optimised and kept consistent throughout a particular experiment. However, after an exhaustive analysis of the effects of various reaction conditions on the fidelity of RAPD phenotype, Weeden *et al.* (1992) came to the conclusion that the amplification processes were not so sensitive to one or more of the parameters tested as to seriously affect reproducibility of results.

Criteria for scoring RAPD markers also forms a vital aspect of the discussion surrounding the optimisation of this technique. Different view points exist as to the nature of faint bands and whether they should be scored or not. One school of thought maintains that faint markers are artifactual as they may represent amplification errors, may arise from characteristics associated with the DNA itself (repetitive sequences, the amplification of *in vitro* recombinants, genetic background) or may be caused by variation in PCR reaction conditions and parameters (Weeden *et al.*, 1992; Lamboy, 1994b). Others hold the view that variation in intensity is merely a result of multiple copies of amplified regions in the template, or is due to the efficiency with which particular regions are amplified, and that faint bands should not be excluded (Caetano-Anollés *et al.*, 1991).

Whether faint bands are adjudged to be artifactual or not, the possibility does exist with the RAPD technique for the generation of artifactual, non-reproducible bands that ideally should not be scored (Weeden *et al.*, 1992). A number of approaches have been adopted towards such artifacts, including discarding faint or inconsistent bands, using only those bands from replicate PCR runs that occur reproducibly, or using all bands and accepting a certain level of error (Lamboy, 1994a). The first approach has the disadvantage of losing information, while the run replication approach is not practical when surveying substantial numbers of strains. The inclusion of all markers, on the other hand, will result in the detection of more genetic diversity than actually exists. A more effective way of negating the effect of artifacts is by the use of computer-based detection of bands, instead of the "eye-balling" methodology which was used in the early stages of PCR and RAPD-PCR development and which is still favoured in small studies. An acceptable margin of error which would take faint bands into account may thus be factored in. A full discussion follows in Chapter 2 (Section 2.4.2) on the use of the band searching filters of the GelCompar computer programme (Applied Maths, Netherlands) for the exclusion of artifactual bands in this study. The bias introduced by artifacts may further be neutralised by using a similarity coefficient which is able to take such bias into account when computing the matrix of similarities (Lamboy, 1994a). Of the three most commonly used similarity coefficients - the simple matching coefficient (*SMC*), Jaccard coefficient (*J*) and Nei and Li's coefficient (*NL*) - it has been demonstrated that the *NL* coefficient (which is a variation of the Dice coefficient of Sneath and Sokal (1973)) is superior as it is the only one that has biological meaning, in that it measures the proportion of bands that two individuals share because they have been inherited from a common ancestor (Lamboy, 1994a). This coefficient has the further advantage of not requiring the recomputing of all the similarities whenever a new

RAPD band is added to the database. This coefficient is thus recommended for use with RAPD-PCR (Lamboy, 1994a). The Dice variation of this coefficient is one of five similarity coefficients available in the GelCompar cluster analysis programme. The procedure whereby it was chosen as the most suitable coefficient for this study follows in Chapter 2 (Section 2.4.2).

RAPD-PCR has been successfully used to measure genetic diversity in a wide variety of microorganisms, such as *Listeria* spp., *E. coli*, *Lactococcus lactis*, *Haemophilus somnus*, *Helicobacter pylori*, *Staphylococcus aureus*, *Trypanosoma congolense*, *Aspergillus fumigatus*, *Histoplasma capsulatum* and *Candida albicans* (Linton *et al.*, 1994). RAPD was demonstrated to be more sensitive and reproducible for typing large numbers of *E. coli* strains, using just a few arbitrary primers, than was multilocus enzyme electrophoresis (MLEE) typing using twenty diagnostic enzymes (Wang *et al.*, 1993). It was thus deserving of consideration for use in the measurement of genetic diversity in *M. tuberculosis*.

1.2.2 Measuring genetic diversity in *M. tuberculosis*

A number of molecular marker techniques have been used for measuring genetic diversity in populations of this organism. An RFLP-based molecular marker technique, which utilises the occurrence of DNA polymorphism among strains due to the diversity in copy number and genomic location of a particular insertion sequence, has found wide acceptance (van Soolingen *et al.*, 1991; van Embden *et al.*, 1993). Insertion sequences have been used widely in the epidemiological study of a variety of pathogenic bacteria, including *Salmonella* species, *Staphylococcus aureus* and *Leptospira borgpetersenii* (Werner *et al.*, 1995). Bacterial insertion sequences are genetic entities able to translocate to new locations either within a replicon or between different replicons of a bacterial cell (Werner *et al.*, 1995). IS elements are typically 0.7 to 2.5 kilobases in size and encode only those genetic functions that relate to their transpositional activity, which distinguishes them from another class of mobile genetic elements called transposons. Such IS elements are present in different copy numbers and at different positions on the bacterial genome which means that restriction fragments carrying them are highly polymorphic, thus providing satisfactory molecular markers for measuring genetic diversity within a bacterial population (van Soolingen *et al.*, 1991). The existence of such insertion sequences in *M. tuberculosis*, and their potential for generating differentiating molecular markers, was recognised independently by Eisenach *et al.* (1988),

Zainuddin (1988) and Zainuddin and Dale (1989). The sequence of one of these elements, called IS*6110*, has been used in a large number of molecular epidemiological studies of *M. tuberculosis* (Thierry *et al.*, 1990; van Soolingen *et al.*, 1991; Chevrel-Dellagi *et al.*, 1993).

Fears that IS fingerprints might prove unsuitable for strain differentiation in *M. tuberculosis* due to the inherent instability of such transposable elements, were dispelled by demonstrating that long-term, repeated serial passages over a period of months, both *in vivo* and *in vitro,* were not able to give rise to changes in position of any IS copies amongst the strains tested (van Soolingen *et al.*, 1991). The stability of mycobacterial insertion sequences was further tested by long-term *in vitro* exposure to the influence of macrophages, as well as by the selection of drug-resistant mutants using serial passage in culture medium containing increasing concentrations of antibiotics (van Soolingen *et al.*, 1991). The acquisition of *in vivo* drug resistance as a result of treatment was also unable to alter the organism's fingerprint (Rigouts and Portaels, 1994). Furthermore, no change of fingerprint was observed in strains of the organism isolated from a patient over a period of one year (van Soolingen *et al.*, 1991). Strässle *et al.* (1997) showed that the IS*6110* pattern of multiple isolates from one particular patient remained stable for up to four years. Thus it would seem that IS*6110* is very well conserved in strains of *M. tuberculosis* (van Soolingen *et al.*, 1991). The stability of IS fingerprints has also been demonstrated in other bacterial species (Werner *et al.*, 1995).

IS*6110*-RFLP typing, henceforth simply called RFLP typing, has thus become a routine tool in certain circles for the investigation of outbreaks of tuberculosis, especially in small communities (Warren *et al.*, 1996) and in institutionalised settings and hospitals (Small and Moss, 1993). It has also been used in a number of large scale molecular epidemiology studies where the existence of identical RFLP patterns indicated infection with the same strain (van Soolingen *et al.*, 1991; Hermans *et al.*, 1995; Safi, 1997). This is usually indicative of recent contact with a common source of infection (Small and Moss, 1993). The proportion of the population that is clustered thus gives an estimation of the degree of recent infection in that community. Some studies have revealed that as many as two-thirds of cases in an epidemic population were due to recent infection (Small and Moss, 1993), while other studies showed that this was the case in only thirty percent of cases (Warren *et al.*, 1996). A high percentage of clustering has also been seen in communities with a high incidence of tuberculosis, and is often associated with risk factors for HIV (Small and Moss, 1993). The non-clustered, unique strains within populations usually represent reactivation of endogenous infection due to past exposure to the organism.

A number of important studies have been performed in which strains from different geographical regions have been characterised using this method. A study incorporating strains from the Netherlands and central Africa demonstrated that central African strains were significantly less diverse genetically than the Dutch strains (van Soolingen *et al.*, 1991). This was unexpected, as the African strains originated from three different countries. A number of other studies showed similar differences between European and African strains (Hermans *et al.*, 1995; Strässle, 1997; Safi, 1997). The level of clustering in developing African countries such as Tunisia, Ethiopia and the Central African republic, where the incidence of tuberculosis is high, has been shown to be as high as 62% (Hermans *et al.*, 1995). This low level of genetic diversity is due to the person to person transmission of a limited number of clones in a geographical region with a high incidence of tuberculosis. However, the level of clustering in Europe, where the incidence of tuberculosis is lower, was shown to be low, with the Netherlands at 19%, followed by France (28%) and Spain (38%) (van Soolingen *et al.*, 1991; Safi *et al.*, 1997). Such high levels of genetic diversity are indicative of a large degree of reactivation disease.

However, there are a number of studies which have produced results that do not fit neatly into the low incidence/high genetic diversity and high incidence/low genetic diversity paradigm. A Taiwanese study revealed a relatively low degree of genetic diversity in an area of low incidence of tuberculosis infection, using the RAPD-PCR technique (Harn *et al.*, 1997). This could have been due in part to the transmission of a small number of strains in a country which has succeeded in controlling this disease to a larger degree than elsewhere. Anomalies were also seen in the study carried out by Hermans *et al.* (1995), where the Dutch and Ethiopian populations adhered to the traditional paradigm, while the degree of diversity in eastern Tunisia was found to be lower than that in Ethiopia, even though the incidence rate in Tunisia was substantially lower (30 per 100 0000) and more in keeping with that of a European country such as Spain (Safi *et al.*, 1997). An earlier study conducted in northern Tunisia also produced contradictory results, with a high degree of genetic diversity being detected in three of the four regions sampled, but an unexpectedly low diversity being found in a fourth region (Chevrel-Dellagi *et al.*, 1993). The authors attributed the high level of diversity to the existence of a large city (Tunis) in one of the three regions, as well as to the existence of a heterogeneous human population in the other two regions. The unexpectedly low diversity in the fourth region was thought to be due to the existence of a more stable population in that area. Further contradictory results were produced by a study conducted in Dar es Salaam, Tanzania, which country had a high incidence rate of 141 per 100 000 in 1996.

Here a surprisingly high degree of diversity was detected, similar to that found in cities like New York and San Francisco (Yang *et al.*, 1995). Once again, divergence from the accepted model was attributed to the dynamics of a city inhabited by a mixed population of people originating from many parts of Tanzania as well as from other parts of the world. Dar es Salaam's position as a major trade centre in central Africa was also thought to be significant (Chevrel-Dellagi *et al.*, 1993).

The RFLP technique has also been used successfully in community-wide monitoring of drug resistant strains in America, where it has been used to determine the degree of acquired, as opposed to primary, drug resistance in a community (Small and Moss, 1993). Acquired resistance is characterised by a high proportion of unique RFLP patterns, whereas limited RFLP pattern diversity indicates primary resistance as a result of clonal dissemination of a few resistant strains within the community (Small and Moss, 1993).

A number of other RFLP-based methods have been developed for use in laboratory diagnosis as well as in the field of tuberculosis epidemiology. A method developed by Eisenach (1994) incorporates PCR amplification of genomic DNA, using primers specific to a 123-bp region within IS*6110*. The advantage of this modification is that the test takes slightly less time to complete, making it suitable for diagnostic application. Haas *et al.* (1993) developed a relatively rapid technique, known as mixed-linker PCR, using a double-stranded oligonucleotide mixed linker. Good correlation with RFLP results was shown (Haas *et al.*, 1993). Zhang *et al.* (1992) used infrequent-cutting restriction endonucleases such as *Dra*I, *Xba*I or *Asn*I to generate a small number of large restriction fragments (LRFs), which could only be visualised using pulsed-field gel electrophoresis (PFGE). LRF typing was shown to be comparable to IS*6110*-RFLP typing in their hands, and obviated the need to work with radioactive isotopes. Plikaytis *et al.* (1993) developed a technique known as ampliprinting, where IS*6110* and major polymorphic tandem repeat (MPTR) primers were used in a unilateral-nested amplification procedure. Once again, good correlation with RFLP typing was demonstrated.

Other molecular marker techniques that have been used to measure genetic diversity in *M. tuberculosis* include a ligation-mediated PCR procedure, which detected polymorphisms by amplifying the flanking sequences on both sides of IS*6110* (Palittapongarnpim *et al.*, 1993a). This method is faster than the traditional RFLP protocol, technically simpler to perform and as sensitive

as RFLP. Goyal *et al.* (1994) developed a strain typing method based on one particularly polymorphic DNA sequence which flanks the *katG* gene, and which was shown to be specific to the *M.tuberculosis* complex. They used this protocol to type 130 *M. tuberculosis* isolates, and were able to distinguish seven different subtypes, fewer than were generated by IS*6110* analysis of the same isolates.

Detection of genetic diversity amongst strains of *M. tuberculosis* on the basis of changes in the actual nucleotide sequence of various regions of the genome provides the most specific way of measuring interstrain diversity. Valuable information on the genetics of the organism has been provided by the recently completed sequencing of the entire circular chromosome of H37Rv, the type strain of *M. tuberculosis* (Cole *et al.*, 1998). The genome of H37Rv consists of 4 411 529 base pairs with a GC content of 65.5%, encoding approximately 4000 genes (Cole *et al.*, 1998). The GC content is distributed relatively evenly throughout the genome, which seems to indicate that horizontally transferred pathogenicity islands of atypical base composition are absent. The genome is rich in repetitive DNA, in particular insertion sequences, and in multigene families and duplicated housekeeping genes (Cole *et al.*, 1998). Sixteen copies of IS*6110* occur in the H37Rv genome, along with copies of another fourteen insertion elements. Sequencing techniques developed before this information was available include direct variable repeat PCR (DVR-PCR), which exploits the high degree of polymorphism situated in the region of the genome known as the direct repeat (DR) cluster (Groenen *et al.*, 1993). The majority of Dutch strains in Groenen's study exhibited marked genetic diversity, while 85% of the Indian strains were identical (Groenen *et al.*, 1993). This method is relatively simple on the level of analysis and generates reproducible results However, the complicated and lengthy methodology is a distinct disadvantage when sampling large numbers. Two large, unrelated multigene families, PE and PPE, may provide new targets for measuring diversity as they have been shown to contain a certain degree of polymorphism (Cole *et al.*, 1998). The gene coding for the PE-PGRS protein Rv0746 of BCG was found to differ from that in H37Rv. Similar variation was seen in the gene for the PPE protein Rv0442. Comparative sequence analysis will have to be performed on larger numbers of strains to substantiate this finding. However, sequencing methods on the whole are not able to provide an accurate indication of the degree of genetic diversity in this organism as they survey only small regions of the genome. Furthermore, these protocols are not always suitable for large scale epidemiological studies, as they are lengthy and complex, involve working with radioactive isotopes and require sophisticated and expensive equipment, as well as a high degree of expertise.

The RAPD-PCR technique has been investigated for application to strain identification and measurement of genetic diversity of *M. tuberculosis* by a number of researchers. Palittapongarnpim *et al.* (1993b) characterised a small population of Canadian strains using three primers singly and in various combinations. It was shown that reliable differentiation of strains could only be achieved by analysing the RAPD markers generated by several sets of primers. Good correlation with RFLP typing was demonstrated, with the same degree of polymorphism being detected by both methods. It was not possible, however, to differentiate between closely related strains, with identical RAPD profiles being obtained for *M. tuberculosis* type strains H37Rv and H37Ra (Palittapongarnpim *et al.*, 1993b). Linton *et al.* (1994) used RAPD to survey strains from the United Kingdom, India, Malawi, Kenya and Saudi Arabia using forty different primers ranging in length from ten to twenty bases. Only four of these generated RAPD profiles that showed good discrimination of strains, with eight producing markers that were able to discriminate slightly between strains. Nine generated bands which were not able to discriminate between strains, while nineteen were unable to generate RAPD markers at all. The primers that were most successful at strain differentiation were twenty nucleotides in length. Once again, sufficient discrimination was only achieved by using a number of primers (Linton *et al.*, 1994).

Linton *et al.* (1995) subsequently undertook a comparative study of the RAPD and RFLP techniques as epidemiological typing methods for differentiating strains of *M. tuberculosis*. Fairly good correlation was shown between the RFLP and RAPD groupings, with most discrepancies being found with strains that had only one copy of IS*6110*. This was expected, as the difficulty of typing strains with a low copy number of IS*6110* is well documented (Hermans *et al.*, 1995). Harn *et al.* (1997) reported the use of RAPD analysis both for strain typing as well as for investigating the pathogenesis of tuberculosis in Taipei City. The majority of cases were due to recently transmitted, exogenous disease, with the smaller percentage attributable to reactivation of endogenous infection. However, the clusters of recent infection occurred over a limited time period only, while strains representing reactivation infection were present throughout the time period surveyed. The clustered strains in this study thus reflected the presence of a series of mini-epidemics, while the unique strains probably represented reactivation in elderly people who came over to the island having previously contracted tuberculosis on the mainland, and who had until that time been passive carriers (G.C. Liu, personal communication).

These studies have indicated the potential of the RAPD technique to provide valuable genetic and epidemiological information on the causative organism of a disease that has once again reached epidemic proportions in many parts of the world.

1.3 Investigating the geographical distribution of *Mycobacterium tuberculosis*

Application of geographical information system (GIS) technology to the field of epidemiology has provided new opportunities to study the associations between environmental exposure to infectious agents and the spatial distribution of the concomitant disease (Vine *et al.*, 1997). This computer-based technology was first developed in the field of geographical science, but is being used increasingly in other scientific disciplines as well. It is of great value in natural resource management, where it is one of the principle technologies available for the investigation of landscapes (Coulson *et al.*, 1991). GIS is able not only to represent landscape features, but can also be used to predict the consequences of a contemplated action, to evaluate the results of actions that have already been taken, and to compare alternative action scenarios. GIS may also be used for carrying out environmental impact assessments, where the integration and analysis of data relating to soils, land use, current vegetation and administrative districts is required (Coulson *et al.*, 1991).

A GIS is a computerized mapping system for capture, storage, management, integration, analysis and display of spatial and descriptive data (Coulson *et al.*, 1991). A GIS programme comprises the four fundamental components of a data input subsystem to collect and process spatial and descriptive data from maps, aerial photographs and other sources; a database management subsystem to store and retrieve data; an analysis subsystem to interpret within and among data themes; and a reporting subsystem or data output to display maps and reports (Coulson *et al.*, 1991). Most GIS systems have the ability to interface with independent databases, mathematical models, evaluation functions and statistical analysis systems. GIS should not be confused with computerised cartographic systems. While the main function of the latter is to generate computer-stored maps, GIS serves to create information by integration of data layers so that the original data is shown in different ways and from different perspectives (Coulson *et al.*, 1991).

The advantages of GIS include the benefits of physically compact data, the ability to speedily store and retrieve large volumes of data, the ability to manipulate and integrate different types of data,

and the ability to perform complex spatial and non-spatial analyses (Coulson *et al.*, 1991). Spatial data usually refer to the geographical location of the entity being studied, while non-spatial data consist of the various attributes or properties of such an entity (Coulson *et al.*, 1991). GIS allows the spatial data to be plotted on a map, with the various attributes of the non-spatial data forming overlays in relation to the spatial distribution of the entity. However, GIS is, not without a number of disadvantages. It is expensive to implement, involves the conversion of data into digital form, and involves a significant overhead cost in terms of maintenance of the system as well as operation by personnel who need a high degree of expertise (Coulson *et al.*, 1991).

In the field of medical geography, GIS has been used in cross-sectional, case-control and cohort studies, contributing much to our knowledge of all types of disease. Guthe's study predicted the populations of children which were at highest risk of lead exposure in New Jersey, USA (Vine *et al.*, 1997). Wartenberg's group investigated populations with potential for high exposure to magnetic fields (Vine *et al.*, 1997). Fifty three environmental variables were investigated by Glass's group, in a study using GIS and case-control methods to determine residential environmental risk factors for Lyme disease, which is a tick-borne disease caused by the bacterium *Borrelia burgdorferi* (Vine *et al.*, 1997). These studies demonstrated the importance of using databases that are as accurate, complete and comparable as possible. Furthermore, it has become apparent that collaboration and communication among researchers in a variety of fields, including epidemiology, medical geography, environmental science and biostatistics, are necessary in order to realise the full potential of GIS technology in epidemiological studies (Vine *et al.*, 1997).

Increasing research is being done in the field of medical geography into the geographical distribution of infectious disease, with particular application to AIDS. The major transmission routes and the geography of AIDS in different countries have been mapped (Gould, 1989; Shannon *et al.*, 1991). However, limited data is available as to the precise geographical distribution of strains of *M. tuberculosis,* particularly in South Africa. The distribution of the disease in a small, high incidence community in the Western Cape has been described using GIS (Beyers *et al.*, 1996). The distribution and number of tuberculosis cases were plotted by GIS according to dwelling unit. This provided a graphic demonstration of the impact of the disease on the community, highlighting the fact that it had the greatest effect in areas that were characterised by overcrowding and poor socio-economic conditions. An accurate database of the geographical distribution of tuberculosis in this community was established, showing that a higher than usual incidence existed in certain districts. It was even possible to identify specific dwelling units in which tuberculosis occurred repeatedly.

Such information would potentially be of great value in focussing the attention of health services on such high risk areas. Such a database could be used by health authorities in conjunction with other epidemiological and demographic tools for the monitoring of treatment and control programmes in communities. However, the Beyers study did not attempt to characterise the genetic makeup of strains in any way, which would have provided valuable additional data. Furthermore, GIS technology had not previously been applied to a genetic study of the organism in the South Africa.

1.4 Research hypothesis

From the above, it can be seen that much valuable information on certain aspects of the genetic makeup of *M. tuberculosis* has been provided by a number of key molecular epidemiological studies performed over the past decade. It has been demonstrated that it is possible to establish the proportion of cases in a community due to recent infection as opposed to those due to reactivation of endogenous infection (van Soolingen *et al.*, 1991; Hermans *et al.*, 1995; Warren *et al.*, 1996). The existence of an inverse correlation between disease incidence and degree of genetic diversity has been demonstrated by some (Strässle, 1997; Safi, 1997), while other studies have shown that the mobility and heterogeneity of human populations may have a significant effect on genetic diversity (Chevrel-Dellagi *et al.*, 1993; Harn *et al.*, 1997). Molecular markers have been of value in distinguishing between cases of acquired, as opposed to primary, drug resistance in a community (Small and Moss, 1993). The demonstration of a high proportion of acquired resistance would alert health authorities to the existence of a high degree of patient non-compliance in the community. Knowledge of the various aspects of tuberculosis epidemiology referred to above may assist in the establishment of effective treatment and control programmes.

The resurgence of tuberculosis in South Africa has made it imperative to address vital unanswered questions pertaining to the genetic diversity and population structure of the organism in areas of South Africa in the throes of a growing tuberculosis epidemic. Of importance is to establish whether strains of *M. tuberculosis* occurring in urban areas differ genetically in any significant way from those in rural areas, as the existence of urban strains that are very different to those circulating in rural areas might represent a significantly more difficult disease control scenario. Also of interest is the question of whether there is any significant degree of genetic differentiation between drug resistant as opposed to drug sensitive strains. Allied to this is the importance of demonstrating whether there is a correlation between high genetic diversity and drug resistance. The type of drug

resistance existing in a community also needs to be established. A predominance of primary resistance might be indicative of the existence of a high incidence of AIDS in a particular community. Conversely, a large amount of acquired resistance would indicate the need to deal with unacceptably high levels of patient non-compliance. Furthermore, information might be provided on the evolution of the organism in a particular community. A comparison of the results of this study with those generated by other national and international studies would also be of value in gaining a broader understanding of the relationship between strains in South Africa and in other countries.

Questions relating to the geographical distribution of this organism have not yet been addressed on a large scale in South Africa. GIS technology would be suited to the swift analysis of large numbers of strains within a particular geographical region. Correlation between genetic type and specific geographical location needs to be investigated in the hope of shedding light on the nature of the disease process in a certain region. Of interest would be to establish whether the spread of the disease in a particular geographical centre is primarily due to a small number of strains or to strains that are widely spread throughout the region. The correlation of specific genetic types with particular routes of travel might provide information related to the modes of transmission of strains in a region. Geographical foci containing strains belonging to large numbers of different genetic types should be identified as these might represent areas that present a difficult disease control scenario. The geographical distribution of drug resistant strains would be of particular interest, with attention being given to any areas with a predominance of resistant as opposed to sensitive strains. Knowledge of the geographical distribution of strains of this organism will contribute significantly to the development of effective treatment and control programmes (Beyers *et al.*, 1996).

The Eastern Cape Province has been identified as an area with a rising incidence of tuberculosis and HIV infection, where serious undernotification of both diseases has made it difficult to obtain an accurate picture of the dual epidemic that is ravaging one of the more economically depressed areas of South Africa. No information on the genetic capability and diversity of *M. tuberculosis* in this area is available, as no previous molecular epidemiological studies have been undertaken. Thus, a large scale study encompassing the sampling of towns spanning a broad area of this province would provide baseline knowledge of the genetic capability of this organism on which future studies could build.

While the IS*6110*-RFLP technique has been used for a number of such studies in other parts of South Africa and elsewhere in the world, it is more ideally suited to smaller scale applications due to its technical complexity and length. Furthermore, it is unable to survey as large a portion of the organism's genome as is possible using the RAPD technique in conjunction with multiple primers. Methods based on DNA sequencing were considered to be unsuitable due to their complexity and high cost factors. The studies by Palittapongarnpim *et al.* (1993b), Linton *et al.* (1994 and 1995) and Harn *et al.* (1997) have shown that the RAPD-PCR technique provides a technically simple, rapid and cost-effective way of measuring genetic diversity in a mycobacterial population, with a relatively large amount of genetic information being generated. Furthermore, the technique is suitable for generating DNA markers from large numbers of strains due to its relative rapidity. This molecular marker technique was thus deemed suitable for use in this large scale study aimed at answering the important questions posed above.

1.5 Research objectives

This study formed part of the South African arm of Glaxo Wellcome's *Action TB* research initiative. The following research objectives were established:

1. The sampling of towns located in Health Regions A, B and C of the Eastern Cape, comprising an area of approximately 85 000 km^2.
2. The inclusion of a small sample of drug resistant isolates from KwaZulu-Natal.
3. The generation of RAPD markers from isolates from both geographical areas, using a number of arbitrary ten-mer primers.
4. The measurement of genetic diversity and population structure from the genetic data provided by these RAPD markers, using cluster analysis and analysis of molecular variance (AMOVA).
5. The analysis of the geographical distribution of strains, with the application of GIS to the analysis of the Eastern Cape Province sample.
6. The comparison of the genetic diversity and population structure of the Eastern Cape sample with that of the KwaZulu-Natal sample. The RFLP typing of this sample was also compared with the grouping of the RAPD technique.

CHAPTER 2

MATERIALS AND METHODS

The Eastern Cape and KwaZulu-Natal samples consisted of a specific number of isolates of *M. tuberculosis*, which had been cultured from clinical specimens received from patients with tuberculosis. At this point it is important to draw a distinction between "sample" and "isolate". Sample is used here in the context of the entire number of isolates of *M. tuberculosis* surveyed in this study, and when associated with a particular province pertains to the isolates received from that region only. An isolate refers to the *in vitro* culturing and identification of *M. tuberculosis* from a clinical specimen received from a patient with active tuberculosis.

The sampling procedures used in this study are described and the geographical locations from which the isolates came are identified according to town and medical facility. The development and optimisation of the RAPD technique as well as the analytical methods used are indicated.

2.1 Sampling procedures

2.1.1 The Eastern Cape Province Sample

Isolates of *M. tuberculosis* were received from the South African Institute for Medical Research (SAIMR) laboratory in Port Elizabeth, which provides a tuberculosis diagnostic service to most of the province, excluding the Transkei area. Identification of the organism was carried out by means of the radiometric *p*-nitro-α-acetyleamino-β-hydroxy-propiophenone (NAP) test in BACTEC 12B medium (Becton Dickenson, UK). Each isolate used in this study was subcultured into BACTEC 12B medium to a growth index approximating to a concentration of 1×10^6 colony-forming units (CFU). A total of 50 cultures a month were received from this laboratory over a one year period, beginning February 1996 and ending January 1997. Seventeen percent of all the isolates received were resistant to one or more of seven antituberculosis antibiotics used regularly in the Eastern Cape, these being streptomycin, rifampicin, pyrazinamide, isoniazid (INH), ethambutol, thiacetazone and ethionamide. Only one isolate per patient was included in the sample.

2.1.2 KwaZulu-Natal Province Sample

A small number of isolates were received from the Department of Medical Microbiology at the University of Natal in Durban, where they had already been typed using the IS*6110*-RFLP method. Once again, only one isolate per patient was included in the sample. The same identification procedure was used as for the Eastern Cape sample. Fifty seven of the isolates received were successfully amplified with two of the four ten-mer primers used for the Eastern Cape population. The profiles obtained were compared with those obtained from the Eastern Cape population using the same two primers. All of the KwaZulu-Natal isolates were drug resistant.

2.2 Geographical location of isolates

2.2.1 The Eastern Cape Province

The majority of clinical specimens received by the SAIMR laboratory for isolation of *M. tuberculosis* are from the greater Port Elizabeth area and the surrounding rural areas, as well as from Uitenhage and Despatch, two urban centres in the close vicinity. It was thus unavoidable that a large number of isolates from these areas were included in the sample. On the other hand, not as many clinical specimens are received from rural clinics in the interior regions, resulting in smaller numbers of isolates from these areas being included in the sample. Thus, it was difficult to meet the sampling requirements set out for the measurement of genetic diversity within a given population, which calls for random sampling that should not be biased in any way and that is truly representative of the population (Ayala, 1982). However, the ability of the RAPD technique to generate a relatively large amount of genetic information would enable a meaningful measurement of the genetic diversity of the rural strains.

Isolates of *M. tuberculosis* were obtained from patients who had attended 110 medical facilities in 39 towns spread over an area stretching from Misgund in the south-west to Elliot in the north-east of the province (Table 2.1). Six of the towns were classified as urban centres and the remaining 33 as rural centres (Figure 2.1). The existence of a substantial degree of industrial development was used as the criterion for classifying a town as "urban" as opposed to "rural". Grahamstown was particularly difficult to classify, as it has relatively little industry compared to Port Elizabeth and is surrounded by a large farming community. However, due to the presence of a certain amount of

industrial development, it was decided to include it in the urban category. In the early stages of the project, it was expected that the urban-rural divide would be of considerable significance. The towns sampled fell into Health Regions A, B and C of this province as can be seen from Figure 2.2.

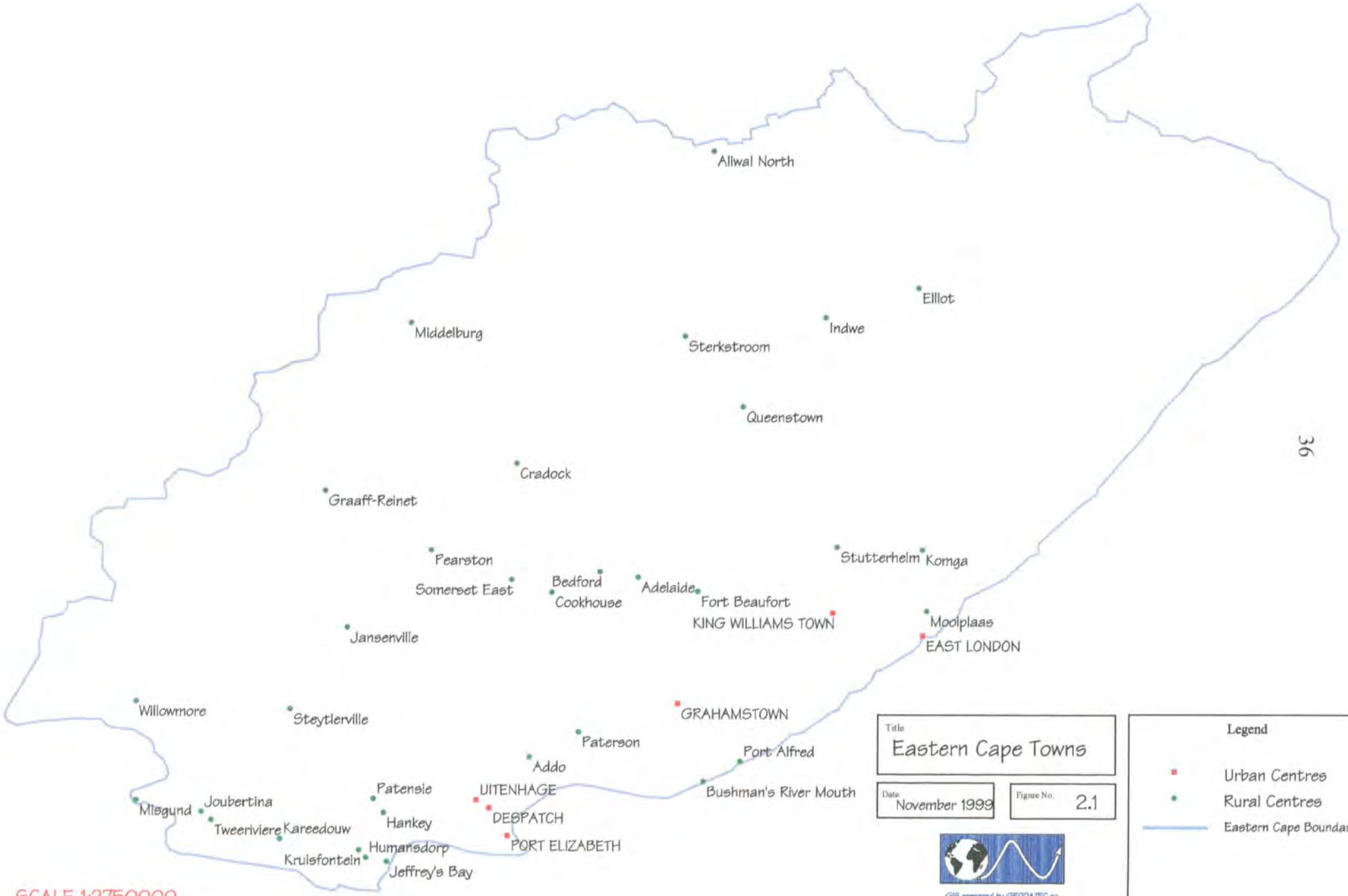
Table 2.1 Medical facilities sampled in the Eastern Cape

Health Region	Town (n=39)	Urban/ Rural	Number of Hospitals (n=20)	Number of Clinics (n=90)
A	Addo	R	0	2
	Cookhouse	R	0	1
	Despatch	U	0	2
	Graaff-Reinet	R	2	1
	Hankey	R	0	1
	Humansdorp	R	2	3
	Jansenville	R	1	0
	Jeffrey's Bay	R	0	1
	Joubertina	R	1	1
	Kareedouw	R	1	1
	Kruisfontein	R	0	1
	Misgund	R	0	1
	Patensie	R	0	2
	Pearston	R	0	1
	Port Elizabeth	U	3	25
	Somerset East	R	0	3
	Steytlerville	R	0	1
	Tweeriviere	R	0	1
	Uitenhage	U	2	7
	Willowmore	R	0	1
B	Aliwal North	R	0	1
	Cradock	R	1	3

Table 2.1 (continued)

Health Region	Town (n=39)	Urban/ Rural	Number of Hospitals (n=20)	Number of Clinics (n=90)
B	Elliot	R	0	1
	Indwe	R	0	1
	Middelburg	R	1	0
	Queenstown	R	0	3
	Sterkstroom	R	0	2
C	Adelaide	R	0	1
	Bedford	R	0	2
	Bushman's River Mouth	R	0	1
	East London	U	2	6
	Fort Beaufort	R	1	2
	Grahamstown	U	1	2
	King William's Town	U	1	0
	Komga	R	0	3
	Mooiplaas	R	0	2
	Paterson	R	0	1
	Port Alfred	R	1	1
	Stutterheim	R	0	2

No personal information was obtained on the patients from which the isolates had been received, as stipulated by the Research Ethics Committee of Rhodes University. However, the name of the medical facility which the patient had attended at the time of the taking of the clinical specimen (usually sputum) from which the organism had subsequently been isolated, was known, as well as its geographical location. It was assumed that the vast majority of patients would attend clinics or be hospitalised in fairly close proximity to their area of residence. However, a substantial proportion of the work force in the urban areas of this province hail from rural areas with which they maintain strong links. Therefore, it is to be expected that some patients might have attended a rural medical facility far removed from the urban areas in which they originally contracted the disease. The migratory nature of large numbers of people in the Eastern Cape was subsequently found to have considerable bearing on the genetic diversity of *M. tuberculosis*.


Aliwal North
Elliot
Middelburg
Indwe
Sterkstroom
Queenstown
Cradock
Graaff-Reinet
Pearston
Stutterheim
Komga
Somerset East
Bedford
Adelaide
Cookhouse
Fort Beaufort
KING WILLIAMS TOWN
Mooiplaas
Jansenville
EAST LONDON
Willowmore
Steytlerville
GRAHAMSTOWN
Paterson
Port Alfred
Addo
Patensie
UITENHAGE
Bushman's River Mouth
Joubertina
Misgund
DESPATCH
Hankey
Tweeriviere
Kareedouw
PORT ELIZABETH
Humansdorp
Kruisfontein
Jeffrey's Bay
SCALE 1:2750000
0 20 40 60 80 100 kilometers
Title
Eastern Cape Towns
Date
November 1999
Figure No.
2.1
GIS prepared by GEODATEC cc
Legend
Urban Centres
Rural Centres
Eastern Cape Boundary

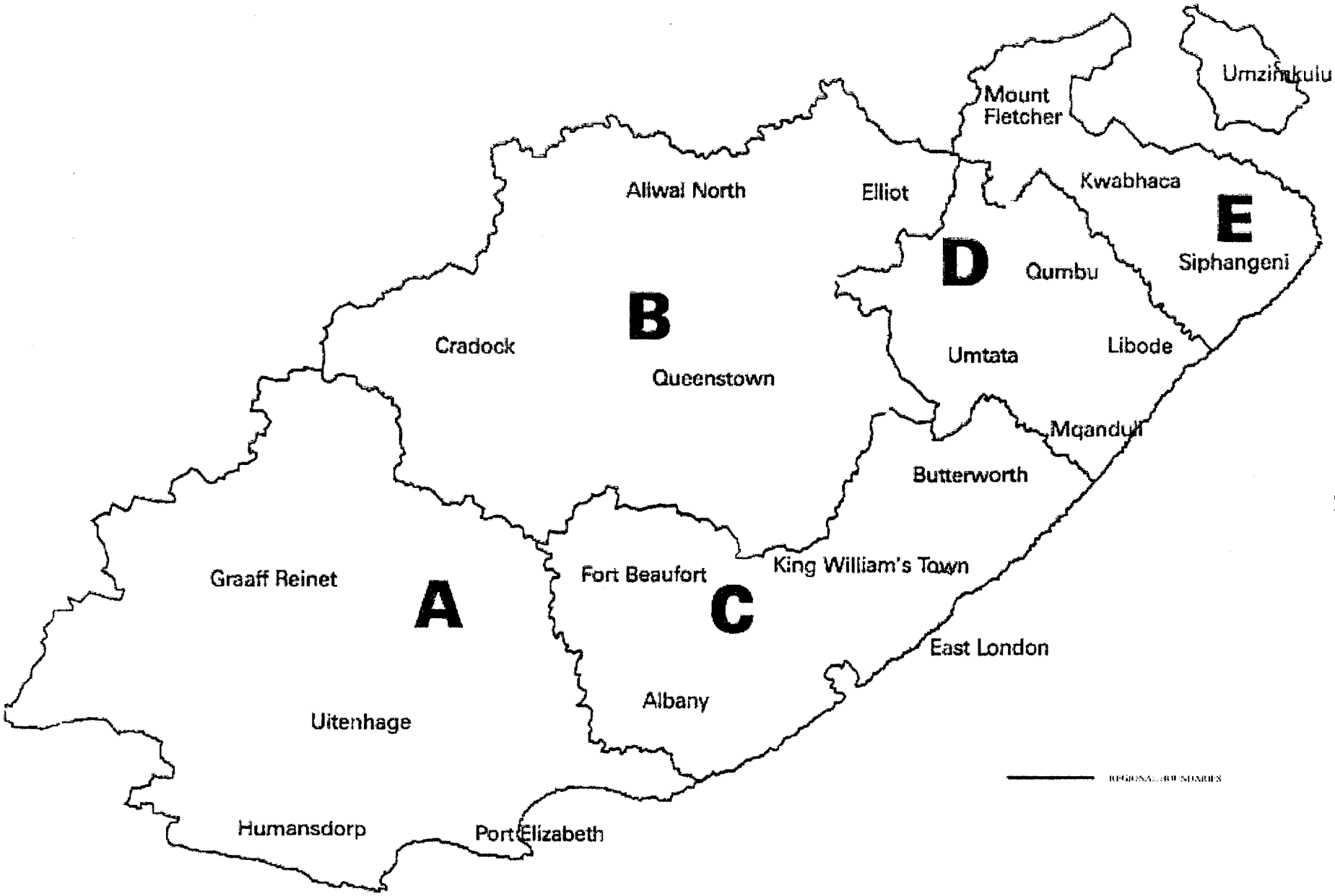


Figure 2.2 Eastern Cape Health Regions

2.2.2 KwaZulu-Natal province

The 57 isolates in this population were obtained from patients attending thirteen medical facilities, ten of which were in the greater Durban area. The other three were from hospitals in Stanger and Hibberdene, and from a clinic in the Mahlabatini area of Zululand, all of which would be classified as rural areas in terms of the definition used in this study. Table 2.2 shows the distribution of hospitals and clinics surveyed in this province.

Table 2.2 Medical facilities sampled in KwaZulu-Natal

Town	Urban/Rural	Number of hospitals (n=5)	Number of clinics (n=8)
Durban	U	3	7
Hibberdene	R	1	0
Stanger	R	1	0
Mahlabatini	R	0	1

Once again, isolates were linked to the geographical location of the medical facility as the home address given is not always a reliable indicator of where the patient spends most of the time, and because of inability to utilise patient data.

2.3 Development and optimisation of RAPD protocols

The development and optimisation of the various protocols for the generation of reproducible and reliable RAPD markers from each isolate was essential. The process commenced with the inactivation of live cultures, was followed by the search for an effective DNA extraction protocol and culminated with the optimisation of the RAPD-PCR parameters.

2.3.1 Inactivation of live cultures

M. tuberculosis live cultures need to be handled under stringently protective conditions in a laboratory setting. Koornhof *et al.* (1995) recommend working in a Class I or II biosafety cabinet. While the facility in which this study was performed did have a Class II biosafety cabinet, it was not possible to situate it in an area of limited access as is also required. For this reason, and because the molecular biology protocols did not call for live organisms, all cultures were heat treated to kill the

organism before any further manipulations were carried out. DNA extraction procedures were then performed in the biosafety cabinet. Heat killing was initially performed at a temperature of 95°C for thirty minutes, with subsequent subculturing confirming the non-viability of the organism. However, in the early stages of the project it was thought that the low yields of DNA obtained were due in part to loss of DNA from bacterial cells ruptured by exposure to such a high temperature. The heat inactivation protocol was thus modified, with exposure of cultures to 80°C for 30 minutes being sufficient to inactivate the organism.

2.3.2 DNA extraction

Extraction of mycobacterial DNA depends on successful lysing of a cell wall that contains a high proportion of lipids known as mycolic acids (Clemens, 1997). These give the organism its characteristic acid-fast staining properties and also result in a cell wall of remarkably low permeability. This makes it possible for the organism to survive under harsh environmental conditions and in hostile host environments (Clemens, 1997). A variety of different DNA extraction protocols have been used with varying degrees of success.

Conventional lysis methods which make use of an enzyme such as lysozyme, and detergents such as sodium dodecyl sulphate (SDS), Tween 20, Nonidet P-40 and Triton X-100 have been used with a measure of success (Hermans *et al.*, 1990; Buck *et al.*, 1992; Folgueira *et al.*, 1993; Haas *et al.*, 1993; Plikaytis *et al.*, 1993; van Soolingen *et al.*, 1994). The amino acid analog, D-cycloserine, has been particularly successful in rendering mycobacteria more susceptible to lysis (Ross and Dwyer, 1993). Mechanical methods have also been used, with a number of protocols utilising glass beads in conjunction either with vortexing (Palittapongarnpim *et al.*, 1993b) or sonicating (Folgueira *et al.*, 1993). Alternatively, high temperatures (particularly boiling) have been used with glass beads (Folgueira *et al.*, 1993) or enzyme treatment (Buck *et al.*, 1992). Extremes of temperature have also been used, with repeated freeze-thaw cycles being followed by boiling (Buck *et al.*, 1992). Yet another protocol utilised the addition of a relatively large volume of water to the bacterial cell suspension to generate osmotic shock, with subsequent bursting of the cell wall (Bose *et al.*, 1993).

Once the cell wall has been lysed, conventional protocols usually involve the removal of cellular debris and bacterial protein, using a nonspecific protease, such as proteinase K, in conjunction with phenol, chloroform and isoamyl alcohol (Hermans *et al.*, 1990). Purified DNA is then precipitated

using ethanol and dissolved in a Tris hydrochloride-EDTA buffer (Hermans *et al.*, 1990). A widely used modification replaces phenol with *N*-cetyl-*N*,*N*,*N*,-trimethyl-ammonium bromide (CTAB), which is able to remove bacterial polysaccharides as well as proteins (Ausubel *et al.*, 1989; van Soolingen *et al.*, 1994). This protocol is much favoured, as phenol is unpleasant to work with and requires extensive purification before it is suitable for molecular biological application. The modification retains the use of chloroform and isoamyl alcohol as a second DNA purification step, followed by an isopropanol wash before the final ethanol precipitation procedure (van Soolingen *et al.*, 1994). Bose *et al.* (1993) developed a DNA purification process using a "GIT cocktail" consisting of guanidium thiocyanate, Tris hydrochloride-EDTA, sodium chloride (NaCl), *N*-lauryl sarcosine and mercaptoethanol in addition to CTAB as a first protein removal step, followed by the usual phenol-chloroform extraction procedure. DNA was then precipitated with ammonium acetate and isopropanol. This protocol is not very satisfactory, however, as it involves the use of a number of toxic and unpleasant chemicals, in addition to phenol.

Commercial DNA extraction kits have been used with varying degrees of success. The Nucleon extraction kit (Scotlab Ltd.) (Linton *et al.*, 1994) and the Bandprep kit (Pharmacia Ltd.) (Linton *et al.*, 1994) both provide purified DNA. Commercial products such as GeneReleaser (Cambio) (Linton *et al.*, 1994) and InstaGene (Bio-Rad) (Richner *et al.*, 1997) extract DNA and bind PCR inhibitors released during the process, thereby providing what is known as PCR-ready DNA in a much shorter time period than the conventional protocols. These latter products have been shown to work well with RAPD-PCR (Linton *et al.*, 1994, Richner *et al.*, 1997).

In other attempts to shorten the extraction protocol and to avoid the use of unpleasant chemical compounds, protocols have been developed which concentrate mainly on the lysis of cells and the release of DNA, using either enzymatic or mechanical means. Such crude lysates have, however, given mixed results. Welsh and McClelland (1990) were able to generate equivalent data using either purified DNA or crude lysate. Linton *et al.* (1994), however, have shown crude lysate to be less than satisfactory for use with the RAPD technique.

The protocol of van Soolingen *et al.* (1994), with one or two minor modifications, was chosen for use in the early stages of this study, as it was felt that purified DNA was needed for the generation of reliable and reproducible RAPD markers. DNA quantification involved agarose gel electrophoresis of DNA samples along with a molecular weight marker, which separated out into

bands of known DNA concentration (Sambrook *et al.*, 1989). Ethidium bromide was used for visualisation as it is sensitive enough to detect as little as 1- 5 ng of DNA (Sambrook *et al.*, 1989). Unfortunately, low concentrations of DNA were detected in a large percentage of the first batch of mycobacterial cultures, while a number of cultures yielded no DNA at all. This was thought to be the result of a number of unfavourable conditions pertaining to these cultures. Firstly, they had been held at 4°C for one to two months before being heat inactivated and extracted. Coutinho *et al.* (1993) showed that culture age affects the amenability of bacterial cells to lysis. This was remedied by ensuring that bacterial cultures were held at 4°C for no longer than two to three weeks prior to extraction. Secondly, the presence of two prophages in the genome of *M. tuberculosis* may result in a persistent low-level of lysis in culture (Cole *et al.*, 1998). Another contributing factor might have been the activity of thermostable nucleases, which are capable of withstanding high temperatures (Gibson and McKee, 1993). Thus, cultures that had been heat inactivated should have been extracted immediately and not held at 4°C for a couple of days, which was occasionally the case. Also, after extraction, the DNA should have been stored at -20°C and not at 4°C. In the first instance, it was possible to make sure that cultures were immediately extracted after being heat inactivated. However, the matter of the storage temperature was a cause for concern, as sample DNA would be amplified with at least four primers. The logistics of RAPD optimisation and primer screening, as well as the large number of isolates to be amplified, called for repeated use of sample DNA, making it undesirable to expose the sample to repeated freezing and thawing.

After having remedied the above-mentioned variables, low concentrations of DNA were still detected and in only 30% of the second batch of cultures. It thus seemed that loss of DNA was occurring during the purification stage of this protocol. It has been shown that CTAB can precipitate a certain amount of DNA if the NaCl concentration is less than 0.5M on addition of the CTAB buffer (Ausubel *et al.*, 1989). It thus became necessary to investigate other DNA extraction protocols in order to find one which was optimal for this organism.

As mentioned earlier, crude lysates may be used for PCR, utilising either enzymatic or mechanical means to prepare them. Two such protocols were developed, based on the first comprising steps 1 to 5 of the above CTAB protocol, after which the crude lysate was centrifuged to pellet cell debris. The second protocol used lysozyme and proteinase K as per the CTAB protocol, but utilised the detergent Sarkosyl instead of SDS. The amount of DNA produced by these two methods was found to be no greater than that obtained with the CTAB protocol and there was evidence of contaminating protein. It was apparent that the traditional enzyme/detergent methods were not very

satisfactory in lysing mycobacterial cells, while centrifugation was not sufficient to remove all contaminating protein. The use of heat was then investigated in an effort to improve cell lysis. A number of methods have been described using high temperatures to lyse bacterial cells and produce a crude lysate (Welsh and McClelland, 1990; Mazurier *et al.*, 1992a). However, as contaminating protein had been shown to be a problem, it was decided that the crude lysate should undergo a number of purification steps. The bacterial cell solution was thus boiled for five minutes, followed by centrifugation at 12 000g for two minutes, after which steps 9 to 14 of the CTAB protocol were followed. The CTAB step was omitted in an attempt to retain as much DNA as possible. Agarose gel electrophoresis showed that DNA extracted in this way was not sheared, as can result from boiling DNA. However, the quantity of DNA was not markedly higher in comparison to that produced by the other two protocols, and there was still evidence of contaminating protein.

After the failure of the CTAB protocol, the two crude lysate protocols and the modified boiling protocol to extract sufficiently high concentrations of mycobacterial DNA, InstaGene Matrix (Bio-Rad, UK), which has been applied successfully to the extraction of DNA from whole blood, cultured cells and bacteria, was tested. The InstaGene extraction protocol involved four steps which resulted in the lysis of bacterial cells and the absorption of cell lysis products that can interfere with the PCR amplification process (Appendix A.3). DNA could be stored at -20°C, and thawed and refrozen for further use. Storing at low temperatures was necessary to inhibit the activity of thermostable nucleases which may be present in the DNA preparation. The shearing of DNA that might result from such multiple freeze-thaw cycles was balanced by the small amount of pipetting involved during the extraction process which minimises the shearing that usually occurs in the early stages of DNA manipulation (Bio-Rad, UK, personal communication). The short periods of vortexing likewise would not produce undue shearing. RAPD-PCR is able to amplify smaller fragments of DNA than are other fingerprinting techniques such as RFLP, making it more tolerant to protocols which may result in a degree of shearing of DNA molecules. No quantification was needed with this protocol, provided that the starting point was kept constant as regards the number of bacterial cells used. A starting concentration of 1×10^6 bacterial cells was used throughout, as recommended by Bio-Rad (personal communication).

In order to determine whether this protocol had been successful in extracting mycobacterial DNA of sufficient quantity and quality, RAPD-PCR was performed using parameters previously optimised in our laboratory for *M. tuberculosis* (Krallis, 1991). The DNA extracts prepared by the

CTAB, the crude lysate and the boiling protocols were also amplified using these conditions. The results were unexpected, with an amplification rate of 100% being demonstrated with the InstaGene protocol, as opposed to only 30% with the other protocols. It was thus decided that the InstaGene Matrix (Bio-Rad, UK) protocol would be used to extract DNA from all isolates in this study.

2.3.3 Optimisation of RAPD-PCR parameters

The optimisation and standardisation of the RAPD technique are extremely important for the production of reproducible and reliable data (Ellsworth *et al.*, 1993; Linton *et al.*, 1994; Harn *et al.*, 1997). Variation in the concentration of the components in the PCR mixture can result in the production of different RAPD markers from the same DNA preparation. Welsh and McClelland (1990) found that magnesium chloride ($MgCl_2$) concentration, primer annealing temperature, template DNA concentration, primer length and primer sequence all affected the number, reproducibility and intensity of bands. Weeden *et al.* (1992) showed that primer concentrations in particular can affect the markers produced. However, unlike Welsh and McClelland's findings, Weeden demonstrated that $MgCl_2$, deoxynucleotide triphosphate (dNTP) and buffer concentrations had no effect on RAPD markers. Park and Kohel (1994) showed that variations in $MgCl_2$ concentration resulted in both quantitative differences in RAPD markers as well as qualitative changes in the band patterns. Variation can also arise from using different thermocyclers and different PCR cycle conditions (Weeden *et al.*, 1992). Schweder *et al.* (1995) showed the importance, on the quantity and size of RAPD markers, of the transition interval between the denaturing and annealing temperatures of the PCR cycle. Schierwater and Ender (1993) found that different DNA polymerases amplified different RAPD products.

The reproducibility of RAPD profiles under the optimised conditions must also be demonstrated. This is easily achieved when small numbers of samples are involved. It is maintained by some that the RAPD technique is best suited to small collections of isolates which can be analysed, each in triplicate, in a single batch (Linton *et al.*, 1994). Schweder *et al.* (1995), on the other hand, demonstrated that the RAPD technique is amenable to the analysis of large numbers of samples. In addition to the problem of reproducibility, great care must be taken to eliminate the effect of contaminating DNA. However, it has been shown that such DNA, because it is usually present in much smaller quantities than the sample DNA, may not be able to compete against the latter for primer and will thus not always amplify sufficiently to be visible as a band in the RAPD profile

(Newbury and Ford-Lloyd, 1993). In spite of this, it is essential that a PCR reaction containing no sample DNA be included with each PCR batch. The appearance of bands in the blank would invalidate the results of such a batch.

The RAPD parameters of Krallis (1991), referred to in Section 2.3.2.4, were used as the starting point for the optimisation of PCR conditions for this study, with varying concentrations of each component in the PCR reaction being tested separately. Magnesium chloride was optimised at 2.0 mM after concentrations of 1.0 and 1.5 mM produced too many bands, and those of 2.5 and 3.5 mM generated too few. The RAPD profiles obtained with concentrations of 100, 200 and 400 µM of each dNTP did not differ greatly, confirming the finding by Weeden *et al.* (1992) of the negligible effect of dNTPs on RAPD markers. They were thus used in this study at a concentration of 100 µM of each dNTP. Arbitrary primers have been optimised at a wide range of concentrations. Hilton *et al.* (1997) used as much as 100 µM of primer per PCR reaction, and Schierwater and Ender (1993) as little as 3 pM. Most studies have used a concentration of 0.2 µM (Williams *et al.*, 1990). Weeden *et al.*(1992) found that most primers have a narrow working range of about 4-fold and were able to get reproducible results with concentrations ranging from 0.2 to 1.0 µM. The starting point for optimisation in this study was 0.2 µM of primer and titrations were performed to ascertain whether it would be possible to use lower concentrations. However, no bands were obtained with 1 nM and 100 pM, and only a few bands were obtained with some primers at concentrations of 0.1 and 0.01 µM. Primers were thus used at a concentration of 0.2 µM in this study .

Schierwater and Ender (1993) showed that different thermostable DNA polymerases amplify different RAPD products, with the differences being both qualitative and quantitative. The enzymes also differed significantly as regards their sensitivity to changes in $MgCl_2$ concentration, due to the fact that the activity and specificity of different polymerases depends on slightly different temperature and reaction preferences. This in turn affects the products amplified in the first cycle, which is the most critical one. Weeden *et al.* (1992), however, found that commercial polymerases derived from two different *Thermus* species gave identical results in their hands. Widely differing concentrations of *Taq*, varying from as little as 0.06 U/µl to as much as 2.5 U/µl, have been used successfully in RAPD-PCR (Sandery *et al.*, 1994; Ramser *et al.*, 1997). Optimisation experiments for this study compared results obtained with 1U/µl (Figure 2.3a) and with 0.5U/µl of *Taq* (Figure 2.3b). As a lower concentration of *Taq* resulted in a reduction in the number of RAPD markers, I U/µl of *Taq* was used in this study.

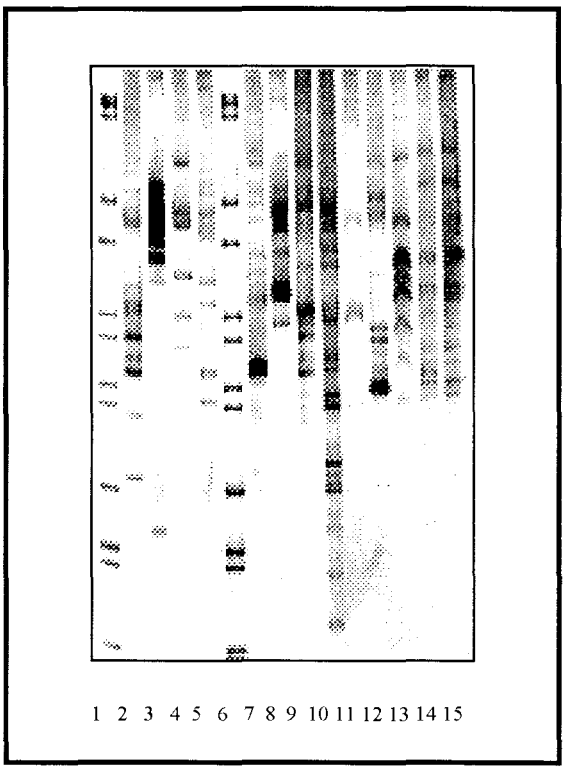


Figure 2.3a Optimisation of *Taq* concentration
Lanes 1 & 6, molecular weight marker; lanes 2 - 5 & 7 - 15, *M. tuberculosis* DNA amplified with 1 unit/µl *Taq*

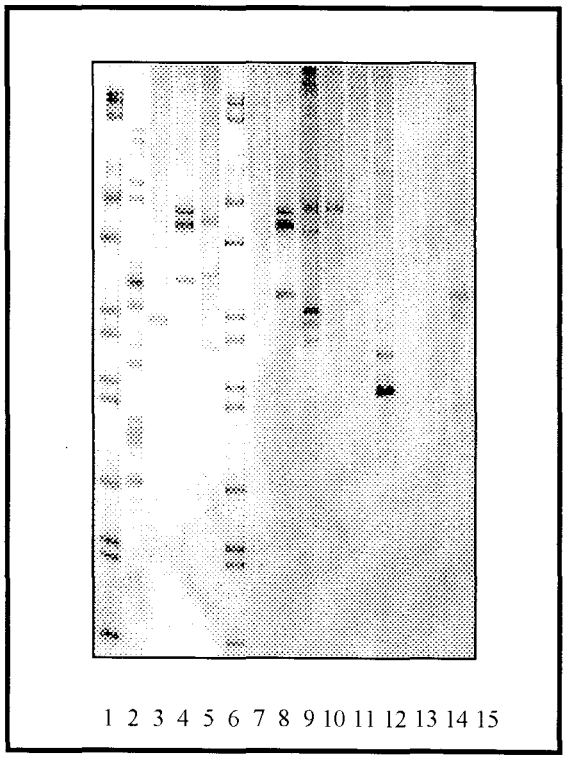


Figure 2.3b Optimisation of *Taq* concentration
Lanes 1 & 6, molecular weight marker; lanes 2 - 5 & 7 - 15, *M. tuberculosis* DNA amplified with 0.5 unit/µl *Taq*

As regards the thermal cycling programme, the denaturing, annealing and extension times, the annealing temperature and the number of cycles need to be optimised. The annealing temperature is particularly important and may vary from primer to primer, as may the annealing time. Annealing temperatures have been optimised from as low as 30°C (Caetano-Anollés *et al.*, 1992) to as high as 60°C (Welsh and McClelland, 1990). However, Welsh and McClelland commenced with 2 low stringency cycles of 40°C, having found that no bands were produced when a high stringency temperature of 60°C was maintained throughout. Weeden *et al.* (1992) found that minor changes in cycle parameters did not significantly alter the markers amplified. They showed that annealing temperatures of 35 to 42°C for certain primer-template combinations gave exactly the same RAPD profiles. Ellsworth *et al.* (1993), on the other hand, showed that RAPD profiles did change when the annealing temperatures were varied across a range from 35°C to 50°C. Annealing times may be as short as 1 second (Caetano-Anollés *et al.*, 1992) and as long as five minutes (Wang *et al.*, 1993), with an average of between 30 and 60 seconds being most common. Yu and Pauls (1992) demonstrated an association between the annealing time and the GC content of the primer. In their hands, primers with a GC content of 70 and 80% gave numerous RAPD markers at an annealing time of five seconds, whereas considerably fewer markers were produced with primers having a GC content of only 50 and 60%.

The denaturing time was usually of the same order as the annealing time, with the extension time often being twice as long. Most programmes start and end with one longer denaturing and extension cycle, respectively. Yu and Pauls (1992) have shown that denaturing times as short as five seconds may yield better RAPD markers, which can be attributed to the fact that *Taq* has a limited lifespan when exposed to high temperatures for longer than a few minutes. As regards the effect of the extension time on RAPD markers, Yu and Pauls (1992) found that shorter times (5 to 30 seconds) resulted in the amplification of products smaller than 1.5 kb, while extension times of 60 seconds were needed to amplify the larger products. Schweder *et al.* (1995) showed the influence of the transition interval between the denaturing and annealing temperatures on the quantity and size of RAPD markers. Experiments with five different transition intervals indicated that a longer transition interval resulted in more stable and numerous RAPD markers.

The number of cycles used ranges from 30 to 45. Yu and Pauls (1992) found no significant differences among RAPD markers obtained with programmes run for 35, 55 or 75 cycles. This might be due to the inactivation of *Taq* over time, or to the limiting activity of other PCR

components at high cycle numbers. Earlier work in our laboratory had optimised the PCR cycles for this organism as follows: 1 cycle of 94°C for 3 minutes; 30 cycles of 94°C for 30 seconds, 36°C for 30 seconds and 72°C for 60 seconds; 1 cycle of 72°C for 4 minutes (Krallis, 1991). The annealing temperature and time were optimised for this study, while the denaturing and extension temperatures and times remained unchanged. Two ten-mer primers previously used were tested at annealing temperatures of 36°C, 38°C and 40°C (Krallis, 1991). The RAPD profiles and number of bands obtained were slightly different for each temperature. The least amount of profile difference was seen with annealing temperatures of 36°C and 38°C, while fewer bands were generated at 40°C. The lower temperatures were thus used as starting points for future optimisation of primers (Section 2.3.4). The number of thermal cycles was also optimised and it was found that 40 cycles produced more bands than 35 cycles, with improved band intensity. Unfortunately, using a greater number of cycles increased the chance of amplifying contaminating DNA. However, this was controlled in each PCR batch by the inclusion of a blank PCR reaction containing no DNA.

Agarose gel electrophoresis and ethidium bromide have been used for the separation and visualisation of RAPD markers in most studies (Williams *et al.*, 1990; Welsh and McClelland, 1990; Linton *et al.*, 1995). The concentration of agarose used is inversely proportional to the size of the DNA molecules being separated (Sambrook *et al.*, 1989). In the case of *M. tuberculosis*, RAPD markers may range in size from 100 to 2 000 base pairs. Such small fragments need to be electrophoresed on a gel with a relatively high concentration of agarose. Polyacrylamide gel electrophoresis (PAGE) and silver staining has also been used in RAPD and is able to resolve fragments of DNA that differ in size by as little as one base pair (Caetano-Anollés *et al.*, 1991). Agarose gel was thought to result in considerable loss of information as it only detects major fragments. As previous RAPD work in our laboratory had optimised the separation and visualisation of RAPD markers using 10% polyacrylamide gel and a silver stain protocol as set out in Appendix C (Krallis, 1991), an attempt was made to optimise the same system for this study. However, the RAPD markers generated from PCR-ready DNA were not quite as sharp, when resolved on polyacrylamide gel, as those obtained from purified DNA would have been (Figure 2.4). However, profiles obtained on agarose gel proved to be far superior, showing little sign of contaminating protein, as can be seen from Figure 2.5. Furthermore, while fewer bands were obtained on agarose, any faint artifactual bands would be excluded. For these two vital reasons, the agarose-ethidium bromide system was used in this study.

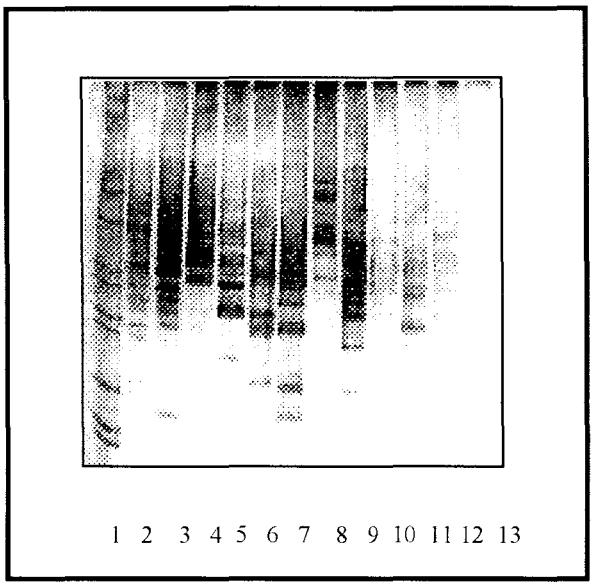


Figure 2.4 PAGE of amplified *M.tuberculosis* DNA extracted with InstaGene Matrix

Lane 1, molecular weight marker; lanes 2 - 12, amplified *M.tuberculosis* DNA; lane 13, blank

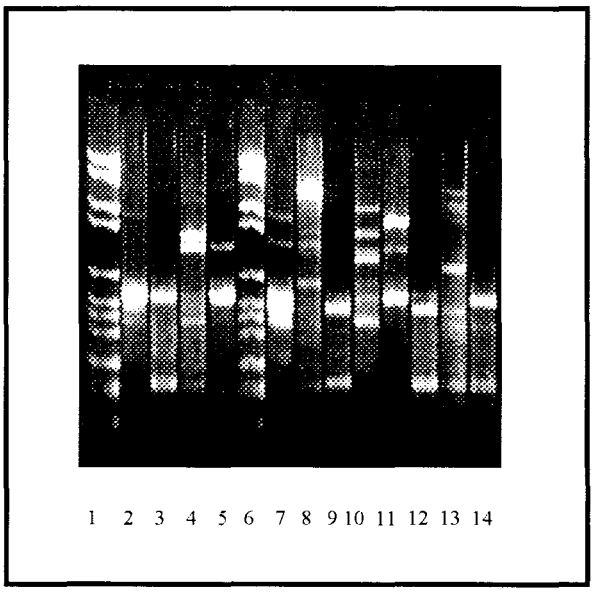


Fig 2.5 Agarose gel electrophoresis of amplified *M. tuberculosis* DNA extracted with InstaGene Matrix

Lanes 1 & 6, molecular weight marker; lanes 2 - 5 & 7 - 14, amplified *M.tuberculosis* DNA

Optimisation of agarose concentration, sample load, electrophoresis voltage and time, and band visualisation was then necessary. As concentrations of 1.6 % and 2% agarose provided similar results, a 2% concentration was chosen, making the gel slab easier to manipulate. This concentration has been used in other mycobacterial studies (Linton *et al.*, 1994). Sample load depends on the width of the well plus the depth of the gel, as well as the number and size of DNA fragments that need to be detected (Sambrook *et al.*, 1989). It was found that10 µl of a mixture of PCR product and loading buffer (Appendix C) resolved sufficient bands of a variety of sizes. Voltages of 80, 90 and 100 Volts (V) were tested, over time periods of 2, 2.5 and 3 hours. A voltage of 90V over a time period of 2.5 hours proved to be most satisfactory. Incorporation of ethidium bromide into the gel proved to be the most satisfactory method for band visualisation (Sambrook *et al.*, 1989).

Gels were photographed under UV transillumination, using Polaroid Type 667 film (ASA 3000), a Wratten 22A filter and a 135 mm lens, with exposure of one second being sufficient to obtain images of even the weakest bands.

The reproducibility of RAPD markers generated in this study was established in two ways. DNA from a single isolate was amplified in triplicate in the same PCR batch, while a more stringent method involved amplifying DNA from a single *M. tuberculosis* isolate in two separate PCR batches. Both of these methods of ensuring reproducibility were performed on randomly chosen DNA samples throughout the study. However, due to the large number of DNA samples being amplified, as well as to the use of multiple primers, it was not logistically possible to amplify every DNA sample in triplicate or to amplify every sample in two separate PCR batches. In addition, the cost factor would have been prohibitive.

Erroneous results due to contaminating DNA were excluded by including a reaction which contained no DNA in each PCR batch. Bands in the so-called "blank" reaction would invalidate that particular batch. In the early stages of the study, problems were experienced with contaminating DNA being introduced via the double-distilled water produced in the laboratory. This was resolved by using commercially prepared sterile cell culture water (Sigma, USA). The use of dedicated automatic pipettes with aerosol-resistant filter tips also contributed to excluding contamination. In addition, DNA extraction, PCR preparation and gel electrophoresis were carried out in three separate enclosures. Sterile plastic 0.2 ml thin-walled omnistrip tubes (Advanced Biotechnology,

UK) were used for PCR reactions. UV irradiation was used to keep surfaces sterile, but could not be used to sterilise any of the components of the PCR reaction, in particular the water and the mineral oil. Exposure of these to UV light resulted in the production of free radicals which have an inhibitory effect on the PCR reaction.

2.3.4 Primer Selection

Once a satisfactory extraction method had been found and the RAPD-PCR parameters optimised, a battery of arbitrary primers was screened in order to identify those suitable for generating RAPD markers from isolates in this study. The ideal arbitrary primer must generate a sufficient number of bands and must be able to amplify polymorphisms, which are a reflection of the genetic diversity within the population. The majority of primers screened were ten bases in length, as these have been used successfully in RAPD work (Williams *et al.*, 1990). However, four twenty-mer primers were also tested, as longer arbitrary primers have also been used successfully (Linton *et al.*, 1994). A total of thirty four ten-mer and four twenty-mer primers were screened.. The sequences and GC content of twenty ten-mer primers obtained from Operon Technologies (Alameda, USA) are set out in Table 2.3. The remaining fourteen ten-mers and the four twenty-mers were custom synthesised by Ransom Hill Bioscience, Inc. (Ramona, USA) (Table 2.4). The first twelve ten-mer primers listed in the table were those described by Williams *et al.* (1990), while the twenty-mer sequences (MTB1, MTB2, MTB3 and MTB4) were used by Linton *et al.* (1994).

Ultimately, four ten-mer primers, OPAI-09, OPAI-13, MBR and NTR, were chosen to amplify all 600 Eastern Cape isolates. The lowest GC content amongst these primers was 60% and the highest 80%. The inability of the first twelve ten-mer primers in Table 2.4 to achieve a sufficiently high rate of amplification of *M. tuberculosis* DNA was probably due to their low GC content, thus substantiating the findings of Yu and Pauls (1992). The four twenty-mer primers gave unsatisfactory results at a concentration of 0.2 µmol and at annealing temperatures of 36°C and 38°C. Further optimisation experiments would have been required but were not performed, as satisfactory results were obtained with four ten-mer primers.

Table 2.3 Ten-mer primers from Operon Kit A1

Primer Name	Primer Sequence (5' - 3')	GC Content (%)
OPAI-01	GGCATCGGCT	70
OPAI-02	AGCCGTTCAG	60
OPAI-03	GGGTCCAAAG	60
OPAI-04	CTATCCTGCC	60
OPAI-05	GTCGTAGCGG	70
OPAI-06	TGCCGCACTT	60
OPAI-07	ACGAGCATGG	60
OPAI-08	AAGCCCCCCA	70
OPAI-09 *	TCGCTGGTGT	60
OPAI-10	TCGGGGCATC	70
OPAI-11	ACGGCGATGA	60
OPAI-12	GACGCGAACC	70
OPAI-13	ACGCTGCGAC	70
OPAI-14	TGGTGCACTC	60
OPAI-15	GACACAGCCC	70
OPAI-16	AAGGCACGAG	60
OPAI-17	CCTCACGTCC	70
OPAI-18	TCGCGGAACC	70
OPAI-19	GGCAAAGCTG	60
OPAI-20	CCTGTTCCCT	60

* Also known as RP15 (Richner *et al.*, 1997)

Table 2.4 Ransom Hill Bioscience custom-synthesised primers

Primer Name	Primer Sequence (5' - 3')	GC Content (%)
RP1	ACGGTACACT	50
RP2	GCAAGTAGCT	50
RP3	CGGCCCCTGT	80
RP4	CACATGCTTC	50
RP5	TGGTGACTGA	50
RP6	TCGTAGCCAA	50
RP7	TCACGATGCA	50
RP8	TCTCGATGCA	50
RP9	CTGTTGCTAC	50
RP10	TGGTCACTGA	50
RP11	GCAAGTAGTG	50
RP12	ATTGCGTCCA	50
MBR	GCCGTTGCCG	80
NTR	GCCGGTGTTG	70
MTB1	CTCGTCCAGCGCCGCTTCGG	75
MTB2	GCGTAGGCGTCGGTGACAAA	60
MTB3	ACGCTCAACGCCAGAGACCA	60
MTB4	GATGAACCACCTGACATGAC	50

DNA from each isolate was amplified with each of the four selected primers, in separate PCR reactions. Thus, four separate RAPD profiles, each containing between one and twelve bands, were generated from each isolate. The profiles produced by these primers were sufficiently polymorphic, especially those generated by OPAI-09 and OPAI-13. The profiles produced by MBR and NTR were somewhat less polymorphic, which is not entirely undesirable. Primers that only select for polymorphisms will result in a biased picture of the genetic diversity within a population.

The number of primers used in RAPD profiling can differ widely. A number of studies have shown that one primer may be sufficient to discriminate between strains (Stephan *et al*, 1994). However, it has been recommended that more than one primer be used, particularly in the context of epidemiological subtyping (Mazurier and Wernars, 1992b). While one primer would not have

provided sufficient genetic information and strain discrimination, the large numbers of isolates in this study would have made it logistically difficult to use more than four.

PCR cycle parameters were optimised for each of the four primers (Table 2.5). It was not possible to achieve a 100% amplification rate and ninety eight of the 600 isolates could not be included in the final analysis of RAPD profiles, due to amplification failure with either one or more of the primers. Amplification failure can be due to a number of reasons, such as sequence differences between the primer and the DNA to be amplified, degradation of DNA, and species misidentification. The first two were the most likely reasons for amplification failure in this study.

Multiplexing of primers has also been used in the generation of RAPD profiles (Caetano-Anollés *et al.*, 1991). Two primers together may produce a profile which is not merely the result of adding the amplification products obtained separately with each individual primer. Certain bands will disappear, while new ones are generated. Multiplexing experiments were carried out in this study with two primers that had already been used individually. OPAI-09 and MBR were used at 0.2 μM and 0.1 μM each. The PCR thermal cycle programme was the same for both of these primers when used individually and so it was decided to use those parameters for the multiplex priming experiment. At a concentration of 0.2 μM each, only 70% of DNA preparations were amplified, with RAPD markers being few and faint. A concentration of 0.1 μmol each amplified only 33% of samples, with even fewer bands being generated. Further optimisation of the PCR reaction and thermal cycle parameters would have been required. However, multiplex priming was not pursued as two more ten-mer primers were subsequently found to gave satisfactory RAPD markers.

The four RAPD profiles generated from each isolate by each of the selected primers were not analysed separately but were combined to form a composite RAPD profile (Stephan *et al.*, 1994), which was then analysed as described in Section 2.4.

Table 2.5 PCR cycle parameters for selected ten-mer primers

Primer Name	PCR Cycle Parameters
OPAI-09	1 cycle : 94°C for 3 mins
	40 cycles : 94°C for 30 secs
	38°C for 30 secs
	72°C for 60 secs
	1 cycle : 72°C for 4 mins
MBR	1 cycle : 94°C for 3 mins
	40 cycles : 94°C for 30 secs
	38°C for 30 secs
	72°C for 60 secs
	1 cycle : 72°C for 4 mins
OPAI-13	1 cycle : 94°C for 3 mins
	40 cycles : 94°C for 30 secs
	36°C for 60 secs
	72°C for 60 secs
	1 cycle : 72°C for 4 mins
NTR	1 cycle : 94°C for 3 mins
	40 cycles : 94°C for 30 secs
	38°C for 45 secs
	72°C for 60 secs
	1 cycle : 72°C for 4 mins

2.3.5 Thermal Cycler

The variability of transition times between thermal cyclers and the effect of this on RAPD reactions is well documented (Hoelzel, 1990). The transition interval between the denaturing and annealing temperatures may be one of the main reasons for this phenomenon (Schweder *et al.*, 1995). Longer transition times resulted in the generation of more amplification products, while the converse was true of shorter transition times. These two situations have different benefits, with a longer transition time affording a better chance of detecting useful polymorphisms, while a shorter time generates more specific products and precludes the synthesis of fragments from mispaired duplexes.

Care was taken in this study to maintain consistent thermal cycler conditions. PCR reactions were performed in an Omnigene Temperature Cycler (Hybaid, UK), with accurate sample temperature control being achieved by the Active Tube Control software. A thermistor probe monitors the temperature in a dummy sample tube filled with 100 µl of mineral oil. Thus, the PCR reaction mixture within the tube, and not merely the wall of the PCR tube, reaches the required temperatures. The machine was kept in an airconditioned room at a constant temperature of 20°C, which ensured consistent transition times.

2.3.6 DNA molecular weight marker

The commercial standard, DNA Molecular Weight Marker type VI (Roche Diagnostics, USA), was used to quantify mycobacterial DNA in the early stages of this study. Consisting of a mixture of pBR328 DNA cleaved with *Bgl* 1, and pBR328 DNA cleaved with *Hinf* 1, electrophoresis in 10% polyacrylamide gel results in the resolution of twelve fragments, varying in size from 2176 to 154 base pairs. On agarose gel ten fragments were resolved, with an eleventh being very faint or occasionally even absent. With the InstaGene protocol, DNA quantification was not necessary. However, the marker was included in duplicate on every agarose gel, which was necessary for the computer-based cluster analysis of RAPD profiles (Section 2.4.2).

2.4 Analytical methods

2.4.1 UVP Gel Documentation System

The UVP Gel Documentation Software SW2000 (Ultra-Violet Products Ltd, UK) was used to convert RAPD markers on agarose gels into a tagged image file (TIF) format, which is the format needed for the cluster analysis programme. A video camera especially designed to resolve low light levels as found on fluorescent patterns like ethidium bromide-stained gels was fitted with an infra-red filter, a 16 mm lens and a specially developed UV filter to improve the contrast. The gel image was captured and converted into digitised optical density values, which were then stored in TIF format.

2.4.2 Cluster analysis

Having generated RAPD profiles from 502 Eastern Cape and 57 KwaZulu-Natal isolates, it was then necessary to determine the degree of similarity amongst them, which would in turn make it possible to gauge the genetic diversity between and within the populations from these two provinces. At this point, it should be noted that a change is made in nomenclature. The term "isolate" has been used up to now to refer to the cultures of *M. tuberculosis* from which DNA was extracted and amplified in order to generate RAPD profiles. The generation of such profiles constitutes typing of these isolates. An isolate, once typed, may be referred to as a strain as will be done from now on.

The degree of similarity may be determined by constructing a triangular similarity matrix of the genetic distances between RAPD profiles of each strain, using a similarity coefficient. Clustering of the similarity matrix is then carried out in order to display the relationships between the strains in a graphic way. Two classes of clustering strategies are available: Sequential, agglomerative, hierarchical, non-overlapping (SAHN) methods, and ordination methods. Principal Components Analysis falls into the latter category. This method, involving two- or even three-dimensional plotting of the similarity matrix, has not proved very useful with large databases (Barker, 1990). SAHN clustering is used more commonly in a wide variety of applications and involves the generation of dendrograms using a particular cluster algorithm, which results in the grouping of strains into clusters according to the similarity of their profiles. A number of computer programmes exist which offer a variety of similarity coefficients and cluster algorithms : NTSYS-pc ver 1.80 (Exeter Software, USA), Phylip ver 3.5 (J.Felsenstein, USA), RAPDistance ver 1.04 (Armstrong *et al.*, Australian National University, Australia) and GelCompar ver. 4.0 (Applied Maths, Belgium). However, of these, only GelCompar was capable of handling large numbers of strains (up to 2500), which made it the software of choice for performing cluster analysis in this study.

The first stage in cluster analysis involved converting gels from TIF into GelCompar format, at which time each strain was identified according to medical facility and town. The second stage of normalization involves a process of alignment of RAPD profiles. Gel to gel variation is inevitable when gels have been generated over a relatively long period of time. Normalization minimises this effect and makes comparison of all RAPD profiles possible. The DNA molecular weight marker

included on all gels was used as an external reference marker for this purpose. Alignment of the marker tracks with one another automatically resulted in the alignment of all RAPD profiles. The third stage involved the assigning of bands to RAPD profiles, using band searching filters which can be set to the desired degree of sensitivity. This is particularly useful with RAPD profiles, as it enables the exclusion of faint bands which may be artifactual. Using the molecular weight marker to determine the optimum settings, the minimal profiling filter was set at 4.5% and the minimal area filter at 2%. Bands were automatically detected using these optimal settings, after which the profiles were compared with the photographic record and edited only where necessary. The tolerance on the positional differences between two bands to be considered as matching also had to be adjusted, to allow for the percentage of error that may result from gel to gel variation. The molecular weight marker bands were once again used to determine the optimal setting to be 2%.

The next stage of cluster analysis involved the numerical comparison of the RAPD profiles based on their similarity. GelCompar offers five band-based similarity coefficients : Jaccard, Dice, a "fuzzy logic" coefficient, Jeffrey's *x* coefficient and an area-sensitive coefficient. The Jaccard coefficient (S_J) calculates the similarity between two profiles by dividing the number of corresponding bands between the two strains by the total number of bands in both strains:

$$\frac{n_{AB}}{n_A + n_B - n_{AB}}$$

n_{AB} is the number of bands in common for A and B, n_A is the total number of bands in A and n_B is the total number of bands in B.

The Dice coefficient (S_D) (a variation of Nei and Li's coefficient) is derived from, and very similar to, Jaccard's coefficient, but gives more weight to matching bands:

$$\frac{2n_{AB}}{n_A + n_B}$$

The "fuzzy logic" coefficient is based on Jaccard but assigns scores to corresponding bands in proportion to their degree of overlap.

Jeffrey's *x* coefficient is very similar to that of Dice:

$$\frac{n_{AB}}{n_A}+\frac{n_{AB}}{n_B}$$

The area-sensitive coefficient takes into account the correspondence of bands as well as the differences of the relative areas under each of the corresponding bands:

$$\frac{\Sigma_{\Delta}S_{AB}}{n_A+n_B-n_{AB}}$$

where $\Sigma_{\Delta}S_{AB}$ is defined as

$$\Sigma\frac{\alpha}{\alpha+|S_{Ai}-S_{Bi}|}$$

α is a constant; S_{Ai}-S_{Bi} is the absolute difference between the areas of band i on A and B, and i ranges from 1 to n_{AB}

The NTSYS-pc programme was used to compare the similarity matrices produced by four of these five coefficients, to determine whether they would generate widely divergent matrices of the same strains. The area-sensitive coefficient was omitted as band intensity was not to be taken into account in this study, only band position. The degree of similarity between any two matrices was expressed as the product-moment correlation, *r*. Correlations were performed on the full database of 502 Eastern Cape strains as well as on the smallest subset of 23 Eastern Cape rural drug resistant strains. Table 2.6 shows the similarity coefficient correlations of the latter database, with *r* being expressed as a percentage. In Table 2.7, the correlations are given for the total Eastern Cape database.

As can be seen from both tables, the similarity matrices generated by the Jaccard, Dice and Jeffrey's *x* coefficients were very similar, while the fuzzy logic coefficient produced a matrix which was slightly different to the other three. The dendrograms produced from the similarity matrices computed by Jaccard, Dice and Jeffrey's *x* coefficients were also very similar, with the fuzzy logic dendrogram showing slight differences in the clustering of strains. The Dice coefficient was held

to be the most suitable for a number of reasons. Firstly, it has been used widely in the field of mycobacterial research (van Soolingen *et al.*, 1994; Hermans *et al.*, 1995; Safi *et al.*, 1997). Secondly, it is preferred to the Jaccard coefficient which causes anomalies through giving equal weight to 1,1 and 1,0 matches but ignoring 0,0 matches (Barker, 1990). Thirdly, the dendrogram of a similarity matrix produced by the Dice coefficient has a higher heuristic value, making it easier to interpret visually (Barker, 1990).

Table 2.6 Correlation of similarity matrices of the Eastern Cape rural drug resistant subpopulation

	Jeffrey's *x*	Fuzzy Logic	Dice
Fuzzy Logic	93.00		
Dice	97.74	94.52	
Jaccard	97.34	93.84	99.55

Table 2.7 Correlation of similarity matrices of the total Eastern Cape population

	Jeffrey's *x*	Fuzzy Logic	Dice
Fuzzy Logic	90.75		
Dice	97.51	92.98	
Jaccard	97.34	92.72	99.56

Having selected a similarity coefficient, it was then necessary to decide which cluster algorithm was optimal for this system. The unweighted pair group method using arithmetic averages (UPGMA), which is widely used, was considered as was the less common algorithm of Ward (1963) which is eminently suited to the analysis of large databases. The Ward algorithm differs from UPGMA in that it minimises the overall deviation of the dendrogram from the original matrix of similarities. The dendrogram generated by a third algorithm, Neighbour Joining (NJ) (Saitou and Nei, 1987), should be interpreted in a different way from those generated by UPGMA and Ward. In the latter two, the level of the branch linking two strains determines the similarity between them. However, in an NJ dendrogram, the summed lengths, in the horizontal direction, of all branches which have to be followed from one track to another, indicate the distance between the two strains. The NJ tree is said to offer a more faithful representation of the original matrix of similarities, albeit one that is more difficult to interpret. It was not used in this study because the NJ algorithm entails a "shortest route" approach, with the result that the branch lengths connecting the strains will differ making

comparison with dendrograms generated by other cluster algorithms difficult. Furthermore, interpretation of the NJ dendrogram of a large database such as this would be problematic.

In attempting to determine which of the remaining two algorithms available in the GelCompar programme was suitable for use in this thesis, the "goodness of fit" of the clustering to the data in the original similarity matrix was investigated. This involved the generation of a cophenetic matrix of the dendrogram which was then compared to the original similarity matrix, resulting in the calculation of a cophenetic correlation value (CCV), expressed as a percentage. The higher the percentage, the better the degree of fit. However, interpretation of the degree of fit can be subjective. Rohlf (1993) adjudges a cophenetic correlation value of > 90% to represent a very good fit, while 80 to 90% is a good fit, 70 to 80% a poor fit, and less than 70% a very poor fit. On the other hand, Lapointe and Legendre (1992) would see a correlation of greater than 50% as acceptable. The use of the CCV as a criterion of optimality is problematic, with the value decreasing as numbers of strains increase, although increasing numbers of characters have no effect on it (Rohlf and Fisher, 1968). Barker (1990) obtained a CCV as low as 30% for a database of 200 strains with 200 characters. Farris (1972) was not satisfied with using the CCV to measure the goodness of fit of a dendrogram, especially as it was believed to be sensitive to cluster sizes, and maintained that the similarity coefficient was the more important criterion of optimality. The CCVs obtained when using UPGMA to cluster a variety of similarity matrices of a number of different databases in this study are set out in Table 2.8, where can be seen that they were far from ideal if one used Rohlf's categories, especially with larger databases. It is evident that the number of strains has a significant effect on the CCV, which is satisfactorily high for the smallest database of 23 strains, but falls by between 7 and 15% with the largest database of 502 strains.

In addition to the low CCV, UPGMA produced dendrograms of a very low heuristic value making them very difficult to interpret, especially when large numbers of strains were involved (Figure 2.6). The reason for this lack of heuristic value is that UPGMA does not recalculate the matrix of similarities after adding a strain to a particular cluster (Barker, 1990).

Table 2.8 Cophenetic correlation values with the UPGMA cluster algorithm

Database	Number of strains	Similarity Coefficient	Cophenetic Correlation (%)
Eastern Cape	502	Jaccard	63.6
		Dice	64.1
		Fuzzy Logic	59.8
		Jeffrey's *x*	60.8
Port Elizabeth	215	Jaccard	66.9
		Dice	66.4
		Fuzzy Logic	62.5
		Jeffrey's *x*	62.1
Urban Resistant	64	Jaccard	69.9
		Dice	69.8
		Fuzzy Logic	68.6
		Jeffrey's *x*	65.3
Rural resistant	23	Jaccard	74.7
		Dice	73.4
		Fuzzy Logic	72.3
		Jeffrey's *x*	71.3

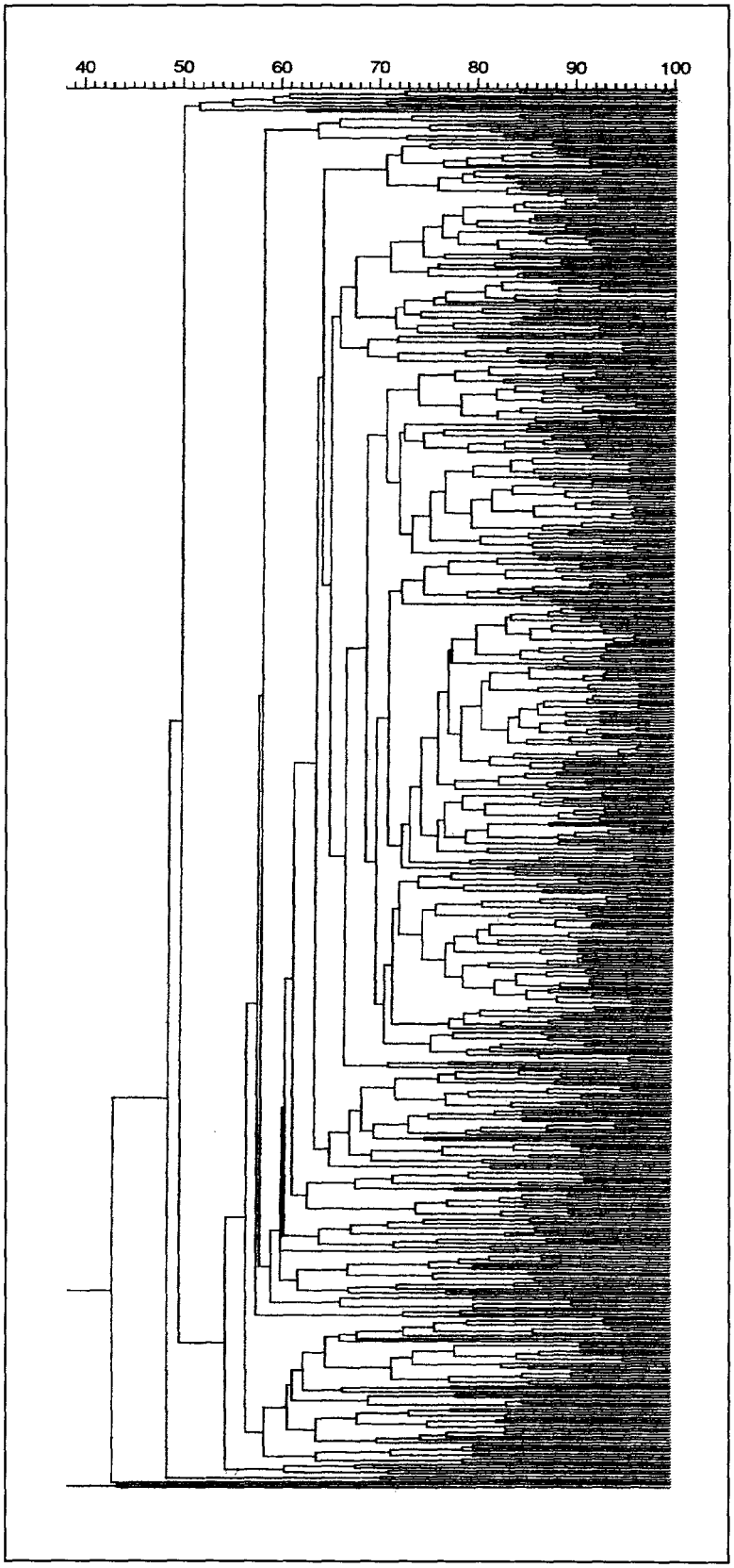


Fig 2.6 Cluster analysis of the Eastern Cape population with the unweighted pair group method using arithmetic averages (UPGMA) algorithm

It is not possible to test the goodness of fit of a dendrogram generated by the Ward algorithm using the cophenetic correlation value, because this algorithm is not based on arithmetic averages (Luc Vauterin, Applied Maths, personal communication). However, the Ward algorithm was considered more suitable for use in this study due to its ability to perform hierarchical grouping on large sets of data with a minimum loss of information (Ward, 1963). Furthermore, in contrast to UPGMA, the matrix of similarities is recalculated by the Ward algorithm after each joining event, which results in a dendrogram of higher heuristic value (Figure 3.1). Clusters are visually apparent, with branches joining strains within the clusters being shorter than the branches that join clusters to each other. Consequently, cluster analysis of all strains included in this study was carried out using the Dice similarity coefficient and the Ward cluster algorithm.

2.4.3 Analysis of molecular variance

Analysis of molecular variance involves the use of specifically designed software, in this case Arlequin ver. 1.1 (Schneider *et al.*, 1997), to determine genetic diversity and population structure. Although originally designed to be used in the analysis of diploid organisms where the two states of the genotype can be identified, it is applicable to haploid organisms such as bacteria and to the RAPD system (L. Excoffier, personal communication). An analysis of this nature is vital in a study of this size, due to its ability to determine the statistical significance of the results obtained.

The genetic information within each RAPD profile was extracted from GelCompar in the form of binary data. A binary table, representing presence (1) and absence (0) of the RAPD markers, was generated, which then formed the basis of Arlequin input files. Each band class created by GelCompar was designated as a locus by Arlequin, with one of two alleles (0 or 1) present at each locus. The sequence of alleles at all defined loci constituted the haplotype of each strain within the population, as follows:

Individual 1 : 1001110001010101
Individual 2 : 0101010111100110

Genetic diversity was then determined by measuring two molecular indices. The first involved calculation of the number of polymorphic sites in each population tested. In the example above, there are a total of sixteen loci and nine of them present more than one allele, making them

polymorphic. The second molecular index measured was the mean number of pairwise differences that existed between all pairs of haplotypes in the database. This was calculated using the following equation :

$$d_{xy} = \sum_{i=1}^{L} \delta_{xy}(i)$$

where δ_{xy} (I) is the Kronecker function, equal to 1 if the alleles of the I-th locus are identical for both haplotypes, and equal to 0 otherwise (Arlequin ver. 1.1 manual, 1997).

Detection of population structure also contributed to determining the degree of genetic diversity amongst the strains in this study. One or more groups consisting of a number of populations, which in turn each consisted of a number of strains, were defined, thus setting up a particular population structure to be tested. Examples of the types of populations that were tested can be found in Table 3.12. Detection of population structure involved the computing of $\boldsymbol{F}_{ST}$ values for all pairs of populations using the following equation:

$$\frac{f_0 - f_1}{1 - f_1} = \frac{t_1 - t_0}{t_1}$$

where f_0 is the probability of identity by descent of two different genes drawn from the same population, f_1 is the probability of identity by descent of two genes from two different populations, t_1 is the mean coalescence time of two genes drawn from two different populations, and t_0 is the mean coalescence time of two genes drawn from the same population (Slatkin, 1991). Such $\boldsymbol{F}_{ST}$ values may then be used to represent short term genetic distances between populations. The smaller the $\boldsymbol{F}_{ST}$ value, the more closely related are the populations, with the converse being true. The statistical significance of the $\boldsymbol{F}_{ST}$ values was calculated, mostly at the 5% level. Another measure of population structure was the migration rate ($\boldsymbol{M}$) between two populations, which formed a corollary to their short term genetic distance. $\boldsymbol{M}$ assumed that two populations exchanged migrants each generation and that the mutation rate was negligible. The larger the value of $\boldsymbol{M}$, the greater the rate of genetic migration between two populations and thus the more similar they would be, with the converse once again being true.

A test was also performed to determine whether there was a significant association, or linkage disequilibrium, between particular loci. This test is an extension of the Fisher exact probability test on contingency tables (Slatkin, 1994). It was of particular application in this study to the examination of drug resistant populations for the existence of unique RAPD markers associated with resistance to particular antibiotics.

2.4.4 Geographical information system analysis

The latest GIS software, TNT-Mips, was used to plot the geographical, and integrate the non-geographical, data produced in this study. The former consisted of the locations of the Eastern Cape towns and medical facilities from which isolates were obtained. A digital map of the Eastern Cape containing most of the 110 medical facilities surveyed, was obtained from the Human Sciences Research Council (HSRC). The facilities not on the map were located and inserted by the staff of Geodatec, a private company which produced the various maps for the geographical analysis of the Eastern Cape sample. The positions of the mobile clinics have been given in relation to the towns in which they are based.

The non-geographical data comprised the genetic data obtained from the cluster analysis of the strains. The distribution of clusters identified in the total Eastern Cape population, and in the drug resistant and urban subpopulations, was plotted according to medical facility. Maps were also generated depicting the total number of clusters occurring in each medical facility.

CHAPTER 3

GENETIC DIVERSITY OF *M. TUBERCULOSIS* IN THE EASTERN CAPE PROVINCE

3.1 Preliminary results obtained with two primers

In order to obtain a preliminary indication of the genetic diversity within this province, genetic and geographical analyses were initially performed on the RAPD markers of 359 isolates (Richner *et al.*, 1997). Strains were from six urban and twenty eight rural locations and 19% were drug resistant. Two of the four selected primers, OPAl-09 (also known as RP15) and MBR, had been used to amplify the DNA extracted from these isolates. The two profiles generated from each strain were combined to form a composite RAPD profile, which provides more genetic information on each strain. The dendrogram of the cluster analysis showed that the population divided into two major groups, with the majority of strains (70%) falling into one. Only eight sets of identical groups occurred, comprising a total of seventeen strains. The degree of similarity amongst the remaining 342 unique RAPD profiles ranged from 70 to 97%.

Analysis of the geographical distribution of strains revealed that only 44% of those that were genetically very similar were also from areas of close geographical proximity. Thus, the majority of strains that clustered together at high similarity indices (96-98%) were from geographically distant areas. There was no clear urban-rural divide, with strains from rural areas grouping together at a relatively high similarity index with those from urban areas. The ratio of urban to rural strains was the same in both groups. An analysis of the drug resistant subpopulation showed that 63% grouped together in the smaller of the two major groups. Four sets of identical groups were found, consisting of a total of nine resistant strains, with a further 59 strains having unique RAPD profiles with a degree of similarity ranging from 87 to 98%. Examination of genetically identical strains for antibiotic phenotype revealed that the strains in only one of the four identical groups shared the same resistance pattern, both being resistant to INH. A relatively large proportion of the population (50.4%) consisted of isolates from the greater P. E. area. Amongst these, too, a large number of unique RAPD profiles occurred, with only four groups of six strains having identical profiles.

These preliminary results were surprising for a number of reasons. Firstly, there was a higher than expected degree of genetic diversity, with 95% of the strains having unique profiles. Secondly, no correlation could be demonstrated between population structure and urban-rural location. Thirdly, closely related strains did not always originate from the same geographical area. Finally, the clustering together of the majority of drug resistant strains seemed to point towards person-to-person transmission as the more important means of acquiring such strains, as opposed to *de novo* development. It remained to be seen whether the results generated from the total population, with four primers, would lead to any significant changes in these observations.

3.2 Results obtained with four primers

The final results of this study were presented at the Fifth International DNA Fingerprinting Conference held in Grahamstown, South Africa (Richner *et al.*, 1999).

3.2.1 Cluster analysis

Cluster analysis of the total Eastern Cape population, as well as the drug resistant and P.E. subpopulations was undertaken in order to determine the degree of genetic diversity amongst these strains.

3.2.1.1 Total Eastern Cape Population

RAPD profiles generated by each of the four selected primers were combined to form a composite profile for each strain. Cluster analysis of the 502 profiles resulted in a dendrogram rooted at a similarity index of 37.6%, with further branching occurring at similarity levels of 53% and 72% (Figure 3.1). Further in-depth analysis required a methodology for defining smaller groups of related strains. The literature reveals the existence of a large degree of subjectivity in the analysis of dendrograms. However, such subjectivity should be complemented by the objectivity that is inherent in the structure of the dendrogram. This principle is seen in the definitive paper on the taxonomy of *Streptomyces* by Williams *et al.* (1983), in which cluster groups were defined at a similarity index of 70.1% and clusters at 77.5%. These indices were selected as they were the highest which gave clear groupings (Williams *et al.*, 1983). Clearly, these researchers were guided by the structure of the dendrogram.

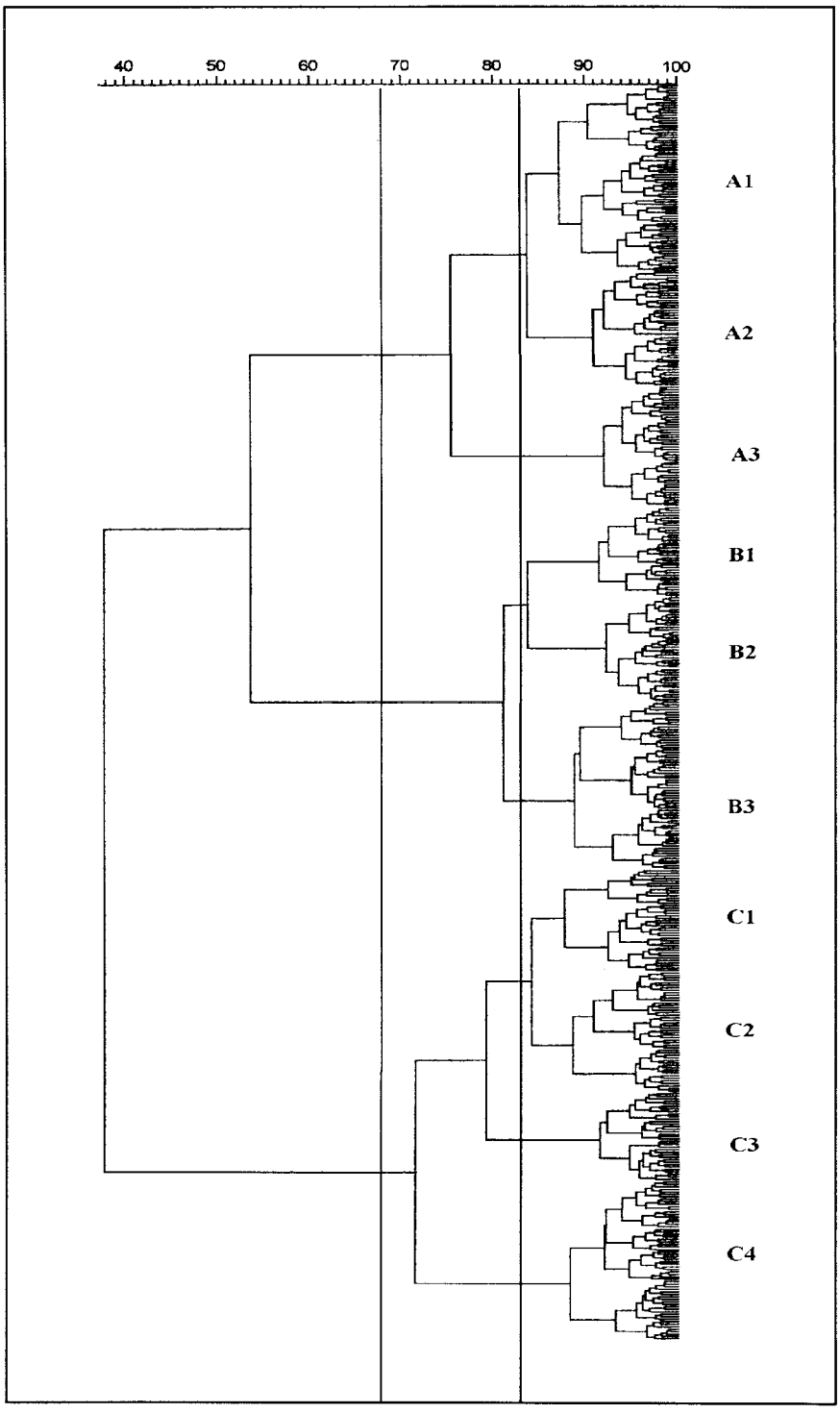


Figure 3.1 Cluster analysis of the Eastern Cape population with the Ward algorithm

Other taxonomic studies have used the Phenetic Ingroup Method (PIM) to determine where to place what is known as a phenon line. This methodology uses a group of strains which are known to cluster together in order to determine the position of the phenon line, which must be drawn down the length of the dendrogram to the left of the known cluster, thus determining the similarity index at which clusters can be defined for that particular population (Barker, 1990). Molecular epidemiological studies of *M. tuberculosis* have used a variety of similarity indices for defining clusters. Van Soolingen *et al.*(1991) and Harn *et al.*(1997) used a 60% similarity index, Yang *et al.* (1995) used 70% and Safi *et al.*(1997) used 75%. In the case of van Soolingen's study, the particular similarity index was chosen because it was at that level that the Dutch and African strains clustered separately. In Harn's and Safi's studies, no reasons were given for using a particular similarity index. In Yang's study, it was clear from the structure of the dendrogram that a similarity index of 70% was the most suitable for defining three large clusters which contained a large proportion of the population.

The PIM method could not be used in this study, as a known group of strains that should always cluster together was not included. However, the high heuristic value of dendrograms created by the Ward algorithm facilitated the use of the structure as the criterion for defining cluster groups. The depths of branches in Figure 3.1 were particularly helpful in determining where the cluster group phenon line should be placed. A phenon line placed at a similarity index of 70% resulted in the definition of three large cluster groups, called "A", "B" and "C". The branch lengths separating cluster groups A and B were clearly different, resulting in 75 to 100% similarity amongst strains in cluster group A and 81 to 99% in cluster group B. Both cluster groups A and B were rooted at the first branching event which occurred at a similarity index of 53%. Cluster group C was very different, however, branching directly from the root of the dendrogram, with a range of 72 to 99% similarity being seen amongst strains (Table 3.1).

The structure of the dendrogram was such that the phenon line for defining clusters could have been placed at a similarity index of either 78% or 85%. Simpson's Index of Diversity, which calculates the probability of two unrelated strains being placed into different groups, has previously been applied to genetic and other methods of typing bacteria and yeasts, and provides an indication of which method is the most discriminatory (Hunter and Gaston, 1988; Dillon *et al.*, 1993).

The numerical index can be calculated using the following equation:

$$D = 1 - \frac{1}{N(N-1)} \sum_{j=1}^{S} n_j(n_j - 1)$$

Positioning of the phenon line at 78% resulted in an index of discrimination of only 77%, compared to that of 90% with the phenon line at 85% similarity. An index of 90% or more is indicative of satisfactory discrimination of strains (Hunter and Gaston, 1988). Thus, RAPD-PCR was shown to provide satisfactory discrimination of the strains in this population when clusters were defined at 85% similarity. Simpson's Index was thus used in this study to determine the positioning of the phenon line for cluster definition. The term cluster is used in this study as it was used by Williams *et al.* (1983), which is in a broader sense than its use in most molecular epidemiological studies, where it refers only to groupings of strains with identical DNA fingerprints.

Analysis of the cluster groups and clusters in Figure 3.1 was performed to determine the distribution of urban as opposed to rural strains, and drug resistant as opposed to drug sensitive strains (Table 3.2). Cluster group C contained the largest number of drug resistant and urban strains, while cluster group B contained the largest proportion of rural strains. As regards the clusters, the largest number of strains (including the type strain, H37Rv) were found in cluster A1 and the smallest number in cluster C3. Cluster B3 contained the largest number of rural strains and cluster A1 the largest number of urban strains. In most clusters, about two-thirds of strains were urban and the rest rural, which was in keeping with the proportions in the population as a whole, with 68% being from urban and 32% from rural areas. The drug resistant subpopulation will be discussed in detail in Section 3.2.1.2.

Only two strains with identical RAPD profiles were found and they occurred in cluster A2. These two strains originated in P.E. and Fort Beaufort, and were both drug sensitive. Very few groups of identical profiles had occurred in the preliminary analysis and the additional genetic information provided by a further two primers resulted in the loss of seven such groups. Such a high degree of genetic diversity, based on the DNA polymorphism of RAPD profiles, was unexpected.

Table 3.1 **Similarity indices of cluster groups and clusters in the Eastern Cape population**

Cluster Group	Similarity index of branching point (%)	Similarity within cluster group (%)	Cluster	Similarity index of branching point (%)	Similarity within cluster (%)
A	53	75 - 100			
			A1	84	87.5 - 99
			A2	84	91 - 100
			A3	75	92 - 99
B	53	81 - 99			
			B1	83	92 - 99
			B2	83	92 - 99
			B3	81	89 - 99
C	37.6	72 - 99			
			C1	84	89 - 99
			C2	84	89 - 99
			C3	79	91.5 - 99
			C4	72	88 - 99

Table 3.2 **Composition of cluster groups and clusters in the Eastern Cape population**

Cluster Group	Cluster	Number of strains	Proportion of population (%)	Urban Strains	Rural Strains	Drug resistant Strains
A		**169**	**34**	**118**	**51**	**19**
	A1	75	14.9	55	20	12
	A2	46	9.2	29	17	2
	A3	48	9.6	34	14	5
B		**145**	**29**	**87**	**58**	**18**
	B1	36	7.2	20	16	2
	B2	42	8.3	30	12	9
	B3	67	13.3	37	30	7
C		**188**	**37**	**129**	**59**	**48**
	C1	41	8.2	29	12	16
	C2	48	9.6	34	14	4
	C3	35	7.0	19	16	17
	C4	64	12.7	47	17	11

A full geographical analysis was performed using GIS technology (Section 3.2.3), but at this point the dendrogram was examined to determine whether strains which clustered together at a high similarity index (between 96 and 100%) were from the same town. There were 204 such groups, consisting of a minimum of two and a maximum of five strains each. Only 62 (30%) of these contained two or more strains from the same town. In 39 (63%) of these 62 groups, the strains were all from urban areas, while the rest were a mixture of urban and rural strains or consisted of only rural strains. This gave an early indication that there was a wide geographical distribution of strain types.

3.2.1.2 Drug resistant subpopulation

The majority of this subpopulation - 89% - was resistant to one, two or three antibiotics, with the rest being resistant to four, five or six. The only antibiotic to which none of the Eastern Cape subpopulation had developed resistance was ethambutol. Table 3.3 indicates the incidence of resistance to individual antibiotics within this subpopulation. As resistant isolates came from only 41% of the towns surveyed, these trends should not be extrapolated to the entire province.

Table 3.3 Incidence of antibiotic resistance in the Eastern Cape drug resistant subpopulation

Antibiotic	Number of resistant strains (n=85)	Percentage
Isoniazid *	82	92
Rifampicin *	40	47
Pyrazinamide	25	29
Ethionamide	12	14
Thiacetazone	11	13
Streptomycin *	9	11
Ethambutol *	0	0

* indicates first line antibiotics

The highest rates of resistance were to two of the four first line antibiotics in common use. The low level of resistance to streptomycin is in all likelihood due to reluctance to use this antibiotic because of its serious side effects.

The antibiotic phenotypes, or the range of antibiotics to which the various strains were resistant, were then identified (Table 3.4). The incidences of the three phenotypes occurring most commonly in this subpopulation are given in Table 3.5. The incidences of the remaining phenotypes varied from 1.2 to 5%. Also indicated is the distribution of these phenotypes amongst the ten clusters. It would seem that resistance to INH is acquired first, followed by rifampicin resistance, and then by resistance to the second line antibiotic, pyrazinamide. No correlation is evident between a particular antibiotic phenotype and specific clusters, as confirmed by the analysis of the antibiotic phenotypes in each cluster (Table 3.6). The largest number of phenotypes occurred in cluster A1, followed by cluster C1 and C3. In four of the ten clusters (A2, A3, B1, C2), each strain had a different phenotype. Examination of closely related pairs of strains in clusters C1 and C3 yielded a very small proportion of correlation between phenotypes and RAPD profiles. This was seen in cluster C1, where two very closely related strains with a similarity index of 97% had the same phenotype, and in cluster C3, where two pairs of very similar strains with similarity indices of 97 and 99.5% respectively, had the same phenotypes.

Of interest would have been to determine whether strains resistant to between one and three antibiotics clustered together in significant numbers, as opposed to those resistant to between 4 and 6 antibiotics. However, it would not have been possible to determine the significance of such grouping, because the subpopulation was numerically biased towards strains resistant to fewer antibiotics. It was interesting to observe that the majority of strains in the two clusters which contained most of the drug resistant strains - fifteen of the sixteen strains in cluster C1 and sixteen of the seventeen strains in cluster C3 - were resistant to between one and three antibiotics. The few strains that were resistant to between four and six antibiotics occurred in clusters A1, A3, B3 and C2.

A high degree of genetic diversity was seen in this subpopulation, with resistant strains occurring in all ten clusters (Table 3.2) However, a noticeable separation was seen, with fifty six percent of all resistant strains occurring in cluster group C, as opposed to only 22% each in cluster groups A and B. Strains in clusters C1 and C3 grouped together at relatively high similarity indices compared to those in the other clusters, where they were interspersed with other drug sensitive strains. The largest number of drug resistant strains belonged to cluster C3, followed closely by cluster C1.

These formed a relatively large proportion (49% and 39% respectively) of the total number of strains in these clusters, compared to that in other clusters, which ranged from 4 to 19%.

Table 3.4 Antibiotic phenotypes in the Eastern Cape drug resistant subpopulation

Phenotype Name	Phenotype Description *	Number of strains
EC1	INH, rifampicin & pyrazinamide	12
EC2	INH	30
EC3	INH, pyrazinamide & streptomycin	3
EC4	INH & rifampicin	15
EC5	INH & pyrazinamide	4
EC6	INH, pyrazinamide, rifampicin & streptomycin	1
EC7	rifampicin	3
EC8	INH, rifampicin & ethionamide	1
EC9	INH, pyrazinamide, rifampicin, streptomycin & thiacetazone	1
EC10	INH, pyrazinamide, rifampicin, streptomycin, thiacetazone & ethionamide	2
EC11	INH, rifampicin, streptomycin, thiacetazone & ethionamide	1
EC12	INH, rifampicin, thiacetazone & ethionamide	2
EC13	INH & ethionamide	2
EC14	INH & thiacetazone	1
EC15	INH, rifampicin & thiacetazone	1
EC16	INH, thiacetazone & ethionamide	2
EC17	INH, rifampicin, streptomycin & thiacetazone	1
EC18	INH & streptomycin	1
EC19	INH, pyrazinamide & ethionamide	1
EC20	INH, pyrazinamide, rifampicin & ethionamide	1

* isolates are resistant to the antibiotic/s indicated in the phenotype description and sensitive to the rest

Table 3.5 Incidence of the most common antibiotic phenotypes in the Eastern Cape

Phenotype Name	Number of antibiotic resistances	Incidence	Number of clusters
EC2	1	35%	8
EC4	2	18%	6
EC1	3	14%	5

Table 3.6 Distribution of Eastern Cape antibiotic phenotypes

Cluster	Phenotypes	Number of phenotypes	Number of strains
A1	EC2, EC4, EC6, EC7, EC10, EC13, EC15	7	12
A2	EC1, EC2	2	2
A3	EC1, EC2, EC10, EC13, EC20	5	5
B1	EC2, EC4	2	2
B2	EC2, EC3, EC4, EC5, EC19	5	9
B3	EC1, EC4, EC12, EC16, EC17	5	7
C1	EC1, EC2, EC4, EC7, EC9, EC14	6	16
C2	EC4, EC5, EC11, EC16	4	4
C3	EC1, EC2, EC4, EC8, EC12, EC18	6	17
C4	EC2, EC3, EC5, EC7	4	11

An attempt was made to identify the existence of unique RAPD markers associated with resistance to a particular antibiotic in this subpopulation. Three databases were generated, using GelCompar's Polymorphism Analysis module, of the markers amplified by each primer from strains resistant to INH, pyrazinamide and rifampicin, the three antibiotics to which there was the highest incidence of resistance in this subpopulation. An equal number of randomly selected strains sensitive to these antibiotics was included in each database. The incidence of each RAPD marker that occurred amongst the resistant strains was compared to its incidence amongst sensitive strains. The INSTAT ver. 2.04a (GraphPad Software) programme was used to determine the significance of these incidences, using Fisher's Exact Test. Seven RAPD markers were detected which occurred either more or less frequently amongst drug resistant as opposed to drug sensitive strains.

However, only the three that occurred significantly more frequently ($p \leq 0.05$) were of interest (Table 3.7). Two were associated with INH resistance and were generated by primers OPA1-09 and OPA1-13, respectively, while the third was linked to rifampicin resistance and was produced by primer OPA1-13.

Table 3.7 RAPD markers linked to antibiotic resistance

Primer	Antibiotic	Marker size (in base pairs)
OPA1-09	INH	550
OPA1-13	INH	520
OPA1-13	Rifampicin	520

These three markers were all relatively small in size compared to some of those that occurred less frequently amongst resistant strains, which were from 540 to 1820bp in length.

3.2.1.3 Port Elizabeth subpopulation

Forty three percent of Eastern Cape strains were received from clinics and hospitals in the P.E. metropolitan area. Table 3.8 shows the distribution of these strains amongst the cluster groups and clusters identified in Figure 3.1. A high degree of genetic diversity was found in this subpopulation as well, with strains from all ten clusters circulating in this urban area. The largest number of P.E. strains occurred in cluster group A and cluster A1, while cluster B1 was less common. The remaining P.E. strains were fairly evenly spread out amongst the other eight clusters.

The comparative distribution of the 214 P.E. strains and the 124 strains received from the five other urban centres demonstrated the existence of a high degree of genetic diversity in these urban areas as well (Table 3.9). This could be seen from the fact that strains belonging to all ten clusters occurred in Uitenhage, even though the sample number was about three times smaller than that of P.E. Genetic diversity was also high amongst the relatively small number of strains from East London, Despatch and Grahamstown.

Table 3.8 Distribution of Port Elizabeth strains

Cluster Group	Cluster	Total number of strains	Number of P.E. strains	Number of P.E. drug resistant strains
A		**169**	**89**	**13**
	A1	75	41	8
	A2	46	19	2
	A3	48	29	3
B		**145**	**45**	7
	B1	36	3	0
	B2	42	22	5
	B3	67	20	2
C		**188**	**80**	**28**
	C1	41	23	9
	C2	48	21	4
	C3	35	11	8
	C4	64	25	7

Table 3.9 Distribution of all urban strains

City	Cluster groups	Clusters	Number of clusters	Number of strains
Port Elizabeth	A, B, C	A1-3, B1-3, C1-4	10	214
Uitenhage	A, B, C	A1-3, B1-3, C1-4	10	70
Despatch	A, B, C	A1, A3, B1, B3, C2, C4	6	11
East London	A, B, C	A1-3, B1-3, C1, C2, C4	9	34
Grahamstown	A, B, C	A2, B1-3, C1, C2	6	8
King William's Town	A	A1	1	1

Of the 48 drug resistant P.E. strains, the largest number were to be found in cluster group C and cluster C1, with no drug resistant P.E. strains occurring in cluster B1. The incidence of the most common phenotypes circulating in this city is given in Table 3.10.

Table 3.10 Incidence of the most common antibiotic phenotypes in Port Elizabeth

Phenotype name	Number of antibiotic resistances	Incidence (%)	Number of clusters
EC2	1	35	6
EC4	2	15	3
EC1	3	13	4

The three antibiotic phenotypes that occurred most commonly in the total resistant subpopulation were also to be found in P.E., with the incidence rates being approximately the same. Fifteen of the 20 phenotypes identified in this province occurred in this subpopulation An investigation of the distribution of antibiotic phenotypes amongst the resistant P.E. strains according to cluster demonstrated that the largest number were to be found in cluster C1 (Table 3.11). In four of the ten clusters (A2, A3, B3, C2), there were as many phenotypes as drug resistant strains, indicating an absence of correlation between antibiotic phenotype and genetic type, as was seen with the full drug resistant subpopulation.

Table 3.11 Distribution of antibiotic phenotypes in the P.E. drug resistant subpopulation

Cluster	Phenotypes	Number of phenotypes	Number of Resistant Strains
A1	EC2, EC6, EC10, EC15	4	8
A2	EC1, EC2	2	2
A3	EC1, EC10, EC13	3	3
B2	EC2, EC3,EC5	3	5
B3	EC16, EC17	2	2
C1	EC1, EC2, EC4, EC7, EC9	5	9
C2	EC4, EC5, EC11, EC16	4	4
C3	EC1, EC2, EC4, EC8,	4	8
C4	EC2, EC3, EC7	3	7

3.2.2 Analysis of molecular variance

Strains from the Eastern Cape were divided into a number of different databases for analysis with the Arlequin programme (Table 3.12). The various databases were designed to test for population structure and genetic diversity amongst specific populations. Databases 1 and 2 tested the cluster groups and clusters defined by the Ward cluster algorithm (Figure 3.1). Database 3 was set up to determine whether there was significant difference between urban as opposed to rural strains. Databases 4 and 5 were designed to determine whether strains from the different health regions varied to any significant degree. Database 6 tested for structure between drug resistant and drug sensitive strains, and Databases 7 and 8 focussed on the differences amongst strains from the P.E. subpopulation. Linkage disequilibrium tests were also performed on Database 6.

3.2.2.1 Total Eastern Cape population

The molecular and population structure indices for Databases 1 to 5 are set out in Appendix D. The three cluster groups contained an equal amount of genetic diversity as evidenced by the number of polymorphic sites and the mean number of pairwise differences (Table D.1, Appendix D). The short term genetic distances (F_{ST}s) between the three cluster groups were not very large, with the mean being 0.025 (Database 1, Appendix D). A study of various conspecific populations of *Bacillus anthracis* has shown genetic distances ranging from 0.03 to 0.22 (Keim *et al.*, 1999), while Jackson *et al.* (1999) obtained a genetic distance of 0.97 for congeneric populations of *B. anthracis* and *B. cereus*. Thus, the genetic distances obtained here clearly reflected the conspecific nature of this population. The lowest migration rate occurred between strains of cluster group B and C.

Of the ten clusters in Database 2, clusters B3 and C4 demonstrated the greatest degree of genetic diversity and cluster C1 the smallest (Table D.2). As regards population structure, relatively large genetic distances were seen between clusters C1 and A2, A3, B1, B2 and B3, and between C3, and A1 and B1 (Database 2, Appendix D). It will be seen in Section 3.2.2.2 that this is probably due to the existence of the relatively large number of drug resistant strains in clusters C1 and C3.

Table 3.12 Arlequin Databases of the Eastern Cape population

Database Number	Description of Database	Number of Strains	Group/s	Populations	Number of strains in population
1	Eastern Cape Cluster Groups	502	Group 1	Cluster Group A	169
				Cluster Group B	145
				Cluster Group C	188
2	Eastern Cape Clusters	502	Group 1	Cluster A1	75
				Cluster A2	46
				Cluster A3	48
			Group 2	Cluster B1	36
				Cluster B2	42
				Cluster B3	67
			Group 3	Cluster C1	41
				Cluster C2	48
				Cluster C3	35
				Cluster C4	64
3	Urban and Rural Strains	502	Group 1	Urban Strain	338
				Rural Strains	164
4	All Strains by Health Region	502	Group 1	Region A	388
				Regions B & C	114
5	Rural Strains by Health Region	162	Group 1	Region A Rural	93
				Region B & C Rural	75
6	Drug Resistant and Drug Sensitive Strains	502	Group 1	Sensitive	417
				Resistant	85
7	Port Elizabeth strains	214	Group 1	Section 1	75
				Section 2	67
				Section 3	72
8	P. E. Drug Resistant & Sensitive strains	214	Group 1	Sensitive	166
				Resistant	48

The mean number of pairwise differences was very similar in the urban and rural subpopulations of Database 3 (Table D.3). The short term genetic distance was very small and the migration rate very high (neither value was statistically significant at the 5% or 10% level), indicating the existence of very little population structure between urban and rural strains (Database 3, Appendix D). There was no significant difference between all strains from the three health regions, as indicated by a high migration rate (Database 4, Appendix D). However, it was interesting to note that the migration rate between rural strains from Health Region A on the one hand, and rural strains from Health Regions B and C on the other, was lower than that of Database 4 (Database 5, Appendix D). However, this evidence of a small degree of population structure was significant at the 10% level only.

3.2.2.2 Drug resistant subpopulation

Molecular indices for this subpopulation indicated the existence of fewer polymorphic loci amongst the resistant strains, which can probably be attributed to the less polymorphic RAPD profiles generated amongst drug resistant strains by primer NTR (Table D.6). The genetic distance between the drug resistant and drug sensitive subpopulations was seen to be a hundred-fold greater than that between the urban and rural populations (Database 6, Appendix D).

Linkage disequilibrium tests were carried out on this database in order to determine whether any RAPD markers in the drug resistant subpopulation could be linked to resistance to specific antibiotics. Tests were performed for resistance to INH, rifampicin and pyrazinamide. A table of all markers obtained with all four primers of the 85 resistant strains was generated in GelCompar. This was converted into binary data, where 1 indicated the presence, and 0 the absence, of a particular marker. The haplotype of each strain initially consisted of 86 loci, with one of two alleles (0 or 1) present at each locus. A further three loci were then added, one each to represent isoniazid, rifampicin and pyrazinamide. Where a strain was resistant to the particular antibiotic, 1 was added, and where it was sensitive 0 was added, thus creating new haplotypes for each strain consisting of a total of 89 loci. A matrix was then produced which indicated all loci that were significantly associated at the 5% level. Only those loci which were associated with the three that represented resistance to the antibiotics were of interest. Table 3.13 indicates the number of such associated loci per primer, the antibiotic with which they were associated and the size, in base pairs, of the

marker represented by the particular locus. The largest number of markers associated with resistance to the three antibiotics was generated by primer OPA1-09. Nine markers associated with resistance to pyrazinamide only were generated by all four primers. Eight markers associated with INH resistance and three with rifampicin resistance were generated by the four primers. One marker associated with resistance to both pyrazinamide and rifampicin was generated by primer MBR. Primer OPA1-09 and OPA1-13 each generated one marker of the same size (750 base pairs) both of which were associated with INH resistance. The markers associated with pyrazinamide resistance covered a wide size range, from 370 to 1800 base pairs, while markers associated with INH and rifampicin resistance covered a narrower size range, from 234 to 950 base pairs, and 440 to 1300 base pairs, respectively. A significant association was also revealed between rifampicin and pyrazinamide resistance, but not between INH and either of the other two antibiotics.

3.2.2.3 Port Elizabeth subpopulation

Isolates were obtained from a total of 28 medical facilities situated in 17 suburbs and townships in the greater P.E. area. They were subdivided into three sections with a view to testing for genetic differences between strains received from facilities in the black townships (Sections 1 and 2), and those from facilities situated in predominantly white and coloured suburbs (Section 3). There was little genetic diversity amongst the three sections (Table D.7). The short term genetic distances were small, and the migration rates large, between Sections 1 and 2, and Sections 1 and 3 (Database 7, Appendix D). However, a small degree of population structure existed between Sections 2 and 3, with a somewhat lower migration rate. However, these tests were not significant at the 5% level.

The P.E. drug resistant strains demonstrated a smaller degree of polymorphism than did the P.E. sensitive strains (Table D.8). The F_{ST} and migration rate of Database 8 was approximately the same as that found with the Database 6 where all drug resistant and drug sensitive strains were compared.

Table 3.13 Linkage disequilibrium data

Primer	Number of associated loci	Antibiotic	Marker size(bp)
OPA1-09	9	Pyrazinamide	1800
		Pyrazinamide	1240
		Pyrazinamide	1120
		Pyrazinamide	520
		Pyrazinamide	370
		INH	750
		INH	700
		INH	570
		Rifampicin	780
MBR	3	Pyrazinamide	700
		Pyrazinamide & Rifampicin	560
		Rifampicin	440
OPA1-13	5	Pyrazinamide	1766
		Rifampicin	1300
		INH	950
		INH	750
		INH	500
NTR	4	Pyrazinamide	590
		Pyrazinamide	394
		INH	350
		INH	234

3.2.3 Geographical distribution

The geographical distribution of the ten clusters identified in Figure 3.1 was mapped according to the medical facilities and towns in which they occurred (Figures 3.2 - 3.5). The distribution of the drug resistant clusters was also mapped (Figures 3.6 - 3.10), as was that of those in P.E. (Figure 3.11), Uitenhage and Despatch (Figure 3.12) and East London (Figure 3.13). The term cluster type will be used in this discussion when referring to the geographical distribution of clusters.

3.2.3.1 Total Eastern Cape population

The number of health regions, towns and medical facilities in which each cluster type occurs is set out in Table 3.14. B3 was the cluster type with the widest distribution, being found in 51% of the towns and in all three health regions. Cluster type A1, however, occurred in the largest number of medical facilities and in almost as many towns as did B3, while C3 was the least commonly occurring cluster type. Eighty percent of cluster types occurred in all three health regions surveyed. Cluster types A2 and B2 were not found in Region B. However, it should be noted that the smallest number of strains came from this region and that wider sampling might reveal the existence of these types of the organism in this region as well.

Figure 3.2 shows the geographical distribution of the three A cluster types - A1, A2 and A3 - amongst the towns surveyed. All three occurred in P.E., Uitenhage and East London. As regards the remaining urban centres, two A cluster types occurred in Despatch and one each in Grahamstown and King William's Town. However, fewer strains were obtained from these latter towns. Further sampling might reveal the presence of all three A cluster types in these centres as well.

Figure 3.2 further demonstrated that there were only three rural centres which contained all three A cluster types - Graaff-Reinet, Fort Beaufort and Stutterheim. The remaining rural centres contained only one or two A cluster types. It must be noted that amongst these were towns from which only one or two isolates were received. Further sampling would be needed to determine whether the other A cluster types exist in these centres or not. There were ten towns in which none of the A cluster types were found.

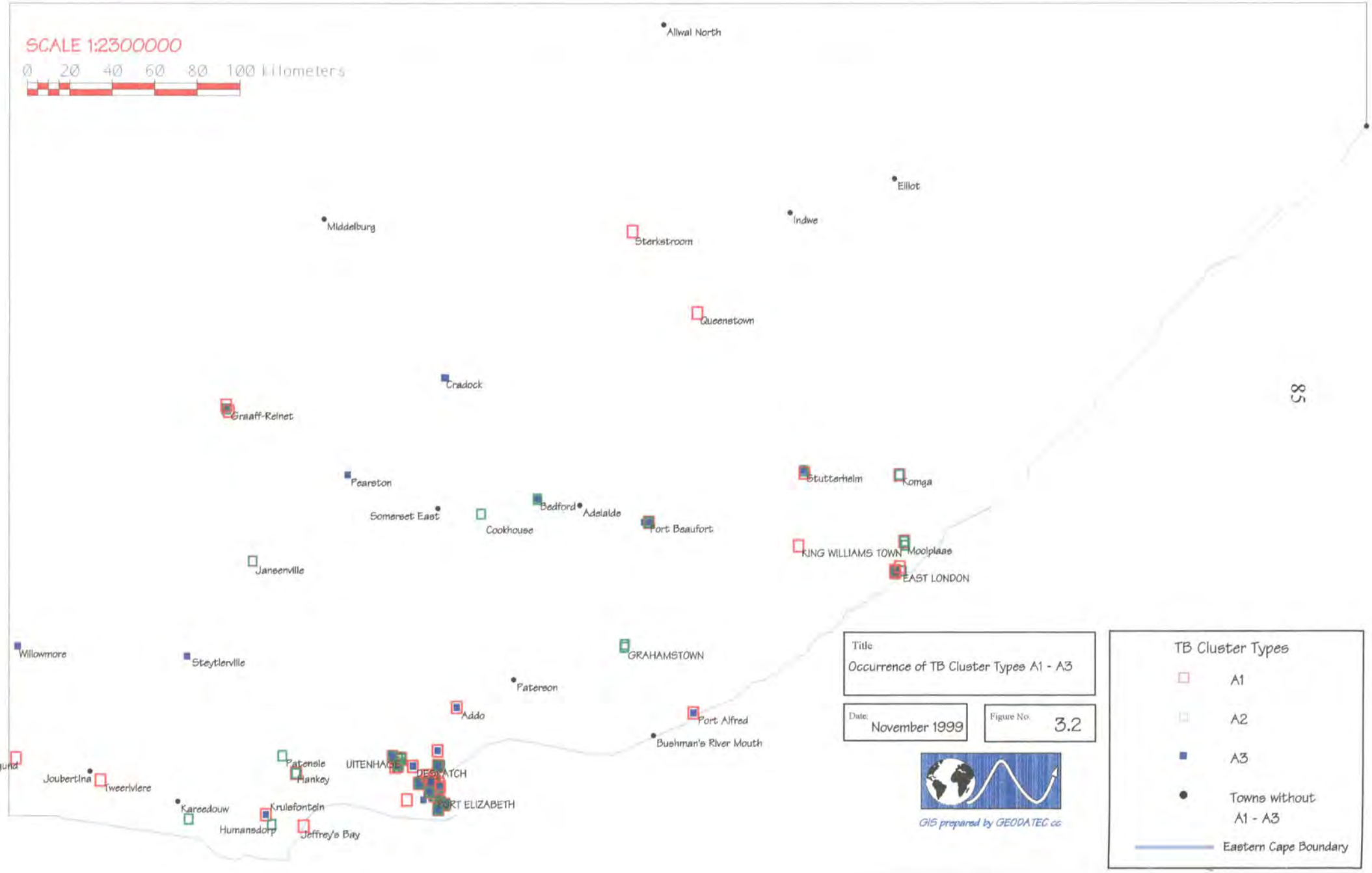
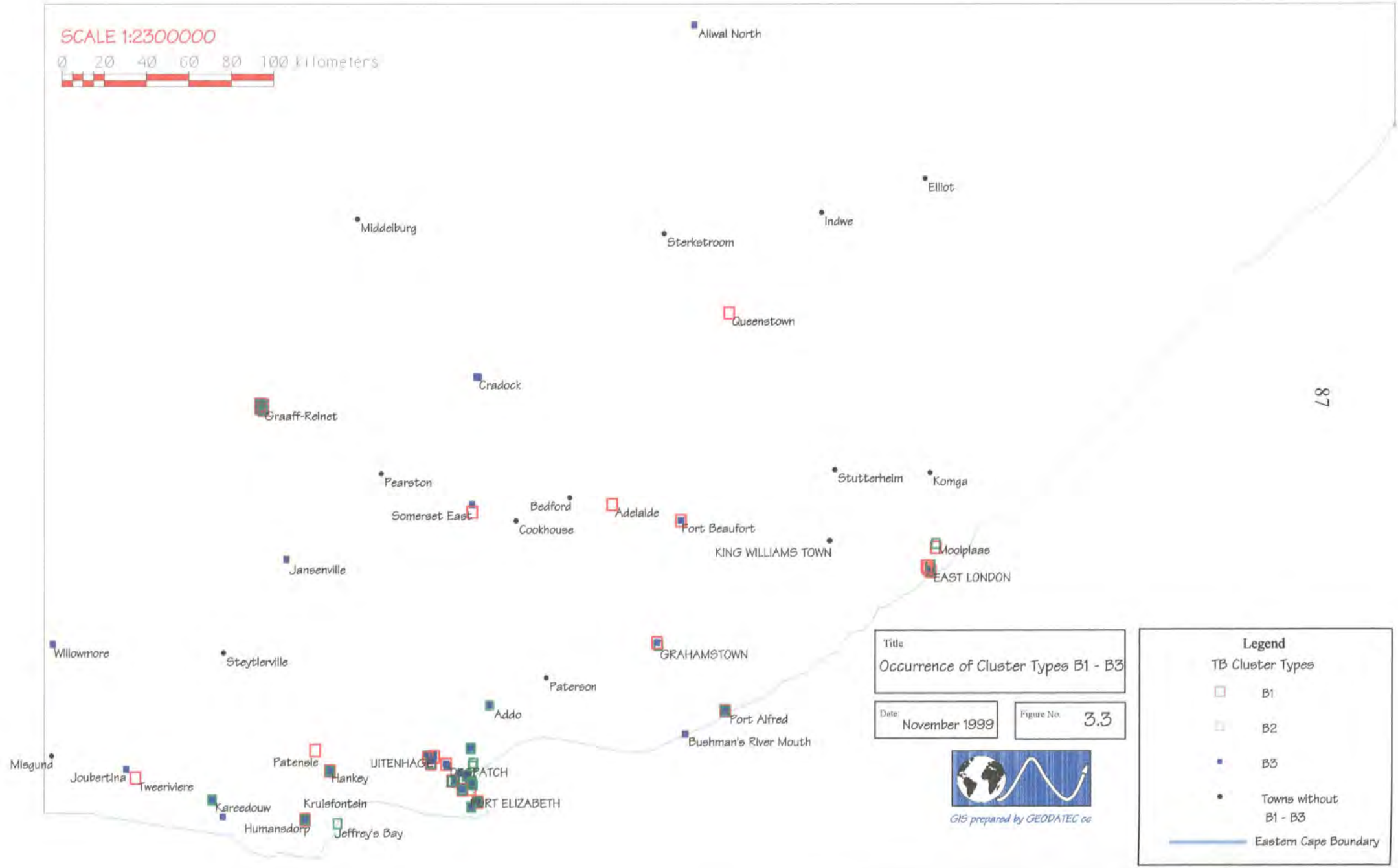

SCALE 1:2300000
0 20 40 60 80 100 kilometers
Allwal North
Elliot
Middelburg
Indwe
Sterkstroom
Queenstown
Cradock
Graaff-Reinet
Pearston
Somerset East
Bedford
Adelaide
Cookhouse
Port Beaufort
Stutterheim
Komga
KING WILLIAMS TOWN
Moolplaas
EAST LONDON
Jansenville
Willowmore
Steytlerville
GRAHAMSTOWN
Paterson
Addo
Port Alfred
Bushman's River Mouth
Misgund
Joubertina
Tweeriviere
Patensie
Hankey
UITENHAGE
DESPATCH
PORT ELIZABETH
Kareedouw
Krulefontein
Humansdorp
Jeffrey's Bay
Title
Occurrence of TB Cluster Types A1 - A3
Date:
November 1999
Figure No:
3.2
GIS prepared by GEODATEC cc
TB Cluster Types
A1
A2
A3
Towns without A1 - A3
Eastern Cape Boundary

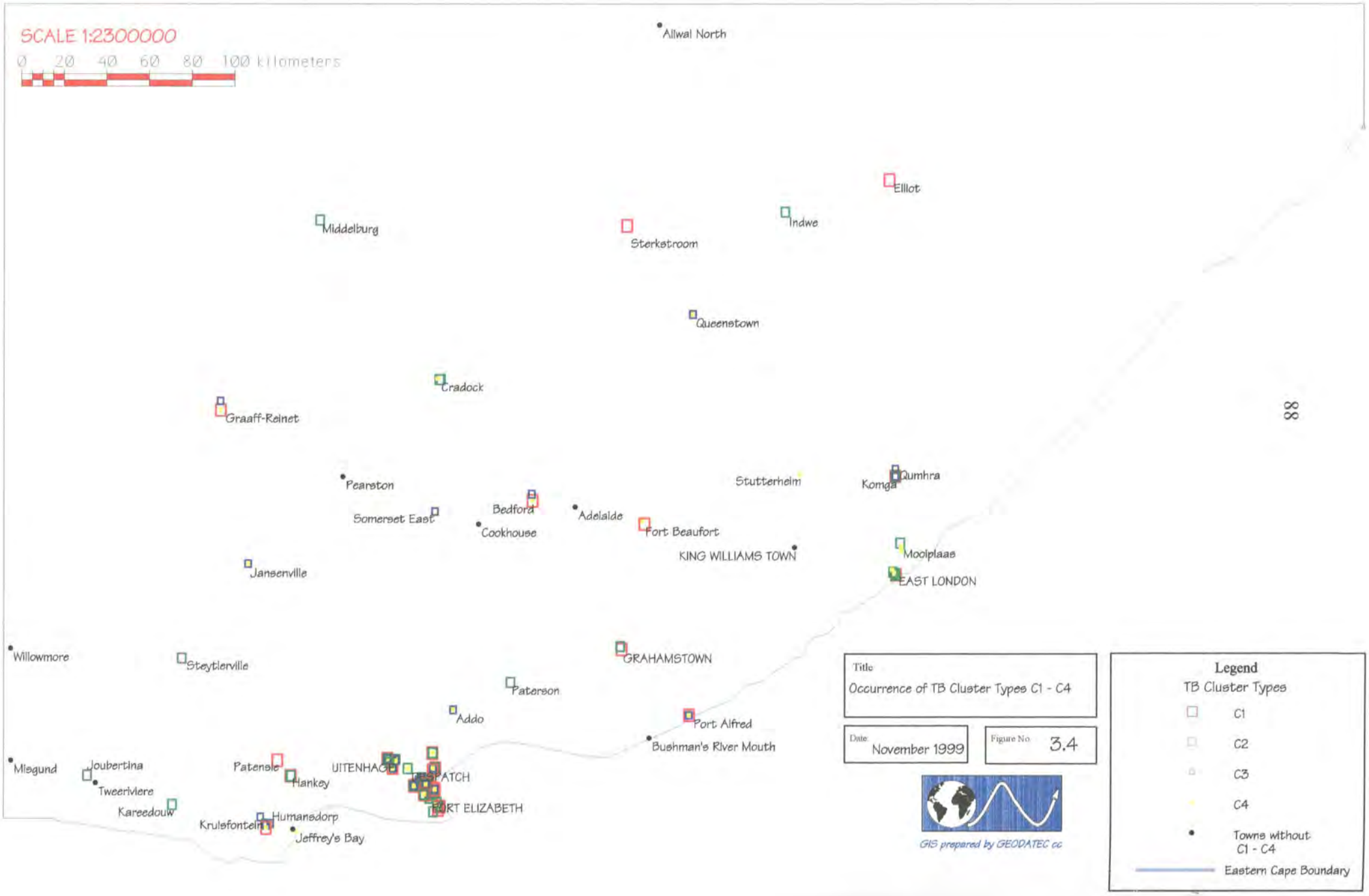
Table 3.14 Distribution of Eastern Cape cluster types

Cluster Type	Number of Strains	Number of Health Regions (n = 3)	Number of Towns (n = 39)	Number of Medical Facilities (n = 110)
A1	75	3	19	43
A2	46	2	15	35
A3	48	3	15	32
B1	36	3	13	23
B2	42	2	12	30
B3	67	3	20	37
C1	41	3	14	30
C2	48	3	16	34
C3	35	3	11	23
C4	64	3	14	37

As regards the distribution of B cluster types, all three were found in P.E., Uitenhage, East London and Grahamstown (Figure 3.3). Two cluster types occurred in Despatch and none in King William's Town from which only one isolate was received. As regards the rural areas, all three types were found in Graaff-Reinet, Hankey and Port Alfred. The remaining rural towns contained only one or two types. Once again, only one or two isolates had been obtained from six of these. None of the B cluster types occurred in 13 towns.

Only two urban centres - P.E. and Uitenhage - contained all four C cluster types (Figure 3.4). Three occurred in East London and two in Grahamstown and Despatch. As regards the rural centres, three types occurred in Graaff-Reinet, Port Alfred, Komga and Bedford. Only one or two C types were found in 18 rural towns. There were nine towns in which none of these types occurred.


SCALE 1:2300000
0 20 40 60 80 100 kilometers
Aliwal North
Elliot
Middelburg
Indwe
Sterkstroom
Queenstown
Cradock
Graaff-Reinet
Pearston
Stutterheim
Komga
Somerset East
Bedford
Adelaide
Cookhouse
Fort Beaufort
KING WILLIAMS TOWN
Moolplaas
Jansenville
EAST LONDON
Willowmore
Steytlerville
GRAHAMSTOWN
Paterson
Addo
Port Alfred
Bushman's River Mouth
Misgund
Joubertina
Tweeriviere
Patensie
Hankey
UITENHAGE
DESPATCH
PORT ELIZABETH
Kareedouw
Krulsfontein
Humansdorp
Jeffrey's Bay
Title
Occurrence of Cluster Types B1 - B3
Date
November 1999
Figure No.
3.3
GIS prepared by GEODATEC cc
Legend
TB Cluster Types
B1
B2
B3
Towns without B1 - B3
Eastern Cape Boundary


SCALE 1:2300000
0 20 40 60 80 100 kilometers
Allwal North
Elliot
Middelburg
Indwe
Sterkstroom
Queenstown
Cradock
Graaff-Reinet
88
Pearston
Stutterheim
Komga
Qumhra
Somerset East
Bedford
Adelaide
Cookhouse
Fort Beaufort
KING WILLIAMS TOWN
Moolplaas
Jansenville
EAST LONDON
Willowmore
Steytlerville
GRAHAMSTOWN
Paterson
Addo
Port Alfred
Bushman's River Mouth
Misgund
Joubertina
Patensie
UITENHAGE
Hankey
Tweerivliere
Kareedouw
Humansdorp
Krulsfontein
Jeffrey's Bay
Title
Occurrence of TB Cluster Types C1 - C4
Date
November 1999
Figure No.
3.4
GIS prepared by GEODATEC cc
Legend
TB Cluster Types
C1
C2
C3
C4
Towns without C1 - C4
Eastern Cape Boundary

These maps were then examined to determine whether it was possible to associate certain main routes of human movement with the transmission of specific genetic types. Of the national roads that traverse South Africa, only one - the N2 - runs through the length of the Eastern Cape, connecting the cities of P.E., Grahamstown, East London and Umtata to the coastal cities of Cape Town in the Western Cape and Durban in KwaZulu-Natal (Route 1, Table 3.15). Other main routes exist which link centres at the coast with those inland. These include the roads from P.E. north to Graaff-Reinet via Uitenhage and Jansenville (Route 2); from East London to Aliwal North via Stutterheim, Queenstown and Sterkstroom (Route 3); from East London to Graaff-Reinet via King William's Town, Fort Beaufort, Bedford and Somerset East (Route 4); and from Port Alfred through Grahamstown and Fort Beaufort to Queenstown (Route 5). Other important routes are the Paterson-Cookhouse-Cradock-Middelburg road (Route 6), and the Graaff-Reinet-Cradock-Queenstown-Elliot axis (Route 7).

The seven routes were examined to determine the total number of cluster types that occurred along each route, as well as those which occurred in all towns on a particular route (Table 3.15).

Table 3.15 Distribution of cluster types along Eastern Cape travel routes

Route Descriptor	Axis	Number of towns	Total number of cluster types	Cluster types in all towns
1	Humansdorp - P.E. - Grahamstown - East London (N2)	4	10	A2, B1, B2, B3, C1, C2
2	P.E. - Graaff-Reinet	4	10	A2, B3, C3, C4
3	East London - Aliwal North	5	10	(A1)*
4	East London - Graaff-Reinet	9	10	(A2, A3, B1)*
5	Port Alfred - Queenstown	4	10	B1
6	Paterson - Middelburg	4	5	(C2)*
7	Graaff-Reinet - Elliot	4	10	(C4)*

* Cluster types that occurred in most towns along route

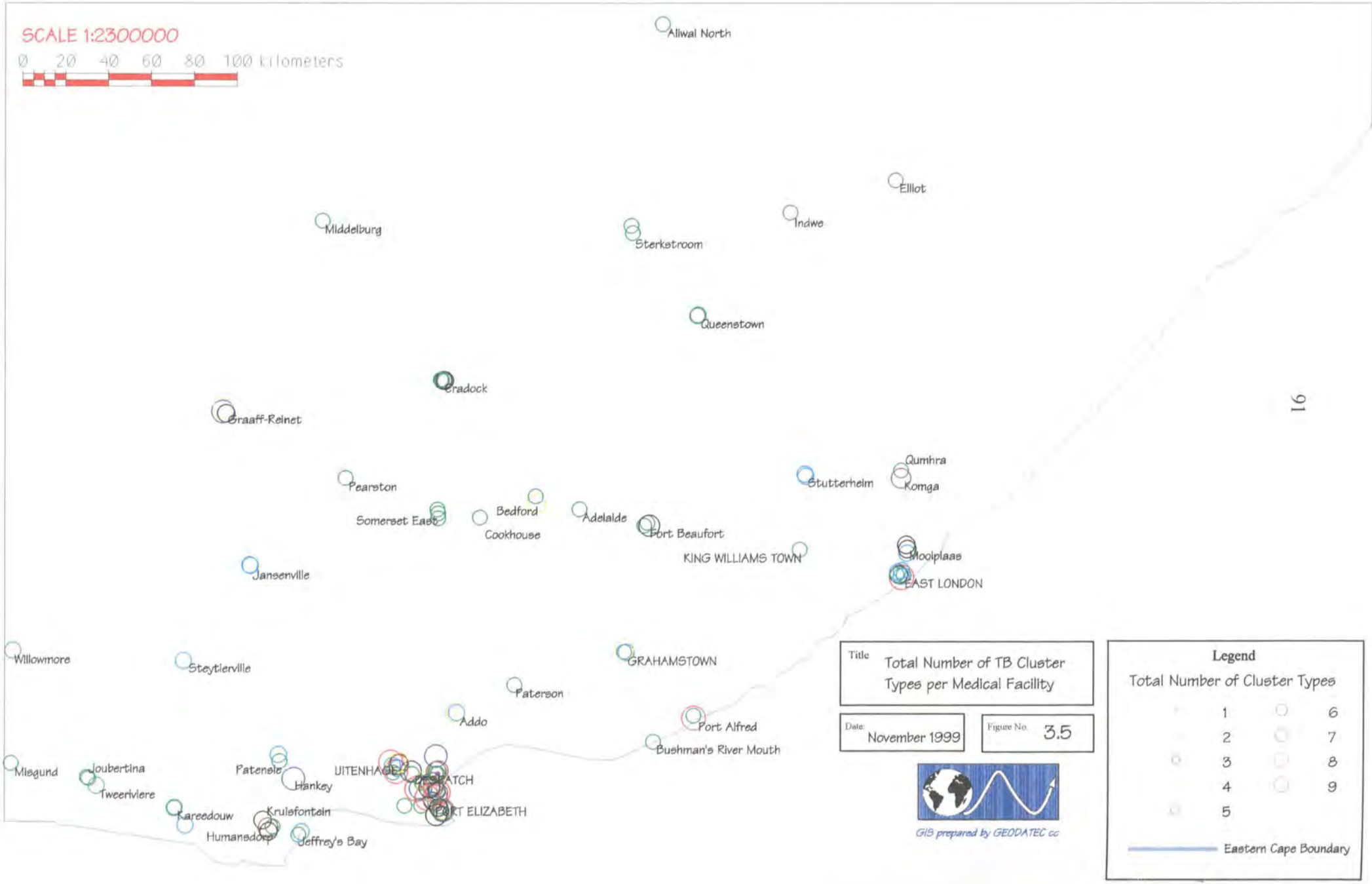
As regards the total number of cluster types occurring along a particular route, it was found that all ten types could be found in one or more towns along six of the seven routes. Very few isolates were received from the four towns along Route 6, which would account for the fact that only 50% of types were found on this route. Further sampling might reveal the existence of more types on this route. As regards the number of cluster types that were found in all towns on a particular route, as many as six were found along Route 1. Only 4 were found in all towns sampled on Route 2 and 1 cluster type was common to all the towns along Route 5. The other routes were then examined to determine which types occurred in the majority of towns. Cluster type A1 occurred in four of the five towns on Route 3, while cluster types A2, A3 and B1 occurred in five of the nine towns on Route 4. Cluster type C2 was predominant on Route 6, being found in two of the four towns. C4 was the predominant type on Route 7, occurring in three out of four towns.

The various medical facilities were then examined to determine the incidence of strains from the ten cluster types in each, giving particular attention to whether the facilities were situated in urban or rural areas (Table 3.16).

Table 3.16 Incidence of Eastern Cape cluster types

Number of Cluster Types (n = 10)	Number of Medical Facilities (n = 110)	Number of Urban Facilities (n=51)	Number of Rural Facilities (n=59)
9	4	4	0
8	4	3	1
7	7	5	2
6	4	3	1
5	8	6	2
4	7	4	3
3	11	5	6
2	22	8	14
1	43	13	30

Figure 3.5 shows the geographical distribution of this cluster type incidence amongst the medical facilities. Of the four clinics in which the largest number of cluster types occurred, three were situated in P.E. and one in East London. Of the three urban facilities in which eight types occurred, one was in P.E. and two in Uitenhage.


SCALE 1:2300000
0 20 40 60 80 100 kilometers
Aliwal North
Elliot
Indwe
Middelburg
Sterkstroom
Queenstown
Cradock
Graaff-Reinet
Qumhra
Komga
Stutterheim
Pearston
Somerset East
Bedford
Cookhouse
Adelaide
Fort Beaufort
KING WILLIAMS TOWN
Moolplaas
EAST LONDON
Jansenville
Willowmore
Steytlerville
GRAHAMSTOWN
Paterson
Addo
Port Alfred
Bushman's River Mouth
Misgund
Joubertina
Tweeriviere
Patensie
Hankey
UITENHAGE
DESPATCH
PORT ELIZABETH
Kareedouw
Kruisfontein
Humansdorp
Jeffrey's Bay
Title
Total Number of TB Cluster Types per Medical Facility
Date:
November 1999
Figure No.
3.5
GIS prepared by GEODATEC cc
Legend
Total Number of Cluster Types
1
2
3
4
5
6
7
8
9
Eastern Cape Boundary

The rural facility in which eight types occurred was the South African National Tuberculosis Association (SANTA) hospital in Port Alfred. The two rural facilities with seven cluster types were situated in Graaff-Reinet and Hankey. A relatively large number of cluster types (6) occurred in the SANTA Tuberculosis hospital in Fort Beaufort. In fact, more than fifty percent of the cluster types were found in six of the eight SANTA hospitals surveyed. It will be noticed that, apart from the medical facilities in the four rural centres mentioned above, the incidence of cluster types was higher in urban medical facilities than in most of the rural facilities. However, this might be due to the fact that the majority of strains in the sample came from urban facilities. Inclusion of larger numbers of strains from rural facilities may result in the discovery of additional cluster types in those areas.

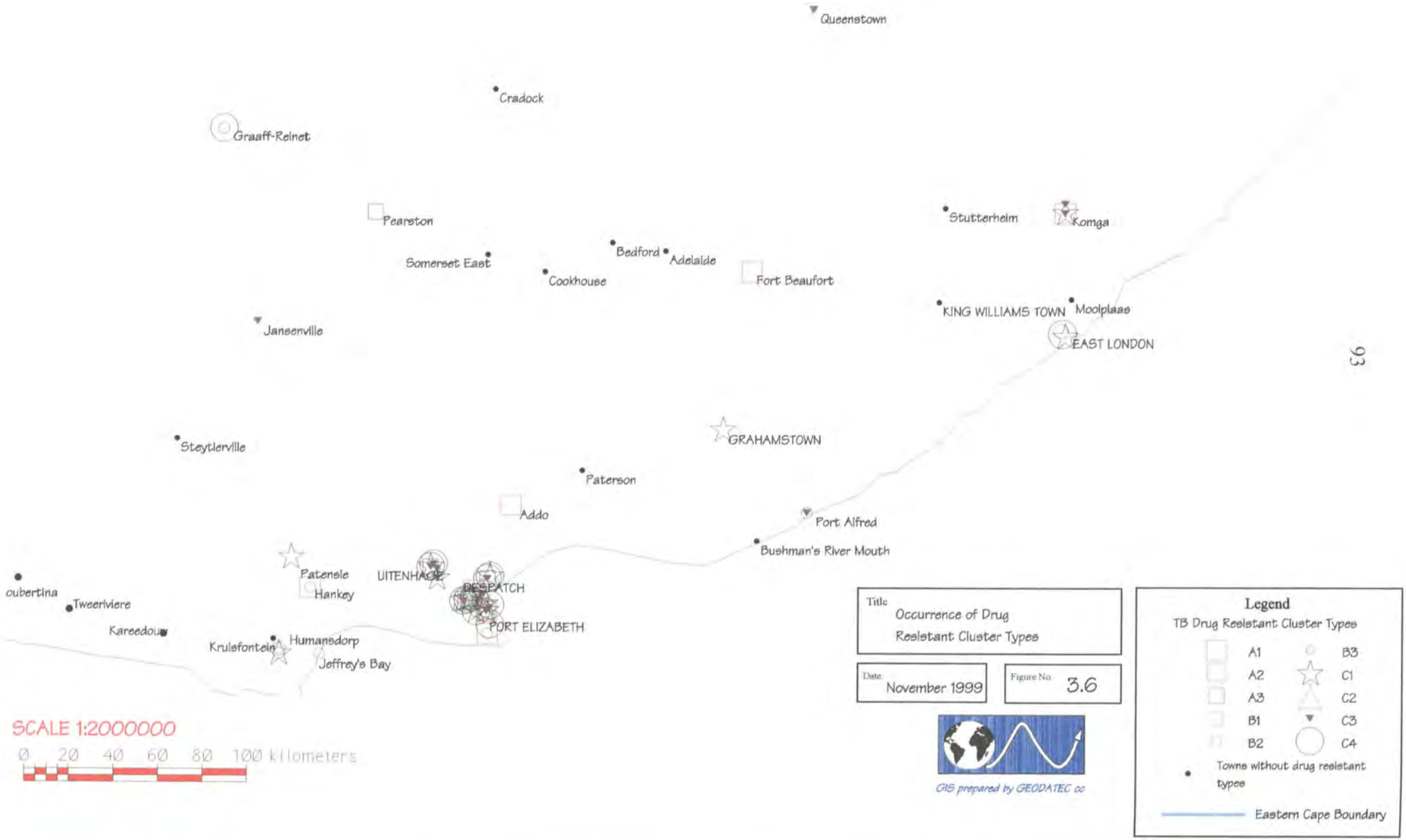
3.2.3.2 Drug resistant subpopulation

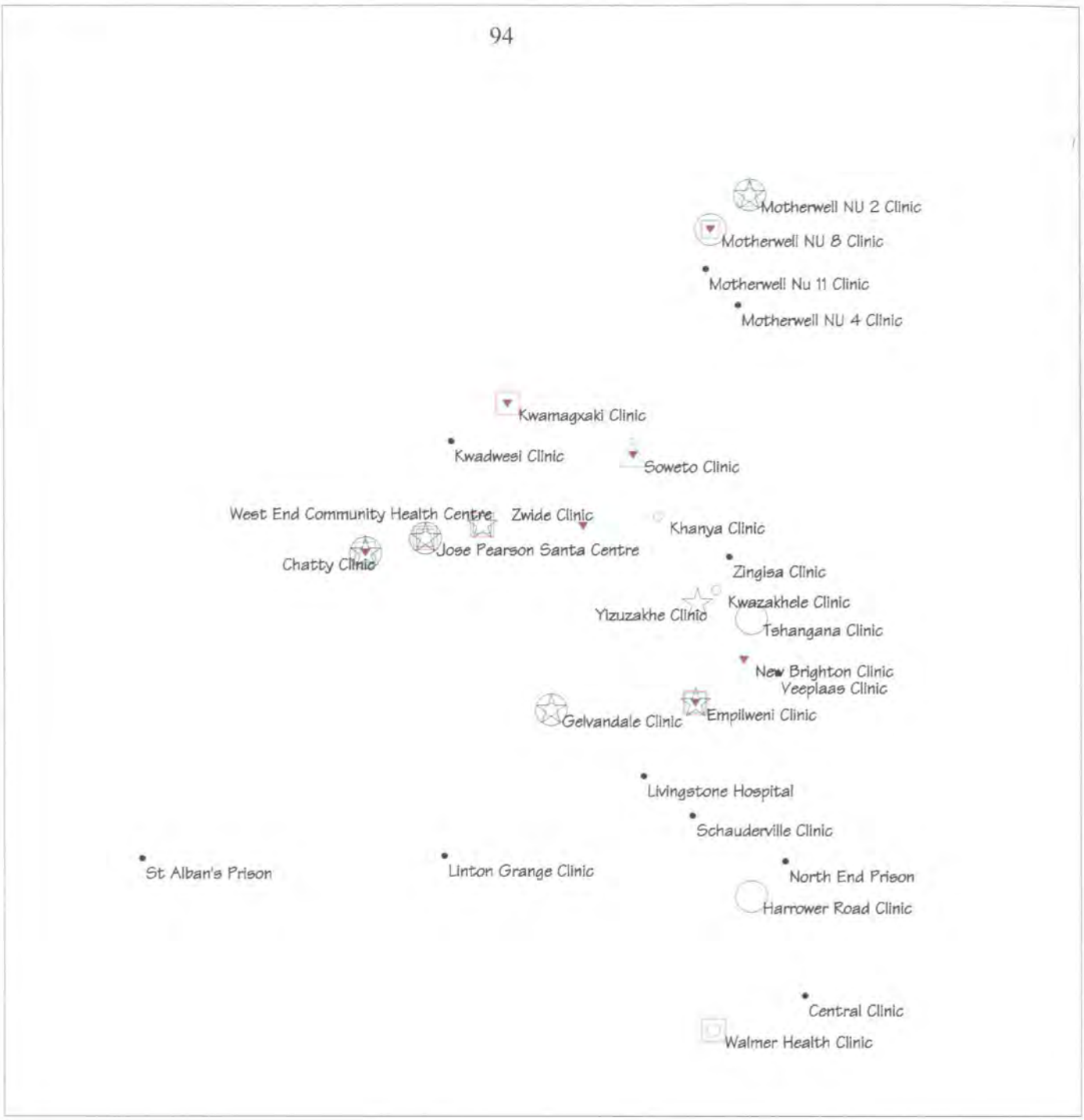
Drug resistant strains were located in 16 towns and 40 medical facilities in all three health regions. The number of health regions, towns and medical facilities in which the cluster types with resistant strains occurred is set out in Table 3.17.

Table 3.17 Distribution of drug resistant cluster types

Drug resistant cluster type	Number of drug resistant strains	Number of Health Regions (n = 3)	Number of Towns (n = 16)	Number of Medical Facilities (n = 40)
A1	12	2	5	9
A2	2	1	1	2
A3	5	1	3	5
B1	2	2	2	2
B2	9	1	4	7
B3	7	2	5	6
C1	16	2	7	14
C2	4	1	1	3
C3	17	3	7	16
C4	11	2	4	11

Cluster types C1 and C3 are of particular importance as they were distributed amongst the largest number of towns and medical facilities. Of the fourteen medical facilities in which type C1 occurred, ten were situated in urban and four in rural areas (Figure 3.6-3.9).


Queenstown
Cradock
Graaff-Reinet
Pearston
Stutterheim
Komga
Somerset East
Bedford
Adelaide
Cookhouse
Fort Beaufort
KING WILLIAMS TOWN
Moolplaas
Jansenville
EAST LONDON
GRAHAMSTOWN
Steytlerville
Paterson
Addo
Port Alfred
Bushman's River Mouth
Patensie
Hankey
UITENHAGE
DESPATCH
PORT ELIZABETH
oubertina
Tweeriviere
Kareedou
Kruisfontein
Humansdorp
Jeffrey's Bay
SCALE 1:2000000
0 20 40 60 80 100 kilometers
Title
Occurrence of Drug Resistant Cluster Types
Date:
November 1999
Figure No.
3.6
GIS prepared by GEODATEC cc
Legend
TB Drug Resistant Cluster Types
A1
A2
A3
B1
B2
B3
C1
C2
C3
C4
Towns without drug resistant types
Eastern Cape Boundary



SCALE 1:150000

0 1 2 3 4 5 kilometers

Title
Port Elizabeth Drug Resistant Medical Facilities: All 10 Cluster Types

Date: November 1999

Figure No. 3.7

GIS by GEODATEC cc

Legend

TB Cluster Types

A1	B3
A2	C1
A3	C2
B1	C3
B2	C4

Medical facilities without drug resistant types

SCALE 1:50000

0 1 2 Kilometers

Title
Uitenhage & Despatch Drug Resistant Medical Facilities: All 10 Cluster Types

Date: November 1999

Figure No. 3.8

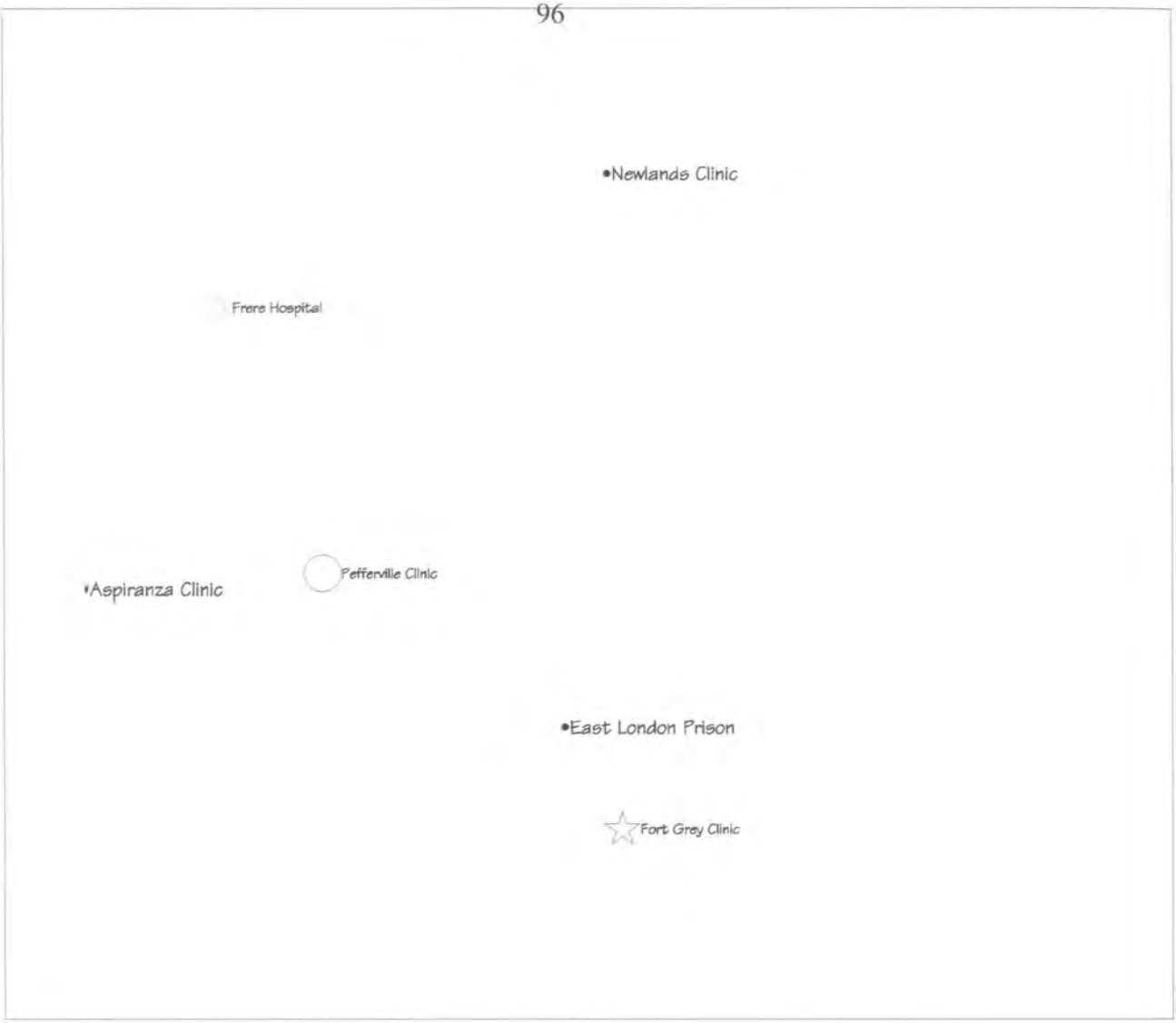
GIS by GEODATEC cc

Legend

TB Cluster Types

	A1		B3
	A2	☆	C1
	A3		C2
	B1	▼	C3
	B2	○	C4

• Medical facilities without drug resistant types



SCALE 1:30000

0 0.5 1 1.5 2 kilometers

Title
East London Drug Resistant Medical Facilities: All 10 Cluster Types

Date: November 1999

Figure No. 3.9

GIS by GEODATEC cc

Legend

TB Cluster Types

A1	B3
A2	C1
A3	C2
B1	C3
B2	C4

Medical facilities without drug resistant types

Type C3 occurred in eleven urban and five rural facilities (Figure 3.6 - 3.8). Resistant strains of types C1 and C3 were found in seven P.E. medical facilities each, although only two medical facilities contained both types (Figure 3.7). Type C1 occurred in two Uitenhage facilities and C3 in four, while two contained both types (Figure 3.8). Only type C1 was found in East London, in one medical facility (Figure 3.9).

The incidence of resistant cluster types amongst medical facilities in this geographical area is set out in Table 3.18.

Table 3.18 Incidence of drug resistant cluster types

Number of Resistant Cluster Types (n = 5)	Number of Medical Facilities (n = 40)	Number of Urban Facilities (n=26)	Number of Rural Facilities (n=14)
5	2	2	0
4	3	3	0
3	4	3	1
2	10	5	5
1	21	13	8

The two clinics with the largest number of resistant types were situated in P.E. (Figure 3.10). The three clinics with four resistant types were located in P.E. and Uitenhage, and the three urban facilities with three resistant types occurred in P.E. Most medical facilities (78%) had only one or two types. However, usually only one or two strains had been received from such facilities.

The distribution of antibiotic phenotypes and resistant cluster types was then examined with particular reference to urban and rural location (Table 3.19). This showed that all ten resistant types and 80% of the antibiotic phenotypes occurred in urban areas, while eight resistant types and 50% of the antibiotic phenotypes were found in the rural areas. It was surprising to find such a wide representation of resistant types, as well as half of the antibiotic phenotypes, in the rural areas in view of the fact that only 26% of resistant strains were from such locations.

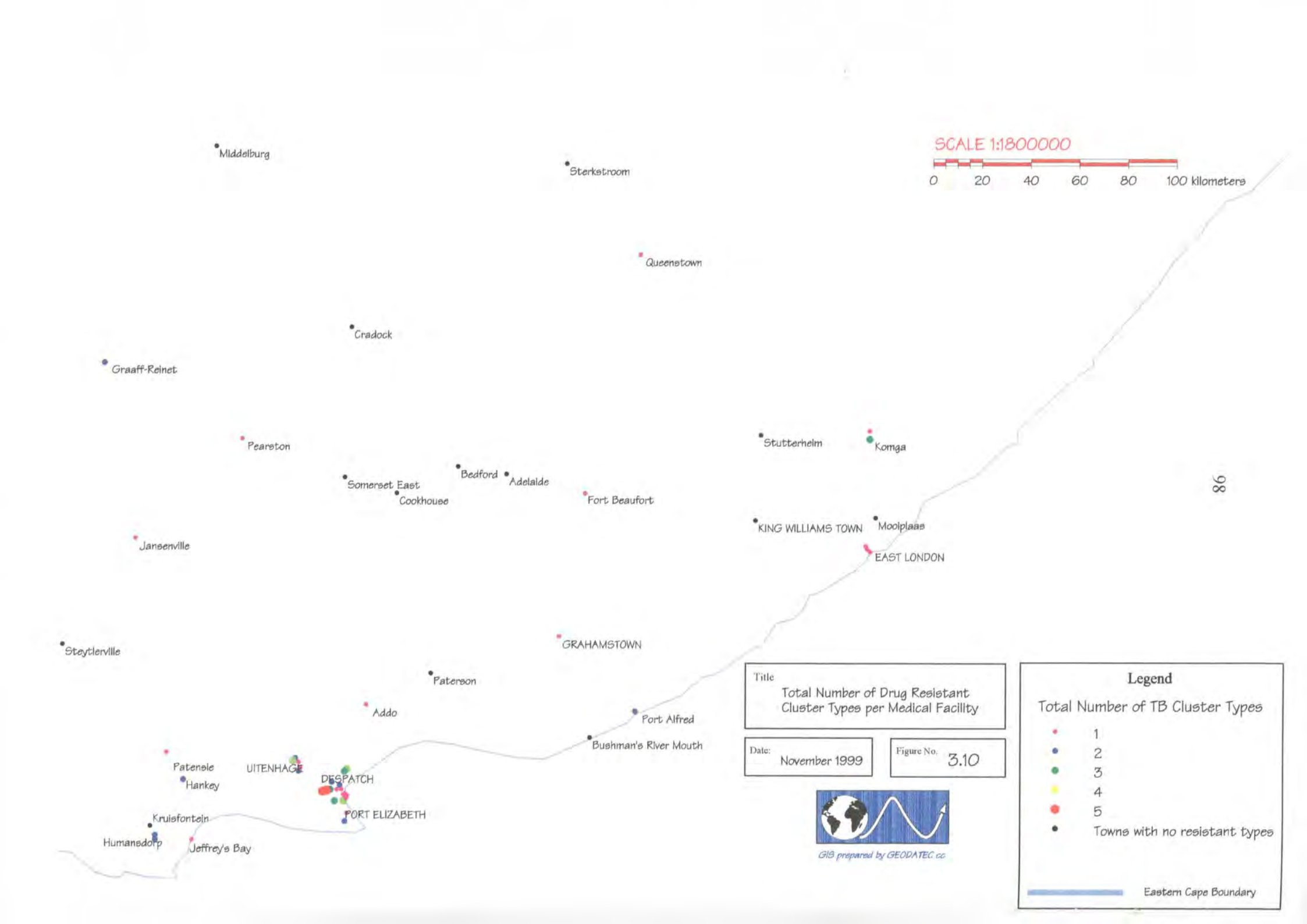

SCALE 1:1800000
0 20 40 60 80 100 kilometers
Middelburg
Sterkstroom
Queenstown
Cradock
Graaff-Reinet
Pearston
Somerset East
Cookhouse
Bedford
Adelaide
Fort Beaufort
Stutterheim
Komga
KING WILLIAMS TOWN
Moolplaas
EAST LONDON
Jansenville
Steytlerville
GRAHAMSTOWN
Paterson
Addo
Port Alfred
Bushman's River Mouth
Patensie
Hankey
UITENHAGE
DESPATCH
PORT ELIZABETH
Kruisfontein
Humansdorp
Jeffrey's Bay
86
Title
Total Number of Drug Resistant Cluster Types per Medical Facility
Date:
November 1999
Figure No.
3.10
GIS prepared by GEODATEC cc
Legend
Total Number of TB Cluster Types
1
2
3
4
5
Towns with no resistant types
Eastern Cape Boundary

Table 3.19 Urban and rural distribution of drug resistant cluster types and antibiotic phenotypes

Urban Centres		Rural Centres	
Resistant Type	**Antibiotic Phenotypes**	**Resistant Type**	**Antibiotic Phenotypes**
A1	EC2, EC6, EC10, EC13, EC15	A1	EC2, EC4, EC7, EC13
A2	EC1, EC2	A2	-
A3	EC1, EC2, EC10, EC13	A3	EC20
B1	EC2	B1	EC4
B2	EC2, EC3, EC5	B2	EC2, EC4, EC19
B3	EC4, EC12, EC16, EC17	B3	EC1
C1	EC1, EC2, EC4, EC7, EC9	C1	EC2, EC14
C2	EC4, EC5, EC11, EC16	C2	-
C3	EC1, EC2, EC4, EC8, EC12	C3	EC1, EC2, EC4, EC18
C4	EC2, EC3, EC7	C4	EC5

3.2.3.3 Port Elizabeth subpopulation

This subpopulation consisted of 214 strains from 28 medical facilities situated in 17 suburbs or townships. Table 3.20 and Figure 3.11 show the distribution of cluster types amongst the various medical facilities in P.E. Type A1 strains occurred in the largest number of P.E. facilities (68%), with the least commonly occurring type being B1. The other eight types were found in from 29 to 57% of the P.E. facilities.

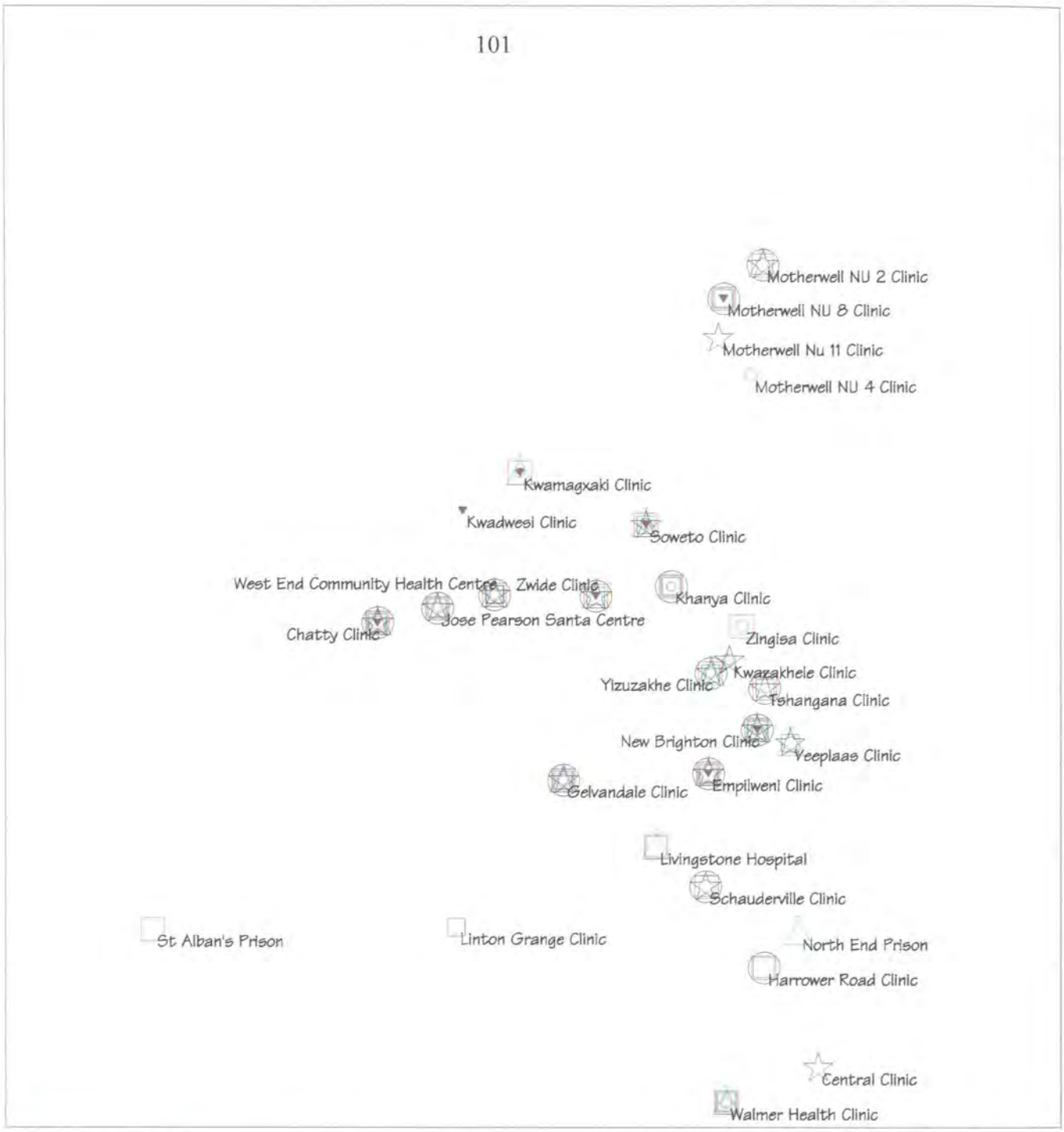
A high proportion of cluster types were found in eight of the 28 P.E. facilities, with the incidence amongst the remaining 20 ranging from 10 to 60% (Table 3.21 and Figure 3.11).

Table 3.20 Distribution of cluster types in Port Elizabeth

Cluster Type	Number of medical facilities (n=28)	Number of strains (n=214)
A1	19	42
A2	12	19
A3	15	29
B1	3	3
B2	14	22
B3	11	20
C1	16	23
C2	14	21
C3	8	11
C4	14	24

Table 3.21 Incidence of cluster types in Port Elizabeth medical facilities

Medical Facility	Suburb/Township	Number of Cluster Types (n=10)	Number of Strains (n=214)
Gelvandale Clinic	Gelvandale	9	11
Chatty Clinic	Chatty	9	18
New Brighton Clinic	New Brighton	9	22
Empilweni Hospital	New Brighton	8	12
West End Community Health Centre	Bethelsdorp	7	22
Jose Pearson SANTA Centre	Bethelsdorp	7	18
Soweto Clinic	Zwide	7	13
Yizuzakhe Clinic	Kwa-Zakhele	7	11



SCALE 1:150000

0 2 4 6 8 10 Kilometers

Title
Port Elizabeth Medical Facilities:
All 10 Cluster Types

Date: November 1999

Figure No. 3.11

GIS by GEODATEC cc

Legend

TB Cluster Types

A1		B3	
A2		C1	
A3		C2	
B1		C3	
B2		C4	

The drug resistant strains from P.E. have already been examined to a certain degree in Section 3.2.3.2. At this point, the distribution of drug resistant types amongst these medical facilities is examined (Table 3.22). Strains of cluster types C1, C3 and C4 were predominant in this city, with one clinic containing all three.

Table 3.22 Distribution of drug resistant cluster types in Port Elizabeth

Resistant Cluster Type (n=9)	Number of Medical Facilities (n = 17)	Number of Strains (n = 48)
A1	5	8
A2	2	2
A3	3	3
B2	4	5
B3	2	2
C1	7	9
C2	3	4
C3	7	8
C4	7	7

Cluster types C1 and C3, which contained the major proportion of drug resistant strains, were distributed amongst a total of 12 medical facilities (71%) in eight suburbs and townships of P.E. Only type B1 resistant strains did not occur in P.E. The incidence of resistant cluster types was highest in the SANTA hospital (50%), where 20% of antibiotic phenotypes were present.

Although only 81 isolates were received from ten medical facilities in the Uitenhage-Despatch area, all ten cluster types were found there as well. Table 3.23 and Figure 3.12 show the distribution of types amongst those medical facilities. Type B3 strains occurred in the largest number of medical facilities, with B2 and C1 being the least common types in this area. Seventy percent of types occurred in the Orsmond SANTA hospital in Uitenhage.

Table 3.23 Distribution of cluster types in Uitenhage and Despatch

Cluster Type	Number of Medical Facilities (n =10)	Number of strains (n=81)
A1	6	10
A2	6	6
A3	4	4
B1	5	10
B2	2	4
B3	7	11
C1	2	3
C2	4	9
C3	4	8
C4	5	16

The distribution of the small sample of 34 East London strains is set out in Table 3.24 and Figure 3.13. Types A1, B1, C2 and C4 strains were the most widespread, with all four being found in the Fort Grey SANTA hospital. Type C3 was not found in the facilities sampled in this city.

Table 3.24 Distribution of cluster types in East London

Cluster Type	Number of Medical Facilities (n = 7)	Number of Strains (n = 34)
A1	3	4
A2	2	4
A3	1	1
B1	3	6
B2	2	4
B3	2	4
C1	1	2
C2	3	3
C3	0	0
C4	3	6

SCALE 1:60000

0 1 2 kilometers

Title
Uitenhage & Despatch Medical Facilities: All 10 Cluster Types

Date: November 1999

Figure No. 3.12

GIS by GEODATEC cc

Legend

TB Cluster Types

	A1		B3
	A2	☆	C1
	A3		C2
	B1	▼	C3
	B2		C4

SCALE 1:30000

0 0.5 1 1.5 2 Kilometers

Title
East London Medical Facilities:
All 10 Cluster Types

Date:
November 1999

Figure No.
3.13

GIS by GEODATEC cc

Legend

TB Cluster Types

A1
A2
A3
B1
B2
B3
C1
C2
C3
C4

CHAPTER 4

GENETIC DIVERSITY OF *M. TUBERCULOSIS* IN THE PROVINCE OF KWAZULU-NATAL

The analysis of the genetic diversity of this sample was undertaken in three sections. In the first, the 57 strains from KwaZulu-Natal were analysed in terms of drug resistance patterns, genetic diversity, population structure and geographical distribution. The second section comprised a comparative analysis of the KwaZulu-Natal population and the Eastern Cape drug resistant subpopulation, giving attention to population structure and geographical distribution. In the last section, all strains from the two provincial populations were compared to determine the degree of genetic difference between them. The same analytical methods and computer programmes were used for this sample as were used for the cluster analysis and AMOVA of the Eastern Cape population. However, GIS technology was not used due to the small size of the sample.

4.1 KwaZulu-Natal Population

Bacterial culture, DNA extraction, RAPD-PCR conditions and gel electrophoresis were performed on 57 isolates, all of which were drug resistant, in exactly the same way as with the Eastern Cape population. However, only two of the four primers were used to generate RAPD markers from these isolates - OPA1-13 and NTR. The RAPD profiles were combined to form a composite profile for each strain. A dendrogram was generated in GelCompar using the same settings as for the Eastern Cape population, which is essential for comparison of the two databases. The dendrogram then formed the basis for cluster analysis, AMOVA and an analysis of the geographical distribution of these strains. However, the first consideration was an analysis of the drug resistance patterns of this population, with a view to comparing them with those of the Eastern Cape sample. .

4.1.1 Drug resistance patterns

Information was received for only six of the seven antibiotics used in the Eastern Cape, as pyrazinamide had not been tested by the laboratory that supplied the KwaZulu-Natal isolates. Table 4.1 indicates the incidence of resistance to individual antibiotics.

Table 4.1 Incidence of antibiotic resistance in the KwaZulu-Natal population

Antibiotic	Number of resistant isolates (%) (n=57)	Comparative incidence ($p \leq 0.05$)
Isoniazid*	51 (91)	0.1568
Rifampicin*	53 (95)	<0.0001
Ethionamide	18 (32)	0.0201
Thiacetazone	14 (25)	0.1143
Streptomycin*	30 (54)	<0.0001
Ethambutol*	13 (23)	<0.0001

* First line antibiotics

Comparison with the Eastern Cape drug resistant subpopulation indicated that the incidence of resistance to rifampicin, ethionamide, streptomycin and ethambutol was significantly higher in the KwaZulu-Natal population (Table 4.1). This could possibly be due to differences in treatment policies, especially as regards first line drugs. Thiacetazone resistance was slightly higher in the KwaZulu-Natal population, but not significantly so. INH resistance was at about the same level in both populations.

Thirteen antibiotic phenotypes were found in this population, with a total of seven phenotypes being common to both populations (Table 4.2). Thirteen phenotypes occurred only in the Eastern Cape subpopulation (Table 3.4) and six only in the KwaZulu-Natal population. The comparative incidence rates of the common phenotypes in these two populations is given in Table 4.3. The incidence of phenotype N2, which occurs in the Eastern Cape subpopulation as EC4, is approximately the same in both populations. Phenotype N8, however, occurred significantly more frequently in the Eastern Cape population (p <0.0001). Phenotypes N5 and N7, which also occur in the Eastern Cape as EC12 and EC7, were found more frequently in the KwaZulu-Natal population, although not significantly so. Phenotype N3 was the second most frequently occurring in the KwaZulu-Natal population, but was not found at all in the Eastern Cape subpopulation. The frequency of this phenotype seems to indicate greater use of streptomycin in KwaZulu-Natal than in the Eastern Cape.

Table 4.2 Antibiotic phenotypes in the KwaZulu-Natal population

Phenotype Name	Phenotype Description*	Number of isolates
N1 (EC17)	INH, rifampicin, streptomycin & thiacetazone	1
N2 (EC4)	INH & rifampicin	11
N3	INH, rifampicin & streptomycin	10
N4 (EC11)	INH, rifampicin, streptomycin, thiacetazone & ethionamide	5
N5 (EC12)	INH, rifampicin, thiacetazone & ethionamide	3
N6	INH, rifampicin, streptomycin, thiacetazone, ethionamide & ethambutol	4
N7 (EC7)	Rifampicin	5
N8 (EC2)	INH	4
N9	INH, rifampicin & ethambutol	3
N10	INH, rifampicin, streptomycin & ethambutol	4
N11 (EC15)	INH, rifampicin & thiacetazone	1
N12	INH, rifampicin, streptomycin & ethionamide	4
N13	INH, rifampicin, streptomycin, ethionamide & ethambutol	2

* isolates are resistant to the antibiotic(s) indicated in the phenotype and sensitive to the rest

Table 4.3 Incidence of antibiotic phenotypes common to both KwaZulu-Natal and the Eastern Cape

Phenotype Name	Incidence in KwaZulu-Natal (%)	Incidence in the Eastern Cape (%)
N1 (EC17)	1.7	1.2
N2 (EC4)	19.0	18.0
N4 (EC11)	9.0	1.2
N5 (EC12)	7.0	2.4
N7 (EC7)	9.0	3.5
N8 (EC2)	7.0	35.0
N11 (EC15)	1.7	1.2

4.1.2 Cluster Analysis

4.1.2.1 RAPD profiling

The dendrogram of the KwaZulu-Natal strains was rooted at a similarity index of 80.4%, with a phenon line placed at 85% similarity defining three cluster groups (D, E and F) (Figure 4.1). The cluster phenon line placed at a similarity index of 89% gave an index of discrimination of 98% using Simpson's Index of Diversity, which indicated that the six clusters thus defined were valid groupings. Table 4.4 sets out the various similarity indices for the cluster groups and clusters in this population.

Table 4.4 Similarity indices of cluster groups and clusters in the KwaZulu-Natal population

Cluster Group	Similarity index of branching point (%)	Similarity within cluster group (%)	Cluster	Similarity index of branching point (%)	Similarity within cluster (%)
D	84.5	87 - 100			
			D1	88.5	90.8 - 97.8
			D2	88.5	93.5 - 100
			D3	87	94 - 100
E	84.5	88.6 - 100			
			E1	88.4	93 - 100
			E2	88.4	95.7 - 100
F	80.4	91 - 99			
			F1	80.4	91 - 99

Each cluster was examined to determine the total number of strains in each, the proportion of the total population this constituted and the urban-rural ratio (Table 4.5).

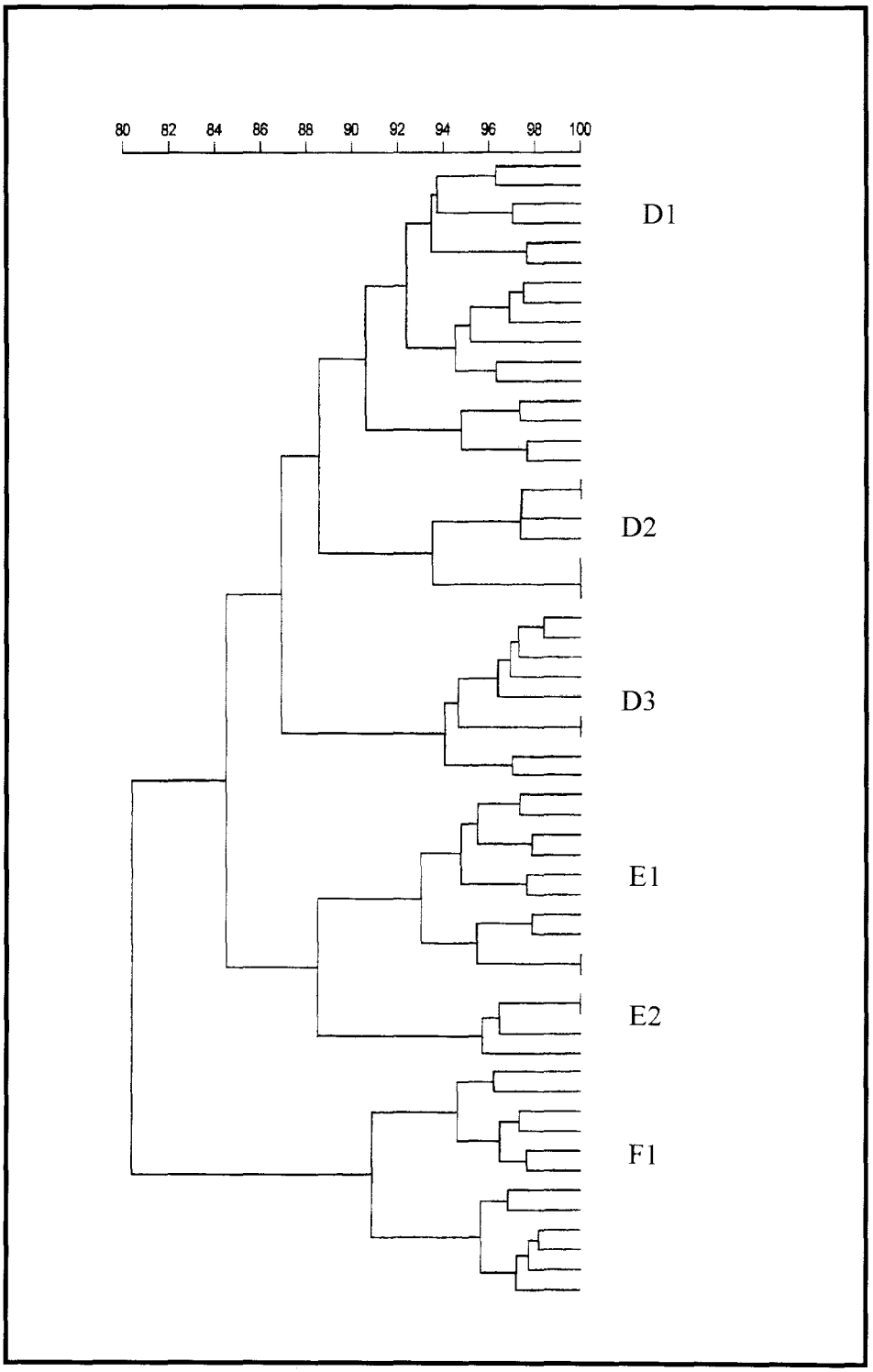


Figure 4.1 Cluster analysis of the KwaZulu-Natal population with the Ward algorithm

Table 4.5 Composition of cluster groups and clusters in the KwaZulu-Natal population

Cluster Group	Cluster	Number of strains	Proportion of population (%)	Number of Urban Strains	Number of Rural Strains
D		**31**	**54**	**29**	**2**
	D1	16	28	15	1
	D2	7	12	7	0
	D3	8	14	7	1
E		**14**	**25**	**11**	**3**
	E1	10	18	7	3
	E2	4	7	4	0
F		**12**	**21**	**12**	**0**
	F1	12	21	12	0

Cluster D1 was the predominant type in this population, while the smallest number of strains belonged to cluster E2. There were five small groups with identical RAPD profiles, comprising a total of eleven strains. Two occurred in cluster D2, one in cluster D3, one in cluster E1 and one in cluster E2. There were a further 46 unique profiles, resulting in a total of 51 RAPD profiles amongst the 57 isolates. The *M. tuberculosis* type strain, H37Rv, grouped in cluster D3.

The distribution of the thirteen antibiotic profiles amongst the six clusters indicated that strains with different antibiotic phenotypes were often closely related genetically, as was seen in the Eastern Cape population (Table 4.6).

Table 4.6 Distribution of KwaZulu-Natal antibiotic phenotypes

Cluster	Phenotypes	Number of phenotypes	Number of strains
D1	N1, N2, N3, N4, N5, N6, N7, N8, N9, N10, N12, N13	12	16
D2	N2, N4, N5, N8, N9	5	7
D3	N2, N3, N5, N8, N10, N12	6	8
E1	N2, N3, N5, N7, N9, N11, N13	7	10
E2	N2, N3, N6	3	4
F1	N2, N3, N4, N6, N7, N10	6	12

In contrast to the Eastern Cape resistant subpopulation, significantly more of the KwaZulu-Natal strains were resistant to between four and six antibiotics (p=0.03). The clusters were examined to determine whether these strains correlated with specific clustering of these strains occurred. In clusters D1, D3 and E2, the majority of strains were resistant to between four and six antibiotics, but the p values were not significant. There was a non-significant predominance of strains resistant to fewer antibiotics in clusters E1 and F1, and cluster D2 had approximately fifty percent of each group.

4.1.2.2 Comparison between RAPD and RFLP typing

Forty of the isolates in this population had initially been typed in the laboratories of the Department of Medical Microbiology at the University of Natal, using the IS*6110*-RFLP technique. TIF files created from the autoradiographs were processed in the same way as the RAPD profiles with all GelCompar settings being the same as for the RAPD profiles. The dendrogram, generated using the Dice similarity coefficient and the Ward cluster algorithm, was rooted at a similarity index of 61.8% with strains separating into two distinct cluster groups, G and H. (Figure 4.2). There were five groups of fifteen strains with identical RFLP profiles - three in group G and two in group H.

A dendrogram was also generated of the RAPD profiles of these 40 strains (Figure 4.3). The dendrogram was rooted at 86.2% similarity, with the strains also separating into two cluster groups, I and J. There were four groups of eight strains with identical RAPD profiles. However, three out of the four groups showed RFLP differences. This suggests that the rate of IS sequence variation is significantly higher than that of RAPDs.

A comparative analysis was performed on the groups comprising these two dendrograms, with a view to determining the degree of similarity in typing of the two techniques (Table 4.7).

Table 4.7 Comparison of RFLP and RAPD typing

RFLP Group	Number of strains	RAPD Group
G	11	I
	15	J
H	5	I
	9	J

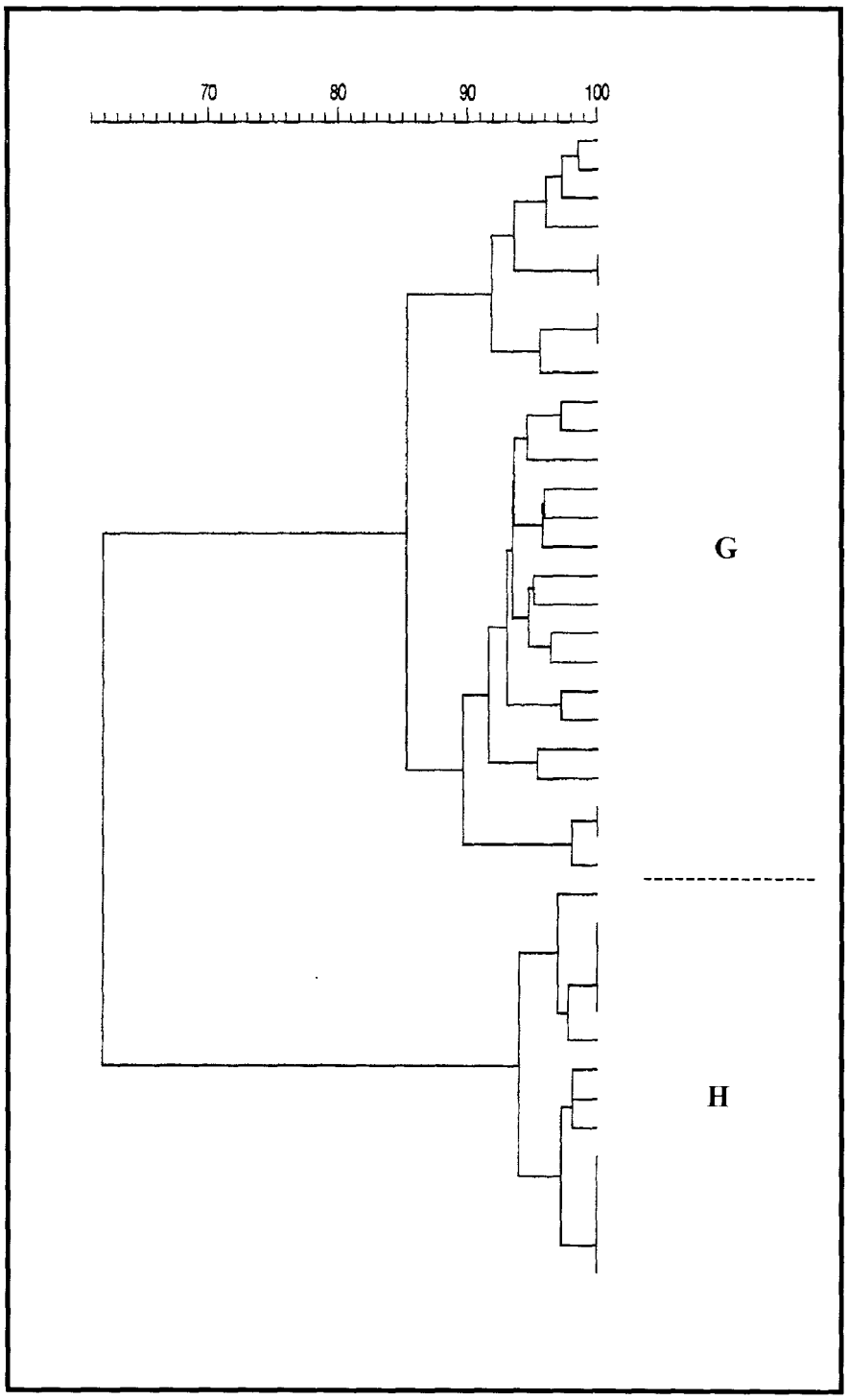


Figure 4.2 Cluster analysis of the RFLP profiles of 40 KwaZulu-Natal strains

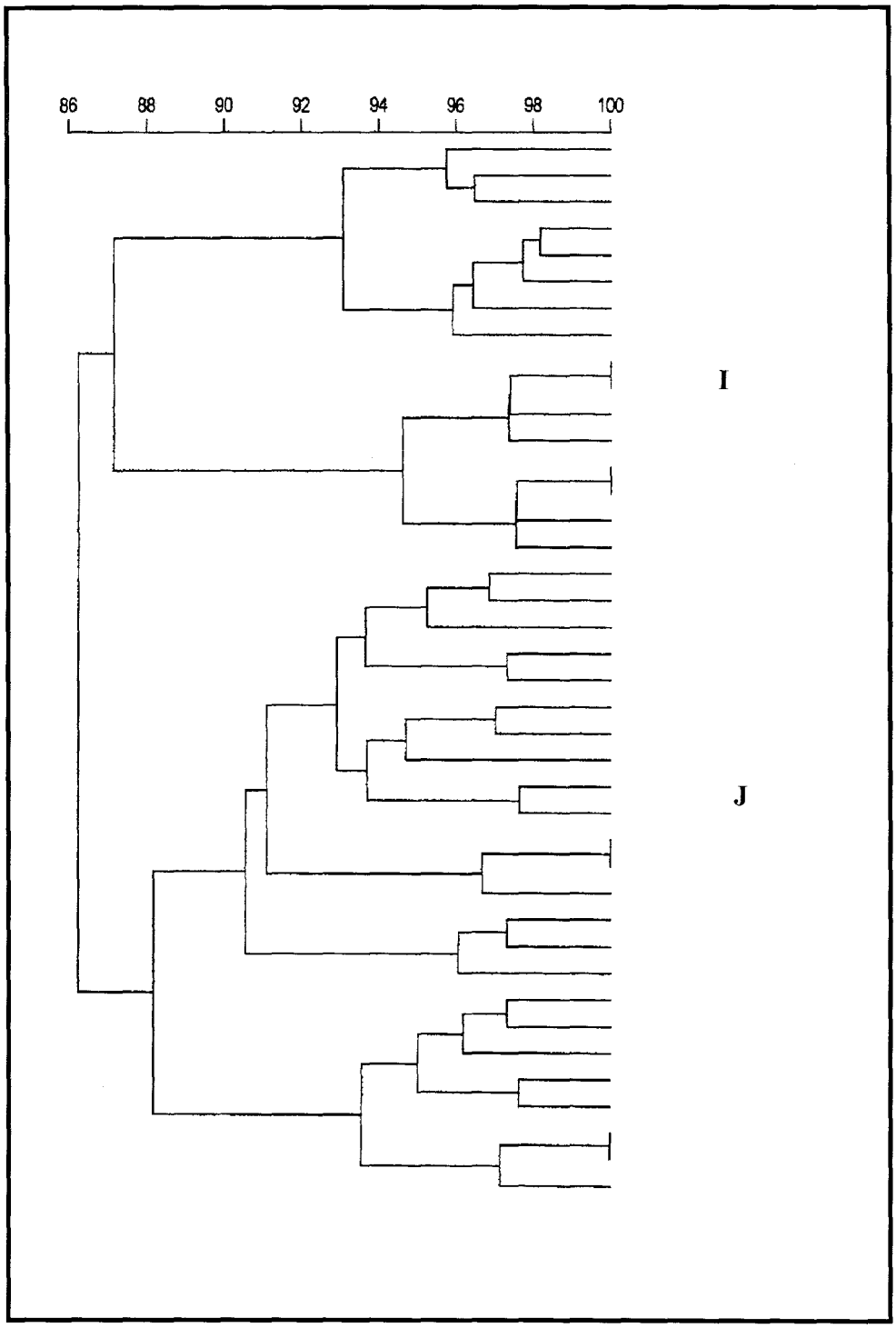


Figure 4.3 Cluster analysis of composite RAPD profiles of 40 KwaZulu-Natal strains

Isolates from both RFLP cluster groups were found in the two RAPD cluster groups. The majority of isolates in RAPD cluster group I were from RFLP cluster group G, but this was not statistically significant. The same was true of RAPD cluster group J. These results seemed to indicate a certain level of genetic independence of RAPD markers and IS*6110* insertion events. The lack of correlation between the two methods of typing for this population was in contrast to the study carried out by Linton *et al.* (1995), in which typing with RFLP and RAPD was compared in a small Swiss population. Here, correlation was much better, with strains from only three of nine RFLP groups breaking up into two different RAPD groups each. It should, however, be noted that the infection events were probably much closer in time in the Swiss population than in the KwaZulu-Natal population, and thus represented mostly recent rather than reactivation infection. It is likely that a significant proportion of the KwaZulu-Natal infection events were due to reactivation, which would explain the lack of correlation between RAPDs and RFLPs.

4.1.3 Analysis of molecular variance

A variety of databases were designed to test the genetic diversity and population structure of this population (Table 4.8). The molecular and population structure indices may be found in Appendix D. The mean genetic distance amongst the three cluster groups comprising Figure 4.1 was relatively large (0.148), compared to that in the Eastern Cape population (0.025) (Database N1, Appendix D). However, this was still within the range for conspecific populations (Keim *et al.*, 1999). Cluster groups E and F had only half of the genetic diversity of cluster group D, as seen from a comparison of the number of polymorphic sites (Table D.9).

Cluster D1 demonstrated the largest degree of genetic diversity relative to the other clusters, with cluster E2 showing very little diversity (Table D.10). The mean genetic distance amongst the cluster was 0.21, which was approximately five times greater than was found in the Eastern Cape population (0.04).

Strains from the various medical facilities surveyed in this population were divided into two sections, with Section 1 consisting of the 27 isolates received from King George V Hospital, which comprised 47% of the sample. The remaining 30 isolates in Section 2 were received from twelve medical facilities, ranging in distance from 10 km to 200 km away from each other. The aim of this division was to determine whether the relatively large King George V Hospital subpopulation was

in any way significantly different to strains from the other medical facilities. The molecular indices for this database (Table D.11), as well as the very low genetic distance of -0.007 between these two populations, indicated that no population structure existed between strains based on the location of medical facility (Database N3, Appendix D). It was not possible to test for urban-rural population structure due to the very small number of rural strains in this sample.

Table 4.8 Arlequin databases of the KwaZulu-Natal Population

Database	Description of Database	Number of strains	Group/s	Populations	Number of strains in population
N1	Ward dendrogram cluster groups of RAPD profiles	57	Group 1	Cluster Group D	31
				Cluster Group E	14
				Cluster Group F	12
N2	Ward dendrogram clusters of RAPD profiles	57	Group 1	Cluster D1	16
				Cluster D2	7
				Cluster D3	8
			Group 2	Cluster E1	10
				Cluster E2	4
			Group 3	Cluster F1	12
N3	Medical Facilities	57	Group 1	Section 1	27
				Section 2	30

4.1.4 Geographical distribution

Figure 4.4 sets out the locations of the medical facilities in the greater Durban area from which most isolates were received, according to suburb. The three towns containing medical facilities from which a small number of isolates were received, are indicated in the inset. The geographical distribution of cluster types demonstrated that, as with the Eastern Cape population, there was no correlation between specific cluster types and geographical location (Table 4.9). Cluster type D1 was the most widespread, stretching from Mahlabatini in the north down to Amanzimtoti. Cluster type D3 was also fairly widespread, being found from Botha's Hill down to Hibberdene. Cluster type E1 occurred in towns to the north and the south of Durban, but only in two suburbs of the city itself.

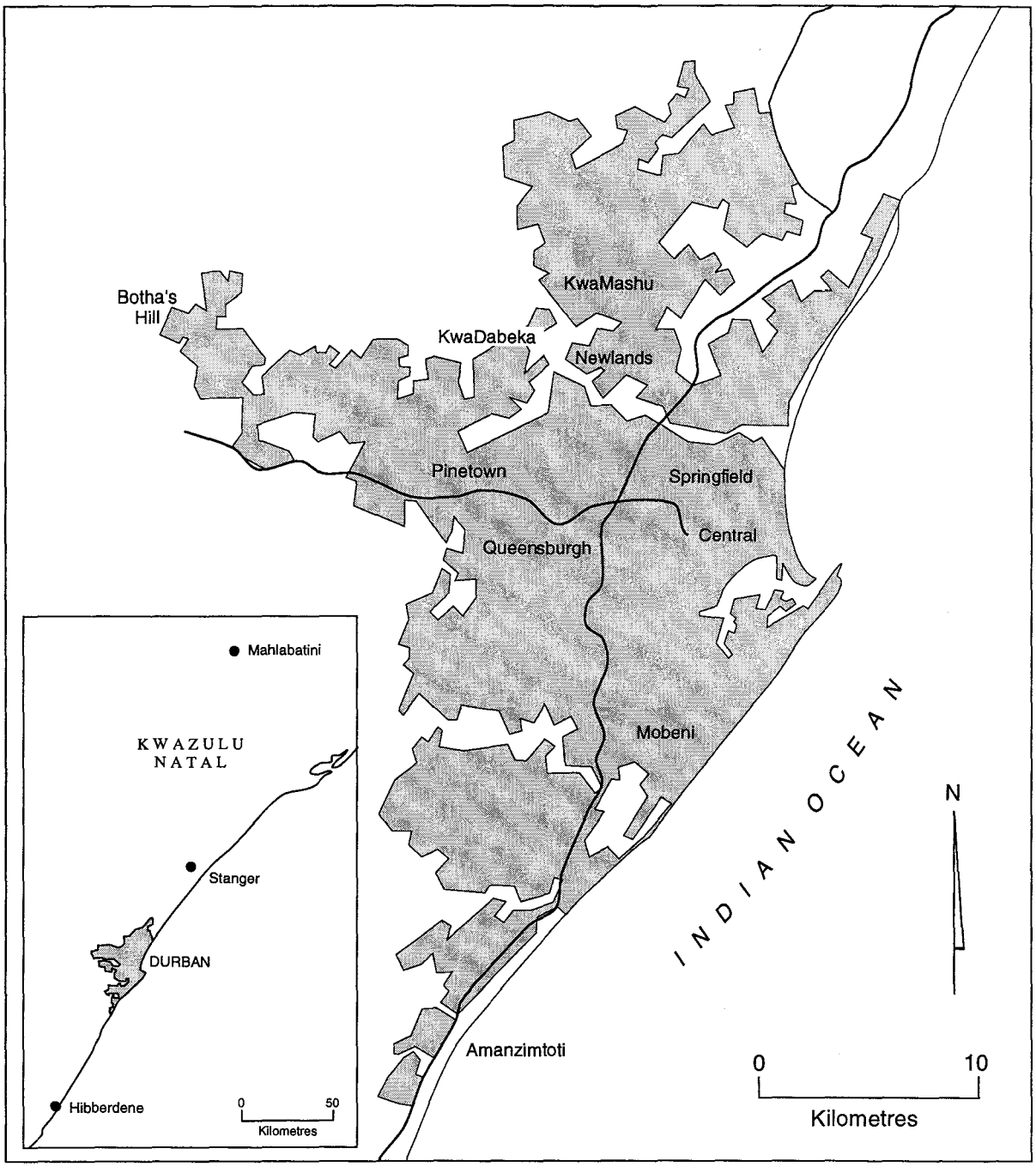


Figure 4.4 Locations of the KwaZulu-Natal medical facilities

The majority of isolates in the population were received from two medical facilities situated in the suburbs of Springfield and Central, which is why strains from all six cluster types occurred in these two areas. The majority of isolates (58%) in cluster type F1 came from Springfield, the suburb in which the King George V Hospital is situated.

Table 4.9 Distribution of KwaZulu-Natal cluster types

Cluster Type	Location of Medical Facilities (n = 13)
D1	Mahlabatini, KwaMashu, Newlands, Pinetown, Springfield, Central, Queensburgh, Amanzimtoti
D2	Pinetown, Newlands, Central, Springfield, KwaMashu
D3	Botha's Hill, Newlands, Springfield, Central, Mobeni, Hibberdene
E1	Stanger, Springfield, Central, Hibberdene
E2	Springfield, Central
F1	KwaDabeka, Springfield, Central

An examination of the incidence of cluster types in each medical facility revealed that all six cluster types occurred in the King George V hospital, from which almost 50% of the isolates had been received (Table 4.10). As many as five cluster types were found in the Durban Chest Clinic, even though only 19% of isolates came from this facility. An examination of pairs of closely related strains (similarity index ranging from 96 to 100%) indicated that strains from these two medical facilities grouped with strains from between five and seven other facilities. In addition, there were eight closely related groups containing strains from King George V Hospital only. Of these, one pair of strains had identical RAPD profiles.

As regards the other medical facilities, there were only three closely related groups containing strains from the same facility - one group each from Durban Chest Clinic, the Friends of the Sick Association (FOSA) clinic and Dunston Farell SANTA Centre. It was clear from this rather limited geographical analysis that cluster types identified in this small population were widely distributed throughout the area sampled.

Table 4.10 Incidence of KwaZulu-Natal cluster types

Medical Facility	Location	Number of cluster types (n = 6)	Number of Strains
King George V Hospital	Springfield	6	27
Durban Chest Clinic	Central	5	11
FOSA clinic	Newlands	3	4
Dunston Farell Santa Centre	Hibberdene	2	3
KwaMashu Chest Clinic	KwaMashu	2	2
KwaDabeka Clinic	KwaDabeka	1	2
Charles James Clinic	Amanzimtoti	1	2
St Francis Clinic	Mahlabatini	1	1
Clairwood Hospital	Mobeni	1	1
Pinetown Clinic	Pinetown	1	1
Shallcross Clinic	Queensburgh	1	1
Don McKenzie Hospital	Botha's Hill	1	1
Stanger Hospital	Stanger	1	1

4.2 KwaZulu-Natal and Eastern Cape drug resistant populations

Cluster analysis, AMOVA and an analysis of geographical distribution were performed on a combined database consisting of the composite RAPD profiles, generated by primers OPA1-13 and NTR, of 85 Eastern Cape drug resistant strains and 57 KwaZulu-Natal drug resistant strains.

4.2.1 Cluster Analysis

The dendrogram generated by the Dice similarity coefficient and the Ward cluster algorithm was rooted at a similarity index of 61.5%, with the strains separating into two cluster groups when a phenon line was placed at 73% similarity (Figure 4.5). Simpson's Index of Diversity dictated that the phenon line for cluster definition should be placed at 90% similarity, giving an index of 89%. Table 4.11 shows the similarity indices of the branching events that gave rise to cluster groups and clusters, as well as the range of similarity amongst strains within each.

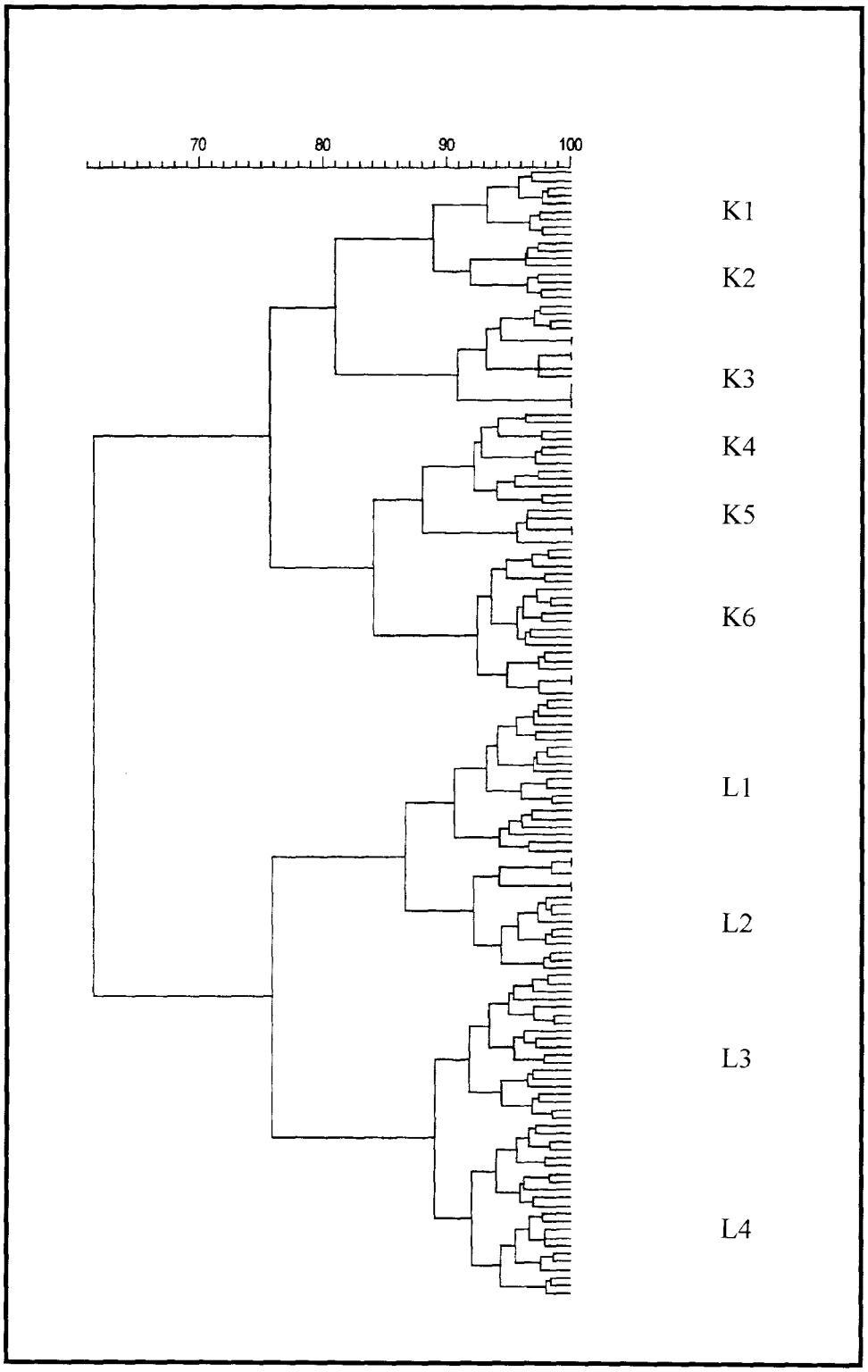


Figure 4.5 Cluster analysis of the drug resistant Eastern Cape and KwaZulu-Natal populations

Examination of the two cluster groups revealed a significant separation between the strains of these two geographically distinct populations, with the majority of KwaZulu-Natal strains (77%) being found in cluster group K and the majority of the Eastern Cape subpopulation (73%) occurring in cluster group L (Table 4.12).

Table 4.11 Similarity indices of cluster groups and clusters in the KwaZulu-Natal and drug resistant Eastern Cape populations

Cluster Group	Similarity index of branching point (%)	Similarity within cluster group (%)	Cluster	Similarity index of branching point (%)	Similarity within cluster (%)
K	61.5	75.5-100			
			K1	89	93.8- 98
			K2	89	92 - 96.5
			K3	81	91 - 100
			K4	88	92 - 98
			K5	88	96 - 100
			K6	84	92.4 - 100
L	61.5	76-100			
			L1	86.4	90.5 - 98
			L2	86.4	91.8 - 100
			L3	89	92 - 98.5
			L4	89	92 - 99

There were seven groups of sixteen strains with identical RAPD profiles in four of the ten clusters. Four of these groups consisted of pairs of KwaZulu-Natal strains, one of a pair of Eastern Cape strains and two of strains from both populations.

Table 4.12 Composition of cluster groups and clusters in the KwaZulu-Natal and drug resistant Eastern Cape populations

Cluster Group	Cluster	Number of KwaZulu-Natal strains	Number of Eastern Cape strains	Comparative incidence ($p \leq 0.05$)	Total number of strains
K		44	23	0.0132	67
	K1	6	3		9
	K2	6	2		8
	K3	9	5		14
	K4	10	2		12
	K5	4	1		5
	K6	9	10		19
L		13	62	<0.0001	75
	L1	4	16		20
	L2	6	8		14
	L3	2	17		19
	L4	1	21		22

The RAPD markers in this database were also analysed to determine whether certain markers occurred significantly more frequently in one geographical population than in the other amongst strains resistant to INH and rifampicin. Table 4.13 sets out the significant markers detected for each antibiotic and each primer. It can be seen that primer OPA1-13 generated only one marker, which could be correlated with INH resistance significantly more frequently amongst Eastern Cape strains than amongst KwaZulu-Natal strains. On the other hand, three markers generated by this primer occurred significantly more frequently amongst Eastern Cape strains resistant to rifampicin. One of these was the same size as the marker for INH resistance. Primer NTR generated two markers which occurred significantly more frequently amongst INH resistant Eastern Cape strains. No unique RAPD markers were generated by either primer amongst strains resistant to streptomycin, ethionamide, thiacetazone and ethambutol.

Table 4.13 RAPD markers linked to antibiotic resistance

Primer	Antibiotic	Frequency of Occurrence	Marker size (in base pairs)
OPA1-13	INH	More in Eastern Cape strains	1990
	Rifampicin	More in Eastern Cape strains	1990
		More in Eastern Cape strains	1600
		More in Eastern Cape strains	1100
NTR	INH	More in Eastern Cape strains	1500
		More in Eastern Cape strains	1000

4.2.2 Analysis of molecular variance

Molecular and population structure indices were calculated on the databases set out in Table 4.14 and may be found in Appendix D. The degree of genetic diversity was relatively high in both cluster groups, as can be seen from the large number of polymorphic sites (Table D.13). A small measure of population structure was demonstrated by the genetic distance between cluster group K (which contained more KwaZulu-Natal strains) and cluster group L (which contained the majority of Eastern Cape resistant strains), which was slightly larger, at 0.07, than the mean genetic distance between the Eastern Cape cluster groups (Database NE1, Appendix D).

The four clusters of cluster group L exhibited a greater degree of genetic diversity than was seen in the K clusters (Table D.14). The mean genetic distance (0.15) amongst the ten clusters was approximately three times higher than that in the Eastern Cape population. The significant degree of population structure between these two provincial samples was evident from the genetic distance between the KwaZulu-Natal and resistant Eastern Cape strains, which was almost a hundred-fold greater than that between the Eastern Cape urban and rural strains, as can be seen from a comparison of Databases 3 and NE3 in Appendix D. However, this was still well within the parameters for conspecific bacterial populations (Keim *et al.* (1999).

Table 4.14 Arlequin Databases of the KwaZulu-Natal and drug resistant Eastern Cape populations

Database	Description of Database	Number of strains in database	Group/s	Populations	Number of strains in population
NE1	Ward dendrogram cluster groups	142	Group 1	Cluster group K	67
				Cluster group L	75
NE2	Ward dendrogram clusters	142	Group 1	Cluster K1	9
				Cluster K2	8
				Cluster K3	14
				Cluster K4	12
				Cluster K5	5
				Cluster K6	19
			Group 2	Cluster L1	20
				Cluster L2	14
				Cluster L3	19
				Cluster L4	22
NE3	Geographical populations	142	Group 1	Eastern Cape drug resistant strains	85
				KwaZulu-Natal strains	57
NE4	Geographical populations	142	Group 1	Eastern Cape drug resistant urban strains	63
				Eastern Cape drug resistant rural strains	22
			Group 2	KwaZulu-Natal strains	57

Linkage disequilibrium tests were carried out to determine whether there were any unique RAPD markers in common between these two populations, which could be linked to resistance to any particular antibiotic. The methodology used was essentially the same as that for the Eastern Cape resistant subpopulation (Section 3.2.2.1). However, for this population, composite RAPD profiles

were used instead of profiles generated by single primers. Two markers, 1 300 and 550 base pairs each in size, were found to correlate with resistance to rifampicin and thiacetazone, and to INH and thiacetazone, respectively, amongst the Eastern Cape strains. In the KwaZulu-Natal population, four markers were associated with rifampicin resistance and two with thiacetazone resistance, but none were associated with resistance to more than one antibiotic. No unique markers were found which were common to both populations. A second set of linkage disequilibrium tests was performed using equal numbers of KwaZulu-Natal strains and resistant strains from Port Elizabeth and Uitenhage. This revealed unique markers common to both populations, but they were not associated with resistance to the same antibiotics (Table 4.15).

Table 4.15 Linkage disequilibrium data

Marker size in base pairs	Antibiotic/s	Population
1700	INH	KwaZulu-Natal
	Rifampicin & thiacetazone	Eastern Cape
750	Thiacetazone	KwaZulu-Natal
	INH	Eastern Cape
350	INH	KwaZulu-Natal
	Rifampicin & thiacetazone	Eastern Cape

4.2.3 Geographical distribution

An analysis of the geographical distribution of the ten cluster types found in this combined population shows their widespread occurrence in the two provinces (Table 4.16). This was to be expected in spite of the significant separation of the two provincial populations demonstrated by cluster analysis and AMOVA, as strains from each province were to be found in each cluster. Of the two cluster types occurring in the southern-most KwaZulu-Natal town sampled (Hibberdene), only one (L2) is found in a health region that is situated in the region of the Eastern Cape closest to that town (Health Region C). Furthermore, two cluster types (K6 and L1), found in the southern region of the Eastern Cape (Health Region A), also occur in the northern KwaZulu-Natal towns of Stanger and Mahlabatini. Cluster types K6 and L4 were the most widely spread in the Eastern Cape.

Table 4.16 Distribution of KwaZulu-Natal and drug resistant Eastern Cape cluster types

Cluster type	KwaZulu-Natal Towns	Eastern Cape Health Regions
K1	Durban	A
K2	Durban	A
K3	Durban	A, C
K4	Durban, Hibberdene	A
K5	Durban	A
K6	Durban, Stanger	A, B, C
L1	Durban, Mahlabatini	A, C
L2	Durban, Hibberdene	A, C
L3	Durban	A, C
L4	Durban	A, B, C

4.3 KwaZulu-Natal and Eastern Cape populations

The RAPD profiles of the 57 KwaZulu-Natal strains, generated by primers OPA1-13 and NTR, were combined with the 502 Eastern Cape profiles generated by the same primers. Cluster analysis, AMOVA and an analysis of geographical distribution were performed on the resultant dendrogram.

4.3.1. Cluster analysis

Cluster analysis was performed using the Dice coefficient and Ward cluster algorithm (Figure 4.6). The dendrogram was rooted at 8%, a phenon line placed at a similarity index of 60% resulting in the definition of four cluster groups. Simpson's Index of Diversity placed the cluster phenon line at 86%, giving an index of discrimination of 94% and defining seventeen clusters. Table 4.17 sets out the similarity indices of the branching events that gave rise to the cluster groups and clusters, as well as the range of similarity amongst strains within each cluster group and cluster. Table 4.18 sets out the composition of the clusters for this combined population.

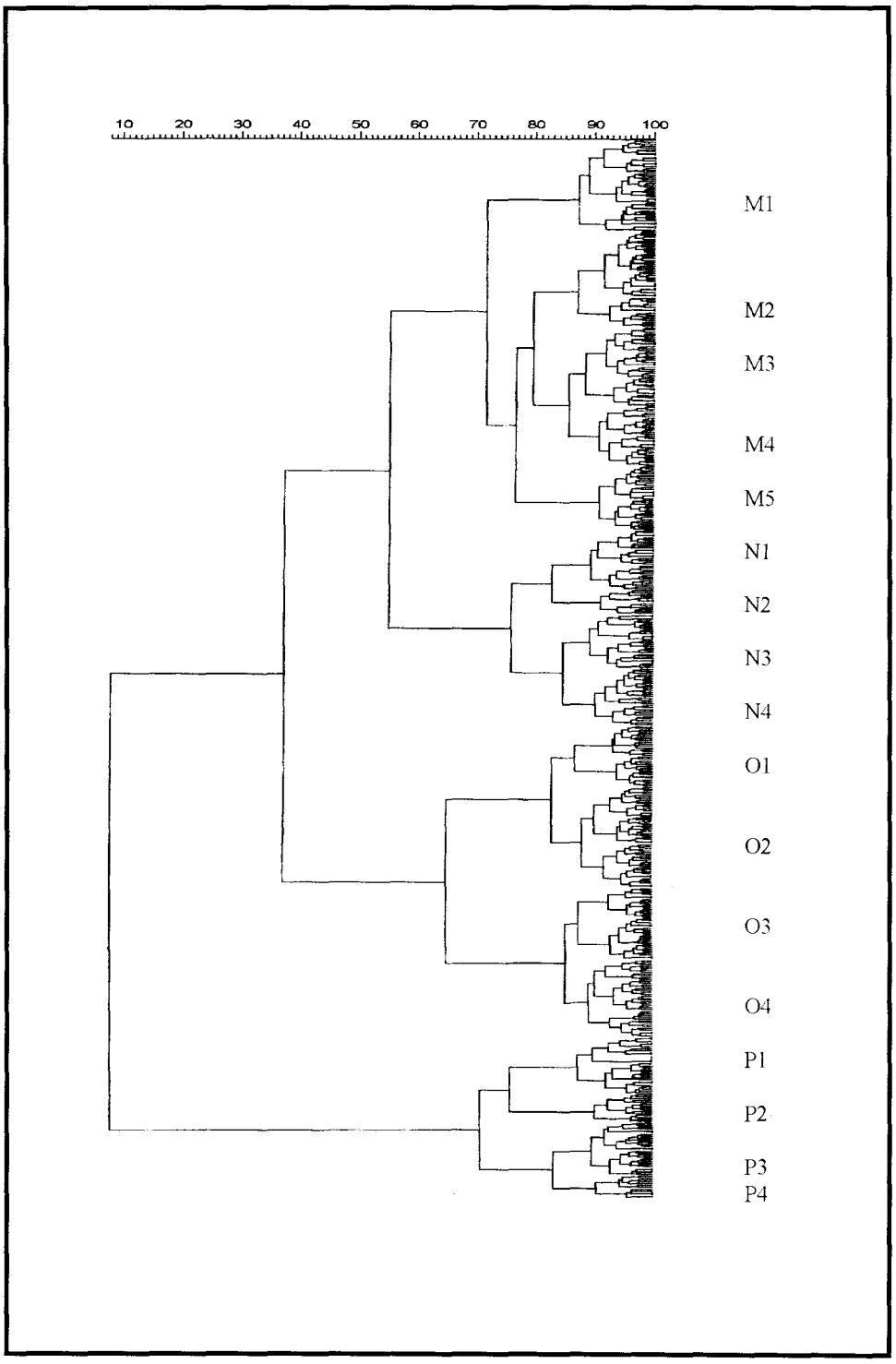


Figure 4.6 Cluster analysis of the Eastern Cape and KwaZulu-Natal populations

Table 4.17 Similarity indices of cluster groups and clusters in the KwaZulu-Natal and Eastern Cape populations

Cluster Group	Similarity index of branching point (%)	Similarity within cluster group	Cluster	Similarity index of branching point (%)	Similarity within cluster (%)
M	55	71.5 - 100			
			M1	71.5	87.1 - 100
			M2	80	87.1 - 100
			M3	85.5	88.5 - 100
			M4	85.5	90.5 - 100
			M5	77	91 - 100
N	55	77 - 100			
			N1	83	90 - 100
			N2	83	92 - 100
			N3	86	90.5 - 100
			N4	86	92 - 98
O	47.5	65 - 100			
			O1	83	87.1 - 100
			O2	83	88.1 - 100
			O3	86.5	87.3 - 100
			O4	86.5	89 - 100
P	8	70 - 100			
			P1	75	87.1 - 100
			P2	75	90.5 - 98
			P3	83	89.5 - 100
			P4	83	91 - 100

Table 4.18 Composition of cluster groups and clusters in the KwaZulu-Natal and Eastern Cape populations

Cluster Group	Cluster	Number of KwaZulu-Natal strains	Number of Eastern Cape strains	Number of Eastern Cape resistant strains	Total number of strains
M		**1**	**206**	**16**	**207**
	M1	0	49	4	49
	M2	0	51	3	51
	M3	0	42	3	42
	M4	0	31	15	31
	M5	1	33	5	34
N		**13**	**90**	**13**	**103**
	N1	1	31	4	32
	N2	4	9	1	13
	N3	7	22	4	29
	N4	1	28	4	29
O		**2**	**162**	**38**	**164**
	O1	1	31	8	32
	O2	0	53	7	53
	O3	0	38	0	38
	O4	1	40	13	41
P		**41**	**44**	**18**	**85**
	P1	12	18	5	30
	P2	12	3	2	15
	P3	6	22	10	28
	P4	11	1	1	12

It had been expected that the relatively few KwaZulu-Natal strains would be dispersed fairly evenly amongst the strains of the larger population. However, the majority (72%) of strains from the KwaZulu-Natal population separated out along with only 9% of the Eastern Cape population to form the smallest cluster group, P, which branched from the root of the dendrogram. This group consisted of only 85 strains, which constituted a mere 15% of the combined Eastern Cape and

KwaZulu-Natal database. It would appear that the majority of KwaZulu-Natal strains were very different to the Eastern Cape strains, while 28% were genetically more closely related to the larger population. This separation of the two geographical populations was statistically significant with a p value of 0.0011. It is interesting to note that, of the 9% of Eastern Cape strains that clustered with the majority of KwaZulu-Natal strains in cluster group P, 41% were antibiotic resistant. However, these antibiotic resistant strains only represented 21% of the Eastern Cape drug resistant subpopulation.

4.3.2 Analysis of molecular variance

Molecular and population structure indices were calculated on the databases in Table 4.19 and are to be found in Appendix D. Cluster group P exhibited a smaller degree of genetic diversity than was found in the other three cluster groups, as was evident from the small mean number of pairwise differences (Table D.17). The larger genetic distances between cluster groups P and cluster groups M, N and O reflected the population structure between strains of the two provinces (Database NEA1, Appendix D).

The distinction between the two populations on the basis of geographical location was also seen in Databases NEA3 and NEA4 (Appendix D). The migration rate between strains from the two provinces was very low at 5.9 (Database NEA3), in comparison with that between the urban and rural Eastern Cape strains at 1117 (Database NEA4). It was interesting to note that the genetic distance between KwaZulu-Natal strains and rural Eastern Cape strains was marginally larger than that between the KwaZulu-Natal and urban Eastern Cape strains, indicating a closer relationship between the urban strains from the two provinces than between urban and rural strains. The mean number of pairwise differences amongst the KwaZulu-Natal strains was almost half that found amongst the Eastern Cape strains as can be seen from Tables D.19 and D.20. This indicated the existence of a smaller degree of genetic diversity in the former population. However, it should be borne in mind that this may be due to the fact that primer NTR generated fewer markers and polymorphisms from resistant strains than it did from drug sensitive strains.

Table 4.19 Arlequin Databases of the KwaZulu-Natal and Eastern Cape populations

Database	Description of Database	Number of strains in Database	Group/s	Populations	Number of strains in population
NEA1	Ward dendrogram cluster groups	559	Group 1	Cluster group M	207
				Cluster group N	103
				Cluster group O	164
				Cluster group P	85
NEA2	Ward dendrogram clusters	559	Group 1	Cluster M1	49
				Cluster M2	51
				Cluster M3	42
				Cluster M4	31
				Cluster M5	34
			Group 2	Cluster N1	32
				Cluster N2	13
				Cluster N3	29
				Cluster N4	29
			Group 3	Cluster O1	32
				Cluster O2	53
				Cluster O3	38
				Cluster O4	41
			Group 4	Cluster P1	30
				Cluster P2	15
				Cluster P3	28
				Cluster P4	12
NEA3	Geographical populations	559	Group 1	Eastern Cape strains	502
				KwaZulu-Natal strains	57
NEA4	Geographical populations	559	Group 1	Eastern Cape urban strains	340
				Eastern Cape rural strains	162
			Group 2	KwaZulu-Natal strains	57

4.3.3 Geographical distribution

The distribution amongst the KwaZulu-Natal towns and the Eastern Cape Health Regions of only those cluster types that contained strains from both geographical populations, is set out in Table 4.20

Table 4.20 Distribution of KwaZulu-Natal and Eastern Cape cluster types

Cluster type	KwaZulu-Natal Towns	Eastern Cape Health Regions
M5	Durban	A, B, C
N1	Durban	A, B, C
N2	Durban, Mahlabatini	A, B
N3	Durban, Hibberdene	A, B, C
N4	Durban	A, C
O1	Durban	A, B, C
O4	Durban	A, B, C
P1	Durban	A, C
P2	Durban	A
P3	Durban, Hibberdene, Stanger	A, B, C
P4	Durban	A

The widespread distribution of the cluster types common to both provinces is evident from this analysis. However, certain of the common cluster types occurred more frequently in one province than in the other. This was seen with cluster types M5, N1, N4, O1 and O4, which occurred more frequently in the Eastern Cape. However, strains of cluster types P2 and P4 occurred less frequently in the Eastern Cape, being limited to Health Region A. Cluster types N3 and P3 were fairly ubiquitous, occurring in all three Health Regions of the Eastern Cape, as well as in a number of locations sampled in KwaZulu-Natal.

CHAPTER 5

DISCUSSION

The results of the genetic, geographical and antibiotic resistance analyses of the *Mycobacterium tuberculosis* populations from the two provinces are first discussed separately (Sections 5.1 and 5.2) and then comparatively (Section 5.3).

5.1 Eastern Cape *M. tuberculosis* population

The significance of the genetic diversity and population structure detected by cluster analysis and AMOVA in the total Eastern Cape population (502 strains), the Eastern Cape drug resistant subpopulation (85 strains) and the Port Elizabeth subpopulation (214 strains) is discussed in Sections 5.1.2, 5.1.3 and 5.1.4, respectively. The usefulness of the GIS analysis of these various populations is considered in Section 5.1.5.

5.1.1 Methodology and optimization of RAPD profiling

Before commencing this discussion, it is necessary to comment on certain methodological considerations. The sample methodology incorporated a distinct bias towards strains from urban centres in general and P.E. in particular. Consequently, this study has measured the genetic diversity of the organism in urban areas more extensively than it has in rural areas. Therefore, conclusions based on the rural data must of necessity be of a provisional nature until substantiated by further sampling.

Optimisation of RAPD-PCR usually involves a sequential investigation of each reaction variable, which requires a prohibitively large number of experiments (Cobb and Clarkson, 1994). However, the RAPD-PCR conditions previously optimized in our laboratory served as a satisfactory starting point for further optimisation (Krallis, 1991; Barnes, 1994; Nxomani, 1997; da Serra, 1997).

The methodology used in determining the percentage similarity for defining clusters of closely related strains was approached in two ways. The structure of dendrograms provided an indication of the separation of strains into cluster groups, due to the high heuristic value of the Ward cluster

algorithm, after which Simpson's Index of Diversity, which calculates the probability of two unrelated strains being placed into different groups, provided a satisfactory means of establishing which strains should be considered as belonging to the same cluster.

5.1.2 Genetic diversity and population structure of the total Eastern Cape population

Cluster analysis revealed an extremely high degree of DNA polymorphism amongst the 502 RAPD profiles generated by the four primers. This had been seen in the earlier Eastern Cape study as well, where as many as 342 unique RAPD profiles and only eight small groups of identical strains were seen amongst 369 strains (Richner *et al.*, 1997). In the full study, seven of these identical groups were lost due to the additional genetic information provided by the another two primers, resulting in the detection of 501 different RAPD profiles amongst 502 strains. A difference was seen in the population structure revealed by cluster analysis, with the smaller population of the earlier study separating into two cluster groups and four clusters, while the total population subsequently separated into three large cluster groups and ten clusters (Richner *et al.*, 1997; Richner *et al.*, 1999). Thus, the additional genetic information provided by the two extra primers and a further 133 isolates was considerable.

This unexpectedly high genetic diversity was surprising for two reasons. Firstly, studies of drug resistance and pathogenesis of this organism have shown that synonymous nucleotide substitution in the structural genes of *M. tuberculosis* is considerably less than in other pathogenic bacteria (Sreevatsan *et al.*, 1997). This limited allelic diversity in an organism that has infected one third of the world's population has led to the hypothesis that *M. tuberculosis* underwent a recent evolutionary bottleneck at the time of speciation, which is estimated to be some 15 000 to 20 000 years ago. However, this characterization of the organism as being relatively inert genetically is at odds with the high degree of chromosomal heterogeneity revealed by RFLP patterns (Warren *et al.*, 1996), and now by the RAPD profiling of this study. This anomaly may be due to the creation of genetic diversity by the transposition of IS*6110* and other mobile elements, giving rise rapidly to new subclones, as well as to genomic changes due to the selective pressure of antibiotic use (Sreevatsan *et al.*, 1997). Variants produced by the latter mechanism would usually be rapidly removed from the gene pool by successful disease treatment. However, given the difficulty in successfully treating drug resistant *M. tuberculosis*, such resistance associated diversity is maintained in the population (Sreevatsan *et al.*, 1997).

Secondly, the high genetic diversity in this population was surprising due to the high incidence of tuberculosis in the Eastern Cape. As previously mentioned, the majority of molecular epidemiological studies in America, Europe and the rest of Africa have shown that a high incidence of tuberculosis is usually accompanied by a low degree of genetic diversity, and *vice versa* (van Soolingen *et al.*, 1991; Hermans *et al.*, 1995). However, it was noted that a number of studies produced results which did not fit this neat model (Hermans *et al.*, 1995; Chevrel-Dellagi *et al.*, 1993; Safi *et al.*, 1997). These studies demonstrated a correlation between population heterogeneity and high genetic diversity. Such population heterogeneity is a reflection of the growing mobility of human populations, which has been a common trend in many parts of the world over the past few decades. Urbanisation and wars are amongst some of the factors that have contributed significantly towards population mobility, especially in Africa.

The results of this study, along with that performed in the Western Cape (Warren *et al.*, 1996), have substantiated the importance of population mobility and heterogeneity for genetic diversity of *M. tuberculosis*. It is well-known that the human population in the Eastern Cape province has been characterized by a high degree of mobility over the past fifty years for three main reasons. Firstly, demand for migrant labour on the mines in provinces to the north began towards the end of the 19th century and was responsible for the continuous movement of people between provinces for most of the 20th century (Davenport, 1987, pp 516-523). Migrancy also occurred between the rural areas of the Eastern Cape and the urban areas of the Eastern and Western Cape. However, the deliberate policy of impermanent housing, especially for migrant mine workers, made it difficult, even illegal, for them to relocate their entire families, thus encouraging the maintenance of ties with their region of origin. Consequently a large amount of commuting between urban and rural areas occurred, in the process of which new strains of tuberculosis contracted while away were introduced into the home community and *vice versa*. Secondly, accelerated industrialisation began in South Africa during World War II and resulted in an increase in the permanent movement of Xhosa-speaking people to the cities (Davenport, 1987, pp 338, 524, 526-529).

Finally, the creation of heterogeneous populations in the cities of the Eastern Cape occurred as a long term result of the forced resettlement of large numbers of people, which was part of the social policy of the pre-1994 government for more than forty years (Platzky and Walker, 1985, p 10). In the period 1960 to 1983, 139 000 people were moved off farms, more than 10 000 off Black Spots (freehold land owned by indigenous Africans or church missions), 12 000 off Informal

Settlements and 90 000 from areas targeted by the State for infra-structural and strategic consolidation. This represented approximately a quarter of the African population of the western part of the Eastern Cape (South African Institute of Race Relations, 1984, pp 99-102).

These factors have all contributed to the breaking up of stable rural communities, with large numbers of people ultimately finding their way into urban squatter areas. At the present time, population mobility continues to be a social reality in the Eastern Cape and in South Africa at large. Obtaining employment in urban areas has become increasingly desirable, as it sustains subsistence farming on ancestral lands in the rural areas. Consequently, a large percentage of urban dwellers have continual links with members of the extended family in the rural areas, which include care of the elderly, and ongoing involvement in the important religious, cultural and social aspects of community life (Mayer, 1961; Wilson, 1972; Pauw, 1973).

The population structure demonstrated by cluster analysis in this study correlated with the findings of the Western Cape study, in which the RFLP technique identified three clonal groups (Paul van Helden, personal communication). As with the preliminary results of the Eastern Cape study, only two groups had initially been characterized in the Western Cape (Warren *et al.*, 1996). Cluster analysis of the various cluster groups and clusters indicated that cluster group C had the highest amount of genetic diversity, while the converse was true of cluster group B (Table 3.1.) The largest proportion of rural individuals belonged to cluster group B, indicating the possible existence of less genetic diversity in the rural areas of this province (Table 3.2). No correlation was demonstrated between population structure and urban-rural location by either the preliminary or final results (Table 3.2). An urban-rural divide had been expected, based on the hypothesis that a limited number of strains would circulate in urban areas due to close living conditions, while larger numbers would be found in rural communities where populations are smaller and more separated by distance. Thus, it was expected that there would be less genetic diversity amongst urban isolates than amongst rural ones. However, this hypothesis failed to take into account the reality of a highly mobile human population, as discussed above.

As regards the possible evolution of *M. tuberculosis* in this region, the difference in branch lengths between the three cluster groups suggests that these groups diverged a significant time in the past (Table 3.1). However, placing a date on this would not be possible without some means of using a molecular clock on the RAPD data. The most that can be observed from the data is that the three

cluster groups may represent clonal groups of *M. tuberculosis* present in the Eastern Cape today, which could have evolved from three progenitor strains introduced into the South African human population in the past from Europe and/or Asia. Europeans first settled in present-day Cape Town in 1652, with the Dutch authorities subsequently using the newly established settlement for exiled political prisoners from Dutch East Asia during the 17th and 18th centuries (Davenport, 1987, pp 28-29; du Plessis, 1947, pp 4-7). It is possible that one of the three parent strains was already present when Europeans reached southern Africa and that the other two were then introduced from Europe and Asia. Cluster group C would be the obvious candidate for the original African progenitor strain, due to the fact that it is the "oldest", as evidenced by its branching directly from the root of the dendrogram (Figure 3.1). Further molecular studies of conserved genes, possibly those coding for the outer membrane protein, would be required to explore this hypothesis.

AMOVA provided a good indication of the degree of genetic diversity occurring at the molecular level in this population. A high percentage of polymorphic sites was found in the various populations which constituted the eight databases tested, with the lowest being 83% and the highest 100%. This was a reflection of the high degree of DNA polymorphism detected by cluster analysis. An examination of the mean number of pairwise differences in Table D.2 resulted in the identification of cluster group B as having the highest level of genetic diversity and cluster group C the lowest, which is directly opposite to the findings of cluster analysis. However, the molecular index provides a better indication of the amount of genetic diversity present in a population, as cluster analysis may suffer from biases inherent in the algorithm used.

AMOVA further indicated that the three cluster groups were statistically significant genetic groupings, even though the short term genetic distances ($\boldsymbol{F}_{ST}$) between them were not very large (Database 1, Appendix D). The mean genetic distance was 0.025, which is in line with other conspecific bacterial populations (Keim *et al.*, 1999). The ten clusters were also statistically supported at a molecular level (Database 2, Appendix D). It is not clear what these clusters represent other than a group of *M. tuberculosis* strains that are more closely related within their cluster than they are to strains in other clusters.

AMOVA supported the findings of cluster analysis as regards the absence of population structure at the urban-rural level. The existence of movement in both directions is reflected in the high migration rate of 8125 between urban and rural strains (Database 3, Appendix D), and of 354

between strains from Health Region A versus Health Regions B and C (Database 4). The existence of a small degree of population structure between the rural individuals of the health regions may be due to a smaller degree of human mobility between different rural areas, particularly over large distances, which would allow the development of strains specific to a particular rural area. Furthermore, in terms of medical services, Health Region A has been much better served historically than Health Regions B and C, and resultant differences in treatment regimens may have affected the genetic makeup of these populations.

5.1.3 Genetic diversity and population structure of the drug resistant subpopulation

Cluster analysis indicated high genetic diversity in the resistant subpopulation, with strains being distributed amongst all ten clusters, albeit in relatively small numbers in six of them (Table 3.2). Furthermore, the resistant strains in these six clusters were not more closely related to each other than they were to the drug sensitive strains in each cluster (Figure 3.1). The slightly larger numbers of strains found in clusters A1 and C4 were also fairly evenly spread out amongst the drug sensitive strains. However, the resistant strains in clusters C1 and C3 clustered together in a number of small groups at high similarity indices ranging from 96% to 99%. Taking into account the margin of error allowed for in the cluster analysis (Section 2.4.2), such strains can be considered to be very closely related and to be approaching 100% identity. The existence of these small groups of very closely related resistant strains might, therefore, constitute evidence of a small amount of person to person transmission in this population. However, the majority of strains (p = 0.0493) have developed resistance to antibiotics as a result of *de novo* mutation. The existence of a relatively large number of antibiotic phenotypes in this small resistant subpopulation is also indicative of a large degree of *de novo* evolution of antibiotic resistance (Table 3.4). It had been expected that primary acquisition of resistance by person to person transmission would predominate in this province as a result of the high incidence of HIV exposure and infection. This might well be the case in a high population density, high HIV incidence area such as a suburb of P.E. However, acquired resistance was shown to be the predominant mechanism of resistance development amongst the strains comprising this large scale study.

The demonstration of a low level of primary drug resistance in the Eastern Cape is in contrast to the situation in the Western Cape, where about two-thirds of drug resistance in the small, high density population was shown to be transmission driven (Paul van Helden, personal

communication). It is also in contrast to the preliminary findings of this study, where the level of drug resistance attributed to transmission was approximately the same as in the Western Cape (Richner *et al.*, 1997). This provides further indication that the inclusion of additional isolates and the use of a number of primers is able to alter the results obtained with this molecular marker technique.

The large proportion of *de novo* mutation detected within the drug resistant subpopulation might be indicative of a higher rate of patient non-compliance in the Eastern Cape than is the case in the Western Cape. The treatment interruption rate of 20% in the Eastern Cape is somewhat higher than that of the Western Cape, which is 15% (Department of Health, 1998a). The lowest treatment interruption rate occurs in the Northern Province (11%) and the highest in the North West Province (29%). A high level of patient non-compliance could point to the fact that implementation of the DOTS programme in the Eastern Cape has not been as successful as it has in the Western Cape. While the DOTS programme is being satisfactorily implemented in urban areas, this is not yet the case in many rural areas of the Eastern Cape. It is clear, therefore, that high priority needs to be given by health authorities to the speedy and extensive implementation of this programme, in order to improve patient compliance and to curtail the increase in MDR strains of *M. tuberculosis*.

Examination of the antibiotic phenotypes of strains in clusters C1 and C3 revealed that these were resistant to relatively few, mostly first line, antibiotics (Table 3.6). The other strains resistant to between one and three first line antibiotics are scattered throughout the remaining clusters. Thus, resistance to between one and three antibiotics may be transmission driven as well as arise from *de novo* mutation. As regards the small proportion of strains resistant to four or more antibiotics, these were not limited to one specific cluster, but could be found in six of the ten. This suggests that these highly dangerous strains are not yet a major problem within the community, having developed their resistance to multiple antibiotics by *de novo* mutation in response to repeated non-compliance with antibiotic treatment regimes. However, the genetic diversity of such strains presents a serious threat in that it increases the possibility of the development of a highly pathogenic, untreatable, and thus potentially fatal, strain.

An examination of the three antibiotic phenotypes that occurred most commonly in this population may throw some light on the question relating to progression in resistance development (Table 3.5). The most common phenotype, EC2, involves resistance to just one antibiotic, isoniazid, while

phenotype EC4 involves development of resistance to isoniazid and rifampicin, and phenotype EC1 involves resistance to isoniazid, rifampicin and pyrazinamide, the first two being first line drugs, and the third a second line drug. Thus, the progression seems to involve development of resistance from first to second line antibiotics, which is to be expected, as well as development of resistance to one, then to two and finally to three or more, antibiotics. This would fit in with the mechanism of resistance acquisition of *Mycobacteria*, which involves stepwise accumulation of individual mutations in several independent genes.

There was very little evidence in this population of correlation between particular antibiotic phenotypes and specific genetic types. Phenotypes were widely distributed amongst the ten clusters, and a variety of different phenotypes were found in any one cluster (Table 3.6). This was true even amongst the closely related strains in clusters C1 and C3. The only evidence of correlation was to be found in five pairs of closely related individuals belonging to clusters A1, C1 and C3. This lack of correlation may be due to the fact that the various structural genes in which mutations occur in response to the pressure of drug use, have not been surveyed by RAPD profiling. The two studies carried out in Tanzania and eastern Tunisia similarly failed to show correlation between genetic type and drug resistance (Chevrel-Dellagi *et al.*, 1993; Yang *et al.*, 1995).

Attempts to correlate specific antibiotic resistance with a unique RAPD marker or markers yielded mixed results. The two methods utilised failed to arrive at the same results (Tables 3.7 and 3.13). One reason for this could be that they did not approach the problem in the same way mathematically. The one method simply determined whether a particular RAPD marker or markers occurred significantly more frequently amongst drug resistant than amongst drug sensitive individuals (Table 3.7). The other detected the existence of a marker or markers that did not segregate randomly in the population at large. Such a marker would be in linkage disequilibrium and thus significantly associated with resistance to a particular antibiotic (Table 3.13). A more likely reason for the discrepancies between the two methods was that the portion of the organism's genome surveyed by the four primers was not large enough to generate a sufficient marker pool from which to detect unique resistance markers.

While cluster analysis indicated the existence of a marked degree of genetic diversity amongst the drug resistant strains, AMOVA demonstrated this to be smaller than that amongst the drug sensitive

strains in the sample (Table D.6, Appendix D). This was due to the less polymorphic, smaller number of RAPD markers generated by primer NTR from the majority of resistant strains. The population structure indices demonstrated a certain degree of population structure between drug resistant and drug sensitive individuals, in a way not evident from the cluster analysis. This was seen from the fact that the genetic distance between the drug resistant and drug sensitive strains was a hundred-fold greater than that between the urban and rural individuals (Database 6, Appendix D).

5.1.4 Genetic diversity and population structure of the Port Elizabeth subpopulation

Apart from considering the genetic diversity and population structure of this subpopulation, it was also important to compare it with strains from the other urban areas. This was difficult to do with any degree of confidence, because of the discrepancy in the numbers sampled. Cluster analysis again demonstrated the existence of high genetic diversity in the P.E. subpopulation, which can be explained by the continual movement of people into and out of the largest metropolitan area in the province. As previously established, people from the rural areas immediately surrounding Port Elizabeth, as well as from further afield, continually flow into the city in search of employment, education and economic betterment. Of the three clonal groups circulating in the province, the A type occurs most commonly in the Port Elizabeth area, in contrast to the more frequent occurrence of type C in the full population (Table 3.8). Type B strains occur less frequently in this subpopulation, as they do in the province as a whole. Strains of cluster A1 are in the majority in P.E., while B1 is represented by only three strains. No one specific genetic type was found to predominate in any one medical facility to the exclusion of others.

Strains in the adjacent city of Uitenhage also displayed a remarkable degree of genetic diversity, given that the number of isolates sampled was only a third of the size of the P.E. sample (Table 3.9). As many as six cluster types occurred in a sample of just eleven isolates in neighbouring Despatch. A high degree of genetic diversity was detected in East London as well, which is the second largest city in the province, situated some 300 kilometres east of Port Elizabeth. All ten cluster types were found amongst only 36 strains. The detection of such a high degree of genetic diversity in these small urban samples serves to reinforce the already established existence of a great degree of human mobility in this province.

Drug resistant strains from P.E. also displayed a relatively large amount of diversity, belonging to nine of the ten clusters (Table 3.8). Fifty six percent and 47% of the drug resistant individuals in clusters C1 and C3, respectively, were from P.E., indicating the existence of a certain degree of transmission driven resistance in this city. However, the majority of resistance development in P.E. was due to *de novo* mutation. This was further substantiated by the high percentage of antibiotic phenotypes (75%) found in this subpopulation (Table 3.11). The incidence of the three most common phenotypes - EC2, EC4 and EC1 - is virtually the same in the P.E. subpopulation as in the total population (Table 3.10). As with the total population, there was no correlation between particular antibiotic phenotypes and specific genetic types in P.E. Another similarity was seen in that the majority of P.E. drug resistant strains (88%) were resistant to between one and three antibiotics. The small group of P.E. strains resistant to between four and six antibiotics also displayed an alarming amount of genetic diversity, being distributed amongst five clusters.

AMOVA demonstrated little difference in genetic diversity between strains received from medical facilities in the black townships of P.E. and those from facilities in predominantly white and so-called coloured suburbs. This was seen in the similarity between the mean number of pairwise differences and number of polymorphic sites (Table D.7, Appendix D). Port Elizabeth resistant strains were also seen to have less genetic diversity than the drug sensitive strains (Table D.8). The same degree of population structure was found to exist between the drug resistant and drug sensitive strains in P.E. as was seen in the province at large.

5.1.5 Geographical distribution of the Eastern Cape cluster types

GIS was of great value in analysing the geographical distribution of the large number of strains in this study. The mapping according to medical facility and town of the ten cluster types into which cluster analysis had grouped the strains provided a graphic means of examining their distribution.

5.1.5.1 Total Eastern Cape population

An examination of Figures 3.2, 3.3 and 3.4, which plotted the distribution of the A, B and C cluster types respectively, indicated a widespread distribution of most cluster types amongst the 39 towns sampled, with eight of the ten occurring in all three health regions. Cluster types that occurred in towns in the extreme north of the province, such as A1, B3, C1 and C2, were also found in the

southern regions, with the same being true of the east-west distribution. The sparse distribution of cluster types in towns in Health Region B was due to the fact that only 4% of the strains making up the study population were received from the seven towns sampled in that area.

The lack of population structure on the urban-rural level was confirmed by GIS analysis, indicating that it is not helpful to think in terms of "urban" and "rural" strains of *M. tuberculosis* in this province.

Three of the seven main routes of human movement in the survey region could be correlated with the transmission of eight cluster types (Table 3.15). It was to be expect that the N2 would play a significant role as a major transmission route of the organism in the province. Sampling of Western Cape and KwaZulu-Natal towns along this route might reveal the existence of the same cluster types detected in this study. As regards the direction of movement of the cluster types in this population, it was not possible to determine whether cluster types radiated outward from the main urban centres on the coast, in northerly, easterly and westerly directions toward the rural areas, or whether they moved from the rural centres towards the urban areas. Movement in both directions is highly probable, due to the ongoing movement of people from the cities to the rural home communities with which they have maintained close ties, and back again. What is clear is that the large metropolitan areas form foci in which all of the genetic types identified in this study are to be found. In addition, certain rural centres, such as Graaff-Reinet, also form important foci of infection from which a variety of cluster types may be disseminated into neighbouring provinces.

The majority of medical facilities and towns with a large number of cluster types were located in urban rather than rural areas (Table 3.16). All ten cluster types were found in P.E., with some medical facilities having as many as nine types (Table 3.23). The close geographical proximity of Uitenhage and Despatch to P.E. would in all likelihood account for the large number of cluster types found there as well (Table 3.21). A specialist tuberculosis hospital in East London also had nine cluster types (Figure 3.13). A total of six cluster types were found in Grahamstown, from a sample of only eight isolates received from three medical facilities. King William's Town can unfortunately not be brought into this discussion, as only one isolate was received from its hospital. In the light of the situation in the other five cities, however, it would seem that there is a correlation between high incidence of genetic type and urban location.

On the whole, the converse was true in rural locations, except for four rural centres which contained a large number of cluster types. An examination of the geographical situations, and the medical facilities found in each of these towns, served to shed light on this surprising situation. Graaff-Reinet is situated at the junction between the Eastern, Western and Northern Cape provinces, which could result in a fair amount of migratory human movement through this town. Furthermore, the town contains a specialist tuberculosis hospital which serves a large catchment area in the northwestern region of the province. Tuberculosis hospitals exist in Fort Beaufort and Port Alfred, two other rural areas of large cluster type diversity. The exceptional degree of cluster diversity in the fourth town, Hankey, was difficult to explain, especially in the absence of a tuberculosis hospital. However, the proximity of this town to P.E., Uitenhage and Despatch might account for the unusual situation. The existence of fewer cluster types in the other rural towns may be due to the fact that smaller numbers of isolates were received from these areas. It is clear that further sampling would be needed in order to confirm the seeming correlation between low cluster diversity and rural location.

5.1.5.2 Drug resistant subpopulation

GIS analysis demonstrated the widespread distribution of this relatively small subpopulation (Figure 3.6). Resistant cluster types found in the western section of the sample area also occurred in the extreme eastern section. The most common resistant types, C1 and C3, were scattered amongst medical facilities ranging from Jansenville in the west to Komga in the east, and from Queenstown in the north, to P.E. in the south. A certain degree of correlation was seen between specific resistant cluster types and medical facilities based in urban locations. This was the case with types C1 and C3, which were found in ten urban facilities as opposed to only four rural ones, as can be seen from a comparison of Figures 3.6, 3.7, 3.8 and 3.9.

A high incidence of resistant cluster types was found in two urban areas, P.E. and Uitenhage. As many as five different resistant cluster types were found circulating amongst a relatively small number of strains in two medical facilities in P.E. (Figure 3.7). One of these facilities was a specialist tuberculosis hospital, in which a high degree of cluster diversity might be expected. Four resistant types occurred in a clinic in Uitenhage, from a sample of only five isolates (Figure 3.8). Urban centres thus seemed to form foci of resistant organisms with a high degree of cluster diversity. More extensive sampling would be required in East London, Grahamstown and King

William's Town, as well as in rural centres, in order to confirm this trend. However, cluster analysis demonstrated that the overall degree of cluster diversity amongst rural resistant strains was not significantly less than amongst urban resistant isolates. In spite of the fact that only 26% of the resistant subpopulation was obtained from rural locations, 80% of cluster types were found to be circulating in these areas (Table 3.19). This figure would in all likelihood rise with further rural sampling.

The relatively high degree of genetic diversity detected amongst rural resistant strains is probably indicative of a high degree of patient non-compliance, which may be due in part to the relative inaccessibility of health care and medical facilities in these more remote areas. The reasons for non-compliance would be different in urban areas, where large numbers of easily accessible medical facilities occur. The main reason for non-compliance in the cities can usually be linked to reluctance to complete the full course of the treatment regimen, which involves the prolonged administration of a number of powerful antibiotics whose side effects often compound the symptoms of the disease. However, one would expect genetic diversity caused by patient non-compliance to be tempered in the urban areas by the overcrowding and close contact which characterise urban life, and which should facilitate the person to person transmission of a limited number of MDR-TB strain types. This does not seem to be the case in this province, however, with all ten resistant cluster types circulating in urban areas and as many as five being identified from amongst eight resistant isolates sampled at a single clinic in P.E. (Figures 3.7 - 3.9). This would seem to support the findings of the cluster analysis, which indicates that acquired resistance is the predominant mechanism for the acquisition of drug resistant tuberculosis in this province.

The smallness of this subpopulation made it difficult to detect any correlation between resistant cluster types and main travel routes. Further sampling would be needed to shed light on the routes of transmission of resistant strains in this province.

5.1.5.3 Port Elizabeth subpopulation

All ten cluster types identified in this province were widely distributed amongst the medical facilities sampled in P.E., with no evidence of geographical correlation between cluster types and specific medical facilities (Figure 3.11). For example, cluster type A2 was found in the Motherwell NU2 clinic, which is the most northerly facility sampled, as well as in the Walmer Health Clinic at the

southern extremity of the sample area. It occurred in another ten facilities spread all over the greater P.E. area. This widespread distribution of most cluster types was evidence of much movement of types between facilities and suburbs. In addition a high degree of cluster diversity was seen in eight medical facilities, with the occurrence of between 70% and 90% of cluster types (Table 3.23).

All cluster types occurred in Uitenhage as well, while 90% were found in East London (Figures 3.12 and 3.13). Only seventy percent occurred in Despatch (Figure 3.12) and sixty percent in Grahamstown, but the sample size from these two cities was much smaller. Cluster types in Uitenhage and East London were randomly distributed amongst the medical facilities surveyed. Two clinics and a specialist tuberculosis hospital in Uitenhage showed a high degree of cluster diversity, with between 70% and 80% of cluster types being found amongst an average of fourteen strains received from each. The tuberculosis hospital in East London yielded as many as nine cluster types from a sample of just fourteen isolates. Thus, GIS substantiated the existence of high genetic diversity in the cities of this province.

5.2 KwaZulu-Natal *M. tuberculosis* population

The analysis of the drug resistance patterns, genetic diversity and geographical distribution of this small sample provided valuable information on *M. tuberculosis* in the area surveyed.

5.2.1 Drug resistance patterns

The higher incidence of resistance in the KwaZulu-Natal population to three first line antibiotics (rifampicin, streptomycin and ethambutol) may be indicative of their more frequent use in that province as opposed to the Eastern Cape (Table 4.1). It is to be expected that overuse of a particular antibiotic will result in a high incidence of resistance amongst strains. The frequent use of isoniazid for the treatment of tuberculosis throughout South Africa over the past 40 years has resulted in a high incidence of resistance to this antibiotic all over the country, as reflected in the two provincial populations examined in this study.

A relatively large number of antibiotic phenotypes was detected in this population, as was the case amongst the Eastern Cape isolates (Table 4.2). The most common phenotype in KwaZulu-Natal involved resistance to two first line antibiotics (isoniazid and rifampicin), whereas resistance to one first line drug (isoniazid) was more common in the Eastern Cape subpopulation (Table 4.3). However, it is not possible to extrapolate findings based on such a small sample. Further work involving a large scale sampling of resistant strains in both provinces would be necessary to fully understand the resistance dynamics of this organism.

5.2.2 Genetic diversity and population structure

Cluster analysis demonstrated a high degree of genetic diversity in this population, as reflected by the amount of DNA polymorphism (Figure 4.1). Eighty one percent of this population consisted of unique RAPD profiles, which was somewhat lower than that found in the Eastern Cape population (99%). However, only two primers were used to generate the KwaZulu-Natal RAPD profiles, as opposed to the four used in the Eastern Cape population. Furthermore, one of those - primer NTR - generated fewer polymorphisms and markers than the other primers. It has been shown that a further two primers, as well as the incorporation of larger numbers of isolates, can result in a considerable amount of additional genetic information. Thus, further work on isolates from KwaZulu-Natal might result in the detection of a level of genetic diversity in this population equal to that of the Eastern Cape.

If the criterion for determining transmission of drug resistant strains in this population is set at 100% similarity, then the five groups of identical strains would constitute evidence of a small amount of such transmission. However, as in the Eastern Cape population, the large majority of strains had unique profiles, which was indicative of the predominance of acquired, *de novo* resistance in this population. Examination of the strains constituting the identical groups showed that they were obtained from either the same medical facility or the same town. However, none of these identical strains had the same antibiotic phenotypes, as was also the case with closely related strains in the Eastern Cape resistant subpopulation. This might indicate that transmission of such strains occurred before development of drug resistance, with the strain subsequently developing resistance independently to different antibiotics.

As in the Eastern Cape population, there was no correlation between antibiotic phenotypes and specific genetic types (Table 4.6). Relatively large numbers of antibiotic phenotypes were present in each cluster, even in those containing small numbers of isolates. In contrast to the Eastern Cape population, significantly more strains in the KwaZulu-Natal population were resistant to four or more antibiotics, with such strains being widely distributed amongst all six clusters. The development of a high degree of multiple resistances is of great concern, being evidence of repeated patient non-compliance. Patient non-compliance in KwaZulu-Natal, as reflected in the treatment interruption rate determined by the Health Department for 1997/1998, was 19%, which is slightly less than that in the Eastern Cape (Department of Health, 1998a). The existence of such high genetic diversity amongst these multiply resistant strains is added cause for concern. However, further sampling would be needed to determine whether such strains predominate in all areas of the province.

It was not possible to test for population structure at the urban-rural level, as only 9% of strains were received from medical facilities in rural areas (Table 4.5). More extensive sampling would probably reveal a similar lack of correlation as was seen in the Eastern Cape study, as human migrancy in KwaZulu-Natal is of the same order as that in the Eastern Cape.

One of the three cluster groups (F) defined by cluster analysis was very different and contained only 21% of the population, all of which were closely related (Figure 4.1). The twelve strains in this group, from only three medical facilities situated in the greater Durban area, may represent a clonal group responsible for a microepidemic in that area (Table 4.9).

The lack of exact correlation between the grouping of strains by the RFLP and RAPD techniques demonstrated in this study probably implies that, for this population, the evolution of IS*6110* insertion elements and RAPD markers occurred independently (Table 4.7). Furthermore, for a number of reasons, direct comparisons should not be drawn with the Linton study in which good correlation between the two typing methods was shown (Linton *et al.*, 1995). The incidence of tuberculosis is lower in Switzerland than in South Africa, and isolates in the Swiss study came from a small geographical area, as compared to the KwaZulu-Natal isolates which originated from an area with a radius of 140 km. Both the lower incidence levels and the limited geographical sample of the Swiss study would increase the chances of detection of first level transmission events. The existence of very few identical RAPD profiles in the KwaZulu-Natal study suggested that such first level transmission had not been detected.

AMOVA demonstrated that the RAPD profiles of strains belonging to cluster groups E and F had substantially fewer DNA polymorphisms than those of cluster group D (Table D.9, Appendix D). While the mean genetic distances for KwaZulu-Natal cluster groups and clusters were five times greater than those seen in the Eastern Cape population, they were still within the range for conspecific populations. There was no evidence of population structure between strains received from King George V Hospital and those from the other medical facilities, with a very small genetic distance that was not significant at the 5% level, being seen (Database N3).

5.2.3 Geographical distribution of the KwaZulu-Natal cluster types

Due to the smallness of the population, computer-based GIS technology was not used to analyse the geographical distribution of these genetic types. As was the case in the Eastern Cape population, no correlation between genetic type and geographical location was detected in the KwaZulu-Natal sample (Table 4.9). Human migrancy is in all likelihood again responsible for the widespread distribution of genetic types in this population, where a continuous flow of migrant labour to the gold mines outside of the province as well as to the coal mines in the northern interior of KwaZulu-Natal takes place (Davenport, 1987, pp 516-523; Duminy and Guest, 1989, pp 222-224). Migration also takes place from the poverty-stricken rural areas into the major cities of Durban and Pietermaritzburg, as well as to centres of industrial development such as Richard's Bay (Davenport, 1987, pp 324, 338, 524, 526-529). The former government's policy of forced removal and resettlement of Black people was carried out in KwaZulu-Natal as well, with 300 000 people being moved off farms, 115 000 off so-called Black Spots, and 18 500 being moved to make way for infrastructural and strategic consolidations by the State (Platzky and Walker, 1985, pp 10). In more recent times, the deliberate policy practised by white farmers of underemployment on farms has contributed to the influx into the cities. As in the Eastern Cape, many of KwaZulu-Natal's migrants maintain ongoing contact with the traditional rural homeland (Pauw, 1973; Hammond-Tooke, 1974, pp 441-472). This high degree of human mobility has undoubtedly resulted in the wide distribution of genetic types, which is reflected in the fact that the same genetic type that occurred in seven urban medical facilities was also found in Mahlabatini in rural Zululand, 190 km north of Durban.

A high degree of cluster diversity was found in a number of medical facilities, with all six types circulating in the King George V Hospital (Table 4.10). This was to be expected, as the largest

number of strains was received from this facility. Furthermore, this major hospital is situated in the largest city in the province and thus serves a large catchment area. The high degree of cluster diversity in the Durban Chest Clinic was probably due to the specialist nature of this clinic, with the majority of people attending it being tuberculosis sufferers. More extensive geographical sampling would provide information related to routes of transmission of this organism in this province, as well as enable more accurate characterisation of strains from rural areas.

5.3 Eastern Cape and KwaZulu-Natal *M. tuberculosis* populations

Cluster analysis and AMOVA of the combined KwaZulu-Natal-Eastern Cape drug resistant population and the combined KwaZulu-Natal-total Eastern Cape population were used to determine whether any significant genetic difference existed between strains from the two provinces. The former population was also examined for the existence of common RAPD markers linked to specific antibiotic resistance in both populations. Finally, the geographical distribution of the cluster types defined for the combined databases was examined to determine whether certain types were predominant in one province as opposed to the other.

5.3.1 Genetic diversity and population structure

A significant degree of population structure was detected between strains from the two provinces, which was particularly evident in the KwaZulu-Natal and the drug resistant Eastern Cape population (Fig 4.5). A statistically significant proportion of KwaZulu-Natal strains separated out into one cluster group, while the majority of Eastern Cape drug resistant strains were found in the other (Table 4.12). In the KwaZulu-Natal-total Eastern Cape population, the separation between the two provinces was again unexpectedly distinct, with a significant majority of the KwaZulu-Natal strains grouping together in one cluster group along with very few Eastern Cape strains (Table 4.18). Thus, even though the KwaZulu-Natal population was relatively small, the majority of strains were not evenly distributed amongst those of the much larger Eastern Cape population as was expected, but clustered together separate from the majority of Eastern Cape strains.

AMOVA substantiated this correlation between population structure and provincial sample. The majority of the KwaZulu-Natal strains were less diverse genetically than the Eastern Cape strains (Table D.13 and D.17). The genetic distances between the cluster groups in both these populations

were more than twice the mean genetic distance of the Eastern Cape cluster groups, with correspondingly small migration rates (Databases NE1 and NEA1, Appendix D). Similar molecular and population structure results were obtained when testing databases composed of all KwaZulu-Natal strains versus Eastern Cape drug resistant strains (Table 4.14). The genetic distance between drug resistant urban and rural Eastern Cape strains was much smaller (0.015) than the genetic distances between the KwaZulu-Natal strains, and the urban and rural resistant Eastern Cape groups (0.072 and 0.068) (Database NE4, Appendix D). The same was true of the KwaZulu-Natal strains compared to the total Eastern Cape population (Database NEA4).

Thus, while the strains from the two provinces clearly belonged to the same species, in that the genetic distances were not as large as would occur between members of different species, it has been shown that a significant degree of population structure exists between the strains of *M. tuberculosis* sampled in these two provinces. Examination of the geographical areas sampled may provide a clue to the reason for this distinct genetic difference. While a large area of the Eastern Cape was surveyed, the majority of isolates were received from the western region of the province. A few isolates were received from towns to the north-east of East London, but nothing was obtained from the Transkei area. Only a small area of KwaZulu-Natal was sampled, with 91% of the isolates coming from the Durban metropolitan area. It has already been established that human migration from the western region of the Eastern Cape is in the direction of the Western Cape and the mining areas of the Witwatersrand and Free State. The majority of the migration from the Transkei is also in the direction of the western region of the Eastern Cape and the Western Cape, as well as to mines in provinces to the north (Wilson, 1972, pg 96). Thus, little movement occurs from the Eastern Cape into KwaZulu-Natal, except possibly from within the Umzimkulu district of the Transkei, which forms an enclave within the borders of the KwaZulu-Natal province. The converse is also true, with most migration from KwaZulu-Natal being in a northerly direction. Therefore, it is probable that the Transkei forms a geographical barrier between the two regions sampled in this study, which would account for the larger genetic distance between strains of the two provinces than is seen amongst strains within the Eastern Cape sample.

Testing of this hypothesis would involve extensive sampling of isolates from the Transkei, as well as more widespread sampling in KwaZulu-Natal. Such a study would reveal whether the *M. tuberculosis* population in the Transkei is genetically more closely related to the population in the western region of the Eastern Cape, or whether it forms a third population distinct from the other

two. On the other hand, sampling of a wider area of KwaZulu-Natal might result in the discovery of strains that are more closely related to those of the western and central Eastern Cape regions.

Tests for unique RAPD markers linked to antibiotic resistance were performed in the hope of detecting markers in common for both provincial populations. However, the two methods used to determine the existence of such markers failed to produce consistent results (Tables 4.13 and 4.15). This might be due partly to the fact that only two primers were used, which would have generated even less genetic data than was produced from the Eastern Cape strains. The method which was based on the frequency of markers in the two populations did detect a number of common markers, all of which occurred significantly more frequently in the Eastern Cape population (Table 4.13). However, the linkage disequilibrium test detected no common resistance markers from a database consisting of all drug resistant Eastern Cape and all KwaZulu-Natal strains. Tests carried out on a database consisting of all KwaZulu-Natal strains and an equal number of drug resistant strains from two major Eastern Cape cities did yield three common markers (Table 4.15). However, none of these was linked to resistance to the same antibiotic. Thus, neither test was able to show correlation between the RAPD markers detected in this study and mutations responsible for the development of resistance to antibiotics.

5.3.2 Geographical distribution

Widespread distribution of most cluster types identified in these two combined populations occurred, due to the fact that the separation of provincial strains by cluster analysis was not absolute (Tables 4.16 and 4.20). However, the small size of the KwaZulu-Natal sample makes it difficult to obtain an accurate idea of the distribution patterns of genetic type in this province. A large scale survey of the KwaZulu-Natal province, as well as of the Transkei region of the Eastern Cape, might reveal the existence of cluster types that occur predominantly in that province.

CHAPTER 6

CONCLUSION

6.1 The Eastern Cape study

This study has produced the largest third world database on the genetic diversity of *Mycobacterium tuberculosis* using RAPD profiling, and in so doing, has provided vital information on the genetic capability of a previously uncharacterised bacterial population. In spite of the fact that direct comparison with studies using RFLP methodology is problematic, a number of observations made in this study were found to be very similar to those of the small Western Cape study (Warren *et al.*, 1996). Furthermore, along with other African studies, this study has demonstrated the effect that the mobility and heterogeneity of human populations have on the genetic diversity *M. tuberculosis* (Chevrel-Dellagi *et al.*, 1993; Hermans *et al.*, 1995.

While the difficulties related to the use of the RAPD technique are acknowledged and have been controlled for as effectively as possible, this technique has substantiated previously known aspects of tuberculosis epidemiology and has also answered a number of vitally important questions pertaining to the genetic diversity, population structure and geographical distribution of the organism in a geographical region in the grip of a growing epidemic. The results of this study thus provide a baseline of knowledge which can be verified or modified by future work.

Cluster analysis was able to establish that the majority of infection represented in this study was due to reactivation disease as opposed to recent infection, as demonstrated by the large amount of DNA polymorphism amongst the RAPD profiles generated. This high degree of genetic diversity was found in all populations and subpopulations. However, a small scale study of strains from one medical facility or suburb of a city might reveal the existence of a higher degree of clustering due to microepidemics of recent infection. This study thus demonstrated the existence of a reservoir of infection that is more diverse genetically than was previously thought to be the case. An unexpected amount of interstrain diversity has also been detected by studies using RFLP fingerprinting, in an organism which has been characterised as genetically young and inert (Warren *et al.*, 1996). However, this genetic diversity does not seem to be related to mutations that lead to the acquisition of drug resistance, which have been shown by others to be located in the structural genes of the

organism (Sreevatsan *et al.*, 1997). Future work would be needed in order to correlate the diversity detected by these two molecular marker methods with other factors such as virulence and pathogenicity.

Cluster analysis provided valuable information on the drug resistant strains of the organism in this province. A relatively high degree of genetic diversity was detected, with no evident correlation between genetic type and antibiotic phenotype. Thus, the molecular markers generated by RAPD are not able to reflect resistance-related changes in the structural genes of the organism. However, RAPD was able to show that the predominant mechanism of drug resistance in this population was due to *de novo* mutation, which has served to highlight the need for priority to be given to improving patient compliance. However, a small scale study of a geographically restricted community in this province, such as a suburb or township of Port Elizabeth, might reveal primary resistance due to person to person transmission of resistant strains to be the predominant mechanism in such a situation, as was shown in the Cape Town study. Sampling of such a community, especially one with a high AIDS incidence, might be of great interest in the light of the correlation that has been shown elsewhere between primary drug resistance and the high incidence of AIDS.

Both cluster analysis and AMOVA demonstrated very clearly that no significant genetic difference is to be found between urban and rural strains, which may be attributed to the high degree of mobility of the human population. However, correlation could be demonstrated between genetic diversity and urban location, with large numbers of cluster types circulating in the majority of urban centres surveyed. A similar situation occurred in a small number of rural centres, where it was associated with the existence of specialist tuberculosis hospitals serving a large catchment area. The smaller degree of genetic diversity demonstrated in the majority of rural centres surveyed would need to be confirmed by more extensive sampling, as it may be related to the relatively small numbers of strains received from such areas.

AMOVA substantiated the high degree of genetic diversity demonstrated by cluster analysis, as well as the existence of three clonal groups of the organism in this province. The relatively high degree of genetic diversity detected in the resistant subpopulation by cluster analysis was shown by

AMOVA to be somewhat less than that of drug sensitive strains. AMOVA was able to detect a small degree of population structure amongst the strains, which was shown to correlate with drug resistance. The absence of population structure at the urban-rural level was substantiated by AMOVA.

GIS provided valuable information on the geographical distribution of cluster types identified in this study, showing them to be widely spread, with no correlation between a particular cluster type and a specific geographical location being demonstrated. Thus, GIS substantiated the lack of significant genetic difference between urban and rural strains demonstrated by cluster analysis and AMOVA. Furthermore, GIS demonstrated very graphically that urban areas were foci of high genetic diversity containing strains from all cluster types. This was to be expected, with the urban centres of this province serving as magnets for a large rural human population in search of employment. The widespread geographical distribution of genetic types of the organism would thus be due largely to the mobility of that sector of the human population which maintains ties with the rural community. More extensive sampling would be of value in confirming this hypothesis, as well as in further elucidating the role played in the transmission of this organism by main routes of human movement.

Resistant strains also displayed a wide geographical dissemination, with no one area showing a predominance of resistant as opposed to sensitive strains. However, more extensive sampling would be needed to confirm this finding, as the majority of strains included in this sample were drug sensitive.

This study also demonstrated that the use of more primers and larger numbers of isolates may modify preliminary results to a significant degree. This was particularly true of the drug resistant subpopulation, where early results, indicating primary acquisition of resistance to be the predominant mechanism in the population, were not borne out by the final results.

A final observation relates to the evolution of *M. tuberculosis* in the Eastern Cape Province. Cluster analysis revealed the existence of three large clonal groups, one of which is genetically more distant from the other two than they are from each other. The ubiquitous nature of this organism, along with the predisposition to infection that accompanies factors such as malnutrition and lowered

immunity, makes it likely that it was present on the subcontinent before the arrival of the European settlers. European strains would undoubtedly have accompanied the various waves of colonists in the 17th, 18th and 19th centuries. Strains from a third geographical region may have been introduced with the arrival of Asian labourers around the middle of the 19th and beginning of the 20th centuries. Thus, the older clonal group (cluster group C) may possibly be derived from the original African ancestor, while the other two clonal groups may be descendants of the original European and Asian ancestors. The racial groupings of patients from whom strains in these three clonal groups were obtained might provide an indication as to the accuracy of this hypothesis.

6.2 Comparison between the Eastern Cape and KwaZulu-Natal strains

A number of similarities between the KwaZulu-Natal and Eastern Cape samples was revealed by cluster analysis and AMOVA. The small, drug resistant KwaZulu-Natal population reflected a similar degree of genetic diversity to that seen in the Eastern Cape sample. Furthermore, acquisition of drug resistance also seemed to be due to *de novo* mutation, as opposed to person to person transmission of a small number of strains. However, it was not possible to determine whether there was an urban-rural divide in the KwaZulu-Natal population, as most of the strains were obtained from medical facilities situated in urban areas. Comparison of the two typing methods (RAPD and RFLP) yielded disparate results, which seems to indicate that IS*6110* insertion elements and RAPD markers have evolved independently in this population.

A comparative analysis of strains from the two provinces revealed an unexpectedly marked degree of population structure, which was statistically significant. This population separation was detected by both cluster analysis and AMOVA. The reason for the marked difference between the two provincial populations is not immediately evident, but further surveying of strains of *M. tuberculosis* in the Transkei, which forms the extreme eastern part of the Eastern Cape, might indicate whether this area forms a geographical barrier.

The genetic differentiation between the two provincial samples did not correlate with the confinement of certain genetic types identified in the combined populations to specific provincial regions, with most types being widely distributed in both provinces. This was due to the fact that the two provincial populations did not separate out totally, with a small proportion of Eastern Cape

strains being genetically similar to the majority of those from KwaZulu-Natal, and *vice versa*. However, this is to be expected with conspecific populations.

From all the above, it is apparent that RAPD profiling has provided considerable information relating to the genetic capability and geographical distribution of strains of *Mycobacterium tuberculosis*, particularly in the Eastern Cape Province. The groundwork has been laid for further work, using other suitable molecular marker techniques and incorporating isolates from areas not yet surveyed, to confirm the genetic diversity, population structure and geographical distribution of strains of *Mycobacterium tuberculosis* in these two provinces, as well as to investigate that occurring in other regions of South Africa.

REFERENCES

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. & Struhl, K. (1989). Current Protocols in Molecular Biology. Greene Publishing Associates & Wiley-Interscience: New York, USA.

Ayala, F.J. (1982). Population and Evolutionary Genetics, Chapter 1. The Benjamin/Cummings Publishing Co., Inc.: Menlo Park, Ca., USA .

Barker, N.P. (1990). The taxonomy of *Pentameris Beauv.* and *Pseudopentameris Conert.* M.Sc. Thesis, University of the Witwatersrand, Johannesburg, South Africa.

Barnes, J.M. (1994). Development of an identification system for *Legionella* species based on Random Amplified Polymorphic DNA fingerprinting. B.SC. (Hons.) Thesis, Rhodes University, Grahamstown, South Africa.

Beyers, N., Gie, R.P., Zietsman, H.L., Kunneke, M., Tatley, M. & Donald, P.R. (1996). The use of a geographical information system (GIS) to evaluate the distribution of tuberculosis in a high-incidence community. *South African Medical Journal*, **86**: 40-44.

Blinkhoff, P., Bukanga, E., Syamalevwe, B. & Williams, G. (1999). Under the Mupundu Tree. ACTIONAID: London, UK.

Bloom, B.R. & Murray, C.J. (1992). Tuberculosis: commentary on a re-emergent killer. *Science*, **257**: 1055-1064.

Bose, M., Chandler, A. & Das, R.H. (1993). A rapid and gentle method for the isolation of genomic DNA for *Mycobacteria*. *Nucleic Acids Research*, **21**: 2529-2530.

Bridge, P.D., Pearce, D.A., Rivera, A. & Rutherford, M.A. (1997). VNTR derived oligonucleotides as PCR primers for population studies in filamentous fungi. *Letters in Applied Microbiology*, **24**: 426-430.

Brower, V. (1996). Combating consumption. *Nature Biotechnology*, **14**: 1097-1099.

Buck, G.E., O'Hara, L.C. & Summersgill, J.T. (1992). Rapid, simple method for treating clinical specimens containing *Mycobacterium tuberculosis* to remove DNA for polymerase chain reaction. *Journal of Clinical Microbiology*, **30** (5): 1331-1334.

Caetano-Anollés, G., Bassam, B.J. & Gresshoff, P.M. (1991). DNA amplification fingerprinting using very short arbitrary oligonucleotide primers. *Bio Technology*, **9**: 553-557.

Caetano-Anollés, G., Bassam, B.J. & Gresshoff, P.M. (1992). Primer-template interactions during DNA amplification fingerprinting with single arbitrary oligonucleotides. *Molecular and General Genetics*, **235**: 157-165.

Caetano-Anollés, G. (1993). Amplifying DNA with arbitrary nucleotide primers. *PCR Methods and Applications*, **3**: 85-94.

Central Statistical Services. (1998). Preliminary results of 1996 South African Census. Pretoria, South Africa.

Chevrel-Dellagi, D., Abderrahman, A., Haltiti, R., Koubaji, H., Gicquel, B. & Dellagi, K. (1993). Large-scale DNA fingerprinting of *Mycobacterium tuberculosis* strains as a tool for epidemiological studies of tuberculosis. *Journal of Clinical Microbiology*, **31** (9): 2446-2450.

City Health Department. (1997). Port Elizabeth Annual Health Report. (July 1996-June 1997). Port Elizabeth, South Africa.

Clemens, D.L. (1997). *Mycobacterium tuberculosis*: bringing down the wall. *Trends in Microbiology*, **5** (10): 383-385.

Cobb, B.D. & Clarkson, J.M. (1994). A simple procedure for optimising the polymerase chain reaction (PCR) using modified Taguchi methods. *Nucleic Acids Research*, **22** (18): 3801-3805.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry III, C.E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.-A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S. & Barrell, B.G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**: 537-544.

Collins, T. (1998). Tuberculosis: What everyone should know about this disease. *SANTA TB and Health News*, **37** (2): 4-5.

Coulson, R.N., Lovelady, C.N., Flamm, R.O., Spradling, S.L. & Saunders, M.C. (1991). Intelligent geographic information systems for natural resource management. *In* Quantitative Methods in Landscape Ecology. Turner, M.G. & Gardner, R.H. (Eds.) Ecological Studies, vol. 82. Springer-Verlag: New York, USA.

Coutinho, H.L.C., Handley, B.A., Kay, H.E., Stevenson, L. & Beringer, J.E. (1993). The effect of colony age on PCR fingerprinting. *Letters in Applied Microbiology*, **17**: 282-284.

Crawford, J. & Bates, J.H. (1984). Phage typing of mycobacteria. *In* The Mycobacteria: a sourcebook, Part A. Kubica, G.P. & Wayne, L.G. (Eds.). Marcel Dekker, Inc.: New York, USA.

Dallas, J.F. (1988). Detection of DNA "fingerprints" of cultivated rice by hybridization with a human minisatellite DNA probe. *Proceedings of the National Academy of Sciences of the United States of America*, **85**: 6831-6835.

Da Serra, M.F. (1997). Fungal and substrate-associated factors affecting lignocellulolytic mushroom cultivation on wood sources available in South Africa. M.Sc. Thesis, Rhodes University, Grahamstown, South Africa.

Davenport, T.R.H. (1987). South Africa: A Modern History. 3rd Edition. MacMillan: Johannesburg, South Africa.

Department of Health. (1996). The South African Tuberculosis Control Programme: Practical Guidelines. Pretoria, South Africa.

Department of Health. (1997). Seventh national HIV survey of women attending antenatal clinics of the public health services in the Republic of South Africa, October/November 1996. *Epidemiological Comments*, **23** (2): 4-16.

Department of Health. (1998a). Annual Report of the National Tuberculosis Control Programme 1997-1998. Pretoria, South Africa.

Department of Health. (1998b). EPI Disease Surveillance Field Guide. Pretoria, South Africa.

Department of Health. (1999). Annual Report of the National Tuberculosis Control Programme 1998-1999. Pretoria, South Africa.

Dillon, J.R., Rahman, M. & Yeung, K. (1993). Discriminatory power of typing schemes based on Simpson's Index of Diversity for *Neisseria gonorrhoeae*. *Journal of Clinical Microbiology*, **31** (10): 2831-2833.

Duminy, A. & Guest, B. (Eds.) (1989). Natal and Zululand from the earliest times to 1910. University of Natal Press & Shuter and Shooter: Pietermaritzburg, South Africa.

du Plessis, I.D. (1947). The Cape Malays. Maskew Miller, Ltd: Cape Town, South Africa.

Eastern Cape Department of Health. (1997). *Eastern Cape Epidemiological Notes*, **1** (1): 1-8.

Eisenach, K.D., Crawford, J.T. and Bates, J.H. (1988). Repetitive DNA sequences as probes for *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology*, **26**: 2240-2245.

Eisenach, K.D. (1994). Use of an insertion sequence for laboratory diagnosis and epidemiologic studies of tuberculosis. *Annals of Emergency Medicine*, **24** (3): 450-453.

Ellsworth, D.L., Rittenhouse, K.D. & Honeycutt, L. (1993). Artifactual variation in randomly amplified polymorphic DNA banding patterns. *BioTechniques*, **14** (2): 214-217.

Farris, J.S. (1972). Estimating phylogenetic trees from distance matrices. *American Naturalist*, **106**: 645-668.

Festenstein, F. & Grange, J.M. (1991). Tuberculosis and the acquired immunodeficiency syndrome. *Journal of Applied Bacteriology*, **71**: 19-30.

Folgueira, L., Delgado, R., Palenque, E. & Noriega, A.R. (1993). Detection of *Mycobacterium tuberculosis* DNA in clinical samples using a simple lysis method and polymerase chain reaction. *Journal of Clinical Microbiology*, **31** (4): 1019-1021.

Foster, P.L. (1995). Adaptive Mutation. *In* Population Genetics of Bacteria. Baumberg, S., Young, J.P.W., Wellington, E.M.H. & Saunders, J.R. (Eds.) Society for General Microbiology, Symposium 52. University Press: Cambridge, UK.

Foulds, J. (1997). Meeting Report: The Comstock Conference on tuberculosis vaccines. *ASM News*, **63** (5): 256-258.

Gibson, J.R. & McKee, R.A. (1993). PCR products generated from unpurified *Salmonella* DNA are degraded by thermostable nuclease activity. *Letters in Applied Microbiology*, **16**: 59-61.

Gillespie, S.H. & McHugh, T.D. (1997). The biological cost of antimicrobial resistance. *Trends in Microbiology*, **5** (9): 337-339.

Gould, P. (1989). Geographic dimensions of the AIDS epidemic. *Progress in Geography*, **49**: 9.

Goyal, M., Young, D., Zhang, Y., Jenkins, P.A. & Shaw, R.J. (1994). PCR amplification of variable sequence upstream of *katG* gene to subdivide strains of *Mycobacterium tuberculosis* complex. *Journal of Clinical Microbiology*, **32** (12): 3070-3071.

Groenen, P.M.A., Bunschoten, A.E., van Soolingen, D. & van Embden, J.D.A. (1993). Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Molecular Microbiology*, **10** (5): 1057-1065.

Haas, W.H., Butler, W.R., Woodley, C.L. & Crawford, J.T. (1993). Mixed-linker polymerase chain reaction: a new method for rapid fingerprinting of isolates of the *Mycobacterium tuberculosis* complex. *Journal of Clinical Microbiology*, **31** (5): 1293-1298.

Hammond-Tooke, W.D. (Ed.) (1974). The Bantu-speaking Peoples of Southern Africa. 2nd Edition. Routledge & Kegan Paul Ltd: London, UK.

Harn, H-J., Shen, K.L., Ho, L.I., Yu, K.W., Liu, G.C., Yueh, K.C. & Lee, J.H. (1997). Evidence of transmission of *Mycobacterium tuberculosis* by random amplified polymorphic DNA (RAPD) fingerprinting in Taipei City, Taiwan. *Journal of Clinical Pathology*, **50**: 505-508.

Hermans, P.W.M., Schuitema, A.R.J., van Soolingen, D., Verstynen, C.P.H.J., Bik, E.M., Thole, J.E.R., Kolk, A.H.J. & van Embden, J.D.A. (1990). Specific detection of *Mycobacterium tuberculosis* complex strains by polymerase chain reaction. *Journal of Clinical Microbiology*, **28** (6): 1204-1213.

Hermans, P.W.M., Messadi, F., Guebrexabher, H., van Soolingen, D., de Haas, P.E.W., Heersma, H., de Neeling, H., Ayoub, A., Portaels, F., Frommel, D., Zribi, M. & van Embden, J.D.A. (1995). Analysis of the population structure of *Mycobacterium tuberculosis* in Ethiopia, Tunisia, and the Netherlands: usefulness of DNA typing for global tuberculosis epidemiology. *The Journal of Infectious Diseases*, **171**: 1504-1513.

Hilton, A.C., Banks, J.G. & Penn, C.W. (1997). Optimization of RAPD for fingerprinting *Salmonella*. *Letters in Applied Microbiology*, **24**: 243-248.

Hoelzel, R. (1990). The trouble with "PCR" machines. *Trends in Genetics*, **6**: 237-238.

Hunter, P.R. & Gaston, M.A. (1988). Numerical index of the discriminatory ability of typing systems: an application of Simpson's Index of Diversity. *Journal of Clinical Microbiology*, **26** (11): 2465-2466.

Jackson. P.J., Hill, K.K., Laker, M.T., Ticknor, L.O. & Keim.P. (1999). Genetic comparison of *Bacillus anthracis* and its close relatives using amplified fragment length polymorphism and polymerase chain reaction analysis. *Journal of Applied Microbiology,* **87**: 263-269.

Jacobs, W.R., Tuckman, M. & Bloom, B. (1987). Introduction of foreign DNA into mycobacteria using a shuttle plasmid. *Nature*, **327**: 532-534.

Karp, A., Edwards, K.J., Bruford, M., Funk, S., Vosman, B., Morgante, M., Seberg, O., Kremer, A., Bourst, P., Arctander, P., Tautz, D. & Hewitt, G.M. (1997). Molecular technologies for biodiversity evaluation: opportunities and challenges. *Nature Biotechnology*, **15**: 625-628.

Keim, P., Klevytska, A.M., Price, L.B., Schupp, J.M., Zinser, G., Smith, K.L., Hugh-Jones, M.E., Okinaka, R., Hill, K.K. & Jackson, P.J. (1999). Molecular diversity in *Bacillus anthracis*. *Journal of Applied Microbiology*, **87**: 215-217.

Kochi, A. (1991). The global tuberculosis situation and the new control strategy of the World Health Organization. *Tubercle*, **72**: 1-6.

Koornhof, H.J., Sirgel, F.A. & Fourie, P.B. (1995). Safety assurance for the laboratory diagnosis and monitoring of tuberculosis. Personal Communication.

Krallis, N. (1991). Rapid identification of *Mycobacterium* spp. using the PCR reaction. B.Sc. (Hons) Thesis, Rhodes University, Grahamstown, South Africa.

Lamboy, W.F. (1994a). Computing genetic similarity coefficients from RAPD data: the effects of PCR artifacts. *PCR Methods and Applications*, **4**: 31-37.

Lamboy, W.F. (1994b). Computing genetic similarity coefficients from RAPD data: correcting for the effects of PCR artifacts caused by variation in experimental conditions. *PCR Methods and Applications*, **4**: 38-43.

Lapointe, F.J. & Legendre, P. (1992). Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees. *Systematic Biology*, **41**: 378-384.

Linton, C.J., Jalal, H., Leeming, J.P. & Millar, M.R. (1994). Rapid discrimination of *Mycobacterium tuberculosis* strains by random amplified polymorphic DNA analysis. *Journal of Clinical Microbiology*, **32** (9): 2169-2174.

Linton, C.J., Smart, A.D., Leeming, J.P., Jalal, H., Telenti, A., Bodmer, T. & Millar, M.R. (1995). Comparison of random amplified polymorphic DNA with restriction fragment length polymorphism as epidemiological typing methods for *Mycobacterium tuberculosis*. *Journal of Clinical Pathology*, **48**: M133-M135.

Mayer, P. (1961). Townsmen or Tribesmen. Oxford University Press: Cape Town, South Africa.

Maynard Smith, J. (1995). Do bacteria have population genetics? *In* Population Genetics of Bacteria. Baumberg, S., Young, J.P.W., Wellington, E.M.H. & Saunders, J.R. (Eds.) Society for General Microbiology, Symposium 52. University Press, Cambridge, UK.

Mazurier, S., van de Giessen, A., Heuvelman, K. & Wernars, K. (1992a). RAPD analysis of *Campylobacter* isolates: DNA fingergerprinting without the need to purify DNA. *Letters in Applied Microbiology*, **14**: 260-262.

Mazurier, S. & Wernars, K. (1992b). Typing of *Listeria* strains by random amplification of polymorphic DNA. *Research in Microbiology*, **143**: 499-505.

Medical Research Council. (1997). Health Impact Report. Corporate Communication Division, Tygerberg, South Africa.

Metcalf, C. (1991). A History of Tuberculosis (Chapter 1). *In* A century of Tuberculosis: South African Perspectives. Coovadia, H.M. and Benatar, S.R. (Eds.) Oxford University Press: Cape Town, South Africa.

Narain, J.P. (1999). TB and HIV: Asian healthcare workers fight on two fronts. *The TB Treatment Observer*,7: 2.

Newbury, H.J. & Ford-Lloyd, B.V. (1993). The use of RAPD for assessing variation in plants. *Plant Growth Regulation*, **12**: 43-51.

NJMS National Tuberculosis Center. (1996). Newsletter. Website: http://www.umdnj.edu/~ntbcweb/history.htm.

Nxomani, C.D. (1997). Genetic characterization of conspecific populations of *Tilapia sparrmanii* in the dolomitic sinkholes and springs of the Western Transvaal (South Africa) and their comparison to *Tilapia guinasana*. Ph.D. Thesis, Rhodes University, Grahamstown, South Africa.

Orme, I.M. & Belisle, J.T. (1999). TB vaccine development: after the flood. *Trends in Microbiology*, 7 (10): 394-395.

Palittapongarnpim, P., Chomyc, S., Fanning, A. & Kunimoto, D. (1993a). DNA fingerprinting of *Mycobacterium tuberculosis* isolates by ligation-mediated polymerase chain reaction. *Nucleic Acids Research*, **21** (3): 761-762.

Palittapongarnpim, P., Chomyc, S., Fanning, A. & Kunimoto, D. (1993b). DNA fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolates by arbitrarily primed polymerase chain reaction. *The Journal of Infectious Diseases*, **167**: 975-978.

Park, Y-H & Kohel, R.J. (1994). Effect of concentration of $MgCl_2$ on random-amplified DNA polymorphism. *BioTechniques*, **16** (4): 652-655.

Pauw, B.A. (1973). The Second Generation. 2nd Edition. Oxford University Press: Cape Town, South Africa.

Peillon, R., Drouet, E.B., Bruneau, S., Panteix, G., Denoyel, G-A. & de Montclos, H.P. (1994). Discrimination of *Mycobacterium avium-Mycobacterium intracellulare* strains by genomic DNA fingerprinting with a 16S rRNA gene probe. *FEMS Microbiology Letters*, **124**: 75-80.

Platzky, L & Walker, C. (1985). The Surplus People: Forced Removals in South Africa. Ravan Press: Johannesburg, South Africa.

Plikaytis, B.B., Crawford, J.T., Woodley, C.L., Butler, W.R., Eisenach, K.D., Cave, M. D. & Shinnick, T.M. (1993). Rapid, amplification-based fingerprinting of *Mycobacterium tuberculosis*. *Journal of General Microbiology*, **139**: 1537-1542.

Ramser, J., Weising, K., Chikaleke, V. & Kahl, G. (1997). Increased informativeness of RAPD analysis by detection of microsatellite motifs. *BioTechniques*, **23**: 285-290.

Richner, S.M., Meiring, J. & Kirby, R. (1997). A study of the genetic diversity of *Mycobacterium tuberculosis* isolated from patients in the Eastern Province of South Africa using random amplified polymorphic DNA profiling. *Electrophoresis*, **18** (9): 1570-1576.

Richner, S., Meiring, J. & Kirby, R. (1999).DNA profiling of *Mycobacterium tuberculosis* from the Eastern Cape Province of South Africa and the detection of a high level of genetic diversity. *Electrophoresis*, **20** (8) : 1800-1806.

Rigouts, L. & Portaels, F. (1994). DNA fingerprints of *Mycobacterium tuberculosis* do not change during the development of resistance to various antituberculosis drugs. *Tuberculosis and Lung Disease*, **75** (2): 160.

Rohlf, F.J. & Fisher, D.R. (1968). Tests for hierarchical structure in random data sets. *Systematic Zoology*, **17**: 407-412.

Rohlf, F.J. (1993). NTSYS-pc: Numerical Taxonomy and Multivariate Analysis System, ver. 1.80. Applied Biostatistics, Inc.: New York, USA.

Ross, B.C. & Dwyer, B. (1993). Rapid, simple method for typing isolates of *Mycobacterium tuberculosis* by using the polymerase chain reaction. *Journal of Clinical Microbiology*, **31** (2): 329-334.

Safi, H., Aznar, J. & Palomares, J.C. (1997). Molecular epidemiology of *Mycobacterium tuberculosis* strains isolated during a 3-year period (1993 to 1995) in Seville, Spain. *Journal of Clinical Microbiology*, **35** (10): 2472-2476.

Saitou, N. & Nei, M. (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**: 406-425.

Sambrook, J., Fritsch, E.F. & Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual. 2nd Edition. Cold Spring Harbour Laboratory: Cold Spring Harbour, New York, USA.

Sandery, M., Coble, J. & McKerise-Donnolley, S. (1994). Random amplified polymorphic DNA (RAPD) profiling of *Legionella pneumophila*. *Letters in Applied Microbiology*, **19**: 184-187.

Schneider, S., Küffer, J-M., Rössli, D. & Excoffier, L. (1997). Arlequin ver. 1.1: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Schweder, M.E., Shatters Jr, R.G., West, S.H. & Smith, R.L. (1995). Effect of transition interval between melting and annealing temperatures on RAPD analyses. *BioTechniques*, **19** (1): 38-42.

Schierwater, B. & Ender, A. (1993). Different thermostable DNA polymerases may amplify different RAPD products. *Nucleic Acids Research*, **21** (19): 4647-4648.

Shannon, G.W., Pyle, G.F. & Bashur, R.L. (1991). The Geography of AIDS: Origins and Course of an Epidemic. The Guildford Press: New York, USA.

Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research Cambridge*, **58**: 167-175.

Slatkin, M. (1994). Linkage disequilibrium in growing and stable populations. *Genetics*, **137**: 331-336.

Small, P.M. & Moss, A. (1993). Molecular epidemiology and the new tuberculosis. *Infectious Agents and Disease*, **2**: 132-138.

Sneath, P.H.A. & Sokal, R.R. (1973). Numerical Taxonomy: The Principles and Practice of Numerical Classification. Freeman: San Francisco, USA.

South African Institute of Race Relations. (1984). Survey of Race Relations in South Africa, 1983. Johannesburg, South Africa.

Spratt, B.G., Smith, N.H., Zhou, J., O'Rourke, M. & Feil, E. (1995). The population genetics of the pathogenic *Neisseria*. *In* Population Genetics of Bacteria. Baumberg, S., Young, J.P.W., Wellington, E.M.H. & Saunders, J.R. (Eds.) Society for General Microbiology, Symposium 52. University Press: Cambridge, UK.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N., Kreiswirth, B.N., Whittam, T.S. & Musser, J.M. (1997). Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates revolutionarily recent global dissemination. *Proceedings of the National Academy of Sciences of the United States of America*, **94**: 9869-9874.

Stephan, R., Schraft, H. & Untermann, F. (1994). Characterization of *Bacillus licheniformis* with the RAPD technique (randomly amplified polymorphic DNA). *Letters in Applied Microbiology*, **18**: 260-263.

Strässle, A., Putnik, J., Weber, R., Fehr-Merhof, A., Wüst, J. & Pfyffer, G. (1997). Molecular epidemiology of *Mycobacterium tuberculosis* strains isolated from patients in a human immunodeficiency virus cohort in Switzerland. *Journal of Clinical Microbiology*, **35** (2): 374-378.

Taylor, G.M., Goyal, M., Legge, A.J., Shaw, R.J. & Young, D. (1999). Genotypic analysis of *Mycobacterium tuberculosis* from medieval human remains. *Microbiology*, **145**: 899-904.

Telenti. A. (1997a). Genetics of drug resistance in tuberculosis. *Clinics in Chest Medicine*, **18** (1): 55-63.

Telenti, A. (1997b). Genotypic assessment of isoniazid and rifampin resistance in *Mycobacterium tuberculosis*: a blind study at reference laboratory level. *Journal of Clinical Microbiology*, **35** (3): 719-723.

Telenti, A. (1997c). The *emb* operon, a gene cluster of *Mycobacterium tuberculosis* involved in resistance to ethambutol. *Nature Medicine*, **3** (5): 567-570.

Thierry, D., Cave, M.D., Eisenach, K.D., Crawford, J.T., Bates, J.H., Gicquel, B. & Guesdon, J.L. (1990). IS*6110*, an IS-like element of *Mycobacterium tuberculosis* complex. *Nucleic Acids Research*, **18**: 188.

van Embden, J.D.A., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P. Martin, C., McAdam, R., Shinnick, T.M. & Small. P.M. (1993). Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *Journal of Clinical Microbiology*, **31** (2): 406-409.

van Soolingen, D., Hermans, P.W.M., de Haas, P.E.W., Soll, D.R. & van Embden, J.D.A. (1991). Occurrence and stability of insertion sequences in *Mycobacterium* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *Journal of Clinical Microbiology*, **29** (11): 2578-2586.

van Soolingen, D., de Haas, P.E.W., Hermans, P.W.M. & van Embden, J.D.A. (1994). DNA fingerprinting of *Mycobacterium tuberculosis*. *Methods in Enzymology*, **235**: 196-205.

Vine, M.F., Degnan, D. & Hanchette, C. (1997). Geographic information systems: their use in environmental epidemiologic research. *Environmental Health Perspectives*, **105** (6): 598-605.

Wang, G., Whittam, T.S., Berg, C.M. & Berg, D.E. (1993). RAPD (arbitrary primer) PCR is more sensitive than multilocus enzyme electrophoresis for distinguishing related bacterial strains. *Nucleic Acids Research*, **21** (25): 5930-5933.

Ward Jr, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**: 236-244.

Warren, R., Hauman, J., Beyers, N., Richardson, M., Schaaf, H.S., Donald, P. & van Helden, P. (1996). Unexpectedly high strain diversity of *Mycobacterium tuberculosis* in a high-incidence community. *South African Medical Journal*, **86** (1): 45-49.

Wayne, L.G. & Kubica, G.P. (1986). Family *Mycobacteriaecae* (Chapter 16). *In* Bergey's Manual of Systematic Bacteriology Volume 2. Sneath, P.H.A., Mair, N.S., Sharpe, M.E. & Holt, J.G. (Eds.). Williams and Wilkins: Baltimore, Md., USA.

Weeden, N.F., Timmerman, G.M., Hemmat, M., Kneen, B.E. & Lodhi, M.A. (1992). Inheritance and reliability of RAPD markers. *In* Applications of RAPD Technology to Plant Breeding, Symposium Proceedings, pp. 12-17. Crop Science Society of America: Madison, WI.., USA.

Welsh, J. & McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research*, **18** (24): 7213-7218.

Werner, S., Jording, D., Simon, R. & Pühler, A. (1995). Insertion sequence (IS) elements as natural constituents of the genomes of the Gram-negative *Rhizobiaceae* and their use as a tool in ecological studies. *In* Population Genetics of Bacteria. Baumberg, S., Young, J.P.W., Wellington, E.M.H. & Saunders, J.R. (Eds.) Society for General Microbiology, Symposium 52. University Press: Cambridge, UK.

Whittam, T.S. (1995). Genetic population structure and pathogenicity in enteric bacteria. *In* Population Genetics of Bacteria. Baumberg, S., Young, J.P.W., Wellington, E.M.H. & Saunders, J.R. (Eds.) Society for General Microbiology, Symposium 52. University Press: Cambridge, UK.

Wiid, I.J.F., Werely, C., Beyers, N., Donald, P. & van Helden, P. (1994). Oligonucleotide (GTG)5 as a marker for *Mycobacterium tuberculosis* strain identification. *Journal of Clinical Microbiology*, **32** (5): 1318-1321.

Williams, S.T., Goodfellow, M., Alderson, G., Wellington, E.M.H., Sneath, P.H.A. & Sackin, M.J. (1983). Numerical classification of *Streptomyces* and related genera. *Journal of General Microbiology*, **129**:1743-1813.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A. & Tingey, S.V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research*, **18** (22): 6531-6535.

Wilson, F. (1972). Migrant Labour in South Africa. The S.A. Council of Churches & SPRO-CAS: Johannesburg, South Africa.

World Health Organization. (1995). WHO Report on the Tuberculosis Epidemic, 1995. WHO: Geneva, Switzerland.

World Health Organization. (1998a). Tuberculosis Fact Sheet. WHO: Geneva, Switzerland.

World Health Organization. (1998b). Report of the Technical Review Group, Global Program for Vaccines and Immunization. WHO: Geneva, Switzerland.

World Health Organization. (1998c). WHO Report on the Tuberculosis Epidemic, 1998. WHO: Geneva, Switzerland.

World Health Organization. (1998d). AIDS Epidemic Update - December 1998. WHO: Geneva, Switzerland.

World Health Organization. (1999). Global Tuberculosis Control: WHO Report, 1999. WHO: Geneva, Switzerland.

Yang, Z.H., Mtoni, I., Chonde, M., Mwasekaga, M., Fuursted, K., Askgård, D.S., Bennedsen, J., de Haas, P.E.W., van Soolingen, D., van Embden, J.D.A. & Andersen, Å.B. (1995). DNA fingerprinting and phenotyping of *Mycobacterium tuberculosis* isolates from human immunodeficiency virus (HIV)-seropositive and HIV-seronegative patients in Tanzania. *Journal of Clinical Microbiology*, 33 (5): 1064-1069.

Young, R.A., Bloom, B.R., Grosskinsky, C.M., Ivanyi, J., Thomas, D. & Davies, R.W. (1985). Dissection of *Mycobacterium tuberculosis* antigens using recombinant DNA. *Proceedings of the National Academy of Science of the United States of America*, **85**: 2583-2587.

Yu, K. & Pauls, K.P. (1992). Optimization of the PCR program for RAPD analysis. *Nucleic Acids Research*, **20** (10): 2606.

Zainuddin, Z.F. (1988). Mycobacterial plastids and related DNA sequences. Ph.D. thesis. University of Surrey, Surrey, UK.

Zainuddin, Z.F. & Dale, J.W. (1989). Polymorphic repetitive DNA sequences in *Mycobacterium tuberculosis* detected with a gene probe from a *Mycobacterium fortuitum* plasmid. *Journal of General Microbiology*, **135**: 2347-2355.

Zhang, Y., Mazurek, G.H., Cave, M.D., Eisenach, K.D., Pang, Y., Murphy, D.T. & Wallace Jr, R.J. (1992). DNA polymorphisms in strains of *Mycobacterium tuberculosis* analyzed by pulsed-field gel electrophoresis: a tool for epidemiology. *Journal of Clinical Microbiology*, **30** (6): 1551-1556.

APPENDIX A

DNA EXTRACTION PROTOCOLS, REAGENTS AND BUFFERS

1. Modification of van Soolingen's CTAB protocol

1.1 Centrifuge entire volume of heat-inactivated liquid bacterial culture (4 ml) for 5 minutes at 12 000g in sterile eppendorf tube, thus pelleting bacterial cells. Decant supernatent fluid.

1.2 Resuspend pellet in 500 µl of Tris-EDTA buffer (pH 7.2).

1.3 Add 50 µl of 10 mg/ml lysozyme and vortex cell suspension.

1.4 Incubate cell suspension at 37°C for 1 hour.

1.5 Add 70 µl of 10% sodium dodecyl sulphate and 3 µl of 20 mg/ml proteinase K. Incubate suspension at 65°C for 10 minutes.

1.6 Add 100 µl of 5M NaCl and mix suspension well.

1.7 Add 80 µl of CTAB/ NaCl solution.

1.8 Mix suspension well and incubate at 65°C for 10 minutes.

1.9 Add 700 µl of chloroform-isoamyl alcohol solution (24:1), mix for 10 seconds and centrifuge suspension at 12 000g for 5 minutes at room temperature. Buffer-saturated chloroform was used.

1.10 Transfer aqueous supernatant fluid to clean eppendorf tube. Add 0.6 times volume of isopropanol to supernatant fluid and hold tube at -70°C for thirty minutes.

1.11 Centrifuge at 12 000g for 15 minutes at room temperature.

1.12 Wash DNA pellet with 70% alcohol by centrifuging at 12 000g for 5 minutes at room temperature.

1.13 Decant alcohol and air-dry pellet.

1.14 Dissolve DNA in 50 µl of Tris-EDTA buffer and store at 4°C.

2. Reagents and buffers for CTAB protocol

2.1 Tris-EDTA buffer, pH 7.2

Tris-HCl, pH 7.2	10 mM
EDTA	1 mM

2.2 Lysozyme Solution

Lysozyme	10 mg
ddH_20	1 ml

2.3 Proteinase K Solution

Proteinase K	20 mg
ddH_20	1 ml

2.4 Sodium dodecyl sulphate (SDS)

SDS	10%

2.5 NaCl Solution

NaCl	5 mM

2.6 CTAB/NaCl solution

CTAB	10%
NaCl	4.1%

3. InstaGene Matrix protocol

3.1 Centrifuge of heat-inactivated liquid bacterial culture for 5 minutes at 12 000g in sterile Eppendorf tube, thus pelleting bacterial cells. Decant supernatant fluid.

3.2 Add 200 µl of InstaGene Matrix to pellet and incubate at 56°C for 15-30 minutes.

3.3 Vortex at high speed for 10 seconds. Place suspension in boiling waterbath for 8 minutes.

3.4 Vortex for 10 seconds. Centrifuge at 12 000g for 2-3 minutes.

3.5 Use 20 µl of supernatant per 50 µl PCR reaction. Store remainder at -20°C. Repeat step 3.4 when reusing DNA preparation.

APPENDIX B

PCR REAGENTS AND BUFFERS

1. 10x Reaction Buffer IV (Advanced Biotechnologies, Ltd)

200 mM $(NH_4)_2SO_4$; 750 mM Tris-HCl pH 9,0 ; 0.1% (w/v) Tween

2. Magnesium chloride (Advanced Biotechnologies)

Stock concentration	25 mM
Final concentration	2 mM

3. Deoxynucleotide triphosphates (Promega)

dATP, dCTP, dTTP and dGTP stock concentration	100 mM
Final concentration	100 µM

4. Primers

Stock solution	10 µM
Final concentration	0.2 µM

5. *Taq* polymerase

Stock solution	5 U/µl
Working concentration	1U/µl

6. Mineral Oil (Sigma)

Add 50 µl per PCR reaction

APPENDIX C

GEL ELECTROPHORESIS PROTOCOLS, REAGENTS AND BUFFERS

1. Polyacrylamide Gel Electrophoresis

Solutions and buffers are as described in Sambrook *et al.*, 1989.

1.1 Reagents and Buffers

1.1.1 Acrylamide Stock Solution

Acrylamide	30%
NN-methylene-bis-acrylamide	0.8%

1.1.2 Laemmli Resolving Gel Buffer pH 8.8

Tris-HCl	1.5 M
SDS	0.4%

1.1.3 Laemmli Stacking Gel Buffer pH 6.8

Tris-HCl	0.5 M
SDS	0.4%

1.1.4 1x TBE Buffer pH 8.0

Tris-borate	0.09 M
EDTA	0.002 M

1.1.5 Loading Buffer

Bromophenol Blue	0.25%
Xylene Cyanol	0.25%

1.2 PAGE procedure

10% resolving and 5% stacking gels were prepared using the Laemmli (1970) buffer systems, and allowed to polymerise separately using TEMED and 10% ammonium persulphate as polymerising agents. The Tall Mighty Small gel apparatus of Hoeffer Scientific Instruments (Ca, USA) was assembled according to manufacturer's instructions.

1.3 Silver Staining

1.3.1 Fixing Buffer

Ethanol	10%
Acetic acid	0.1%

1.3.2 Silver stain

Silver nitrate solution	0.1%

1.3.3 Developing solution

NaOH	1.5%
Formaldehyde	0.15%

1.4 Silver Stain procedure

Submerge in fixing buffer for 6 minutes. Discard fixant and stain in silver stain for 10 minutes. Pour off silver stain and wash gel twice with double-distilled water. Immerse gel in developing solution until bands become visible. Do not develop for longer than 40 minutes. Stop development by washing the gel in fixing buffer.

2. Agarose Gel Electrophoresis

2.1 Reagents and Buffers

2.1.1 Agarose Gel

Agarose (molecular grade)	2%
TBE pH 8.0	1x

2.1.2 Ethidium bromide

Ethidium bromide	10 mg/ml

2.2 Agarose gel electrophoresis procedure

Prepare a 2% agarose solution in 1x TBE (pH 8,0) by melting in microwave oven. Add 0.5 µg/ml of ethidium bromide to agarose before pouring mixture into horizontal gel casting tray. Insert well combs. Once gel slab has set, remove gel combs and place casting tray in horizontal buffer chamber filled with 1x TBE. Dilute RAPD product with loading buffer and electrophorese at 90 volts for 2.5 hours. Visualise markers on UV transilluminator.

APPENDIX D

AMOVA RESULTS

MOLECULAR STRUCTURE INDICES

Table D.1 Molecular indices for Database 1

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Cluster Group A	110 (96)	30.33
Cluster Group B	110 (96)	34.20
Cluster Group C	110 (96)	27.92

* Out of a total of 115 loci

Table D.2 Molecular indices for Database 2

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Cluster A1	106 (92)	28.11
Cluster A2	98 (85)	30.43
Cluster A3	102 (89)	32.39
Cluster B1	101 (88)	33.40
Cluster B2	102 (89)	35.14
Cluster B3	107 (93)	33.23
Cluster C1	96 (83)	23.29
Cluster C2	100 (87)	28.45
Cluster C3	100 (87)	26.86
Cluster C4	107 (93)	29.14

Table D.3 Molecular indices for Database 3

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Urban strains	113 (98)	30.89
Rural strains	111 (97)	31.41

Table D.4 Molecular indices for Database 4

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Region A	113 (98)	30.89
Regions B & C	111 (97)	31.60

Table D.5 Molecular indices for Database 5

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Region A Rural	108 (94)	31.19
Regions B & C Rural	109 (95)	31.53

Table D.6 Molecular indices for Database 6

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Drug Sensitive strains	115 (100)	31.27
Drug Resistant strains	106 (92)	28.85

* Out of a total of 115 loci

Table D.7 Molecular indices for Database 7

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Section 1	92 (96)	28.34
Section 2	94 (98)	28.62
Section 3	91 (95)	27.97

* Out of a total of 96 loci

Table D.8 Molecular indices for Database 8

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
P.E. drug sensitive strains	110 (96)	30.78
P.E. drug resistant strains	101 (88)	29.16

*Out of a total of 115 loci

Table D.9 Molecular indices for Database N1

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Cluster group D	34 (100)	8.83
Cluster group E	18 (53)	5.95
Cluster Group F	17 (50)	5.88

* Out of a total of 34 loci

Table D.10 Molecular indices for Database N2

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Cluster D1	31 (91)	8.89
Cluster D2	16 (47)	5.52
Cluster D3	19 (56)	7.18
Cluster E1	17 (50)	6.24
Cluster E2	7 (21)	3.67
Cluster F1	17 (50)	5.88

Table D.11 Molecular indices for Database N3

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Section 1	30 (88)	8.09
Section 2	32 (94)	8.47

Table D.12 Molecular indices for Database N4

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Cluster group G	37 (100)	13.43
Cluster group H	15 (41)	3.34

*Out of a total of 37 loci

Table D.13 Molecular indices for Database NE1

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Cluster group K	41 (84)	9.09
Cluster group L	49 (100)	13.67

* Out of a total of 49 loci

Table D.14 Molecular indices for Database NE2

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Cluster K1	28 (57)	8.56
Cluster K2	20 (41)	6.54
Cluster K3	26 (53)	8.14
Cluster K4	26 (53)	7.53
Cluster K5	8 (16)	3.80
Cluster K6	30 (61)	9.02
Cluster L1	44 (90)	14.89
Cluster L2	32 (65)	9.08
Cluster L3	44 (90)	14.62
Cluster L4	34 (69)	11.75

Table D.15 Molecular indices for Database NE3

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Eastern Cape drug resistant strains	48 (98)	13.51
KwaZulu-Natal strains	44 (90)	8.70

Table D.16 Molecular indices for Database NE4

Population	*Number of polymorphic sites (%)	Mean number of pairwise differences
Eastern Cape drug resistant urban strains	48 (98)	13.43
Eastern Cape drug resistant rural strains	42 (86)	13.45
KwaZulu-Natal strains	44 (90)	8.70

Table D.17 Molecular indices for Database NEA1

Population	*Number of polymorphic sites(%)	Mean number of pairwise differences
Cluster group M	54 (95)	15.18
Cluster group N	54 (95)	13.03
Cluster group O	51 (89)	14.97
Cluster group P	48 (84)	9.86

*Out of a total of 57 loci

Table D.18 Molecular indices for Database NEA2

Population	*Number of polymorphic sites(%)	Mean number of pairwise differences
Cluster M1	46 (81)	12.43
Cluster M2	51 (89)	16.77
Cluster M3	46 (81)	14.85
Cluster M4	41 (72)	13.28
Cluster M5	49 (86)	15.26
Cluster N1	46 (81)	12.51
Cluster N2	27 (47)	8.64
Cluster N3	45 (79)	12.48
Cluster N4	47 (82)	14.43
Cluster O1	44 (77)	14.56
Cluster O2	50 (88)	15.99
Cluster O3	44 (77)	14.32
Cluster O4	47 (82)	12.34
Cluster P1	40 (70)	10.30
Cluster P2	28 (49)	7.75
Cluster P3	40 (70)	9.90
Cluster P4	23 (40)	5.88

Table D.19 Molecular indices for Database NEA3

Population	*Number of polymorphic sites(%)	Mean number of pairwise differences
Eastern Cape strains	55 (96)	14.89
KwaZulu-Natal strains	48 (84)	8.91

Table D.20 Molecular indices for Database NEA4

Population	*Number of polymorphic sites(%)	Mean number of pairwise differences
Eastern Cape urban strains	55 (96)	14.95
Eastern Cape rural strains	54 (95)	14.75
KwaZulu-Natal strains	48 (84)	8.91

POPULATION STRUCTURE INDICES

Eastern Cape population

Database 1

Population Pairwise F_{ST}s (Significance)

	Cluster Group A	Cluster Group B
Cluster Group B	0.017 (+)*	
Cluster Group C	0.023 (+)	0.035 (+)

Migration Rate

	Cluster Group A	Cluster Group B
Cluster Group B	28.99	
Cluster Group C	21.52	13.81

Database 2

Population Pairwise F_{ST} s

	Cluster A1	Cluster A2	Cluster A3	Cluster B1	Cluster B2	Cluster B3	Cluster C1	Cluster C2	Cluster C3
Cluster A2	0.014								
Cluster A3	0.021	0.016							
Cluster B1	0.042	0.048	0.038						
Cluster B2	0.040	0.033	0.014	0.018					
Cluster B3	0.023	0.012	0.021	0.018	0.016				
Cluster C1	0.048	0.074	0.074	0.106	0.076	0.085			
Cluster C2	0.021	0.036	0.024	0.052	0.037	0.039	0.022		
Cluster C3	0.063	0.053	0.058	0.092	0.041	0.058	0.047	0.038	
Cluster C4	0.015	0.028	0.029	0.032	0.023	0.026	0.030	0.016	0.046

All the above values were significant.

* "+" indicates values that are statistically significant

"-" indicates values that are not statistically significant

Migration Rate

	Cluster A1	Cluster A2	Cluster A3	Cluster B1	Cluster B2	Cluster B3	Cluster C1	Cluster C2	Cluster C3
Cluster A2	35.64								
Cluster A3	23.14	30.31							
Cluster B1	11.50	9.93	12.65						
Cluster B2	12.15	14.95	36.27	28.02					
Cluster B3	21.13	40.68	23.82	27.31	30.91				
Cluster C1	9.89	6.28	6.24	4.20	6.10	5.38			
Cluster C2	23.59	13.48	19.99	9.10	13.17	12.41	22.62		
Cluster C3	7.41	8.93	8.07	4.93	11.67	8.12	10.13	12.80	
Cluster C4	32.97	17.23	16.58	15.29	21.71	19.06	16.07	30.79	10.45

Database 3

Population Pairwise F_{ST} s (Significance)

	Urban
Rural	0.0001 (-)

Migration rate

	Urban
Rural	8125.53

Database 4

Population Pairwise F_{ST}s (Significance)

	Reg A
Reg B & C	0.001 (-)

Migration rate

	Reg A
Reg B & C	353.71

Database 5

Population Pairwise F_{ST}s (Significance)

	Reg A Rural
Reg B & C Rural	0.005 (-)

Migration rate

	Reg A Rural
Reg B & C Rural	109.73

Database 6

Population Pairwise F_{ST}s (Significance)

	Drug Sensitive
Drug Resistant	0.022 (+)

Migration rate

	Drug Sensitive
Drug Resistant	22.18

Table D.19 Molecular indices for Database NEA3

Population	*Number of polymorphic sites(%)	Mean number of pairwise differences
Eastern Cape strains	55 (96)	14.89
KwaZulu-Natal strains	48 (84)	8.91

Table D.20 Molecular indices for Database NEA4

Population	*Number of polymorphic sites(%)	Mean number of pairwise differences
Eastern Cape urban strains	55 (96)	14.95
Eastern Cape rural strains	54 (95)	14.75
KwaZulu-Natal strains	48 (84)	8.91

POPULATION STRUCTURE INDICES

Eastern Cape population

Database 1

Population Pairwise F_{ST}s (Significance)

	Cluster Group A	Cluster Group B
Cluster Group B	0.017 (+)*	
Cluster Group C	0.023 (+)	0.035 (+)

Migration Rate

	Cluster Group A	Cluster Group B
Cluster Group B	28.99	
Cluster Group C	21.52	13.81

Database 2

Population Pairwise F_{ST} s

	Cluster A1	Cluster A2	Cluster A3	Cluster B1	Cluster B2	Cluster B3	Cluster C1	Cluster C2	Cluster C3
Cluster A2	0.014								
Cluster A3	0.021	0.016							
Cluster B1	0.042	0.048	0.038						
Cluster B2	0.040	0.033	0.014	0.018					
Cluster B3	0.023	0.012	0.021	0.018	0.016				
Cluster C1	0.048	0.074	0.074	0.106	0.076	0.085			
Cluster C2	0.021	0.036	0.024	0.052	0.037	0.039	0.022		
Cluster C3	0.063	0.053	0.058	0.092	0.041	0.058	0.047	0.038	
Cluster C4	0.015	0.028	0.029	0.032	0.023	0.026	0.030	0.016	0.046

All the above values were significant.

* "+" indicates values that are statistically significant

"-" indicates values that are not statistically significant

Migration Rate

	Cluster A1	Cluster A2	Cluster A3	Cluster B1	Cluster B2	Cluster B3	Cluster C1	Cluster C2	Cluster C3
Cluster A2	35.64								
Cluster A3	23.14	30.31							
Cluster B1	11.50	9.93	12.65						
Cluster B2	12.15	14.95	36.27	28.02					
Cluster B3	21.13	40.68	23.82	27.31	30.91				
Cluster C1	9.89	6.28	6.24	4.20	6.10	5.38			
Cluster C2	23.59	13.48	19.99	9.10	13.17	12.41	22.62		
Cluster C3	7.41	8.93	8.07	4.93	11.67	8.12	10.13	12.80	
Cluster C4	32.97	17.23	16.58	15.29	21.71	19.06	16.07	30.79	10.45

Database 3

Population Pairwise F_{ST} s (Significance)

	Urban
Rural	0.0001 (-)

Migration rate

	Urban
Rural	8125.53

Database 4

Population Pairwise F_{ST} s (Significance)

	Reg A
Reg B & C	0.001 (-)

Migration rate

	Reg A
Reg B & C	353.71

Database 5

Population Pairwise F_{ST} s (Significance)

	Reg A Rural
Reg B & C Rural	0.005 (-)

Migration rate

	Reg A Rural
Reg B & C Rural	109.73

Database 6

Population Pairwise F_{ST} s (Significance)

	Drug Sensitive
Drug Resistant	0.022 (+)

Migration rate

	Drug Sensitive
Drug Resistant	22.18

Database 7

Population Pairwise F_{ST}s (Significance)

	Section 1	Section 2
Section 2	-0.001 (-)	
Section 3	0.0003 (-)	0.001 (-)

Migration rate

	Section 1	Section 2
Section 2	inf	
Section 3	1863.67	515.22

Database 8

Population Pairwise F_{ST}s (Significance)

	P.E. Drug Resistant
P.E. Drug Sensitive	0.021 (+)

Migration rate

	P.E. Drug Resistant
P.E. Drug Sensitive	23.57

Natal population

Database N1

Population Pairwise F_{ST}s (Significance)

	Cluster Group D	Cluster Group E
Cluster Group E	0.076(+)	
Cluster Group F	0.179(+)	0.188(+)

Migration Rate

	Cluster Group D	Cluster Group E
Cluster Group E	6.06	
Cluster Group F	2.29	2.15

Database N2

Population Pairwise F_{ST}s (Significance)

	Cluster D1	Cluster D2	Cluster D3	Cluster E1	Cluster E2
Cluster D2	0.150(+)				
Cluster D3	0.167(+)	0.275(+)			
Cluster E1	0.089(+)	0.333(+)	0.213(+)		
Cluster E2	0.025(-)	0.319(+)	0.193(+)	0.123(+)	
Cluster F1	0.194(+)	0.343(+)	0.317(+)	0.198(+)	0.234(+)

Migration Rate

	Cluster D1	Cluster D2	Cluster D3	Cluster E1	Cluster E2
Cluster D2	2.89				
Cluster D3	2.49	1.32			
Cluster E1	5.14	1.03	1.85		
Cluster E2	19.68	1.07	2.10	3.57	
Cluster F1	2.08	0.96	1.08	2.03	1.63

Database N3

Population Pairwise F_{ST}s (Significance)

	Section 1
Section 2	-0.007(-)

Migration Rate

	Section 1
Section 2	inf

Database N4

Population Pairwise F_{ST}s (Significance)

	Cluster Group G
Cluster Group H	0.423(+)

Migration Rate

	Cluster Group G
Cluster Group H	0.73

Natal and Eastern Cape resistant populations

Database NE1

Population Pairwise F_{ST}s (Significance)

	Cluster Group K
Cluster Group L	0.071 (+)

Migration Rate

	Cluster Group K
Cluster Group L	6.51

Database NE2

Population Pairwise F_{ST}s (Significance)

	Cluster K1	Cluster K2	Cluster K3	Cluster K4	Cluster K5	Cluster K6	Cluster L1	Cluster L2	Cluster L3
Cluster K2	0.134								
Cluster K3	0.139	0.278							
Cluster K4	0.155	0.128	0.181						
Cluster K5	0.213	0.195	0.259*	0.018					
Cluster K6	0.147	0.129	0.209	0.058*	0.065				
Cluster L1	0.113	0.179	0.073	0.083	0.109	0.125			
Cluster L2	0.164	0.311	0.227	0.192	0.213	0.206	0.076		
Cluster L3	0.111	0.151	0.131	0.094	0.126	0.089	0.043	0.116	
Cluster L4	0.194	0.258	0.184	0.111	0.147	0.101	0.082	0.167	0.038

All the above values were significant, except those marked *.

Migration Rate

	Cluster K1	Cluster K2	Cluster K3	Cluster K4	Cluster K5	Cluster K6	Cluster L1	Cluster L2	Cluster L3
Cluster K2	3.24								
Cluster K3	3.11	1.30							
Cluster K4	2.72	3.40	2.27						
Cluster K5	1.85	2.06	1.43	27.37					
Cluster K6	2.90	3.39	1.89	8.09	7.23				
Cluster L1	3.94	2.30	6.35	5.53	4.09	3.51			
Cluster L2	2.55	1.11	1.71	2.11	1.85	1.93	6.05		
Cluster L3	4.02	2.80	3.32	4.82	3.48	5.15	11.04	3.82	
Cluster L4	2.07	1.44	2.22	3.99	2.89	4.46	5.60	2.50	12.81

Database NE3

Population Pairwise F_{ST}s (Significance)

	Natal
Eastern Cape	0.062(+)

Migration Rate

	Natal
Eastern Cape	7.58

Database NE4

Population Pairwise F_{ST}s (Significance)

	Natal	Eastern Cape Urban
Eastern Cape Urban	0.072(+)	
Eastern Cape Rural	0.068(+)	0.015(+)

Migration Rate

	Natal	Eastern Cape Urban
Eastern Cape Urban	6.46	
Eastern Cape Rural	6.81	33.29

Natal and Eastern Cape populations

Database NEA1

Population Pairwise F_{ST}s (Significance)

	Cluster Group M	Cluster Group N	Cluster Group O
Cluster Group N	0.038(+)		
Cluster Group O	0.036(+)	0.041(+)	
Cluster Group P	0.116(+)	0.052(+)	0.080(+)

Migration Rate

	Cluster Group M	Cluster Group N	Cluster Group O
Cluster Group N	12.57		
Cluster Group O	13.56	11.78	
Cluster Group P	3.80	9.08	5.74

Database NEA2

Population Pairwise F_{ST}s (Significance)

	M1	M2	M3	M4	M5	N1	N2	N3	N4	O1	O2	O3	O4	P1	P2	P3
Cluster M2	0.048															
Cluster M3	0.053	0.028														
Cluster M4	0.064	0.039	0.014													
Cluster M5	0.086	0.045	0.050	0.078												
Cluster N1	0.062	0.065	0.087	0.087	0.111											
Cluster N2	0.068	0.102	0.086	0.093	0.131	0.042										
Cluster N3	0.082	0.080	0.061	0.077	0.113	0.072	0.080									
Cluster N4	0.023	0.049	0.040	0.060	0.064	0.030	0.027	0.027								
Cluster O1	0.119	0.066	0.035	0.060	0.099	0.074	0.106	0.096	0.083							
Cluster O2	0.085	0.029	0.030	0.042	0.049	0.065	0.118	0.071	0.056	0.036						
Cluster O3	0.093	0.041	0.047	0.060	0.116	0.058	0.089	0.064	0.047	0.058	0.048					
Cluster O4	0.131	0.081	0.090	0.067	0.162	0.055	0.090	0.092	0.083	0.057	0.072	0.018				
Cluster P1	0.118	0.136	0.069	0.100	0.131	0.119	0.060	0.076	0.057	0.085	0.093	0.098	0.111			
Cluster P2	0.210	0.213	0.174	0.210	0.210	0.172	0.162	0.136	0.110	0.175	0.170	0.184	0.196	0.106		
Cluster P3	0.193	0.155	0.141	0.149	0.190	0.087	0.106	0.078	0.093	0.095	0.115	0.101	0.066	0.092	0.108	
Cluster P4	0.190	0.200	0.163	0.185	0.210	0.097	0.085	0.109	0.094	0.150	0.161	0.159	0.155	0.100	0.132	0.068

Migration Rate

	M1	M2	M3	M4	M5	N1	N2	N3	N4	O1	O2	O3	O4	P1	P2	P3
Cluster M2	9.97															
Cluster M3	8.99	17.49														
Cluster M4	7.36	12.46	35.37													
Cluster M5	5.35	10.69	9.92	5.88												
Cluster N1	7.59	7.23	5.25	5.23	4.02											
Cluster N2	6.84	4.40	5.35	4.89	3.32	11.30										
Cluster N3	5.58	5.79	7.73	6.01	3.94	6.46	5.76									
Cluster N4	20.91	9.78	12.06	7.99	7.26	16.04	18.16	18.14								
Cluster O1	3.71	7.08	13.86	7.89	4.58	6.31	4.21	4.69	5.54							
Cluster O2	5.39	16.73	15.96	11.37	9.78	7.19	3.74	6.54	8.14	13.37						
Cluster O3	4.92	11.76	10.23	7.85	3.80	8.16	5.10	7.33	10.10	8.07	9.86					
Cluster O4	3.35	5.68	5.06	7.00	2.58	8.62	5.03	4.95	5.53	8.35	6.49	26.79				
Cluster P1	3.72	3.18	6.79	4.54	3.31	3.69	7.80	6.10	8.30	5.41	4.88	4.63	4.02			
Cluster P2	1.88	1.85	2.38	1.88	1.88	2.41	2.59	3.17	4.04	2.35	2.45	2.22	2.06	4.23		
Cluster P3	2.09	2.74	3.05	2.86	2.13	5.23	4.21	5.94	4.88	4.77	3.85	4.46	7.11	4.93	4.12	
Cluster P4	2.14	2.00	2.57	2.20	1.88	4.67	5.39	4.09	4.85	2.84	2.61	2.64	2.74	4.49	3.38	6.82

Database NEA3

Population Pairwise F_{ST}s (Significance)

	Natal
Eastern Cape	0.077(+)

Migration Rate

	Natal
Eastern Cape	5.99

Database NEA4

Population Pairwise F_{ST}s (Significance)

	Natal	Eastern Cape Urban
Eastern Cape Urban	0.077(+)	
Eastern Cape Rural	0.086(+)	0.001(-)

Migration Rate

	Natal	Eastern Cape Urban
Eastern Cape Urban	5.98	
Eastern Cape Rural	5.32	1117.37